

# Automated 3-D Ultrasound Image Analysis for First Trimester Assessment of Fetal Health

Hosuk Ryou<sup>1</sup>, Mohammad Yaqub<sup>1</sup>, Angelo Cavallaro<sup>2</sup>, Aris T. Papageorghiou<sup>2</sup> & J. Alison Noble<sup>1</sup>

<sup>1</sup> Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK

<sup>2</sup> Nuffield Department of Women's and Reproductive Health, University of Oxford, UK

E-mail: hosuk.ryou@eng.ox.ac.uk

July 2019

**Abstract.** The first trimester fetal ultrasound scan is important to confirm fetal viability, to estimate the gestational age of the fetus, and to detect fetal anomalies early in pregnancy. First trimester ultrasound images have a different appearance than for the second trimester scan, reflecting the different stage of fetal development. There is limited literature on automation of image-based assessment for this earlier trimester, and most of the literature is focused on one specific fetal anatomy. In this paper, we consider automation to support first trimester fetal assessment of multiple fetal anatomies including both visualization and the measurements from a single 3-D ultrasound scan. We present a deep learning and image processing solution i) to perform semantic segmentation of the whole fetus, ii) to estimate plane orientation for standard biometry views, iii) to localize and automatically estimate biometry, and iv) to detect fetal limbs from a 3-D first trimester volume. Computational analysis methods were built using a real-world dataset (n=44 volumes). An evaluation on a further independent clinical dataset (n=21 volumes) showed that the automated methods approached human expert assessment of a 3D volume.

*Keywords:* 3-D Fetal Ultrasound, Fetal Health Assessment, Fetal Biometry, First Trimester Scan, Convolutional Neural Networks, Deep Learning

## 1. Introduction

Fetal growth can be affected by many factors such as nutrition, smoking, genetic factors and the social environment (WHO 2005). Due to these factors, it is important to monitor the fetus during a pregnancy. This is especially important during early fetal development. In the first trimester of pregnancy the miscarriage rate is higher than for later trimesters. Typically, this is also the stage when a medical professional confirms fetal viability and begins to monitor fetal growth rate (Salomon et al. 2013). However, most of the fetal anatomies are early in development in the first trimester and for this reason, the traditional first trimester assessment mostly focus on the detection of chromosomal defects/genetic syndromes whereas anatomical anomaly assessment is performed at a second trimester ultrasound scan. If more pregnancy problems could be picked up earlier via a more comprehensive first trimester ultrasound examination, there is the potential to reduce cost within a healthcare system. In this paper, we consider this possibility by considering how to automate biometry from a 3-D ultrasound first trimester fetal scan. With our approach, all the sonographer needs to do is acquire a 3D scan following a simple standard acquisition guideline. Hence, our solution aims to speed up assessment for skilled sonographers, and is potentially suitable for non-expert sonographers.

Sonographers use ultrasound imaging to visualize anatomical planes for fetal anatomical assessment and biometric measurement in second and third trimester scans. However, this is recognized as a highly skilled task, and the quality of ultrasound decision-making is heavily dependent on sonographer skill. Even when a sonographer is trained to a high level, there can be significant inter- and intra-observer acquisition variability (Dudley & Chapman 2002, Sarris et al. 2012). This variability can potentially lead to a mis-diagnosis (for instance, inaccurate estimation of fetal growth) or worse, that conditions are missed. Measurement inaccuracy in turn affects the accuracy of estimation of fetal weight. For instance in a comprehensive systematic review, it was estimated that the size of the 95% confidence intervals of random errors exceed 14 % of birth weight (Dudley 2005). Further, that paper claimed that measurement methods and observer variability are major contributors to systematic and random error; standardization and excellent training are necessary. To reduce this dependency, there is active current interest in medical image analysis research to support automated clinical decision-making (Yaqub et al. 2014, Ryou et al. 2016, Yang et al. 2017).

We briefly place our work in the context of related literature that has considered automated image analysis of 3-D fetal ultrasound volumes. The most closely related work has considered automatic detection of fetal anatomies. In our earlier work, (Ryou et al. 2016), we developed consecutive Random Forest models to localize the whole fetus from a 3-D first trimester volume. Based on this localized whole fetus, a *Convolutional Neural Network* (CNN) model classified each 2-D slice extracted from a volume into head, torso and non-fetal tissue classes. However, that method assumed that the orientations of the head and abdomen plane with the acceptable quality are in the

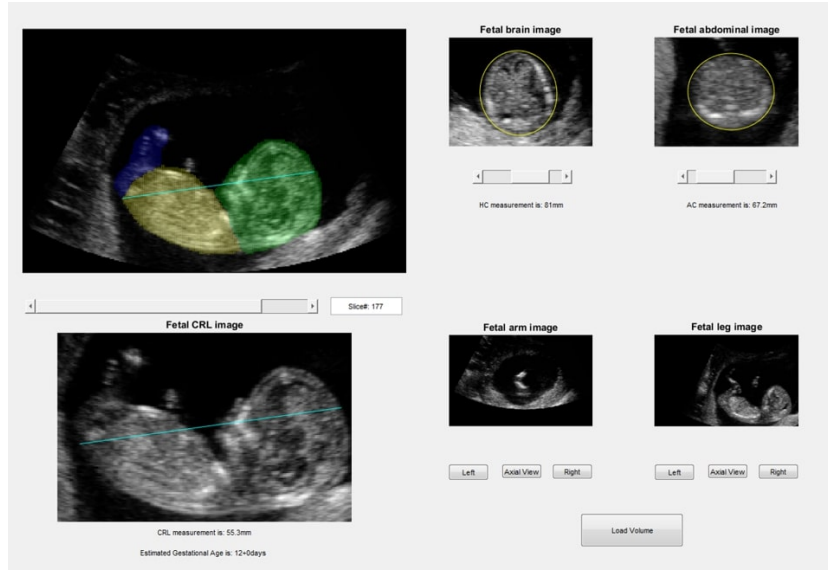


Figure 1: The integrated software interface for automated biometry assessment.

same orientation as the fetus. Most other related work has considered assessment of a single fetal anatomy (Yaqub et al. 2012, Yaqub et al. 2014, Nie et al. 2017, Chen et al. 2015). For example, (Yaqub et al. 2012) proposed a Random Forests framework to detect local brain structures in 3-D ultrasound fetal brain images. That approach was extended to segment the fetal femur from 3-D ultrasound fetal femur volumes (Yaqub et al. 2014). Other methods have considered segmentation of fetal structures using a Fully Convolutional Network (FCN) (Yang et al. 2017, Raynaud et al. 2017). For example, (Yang et al. 2017) performed 3-D segmentation of the whole fetus, gestational sac and the placenta from first trimester fetal volumes using a 3-D FCN as the initial segmentation of each part and an Recurrent Neural Network (RNN) to refine the segmentation result. Although this interesting work proposed volumetric segmentation of the whole fetus, only whole fetus segmentation was performed, but not segmentation of each fetal part. (Raynaud et al. 2017) used an FCN combined with morphological filters for fetal spine segmentation to automatically align the 3-D ultrasound volume to localize multiple organs such as the heart, the stomach, the umbilical vein and the bladder. However, that work only considered fetal abdominal volumes.

In this paper, we describe original automated image analysis algorithms developed to assist first trimester fetal assessment from a single 3-D ultrasound scan. The developed automated image analysis methods provide a means to visualize key fetal anatomy such as the head, abdomen and fetal limbs, and to perform automated biometry as illustrated in figure 1.

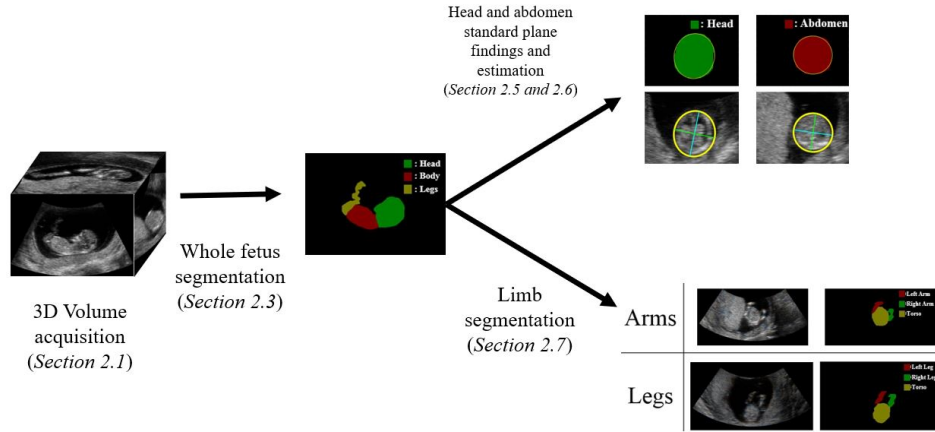


Figure 2: The overall components of the proposed method. A slice of the whole fetus in the sagittal view is automatically extracted. Automated segmentation of the whole fetus into the head, the abdomen and the lower limbs is then performed. Based on the detected region of the whole fetus, the segmentation of the head, the abdomen and the individual limbs in the axial view is performed. Finally, from the segmentation results, the automated measurement of the head and the abdomen is performed.

## 2. Materials and Methods

During the first trimester fetal ultrasound scan, anatomical assessment aims to check whether there are any fetal structural abnormalities based on visual appearance and to confirm the presence of internal structures. According to International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) guidelines (Salomon et al. 2013), there are a number of anatomical structures to check in the first trimester scan. For the purpose of this study, we have focused on the whole fetus profile as observed from the crown-rump-length (CRL) view, and the head, the abdomen and the limbs from the axial view.

The overall components of the proposed method are shown in figure 2. A slice of the whole fetus in the sagittal view (CRL view) from a 3-D volume is automatically extracted (Section 2.3). Automated segmentation of the whole fetus into the head, the abdomen and the lower limbs in sagittal view is then performed (Section 2.3). Based on the detected region of the whole fetus, the segmentation of the head, the abdomen and the individual limbs in the axial view is performed (Section 2.5 and 2.7). Finally, from the segmentation results, automated measurement of the head and the abdomen is performed (Section 2.6). Before describing the method in detail, we first describe the acquisition protocol and data preparation for analysis.

### 2.1. Acquisition Protocol

First trimester ultrasound screening is typically performed between 11 and 14 weeks of gestation and our dataset covers this range. All data was from healthy pregnancies.

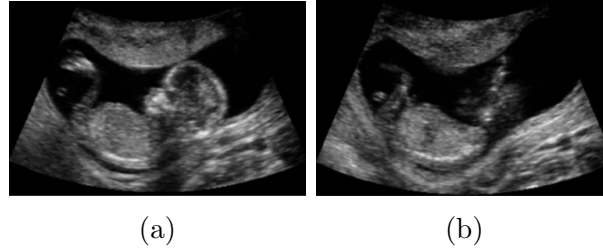


Figure 3: Typical example of a) acceptable and b) unacceptable sagittal planes. They were labelled based on whether the slice satisfied the CRL criteria or not.

The data was acquired with a Philips HD9 ultrasound machine (Bothell, WA 98021, USA) equipped with a V7-3 transducer with a voxel resolution of  $0.33\text{mm} \times 0.33\text{mm} \times 0.33\text{mm}$ . In total 65 3-D fetal volumes were acquired. Acquisitions were made following local ethical committee guidelines. This study was approved by the institutional review board and all women had written informed consent. The acquisition protocol followed the crown-rump-length (CRL) criteria:

- The fetus should be horizontal (at  $90^\circ$  to the angle of insonation) with a posterior spine position;
- The fetus should fill at least 30% of the screen;
- The crown and rump should both be clearly visible (the whole fetus must be present); and,
- The fetus should be in a neutral position (not hyperextended or flexed).

## 2.2. Data Preparation for Automated Image Analysis

The data was separated into training and test sets such that every fetal volume was included in one set only. In total 65 3-D fetal volumes were used; 44 as the training set and 21 as the test set. 2-D slices for each whole fetus (sagittal view), head, abdomen and limbs (axial view) were manually extracted from each volume. From each slice, the fetal parts were manually annotated as groundtruth. We also asked clinicians for the manual biometry measurements and to check the quality of the images to validate our automated biometry.

For CRL plane classification, extracted sagittal slices (planes) that satisfied the CRL criteria were manually classified as *acceptable*. Sagittal slices that did not satisfy the CRL criteria were considered *unacceptable*. Figure 3 shows typical acceptable and unacceptable examples. Whole fetus segmentation was performed in images labelled as acceptable sagittal slices. In this case, the image was manually annotated into 4 classes: head, torso, lower limb and background as illustrated in figure 4.

Manual annotation was also performed in other sagittal slice images that contained a clear visualization of one of the classes. This is because such an image still contains important fetal features but if annotated as background, can confuse the training algorithm.

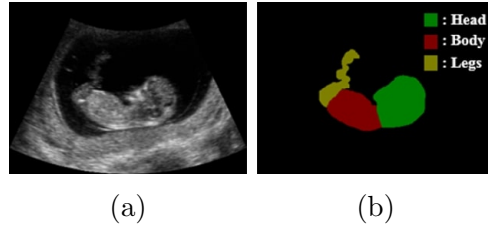


Figure 4: Example original images (column a) and corresponding labels (column b).

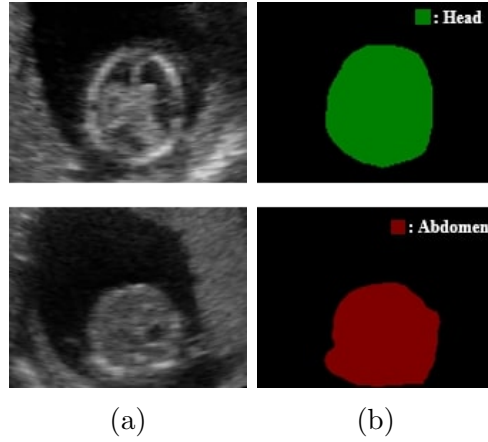


Figure 5: Typical examples of head and abdomen automatic annotation; a) Original image, and b) corresponding labels.

The head and abdomen slices were first cropped based on the segmentation of the whole fetus in the sagittal plane following (Ryou et al. 2016) and then manually labelled as shown in figure 5. Since the whole fetus had already been localized, the unrelated non-fetal parts could be removed as shown in figure 5 (a). This allowed the analysis to focus on the fetal parts. For training, the image was first cropped based on a manual segmentation of the whole fetus. Then, separately, each head and each abdomen were manually annotated. However, note that for testing, the image was automatically cropped following (Ryou et al. 2016).

Compared to the head and the abdomen, the arms and legs have significantly greater variance in appearance due to their articulation (variable pose). Therefore, instead of cropping the image as for the head and abdomen, the original, full size image was used in both the training and test set. Both limbs and the torso were manually annotated as shown in figure 6. Data augmentation was used to increase training data size and help the trained model cope with variabilities that are not in the original training set. In this case, data augmentation was performed by flipping left-right and the random image rotation about the centre ( $-30^\circ$  to  $+30^\circ$ ). For the limbs, however, only image rotation was performed since in our work, the left and right limbs are considered different classes for limb segmentation. Intensity normalization was performed on both training and test data prior to automated analysis.

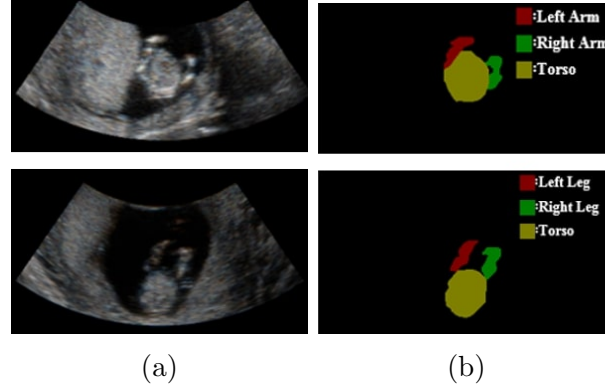


Figure 6: The example of (a) axial slice of each left and right arms and legs, and torso with (b) corresponding labels.

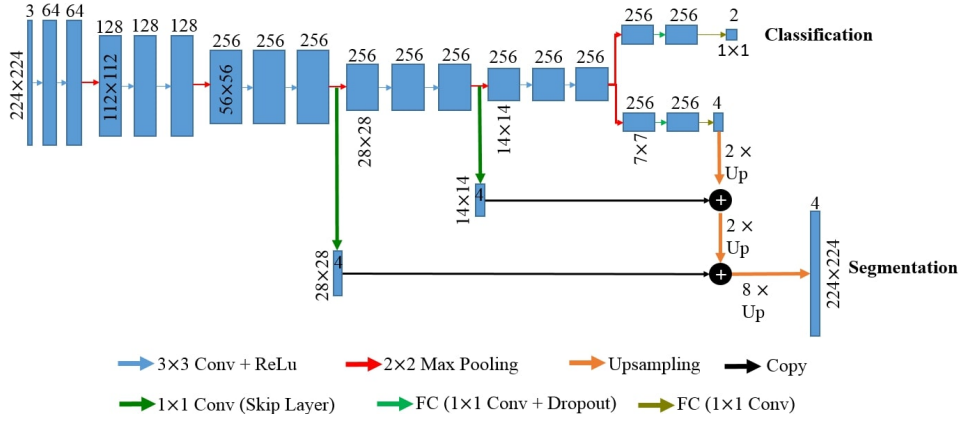


Figure 7: The multi-task FCN architecture for classification and segmentation of the whole fetus in the sagittal view. The fully connected layers (FC) are implemented to perform both classification and segmentation.

### 2.3. Whole Fetus Segmentation Algorithm

The first step in our approach is to perform the automatic segmentation and partitioning of the whole fetus. To do this, we developed a Fully Convolutional Network (FCN) architecture (Long et al. 2015). A basic FCN architecture is similar to that of a Convolutional Neural Network (CNN), but it allows prediction of the class for each pixel which makes it suitable for image segmentation. For the basic FCN architecture, a VGG-16 networks (Simonyan & Zisserman 2014) was adopted. Specifically, we proposed the solution as a *Multi-task Network* that outputs both image plane classification and whole-fetus segmentation predictions as illustrated in figure 7.

The classification component of the multi-task network estimates the position of the sagittal plane of the whole fetus based on whether the whole fetus is visualized in the image or not. This classification network uses the same images that are used for whole fetus segmentation in the sagittal view. Thus, the classification and segmentation networks have shared features.

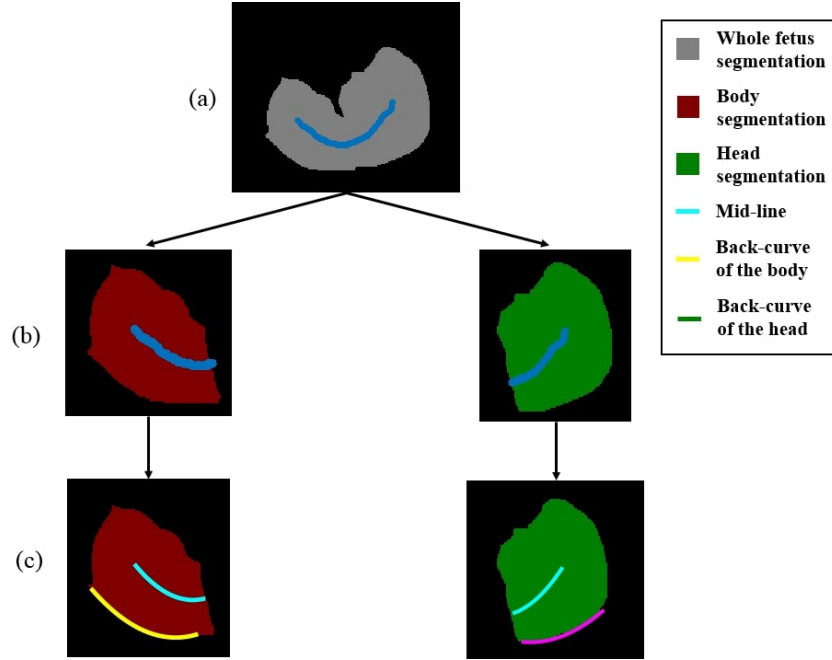


Figure 8: Schematic summarizing how head and torso orientation are automatically determined. (a) Based on automatic segmentation, the mid-line of the whole fetus is obtained using the skeletonization transform. (b) The whole fetus is automatically partitioned into the head and the torso. (c) Polynomial fitting is applied to each skeletonization to get the mid-line head and torso curves. The back-curve (the boundary of the backside of each region), is also obtained. The orientation of the head and the torso are calculated as the average slope of the mid-curve and back-curve in each case.

Note that in this work, we used a 2-D FCN-based architecture applied to slices of a 3-D volume rather than a 3-D FCN model as we were constrained by data availability. 3-D scans are not typically acquired in the first trimester. The number of available clinical 3-D datasets was too small for a 3-D FCN implementation even with data augmentation.

#### 2.4. Estimation of Head and Abdomen Standardized Plane Orientations

To obtain a good visualization and measurement of each anatomical structure, the image plane needs to be estimated accurately. However, since the fetal 3-D volume is acquired based on whole fetus criteria, the extracted anatomical plane from a 3-D volume may be sub-optimal relative to one obtained by a 2-D acquisition. To assist automatic analysis, a method to estimate the orientations of the head and the abdomen has been developed. Since the segmentation of the head and torso in the sagittal view have been performed, the orientation of each part can be approximated by the mid-line of each structure and the back of the whole fetus based on the segmented region as shown in figure 8.

To obtain the mid-line, skeletonization is used (Telea & van Wijk 2002). The mid-line is extracted in two steps. First, the whole fetus segmentation is skeletonized to obtain the mid-line of the whole fetus. Second, the whole fetus segmentation



is partitioned into the head and torso, and a second-order polynomial curve fit is performed on each skeleton. However, since the mid-line is obtained from (possibly noisy) skeletonized points of the whole fetus, it might not accurately represent the mid-line of the head and body. To increase robustness, the back boundary of the head and torso is also extracted. A second order polynomial is fit to the boundary of the back of the fetus to define the back-curve. Then, the slopes of each mid-line curve and back-curve are averaged to estimate the final orientations of the head and the torso, respectively.

### 2.5. Diagnostic Biometry Plane Estimation for the Head and Abdomen

In this work, we assumed that the positions of the diagnostic plane of the head and the abdomen with respect to the fetal head and torso length are similar among the normal fetus of similar gestational age. Therefore, the head and abdomen fetal biometry planes are estimated based on a linear regression of the head and torso length following (Ryou et al. 2016). Specifically, to estimate the head plane, a linear regression is performed to predict the distance (in mm) of the head plane from the approximate fetal crown as a function of the length of the head. A second linear regression is performed for the abdominal plane from the approximate fetal rump as a function of the length of the torso.

### 2.6. Biometric Measurement and Gestational Age Estimation

Three ultrasound-based-biometric measurements are computed in this work: Crown-Rump-Length (CRL), Head Circumference (HC) and Abdomen Circumference (AC). The Gestational Age (GA) is derived from the CRL measurement using a parametric equation. The **Crown-Rump-Length (CRL)** is the length between two points placed at the crown and the rump respectively. To estimate these two points, the rump, for example, is the farthest point of the segmented region of the lower part of the torso after correcting the orientation.

For the head circumference and the abdomen circumference, the head and abdomen are segmented using the FCN algorithm similar to the architecture shown in figure 7 except the classification branch. Then an ellipse is fit to the boundary of each segmented region using a least-squares fitting method by having the minimum sum of the squares of the distances to the given points. After the ellipse parameters are estimated, the longest and shortest outer-to-outer diameter is estimated. Typical examples of automatic HC and AC measurement are shown in Figure 9. The **Head Circumference (HC)** can be calculated from the Biparietal Diameter (BPD), the shortest radius, and Occipitofrontal Diameter (OFD), the longest radius, using equation 1,

$$HC = \frac{\pi \times (BPD + OFD)}{2} \quad (1)$$

The **Abdomen Circumference (AC)** measurement is calculated from the *Antero-posterior Abdominal Diameter* (APAD), the shortest radius, and *Transverse Abdominal*

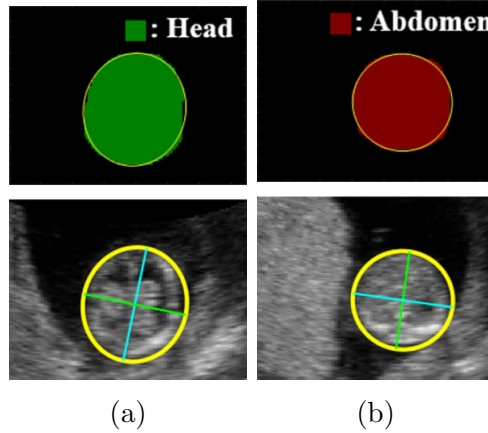


Figure 9: Typical examples of automatic measurement of (a) the HC and (b) the AC. When the segmentation result is obtained, fit the ellipse to the segmented region. The short and long axis of the resulted ellipse is then obtained to calculate the circumferences.

*Diameter* (TAD), the longest radius, using equation 2,

$$AC = \frac{\pi \times (APAD + TAD)}{2} \quad (2)$$

**Gestational Age (GA)** is calculated from the CRL using a standard formula(Papageorghiou et al. 2014),

$$GA(Days) = 40.9041 + 3.21585 \times CRL^{0.5} + 0.348956 \times CRL \quad (3)$$

### 2.7. Diagnostic Plane Extraction for the Limbs

Automatic limb detection is quite challenging. In this work, we assumed that the location of the limbs can be found approximately, based on the location of the head and torso using the anatomical constant that the arms are located in the upper torso and the legs are located at the end of the torso. For limb segmentation, we utilize a U-Net (Ronneberger et al. 2015). The U-Net is similar to a FCN, but merges the features from the higher layers with the features from the lower layers by using *skip connections*. In our application, this helps the network to recognize boundary details which is important in identifying limbs as they are small and well-characterized by boundary definition. The architecture used for limb segmentation in this work is shown in figure 10. *Batch normalization* (BN) is employed to reduce *internal covariate shift* and to encourage a higher learning rate (Ioffe & Szegedy 2015).

Based on the limb segmentation, the orientation of each of the four limbs (both arms and legs) is estimated following the method shown in figure 11. Since not all limbs can be viewed in a single slice, the segmented regions from each slice are stacked (figure 10(a)). Then by using the averaged class probability of each of the stacked regions, the region pixels that receive more than 50% probability score are extracted as the final segmentation of the limbs (figure 10(b)). With this segmentation region, the mid-line is

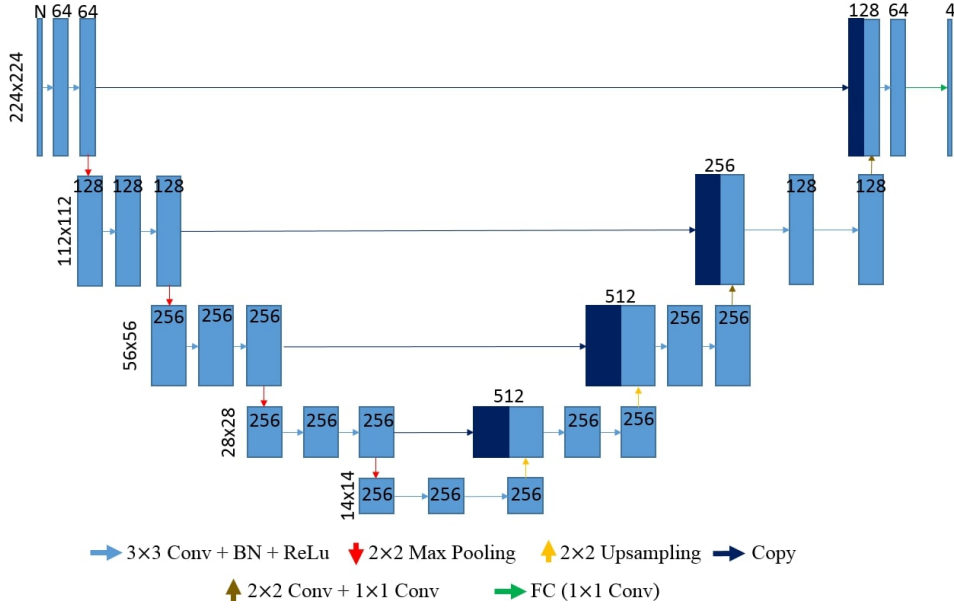


Figure 10: The U-Net architecture for the limb segmentation. Compare to the FCN-8s shown in figure 8, a U-Net combines all the previous layers. The differences between FCN-8s and U-Net are that FCN-8s sums the predictions whereas a U-Net concatenates the features from the previous layers.

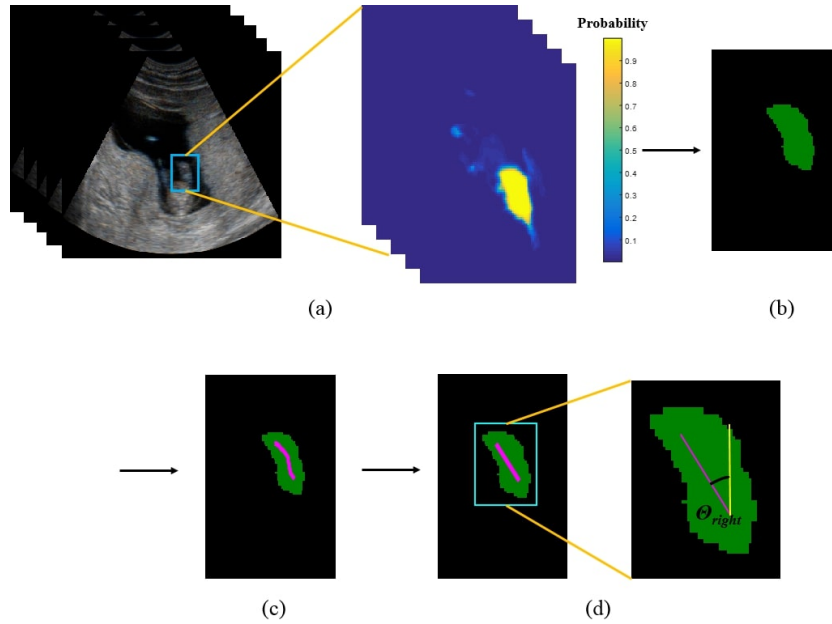


Figure 11: The process of determining the orientation for the right leg. (a) Based on the probability map for the segmentation of the right leg, the pixels with the averaged probability larger than 50% are extracted. (b) Based on the segmentation results, (c) use skeletonization transform to get the mid-line of the right arm. (d) Linear polynomial fitting is then applied to obtain the orientation angle.

obtained using the skeletonization transform (figure 10(c)). Finally, using polynomial curve fitting, the mid-line of the segmentation region and the orientation are obtained (figure 10(d)).

## 2.8. Implementation Details

The image processing and deep learning algorithms were developed in *Matlab 2016* (The MathWorks, Inc., Natick, Massachusetts, United States). All the networks follow the *VGG-16* architecture. MatConvnet (Vedaldi & Lenc 2015) was used to train all networks and training and testing were performed on a Computer Processing Unit (CPU) (Intel Xeon CPU at 3.50 GHz with 16.0 GB RAM). For FCN-8s, the parameters were first learned by training FCN-32s, and then fine-tuned while training FCN-8s (Long et al. 2015). All FCN parameters were empirically chosen (filter size, number of channels, stride and max pool size). The *stochastic gradient descent (SGD)* and *softmax loss* were used in all FCNs in this study. An initial learning rate of  $10^{-3}$  was chosen for all the FCNs and then gradually decreased. A higher learning rate ( $10^{-2}$ ) was chosen for the U-Net due to the batch normalization (Ioffe & Szegedy 2015).

*2.8.1. Algorithm Performance Metrics.* Pixel accuracy and  $IoU_{seg}$  were used as performance metrics for the FCN. Pixel accuracy defines how many pixels are correctly classified compared to the ground truth and  $IoU_{seg}$  defined how much the automatic results overlap the ground truth.

## 2.9. Network Architecture Variants

To validate and justify network architectures, a number of experiments were conducted. For whole fetus segmentation, we compared three different networks: FCN-32s with the original number of nodes, FCN-8s with reduced nodes, and a multi-task FCN-8s. For head and abdomen segmentation, we first considered separate FCNs for head and abdomen segmentation. However, it is also possible to have one FCN to simultaneously perform both segmentation tasks. Thus three FCNs were implemented and compared: FCN-8s for head segmentation, FCN-8s for abdomen segmentation and FCN-8s for both head and abdomen segmentation. For limb segmentation, we compared two different networks: FCN-8s and U-Net.

## 2.10. Manual versus Automatic Comparison

Comparison was made between the automatic method (A) and results obtained by two experts ( $M_1$  and  $M_2$ ). Specifically, we computed  $(A-M_1)$ ,  $(A-M_2)$  and the inter-observer variability ( $M_2-M_1$ ).

Table 1: Evaluation of the whole fetus segmentation based on the pixel accuracy.

	Mean Accuracy (%)	Torso (%)	Head (%)	Lower Limb (%)	BG (%)
FCN-32s	85.8±13.6	91.7	90.8	65.6	95.1
FCN-8s	88.6±11.7	91.7	92.5	71.6	98.6
Multitask	89.4±11.4	91.9	94.2	72.9	98.8

### 2.11. Plane Estimation Accuracy Metric

Two metrics were used to validate the accuracy of estimation of the plane localization and orientation: (1) the distance (in mm) between the centres of the extracted planes; and (2) the angle difference (in degrees) between the manually and automatically chosen angle. The distance between planes and the angle differences quantify how far the centre of the two planes are from each other. The smaller the distance and angle, the higher the accuracy.

### 2.12. Qualitative Evaluation of Extracted Planes

As a qualitative evaluation, two experts were blindly given the manually and automatically extracted planes for each fetal part from each data. They were asked to decide whether the presented image was suitable for anatomical assessment based on the ISUOG guidelines (Salomon et al. 2013).

## 3. Results

In this section, we summarize experimental results to demonstrate the performance of the proposed methods of whole fetus segmentation, head and abdomen segmentation, plane extraction, measurement and limb detection.

### 3.1. Whole Fetus Classification and Segmentation

Tables 1 and 2 summarize the pixel accuracy and  $IoU_{seg}$ . For whole fetus classification and segmentation, the multi-task FCN-8s achieved the best performance with a pixel accuracy of 89.4% and  $IoU_{seg}$  of 0.76. It also achieved 98.9% slice classification accuracy.

### 3.2. Diagnostic Biometry Plane Estimation for the Head and Abdomen

By performing the linear regression as mentioned in Section 2.5, the best planes for the head and abdomen were defined by,

$$Plane_{Head} = 0.30 \times Head\ Length + 0.19 \quad (4)$$

$$Plane_{Abdomen} = 0.50 \times Torso\ Length + 0.01 \quad (5)$$

Table 2: Evaluation of the whole fetus segmentation based on IoU.

	Mean IoU	Torso	Head	Lower Limb	BG
FCN-32s	0.58±0.33	0.59	0.62	0.15	0.95
FCN-8s	0.74±0.24	0.79	0.81	0.41	0.98
Multitask	0.76±0.23	0.80	0.82	0.43	0.98

Table 3: Segmentation results of the head and abdomen: comparing the performance between different FCNs.

	Head pixel acc	Head IoU	Abdomen pixel acc	Abdomen IoU
FCN (Head)	95.4	0.90		
FCN (Abdomen)			96.6	0.94
FCN (Both)	95.0	0.92	96.3	0.94

acc = accuracy (%)

Table 4: Performance of arm segmentation assessed by pixel accuracy: comparison of FCN-8s and U-Net architecture.

	Mean Accuracy (%)	Left (%)	Right (%)	Torso (%)	BG (%)
FCN-8s	80.1±15.3	67.9	66.9	87.5	98.2
U-Net	85.0±12.7	72.3	76.5	92.1	99.2

with  $R^2$  values 0.72 and 0.89, respectively.

### 3.3. Segmentation of the Head and the Abdomen

Table 3 reports the segmentation accuracy and  $IoU_{seg}$  of the three different networks for the head and abdomen segmentation task.

### 3.4. Limb Segmentation

Tables 4-7 report the segmentation accuracy of FCN and U-Net architectures for the limb segmentation task. The U-Net for arm and leg segmentation achieved a pixel accuracy of 85.0% and 88.3%, and  $IoU_{seg}$  of 0.70 and 0.63, respectively.

### 3.5. Automatic-Manual and Manual-Manual

The average distance and the average orientation difference between automatically and manually extracted planes are reported in Table 8 and Table 9. The average distances between A-M<sub>1</sub> for the head and abdomen are 1.98mm ±1.19 and 2.51mm ±1.58 where

Table 5: The performance of arm segmentation and the comparison based on overall mean IoU with standard deviations. IoU for left, right arms and torso are also presented.

	Mean IoU	Left	Right	Torso	BG
FCN-8s	0.57±0.32	0.32	0.32	0.67	0.98
U-Net	0.70±0.24	0.52	0.50	0.81	0.99

Table 6: Performance of leg segmentation assessed by pixel accuracy: comparison of FCN-8s and U-Net architecture.

	Mean Accuracy (%)	Left (%)	Right (%)	Torso (%)	BG (%)
FCN-8s	83.4±11.9	71.4	76.9	86.7	98.6
U-Net	88.3±8.0	81.8	82.6	90.0	99.1

Table 7: Evaluation of the whole fetus segmentation based on IoU.

	Mean IoU	Left	Right	Torso	BG
FCN-8s	0.60±0.27	0.43	0.42	0.56	0.99
U-Net	0.63±0.25	0.45	0.46	0.63	0.99

Table 8: The averaged distance between A-M<sub>1</sub>, A-M<sub>2</sub> and M<sub>2</sub>-M<sub>1</sub> (A: Automatically extracted plane, M<sub>1</sub> and M<sub>2</sub>: Manually extracted plane by two experts). The averaged distance between both A-M<sub>1</sub> and A-M<sub>2</sub> have shown similar results compared to M<sub>2</sub>-M<sub>1</sub> in head and abdomen plane extraction.

Comparison	Head			Abdomen		
	A-M <sub>1</sub>	A-M <sub>2</sub>	M <sub>2</sub> -M <sub>1</sub>	A-M <sub>1</sub>	A-M <sub>2</sub>	M <sub>2</sub> -M <sub>1</sub>
AD (mm)	1.98 ± 1.19	2.43 ± 0.94	2.11 ± 1.06	2.51 ± 1.58	2.34 ± 1.60	1.92 ± 1.69

AD = Averaged Distance

M<sub>2</sub>-M<sub>1</sub> are 2.11mm ±1.06 and 1.92mm ±1.69, respectively. Also, as shown in Table 9, A-M<sub>1</sub> orientation differences are 9.45±7.34 and 4.32 ±2.42 for the head and abdomen, respectively.

The difference in biometry estimation is compared in Table 10. A-M<sub>1</sub> is 6.03mm ±3.62 and 3.34mm ±2.84 whereas M<sub>2</sub>-M<sub>1</sub> is 1.00mm ±3.43 and -1.40mm ±4.22 for the HC and AC, respectively. For the CRL, A-M<sub>1</sub> is -1.65mm ±5.21 and M<sub>2</sub>-M<sub>1</sub> is -0.93mm ±1.96. GA estimation differences is within 1 day of the manual prediction.

Table 9: Orientation differences in A-M<sub>1</sub>, A-M<sub>2</sub> and M<sub>2</sub>-M<sub>1</sub>. Note that the orientation differences are larger in M<sub>2</sub>-M<sub>1</sub> for the head than for the abdomen. This might show that the head plane extraction is more subjective than the abdomen plane extraction.

Comparison	Head			Abdomen		
	A-M <sub>1</sub>	A-M <sub>2</sub>	M <sub>2</sub> -M <sub>1</sub>	A-M <sub>1</sub>	A-M <sub>2</sub>	M <sub>2</sub> -M <sub>1</sub>
OD (°)	$9.45 \pm 7.34$	$20.30 \pm 9.39$	$20.13 \pm 12.65$	$4.32 \pm 2.42$	$5.93 \pm 4.78$	$7.12 \pm 5.13$

OD = Orientation Difference

Table 10: The averaged difference in the measurements: A-M<sub>1</sub>, A-M<sub>2</sub> and M<sub>2</sub>-M<sub>1</sub>. The automatic measurements for both HC and AC are larger than the corresponding manual measurement (as shown in A-M<sub>1</sub> and A-M<sub>2</sub>) whereas in M<sub>2</sub>-M<sub>1</sub>, the difference between two manually extracted planes is small. The CRL and GA, however, showed similar results.

	A-M <sub>1</sub>	A-M <sub>2</sub>	M <sub>2</sub> -M <sub>1</sub>
HC (mm)	$6.03 \pm 3.62$	$5.04 \pm 3.15$	$1.00 \pm 3.43$
AC (mm)	$3.34 \pm 2.84$	$4.57 \pm 4.34$	$-1.40 \pm 4.22$
CRL (mm)	$-1.65 \pm 5.21$	$-0.69 \pm 4.38$	$-0.93 \pm 1.96$
GA (days)	$-0.91 \pm 2.83$	$-0.41 \pm 2.49$	$-0.51 \pm 1.08$

### 3.6. Assessment of Limb Detection

The detection rate for each of the limbs (left arm, right arm, left leg and right leg) are 0.65, 0.50, 0.55 and 0.35 respectively. These are relatively low but reflect the difficulty of the task.

### 3.7. Qualitative Assessment

Qualitative analysis of the head and abdomen images is reported in Table 11. 19 of 21 automatically extracted head planes and 15 of 21 abdomen planes were selected as good image planes by the clinician. The qualitative analysis of automatic limb detection is reported in Table 12. Of 21 examples in each case, 7 left-arm-planes, 5 right-arm-planes, 6 left-leg-planes and 4 right-leg-planes were selected as good image planes.

## 4. Discussion

From the results in Table 1 and 2, the FCN-8s achieved better results than the FCN-32s, even though the FCN-8s has less parameters. We explain this as follows. The prediction from the FCN-32s is up-sampled from the lower resolution-prediction. Because of this, the boundary of the prediction appears coarse after up-sampling. Compared to the FCN-32s, the FCN-8s gradually up-samples the prediction and combines the output of



Table 11: The qualitative analysis of head and abdomen plane detection. As explained in Section 2.12, two experts selected manually and automatically planes that are suitable for anatomical assessment. Both automatically and manually extracted head plane obtained similar results which verifies that the automated method can extract the head plane with a good visualization for the anatomical assessment. Automatically extracted abdomen plane also showed promising results.

	$Head_{Auto}$	$Head_{Manual}$	$Abdomen_{Auto}$	$Abdomen_{Manual}$
QA (n=21)	19	20	15	19

QA = Qualitative Assessment

Table 12: The qualitative assessment for the limbs. Both arms and legs have shown low qualitative analysis results due to the low detection rate of left arm, right arm, left leg and right leg (0.65, 0.50, 0.55 and 0.35, respectively).

	Auto	Manual
Left Arm (n=21)	7	17
Right Arm (n=21)	5	13
Left Leg (n=21)	6	21
Right Leg (n=21)	4	16

lower layers to make local predictions. This yields a finer boundary. Because of this, both pixel accuracy and  $IoU_{seg}$  are higher for the FCN-8s architecture. Further, observe from Table 1 and 2 that the multi-task network achieved slightly better accuracy and  $IoU_{seg}$  than FCN-8s. This improvement might be due to the fact that the multi-task network is fine-tuned further from FCN-8s. This might also indicate that learning the classification task provided additional relevant features to have better performance in the segmentation task. More importantly, the multi-task network can perform both tasks simultaneously which potentially indicates faster computation compared to a cascade of two single-task network. As shown in Table 3, the segmentation results show negligible difference between the FCN for two individual segmentations and each FCN for either head or abdomen segmentation. This might be due to their different internal structures image patterns, and the FCN can recognize these differences. For limb segmentation, the U-Net produced better segmentation results compared to the FCN-8s as shown in Table 4-7. Our hypothesis for this better performance is that the U-Net was able to utilize information from the lower layers to improve segmentation accuracy. As mentioned before, using the U-Net architecture is more beneficial for the limb segmentation since the limb is small and requires finer boundary details compared to the head and the abdomen where FCN networks provided excellent segmentation.

As shown in Table 8, the average distances from A-M<sub>1</sub> and A-M<sub>2</sub> are similar to

those of  $M_2-M_1$ . This shows that automatic plane extraction is within the range of human inter-variability as shown in Table 9. For the head, the inter-observer difference in orientation estimation was  $20.13^\circ$  whereas for the abdomen, the difference was  $7.12^\circ$ . Note that the  $A-M_1$  variance in orientation estimation for the head is larger than for the abdomen. The characteristic internal structure of the head plane is the choroid plexus which covers a large area in the head. Hence the plane can be difficult to localize exactly. For the abdomen, however, the only characteristic is the stomach bubble, which is relatively smaller than the choroid plexus, and is more likely to be visible in only few (and possible one) candidate planes. Future work may look to improve the accuracy of head plane extraction but this may require a 3-D analysis approach (or locally comparing slices). The differences in manual and automated estimation were small compared to the size of the corresponding fetal parts as shown in Table 10. However, both automatic HC and AC measurements were found to over-estimate compared to manual measurement. The auto-manual differences in the HC and AC are mainly caused by segmentation prediction that the true boundary around the fetal head can be unclear which may have led to over-segmentation. In this case, the ellipse fit to the segmentation was larger than the manually-placed ellipse. Since the FCN over-segments fetal parts, the automatic measurements are larger than the manual measurements. By contrast, the inter-observer difference ( $M_2-M_1$ ) is small. Unlike the HC and AC, however, the manual and computer-based CRL measurements were found to be similar as shown in Table 10. This resulted in an automatic estimation of the GA within 1 day of the manual prediction. For limb assessment, the detection rate was found to be relatively low, but results are encouraging. The qualitative analysis of the head and abdomen shows promising results as shown in table 11. Compared to the head, the abdomen is more challenging and can be difficult to visualize due to the small size of the stomach bubble. By contrast, the choroid plexus in the head is large which is easier to visualize. As shown in table 12, the limb qualitative assessment was lower than for the head and the abdomen. This is not unexpected as this task is considerably hard for sonographers. Our analysis shows that there needs to be further prior knowledge incorporated in automatic detection such as the spatial relationship between localized fetal parts in 3-D space. In clinical practice, high maternal body mass index (BMI) can affect the quality of fetal images. We have not investigated how our proposed technique perform on high BMI subjects. Fetal position is another factor known to affect image quality that would require further study in future work. A trans-abdominal transducer was used to acquire data for this work. An endovaginal ultrasound transducer may produce higher quality images than for a trans-abdominal transducer. However, there is typically a more limited field of view which may result in it being difficult to image the whole fetus in one 3-D volume.

## **Conclusions**

In this paper, we have described original methods for automatic first trimester fetal biometry assessment based on FCN-based neural networks which is, to our knowledge, the first attempt to do multiple automatic biometry on a single first trimester 3-D volume. We believe that our work provides a stepping stone towards fully automated biometry which is comparable to human measurements in 3-D fetal ultrasound for first trimester scans.

## **5. Acknowledgements**

This work was supported by the NIHR Oxford Biomedical Research Centre, EPSRC grants EP/L505316/1, EP/RO13853/1 and Innovate UK grant 101684. Aris T. Papageorghiou is supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC)

## References

- Chen, H., Dou, Q., Ni, D., Cheng, J.-Z., Qin, J., Li, S. & Heng, P.-A. (2015), Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks, *in* N. Navab, J. Hornegger, W. M. Wells & A. Frangi, eds, ‘Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015’, Springer International Publishing, Cham, pp. 507–514.
- Dudley, N. (2005), ‘A systematic review of the ultrasound estimation of fetal weight’, *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* **25**(1), 80–89.
- Dudley, N. & Chapman, E. (2002), ‘The importance of quality management in fetal measurement’, *Ultrasound in obstetrics & gynecology* **19**(2), 190–196.
- Ioffe, S. & Szegedy, C. (2015), Batch normalization: Accelerating deep network training by reducing internal covariate shift, *in* ‘Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37’, ICML’15, JMLR.org, pp. 448–456.
- Long, J., Shelhamer, E. & Darrell, T. (2015), Fully convolutional networks for semantic segmentation, *in* ‘Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition’, pp. 3431–3440.
- Nie, S., Yu, J., Chen, P., Wang, Y. & Zhang, J. Q. (2017), ‘Automatic detection of standard sagittal plane in the first trimester of pregnancy using 3-d ultrasound data’, *Ultrasound in Medicine & Biology* **43**(1), 286 – 300.
- Papageorghiou, A. T., Kennedy, S. H., Salomon, L. J., Ohuma, E. O., Cheikh Ismail, L., Barros, F. C., Lambert, A., Carvalho, M., Jaffer, Y. A., Bertino, E., Gravett, M. G., Altman, D. G., Purwar, M., Noble, J. A., Pang, R., Vitoria, C. G., Bhutta, Z. A., Villar, J., for the International Fetal & for the 21st Century (INTERGROWTH-21st), N. G. C. (2014), ‘International standards for early fetal size and pregnancy dating based on ultrasound measurement of crownrump length in the first trimester of pregnancy’, *Ultrasound in Obstetrics & Gynecology* **44**(6), 641–648.
- Raynaud, C., Ciofolo-Veit, C., Lefèvre, T., Ardon, R., Cavallaro, A., Salim, I., Papageorghiou, A. & Rouet, L. (2017), Multi-organ detection in 3d fetal ultrasound with machine learning, *in* M. J. Cardoso, T. Arbel, A. Melbourne, H. Bogunovic, P. Moeskops, X. Chen, E. Schwartz, M. Garvin, E. Robinson, E. Trucco, M. Ebner, Y. Xu, A. Makropoulos, A. Desjardin & T. Vercauteren, eds, ‘Fetal, Infant and Ophthalmic Medical Image Analysis’, Springer International Publishing, Cham, pp. 62–72.
- Ronneberger, O., Fischer, P. & Brox, T. (2015), U-net: Convolutional networks for biomedical image segmentation, *in* ‘International Conference on Medical image computing and computer-assisted intervention’, Springer, pp. 234–241.
- Ryou, H., Yaqub, M., Cavallaro, A., Roseman, F., Papageorghiou, A. & Noble, J. A. (2016), Automated 3d ultrasound biometry planes extraction for first trimester fetal assessment, *in* ‘International Workshop on Machine Learning in Medical Imaging’, Springer, pp. 196–204.
- Salomon, L., Alfrevic, Z., Bilardo, C., Chalouhi, G., Ghi, T., Kagan, K., Lau, T., Papageorghiou, A., Raine-Fenning, N., Stirnemann, J., Suresh, S., Tabor, A., Timor-Tritsch, I., Toi, A. & Yeo, G. (2013), ‘Isuog practice guidelines: performance of first-trimester fetal ultrasound scan’, *Ultrasound in Obstetrics & Gynecology* **41**(1), 102–113.
- Sarris, I., Ioannou, C., Chamberlain, P., Ohuma, E., Roseman, F., Hoch, L., Altman, D. & Papageorghiou, A. (2012), ‘Intra-and interobserver variability in fetal ultrasound measurements’, *Ultrasound in Obstetrics & Gynecology* **39**(3), 266–273.
- Simonyan, K. & Zisserman, A. (2014), ‘Very deep convolutional networks for large-scale image recognition’, *arXiv preprint arXiv:1409.1556*.
- Telea, A. & van Wijk, J. J. (2002), An augmented fast marching method for computing skeletons and centerlines, *in* ‘Proceedings of the Symposium on Data Visualisation 2002’, VISSYM ’02, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, pp. 251–ff.
- Vedaldi, A. & Lenc, K. (2015), ‘Matconvnet: convolutional neural networks for matlab’.

- WHO (2005), ‘Report on the regional consultation towards the development of a strategy for optimizing fetal growth and development’, *WHO Regional Office for the Eastern Mediterranean: Cairo*.
- Yang, X., Yu, L., Li, S., Wang, X., Wang, N., Qin, J., Ni, D. & Heng, P.-A. (2017), Towards automatic semantic segmentation in volumetric ultrasound, *in* M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins & S. Duchesne, eds, ‘Medical Image Computing and Computer Assisted Intervention MICCAI 2017’, Springer International Publishing, Cham, pp. 711–719.
- Yaqub, M., Javaid, M. K., Cooper, C. & Noble, J. A. (2014), ‘Investigation of the role of feature selection and weighted voting in random forests for 3-d volumetric segmentation’, *IEEE transactions on medical imaging* **33**(2), 258–271.
- Yaqub, M., Napolitano, R., Ioannou, C., Papageorgiou, A. T. & Noble, J. A. (2012), Automatic detection of local fetal brain structures in ultrasound images, *in* ‘2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)’, pp. 1555–1558.