Towards Automatic Semantic Segmentation in Volumetric Ultrasound

Xin Yang[1], Lequan Yu[1], Shengli Li[2], Xu Wang[3], Na Wang[3], Jing Qin[4], Dong Ni[3(✉)], and Pheng-Ann Heng[1,5]

[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
[2] Department of Ultrasound, Affliated Shenzhen Maternal and Child Healthcare Hospital of Nanfang Medical University, Shenzhen, China
[3] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China
nidong@szu.edu.cn
[4] School of Nursing, Centre for Smart Health, The Hong Kong Polytechnic University, Hong Kong, China
[5] Shenzhen Key Laboratory of Virtual Reality and Human Interaction Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

Abstract. 3D ultrasound is rapidly emerging as a viable imaging modality for routine prenatal examinations. However, lacking of efficient tools to decompose the volumetric data greatly limits its widespread. In this paper, we are looking at the problem of volumetric segmentation in ultrasound to promote the volume-based, precise maternal and fetal health monitoring. Our contribution is threefold. First, we propose the first and fully automatic framework for the simultaneous segmentation of multiple objects, including fetus, gestational sac and placenta, in ultrasound volumes, which remains as a rarely-studied but great challenge. Second, based on our customized 3D Fully Convolutional Network, we propose to inject a Recurrent Neural Network (RNN) to flexibly explore 3D semantic knowledge from a novel, sequential perspective, and therefore significantly refine the local segmentation result which is initially corrupted by the ubiquitous boundary uncertainty in ultrasound volumes. Third, considering sequence hierarchy, we introduce a hierarchical deep supervision mechanism to effectively boost the information flow within RNN and further improve the semantic segmentation results. Extensively validated on our in-house large datasets, our approach achieves superior performance and presents to be promising in boosting the interpretation of prenatal ultrasound volumes. Our framework is general and can be easily extended to other volumetric ultrasound segmentation tasks.

1 Introduction

Ultrasound imaging is a dominant modality for maternal and fetal health monitoring during pregnancy. However, traditional 2D planar ultrasound scanning is

M. Descoteaux et al. (Eds.): MICCAI 2017, Part I, LNCS 10433, pp. 711–719, 2017.
DOI: 10.1007/978-3-319-66182-7_81

implicit to inspect anatomies and thus brings about inevitable user-dependency and diagnosis error. With broad volumetric field of views, 3D prenatal ultrasound is rapidly emerging as a viable alternative. Volumetric biometrics have been proposed and attached great desire for more accurate fetal growth evaluation [11]. Versatile as it is, the widespread of 3D prenatal ultrasound is still limited due to the great lack of efficient ways to be decomposed. Semi-automatic segmentation systems, like VOCAL [14], have been applied in clinic. However, these systems often involve cumbersome interactions and result in diagnosis discrepancy. Under this situation, automated volumetric segmentation techniques are highly demanded to accurately interpret prenatal ultrasound volumes.

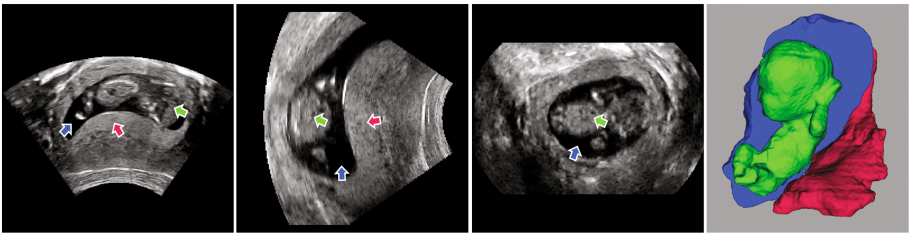

Fig. 1. From left to right: sagittal, traverse, coronal plane and a cutaway view of volumetric segmentation of a prenatal ultrasound volume. Fetus, gestational sac and placenta in planes and segmentation are denoted with green, ocean blue and red color.

As depicted in Fig. 1, simultaneously segmenting multiple objects, including fetus, gestational sac and placenta, in prenatal ultrasound volumes remains as a very arduous task. Firstly, speckle noise, acoustic shadow and low contrast between tissues conspire towards the ubiquitous boundary ambiguity and deficiency. Secondly, the spatial consistency of objects in ultrasound volume is degraded along the directions which are perpendicular to the acoustic beam. Thirdly, fetus, gestational sac and placenta present large appearance variances, highly irregular shapes and floating spatial relationships.

Utilizing tissue intensity distribution, Anquez et al. [1] made early attempt to segment utero-fetal volume unit. Stevenson et al. [15] proposed a semi-automatic method to extract placenta volume. Intensity priors exploited in these methods degrade their robustness against appearance diversity across subjects. Lee et al. [7] built boundary traces to extract limb volume for fetal weight estimation. Recently, Andrea et al. [3] explored statistical shape model to analyze fetal facial morphology. However, confined by limited training data, shape models can not tackle highly varying objects, like fetus and placenta in Fig. 1. The huge surge of deep learning [9] are taking the dominant role over traditional methods [12] for ultrasound image segmentation. However, the limited receptive field degrades the capability of deep networks, like Convolutional Neural Networks, in conquering arbitrary sized boundary incompleteness [2].

In this paper, we are looking at the problem of volumetric segmentation in prenatal ultrasound. Our contribution is threefold. First, we propose a general

framework for simultaneous segmentation of multiple complex objects in ultrasound volumes, including fetus, gestational sac and placenta, which remains a rarely-studied but great challenge currently. To the best of our knowledge, this is the first fully automatic solution in the field. Second, based on our customized 3D Fully Convolutional Network, we propose to inject a Recurrent Neural Network (RNN) to flexibly explore 3D semantic knowledge from a novel, sequential perspective and therefore significantly refine the local segmentation result. Coupled with an effective serialization strategy, our RNN proves to successfully tackle the ubiquitous boundary uncertainty in ultrasound volume. Third, to attack the gradient vanishing problem and consider the latent hierarchy in sequence, we introduce a hierarchical deep supervision mechanism (HiDS) to effectively boost the information flow within RNN and further improve the semantic segmentation. Validated on a large dataset, our approach achieves superior performance and presents to be promising in decomposing prenatal ultrasound volumes.

2 Methodology

Figure 2 is the schematic view of our proposed framework. System input is an ultrasound volume. Our customized 3D FCN firstly conducts dense voxel-wise semantic labeling and generates intermediate probability volumes for different classes. The RNN trained with hierarchical deep supervision then explores contextual information within multiple volume channels to refine the semantic labeling. System output are extracted volumes of fetus, gestational sac and placenta.

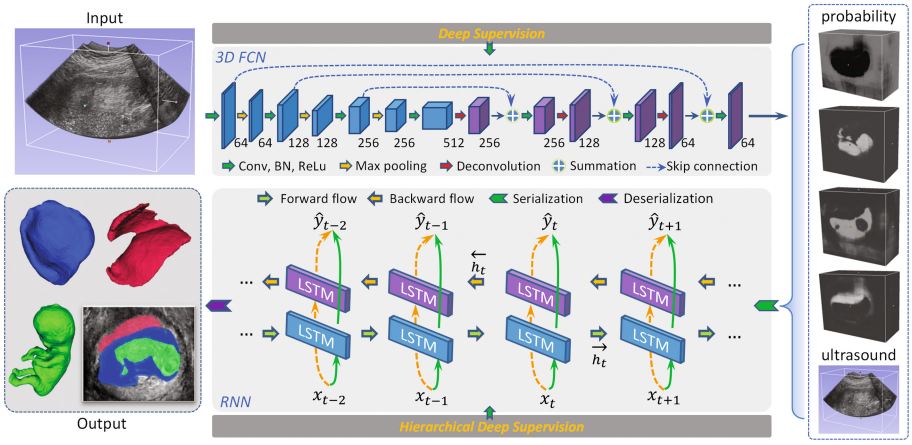

Fig. 2. Schematic view of our proposed framework. For probability volumes, from top to bottom: background, fetus, gestational sac and placenta.

2.1 Initial Dense Semantic Labeling with 3D FCN

Fully Convolutional Network (FCN) [10] is popular in semantic segmentation for its capability in end-to-end mapping. U-net [13] promotes FCN by adding skip connections to merge feature maps from different semantic levels. Skip connections are critical for network to recognize possible boundary details in ultrasound image. Since volumetric data inherently provide more complete stereo information than 2D planar images, it's also desired if the network can digest 3D data directly [4]. Therefore, as shown in Fig. 2, by equipping all layers with 3D operators, we customize a 3D FCN with long skip connections to efficiently conduct dense semantic labeling on prenatal ultrasound volumes. Specifically, we take element-wise sum operator to merge feature volumes from different resolutions and thus smooth the gradient flow. To suppress computation cost, we adopt small convolution kernels with size of $3 \times 3 \times 3$ in convolutional layers (Conv). Each Conv layer is followed by a batch normalization (BN) layer and a rectified linear unit (ReLU). 3D FCN outputs probability volumes for different classes.

2.2 Semantic Labeling Refinement with RNN

As we observe, local boundary deficiency in ultrasound volumes tend to corrupt 3D FCN's semantic predictions. Leveraging contextual information is effective in addressing boundary incompleteness. Motivated by [2,17], and being different from using traditional, fixed structures to collect context cues [16], we propose to explore Recurrent Neural Networks (RNNs) to flexibly encode contextual knowledge and refine the semantic labeling from a novel, sequential perspective. With internal memory cells, RNNs infer current timestep output by considering current input and historical information accumulated in hidden state. In our case, RNN will sequentially run over the local space, the dynamic hidden states can thus be interpreted as local contextual knowledge and be utilized to recover corrupted boundary. Our RNN is trained after the training of 3D-FCN. Shown as Fig. 2, by taking the concatenation of probability volumes and raw ultrasound volume, RNN can distill rich context information for prediction enhancement. Specifically, we propose to exploit Bidirectional Long-Short Term Memory (BiLSTM) [5] network, a popular RNN variation, in our framework to capture long range spatial dependencies and arouse interactions between sequential information flows from different directions, shown as Fig. 2. Mathematically, 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 the following equations:

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

$$\overleftarrow{h}_t = \overleftarrow{\mathcal{H}}(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}) \tag{2}$$

$$\hat{y}_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y, \tag{3}$$

where W terms denote weight matrices, h terms denote internal hidden states which are controlled by tunable gates, b terms denote bias vectors. $\overrightarrow{\mathcal{H}}$ and $\overleftarrow{\mathcal{H}}$ are

hidden layer functions. By serializing volumes into sequences and trained with cross-entropy loss function, our BiLSTM conducts direct sequence-to-sequence mapping and output the refined voxel labeling results with a *softmax* layer.

Different serialization manners differ in mining the sequentiality of volumetric data. We find that, by choosing proper size for sequence primitives, serializing a volume into a sequence of overlapping cubes can provide better capability than the slice based serialization manner in [2]. With this manner, a $50 \times 50 \times 50$ volume can be evenly divided into more than 1000 overlapped $7 \times 7 \times 7$ cubes, these cubes are then sequentially concatenated to form a sequence. Deserialization is the inverse. BiLSTM captures context cues over the long sequence and significantly refines the labeling result, and, as detailed in Sect. 2.3, we can get further improvement by coupling our RNN with a profound training mechanism.

2.3 Network-Specific Deep Supervision Mechanism

Subject to gradient vanishing issue, the parameter tuning processes of our 3D-FCN and RNN are at high risks of low efficiency and overfitting. In this paper, we propose a network-specific deep supervision strategy to facilitate the system training. For the 3D FCN part, we adopt the deep supervision strategy introduced in [4,6], which promotes training by exposing shallow convolutional layers to the direct supervision of $\mathcal{M}$ auxiliary classifiers. The final loss function for our deeply supervised 3D FCN is formulated as Eq. 4, where $\mathcal{X}$, $\mathcal{Y}$ are training pairs, $\mathcal{W}$ is the weight of main network. $w = (w^1, w^2, .., w^m)$ are the weights of auxiliary classifiers, α_m is the corresponding ratio in final loss. $m = 2$ in this paper. Cross entropy is used as a metric for main loss $\mathcal{L}$ and auxiliary $\mathcal{L}_m$.

$$\mathcal{L}(\mathcal{X}, \mathcal{Y}; \mathcal{W}, w) = \mathcal{L}(\mathcal{X}, \mathcal{Y}; \mathcal{W}) + \sum_{m \in \mathcal{M}} \alpha_m \mathcal{L}_m(\mathcal{X}, \mathcal{Y}; \mathcal{W}, w^m) + \lambda(||\mathcal{W}||^2) \quad (4)$$

Hierarchical Deep Supervision for RNN. Although BiLSTM has gating functions to guide gradient flow, it's nontrivial for BiLSTM to effectively tune gates and parameters for early timesteps. BiLSTM may be over-tuned to fit latter part of sequences for convergence, especially when tackling sequences with extreme length ($\geq$1000), which is exactly our case. Traditional training strategy for RNN is to attach a loss function at the end of the chain, and rare studies have been reported for deep supervision mechanisms in RNN. The target label replication strategy proposed in [8] is intractable for our sequence-to-sequence mapping task. A proper deep supervision strategy for RNN should consider the following two facts: (i) auxiliary supervision should be injected in early timesteps to shorten the gradient backpropagation path; (ii) the locations to trigger auxiliary supervision should consider latent, hierarchical context dependencies in the sequence. Rooting in these thoughts, we propose a novel, hierarchical deep supervision mechanism to boost the training efficiency and generalization of RNN, denoted as HiDS, shown in Fig. 3(a). Sharing same anchor point, with a main loss function for whole sequence, HiDS attaches auxiliary loss functions along the sequence with gradually increasing scopes. Equation 5 illustrates the

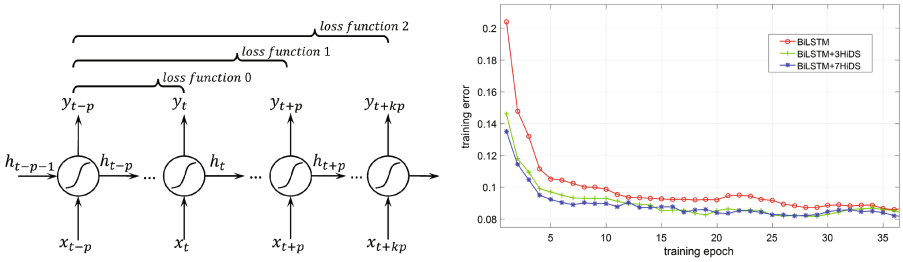

Fig. 3. Illustration of the hierarchical deep supervision mechanism for RNNs.

final loss function with HiDS, where X, Y are input and output sequences with length T and $T = \mathcal{N}p$. W is weight matrix of RNN shared by all timesteps. $\mathscr{L}_N$ is the main loss function charging the complete sequence, $\mathscr{L}_n$ are auxiliary loss functions, β_n are the associated ratio in final loss $\mathscr{L}$. $\beta_n = 1$ in this paper.

$$\mathscr{L}(\mathrm{X}, \mathrm{Y}; W) = \mathscr{L}_N(\mathrm{X}, \mathrm{Y}; W) + \sum_{n=1}^{\mathcal{N}-1} \beta_n \mathscr{L}_n(\mathrm{X}_{1 \le t < np}, \mathrm{Y}_{1 \le t < np}; W) \tag{5}$$

Figure 3(b) provides a proof about cross entropy based HiDS in boosting the training of BiLSTM over sequences with 1000 timesteps. BiLSTM equipped with 3, 7 auxiliary loss functions get faster convergence speeds and lower training errors than that in BiLSTM only with main loss function. Improvement in generalization ability brought by HiDS will be elaborated in Sect. 3.

3 Experimental Results

Materials and Implementation Details: We firstly built a dataset consisting of 104 anonymized prenatal ultrasound volumes acquired from 104 pregnant women volunteers with gestational age 10–14 weeks. Our dataset is the largest one reported in the field. The average size of volume is $221 \times 198 \times 283$ with a voxel size of $0.5 \times 0.5 \times 0.5$ mm. Approved by local IRB, all volumes were obtained by Mindray DC-8 ultrasound systems with integrated 3D probes. 10 experienced radiologist provided annotations for all volumes under strict quality control. The dataset is randomly split into 50, 10 and 44 volumes for training, validation and testing. We further augmented the training dataset to 150 volumes by flipping and rotation. 3D FCN is implemented in *caffe*, BiLSTM in *Theano*. Restricted by GPU memory, 3D FCN takes $64 \times 64 \times 64$ sub-volume as input. 800 internal memory cells are allocated for forward and backward branch each in BiLSTM. The input of BiLSTM are 5 sub-volumes ($50 \times 50 \times 50$) cropped in ultrasound and 4 probability volumes. These sub-volumes are firstly serialized into sequences of overlapped cubes with size $7 \times 7 \times 7$, and then flattened and concatenated, and finally input into BiLSTM step by step. We adopt sliding

Table 1. Quantitative evaluation of our proposed framework

Method	Fetal				Gestational sac				Placenta			
	DSC	Conf	Adb	Hdb	DSC	Conf	Adb	Hdb	DSC	Conf	Adb	Hdb
3D-F	0.876	0.714	0.88	8.77	0.870	0.699	1.12	10.01	0.590	−0.58	3.45	26.48
FB	0.876	0.713	0.76	8.06	0.887	0.741	0.93	9.85	0.624	−0.33	3.24	25.80
FB3Hi	0.880	0.724	0.79	8.36	0.889	0.746	0.96	10.15	**0.643**	**−0.25**	2.68	**24.54**
FB7Hi	**0.882**	**0.730**	**0.74**	**7.90**	**0.890**	**0.749**	**0.92**	**9.74**	0.622	−0.38	**2.56**	25.02

window and overlap-tiling stitching strategies to generate predictions for whole ultrasound volumes.

Quantitative and Qualitative Analysis: To consider both region and boundary similarities, we adopt 4 metrics to evaluate the proposed framework on segmentation, including Dice coefficient ($DSC = 2(A \cap B)/(A + B)$), Conformity ($Conf = (3DSC - 2)/DSC$), Hausdorff Distance of Boundaries (Hdb[mm]) and Average Distance of Boundaries (Adb[mm]). Ablation study is conducted on our framework. Deeply supervised 3D FCN is taken as a competitive baseline and denoted as 3D-F. 3D-F with basic BiLSTM for refinement is denoted as FB. FB equipped with 3, 7 HiDS auxiliary loss functions are denoted as FB3Hi and FB7Hi. Table 1 illustrates the extensive comparison between these methods in segmentation. Obvious improvement in both 3 classes occur when our modules are applied successively. It is most challenging to segment placenta, but with

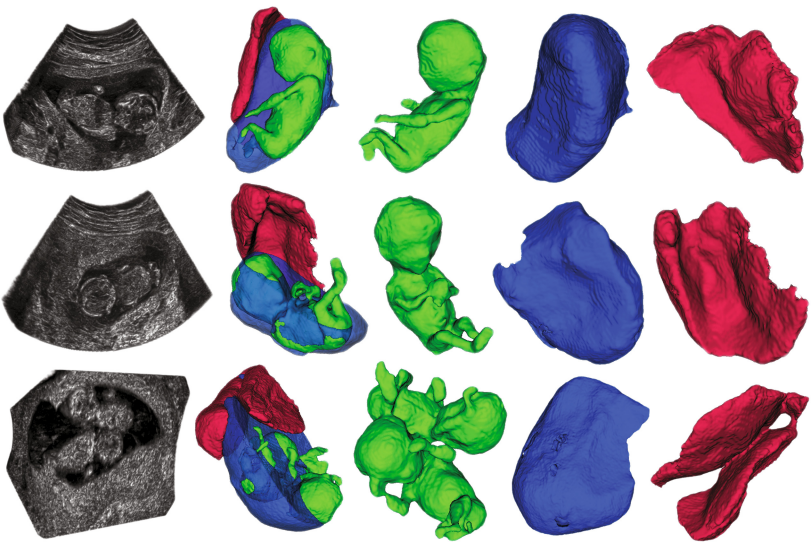

Fig. 4. From left to right: cutaway view of ultrasound volume, cutaway view of complete segmentation to show spatial relationship, volume of fetus, gestational sac and placenta.

the context information contributed by our BiLSTM, we obtained the improvement of more than 4% in DSC and 1.9 mm in Hdb. Enhancing the generalization abilities, our deep supervision mechanism HiDS boosts the segmentation performance in all metrics. Explicit visualization of 3 semantic segmentation results produced by FB7Hi for fetus, gestational sac and placenta are shown in Fig. 4. Our method conquers the poor image quality, complicated spatial configuration (even twins), boundary deficiency and spatial inconsistency in ultrasound volumes and presents smooth, promising segmentation for both 3 classes.

4 Conclusions

We present the first fully automatic framework for semantic segmentation in ultrasound volumes, which would potentially promote fetal health monitoring and open opportunities for many crucial clinical studies which can not be achieved via 2D planar ultrasound. We explore RNN to flexibly encode local contextual knowledge and therefore refine the corrupted predictions from a novel, sequential perspective. By closely coupling the RNN with a hierarchical deep supervision mechanism, the latent hierarchy in sequence is distilled to further boost segmentation performance. Promising quantitative and qualitative results are achieved on a large dataset. More clinical studies will be conducted in the near future.

Acknowledgments. The work in this paper was supported by the grant from National Natural Science Foundation of China under Grant 61571304, and grants from Hong Kong Research Grants Council (Project no. CUHK 14202514) and Hong Kong Innovation and Technology Fund (Project no. GHP/002/13SZ).

References

1. Anquez, J., Angelini, E.D., Grangé, G., Bloch, I.: Automatic segmentation of antenatal 3-d ultrasound images. IEEE TBME **60**(5), 1388–1400 (2013)
2. Chen, J., et al.: Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. In: NIPS, pp. 3036–3044 (2016)
3. DallAsta, A., et al.: Quantitative analysis of fetal facial morphology using 3d ultrasound and statistical shape modeling: a feasibility study. AJOG **217**(1), 76.e1–76.e8 (2017)
4. Dou, Q., Yu, L., et al.: 3d deeply supervised network for automated segmentation of volumetric medical images. Med. Image Anal. **S1361-8415**(17), 30072–30075 (2017)
5. Graves, A., Jaitly, N., Mohamed, A.R.: Hybrid speech recognition with deep bidirectional LSTM. In: ASRU, pp. 273–278 (2013)
6. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: Artificial Intelligence and Statistics, pp. 562–570 (2015)
7. Lee, W., et al.: Prospective validation of fetal weight estimation using fractional limb volume. UOG **41**(2), 198–203 (2013)
8. Lipton, Z.C., Kale, D.C., Elkan, C., Wetzell, R.: Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:1511.03677 (2015)

9. Litjens, G., et al.: A survey on deep learning in medical image analysis. arXiv preprint arXiv:1702.05747 (2017)
10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 (2015)
11. Mercé, L.T., Barco, M.J., Bau, S.: Three-dimensional volume sonographic study of fetal anatomy. JUM **27**(7), 1053–1063 (2008)
12. Mozaffari, M.H., Lee, W.: 3d ultrasound image segmentation: a survey. arXiv preprint arXiv:1611.09811 (2016)
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). doi:10.1007/978-3-319-24574-4_28
14. Simcox, L.E., et al.: Intraexaminer and interexaminer variability in 3d fetal volume measurements during the second and third trimesters of pregnancy. JUM **36**(7), 1291–1530 (2017)
15. Stevenson, G.N., et al.: 3-d ultrasound segmentation of the placenta using the random walker algorithm: reliability and agreement. UMB **41**(12), 3182–3193 (2015)
16. Tu, Z.: Auto-context and its application to high-level vision tasks. In: CVPR, pp. 1–8 (2008)
17. Yang, X., Yu, L., Wu, L., Wang, Y., Ni, D., Qin, J., Heng, P.-A.: Fine-grained recurrent neural networks for automatic prostate segmentation in ultrasound images. In: AAAI, pp. 1633–1639 (2017)