Ultrasonic Imaging
2021, Vol. 43(6) 308–319
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01617346211042526
journals.sagepub.com/home/uix
SAGE

# Simultaneous Segmentation of Fetal Hearts and Lungs for Medical Ultrasound Images via an Efficient Multi-scale Model Integrated With Attention Mechanism

**Jianing Xi[1]*, Jiangang Chen[2]*, Zhao Wang[3], Dean Ta[4], Bing Lu[5], Xuedong Deng[5], Xuelong Li[1], and Qinghua Huang[1]**

## Abstract

Large scale early scanning of fetuses via ultrasound imaging is widely used to alleviate the morbidity or mortality caused by congenital anomalies in fetal hearts and lungs. To reduce the intensive cost during manual recognition of organ regions, many automatic segmentation methods have been proposed. However, the existing methods still encounter multi-scale problem at a larger range of receptive fields of organs in images, resolution problem of segmentation mask, and interference problem of task-irrelevant features, obscuring the attainment of accurate segmentations. To achieve semantic segmentation with functions of (1) extracting multi-scale features from images, (2) compensating information of high resolution, and (3) eliminating the task-irrelevant features, we propose a multi-scale model with skip connection framework and attention mechanism integrated. The multi-scale feature extraction modules are incorporated with additive attention gate units for irrelevant feature elimination, through a U-Net framework with skip connections for information compensation. The performance of fetal heart and lung segmentation indicates the superiority of our method over the existing deep learning based approaches. Our method also shows competitive performance stability during the task of semantic segmentations, showing a promising contribution on ultrasound based prognosis of congenital anomaly in the early intervention, and alleviating the negative effects caused by congenital anomaly.

## Keywords

ultrasound image, fetal heart, fetal lung, image segmentation, data mining

## Introduction

Congenital anomaly is one of the most important reasons for infant mortality, and perinatal care or intervention can significantly reduce the infant mortality caused by congenital anomaly.[1] One of the most prevalent congenital defects of infant is congenital heart disease, of which the incidence rate ranges from 0.8% to 1.1% and is ranked at the top of the list of congenital defects.[2] Through the application of early scanning and subsequent prenatal intervention, the infant morbidity due to congenital heart disease can be largely alleviated after the birth of fetuses.[3] Meanwhile, a series of anomalies associated with fetal lung are also at risk of becoming irreversible chronic lung disease.[4] If these congenital anomalies of fetuses could be detected before their delivery, early intervention with long-term care for individuals would be used to prevent such type of risk.[5] Consequently, early scanning of fetuses is the key for prognosis of congenital anomaly, and provides the basis for selection of intervention to decrease the morbidity or mortality caused by congenital anomaly.[6]

Compared with Computed Tomography (CT) imaging and Magnetic Resonance Imaging (MRI), ultrasound imaging is noninvasive and nonradioactive, harmless to both pregnant women and fetuses during the scanning process.[7]

[1]School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China
[2]Shanghai Key Laboratory of Multidimensional Information Processing, School of Communication & Electronic Engineering, East China Normal University, Shanghai, China
[3]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[4]Department of Electronic Engineering, Fudan University, Shanghai, China
[5]Center for Medical Ultrasound, Nanjing Medical University Affiliated Suzhou Hospital, Suzhou, China

*These authors contributed equally to this work.

**Corresponding Author:**
Qinghua Huang, School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, P.O. Box 64, 127 West Youyi Road, Xi'an 710072, China.
Email: qhhuang@nwpu.edu.cn

Also, the high speediness and low cost also prompt ultrasound imaging to be the most common imaging method in fetal scanning.[8] After the ultrasound scanning, experienced physicians can prognose whether there is any congenital anomaly via the regions of fetal hearts and lungs in ultrasound images.[9] Nevertheless, it should be noted that ultrasound images contain speckle noises.[10] Furthermore, unlike CT and MRI, the collection of fetal scanning is non-standardized and irregular.[11] The existence of both speckle noise and non-standardization indeed enlarges the difficulty of recognizing the regions of fetal hearts and lungs from ultrasound images.[9] For large scale ultrasound scanning of fetal anomaly, manually recognizing the regions of fetal hearts and lungs requires a great number of costs of labor of experienced physicians and long period of their time.[12] In order to reduce the cost, it is a clear need to establish an accurate and automatic computer aided diagnosis method for segmenting the regions of fetal hearts and lungs from ultrasound images.[12]

For organ segmentation of medical ultrasound images, a noticeable number of methods has emerged during the last decade.[13] Regarding ultrasound images as monochromes, various types of ultrasound image segmentation methods have been applied in early approaches, such as thresholding,[14] clustering,[15] watershed transformation,[16] graph-based segmentation,[17] level set based active contour model,[18] and Markov random field.[19] Despite the success obtained by these aforementioned methods, the shortage of semantics of organs in segmented regions considerably limits their applications for segmentations of specific organs.[20] Recently, with the unprecedented growth of artificial intelligence in recent years, deep learning technique has been widely adopted in various fields, showing a great advantage of automatic extraction of features from images when compared with classical machine learning methods.[21] With the flourishing of deep learning technique, to achieve semantic segmentation of specific organs from medical images, a series of deep neural network based methods have been proposed.[22]

For deep learning of organ-specific semantic segmentations, fully connected network (FCN) is utilized as the baseline method, which can obtain a considerable performance in applications.[23] Based on FCN framework, U-Net is constructed in consideration of the encoding-decoding structure, where the encoding part is used for contracting the images into high-level features and the decoding part is for expanding the features into pixels of segmentation masks.[24] To compensate the textural information of high resolution into the high-level features with relatively low resolution in expansive network, skip connections are also introduced between the contracting and expansive networks.[24] To accelerate the processing speed during segmentation, E-Net introduced convolutions in bottleneck module, including both full convolution and dilated convolution.[25] Furthermore, for alleviating gradient vanishing problem, DenseNet connected each pair of layers via a feed-forward fashion, strengthening the propagation between different layers as feature reuse.[26] Generally, in semantic segmentation, there are noteworthy successes achieved by deep learning base approaches.[22]

It should be noted that, different organs in fetal ultrasound images demonstrate various scales.[27] For example, the scales of textures of different organs vary distinctly, and the scales of region sizes of organs also differ from each other.[27] Moreover, when we observe the ultrasound image of a specific organ, we can also note that there are recognizable distinctions of shapes, textures, and structures among different scales.[28] Nevertheless, the aforementioned deep learning based methods lack of the module of multi-scale feature extraction at a larger range of receptive fields. Accordingly, PSP-Net further aggregates contexts of different regions through a pyramid scene parsing network to extract global context information, which can ensemble multi-scale feature of images.[29] However, in comparison of U-Net and DenseNet, the absence of skip connections in PSP-Net leads to the incapability of compensation of information of high resolution in features for segmentation. In addition to the deficiencies above, there is also another issue that the features extracted by deep learning methods contain not only the features of shapes and textures of the organs to be segment, but also the features of task-irrelevant organs or background of the ultrasound images, which dilutes the power of segmentation task.[30] Consequently, there is a clear need for establishing an efficient semantic segmentation method with the functions of (1) extracting multi-scale features from images, (2) compensating information of high resolution, and (3) eliminating the task-irrelevant features.

To achieve the feature extraction of multi-scales at a larger range of receptive fields, the compensation of high resolution information, and the elimination of irrelevant features in fetal heart and lung segmentation from ultrasound images, in this paper we propose a deep learning based multi-scale model with skip connection framework and feature attention integrated. To achieve the integration of both multi-scale feature extractions and high resolution information compensation, we exploit the U-Net framework with skip connections between contracting and expansive networks, and introduce the scheme of network in network.[31] by inserting multi-scale feature extractor modules into the framework.[32] In order to eliminate the task-irrelevant features for segmentation, we also incorporate attention mechanism into our integrated framework via assigning higher weights on segmentation relevant features.[33] When we evaluate the performance of our proposed multi-scale model integrated with attention mechanism, we observe a clear advantage of our method over the existing methods in the fetal heart and lung segmentation of ultrasound images. Moreover, our method also shows competitive performance stability in semantic segmentation task. In summary, our proposed method demonstrates a remarkable capability in semantic segmentations of fetal hearts and lungs from ultrasound images, showing a promising potential for applications of early scanning of congenital anomaly.

## Methods

### Fetal Ultrasound Image Collections

The fetal ultrasound image data used in this study are collected from Center for Medical Ultrasound, Nanjing Medical University Affiliated Suzhou Hospital, Suzhou, China. These ultrasound images are fetal ultrasound scanning from pregnant women at 37 weeks of gestation, for the observation of heart and lungs of the fetuses. Ethical approval was obtained from the Ethics Committee of Nanjing Medical University Affiliated Suzhou Hospital (approval No. K2016038) and informed consent requirements were waived. Since the collection procedure may suffer unavoidable disruption that leads to the low quality problems of the related images, we also eliminate the low quality images from the collected data for quality control. For example, in a fraction of images, the gray scales of organs between fetus and pregnant woman are undistinguished, or the whole picture are purely black in a few of them.

Specifically, All the images were acquired in the same hospital, and they were collected from an ultrasound equipment WS80A with Elite (Samsung Medison, Seoul, Korea) equipped with a curved array ultrasound probe (CA1-7A), where all images were acquired with the same scanner, by the same operator, and with the same acquisition settings. There are more than 350 pregnant women involved in the image collections, and 312 images from these women are selected as qualified images of fetal hearts and lungs. After quality control, we finally obtain a dataset containing totally 312 qualified images of fetal hearts and lungs.

### Data Preprocessing and Augmentation

The original format of the collected images is Digital Imaging and Communications in Medicine (DICOM), which is the most widely used standard of medical images. Considering that DICOM does not fit the input format for segmentation model, we firstly use a python package called PyDicom to convert the DICOM files into bitmap (BMP) files. Next, through an open source annotation tool called LabelMe, the contours of regions of fetal hearts and lungs are manually delineated by a professional sonographer physician with $\geq 10$ years of experience, which are then used as the ground truth masks of the two fetal organs. Then, we also cut the raw images with size of $1280 \times 872$ into patches with size of $384 \times 384$, where the selections of $384 \times 384$ regions are annotated by a sonographer physician, and the regions of the patches are regarded as Region Of Interest (details in Supplemental Information). In this size, the view is just almost filled by the fetal hearts or lungs, leading to the preservation of computational cost during the segmentation task. Finally, due to the requirement of data amount of deep learning framework, the images utilized as training data of our method are further augmented through rotating and flipping,[34] resulting in an augmented dataset whose amount is eight times greater than that of the collected dataset.

### The Proposed Integrated Segmentation Model

*Segmentation framework of U-Net.* In image segmentation task, the aimed output is a mask containing subsets of pixels with locations, and each set is assigned to a specific type to be predicted. With the unprecedented opportunity offered by the development of deep learning technique, convolutional neural network based framework has become the de facto standard for most image analysis task including images segmentation of course.[23] Specifically, fully convolutional network (FCN) is constructed by convolutional layers progressively, which can separate pixels with different semantics through high dimensional image representations of local information that are extracted layer by layer.[23] Although the sequential process of FCN yields noticeable achievements in image segmentation, the power of the basic version of FCN is still limited by parameter efficiency.[24] For example, while low-level features in former layers tend to preserve more textural information of high resolution, the detailed information vanishes in a certain extent in high-level features in later layers,[35] diluting the resolution of the output segmentation masks.

To compensate the information of high resolution in the later layers in FCN, a more elegant segmentation framework called U-Net is proposed.[24] The U-Net framework consists of a contracting path for high-level features, and an expansive path for low-level features.[24] In the contracting path, the convolutions followed by a Rectified Linear Unit (ReLU) and a max pooling operation for down-sampling are repeatedly adopted, as a typical architecture of a convolutional network. On the contrary, in the expansive path, the max pooling operation for down-sampling is replaced by an up-convolution for up-sampling. At each step, the number of feature channels is doubled in the contracting path, and is halved in the expansive path. Notably, to compensate the vanished information of details, when the high-level features closer to the outputs are expanded into low-level features by up-sampling, information of high resolution in former layers are passed through the skip connections. The features from both former and later layers are concatenated to accomplish a segmentation prediction of higher resolution.

To take the advantage of compensation of information of high resolution, in our proposed method, we also use the framework of contracting and expansive paths with skip connections as U-Net.[24] Inspired by the structure of U-Net, in our proposed method we incorporate the contracting path to extract the high-level features related to different semantics, and utilize the expansive path to predict the pixels sharing certain semantics from the high-level features. At the same time, we also employ skip connections in our proposed

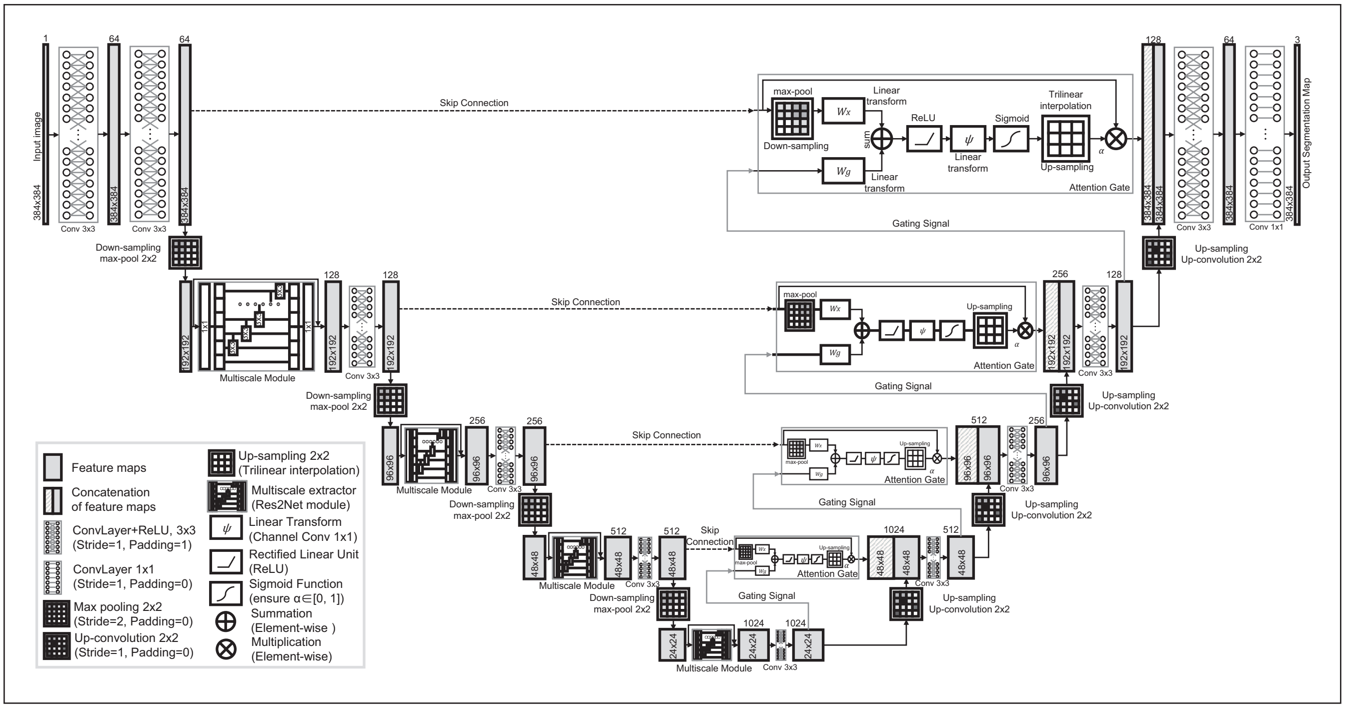

**Figure 1.** The schematic plot of our proposed multi-scale model integrated with attention mechanism. The model is built based on the architecture of U-Net, where the left parts are contracting networks for feature extraction, and the right parts are expansive networks for segmentation mask prediction. In contracting networks, the Res2Net modules are inserted to extract multi-scale features. In expansive networks, the additive attention gates are utilized to eliminate task-irrelevant features for mask prediction. A series of skip connections are also established from contracting networks to expansive networks, compensating high-resolution information for mask prediction. The figure legend is at the left-bottom corner.

method (Figure 1). By concatenating the low-level features through the skip connections, we can compensate the information of high resolution from the contracting paths to the expansive paths. In addition to compensation of information of high resolution, our proposed method also involves two abilities that are not leveraged in U-Net, that is, feature extraction of patterns occurring at multi-scales at a larger range of receptive fields, and suppression of the task-irrelevant features of images. The details of technical implementations of the two additional abilities in our methods are depicted in the following subsections.

*Multi-scale module of Res2Net.* In order to enhance the multi-scale feature extraction ability of our proposed method at a larger range of receptive fields, we further implement the state-of-the-art multi-scale module of Res2Net.[32] into the U-Net framework. Through hierarchical residual-like connections within a single residual block, Res2Net module can efficiently extract multi-scale representation orthogonal to the classical layer-wise feature aggregation models. In Res2Net module, the input feature map $\boldsymbol{X}$ are evenly separated into $s$ subsets of feature maps, where the spatial size of these feature subsets are the same with the input feature maps, but the number of channels is $1/s$ compared with that of the input feature maps.[32] Here we denote the $i$-th feature subset as $\boldsymbol{x}_i$, $i = 1, 2, \ldots, s$, where $\boldsymbol{X} = \boldsymbol{x}_1 \oplus \boldsymbol{x}_2 \oplus \ldots \boldsymbol{x}_i \ldots \oplus \boldsymbol{x}_s$ (here $\oplus$ represents for concatenation). Next, these input features of separated subsets are processed through a $1 \times 1$ convolution for dimension reduction, which is equivalent to a linear projection, that is, $\tilde{\boldsymbol{x}}_i = \boldsymbol{W}_{\text{in}} \boldsymbol{x}_i$, where $\boldsymbol{W}_{\text{in}}$ is parameter matrix of the linear projection for dimension reduction.

For any $i$ from 1 to $s$, for $\tilde{\boldsymbol{x}}_i$ from the $s$ subsets in the second layer, a corresponding feature set $\tilde{\boldsymbol{z}}_i$ is also established in the second to last layer of the module, which is used for restoring the features after the multi-scale extraction from $\tilde{\boldsymbol{x}}_i$. In particular, for preserving the scale of the input features maps, the $\tilde{\boldsymbol{z}}_1$ is set to be the same as $\tilde{\boldsymbol{x}}_1$ through a direct link between the second layer and the second to last layer. For $\tilde{\boldsymbol{x}}_i$ where $i \geq 2$, there are totally $s-1$ of $3 \times 3$ convolutional kernels introduced for each subset of features $\tilde{\boldsymbol{x}}_i$, denoted as $\mathbf{K}_i(\cdot)$. When $2 < i \leq s$, the inputs of convolutional kernel $\mathbf{K}_i(\cdot)$ is the addition of the subset of feature maps $\tilde{\boldsymbol{x}}_i$ and the outputs of the former kernel $\mathbf{K}_{i-1}(\cdot)$. Consequently, the feature subset $\tilde{\boldsymbol{z}}_i$ in the second to last layer of the module can be calculated as:

$$\tilde{\boldsymbol{z}}_i = \begin{cases} \tilde{\boldsymbol{x}}_i, & i = 1; \\ \mathbf{K}_i(\tilde{\boldsymbol{x}}_i), & i = 2; \\ \mathbf{K}_i(\tilde{\boldsymbol{x}}_i + \tilde{\boldsymbol{z}}_{i-1}), & 2 < i \leq s. \end{cases} \quad (1)$$

According to the formula above, we can notice that, for the $i$-th convolutional kernel $\mathbf{K}_i(\cdot)$, the information of feature subsets from $\tilde{\boldsymbol{x}}_1$ to $\tilde{\boldsymbol{x}}_i$ are all fed into this kernel. This design ensures a larger receptive field of the outputs of convolution kernel $\mathbf{K}_i(\cdot)$ than that of the feature subset $\tilde{\boldsymbol{x}}_i$. For the feature subset in the second to last layer $\tilde{z}_i$, the larger the index $i$ becomes, the more information at different scales the $\tilde{z}_i$ contains. Consequently, the concatenation of all the $s$ subset of feature maps $\tilde{z}_i$ $(1 \leq i \leq s)$ in the second to last layer, incorporating various numbers of different combinations of receptive field scales, thanks to the combinatorial explosion effect. Finally, the feature maps involving the information of different scales are fed into the last layer of $1 \times 1$ convolution for dimension recovery, achieved by linear projection $z_i = \boldsymbol{W}_{\text{out}} \tilde{z}_i$, where $\boldsymbol{W}_{\text{out}}$ is parameter matrix of the linear projection for dimension recovery. Their concatenation $\boldsymbol{Z} = z_1 \oplus z_2 \oplus \ldots \oplus z_i \oplus \ldots \oplus z_s$ is the outputs of the internal layers of Res2Net module. The procedures of internal layers of the module can be summarized as a function $\mathcal{F}_{\text{module}}$ with input features $\{\boldsymbol{x}_i\}$, kernels $\{\mathbf{K}_i\}$, and parameter matrices $\boldsymbol{W}_{\text{in}}$ and $\boldsymbol{W}_{\text{out}}$:

$$\begin{aligned}\boldsymbol{Z} = \mathcal{F}_{\text{module}}\left(\{\boldsymbol{x}_i\},\{\mathbf{K}_i\},\boldsymbol{W}_{\text{in}},\boldsymbol{W}_{\text{out}}\right) &= \boldsymbol{W}_{\text{out}}[(\boldsymbol{W}_{\text{in}}\boldsymbol{x}_1) \\ &\oplus \mathbf{K}_2(\boldsymbol{W}_{\text{in}}\boldsymbol{x}_2) \oplus \mathbf{K}_3(\boldsymbol{W}_{\text{in}}\boldsymbol{x}_3 + \mathbf{K}_2(\boldsymbol{W}_{\text{in}}\boldsymbol{x}_2)) \oplus \ldots \\ &\oplus \mathbf{K}_s(\boldsymbol{W}_{\text{in}}\boldsymbol{x}_s + \mathbf{K}_{s-1}(\boldsymbol{W}_{\text{in}}\boldsymbol{x}_{s-1} + \ldots))].\end{aligned} \tag{2}$$

In addition to the internal layers, the Res2Net also has an external composition of skip connection between the inputs and the outputs of the module.[32] Regarding the output feature maps of the module as the element-wise addition of the input feature maps and residuals, this skip connection is used to compel the outputs of the internal layers $\boldsymbol{Z}$ to be the residuals between the inputs $\boldsymbol{X}$ and the final outputs $\boldsymbol{Y}$, that is, $\boldsymbol{Z} = \boldsymbol{Y} - \boldsymbol{X}$. Since the residual maps is easier to be optimized than the unreferenced feature maps in most situations, introducing the skip connection is conducive for addressing gradient vanishing and exploding problems. Considering that the dimensions of inputs and outputs are the same, we directly use the identity skip connection without any extra parameter burden. Finally, the outputs of the internal layers are added to the inputs identically copied from the skip connection, as the final outputs $\boldsymbol{Y}$ in the end of the Res2Net module[32]:

$$\boldsymbol{Y} = \mathcal{F}_{\text{module}}\left(\{\boldsymbol{x}_i\},\{\mathbf{K}_i\},\boldsymbol{W}_{\text{in}},\boldsymbol{W}_{\text{out}}\right) + \boldsymbol{X}. \tag{3}$$

To empower the multi-scale feature extraction capability of our proposed method, we insert the Res2Net modules into the skip connection framework of U-Net as Network in Network (NIN).[31] Considering that the contracting paths are responsible for feature extraction in U-Net framework,[32] we replace the traditional convolution of the input layers of each contracting path with the multi-scale Res2Net modules, which is conducive for the capability of extracting multi-scale information from the input feature maps. Finally, we obtain the contracting paths of extracting high-level feature at multi-scales, as shown in the left part of Figure 1.

*Feature highlight of attention mechanism.* It should be noted that during the procedure of multi-scale feature extraction, the contracting paths are likely to capture both the task-specific features of the regions of interested organs, and the irrelevant features of the regions of unconcerned organs or backgrounds as well.[33] Therefore, in addition to the extraction of multi-scale features, it is also warranted to eliminate the irrelevant features of the fetal ultrasound image segmentation task. In favor of focusing on the task-specific target regions of feature maps, the attention mechanism is further introduced into our proposed method, which has also been commonly employed in various deep learning applications.[36] Here we incorporate a feature map grid based gating attention via additive attention gate (AAG) model into our proposed method, which is used for suppressing the activations of irrelevant features and highlighting the task-specific features.[37] Through the attention coefficients from high-level features with semantics information, attention mechanism can reweight the contributions of features for segmentation task.

In the feature map grid based gating attention, the input feature maps are firstly partitioned as a series of grids, and each grid covers a certain region of the input feature map. For the $j$-th grid, the features related to this region are down-sampled by max pooling of a $2 \times 2$ convolution with stride of 2. All the elements of features of $j$-th grid after down-sampling are then vectorized as a feature vector $\boldsymbol{v}_j$, which is used as the inputs of AAG model for attention. AAG is constructed based on probabilistic scoring, known as soft attention, leading to the facilitation of employing standard back-propagation without extra sampling. The additive attention[38] in AAG is a simple but efficient way to obtain the attention coefficient $\alpha_j$ from the grid feature vector $\boldsymbol{v}_j$ and the gating signal $\boldsymbol{g}_j$:

$$\alpha_j = \sigma(\phi^{\mathrm{T}}(\mathrm{ReLU}(\boldsymbol{W}_v^{\mathrm{T}}\boldsymbol{v}_j + \boldsymbol{W}_g^{\mathrm{T}}\boldsymbol{g}_j + \boldsymbol{b}_g)) + \boldsymbol{b}_\phi), \tag{4}$$

where the gating signal $\boldsymbol{g}_j$ is also a vectorization of features that are collected from the maps of the higher level.[33] Here we denote the dimensions of $\boldsymbol{v}_j$ and $\boldsymbol{g}_j$ as $d_v$ and $d_g$ respectively. Also, $\boldsymbol{W}_v$ and $\boldsymbol{W}_g$ are $d_v \times d_t$ and $d_g \times d_t$ matrices of linear transformations respectively, and the bias term of the transformations is a $d_t \times 1$ vector, where $d_t$ is the dimension of the intermediate variable. The intermediate variable is then activated through a ReLU unit, and is further converted by a linear transformation with vector $\phi$ and bias $\boldsymbol{b}_\phi$ to calculate the attention coefficient of the $j$-th grid.

For image segmentation, in order to highlight the task-specific features and to eliminate the task-irrelevant features, we also regard the AAG unit as the concept of NIN,[31] and insert this module into the framework of our proposed method. In contrast to contracting paths, the expansive paths in U-Net framework are dominant for generating the predicted segmentation mask, which are more sensitive to the disturbance of task-irrelevant features.[24] Consequently, we insert AAG modules into different levels of the expansive paths with inputs from both the former layers and the skip connections, with the gating signal from the former layer, as shown in the right part of Figure 1. Since the inputs of AAG are the low-level features containing higher resolution at multi-scales from skip connections, the AAG modules in the expansive paths are promotive for focusing on the task-specific features of segmentations from the viewpoints of both resolutions and scales (Figure 1). Finally, the outputs of AAG models are the concatenated with the high-level features involved with semantics information,[33] which are prone to increase the predicted resolution of semantic segmentation.

*Multi-scale model integrated with attention mechanism.* To integrate the advantages of high resolutions, multi-scales, and attention mechanism, we propose the multi-scale model integrated with attention mechanism based on the U-Net framework. Specifically, in our proposed method, we set totally five levels for both the contracting paths and the expansive paths. In the contracting paths, the first level is constructed by two cascaded layers of $3\times3$ convolutions with stride 1 and padding 1 and followed by ReLU activations. Subsequently, the second to the fifth levels are constructed by the same structure consisting of a multi-scale Res2Net module followed by a $3\times3$ convolution and ReLU activation. After the feature extraction in the first to the fourth levels, the output feature maps are down-sampled by a $2\times2$ max pooling operation with stride 2 and are then fed into the next level. In the five levels of contracting paths, the sizes of feature maps are gradually contracted as 384, 192, 96, 48, and 24, while the channels in the five levels are 64, 128, 256, 512, and 1024, respectively.

At the same time, in the expansive paths of our proposed method, the fourth to the first levels are constructed by the same structure consisting of three parts: an AAG unit, a concatenation operator, and a traditional convolution layer. Here the AAG unit is used to reweight the feature maps from skip connections of contracting paths at the same level, where the feature maps of former level are regarded as the gating signal for attention. After the feature maps with attentions are obtained from AAG unit at the current level, these maps are further concatenated with the up-sampled feature maps from the former level.[33] Here the up-sampling process is achieved by a $2\times2$ up-convolution with stride 1 as suggested in the traditional U-Net.[24] Through the concatenation operation, the channels of features from both the high resolution maps at the current level and the semantic maps at the former level are doubled. The doubled features are fed into a layer of $3\times3$ convolution with stride 1 followed by ReLU activation. Specially, at the end of the first level in expansive paths, the last layer of our proposed model is a $1\times1$ convolution to project the features in all the channels to the scores of the three classes (fetal heart region, fetal lung region, and background region respectively) of pixels in output segmentation map (3-value map).

Based on the aforementioned architecture of our proposed method, we utilize the cross entropy between the pixels of the ground truth masks and of the predicted segmentation map,[24] as the loss function to be minimized at model training. We also introduce the L2-norm regularization term on the parameters, where the value of the corresponding tuning parameter is set to 0.001 empirically. By strictly following the previous study of original U-Net, we choose 5 as the number of layers of different levels for the U-Net skip connections.[24] According to previously published researches, the choice of layer number is important for the network, since a small layer number might lead to the undermining of high-level features, while a larger layer number is very likely to cause over-compression of high-level features and the degeneration of the model.[39]

The parameters of our proposed method are trained with stochastic gradient descent optimizer, where the gradient information is calculated by back-propagation. The initial value of learning rate during the training process is set to be 0.001, and the learning rate is decayed by multiplicative factor of 0.1 at epochs of multistep. The batch sizes during training is set to 8, and the maximum number of iterations is set to 400. Our proposed method is implemented through the open source machine learning framework PyTorch, which is established in Python environment. In summary, our method integrates the skip connections from U-Net framework for preserving the high resolution information, the Res2Net module for extracting features at multi-scale, and the AAG unit for highlighting the task-specific features of segmentations, showing a strong potential for organ segmentations of fetal ultrasound images.

## Evaluation Metrics

To quantitatively evaluate the segmentation performance of fetal hearts and lungs from ultrasound images, we employ two widely used measurements to calculate the similarity between ground truth masks and segmentation results. For the sake of convenience, we denote that $\mathbf{S}_{\text{pred}}$ is the set of pixels predicted as the regions of the target organs, while $\mathbf{S}_{\text{truth}}$ is the set of pixels of the ground truth masks. The first measurement is Dice coefficient, defined as the faction of two times the area of overlap between the prediction regions and the ground truth regions in the sum of the areas of the two regions, of which the formula is given as below,

$$\text{Dice} = 2\,|\,\mathbf{S}_{\text{truth}} \cap \mathbf{S}_{\text{pred}}\,|\,/(|\,\mathbf{S}_{\text{truth}}\,| + |\,\mathbf{S}_{\text{pred}}\,|). \tag{5}$$

The second measurement is Intersection over Union (IoU), defined as the fraction of the area of overlap between the prediction regions and the ground truth regions in the area of union between the two regions, which is calculated as

$$\mathrm{IoU} = (\mathbf{S}_{\mathrm{truth}} \cap \mathbf{S}_{\mathrm{pred}}) / (\mathbf{S}_{\mathrm{truth}} \cup \mathbf{S}_{\mathrm{pred}}). \tag{6}$$

According to the formulas of the two measurements, we can observe that if there is no intersection between the set of predicted mask $\mathbf{S}_{\mathrm{pred}}$ and the set of ground truth $\mathbf{S}_{\mathrm{truth}}$, then both the Dice coefficients and the IoU equal to zero. On the contrary, if the two sets are identical, then both the two measurements equal one. In intermediate situations between the two extreme situations above, the values of both the two measurements are range from zero to one.

### McNemar's Exact Test

To further evaluate the significance of the differences between the results of the proposed method and other competing methods, we have also adopted McNemar's exact test. McNemar's exact test is a well-known statistical test to analyze statistical significance of the differences in classifier performances. McNemar's exact test is a non-parametric test, and is applicable with any sample size. Here we use McNemar's exact test to test whether there is a significant difference between the results of two methods through the contingency table of (1) foreground overlap for both methods, (2) background overlap for both methods, (3) the overlap of foreground for method 1 and background for method 2, and (4) the overlap of foreground for method 2 and background for method 1. In the McNemar's exact test for both the proposed method and the competing methods, we compare their segmentations regions of predicted masks, rather than their Dice coefficients or IoUs. Therefore, a low $p$-value from McNemar's exact test indicates a significant difference in their segmentation regions, but not their performance measurements.

### Implementation

Our proposed method is implemented on GPU of NVIDIA TITAN V 24G in a high performance server. The code of our method is implemented in Python 3.6.9 using deep learning framework of PyTorch 1.0.1, in the operating system environment of Ubuntu 16.04 LTS. The dataset used in the implementation of our method is 312 fetal ultrasound images as mentioned in section 2.1, along with their corresponding masks of fetal hearts and lungs manually delineated by a professional sonographer physician. Here the sizes of both the fetal ultrasound images and their corresponding masks of labels are 384 by 384 pixels. For each mask, the regions of fetal hearts and fetal lungs are labeled as red and green pixels respectively, while the regions of background are labeled as black pixels. For evaluation of our method, the fetal ultrasound dataset is evenly split into eight subsets, and each subset is used as testing data alternately and the seven remaining subsets are employed as the training data, known as eight fold cross validation.

### Experiment Setting

To evaluate the advantage of our proposed method, we assess the performances of our method in contrast to those of the previously published convolutional segmentation approaches. The comparison approaches include standard FCN,[23] standard U-Net,[24] E-Net,[25] PSP-Net,[29] and DenseNet.[26] The FCNs used in the comparison study are the popular FCN-8s, FCN-16s, and FCN-32s, defined in previous studies.[23] Meanwhile, we also adopt ablation study to evaluate whether and how much the Res2Net modules contribute to the performance of our method. In the experiments of ablation study, we remove the Res2Net modules from our method and compare its results with those of the full version. In all the experiments, the learning rate of these comparison methods are also set to be decayed at epochs of multistep as that of our method, where the multiplicative factor is set to 0.1 and the initial value is 0.001. Furthermore, the maximum numbers of iterations of these comparison methods are also set to 400, and the batch sizes during training is set to 8. Similar to our proposed method, these comparison methods are also implemented on GPU of NVIDIA TITAN V 24G in the same high performance server with the operating system environment of Ubuntu 16.04 LTS. For the training data and testing data of the comparison approaches, we also employ the evenly separated subsets of the fetal ultrasound dataset that are the same as those of our proposed method.

## Results

### Comparison Performance

When we assess of the predicted masks of these comparison approaches and our proposed method, we can observe that for the measurements of both Dice coefficient and IoU, our method yields the biggest overlaps of the ground truth regions of both fetal hearts and fetal lungs (Figure 2). Taking fetal heart segmentation as an example, for all the folds of testing data in cross validation, the average Dice coefficients achieved by FCN-8s, FCN-16s, and FCN-32s are 0.892, 0.890, and 0.887, respectively. As for U-Net that is revised based on FCN with skip connections, its corresponding average Dice coefficient is 0.891, showing a slight improvement over those of FCNs. For the results of the state-of-the-art methods E-Net and PSP-Net, their related average Dice coefficients are 0.895 and 0.896, respectively. In comparison, comparable with the results of DenseNet, our proposed method obtains an average Dice coefficient of 0.902, demonstrating better performance than those of the other methods (Figure 2(A)). For the measurement of IoU for fetal heart segmentation, our method yields an average IoU of 0.822 (Figure 2(C)), which is also comparable or higher than those of the comparison methods.

At the same time, the evaluation results of fetal lung segmentation also display similar phenomenon that our proposed

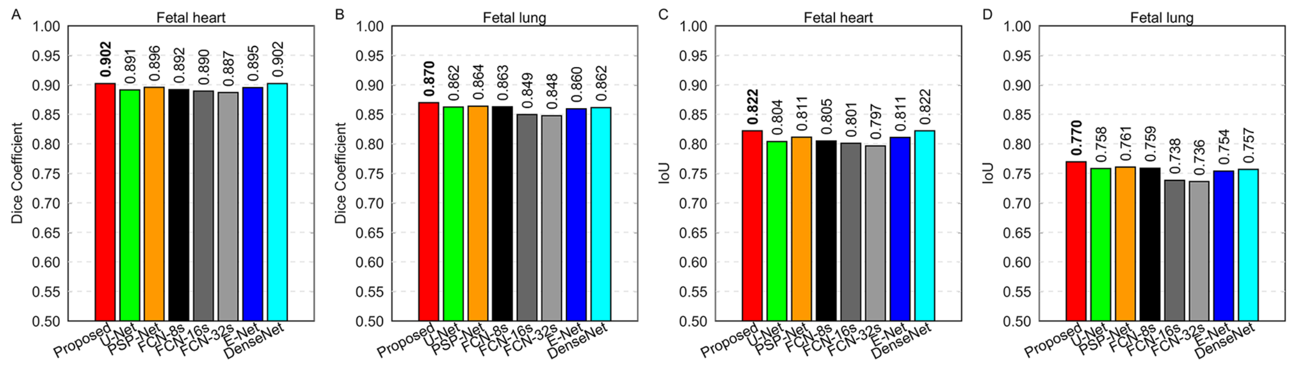

**Figure 2.** The semantic segmentation performances of our proposed method against the competing methods as bar plot. (A) Dice coefficients for fetal heart. (B) Dice coefficients for fetal lung. (C) IoU for fetal heart. (D) IoU for fetal lung.

method outperforms the other comparison methods. For example, among the all subsets of cross validation, the average Dice coefficient of FCN-8s, FCN-16s, and FCN-32s are 0.863, 0.849, and 0.848, respectively. Meanwhile, U-Net, E-Net, and PSP-Net achieve comparable results of average Dice coefficients, ranging from 0.860 to 0.864. DenseNet yields an average Dice coefficient of 0.862. In comparison, the average Dice coefficient of the results of our proposed method is 0.870 (Figure 2(B)), which is the highest among those of the evaluated methods. As for IoU of fetal lung segmentations, the values of segmentation results of the comparison methods vary from 0.736 to 0.761, and the value of IoU of results of our method achieve 0.770 (Figure 2(D)), indicating a superior performance than those of all the investigated methods. Overall, we can observe the advantages of our method on the segmentation for both fetal hearts and lungs from fetal ultrasound images.

Furthermore, we also adopt McNemar's exact tests on the results of both our method and the competing methods to validate whether there are significant difference between these methods. Through the McNemar's exact tests, the *p*-values demonstrate that, although the values of Dice coefficients and IoUs of these methods are close to each other, the differences between the results of the proposed method and the other competing methods are significant (details in Supplemental Table S1 and Supplemental Information). Specifically, although the Dice coefficients and IoUs for the fetal heart achieved by the proposed method and DenseNet approximately equal to each other, since the McNemar's exact test is not relate to the Dice coefficients and IoUs, the *p*-value from McNemar's exact test between the two methods is $<1.00\times10^{-15}$. This result indicates that the segmentation regions of their predicted masks are distinct from each other, even though the two methods yield similar performances.

To intuitively demonstrate the prediction masks of our method on fetal hearts and lungs, we further display an example of the predicted segmentation masks of these investigated methods along with the ground truth mask. As shown in Figure 3 (details in Supplemental Figures S6–S14 and Supplemental Information), we can find that for all the evaluated methods, most regions of the ground truth masks are covered by their predicted segmentation masks, for both fetal hearts and lungs. Specifically, we can observe that the predicted masks of FCNs-8s, FCN-16s, and FCN-32s show similar regions, where the predicted regions of fetal hearts are consistently left biased in comparison to the ground truth regions. For the regions of fetal lungs predicted by U-Net and PSP-Net, we can notice the discontinuity at the contours of the regions, and parts of the regions are even isolated from the rest parts. Despite the high values of numerical evaluation measurements achieved by DenseNet, in some cases, the predicted regions of the two organs are diffused into the regions of each other (here specific to the representative examples in Figure 3). Moreover, we can note that E-Net yields considerable recovery of both fetal hearts and lungs, but the gap between the predicted regions of two organs is slightly larger than that in the ground truth regions. In comparison, although the predicted regions of fetal lung is slight bulging, the regions obtained by our proposed method show satisfactory coincidence and continuity of contours, and the width of the gap between the two predicted regions also fits the ground truth regions.

## *Ablation Study*

To illustrate to which extent the Res2Net module contributes the segmentation performance, we have also provided the results without Res2Net, but still including the attention gates in U-net. Specifically, we conduct an ablation study to analyze the performances of our proposed method, our proposed method without Res2Net, and original U-Net (Supplemental Figure S3–S5). As demonstrated in Supplemental Figure S3(A), our proposed method without

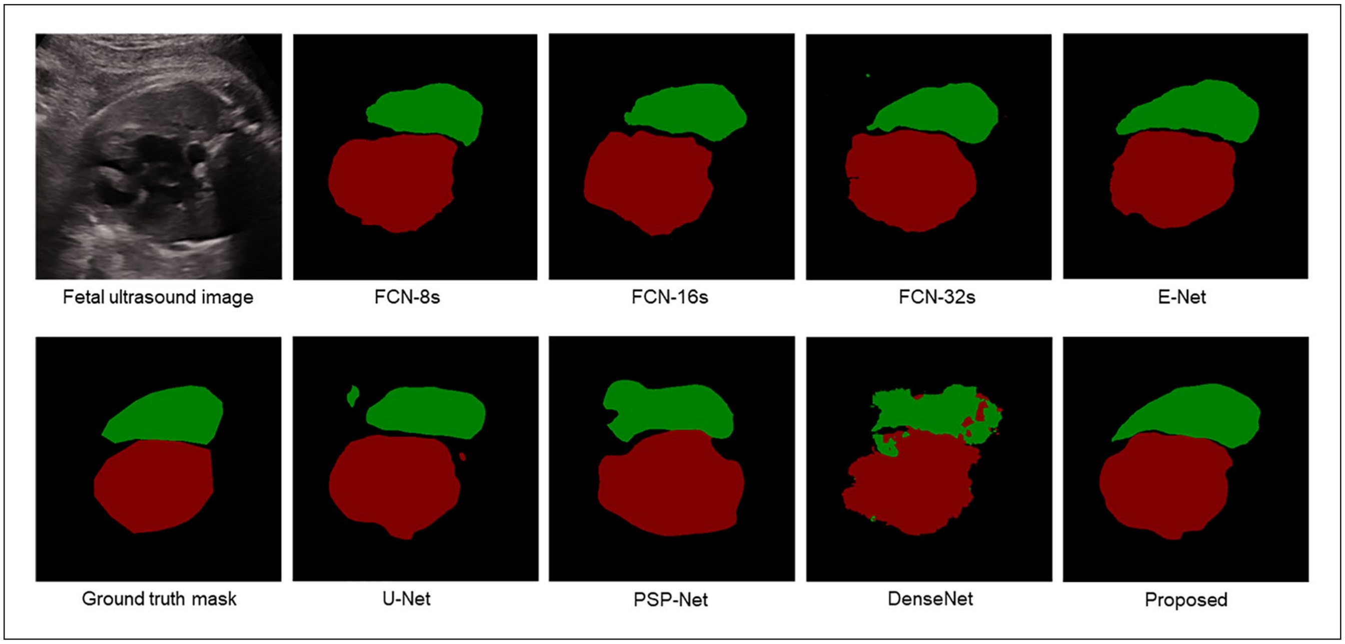

**Figure 3.** An intuitive illustration of predicted segmentation masks of our proposed method and the competing methods. Specifically, the left-top panel is the input ultrasound image of fetus, and the left-bottom panel is the ground truth mask.

Res2Net yields a Dice coefficient of 0.896 for fetal heart segmentation, contributing 45.5% of the performance increment from original U-Net to our method with Res2Net. This observation indicates that removing Res2Net modules might cause 54.5% performance loss, and thus this fraction can also reflect to which the extent Res2Net contributes to the performance. From Supplemental Figure S3(B), we can observe that Res2Net contributes half of the performance increment of for fetal lung segmentation. Similar phenomena for IoU of the two organs are also illustrated in Supplemental Figure S3(C) and (D) that the performance increments of our method include the contribution of Res2Net modules.

## Iterative Performance

To evaluate the iterative performance of our proposed method, we also demonstrate the performance among the iteration steps. After the first 100 iterations where the performances of all competing methods increase intensively, the performances of all the investigated methods vary in narrow ranges (Figure 4(A) and (C), Supplemental Figure S2 for scale zoomed in on the $y$-axis). Taking fetal heart segmentation as an example (Figure 4(A)), for FCN-8s, FCN-16s, and FCN-32s, their Dice coefficients oscillate in the range of 0.817 to 0.893 when the iteration steps are from 100 to 400. Meanwhile, the Dice coefficients yielded by U-Net, PSP-Net, E-Net, and DenseNet at steps of 100 to 400, fluctuate in the ranges of 0.824 to 0.894, 0.826 to 0.901, 0.892 to 0.904, and 0.830 to 0.907, respectively. In comparison, for the results in iteration of 100 to 400, the Dice coefficients of fetal heart segmentation obtained by our proposed method vary in the range of 0.878 to 0.904, which is comparable or smaller than those of the other competing methods, indicating the performance stability of our proposed method. Indeed, in the iteration performance analysis, many other methods also show stable performance when the number of iterations grows. Therefore, the result can only prove that, the range that the performance varies of our method is comparable or narrower than those of some comparison methods, and at the same time shows comparable stability with those of the other comparison methods.

As for the iterative performances for fetal lung segmentation, we can also observe that our method demonstrates a better performance stability than those of the other investigated methods (Figure 4(C)). Specifically, in iteration step in 100 to 400, the Dice coefficients of fetal lung for FCN-8s, FCN-16s, and FCN-32s fluctuate in the range of 0.784 to 0.872. At the same time, the value of Dice coefficients obtained by U-Net vibrate from 0.778 to 0.863 in iteration of 100 to 400. Furthermore, PSP-Net, E-Net, and DenseNet achieve Dice coefficients ranged from 0.791 to 0.872, 0.852 to 0.872, and 0.790 to 0.868, respectively. In contrast, in iteration from 100 to 400, the Dice coefficients of our proposed method for fetal lung vary in the range of 0.840 to 0.873. The widths of the intervals between the minimum and maximum values of Dice coefficients obtained by these comparison methods, indicate a similar phenomenon that when compared with the others, our proposed method shows an advantage of performance stability.

To further demonstrate the superiority of our method during the iterations, we also illustrate the occupation fractions of all the competing methods as best performance at each

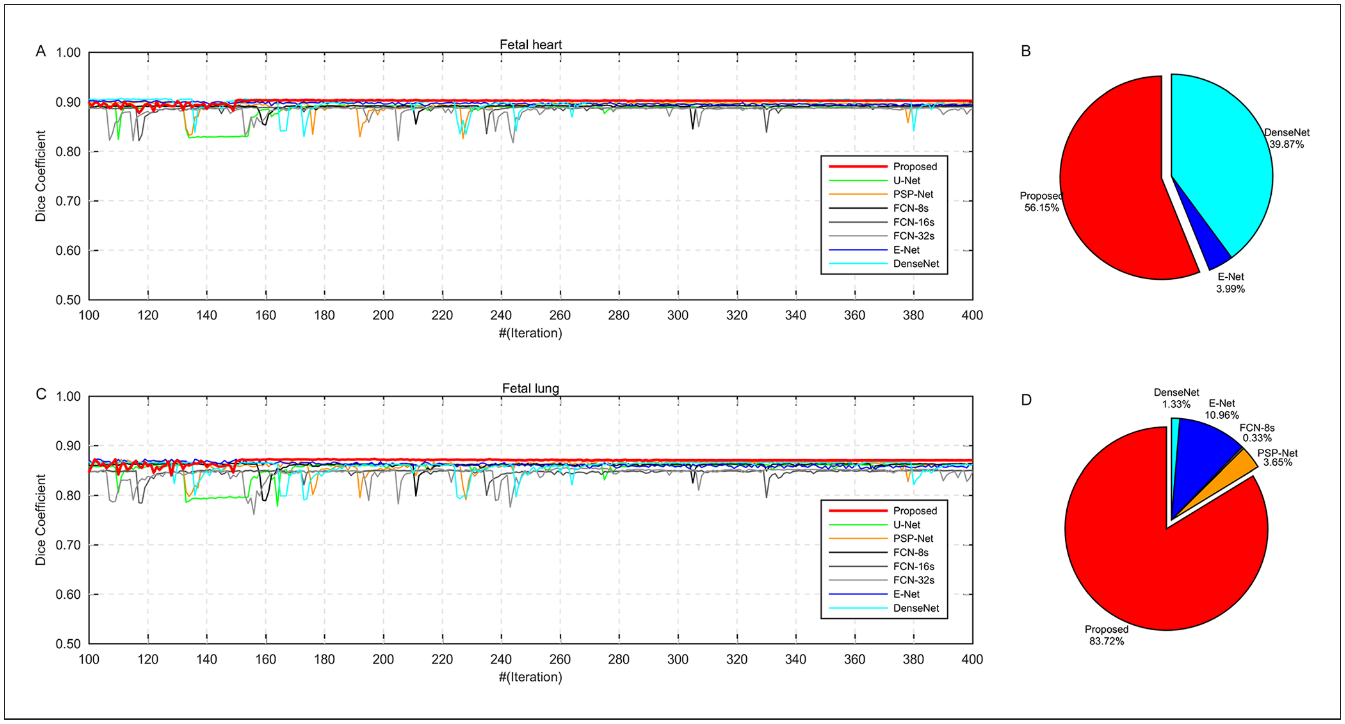

**Figure 4.** The iterative performances of our proposed method against the competing methods at steps of 100 to 400. (A) Dice coefficients of these methods for fetal heart segmentation at each iteration step. (B) Pie chart of occupation fractions of these methods as best performance for fetal heart segmentation at each iteration step. (C) Dice coefficients of these methods for fetal lung segmentation at each iteration step. (D) Pie chart of occupation fractions of these methods as best performance for fetal lung segmentation at each iteration step.

iteration step, as shown in Figure 4(B) and (D). Detailly, we collect whether a method achieves the best performance among all the comparison method for each step in 100 to 400, and calculate the fraction of the steps of their best performance in all investigated steps. By drawing the pie charts of their occupation fractions, we can intuitively observe that for both fetal hearts and lungs, the regions of segmentation performance as Dice coefficient of our method cover the biggest area among those of all methods. For fetal heart segmentation, in iterations from 100 to 400, the Dice coefficients of our methods occupy 56.15% of these iterations as best performance among all the investigated methods. For fetal lung segmentation, our method covers 83.72% iterations from 100 to 400 as best result among those of all competing methods, showing a clear dominance in iteration performance.

## Discussion

Infant mortality due to congenital anomaly can be largely reduced by fetal anomaly ultrasound scanning and subsequent intervention, but the recognition of fetal heart and lung regions manually would cost the labor of experienced physicians intensively. Still, for ultrasound image segmentation, the existing automatic segmentation approaches confront the multi-scale problem at a larger range of receptive fields of organs in images, resolution problem of segmentation mask, and interference problem of task-irrelevant features, obscuring the attainment of accurate segmentation of fetal hearts and lungs. Accordingly, we establish a deep learning based method for ultrasound image semantic segmentation of both fetal hearts and lungs, in which the multi-scale module, attention mechanism, and skip connection framework are integrated into one unified model. A systematic evaluation study also demonstrates the superiority of our method in the performances of fetal heart and lung segmentation when compared with previously published deep learning based semantic segmentation approaches. Generally, our proposed method illustrates a beneficial improvement to the elevation of the ultrasound based early scanning of congenital anomaly.

The main perspectives which might be responsible for the accomplishment of our proposed method can be summarized in three folds. The first perspective is the ability of extracting multi-scale features at a larger range of receptive fields. In our proposed method, the multi-scale modules play the roles as feature extractor to capture features at multi-scales from the input ultrasound images. The second perspective is the implementation of attention mechanism. In the task of fetal heart and lung segmentation, we incorporate the attention gate unit for highlighting the task-specific features and suppress the task-irrelevant features at the same time. The third perspective is embracing the advantage of skip connections between layers. Through the skip connections across contracting and expansive networks in U-Net architecture, high resolution information from the contracting network can be efficiently compensated to the expansive network for output, enhancing the resolution of predicted segmentation masks. The integration of skip connection framework, multi-scale module, and attention mechanism together lays the foundation for the achievement of our method in fetal heart and lung segmentation from ultrasound images.

In addition to the accomplishment obtained by our proposed method, there are also a bunch of directions that is worthwhile for further investigation. One promising direction is to introduce the multi-task learning technique into the fetal heart and lung segmentation of ultrasound images. Considering that the segmentation objectives include more than one organ, we can certainly regard the semantic segmentation of both fetal hearts and lungs as two simultaneous tasks. Furthermore, despite the usage of data augmentation on our dataset, we can still enhance the segmentation performance of our method by expecting the collection of a larger amount of data of ultrasound images of fetuses. Our proposed method also has the potential to be deployed in these applications. Moreover, since the networks heavily relied on the data collection procedure, our network was trained by the data from a single source, and it might face the overfitting problem influenced by the issue of single source data. When the network was applied on the data from a different medical center, since re-training the network from scratch would cause additional data collection of too many images, we highly recommended use transfer learning and fine-tuning on our network with data collection of an appropriate scale of images. This scheme is a promising strategy to expand the application range of our model. In summary, we propose a multi-scale model integrated with attention mechanism, which can efficiently segment the regions of fetal heart and lung from ultrasound images via extracting multi-scale features at a larger range of receptive fields, compensating high resolution information, and eliminating task-irrelevant features, showing a promising contribution on ultrasound based prognosis of congenital anomaly, facilitating the subsequent early intervention, and alleviating of the negative effects caused by congenital anomaly.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is partially supported by National Key Research and Development Program of China (No. 2018AAA0102100), partially by the National Natural Science Foundation of China (No. 61901322, 62071382, 82030047), and partially by China Postdoctoral Science Foundation (No. 2020M673494), and partially by Shaanxi Provincial Foundation for Distinguished Young Scholars (No. 2019JC-13). We would like to thank Dr. Zhaoji Miao for his helpful efforts on preparing the materials of some figures.

## ORCID iDs

Jianing Xi https://orcid.org/0000-0001-6785-5618

Dean Ta https://orcid.org/0000-0001-6651-4491

Qinghua Huang https://orcid.org/0000-0003-1080-6940

## Supplemental Material

Supplemental material for this article is available online.

## References

1. Wen SW, Liu S, Joseph KS, Rouleau J, Allen A. Patterns of infant mortality caused by major congenital anomalies. Teratology. 2000;61(5):342-6.
2. van der Linde D, Konings EE, Slager MA, Witsenburg M, Helbing WA, Takkenberg JJ, et al. Birth prevalence of congenital heart disease worldwide a systematic review and meta-analysis. J Am Coll Cardiol. 2011;58(21):2241-7.
3. Yun SW. Congenital heart disease in the newborn requiring early intervention. Korean J Pediatr. 2011;54(5):183-91.
4. Smith GC, Wood AM, White IR, Pell JP, Cameron AD, Dobbie R. Neonatal respiratory morbidity at term and the risk of childhood asthma. Arch Dis Child. 2004;89(10):956-60.
5. Wurzel DF, Chang AB. An update on pediatric bronchiectasis. Expert Rev Respir Med. 2017;11(7):517-32.
6. Thébaud B. Update in pediatric lung disease 2010. Am J Respir Crit Care Med. 2011;183(11):1477-81.
7. Huang Q, Chen Y, Liu L, Tao D, Li X. On combining biclustering mining and adaboost for breast tumor classification. IEEE Trans Knowl Data Eng. 2019;32(4):728-38.
8. Clough JR, Khanal B, van Poppel MP, Skelton E, Matthews J, Schnabel JA. Image reconstruction in a manifold of image patches: application to whole-fetus ultrasound imaging. In: Machine Learning for Medical Image Reconstruction: Second International Workshop, MLMIR 2019, Held in Conjunction with MICCAI 2019, October 17, 2019, Shenzhen, China, vol 11905, p. 226. Springer Nature.
9. Sundaresan V, Bridge CP, Ioannou C, Noble JA. Automated characterization of the fetal heart in ultrasound images using fully convolutional neural networks. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), April 18 - 21, 2017, Melbourne, VIC, pp. 671–74. IEEE.
10. Liang S, Yang F, Wen T, Yao Z, Huang Q, Ye C. Nonlocal total variation based on symmetric kullback-leibler divergence for the ultrasound image despeckling. BMC Med Imaging. 2017;17(1):57.
11. Yu Z, Tan EL, Ni D, Qin J, Chen S, Li S, et al. A deep convolutional neural network-based framework for automatic fetal facial standard plane recognition. IEEE J Biomed Health Inform. 2018;22(3):874-85.
12. Huang Q, Huang Y, Luo Y, Yuan F, Li X. Segmentation of breast ultrasound image with semantic classification of superpixels. Med Image Anal. 2020;61:101657.
13. Jardim SMG, Figueiredo MAT. Segmentation of fetal ultrasound images. Ultrasound Med Biol. 2005;31(2):243-50.

14. Xian M, Zhang Y, Cheng HD. Fully automatic segmentation of breast ultrasound images based on breast characteristics in space and frequency domains. Pattern Recognit. 2015;48(2):485-97.
15. Moon WK, Lo CM, Chen RT, Shen YW, Chang JM, Huang CS, et al. Tumor detection in automated breast ultrasound images using quantitative tissue clustering. Med Phys. 2014;41(4): 042901.
16. Lo CM, Chen RT, Chang YC, Yang YW, Hung MJ, Huang CS, et al. Multi-dimensional tumor detection in automated whole breast ultrasound using topographic watershed. IEEE Trans Med Imaging. 2014;33(7):1503-11.
17. Chang H, Chen Z, Huang Q, Shi J, Li X. Graph-based learning for segmentation of 3d ultrasound images. Neurocomputing. 2015;151:632-44.
18. Gao L, Liu X, Chen W. Phase-and GVF-based level set segmentation of ultrasonic breast tumors. J Appl Math. 2012;2012: 1–22.
19. Xian M, Huang J, Zhang Y, Tang X. Multiple-domain knowledge based mrf model for tumor segmentation in breast ultrasound images. In: 2012 19th IEEE International Conference on Image Processing, September 30 - October 3, 2012, Orlando, FL, pp. 2021–24. IEEE.
20. Huang Q, Luo Y, Zhang Q. Breast ultrasound image segmentation: a survey. Int J Comput Assist Radiol Surg. 2017;12(3):493-507.
21. Luo F, Wang M, Liu Y, Zhao XM, Li A. Deepphos: prediction of protein phosphorylation sites with deep learning. Bioinformatics. 2019;35(16):2766-73.
22. Asgari Taghanaki S, Abhishek K, Cohen JP, Cohen-Adad J, Hamarneh G. Deep semantic segmentation of natural and medical images: a review. Artif Intell Rev. 2021;54:137-78.
23. Yap MH, Goyal M, Osman FM, Martí R, Denton E, Juette A, et al. Breast ultrasound lesions recognition: end-to-end deep learning approaches. J Med Imaging. 2019;6(1):011007.
24. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, October 5 - 9, 2015, Munich, Germany, pp. 234–41. Springer.
25. Paszke A, Chaurasia A, Kim S, Culurciello E. Enet: a deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147, 2016.
26. Jégou S, Drozdzal M, Vazquez D, Romero A, Bengio Y. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, July 21–26, 2017, Honolulu, HI, pp. 11–9.
27. Shi J, Zhou S, Liu X, Zhang Q, Lu M, Wang T. Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset. Neurocomputing. 2016;194:87-94.
28. Vásquez P, Arana N, Izaguirre A, Burgos J. Labor induction failure prediction based on b-mode ultrasound image processing using multiscale local binary patterns. In: 2016 International Conference on Optoelectronics and Image Processing (ICOIP), June 10–12, 2016, Warsaw, Poland, pp. 25–9. IEEE.
29. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21 - 26, 2017, Honolulu, HI, pp. 2881–90.
30. Athira PK, Mathew LS. Fetal anomaly detection in ultrasound image. Int J Comput Appl. 2015;975:8887.
31. Lin M, Chen Q, Yan S. Network in network. arXiv preprint arXiv:1312.4400, 2013.
32. Gao SH, Cheng MM, Zhao K, Zhang XY, Yang MH, Torr P. Res2net: a new multi-scale backbone architecture. IEEE Trans Pattern Anal Mach Intell. 2021;43(2):652-62.
33. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention U-net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999, 2018.
34. Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA. Albumentations: fast and flexible image augmentations. Information. 2020;11(2):125.
35. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25:1097–105.
36. Chen LC, Yang Y, Wang J, Xu W, Yuille AL. (2016). Attention to scale: scale-aware semantic image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 27 - 30, 2016, Las Vegas, NV, pp. 3640–9.
37. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, et al. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 18 - June 23, 2018, Salt Lake City, UT, pp. 6077–86.
38. Shen T, Zhou T, Long G, Jiang J, Pan S, Zhang C. Disan: directional self-attention network for RNN/CNN-free language understanding. arXiv preprint arXiv:1709.04696, 2017.
39. Zhou D, Li M, Li Y, Qi J, Liu K, Cong X, et al. Detection of ground straw coverage under conservation tillage based on deep learning. Comput Electron Agric. 2020;172:105369.