This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2018.2858779, IEEE Transactions on Medical Imaging

Towards Automated Semantic Segmentation in Prenatal Volumetric Ultrasound

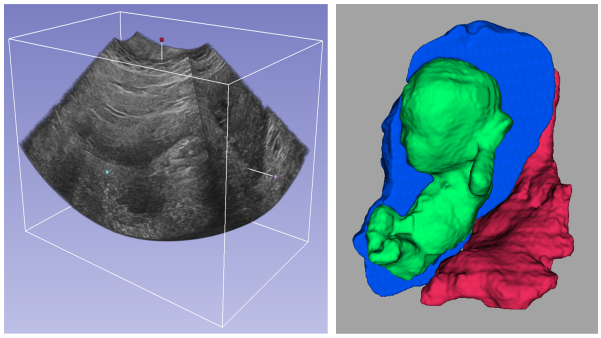
Xin Yang, Lequan Yu, Shengli Li*, Huaxuan Wen, Dandan Luo, Cheng Bian, Jing Qin, Dong Ni*, and Pheng-Ann Heng

***Abstract*—Volumetric ultrasound is rapidly emerging as a viable imaging modality for routine prenatal examinations. Biometrics obtained from the volumetric segmentation shed light on reformation of precise maternal and fetal health monitoring. However, the poor image quality, low contrast, boundary ambiguity and complex anatomy shapes conspire towards a great lack of efficient tools for the segmentation. It makes 3D ultrasound difficult to interpret and hinders the widespread of 3D ultrasound in obstetrics. In this paper, we are looking at the problem of semantic segmentation in prenatal ultrasound volumes. Our contribution is threefold: *i)* we propose the first and fully automatic framework to simultaneously segment multiple anatomical structures with intensive clinical interest, including fetus, gestational sac and placenta, which remains a rarely-studied and arduous challenge; *ii)* we propose a composite architecture for dense labeling, in which a customized 3D fully convolutional network explores spatial intensity concurrency for initial labeling, while a multi-directional recurrent neural network (RNN) encodes spatial sequentiality to combat boundary ambiguity for significant refinement; *iii)* we introduce a hierarchical deep supervision mechanism to boost the information flow within RNN and fit the latent sequence hierarchy in fine scales, and further improve the segmentation results. Extensively verified on in-house large datasets, our method illustrates superior segmentation performance, decent agreements with expert measurements and high reproducibilities against scanning variations, and thus is promising in advancing the prenatal ultrasound examinations.**

***Index Terms*—Prenatal examination, volumetric ultrasound, semantic segmentation, fully convolutional networks, recurrent neural networks.**

I. Introduction

FEATURED with fascinating benefits, such as low cost, free-hand scanning, real-time and free of radiation, ultrasound is a dominant imaging modality for maternal and fetal health monitoring during pregnancy [1]. Population studies based on biometrics extracted from traditional 2D ultrasound images have been conducted to build fetal growth reference charts. Periodical and comprehensive ultrasound screenings during different trimesters are therefore strongly recommended in obstetrics to track fetal growth for prognosis [2].



(a) (b)

Fig. 1. Illustration of (a) a prenatal ultrasound volume and (b) semantic segmentation of fetus, gestational sac and placenta which are denoted with green, ocean blue and red color, respectively.

Currently, standard prenatal ultrasound screening pipeline consists of two chained steps: manually localizing standard planes among consecutive 2D frames [3] and then annotating a series of biometrics [4]. The final diagnoses are then an interpretation of these biometrics. Although this pipeline is tractable, the resulting diagnoses are at high risk of suffering from several drawbacks. First, confined by the planar ultrasound imaging and low image resolution, standard plane selection and manual measurement are highly view- and skill-dependent [3], [4]. Subjected to intra- and inter-expert variability, diagnoses often present low reproducibilities. Second, 2D measures are inadequate to capture deterministic metrics against nonrigid deformation, e.g. fetal abdominal circumference (AC) and crown-rump length (CRL) values become ambiguous when a fetus is curled up or stretched. Third, given the error propagation along the pipeline, final evaluations are often biased by planar metrics which can only partially describe anatomical geometry.

The advent of 3D ultrasound is promising in circumventing aforementioned problems. As shown in Fig. 1(a), 3D ultrasound provides a broad volumetric view for anatomy inspection. Volumetric scanning is less user- and view-dependent than 2D scanning, which also makes off-line analysis more feasible [5]. Biometrics extracted from 3D ultrasound, such as volume, are more comprehensive and explicit than 2D ones for anatomy evaluation. Notably, 3D ultrasound paves new paths for many crucial studies that can not be approached by 2D

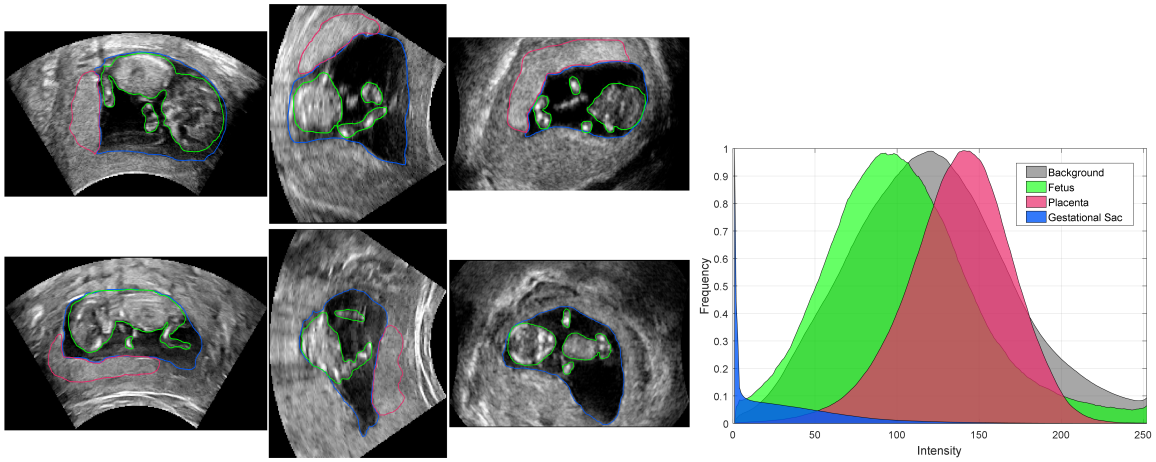
Xin Yang, Lequan Yu and Pheng-Ann Heng are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China.

Dong Ni (corresponding author, *nidong@szu.edu.cn*) and Cheng Bian are with National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China.

Shengli Li (corresponding author, *lishengli63@126.com*), Huaxuan Wen and Dandan Luo are with Department of Ultrasound, Affiliated Shenzhen Maternal and Child Healthcare Hospital of Nanfang Medical University, Shenzhen, China.

Jing Qin is with the Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2018.2858779, IEEE Transactions on Medical Imaging



(a) (b)

Fig. 2. (a) The first and second row are sagittal, traverse and coronal planes from two subjects with different anatomy shapes, scales, orientations and spatial relationships. Fetus, gestational sac and placenta are annotated with green, ocean blue and red color, respectively. (b) Intensity distributions of background, fetus, gestational sac and placenta. Fetus, placenta and background have large overlaps, while gestational sac focuses on low intensity but still has intersections with other anatomies in an extensive range. Best view in color version.

ultrasound, such as the volumes of fetus, fetal lung, gestational sac and placenta are investigated to capture early indicators of growth morbidity [6]–[9]. However, facing the explosive increase of volumetric data and expensive manual delineation, these studies often resort to semi-automatic methods, e.g. VOCAL (GE Healthcare) [9]. Since these methods still involve cumbersome, subjective interactions and run by sacrificing key substructures, they are inappropriate to tackle objects with highly irregular shapes, like the fetus and placenta shown in Fig. 1(b), and therefore tend to output conflicting research conclusions [10], [11]. In this regard, automated volumetric segmentation solutions are highly demanded to decompose prenatal ultrasound volumes efficiently and accurately.

In clinical practice, for prenatal care, doctors attach great desire to segment three key anatomical structures in volumetric ultrasound, including fetus, gestational sac and placenta (Fig. 1(b)). Fetal volume supposes to be the most intuitive indicator in assessing fetal growth [12]. Localizing fetus also facilitates plenty of automated applications, such as key plane detection, volume stitching and longitudinal analysis [3], [13], [14]. Gestational sac is the fluid environment surrounding the fetus, its volume has high potential in affecting fetal weight estimation [15]. Placenta is vital in processing blood-gas exchange and nutrient transport between maternal and fetal circulatory systems. Volume of placenta can be sensitive in scoring its maturity and a possible indicator for complex diseases, e.g. the intrauterine growth restriction (IUGR) syndrome [16], [17].

Simultaneously segmenting fetus, gestational sac and placenta in ultrasound volumes remains a nontrivial task for automated solutions. Firstly, as shown in Fig. 2(a), the ubiquitous speckle noise, acoustic attenuation and inhomogeneous resolutions severely corrupt the structural appearance and blur the boundaries. Secondly, subjected to the low contrasts and analogous intensity distributions among different anatomical structures (Fig. 2(b)), it is tough to estimate the deficient boundaries and slice the touching boundaries from the noisy background. Thirdly, these anatomical structures often present highly varying shapes, scales and orientations across different subjects and time points. The floating spatial relationships among them are hard to capture (Fig. 2(a)). Situation becomes more complicated when scanning orientation variation are considered. In summary, poor image quality, boundary deficiency and ambiguity, and the large anatomical appearance variations are the main concerns for automatic segmentation solutions.

Plenty of effort has been dedicated to meet the challenges. Anquez et al. [18] made early attempts to segment utero-fetus unit in ultrasound volumes utilizing statistical model of intensity distributions. This study treated the fetus, placenta and other utero tissues as the same class. The achieved average Dice was 0.89. However, differentiating these three anatomical structures is very important for the fetal growth evaluation and is the most difficult part as proved in our study. Stevenson et al. [19] proposed a semi-automatic method with Random Walker algorithm to segment placenta and presented an averaged Dice of 0.86. Their proposed algorithm needs very strong and explicit constraints from users. Sonia et al. [20] extended the scheme in [18] by incorporating shape priors of fetal envelope and back. As depicted in Fig. 2(b), these intensity prior based methods tend to be degraded in classifying the points with overlapped intensity distributions. Formulating segmentation as a global model fitting process, Feng et al. [21] constructed boundary traces to extract fetal limb volume for weight estimation. Namburete et al. [14] built a B-spline surface model to parametrize fetal skull volume for neural development prediction. Recently, Andrea et al. [22] explored statistical shape model for fetal facial morphology analysis. Although statistical shape models provide structure and appearance constraints to tackle artifacts and perturbations, they are initialization-dependent and inhibited in modeling multiple highly varying structures [23], like the fetus and placenta in Fig. 1(b). From a pixel-wise classification perspective, random forests are leveraged to segment fetal brain structures and femur in 3D ultrasound [24], [25]. However, the handcrafted local descriptors and limited training data confine the clas-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2018.2858779, IEEE Transactions on Medical Imaging

sifiers in tackling complex cases. We refer readers to [26] for further review on conventional methods for volumetric ultrasound segmentation.

Deep neural networks (DNNs) have brought about profound change to the medical image segmentation field by seamlessly integrating hierarchical feature extraction and discriminative model learning [27]. Leveraging 3D multi-scale convolutional neural networks (CNNs) [28], Alansary et al. [17] focused on segmenting placenta from prenatal MRI scans which present high resolutions but are much more expensive than ultrasound. They reported an averaged Dice of 0.72. Popularized with end-to-end learning and dense prediction, fully convolutional network (FCN) [29] and its variants, like u-net [30] and 3D FCN [31], presented superior performance in segmenting medical images. However, apart from the potential difficulties in training DNNs, relying on fixed convolution manner and receptive field to capture spatial intensity concurrency degrades the capability of convolutional networks to conquer arbitrary-sized boundary deficiency and ambiguity [4], [32]. Exploiting contextual information is helpful in refining semantic labeling, like the Conditional Random Field (CRF) [33], Auto-context [34] and conditional Generative Adversarial Networks (cGAN) [35]. However, driven by engineered features and links, CRF is modest in distinguishing ambiguous classes. Auto-context is also limited by pre-defined context structure. cGAN is expected to improve labelling with adversarial training. Whereas, it is hard for discriminator network to discriminate local details and irregular objects only with the real/fake label based weak supervision. Recently, endowed with internal memory, recurrent neural networks (RNNs) prove to be promising alternatives in improving semantic segmentation [36].

In this paper, as an extension of [37], we are looking at the problem of semantic segmentation in prenatal ultrasound volumes. Our contribution is threefold: *i)* we propose the first and fully automatic framework in the field to simultaneously segment fetus, gestational sac and placenta, which are anatomical structures with intensive interest in obstetrics. *ii)* Based on a 3D FCN with tailored pre-training and auxiliary supervisions to explore spatial intensity concurrency for initial labeling, we further adapt a multi-directional RNN module to encode the complementary, sequential dependency in volume in order to combat boundary ambiguity. With our work, for the first time, we prove that RNN is beneficial and promising to be a general strategy to significantly refine volumetric ultrasound segmentation, especially for weak boundary. *iii)* We introduce a hierarchical deep supervision mechanism to replenish the information flow within RNN and fit the latent sequence hierarchy in fine scales, and thus further improve the segmentation results. Extensively verified on in-house large datasets and compared with competitive solutions, our method illustrates preferable segmentation performance, decent agreements with expert measurements and high reproducibilities against scanning variations. The proposed method has great potentials to advance the quantitative analysis of prenatal scannings and thus motivate volumetric ultrasound based prenatal care.

II. Methodology

Fig. 3 is the schematic view of our proposed framework. The input is an ultrasound volume. Our customized 3D FCN conducts dense voxel-wise semantic labeling on the raw volume, and generates intermediate probability volumes for four classes, namely background, fetus, gestational sac and placenta. With serialization, our adapted RNN then blends contexture and appearance information from multiple volume channels to refine the semantic labeling. System output are the extracted volumes of fetus, gestational sac and placenta.

A. Initial Dense Semantic Labeling with 3D FCN

By interleaving convolutional layer, pooling layer and non-linearity layer in a bionic fashion and then learning feature abstracts and discriminator with a seamless manner, DNNs have brought profound change to computer vision field [38]. FCN [29] is popular in semantic segmentation for its capability in end-to-end mapping. To alleviate the irreversible spatial information loss due to the consecutive down-sampling operations, U-net [30] promotes FCN by establishing skip connections to merge feature maps from different semantic levels. Given the limited image quality, skip connections are critical for networks to recover boundary details in ultrasound images [4]. Additionally, since volumetric data inherently provide more complete spatial information than 2D planar images, it is also desired if networks can digest 3D data directly [31]. Therefore, as shown in Fig. 3 and Table I, by equipping all layers with 3D operators, we customize a 3D FCN with long skip connections bridging down-sampling and up-sampling paths to efficiently label the ultrasound volumes. Specifically, to save GPU memory, we take summation operator to merge feature volumes from different resolutions and smooth the gradient flow. Each convolutional layer (Conv) is followed by a batch normalization (BN) layer and a rectified linear unit (ReLU).

TABLE I
CONFIGURATION OF OUR CUSTOMIZED 3D FCN. LAYERS IN BOLD ARE INITIALIZED WITH TRANSFER LEARNING. STARS DENOTE LAYERS WHERE THE AUXILIARY LOSS FUNCTIONS ARE INJECTED.

Layer	Kernel size	Output size	Layer	Kernel size	Output size
Conv 1:	$3\times3\times3$	$64\times64\times64\times64$	Conv 5:	$1\times1\times1$	$16\times16\times16\times256$
Pooling 1:	$2\times2\times2$	$32\times32\times32\times64$	*Conv 6:	$3\times3\times3$	$16\times16\times16\times256$
Conv 2:	$3\times3\times3$	$32\times32\times32\times128$	DeConv 2:	$4\times4\times4$	$32\times32\times32\times128$
Pooling 2:	$2\times2\times2$	$16\times16\times16\times128$	Merge 2:	-	$32\times32\times32\times128$
Conv 3a:	$3\times3\times3$	$16\times16\times16\times256$	Conv 7:	$1\times1\times1$	$32\times32\times32\times128$
Conv 3b:	$3\times3\times3$	$16\times16\times16\times256$	*Conv 8:	$3\times3\times3$	$32\times32\times32\times128$
Pooling 3:	$2\times2\times2$	$8\times8\times8\times256$	DeConv 3:	$4\times4\times4$	$64\times64\times64\times64$
Conv 4a:	$3\times3\times3$	$8\times8\times8\times512$	Merge 3:	-	$64\times64\times64\times64$
Conv 4b:	$3\times3\times3$	$8\times8\times8\times512$	Conv 9:	$1\times1\times1$	$64\times64\times64\times64$
DeConv 1:	$4\times4\times4$	$16\times16\times16\times256$	Conv 10:	$3\times3\times3$	$64\times64\times64\times64$
Merge 1:	-	$16\times16\times16\times256$	Conv 11:	$1\times1\times1$	$64\times64\times64\times4$

Limited training data is criticized in making DNNs prone to overfitting. Equipped with 3D operators, our 3D FCN contains orders of magnitude parameters than 2D version, which intensifies the risk and asks for more tricky initialization. For vision tasks, the features learned by shallow layers in DNNs can be generic across different tasks. Sharing parameters with models that are well-trained on large scale datasets, denoted as *transfer learning*, proves to be beneficial in avoiding improper initialization and combating overfitting for better generalization ability [39]. However, prevailing models, like ImageNet

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2018.2858779, IEEE Transactions on Medical Imaging

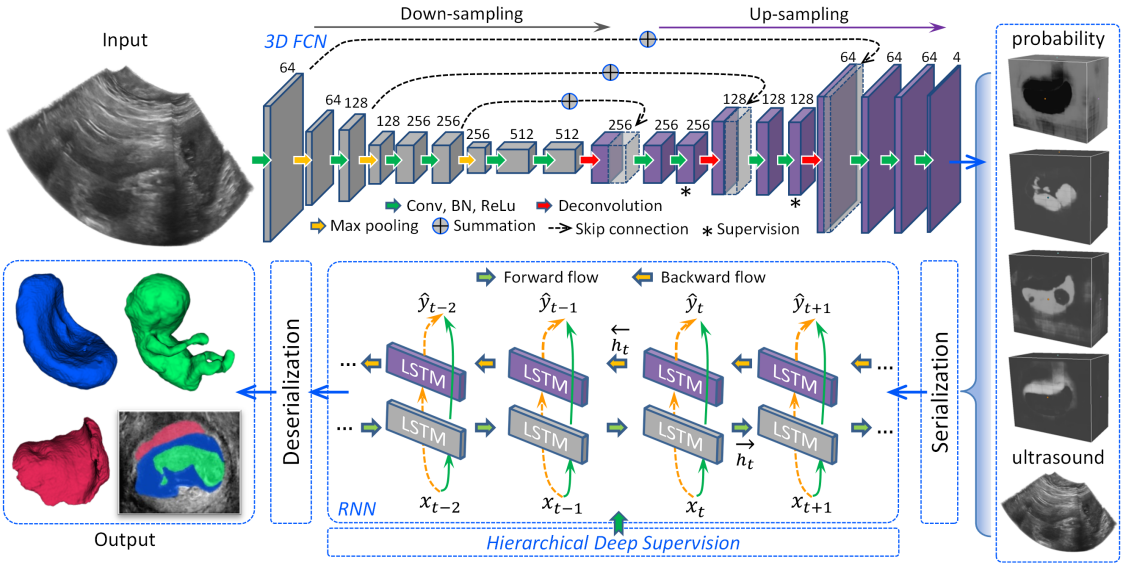


Fig. 3. Schematic view of our proposed framework. Digits denote the number of feature volume channels. For probability volumes, from top to bottom: background, fetus, gestational sac and placenta.

[38] and VGG16 [40], are designed to interpret 2D spatial information and thus are intractable to be transferred to 3D applications. Recently, the *C3D* architecture introduced in [41] gives impetus to the transfer learning in 3D DNNs. Trained on large scale video datasets, 3D convolutions enable *C3D* to simultaneously extract spatial and temporal gists across consecutive frames and thus achieve high performance on video action recognition. By adapting spatial-temporal knowledge encoded by 3D convolutional kernels to isotropic volumetric data, *C3D* model can be transferred to promote volumetric segmentation tasks, especially for anatomical structures with weak boundaries (see Section II). Specifically, we initialize the shallow layers *conv1*, *conv2*, *conv3a*, *conv3b*, *conv4a* and *conv4b* in our down-sampling path same with the layers in *C3D* model (denoted in Table I). During fine-tuning, we set small learning rates for transferred layers to avoid overtuning. Configuration of up-sampling path is symmetric with the down-sampling path but initialized from uniform distributions.

Gradient vanishing problem often adversely affects the learning process of FCNs [42]. Shallow layers in FCNs tend to be under-tuned due to the lack of strong gradient update, which results in the low training efficiency and efficacy. Shortening the backpropagation path of gradient flow for shallow layers is a direct and economical method to alleviate the problem. In this paper, we adopt the deep supervision strategy introduced in [43], which promotes training by exposing shallow layers to the extra, composite supervision of $\mathcal{M}$ auxiliary loss functions via side paths, as shown in Table I. These auxiliary paths share the ground truth with the main network. They consist of successive learnable deconvolutions for upsampling. The number of deconvolution layers depends on the size of the feature map input into the auxiliary path. Last layers of these paths finally output the predictions which have same sizes as the ground truth to inquire losses. The final loss function for our deeply supervised 3D FCN is formulated as Eq. 1,

$$\mathcal{L}(\mathcal{X},\mathcal{Y};W,w) = \mathcal{L}(\mathcal{X},\mathcal{Y};W) + \sum_{m\in\mathcal{M}} \alpha_m \mathcal{L}_m(\mathcal{X},\mathcal{Y};W,w^m) + \lambda(||W||^2) \quad (1)$$

where $\mathcal{X}$, $\mathcal{Y}$ are training pairs, W is the weight of main network, $w = (w^1, w^2, .., w^m)$ are the weights of auxiliary classifiers. Cross entropy is taken as a metric for the main loss function $\mathcal{L}$ and auxiliary $\mathcal{L}_m$. α_m is the corresponding ratio of $\mathcal{L}_m$ in final loss. Auxiliary classifiers with higher semantic levels obtain larger ratios, since their predictions are recognized as more reliable [31]. Therefore, the auxiliary classifiers attached on *conv6* and *conv8* get the ratio 0.3 and 0.6, respectively. We do not assign auxiliary loss functions in down-sampling branch, since the predictions based on shallow feature volumes are rough and it is hard to determine their ratios in final loss.

B. Semantic Labeling Refinement with RNN

Stack of convolution enables FCNs to capture features with hierarchical organization, and be advantageous in semantic labeling. However, rooting in convolution with fixed regular grid to manage data, FCNs severely resort to spatial concurrency of key features and depend on inflexible receptive field. FCNs are thus bottlenecked in reasoning ambiguous and arbitrary-sized deficient boundary and tend to output unexpected estimations [4], [32]. Motivated by [36], [44], we propose to adapt a complementary RNN in order to refine the semantic labeling from a novel, sequential perspective. Notably, this is the first work proving that RNN can be customized as a general module to significantly refine volumetric ultrasound segmentation.

RNNs are originally invented to learn from sequential stream [45]. Shown as the recursive Eq. 2, the core of RNN is the hidden state h_{t-1} stored by internal memory cells, which dynamically maintains the historical encoding of sequence

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2018.2858779, IEEE Transactions on Medical Imaging

before time $t-1$. With recurrent connections W_{hh} to reach h_{t-1}, RNN enables unlimited access from current timestep t to contextual dependencies in varying ranges. RNN then blends both the historical information h_{t-1} and excitation from current timestep x_t to update the hidden state and infer the following output $\hat{y}_t$ (Eq. 3):

$$h_t = \mathcal{H}(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{2}$$

$$\hat{y}_t = W_{hy}h_t + b_y. \tag{3}$$

By reformulating data with grid format into sequence format, RNN can be adapted to flexibly encode contextual knowledge with dynamic hidden states and enhance local predictions [46]. In this work, by serializing volumetric data into sequence, RNN sequentially runs over the local space and provides a complementary perspective to reveal volumetric contextual dependency, which proves to be crucial in combating weak and deficient boundaries in ultrasound volumes. Specifically, we propose to exploit the Bidirectional Long-Short Term Memory (BiLSTM) [45] network, a popular RNN variation, in our framework to emphasize the long range spatial dependencies and arouse interactions between sequential information flows from multiple directions, shown as Fig. 3. Given an input sequence $\boldsymbol{x} = (x_1, ..., x_T)$ and a target sequence $\boldsymbol{y} = (y_1, ..., y_K)$, BiLSTM models the probability of current timestep output by extending Eq. 2 and Eq. 3 into the following equations:

$$\overrightarrow{h}_t = \overrightarrow{\mathcal{H}}(W_{x\overrightarrow{h}}x_t + W_{\overrightarrow{h}\overrightarrow{h}}\overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}) \tag{4}$$

$$\overleftarrow{h}_t = \overleftarrow{\mathcal{H}}(W_{x\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}) \tag{5}$$

$$\hat{y}_t = W_{\overrightarrow{h}y}\overrightarrow{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y, \tag{6}$$

where W terms denote weight matrices, internal hidden states h here are controlled by tunable gates to relieve the gradient vanishing problem and preserve access to long-range dependencies, b terms denote bias vectors. $\overrightarrow{\mathcal{H}}$ and $\overleftarrow{\mathcal{H}}$ are hidden layer functions for forward and backward branch. By serializing volumes into sequences and trained with cross-entropy loss function, our BiLSTM learns a direct sequence-to-sequence mapping and outputs the improved voxel labeling for the four classes, i.e. the final segmentation results.

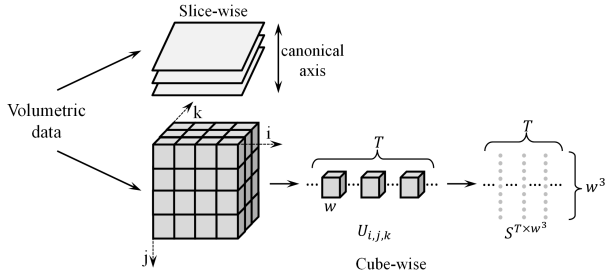


Fig. 4. Illustration of different serialization strategies.

As shown in Fig. 4, there are two intuitive strategies to serialize volumetric data into a sequence format. Slice-wise strategy as proposed in [36], [47] serializes a volume into a stack of slices along a canonical axis. Our proposed cube-wise strategy first evenly partitions the volume into $T = R \times C \times D$ overlapping cubes with the dimension $w \times w \times w$. Then, it casts the flattened cube $U_{i,j,k}$ $(i \leq C, j \leq R, k \leq D)$ as an element of a sequence $\mathcal{S} \in \mathbb{R}^{T\times w^3}$ with an index $k \times C \times R + j \times C + i$. With experiments, we find that, by choosing reasonable cube size w, cube-wise strategy performs much better than slice-wise strategy in mining the sequentiality in ultrasound volumes. With overlap-tiling stitching [48], deserialization procedure is the inverse of the cube-wise serialization.

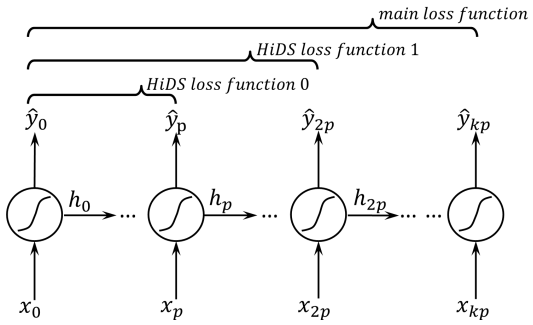


Fig. 5. The proposed hierarchical deep supervision mechanism for RNN in an unfolded version.

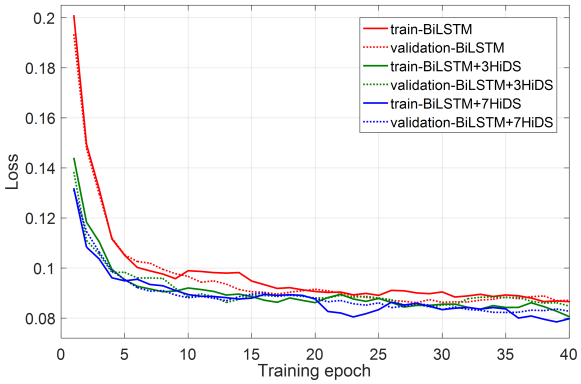


Fig. 6. The effectiveness of HiDS loss function in boosting the training procedure of BiLSTM. Loss curves of training and validation are denoted with solid and dotted lines, respectively.

Based on cube-wise serialization, to provide initial neighboring prediction context, we serialize the probability volumes of background, fetus, gestational sac and placenta as sequences S^b, S^f, S^g and S^p with dimension $T \times w^3$ respectively. We also serialize the raw ultrasound volume into a sequence S^u to provide extra appearance information (Fig. 3). Ground truth volume is serialized to serve as the target sequence for training. We then align S^b, S^f, S^g, S^p and S^u from head to head to form a sequence $S \in \mathbb{R}^{T\times 5w^3}$. Our BiLSTM computes forward hidden state $\overrightarrow{h}_t$ and backward hidden state $\overleftarrow{h}_t$ by iterating the forward branch over S from $t = 1$ to $\tilde{t}$, the backward branch over S from $t = T$ to $\tilde{t}$, where $1 < \tilde{t} \leq T = R \times C \times D$. BiLSTM can capture contextual cues over S and significantly refine the labeling result, and, as detailed in section II-C, we can get further improvement by coupling our RNN with a new supervision mechanism.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2018.2858779, IEEE Transactions on Medical Imaging

C. Hierarchical Deep Supervision for RNN

Although BiLSTM provides gating functions in order to maintain access to long-range spatial dependencies, the training of BiLSTM may be stuck when tackling sequences with extreme length ($\geq$ 1000), which is beyond common cases but exactly our scenario. With cube-wise serialization, a 50×50×50 volume can be serialized into a sequence with more than 1000 overlapping 7×7×7 cubes. The extremely long sequence not only increases the risk of gradient vanishing for early sequence steps, thus leading to the low training efficiency and efficacy, but also makes the global loss function over the whole sequence difficult to fit all local sequence segments. Minimizing the sole loss function for the whole sequence may not suit local prediction well. Few studies have been reported for effective deep supervision mechanisms in RNNs. Lipton et al. [49] proposes to replicate the target label over each timestep to boost classification task, whereas it is intractable for our sequence-to-sequence mapping task. A proper deep supervision strategy for RNNs needs to consider the following two factors: (i) auxiliary supervision should be injected in early timesteps to replenish the gradient flow; (ii) the locations to trigger auxiliary supervision should consider the multi-scale, hierarchical context dependencies in the sequence. Rooting in these thoughts, we propose a novel, hierarchical deep supervision mechanism (HiDS) to promote the training and generalization of RNN, as shown in Fig. 5. Sharing the same anchor point, addition to the main loss function at the chain end, HiDS attaches auxiliary loss functions along the sequence with gradually increasing scopes. The final loss function with $\mathcal{N}-1$ HiDS losses is defined as Eq. 7, where X, Y are input and output sequences with length T and $\mathcal{N}p = T$. W is the weight matrix of RNN shared by all timesteps. $\mathscr{L}_N$ is the main loss function charging the whole sequence, $\mathscr{L}_n$ are auxiliary loss functions which focus on different scales, β_n are the associated ratio in final loss $\mathscr{L}$. Because the sequence fragments with different lengths share the same network, predictions and losses from them present the same reliabilities. Thus, we choose to equally set β_n to 1.0 in this work.

$$\begin{aligned}&\mathscr{L}(\mathrm{X},\mathrm{Y};W) = \\ &\mathscr{L}_N(\mathrm{X},\mathrm{Y};W) + \sum_{n=1}^{\mathcal{N}-1}\beta_n\mathscr{L}_n(\mathrm{X}_{1\leq t<np},\mathrm{Y}_{1\leq t<np};W)\end{aligned} \tag{7}$$

Fig. 6 provides a proof about HiDS in boosting the training procedure of BiLSTM for segmentation running over a sequence with 1000 timesteps. BiLSTM equipped with 3, 7 auxiliary HiDS loss functions get faster convergence speeds and lower training errors than BiLSTM with only a main loss function. All these three models are trained with 120 epochs for convergence and only 40 epochs are shown in Fig. 6 to highlight significant improvements. Results with 120 training epochs and improvement in generalization ability brought by HiDS are further elaborated in section III.

III. Experimental Results

A. Implementation Details

We implemented our framework with *Caffe* [50] for the customized 3D FCN part, *Theano* [51] for the BiLSTM part. *Codes for these two parts are publicly available now* [1]. Model training and testing are run on a NVIDIA GeForce GTX TITAN X GPU (12GB). Although training FCN and RNN in an end-to-end fashion is expected, it's very challenging when there are no good initialization and gradient update to synchronize the 3D FCN and BiLSTM [52]. Alternatively, we choose a two-stage training scheme, i.e., firstly train the 3D FCN with transfer learning for 100 epochs, then train the BiLSTM from scratch with 120 epochs. We set the learning rate for transferred layers in 3D FCN as *1e-6* while other layers as *1e-3*. Learning rate is halved every 20 epochs. In each epoch, limited by GPU memory, 3D FCN takes randomly cropped 64×64×64 sub-volumes as input. We allocate 800 internal memory cells for forward and backward branch each in BiLSTM. Our BiLSTM is trained with an initial learning rate *1e-3*. BiLSTM takes randomly cropped 50×50×50 sub-volumes from the ultrasound volume and 4 probability volumes at same position as input. These sub-volumes are serialized into sequences of overlapping cubes with size 7×7×7. All the sampled sub-volumes are normalized as zero mean and unit variance before fed into networks. Limited by GPU memory, we adopt sliding window and overlap-tiling stitching strategies [48] for 3D FCN and BiLSTM to generate predictions for the whole volume. The final probability map of the overlapped voxels are the mean of probability maps of the overlapped sub-volumes.

B. Datasets and Evaluation Criteria

We built a dataset consisting of 104 prenatal ultrasound volumes acquired from 104 pregnant women volunteers. Approved by local Institutional Review Board, all volumes were anonymized and obtained by an expert using a Mindray DC-8 ultrasound system with an integrated 3D probe, which has a wide transmit frequency range from 3.8 to 8.2 MHz. The 3D probe has a large field of view (75 degree) to capture the complete structures (gestational sac, fetus and placenta) in early gestational ages. Free fetal poses are allowed during scanning. Gestational age ranges from 10 to 14 weeks during which the field of view of the 3D probe can completely cover all the three anatomical structures. Volume size in our dataset varies from subject to subject, with average size of 221×198×283 and unified voxel size of 0.5×0.5×0.5 mm^3. Our solution processes with the original volume without any rescaling operations. Ten sonographers with 5-year experiences provided manual delineation for all volumes as segmentation ground truth. Being proficient in using the annotation software *ITK-SNAP* [53], a sonographer needs about 4 hours to delineate the three anatomies in one volume. All the annotation results are double-checked under strict quality control from a senior expert with 20-year experience. This dataset is currently the

[1] 3D FCN: https://github.com/xy0806/miccai17-mmwhs-hybrid, BiLSTM: https://github.com/xy0806/miccai17-rnn-theano

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2018.2858779, IEEE Transactions on Medical Imaging

TABLE II
ABLATION STUDY ABOUT TRANSFER LEARNING AND DEEP SUPERVISION ON 3D FCN

Method	Gestational sac				Fetus				Placenta			
	Dice	Conf	Adb	Hdb	Dice	Conf	Adb	Hdb	Dice	Conf	Adb	Hdb
3D-F	0.866	0.685	1.118	10.61	0.846	0.629	1.264	9.946	0.365	-4.843	4.769	33.79
3DF-TL	0.882	0.729	1.010	10.82	0.873	0.706	0.879	**8.442**	0.584	-0.637	3.804	27.30
3DF-TL-DS	**0.884**	**0.735**	**0.963**	**10.31**	**0.876**	**0.714**	**0.875**	8.769	**0.590**	**-0.583**	**3.449**	**26.48**

TABLE III
QUANTITATIVE COMPARISON OF DIFFERENT REFINEMENT STRATEGIES

Method	Gestational sac				Fetus				Placenta			
	Dice	Conf	Adb	Hdb	Dice	Conf	Adb	Hdb	Dice	Conf	Adb	Hdb
AutoCtxt-L0	0.884	0.735	0.963	10.31	0.876	0.714	0.875	8.769	0.590	-0.583	3.449	26.48
AutoCtxt-L1	0.884	0.733	1.019	10.79	0.878	0.719	0.843	8.559	0.602	-0.542	3.338	26.83
AutoCtxt-L2	0.881	0.725	1.015	11.03	0.872	0.703	0.931	8.869	0.540	-1.088	2.960	25.53
DenseCRF	0.876	0.706	1.814	12.61	0.875	0.710	1.628	10.72	0.604	-0.472	6.312	32.49
FB-Slice	0.872	0.702	0.993	10.57	0.841	0.620	0.916	11.68	0.610	-0.450	3.636	25.60
3DF-cGAN	0.889	0.745	0.951	10.26	0.880	0.726	**0.671**	**7.345**	0.571	-1.112	**2.422**	24.95
FB-0HiDS	0.887	0.741	0.971	10.38	0.875	0.711	0.776	8.285	0.619	-0.371	3.296	25.97
FB-3HiDS	0.889	0.746	0.936	10.13	0.879	0.723	0.758	8.015	**0.635**	**-0.301**	2.597	**24.18**
FB-7HiDS	**0.890**	**0.749**	**0.925**	**9.788**	**0.882**	**0.729**	0.743	7.904	0.631	-0.315	2.480	24.71

largest one reported in the field. We then randomly split the dataset into 50, 10 and 44 volumes for training, validation and testing, respectively. Considering fetal pose variation, we further augmented the training dataset into 150 volumes via horizontal and vertical flipping. Both of our 3D FCN and RNN parts are trained with voxel-wise cross-entropy losses for end-to-end mapping. This setting enables the network to utilize every voxel in the volume for training and thus alleviate the risk of overfitting [31]. Also, the randomly cropped sub-volumes in each training epoch further augment our training corpora to a considerable amount.

The detailed training and validation curves of BiLSTM are shown in Fig. 6. For those three models, the losses of training and validation gradually decrease with slight fluctuations. Validation curves are similar to the training and no obvious inflection point is observed along all epochs. This kind of phenomenon is also reported in [31] when the domain shift between training and validation sets is small. Patch-based training strategy generates many distinctive training samples and further reduces the risk of overfitting. We stop the training of all models at epoch 120 where all of them achieve convergences.

For segmentation evaluation, we target to assess both the region and boundary similarities with 4 criteria. Dice similarity coefficient ($Dice = 2(A \cap B)/(A + B)$) indicates the mutual overlap between segmentation and ground truth. A new measure coefficient Conformity is adopted ($Conf = (3Dice - 2)/Dice$) which provides wider range and can be more sensitive and rigorous than Dice [54]. Average Distance of Boundaries (Adb[mm]) is used to describe the average distance from segmentation to ground truth. Hausdorff Distance of Boundaries (Hdb[mm]) is sensitive to boundary outliers and emphasizes the worst labeling cases [55].

C. Quantitative and Qualitative Analysis

*1) **Ablation Study on Transfer Learning and Deep Supervision**:* We firstly conduct ablation study to verify the significance of transfer learning and deep supervision. We take the vanilla 3D FCN as a baseline (denoted as 3D-F). The 3D FCN trained with the transfer learning is denoted as 3DF-TL. The method further equipped with deep supervision mechanism by adding two auxiliary loss functions with upsampling layers on *Conv6* and *Conv8* is denoted as 3DF-TL-DS. Table II compares the performance of these three methods in segmenting gestational sac, fetus and placenta. Conquering shape and size variations, our basic 3D FCN presents decent ability in segmenting gestational sac and fetus, but poor in slicing the placenta. Transferring the spatial-temporal knowledge in *C3D* model to volumetric segmentation prominently benefits the segmentation of three anatomies in both area and distance criteria, especially in recognizing the weak placenta (with 20 percent improvement in Dice). Boosted training contributed by auxiliary supervision further improves the segmentation on almost all the metrics. In the following experiments, we will take the predictions from 3DF-TL-DS as a starting point.

*2) **Comparison with Auto-Context and DenseCRF**:* Based on the initial prediction from 3DF-TL-DS, we introduce our HiDS promoted BiLSTM module (denoted as FB-*n*HiDS, where n is the number of HiDS loss function). We also compare our method with some classical semantic labeling refinement schemes, such as Auto-Context and fully connected CRF (denoted as DenseCRF). Table III lists the quantitative comparisons among FB-*n*HiDS, Auto-Context and DenseCRF. Auto-Context [34] cascades a series of models in a way that, the model at level k can revisit the prediction maps in level $k-1$ to collect contexture for refinement. As a standard implementation, we construct our Auto-Context by stacking n 3D FCN networks in successive levels, denoted as AutoCtxt-Ln ($n \geq 0$). All levels have the same configurations

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2018.2858779, IEEE Transactions on Medical Imaging

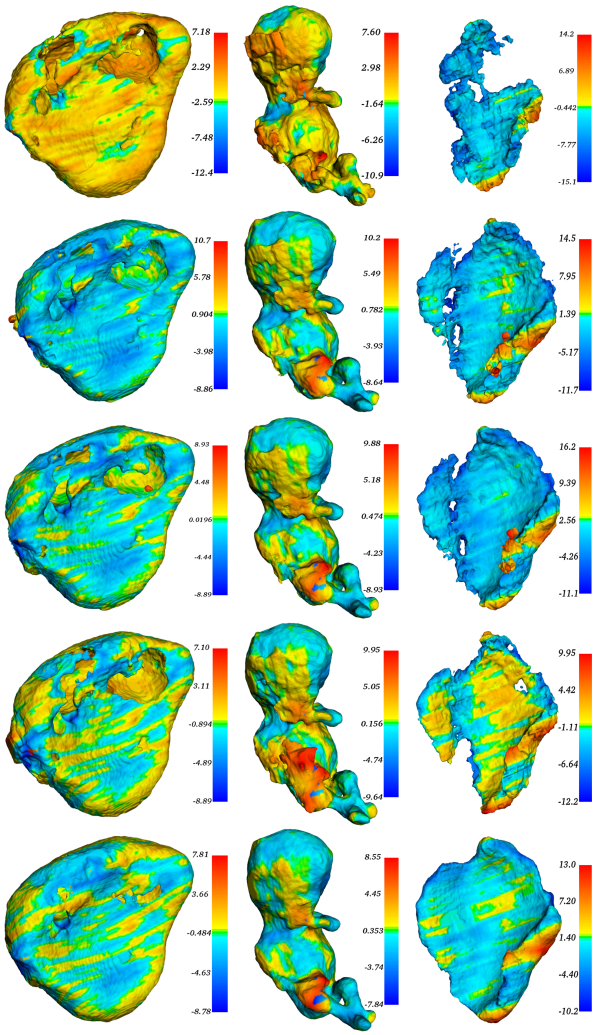


Fig. 7. Comparison of Hausdorff distance [mm] of different methods. From left to right, gestational sac, fetus and placenta. From top to bottom, *3D-F*, *3DF-TL-DS*, *DenseCRF*, *AutoCtxt-L2* and *FB-7HiDS*. The color bar is annotated with *mean* in the center, *min* and *max* on the ends.

with 3DF-TL-DS, but the input for AutoCtxt-Ln ($n \geq 1$) are 5 channels. As illustrated in Table III, we did not observe obvious improvement in AutoCtxt-L1 and AutoCtxt-L2, and sometimes the performance gets even worse. We attribute this to the fact that, 3D FCNs in different levels only apply the same fixed convolution to their predecessors' output and thus can not inject obviously different information for significant refinement. Balancing unary intensity features and pairwise diversity to determine pixel clustering, DenseCRF [1] with high efficiency becomes a popular module to achieve structure-refined segmentation [33], [56]. However, even with arbitrarily large neighborhoods and careful parameter setting, DenseCRF in 3D fashion only brings slight improvement in Dice over placenta while greatly sacrifices boundary similarities over all anatomies. The reason may be that, the analogous intensity distributions among different anatomical structures (Fig. 2(b)) are difficult for pairwise terms in DenseCRF to capture and thus corrupt the final predictions.

[1]Code source: https://github.com/lucasb-eyer/pydensecrf

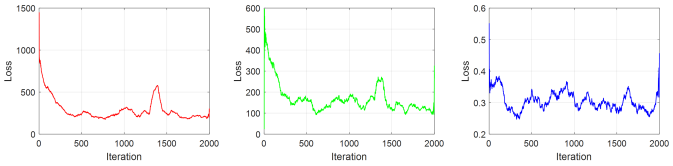


Fig. 8. Training details of 3DF-cGAN. From left to right, loss curves for discriminator, generator and segmentation.

*3) **Performance of HiDS**:* Compared with Auto-Context and DenseCRF, our FB-nHiDS presents improvement in all criteria. Trained only with the main loss function, FB-0HiDS already improves the segmentation of all structures, especially the placenta which is the most difficult to segment. This corroborates the efficacy of BiLSTM in exploring varying context. When fine-scaled HiDS loss functions are embedded to enhance the gradient flow within BiLSTM and provide more fragment-specific supervision over the ultra long sequence, we can observe continuous refinement from FB-3HiDS and FB-7HiDS. Compared with 3DF-TL-DS, FB-3HiDS performs preferably in segmenting all anatomies with an improvement about 0.6 percent in Dice for gestational sac and 0.3 percent for fetus, 4 percent in Dice and 1.7mm in Hdb for placenta. However, we encountered a performance drop in segmenting placenta when moving from FB-3HiDS to FB-7HiDS. This reveals the requirement of a more proper configuration of HiDS and the optimization is left to our future research.

*4) **Paired t-test between Methods**:* To verify whether the performance of different methods is significantly different, we take the Dice as a measure and conduct paired t-tests among FB-7HiDS, DenseCRF and AutoCtxt-L1 methods. Fig. 9 shows the detailed Dice results generated by these methods. In our paired t-tests, the significance level is set as 0.05. The detailed p-values for the paired t-tests among FB-7HiDS, DenseCRF and AutoCtxt-L1 for gestational sac, fetus and placenta segmentation are elaborated in Table IV. All of these paired t-tests present p-values smaller than 0.05 and thus prove that FB-7HiDS performs significantly better than other methods, rather than by chance.

*5) **Comparison of Different Serialization Strategies**:* We further validate the impact of different serialization strategies. In Table III, based on 3DF-TL-DS, FB-Slice denotes the BiLSTM, which takes input with a slice-wise serialization strategy. We can see that, slice-wise serialization can encode the inter-slice dependency in volume to an extent and thus facilitates FB-Slice in segmenting placenta. However, FB-Slice can not get performance gain in segmenting fetus and gestational sac when compared with FB-0HiDS, which only differs in the serialization strategy.

*6) **Comparison with Conditional GAN**:* As the most recent cGAN proves to be effective in refining the organ segmentation in Chest X-rays [35], we also reimplement a cGAN for extensive comparison, denoted as *3DF-cGAN*. We take the pre-trained 3DF-TL-DS as the segmentation generator. The discriminator contains 4 Conv-BN-ReLU layers and a fully connected layer. The generated segmentation is then concatenated with the input $64 \times 64 \times 64$ ultrasound patch which serves as a condition constraint. The discriminator classifies

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2018.2858779, IEEE Transactions on Medical Imaging

TABLE IV
PAIRED T-TESTS (P-VALUE) BETWEEN DIFFERENT METHODS

Method	Gestational sac		Fetus		Placenta	
	DenseCRF	AutoCtxt-L1	DenseCRF	AutoCtxt-L1	DenseCRF	AutoCtxt-L1
FB-7HiDS	0.0241	0.0021	0.0294	0.0180	2.5839*e-5*	1.8174*e-7*

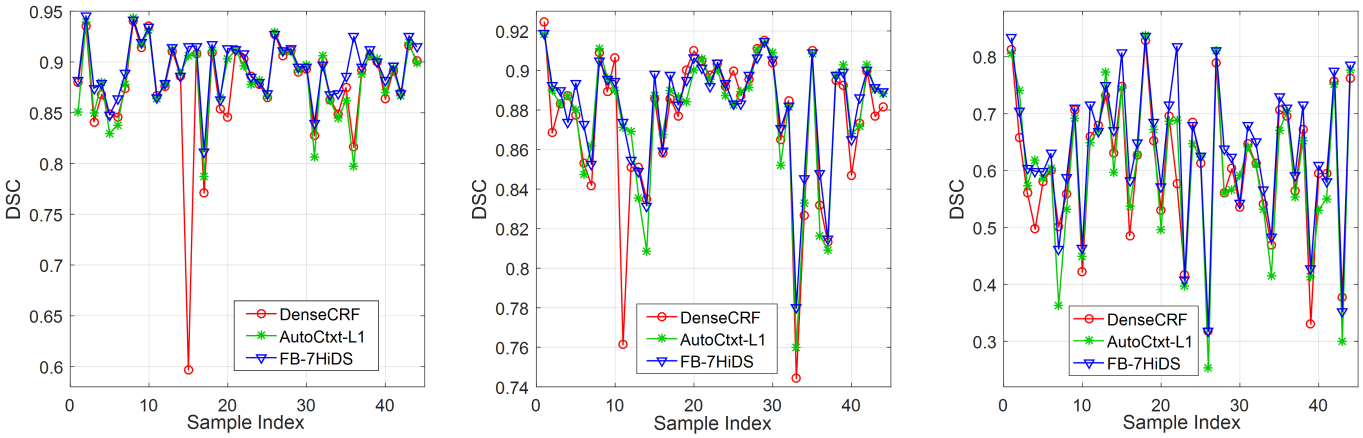


Fig. 9. From left to right, Dice results of segmentation on gestational sac, fetus and placenta by three methods.

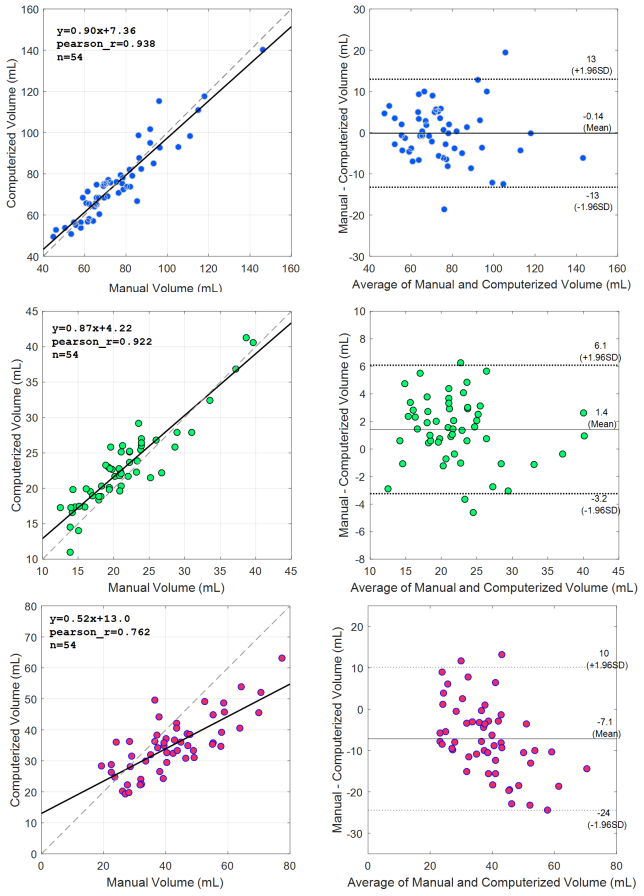


Fig. 10. Correlation and Bland-Altman agreement test on measuring three anatomies. From top to bottom, gestational sac, fetus and placenta.

the concatenated pair as real or fake. *We refer readers to our 3D cGAN implementation online available*[2] *for more details.* The loss curves for discriminator, generator and segmentation during training are shown in Fig. 8. Convergences are achieved with fluctuations. Based on the 3DF-TL-DS, the 3DF-cGAN comes to be a very comparable competitor in segmenting gestational sac and fetus. However, it brings about severe drop in placenta segmentation, which may be because of the irregular shape and weak boundaries of the placenta that are hard for the discriminator to evaluate.

7) ***Visualization of Different Segmentation Results****:* Fig. 7 visualizes the Hausdorff distance from different segmentation surfaces to the ground truth. Our proposed FB-7HiDS presents more accurate and concrete segmentation than other methods, concerning the placenta. More explicit visualization of the semantic segmentations produced by FB-7HiDS for fetus, gestational sac and placenta are shown in Fig. 13. Our method conquers the poor image quality, scale and pose variation, complicated spatial configuration (even twins) and boundary ambiguities, and presents satisfying segmentation for all the three anatomical structures.

8) ***Variability in Manual Labeling****:* Manual delineation of three anatomical structures in ultrasound volume is extremely labor-intensive and expensive. We invited 10 experts with more than 5-year diagnosis experiences to label the dataset used in this study. Each expert was only responsible for the labeling of about 10 volumes. We further invited one senior expert with 20-year diagnosis experience for strict quality control. All the annotations were finalized after two rounds of modifications guided by the senior expert. The annotation process took about 30 days. To evaluate the variability in manual ground truth, we randomly selected 12 volumes in the testing dataset (the

[2]Code source: https://github.com/xy0806/3D-cGAN

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2018.2858779, IEEE Transactions on Medical Imaging

TABLE V
COMPARISON BETWEEN DIFFERENT CUBE SIZES FOR REFINEMENT

Method	Gestational sac				Fetus				Placenta			
	Dice	Conf	Adb	Hdb	Dice	Conf	Adb	Hdb	Dice	Conf	Adb	Hdb
FB-7HiDS-w5	0.889	0.748	0.948	10.43	**0.883**	**0.734**	0.765	7.985	**0.638**	**-0.282**	2.778	**24.57**
FB-7HiDS-w7	**0.890**	**0.749**	0.925	9.788	0.882	0.729	**0.743**	**7.904**	0.631	-0.315	**2.480**	24.71
FB-7HiDS-w9	0.889	0.748	0.918	9.661	0.879	0.722	0.755	8.012	0.626	-0.344	2.609	25.07
FB-7HiDS-w11	0.888	0.745	**0.910**	**9.617**	0.876	0.713	0.762	8.132	0.631	-0.312	3.162	26.16

TABLE VI
VARIABILITY OF DIFFERENT DELINEATIONS ON 12 RANDOMLY SELECTED TESTING SAMPLES

Metric	Gestational sac			Fetus			Placenta		
	E1 vs E2	Alg vs E1	Alg vs E2	E1 vs E2	Alg vs E1	Alg vs E2	E1 vs E2	Alg vs E1	Alg vs E2
Dice	0.937±0.04	0.887±0.04	0.875±0.04	0.936±0.05	0.854±0.04	0.829±0.06	0.929±0.04	0.601±0.11	0.587±0.11
Conf	0.862±0.09	0.741±0.10	0.708±0.12	0.859±0.12	0.653±0.12	0.574±0.20	0.847±0.09	-0.444±0.71	-0.545±0.79
Adb	0.467±0.27	1.010±0.25	1.044±0.25	0.425±0.35	1.122±0.36	1.339±0.55	0.541±0.32	3.750±1.23	3.959±1.39
Hdb	5.127±2.30	8.111±2.12	8.311±2.60	1.929±1.56	9.292±3.60	9.275±3.98	5.328±1.71	26.41±10.85	26.51±10.88

original labeling is denoted as E1) and invited one extra expert (denoted as E2) to manually label them again with the same quality control. E2 was blinded to the manual label of E1. Table VI illustrates the labeling variabilities between E1 and E2 and the agreement between our proposed FB-7HiDS (Alg) and the experts. High reproducibility is presented between E1 and E2 on all three structures, which proves the efficacy of our quality control strategy. Our proposed Alg shows similar agreement with E1 and E2. Alg performs better in segmenting fetus and gestational sac than placenta. As for the difference in volume measurement, we conduct the Bland-Altman tests on the volume size obtained from E1 and E2 annotations. Details of the tests are shown in Fig. 11. As we can see, for the 12 selected samples, volume measurements for three anatomical structures obtained from E1 and E2 show high agreements, which also proves the low variability in the manual labeling.

*9) **Impact of Different Sizes of** w:* As mentioned in Section II-B, based on our proposed serialization strategy, the cube size w determines the neighboring range encoded by each element in the sequence and consequently has impact on RNN based refinement. Table V compares the performance of the FB-7HiDS with different w (FB-7HiDS-wi, i is cube size). The shape similarity metrics, including Dice and Conf, gradually drop as w increases for all three structures. It indicates that packing too much information in one cube prohibits the RNN to effectively explore local sequentiality. For the boundary similarity metrics, the performance depends on the specific anatomical structure as w varies. Large w causes the Adb and Hdb to decrease when segmenting gestational sac, but causes larger boundary distances when segmenting fetus and placenta. Such opposite change may be due to the difference in shape between different anatomical structures, such as the concrete shape of gestational sac and the complicated geometry with branchy details of fetus and placenta. In this study, we set w to 7 as a compromise between shape and boundary similarities.

*10) **Agreements with Expert Measurements**:* For clinical practice, volume size extracted from segmentation is preferred. To compare the volume size derived from experts' annotation and FB-7HiDS, we adopt the correlation coefficient and Bland-Altman agreement [55] for a comprehensive assessment. As shown in Fig. 10, tested on the 54 volumes (testing and validation sets), our algorithm achieves high correlations (0.938 and 0.922) and agreements (-0.14±13 mL and 1.4±4.7 mL) on measuring the volume size of gestational sac and fetus when compared to experts. FB-7HiDS presents a modest correlation in measuring placenta volume size, but 95% of the measurements still locate in the ±1.96 standard deviation in Bland-Altman plot.

*11) **Reproducibility against Scanning Variation**:* Different probe scanning directions can obviously change the appearance of ultrasound volume, therefore being robust against scanning direction variation becomes vital for automated algorithms. Accordingly, to verify the reproducibility of our algorithm, we newly collected 75 volumes from 15 volunteers. Each volunteer was scanned with five predefined, distinctive scanning directions. Taking fetal face as the anterior direction, the rest four are defined from left, right, posterior and superior directions. Five scannings from the same volunteer are analyzed as a group. The gestational age of this reproducibility dataset ranges from 11 to 14 weeks. The detailed distribution of the gestational ages is illustrated in Fig. 12. Most data were acquired when the fetus was in static state, while few groups presented obvious pose variations, especially the limbs. Fig. 12 shows the box-plots of volume measurements generated by FB-7HiDS in each group. We can observe that, our algorithm attains high reproducibilities over all groups in measuring these three anatomical structures. Suffering less from acoustic shadow and intensive ambiguity caused by noisy background, the reproducibility on fetus (mean standard deviation is 1.425 mL) is higher than that on gestational sac and placenta (mean standard deviation are 7.894 mL and 6.098 mL, respectively). Because the volume of gestational sac depends on its engorgement, which varies across different subjects and timepoints, therefore the variation in gestational sac volume among different groups is much larger than that in fetus and placenta. Since placenta is the most difficult part to segment and suffers severely from imaging conditions, the variation in placenta volume among different groups is much

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2018.2858779, IEEE Transactions on Medical Imaging

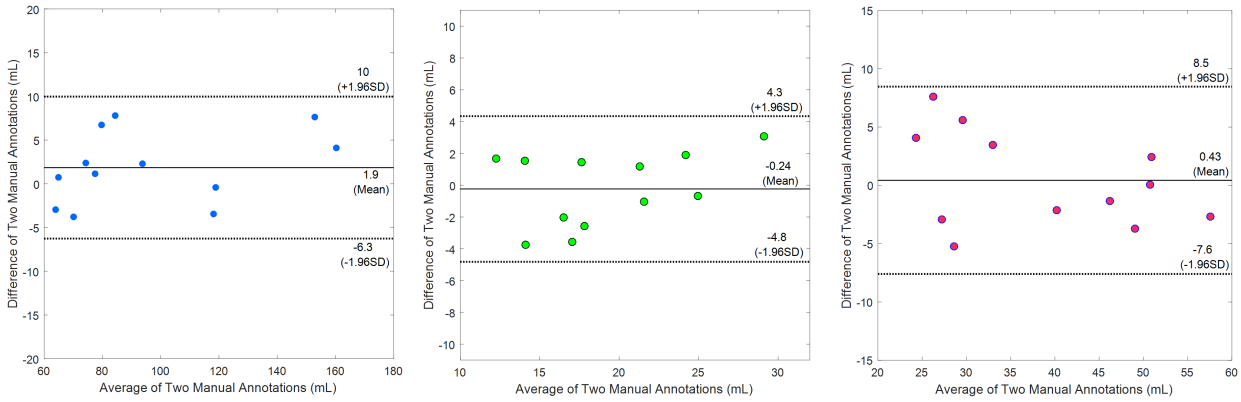


Fig. 11. Bland-Altman agreement between the manual measurements from E1 and E2. From left to right: plot for gestational sac, fetus and placenta.

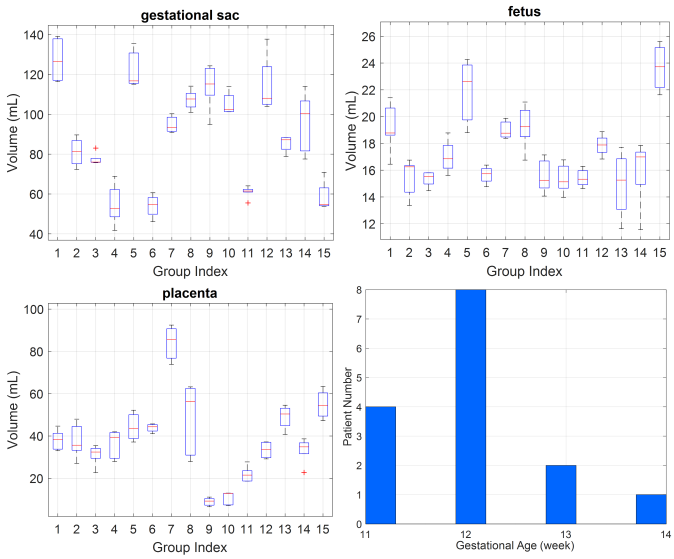


Fig. 12. Measurement reproducibility against different scanning directions.

larger than that in fetus measurement.

IV. Limitation and Discussion

Our proposed framework provides a fundamental and novel tool for prenatal volumetric ultrasound analysis, and may facilitate many advanced applications. Whereas, there still exist several key points for future study. Currently, our framework consists of 3D FCN and RNN, which are two complementary but separately trained parts. Leveraging the feedback from RNN which encodes spatial sequentiality to remedy the preference of FCN in capturing spatial intensity concurrency should further benefit our segmentation. As an advanced mixture of FCN and RNN, ConvLSTM [57] is a promising design which can seamlessly blend concurrency and sequentiality for inference. However, currently, only ConvLSTM in 2D format has been reported for biomedical image segmentation [36], [47]. Implementing efficient 3D ConvLSTM is still pending. Additionally, in our current framework, RNN takes the flattened format of cube as input (Fig. 4) which may slightly destroy the spatial information, while ConvLSTM alleviates this problem by digesting grid-structured data with a convolutional fashion at each time step.

High time efficiency is expected for our task. However, confined by GPU memory, we resort to the overlap-tiling stitching strategy for both FCN and RNN to generate the prediction of whole volume. Also, as a sacrifice to encoding the volumetric sequentiality well, our cube-wise serialization strategy produces extreme long sequences. These two factors render our system with a low efficiency (about 2 minutes for FCN, 3 minutes for RNN to process a volume). Making DNNs more computationally economical via layout design should be considered in the future, such as the densely connected networks [58] and pseudo-3D networks [59]. Reducing computational burden also enables us to operate with larger input volume size which is crucial to involve broader context.

From the reproducibility experiments, we can see that different scanning directions and poses have impacts on the segmentation. Major reason of the problem may be the patch-based learning and prediction, which discard the strong global information during partition. Another possible reason is that the grid-wise convolution is not robust against rotation. This may degrade the segmentation when gestational sac, fetus and placenta have different spatial configurations, poses and orientations. Data augmentation is insufficient to combat this problem. Although challenging, enabling networks to handle large volumes and be invariant against geometry transformation are important research directions in future.

We build our segmentation solution in the scenario of early gestational ages (<14 weeks), where the 3D ultrasound probe can completely capture the gestational sac, fetus and placenta. For the volumetric analysis of these three anatomical structures in larger gestational ages where only partial information can be obtained in each scanning, our solution can still serve as a backbone and support complex techniques, such as volume stitching [13] to achieve the complete segmentation.

Based on the detailed Dice evaluation in Fig. 9, we can see our dataset contains considerable variations in appearance. Regarding the limited imaging resolution, complicated surroundings, small scales and weak acoustic signals of fetal limb, segmenting fetal limbs is much more difficult than segmenting fetal torso. As the two worst cases shown in Fig. 14, discontinuities tend to occur around the elbow and twist.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2018.2858779, IEEE Transactions on Medical Imaging

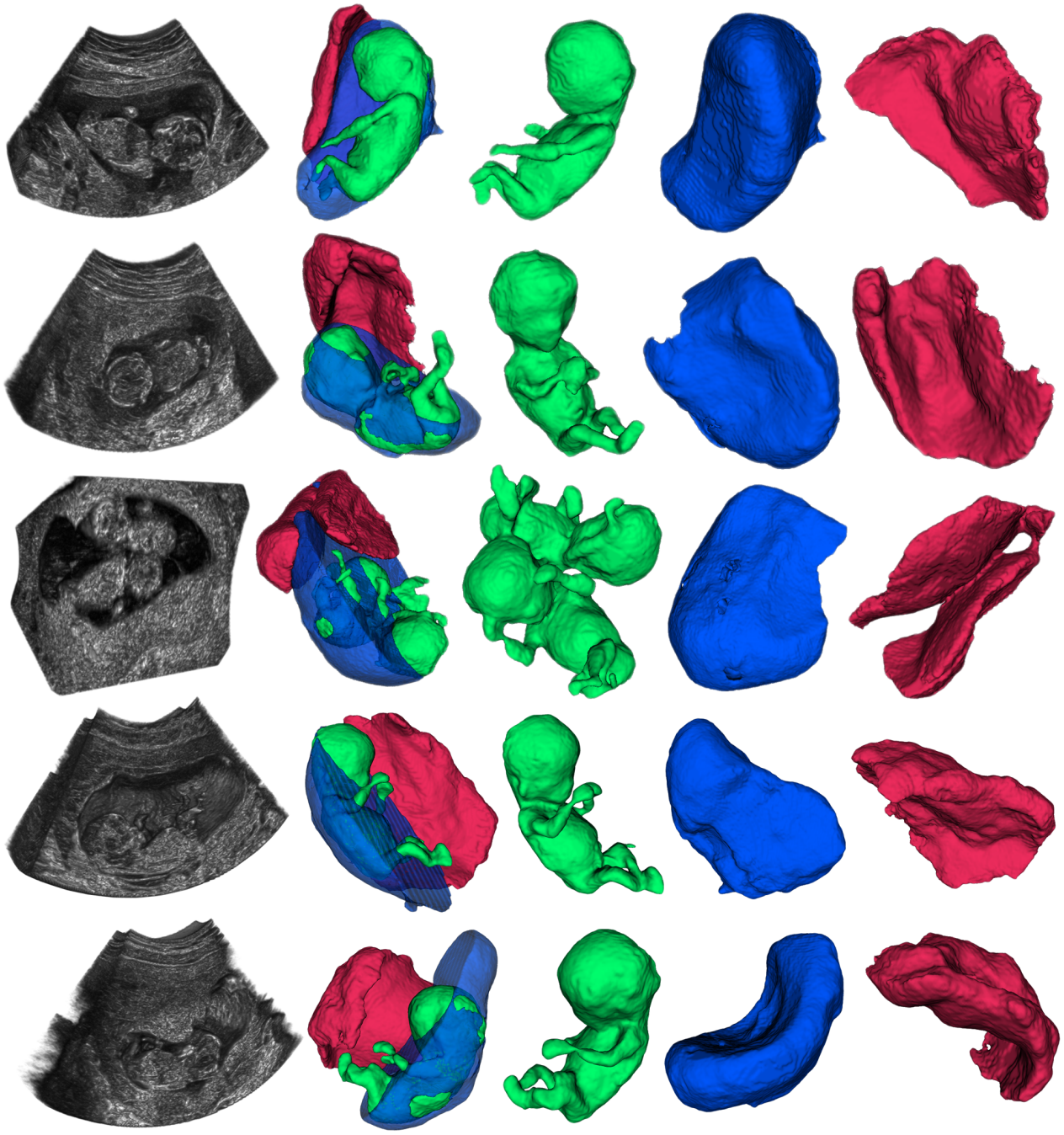


Fig. 13. From left to right: cutaway view of prenatal ultrasound volume, cutaway view of complete segmentation to show spatial relationship among three anatomical structures, volume segmentation of fetus, gestational sac and placenta. Our methods can successfully segment the two placentas for the twins.

The upper arm also presents to be thin and under-segmented around the elbow. It is interesting to exploit direct shape priors for guidance [20] or scale-adaptive networks for fine-grained refinement [44] for the purpose of eliminating discontinuities and enhancing local segmentation.

We also present the analyses about worst cases in segmenting gestational sac and placenta. For gestational sac segmentation (Fig. 15), our method achieves high overlap ratio with ground truth even in the worst cases. Discrepancies mainly occur around boundary where acoustic shadow often causes strong ambiguities. For placenta segmentation (Fig. 16), although with low overlap ratio, our segmentation results still present similar geometries with the ground truth. The highly varying shape and low contrast of placenta lead to the differences in thickness and shape completeness between our segmentation and ground truth, which remains as the problems we need to address in future work.

V. Conclusion

In this work, we propose the first fully automated solution for semantic segmentation in prenatal ultrasound volumes, which would potentially promote fetal health monitoring and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2018.2858779, IEEE Transactions on Medical Imaging

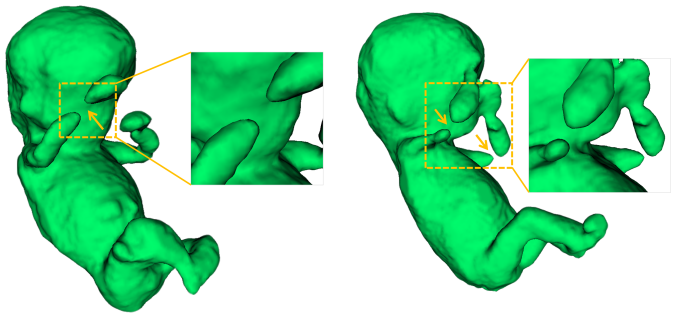


Fig. 14. Two worst cases of fetus segmentation showing the discontinuities and under-segmentation around the elbow and twist (as yellow arrow denotes).

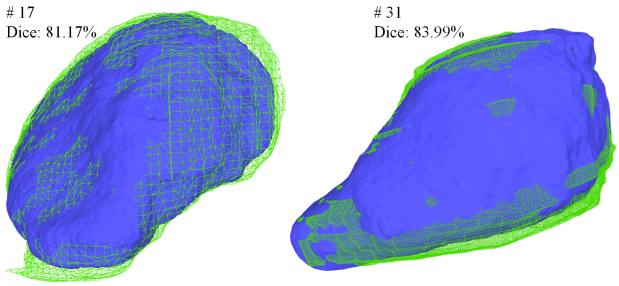


Fig. 15. Two worst cases of gestational sac segmentation. Green mesh denotes the ground truth, blue surface denotes our segmentation.

open new opportunities for many crucial clinical studies which can not be achieved with traditional planar scanning. We propose a 3D FCN enhanced with transfer learning and deep supervision to address the challenge of image quality, scale and shape variations. As complementary to the spatial concurrency exploited in 3D FCN, we propose to adapt RNN as a general module to flexibly encode varying contextual dependency and refine the corrupted predictions from a sequential perspective. Specifically, to attack the training difficulties over long sequences and consider latent sequence hierarchy with fine scales, we closely couple the adapted RNN with a hierarchical deep supervision mechanism. The efficacy of the proposed modules are extensively validated by comparing with other competitive methods. Promising quantitative (about 0.9 in Dice for gestational sac and fetus, 0.64 in Dice for placenta) and qualitative results are achieved on our large datasets. More effort will be involved to improve the segmentation, especially

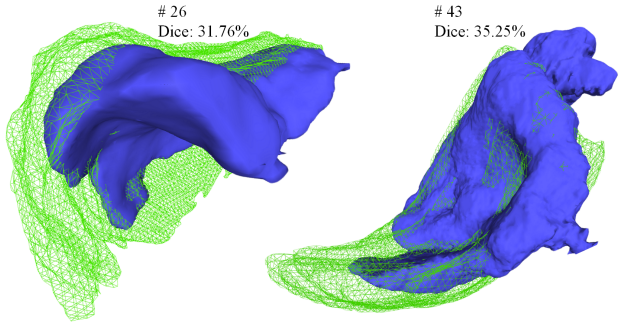


Fig. 16. Two worst cases of placenta segmentation. Green mesh denotes the ground truth, blue surface denotes our segmentation.

for the placenta. More clinical studies will be conducted in the near future, and we believe that coalescing 3D prenatal ultrasound with deep learning will bring profound change to prenatal ultrasound examination.

ACKNOWLEDGMENT

The work described in this paper was supported by a grant from National Natural Science Foundation of China under Grant 61571304, Grant 81771598 and Shenzhen Peacock Plan under Grant KQTD2016053112051497. The work was also supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project no. GRF 14225616) and a grant from Hong Kong Innovation and Technology Commission (Project no. GHP/002/13SZ). We also acknowledge all the volunteers and experts who made great effort in acquiring and annotating the precious data.

REFERENCES

[1] P. W. Callen, *Ultrasonography in Obstetrics and Gynecology E-Book*. Elsevier Health Sciences, 2011.
[2] T. Kiserud, G. Piaggio, G. Carroli *et al.*, "The world health organization fetal growth charts: a multinational longitudinal study of ultrasound biometric measurements and estimated fetal weight," *PLoS medicine*, vol. 14, no. 1, p. e1002220, 2017.
[3] D. Ni, X. Yang, X. Chen, C.-T. Chin, S. Chen, P. A. Heng, S. Li, J. Qin, and T. Wang, "Standard plane localization in ultrasound by radial component model and selective search," *Ultrasound in medicine & biology*, vol. 40, no. 11, pp. 2728–2742, 2014.
[4] L. Wu, Y. Xin, S. Li, T. Wang, P.-A. Heng, and D. Ni, "Cascaded fully convolutional networks for automatic prenatal ultrasound image segmentation," in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE, 2017, pp. 663–666.
[5] B. Tutschek *et al.*, "Three-dimensional ultrasound imaging of the fetal skull and face," *Ultrasound in Obstetrics & Gynecology*, 2017.
[6] L. E. Simcox, L. E. Higgins *et al.*, "Intraexaminer and interexaminer variability in 3d fetal volume measurements during the second and third trimesters of pregnancy," *Journal of Ultrasound in Medicine*, 2017.
[7] R. Ruano, J. Martinovic, M. Dommergues *et al.*, "Accuracy of fetal lung volume assessed by three-dimensional sonography," *Ultrasound in obstetrics & gynecology*, vol. 26, no. 7, pp. 725–730, 2005.
[8] M. Odeh, Y. Hirsh *et al.*, "Three-dimensional sonographic volumetry of the gestational sac and the amniotic sac in the first trimester," *Journal of Ultrasound in Medicine*, vol. 27, no. 3, pp. 373–378, 2008.
[9] D. Meengeonthong, S. Luewan, S. Sirichotiyakul, and T. Tongsong, "Reference ranges of placental volume measured by virtual organ computer-aided analysis between 10 and 14 weeks of gestation," *Journal of Clinical Ultrasound*, vol. 45, no. 4, pp. 185–191, 2017.
[10] N. A. Smeets *et al.*, "The predictive value of first trimester fetal volume measurements, a prospective cohort study," *Early human development*, vol. 89, no. 5, pp. 321–326, 2013.
[11] W. P. Martins, R. A. Ferriani, C. O. Nastri, and F. Mauad Filho, "First trimester fetal volume and crown-rump length: comparison between singletons and twins conceived by in vitro fertilization," *Ultrasound in medicine & biology*, vol. 34, no. 9, pp. 1360–1364, 2008.
[12] R. Aviram, D. K. Shpan, O. Markovitch, A. Fishman *et al.*, "Three-dimensional first trimester fetal volumetry: comparison with crown rump length," *Early human development*, vol. 80, no. 1, pp. 1–5, 2004.
[13] A. Gomez, K. Bhatia *et al.*, "Fast registration of 3d fetal ultrasound images using learned corresponding salient points," in *Fetal, Infant and Ophthalmic Medical Image Analysis*. Springer, 2017, pp. 33–41.
[14] A. I. Namburete, R. V. Stebbing, J. A. Noble *et al.*, "Learning-based prediction of gestational age from ultrasound images of the fetal brain," *Medical image analysis*, vol. 21, no. 1, pp. 72–86, 2015.
[15] O. Barel, R. Maymon, Z. Vaknin, J. Tovbin, and N. Smorgick, "Sonographic fetal weight estimation–is there more to it than just fetal measurements?" *Prenatal diagnosis*, vol. 34, no. 1, pp. 50–55, 2014.
[16] T. Hata, K. Kanenishi, K. Yamamoto, M. A. M. AboEllail, M. Mashima, and N. Mori, "Microvascular imaging of thick placenta with fetal growth restriction," *Ultrasound in Obstetrics & Gynecology*, 2017.

[17] A. Alansary, K. Kamnitsas *et al.*, "Fast fully automatic segmentation of the human placenta from motion corrupted mri," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 589–597.

[18] J. Anquez, E. D. Angelini, G. Grangé, and I. Bloch, "Automatic segmentation of antenatal 3-d ultrasound images," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 5, pp. 1388–1400, 2013.

[19] G. N. Stevenson *et al.*, "3-d ultrasound segmentation of the placenta using the random walker algorithm: Reliability and agreement," *Ultrasound in medicine & biology*, vol. 41, no. 12, pp. 3182–3193, 2015.

[20] S. Dahdouh, E. D. Angelini *et al.*, "Segmentation of embryonic and fetal 3d ultrasound images based on pixel intensity distributions and shape priors," *Medical image analysis*, vol. 24, no. 1, pp. 255–268, 2015.

[21] S. Feng, K. S. Zhou *et al.*, "Automatic fetal weight estimation using 3d ultrasonography," in *Proc. of SPIE Vol*, vol. 8315, 2012, pp. 83 150I–1.

[22] A. DallAsta, S. Schievano, J. L. Bruse, G. Paramasivam, C. T. Kaihura, D. Dunaway, and C. C. Lees, "Quantitative analysis of fetal facial morphology using 3d ultrasound and statistical shape modeling: a feasibility study," *American Journal of Obstetrics and Gynecology*, 2017.

[23] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3d medical image segmentation: a review," *Medical image analysis*, vol. 13, no. 4, pp. 543–563, 2009.

[24] M. Yaqub *et al.*, "Volumetric segmentation of key fetal brain structures in 3d ultrasound," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2013, pp. 25–32.

[25] M. Yaqub, M. K. Javaid, C. Cooper, and J. A. Noble, "Investigation of the role of feature selection and weighted voting in random forests for 3-d volumetric segmentation," *IEEE transactions on medical imaging*, vol. 33, no. 2, pp. 258–271, 2014.

[26] M. H. Mozaffari and W. Lee, "3d ultrasound image segmentation: A survey," *arXiv preprint arXiv:1611.09811*, 2016.

[27] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, no. 0, 2017.

[28] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[29] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[31] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, and P.-A. Heng, "3d deeply supervised network for automated segmentation of volumetric medical images," *Medical Image Analysis*, 2017.

[32] K. Li and J. Malik, "Amodal instance segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 677–693.

[33] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117.

[34] Z. Tu, "Auto-context and its application to high-level vision tasks," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[35] W. Dai, J. Doyle, X. Liang, H. Zhang, N. Dong, Y. Li, and E. P. Xing, "Scan: Structure correcting adversarial network for chest x-rays organ segmentation," *arXiv preprint arXiv:1703.08770*, 2017.

[36] J. Chen, L. Yang *et al.*, "Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation," in *Advances in Neural Information Processing Systems*, 2016, pp. 3036–3044.

[37] X. Yang *et al.*, "Towards automatic semantic segmentation in volumetric ultrasound," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 711–719.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[39] H. Chen, D. Ni *et al.*, "Standard plane localization in fetal ultrasound via domain transferred deep neural networks," *IEEE journal of biomedical and health informatics*, vol. 19, no. 5, pp. 1627–1636, 2015.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[41] D. Tran, L. Bourdev, R. Fergus *et al.*, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[42] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[43] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial Intelligence and Statistics*, 2015, pp. 562–570.

[44] X. Yang, L. Yu, L. Wu, Y. Wang, D. Ni, J. Qin, and P.-A. Heng, "Fine-grained recurrent neural networks for automatic prostate segmentation in ultrasound images." in *AAAI*, 2017, pp. 1633–1639.

[45] A. Graves *et al.*, "Hybrid speech recognition with deep bidirectional lstm," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 273–278.

[46] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki, "Scene labeling with lstm recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3547–3555.

[47] K.-L. Tseng, Y.-L. Lin, W. Hsu, and C.-Y. Huang, "Joint sequence learning and cross-modality convolution for 3d biomedical segmentation," *arXiv preprint arXiv:1704.07754*, 2017.

[48] P. Kontschieder *et al.*, "Structured class-labels in random forests for semantic image labelling," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2190–2197.

[49] Z. C. Lipton *et al.*, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2015.

[50] Y. Jia, E. Shelhamer *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[51] J. Bergstra, O. Breuleux *et al.*, "Theano: A cpu and gpu math compiler in python," in *Proc. 9th Python in Science Conf*, 2010, pp. 1–7.

[52] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[53] P. A. Yushkevich, J. Piven *et al.*, "User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[54] H.-H. Chang, A. H. Zhuang, D. J. Valentino, and W.-C. Chu, "Performance measure characterization for evaluating neuroimage segmentation algorithms," *Neuroimage*, vol. 47, no. 1, pp. 122–135, 2009.

[55] S. Rueda, S. Fathima *et al.*, "Evaluation and comparison of current fetal ultrasound image segmentation methods for biometric measurements: a grand challenge," *IEEE Transactions on medical imaging*, vol. 33, no. 4, pp. 797–813, 2014.

[56] K. Kamnitsas *et al.*, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical image analysis*, vol. 36, pp. 61–78, 2017.

[57] S. Xingjian *et al.*, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[58] L. Yu, J.-Z. Cheng, Q. Dou, X. Yang *et al.*, "Automatic 3d cardiovascular mr segmentation with densely-connected volumetric convnets," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 287–295.

[59] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 5534–5542.