Multi-task CNN for structural semantic segmentation in 3D fetal brain ultrasound

L. Venturini, J.A. Noble, A.I.L. Namburete

BioMedIA Lab, Institute of Biomedical Engineering, University of Oxford

Abstract. The fetal brain undergoes extensive morphological changes throughout pregnancy, which can be visually seen in ultrasound acquisitions. We explore the use of convolutional neural networks (CNNs) for the segmentation of multiple fetal brain structures in 3D ultrasound images. Accurate automatic segmentation of brain structures in fetal ultrasound images can track brain development through gestation, and can provide useful information that can help predict fetal health outcomes. We propose a multi-task CNN to produce automatic segmentations from atlas-generated labels of the white matter, thalamus, brainstem, and cerebellum. The network as trained on 480 volumes produced accurate 3D segmentations on 48 test volumes, with Dice coefficient of 0.93 on the white matter and over 0.77 on segmentations of thalamus, brainstem and cerebellum.

1 Introduction

Fetal ultrasound scanning is a routine procedure during prenatal care, and is in many countries part of standard obstetric care. The scans are visually inspected to verify normal fetal development and to screen for disorders visible at specific gestational timepoints. These scans can be used to discern anatomical structures and track brain development [1]. Cortical structures first become visible within the fetal brain around 14 weeks of gestation and progressively develop throughout pregnancy [1].

Most studies that have analysed brain development have relied on MR imaging to perform segmentation and make quantitative measurements, due to its higher image resolution and signal-to-noise ratio [2]. However, MRI scans are relatively expensive and inaccessible, while ultrasound scans are a routine and widely available modality. Ultrasound displays artifacts which are difficult to interpret. Furthermore, due to the effects of increasing cranial ossification, the cerebral hemisphere proximal to the probe tends to be indistinct and it is difficult to discern structural boundaries [3], while the distal hemisphere is more detailed.

A number of atlases have been constructed for fetal and neonatal brains using MRI. Kuklisova-Murgasova et al [4] generated a publicly available 4D probabilistic atlas over a wider range of gestational ages ($29-44$ gestational weeks (GW)) that could be used to segment specific structures within the brain; however, this atlas was constructed using images of neonatal brains born preterm, and is therefore anatomically distinct from fetal brains. Most recently, Gholipour et al [5] proposed a 4D spatiotemporal atlas of the fetal brain spanning $19 - 39$ GW, using 3D MRI scans of fetuses and producing atlas labels of tissue type and structure.

Previous work on segmentation of fetal brain structures has proposed methods based on regression forests [6], and segmentation based on image atlases [7]. Machine-learning techniques such as convolutional neural networks (CNNs) can learn to distinguish important boundaries and artifacts and are increasingly popular in the segmentation of fetal ultrasound images [8], as they can learn to disregard some of the artifacts presented by ultrasound imaging and independently learn important segmentation features. Ronneberger et al have developed the U-net [9], a CNN architecture for the segmentation of biomedical images.

We propose a machine learning-based method for automated segmentation of multiple fetal brain structures. We implement a CNN structure based on the U-net structure to perform multiple segmentations on 3D ultrasound volumes. To the best of our knowledge, this is the first work that demonstrates a CNN-based segmentation of individual fetal brain structures in 3D ultrasound. This is also the first work to demonstrate that segmentation in ultrasound can be achieved using a network trained exclusively on auxiliary generated labels.

2 Methods

2.1 Network design

A 3D encoder-decoder network architecture based on the U-net architecture was used to perform multi-task segmentation. The size of the network was limited by memory constraints, so the top-level layer learned 16 3×3×3 feature maps. To satisfy memory constraints, the V-net architecture [10] was used. A softmax activation function was used at the output of the final convolutional layer to classify each voxel. The output was a five-class segmentation $\mathbb{Y} \in \mathbb{R}^{n \times N_x \times N_y \times N_z \times 5}$where n is the number of volumes, and all volumes have dimensions $N_x \times N_y \times N_z$. Segmentation maps for the thalamus, white matter, brainstem, cerebellum and background were generated[1]. Multi-label Dice coefficient, the sum of the Dice coefficients of all classes, was used as the loss function, as this led to what visually appeared to be the best results. Multi-label Dice is given by

$$DSC_{ml} = \sum_i \frac{2\,(GT_i \cap Seg_i)}{GT_i + Seg_i}$$

where GT and Seg are mappings of voxels corresponding to the ground truth and generated segmentation, respectively. The other parameters for the network's training were replicated from Milletari et al's V-net study [10] , but ReLU activation functions were used instead of PReLU for simplicity.

[1] Due to the size of this network, we used a batch size of 1 volume for training.

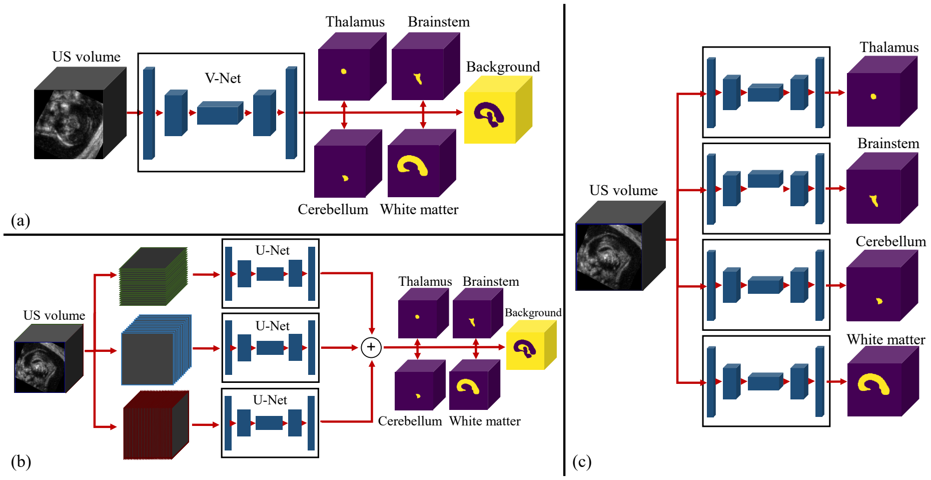

Fig. 1. The different network pipelines proposed. (a) The proposed 3D multi-task architecture, based on V-net. (b) A 2D multi-task framework, based on U-net, with QuickNAT-style merging of the different views. (c) A 3D single-task architecture, where a different network is trained per structure.

The validation set was comprised of eight volumes per gestational week (for a total of 48 volumes). The remaining 480 volumes were used for training. A similar, single-task version of this network was also implemented for comparison. The architecture was identical, but the final layer was given a sigmoid activation function, similar to the original U-net architecture. Another technique which was found to slightly improve performance further was the application of a simple morphological operation (a $3\times3\times3$ morphological closing followed by an opening) to the resulting segmentation. This operation removes any small gaps from the segmentation, and weakly enforces smoothness near the edges of the segmentation. The edges are where the trained network shows the most uncertainty in its segmentations: figure 3 shows that the most misclassifications occur near the edge of the tissues of interest.

A classical 2D U-net architecture was also implemented for comparison. This network took 2D slices as input, and output a segmentation map for each slice. The segmentations of each slice were then stacked to obtain a full 3D semantic segmentation. To incorporate contextual information from other views, the data was sliced in 3 different ways, corresponding to the 3 canonical views, in a strategy similar to QuickNAT [11]. Each 2D network outputs “soft” segmentation masks for each structure, with each voxel given a value between 0 and 1 for each structure corresponding to the network’s confidence. Combining the output of each network could exploit 3D information for segmentation, and therefore lead to a better accuracy than networks trained on individual views. Each network’s output for every voxel was averaged and a threshold was applied to obtain a joint segmentation.

A comparison of all proposed network architectures can be seen in Figure 1. All training was done on an Nvidia GeForce GTX 1080 GPU. The CNN converged to its highest Dice coefficient after 20 epochs, and after training each new volume could be segmented in 250ms.

A large ultrasound dataset is available from the INTERGROWTH-21st study, a longitudinal multi-centre study [12] that collected data from optimally healthy pregnancies carried to term. The data obtained from this study was used to provide a dataset to train and test a CNN-based solution.

The volumes used in this investigation are all of healthy fetuses between 20 and 25 weeks' gestation. A total of 528 3D ultrasound volumes were selected within this age range, based on a visual inspection of the anatomy and the subjective visibility of brain structures of interest within each scan. This narrow range of gestational ages is of particular interest, as women have a routine ultrasound scan at 20 weeks of gestation, and sulci and gyri in the cortex become visible around this time in pregnancy [1].

All volumes were manually cropped to include just the cranium, and rotated to a canonical reference space [13]. Each brain was centered and resampled to a $160 \times 160 \times 160$ volume, with the mean voxel sampled at $0.6 \times 0.6 \times 0.6$ mm. The hemisphere distal to the ultrasound probe is always more detailed than the proximal hemisphere due to interactions between the concave skull and the ultrasound signal, but the acquisition protocol for this data was agnostic to which hemisphere would be more visible.

2.2 Label generation

Given the size of this dataset and the visual artifacts and subject-specific characteristics inherent in ultrasound imaging, it is challenging and time-consuming for human experts to manually segment this dataset. We used an atlas-based method to generate a large amount of weak labels to compensate for this.

Gholipour et al [5] recently proposed a 4D spatiotemporal atlas of the fetal brain spanning $21 - 37$ GW, using 3D MRI scans of fetuses and producing atlas labels of tissue type and structure. This atlas can achieve segmentation quality comparable to human experts based on Dice coefficient [5]. This atlas was used to generate auxiliary labels for this dataset, similar to what was done by Guha Roy et al [11] for the segmentation of brain structures in MRI with limited annotations.

To propagate the atlas labels to individual ultrasound volumes, a mask of the skull was manually fitted to each ultrasound volume: since the skull is a strong ultrasound reflector and has a predictable ellipsoidal shape, this could be done quickly. Registration based on a similarity transform (comprising translation, rotation and scaling) was then performed to find the transformation between the skull mask of an age-matched atlas and the manually labeled skull in each volume. The atlas-based segmentations of four structures, namely the thalamus, brainstem, cerebellum, and white matter were generated in this way. These structures were chosen because they are large and can be seen in ultrasound acquisitions: some, such as the cerebellum, are also inspected as part of routine

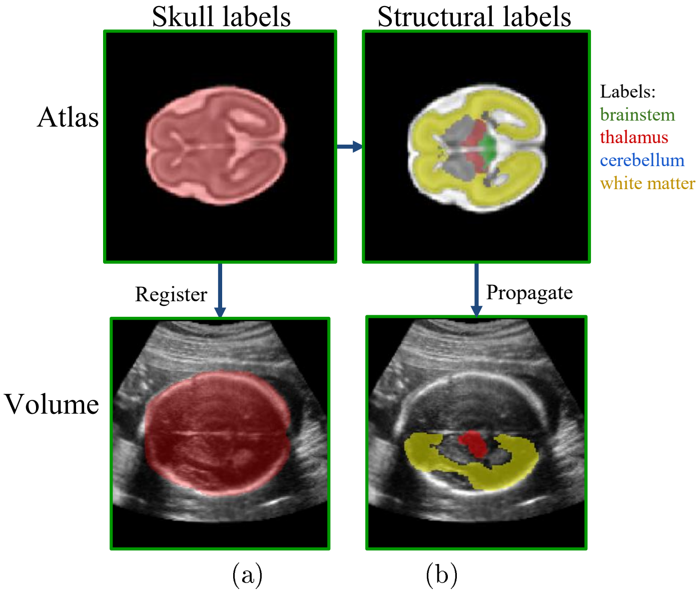

Fig. 2. The pipeline used to generate segmentation labels from the MRI atlas. The skull was segmented in each volume, a similarity transform - based registration was performed to find the correspondence, and then the structural labels were propagated.

clinical scans [14]. The transformation was applied to each of those structures, using nearest-neighbor interpolation to adjust to the new coordinate system. A schematic of the atlas-based segmentation framework can be seen in Figure 2.

Since only the hemisphere distal to the ultrasound probe can be seen in any detail, for structures that extend far from the midsagittal plane (the white matter and the thalamus) only the label distal to the probe was segmented. The cerebellum and brainstem do cross the midsagittal plane, so the entire label was segmented.

3 Results

Each single view was trained for 20 epochs. After this the network showed a tendency to overfit and reduce validation accuracy.

Table 1 shows the improvement in performance when doing multi-task segmentation compared to a single-task framework for each brain structure studied. For every brain structure, the multi-task segmentation framework performed better, with a mean improvement in Dice coefficient improvement of more than 33% over identical network trained with the same data on single-task segmentation. This is a substantial performance improvement, likely due to the fact that the brain structures analysed are spatially near to each other and often share anatomical boundaries, meaning that the same features are useful to extract

Network	DSC	ED (mm)	HD (mm)
Thalamus			
3D multi-task	**0.811 ± 0.061**	2.17 ± 1.35	**3.80 ± 1.95**
3D single-task	0.708 ± 0.070	2.82± 1.65	4.16 ± 2.39
2D	0.664 ± 0.081	**2.09 ± 1.74**	4.18 ± 2.87
Brainstem			
3D multi-task	**0.820 ± 0.081**	2.09 ± 1.26	**4.14 ± 1.29**
3D single-task	0.723 ± 0.098	1.96 ± 1.76	5.47 ± 2.37
2D	0.716 ± 0.066	**2.07 ± 1.95**	4.95 ± 3.20
Cerebellum			
3D multi-task	**0.773 ± 0.149**	2.42 ± 1.32	4.20 ± 2.39
3D single-task	0.689 ± 0.165	2.20 ± 1.72	4.42 ± 1.66
2D	0.681 ± 0.089	**2.15 ± 1.96**	**3.78 ± 2.77**
White matter			
3D multi-task	**0.921 ± 0.033**	2.27 ± 1.46	**5.93 ± 2.28**
3D single-task	0.865 ± 0.036	2.32 ± 1.72	5.90 ± 2.04
2D	0.819 ± 0.040	**2.23 ± 1.89**	14.40 ± 8.21

Table 1. Segmentation performance of single-task and multi-task segmentation architectures, as measured by Dice coefficient (DSC), Euclidean distance of the centres of mass (ED) and Hausdorff distance (HD). Across measures and brain structures, the multi-task architecture outperforms the single-task network.

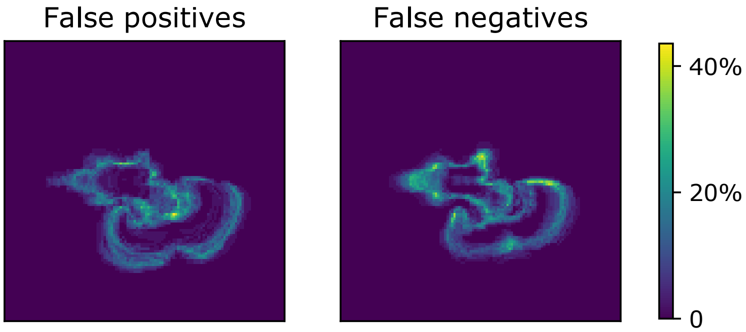

Fig. 3. A schematic showing the position of false negatives and false positives at a given axial slice for this data.

them. A richer training label effectively increases the amount of training data available, by providing important contextual information [15]. We expect that with larger training datasets, this difference should therefore decrease.

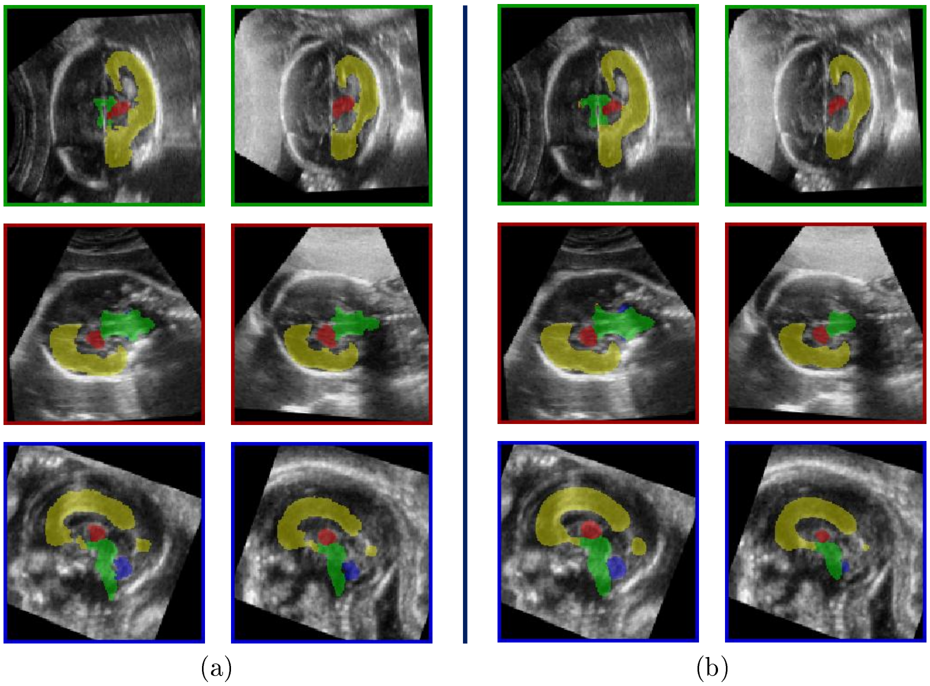

Fig. 4. (a) the atlas-generated labels used to train the CNN. (b) the resulting predictions on the same volumes (from the test set).

It is notable that segmentation of smaller structures, such as the thalamus, results in a significantly lower Dice coefficient than segmentation of the white matter on the same network. This can be explained by their differing physical characteristics: the thalamus is physically much smaller than the white matter label. In the dataset used, the white matter typically has a volume 15 times greater than the thalamus at 20 weeks, and 20 times greater at 24 weeks. The Dice coefficient is therefore biased by the larger number of interior voxels that can be predicted with high confidence, compared to voxels near the surface for which classification is more uncertain.. On the other hand, measures such as the Hausdorff distance are lower on smaller structures, showing that the overall subjective segmentation quality is similar across all structures.

Some examples of the resulting segmentations can be seen in Figure 4. Where anatomical features are clearly visible in the ultrasound image, such as the boundaries of the white matter near the skull, the CNN appears to improve on the atlas-based labels: this is expected, as (beyond gestational age and skull shape) the atlas-based labels do not take individual variation into account. On the other hand, in regions where the ultrasound image is poor or subject to shadowing artifacts, such as the base of the medulla, the CNN appears to perform worse than the atlas.

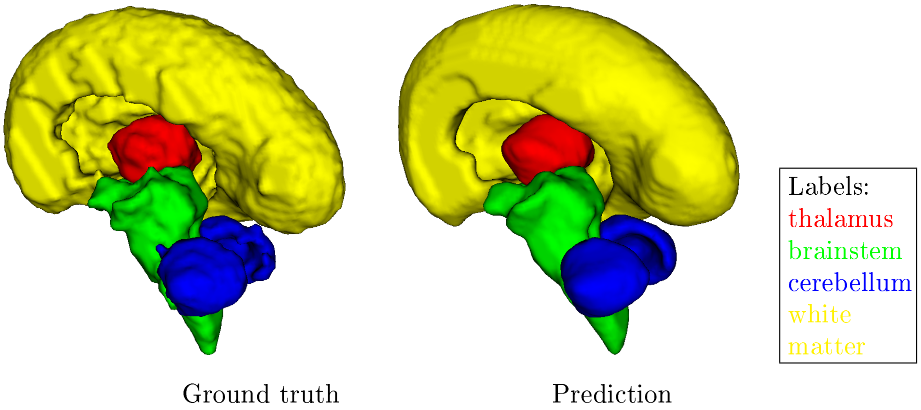

Fig. 5. Comparison of the visual appearance in 3D of the atlas-based ground truth labels and the prediction for a volume.

Visually, the prediction seems to be significantly smoother than the atlas-based ground truth labels used for training, as seen in Figure 5. This is likely due to the roughness of the original atlas-based segmentation: since nearest-neighbour interpolation is necessary, aliasing artifacts are likely to be introduced into the image. The resulting learned images, while smoother, do also appear to lose some of the detail available.

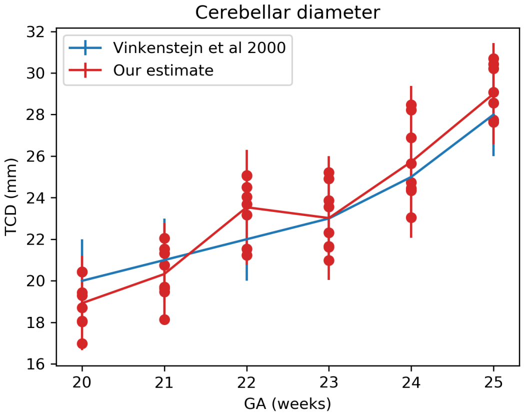

Fig. 6. Estimates of lengths, such as the transcerebellar diameter (TCD) derived from our data are in general agreement with the literature [14].

It is also possible to compare the measurements we obtained to previous results in the literature. Figure 6 shows the transcerebellar diameter (TCD), a clinical biomarker often measured in scans [14]. Our proposed method finds segmentations with lengths that seem to be in agreement with others that have previously been done.

4 Conclusion

In this paper, we obtained multi-task segmentation maps of several brain structures from 3D ultrasound acquisitions, using only coarse atlas-based segmentations for training. The results show that a CNN can learn to segment these structures even from weak labels, and visually improve on the quality of the segmentation. A multi-task segmentation framework was also proposed that improves on the performance of a similar single-task network, and we showed that a natively 3D architecture outperforms a 2D architecture. The methods developed here are an interesting proof of concept, showing that this problem can be tackled with the proposed approach.

Acknowledgment

This work is supported by funding from the Engineering and Physical Sciences Research Council (EPSRC) and Medical Research Council (MRC) [grant number EP/L016052/1].

Bibliography

[1] Pistorius, L.R., Stoutenbeek, P., Groenendaal, F., De Vries, L., Manten, G., Mulder, E., Visser, G.: Grade and symmetry of normal fetal cortical development: a longitudinal two-and three-dimensional ultrasound study. Ultrasound in Obstetrics & Gynecology **36**(6) (2010) 700–708

[2] Studholme, C.: Mapping Fetal Brain Development In Utero Using Magnetic Resonance Imaging: The Big Bang of Brain Mapping. Annual Review of Biomedical Engineering **13**(1) (2011) 345–368

[3] Namburete, A.I.L., Stebbing, R.V., Kemp, B., Yaqub, M., Papageorghiou, A.T., Noble, J.A.: Learning-based prediction of gestational age from ultrasound images of the fetal brain. Medical image analysis **21**(1) (2015) 72–86

[4] Kuklisova-Murgasova, M., Aljabar, P., Srinivasan, L., Counsell, S.J., Doria, V., Serag, A., Gousias, I.S., Boardman, J.P., Rutherford, M.A., Edwards, A.D., Others: A dynamic 4D probabilistic atlas of the developing brain. NeuroImage **54**(4) (2011) 2750–2763

[5] Gholipour, A., Rollins, C.K., Velasco-Annis, C., Ouaalam, A., Akhondi-Asl, A., Afacan, O., Ortinau, C.M., Clancy, S., Limperopoulos, C., Yang, E., Others: A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth. Scientific Reports **7** (2017)

[6] Yaqub, M., Cuingnet, R., Napolitano, R., Roundhill, D., Papageorghiou, A.T., Ardon, R., Noble, J.A.: Volumetric segmentation of key fetal brain structures in 3D ultrasound. In: International Workshop on Machine Learning in Medical Imaging, Springer (2013) 25–32

[7] Habas, P.A., Kim, K., Corbett-Detig, J.M., Rousseau, F., Glenn, O.A., Barkovich, A.J., Studholme, C.: A spatiotemporal atlas of MR intensity, tissue probability and shape of the fetal brain with application to segmentation. Neuroimage **53**(2) (2010) 460–470

[8] Schmidt-Richberg, A., Brosch, T., Schadewaldt, N., Klinder, T., Cavallaro, A., Salim, I., Roundhill, D., Papageorghiou, A.T., Lorenz, C.: Abdomen Segmentation in 3D Fetal Ultrasound Using CNN-powered Deformable Models. In: Fetal, Infant and Ophthalmic Medical Image Analysis. Springer (2017) 52–61

[9] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI, Springer (2015) 234–241

[10] Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. (2016)

[11] Guha Roy, A., Conjeti, S., Navab, N., Wachinger, C.: QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy. NeuroImage **186** (2019) 713–727

[12] Papageorghiou, A.T., Ohuma, E.O., Altman, D.G., Todros, T., Ismail, L.C., Lambert, A., Jaffer, Y.A., Bertino, E., Gravett, M.G., Purwar, M., Noble, J.A., Pang, R., Victora, C.G., Barros, F.C., Carvalho, M., Salomon, L.J., Bhutta, Z.A., Kennedy, S.H., Villar, J.: International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. The Lancet **384**(9946) (2014) 869–879

[13] Namburete, A.I., Xie, W., Yaqub, M., Zisserman, A., Noble, J.A.: Fully-automated alignment of 3D fetal brain ultrasound to a canonical reference space using multi-task learning. Medical Image Analysis **46** (2018) 1–14

[14] Vinkesteijn, A., Mulder, P., Wladimiroff, J.: Fetal transverse cerebellar diameter measurements in normal and reduced fetal growth. Ultrasound in Obstetrics and Gynecology **15**(1) (2000) 47–51

[15] Moeskops, P., Wolterink, J.M., van der Velden, B.H., Gilhuijs, K.G., Leiner, T., Viergever, M.A., Išgum, I.: Deep learning for multi-task medical image segmentation in multiple modalities. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Volume 9901 LNCS. Springer, Cham (2016) 478–486