Ultrasound in Medicine & Biology 50 (2024) 805–816

Contents lists available at ScienceDirect

Ultrasound in Medicine & Biology

journal homepage: www.elsevier.com/locate/ultrasmedbio

Original Contribution

Automating the Human Action of First-Trimester Biometry Measurement from Real-World Freehand Ultrasound

Robail Yasrab [a,b,*], He Zhao [a], Zeyu Fu [a], Lior Drukker [a,c], Aris T. Papageorghiou [d], J. Alison Noble [a]

[a] Department of Engineering Science, University of Oxford, Oxford, UK
[b] School of Clinical Medicine, University of Cambridge, Cambridge, UK
[c] Sackler Faculty of Medicine, Rabin Medical Center, Tel-Aviv University, Tel-Aviv, Israel
[d] Nuffield Department of Women's Reproductive Health, University of Oxford, Oxford, UK

ARTICLE INFO

Keywords:
First-trimester ultrasound
Crown–rump length
Fetal biometry estimation
Machine learning

ABSTRACT

Objective: Automated medical image analysis solutions should closely mimic complete human actions to be useful in clinical practice. However, more often an automated image analysis solution represents only part of a human task, which restricts its practical utility. In the case of ultrasound-based fetal biometry, an automated solution should ideally recognize key fetal structures in freehand video guidance, select a standard plane from a video stream and perform biometry. A complete automated solution should automate all three subactions.

Methods: In this article, we consider how to automate the complete human action of first-trimester biometry measurement from real-world freehand ultrasound. In the proposed hybrid convolutional neural network (CNN) architecture design, a classification regression-based guidance model detects and tracks fetal anatomical structures (using visual cues) in the ultrasound video. Several high-quality standard planes that contain the mid-sagittal view of the fetus are sampled at multiple time stamps (using a custom-designed confident-frame detector) based on the estimated probability values associated with predicted anatomical structures that define the biometry plane. Automated semantic segmentation is performed on the selected frames to extract fetal anatomical landmarks. A crown–rump length (CRL) estimate is calculated as the mean CRL from these multiple frames.

Results: Our fully automated method has a high correlation with clinical expert CRL measurement (Pearson's $p = 0.92$, R-squared [R^2] $= 0.84$) and a low mean absolute error of 0.834 (weeks) for fetal age estimation on a test data set of 42 videos.

Conclusion: A novel algorithm for standard plane detection employs a quality detection mechanism defined by clinical standards, ensuring precise biometric measurements.

Introduction

Worldwide, pregnant women are routinely offered a first-trimester ultrasound (US) scan to confirm pregnancy viability, establish fetal gestational age and assess chromosomal anomaly risk [1]. Sonographers carry out the first-trimester scan as a freehand fetal US examination in which they acquire a mid-sagittal (MS) imaging plane to perform measurement of the fetal crown (head)–rump (bottom) length (CRL). This measurement is usually carried out between 11^{+0} to 13^{+6} wk^{+d} of gestation. CRL is highly correlated with fetal age and is used to estimate the expected due date [2,3]. To increase the accuracy of measurement and because of frequent fetal movement, sonographers are required to perform multiple CRL measurements [4,5]. In practice, the first-trimester scan is highly operator dependent, meaning that the quality and accuracy of the scan can vary depending on the skill and experience of the person performing the scan. The interpretation of US images requires a good understanding of anatomy and experience in recognizing normal and abnormal structures. The operator needs to have a good understanding of the technology and be able to identify and interpret subtle signs, such as nuchal thickness (nuchal translucency [NT]), that are used to assess the risk of certain conditions.

Automated US guidance and fetal biometry estimation are aimed at supporting the streamline of clinical workflow and assisting in image quality assurance. Previous image analysis methods reported in the literature consider only the fetal biometry step for already selected (frozen) single-shot standard planes [6–8] and have not considered how temporal information, such as a change in fetal appearance and position over time, may benefit analysis. We describe how incorporating temporal information to estimate fetal biometry based on multiple high-quality standard planes can improve the overall quality of fetal biometry estimation.

Several studies have considered automated fetal US standard plane detection [6,9,10] or automated estimation of fetal biometry [11]. However, none of these methods have offered an end-to-end solution for the combined tasks of plane guidance, quality assurance and automated

* Corresponding author. Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, UK.
E-mail addresses: ry302@cam.ac.uk, robail.yasrab@eng.ox.ac.uk (R. Yasrab).

https://doi.org/10.1016/j.ultrasmedbio.2024.01.018
Received 26 October 2023; Revised 10 January 2024; Accepted 25 January 2024

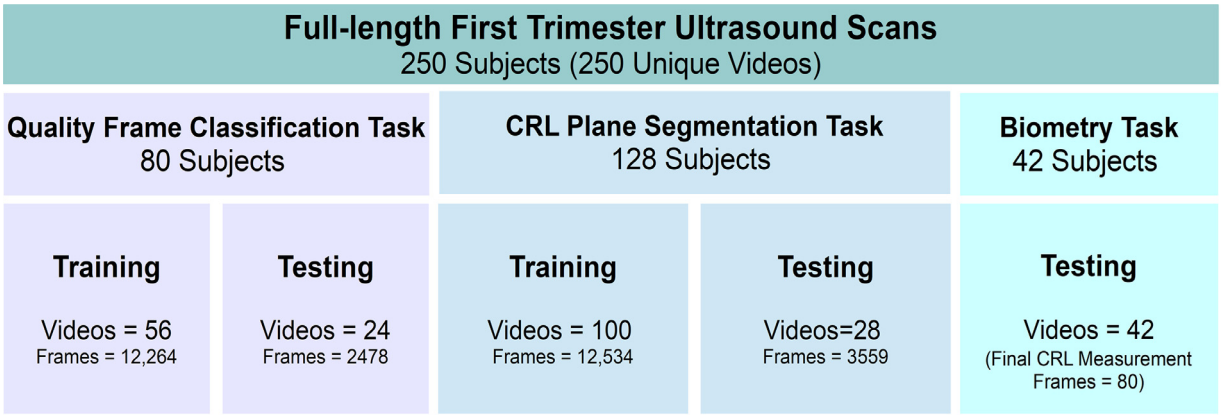

Figure 1. Data set distribution for experiments. CRL, crown–rump length.

fetal biometry assessment. Further, most of the prior literature has focused on the second- and third-trimester US scans for which it is easier to navigate between different anatomical structures. The first trimester is the least explored obstetric trimester in both the clinical and the image analysis literature possibly because of the paucity of clinical guidelines [12] and the limited availability of data for machine learning-based research.

Methodology contribution

The proposed fully automated biometry method automates the human action of first-trimester biometry from a real-world freehand US acquisition video stream. An original step of the algorithm is to take into account multiple fetal appearances and positions to select frames with a neutral fetal position. In addition, the final biometry estimate is based on estimates from multiple video frames, similar to the clinical practice of a sonographer. The design has been chosen to be suitable for real-time implementation. The proposed system could evolve toward a fully automated US biometry system by using deep learning techniques and serving as a guidance system for operators.

The article is structured as follows. Under Methods, we introduce the first-trimester US image analysis and biometry, including the US data (Fig. 1), the acquisition protocol and the steps in the proposed method. The Results, Discussion and Conclusion sections follow.

Methods

Figure 2 is an overview of the proposed multistage method:

1. The first stage is responsible for high-quality standard plane detection:

   (a) Figure 2a illustrates the first module in which freehand US video frames are input into a quality frame classifier to predict

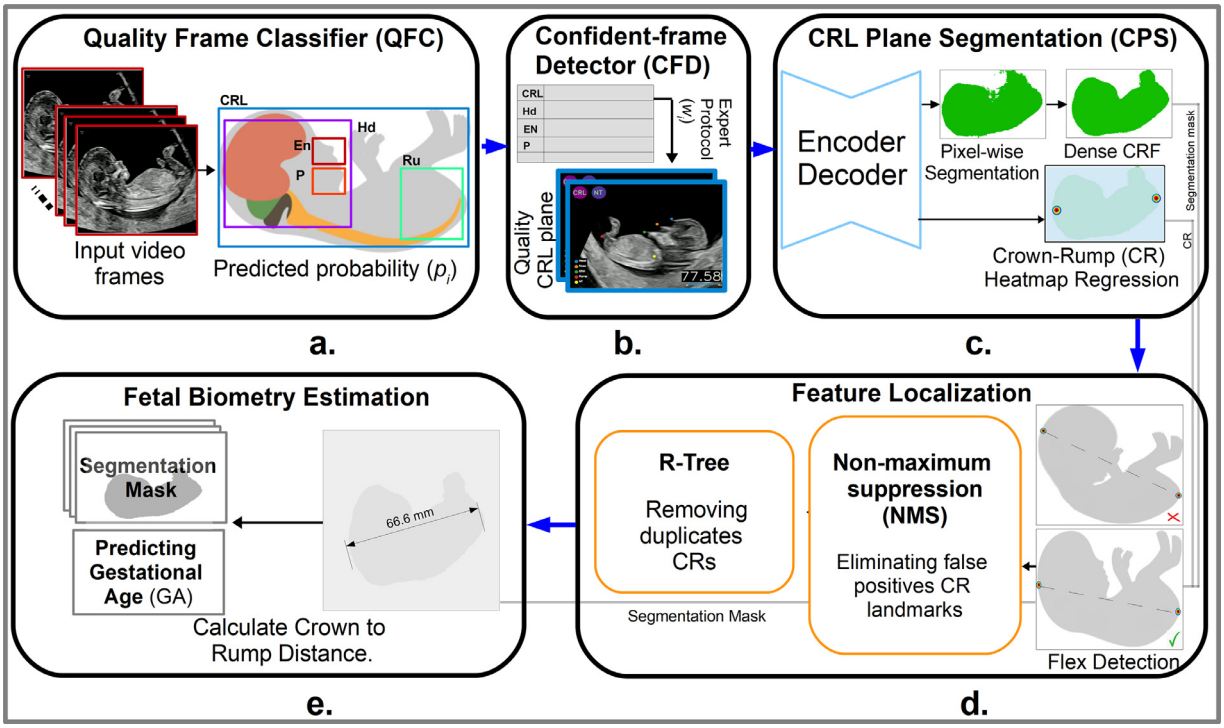

Figure 2. Automated analysis framework. (a) Freehand ultrasound video frames are input into a QFC to predict anatomical structures and their associated probabilities. (b) On the basis of the probability scores, the CFD selects multiple high-quality frames and passes them to (c) a CNN for segmentation of CRL structures. (d) Final predictions are refined using a non-maximum suppression and an R-Tree algorithm to (e) estimate fetal biometry. CNN, convolutional neural network; CRF, conditional random field.

anatomical structures and their associated probabilities. It was achieved through Single Shot MultiBox Detector (SSD) convolutional neural network (CNN).

(b) The next module (Fig. 2b) is a confident-frame detector (CFD) that selects multiple high-quality frames and passes them to the next stage. It was achieved through Confident-Frame Detector, a hybrid global and local classification method that uses CNN-predicted quality scores and expert weights to select the best standard plane frame for the next stage.

2. The second stage, the standard plane segmentation stage called CRL plane segmentation (CPS), is responsible for:

(a) Segmentation of fetal structure (CRL) mask and crown and rump (CR) landmark locations (Fig. 2c). It was achieved through an encoder–decoder CNN that takes high-quality RGB frames from the previous stage and performs pixelwise semantic segmentation of fetal structure.

3. The last stage is feature localization and extraction of fetal biometry:

(a) It involves locating the correct CR landmarks (Fig. 2d) and is the final stage for automatic fetal biometry estimation (Fig. 2e). This was achieved by taking a mean Euclidean distance value of multiple frames for the CRL measure.

Each step will be described, but we start with the data description.

Data description

We used 250 first-trimester fetal US videos of the MS biometry view from a large-scale clinical US study called PULSE [13]. For a more detailed description of the full PULSE acquisition protocol, the reader is referred to Drukker et al. [13]. The study was approved by the UK Research Ethics Committee (Reference 18/WS/0051). Nine different operators were involved in acquisition of the US data.

US system

Full-length first-trimester transabdominal US scans were performed at the Oxford University Hospitals NHS Foundation Trust. The examinations were performed using General Electric (GE) Healthcare Voluson (GE, Vienna, Austria) US machines equipped with standard curvilinear (C2-9-D, C1-6-D, C1-5-D) and 3-D/4-D transducers (RAB6-D, RC6M) operating at 2–8 MHz. The vaginal US scan is not part of the study. The pre-sets for scans were as follows: dynamic depth range during the scan aiming to adjust the field of view to best visualize the structure of interest, single focal zone nearer the structure of interest, no gray-scale gain by default, harmonic imaging always employed, dynamic contrast 7, gray map number 7, compound resolution imaging filter on and speckle reduction imaging filter on. However, as our aim was to capture "real-world" US being undertaken in a clinical (not laboratory) setting, these parameters could be altered by operators.

Video recording

The system is equipped with a customized US video recording card (DVI2PCIe, Epiphany Video, Palo Alto, CA, USA) and purpose-built software to capture the secondary video output from the US machine. The software also ensures real-time anonymization to hide all personal details of participants. Full-length US videos recorded using lossless compression with an HD resolution of 1920 × 1080 pixels and refresh rate of 30 Hz. We used 250 first-trimester fetal US videos scans; the average duration of a scan was 13.73 ± 4.18 min (24,720 ± 7534 frames). The data distribution used in this work is summarized in Figure 1.

Frame extraction

A sonographer takes biometry measurements in the first trimester by scanning the fetal anatomy and locating key anatomical structures. A freeze-frame (FF) video segment is recorded when a sonographer is satisfied that it is a standard plane. Each time sonographers freeze the view, they select the particular part of the fetal anatomy they wish to measure from a dropdown menu. This particular labeled segment of the fetal anatomy is detected and extracted using the optical character recognition (OCR) method. We trained and tested the algorithms using a full FF segment as well as pre-frozen video (90 frames) to ensure that these achieve similar performance on standard and non-standard images.

In CRL measurement during a first-trimester US scan, the fetus should be in a neutral position. A neutral position in this context is a standardized, reproducible position that allows for accurate measurement of the CRL. This measurement is used to estimate gestational age, and the neutral position helps to ensure consistent and reliable results. In a commonly used neutral position, the fetus lies on its back with legs extended straight up toward the head. We selected data set testing (for the biometry task [Fig. 1]) samples with neutral positions for appropriate CRL measurement.

Quality frame classifier

This module is designed to filter unnecessary video frames with less anatomical information. It works as a frame quality control necessary to focus only on high-quality video frames containing all critical anatomical information. The next stage triggers an action only if the quality frame classifier (QFC) detects any high-quality frames desirable for further action.

Referring to Figure 2a, the first stage in our approach is a QFC, which is posed as a regression and classification problem. An engineer and fetal medicine clinician performed manual bounding box annotation for five anatomical structures: (i) head (Hd), (ii) horizontal sagittal section of the fetus (HS), (iii) echogenic tip of the nose (EN), (iv) rump (Ru) and (v) translucent diencephalon (TD). Subsequently, a Single Shot MultiBox Detector CNN Yolo-v5 [14] is used to detect these anatomical structures; it returns class labels (c_i) and associated probability scores (P_i). An attraction of this approach is that predicted scores are computed on the video in real time and can be used as input to detect high-quality CRL frames as described next.

The loss function for the QFC [15] is calculated as

$$l_{QFC} = l_{cls} + l_{box} + l_{obj} \tag{1}$$

where l_{cls} is the anatomical structure classification loss, l_{box} denotes the bounding box regression loss and l_{obj} is the confidence loss for anatomical structures. In turn, these terms are defined as follows:

$$l_{cls} = \lambda_{class} \sum_{i=0}^{S^2} 1_i^{obj} \sum_{C \in classes} (p_i(c) - \hat{p}_i(c))^2 \tag{2}$$

$$l_{obj} = \lambda_{obj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} (c_i - \hat{c}_i)^2 + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{noobj} (c_i - \hat{c}_i)^2 \tag{3}$$

$$l_{box} = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left[(x_i - \hat{x}_i)^2 + \left(y_i - \hat{y}^i\right)^2 \right] + \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right] \tag{4}$$

Here, λ_{class} is the classification loss coefficient and λ_{coord} is the position loss coefficient; $p_i(c)$ represents the predicted class probability, and $\hat{p}_i(c)$ is the true value of the class; x and y are coordinates, and w and h are the width and height. 1_i^{obj} denotes that the object appears in cell i and has a value of "one," and 1_{ij}^{obj} denotes that the jth bounding box predictor is in cell i.

Confident-frame detector

The second stage, and an original contribution of the work described here, is the CFD. The CFD uses a clinically defined quality detection mechanism to determine the best fetal appearance from multiple time stamps. This is a hybrid global and local classification model developed to detect an MS standard plane (Fig. 2b). There are two components to the CFD module: a global component and a local component.

Global component module

This module aims to select more clinically significant frames from the prospects of clinical experts. In the module, expert weights are embedded to select frames containing the most visually significant anatomical landmarks, such as the sagittal view, head and rump. The quality frame classifier module is used to select the potential candidate events, and then the global module is used to decide the best among the candidate events.

We define a high-quality frame as a frame that contains the highest content of clinically significant anatomical structures as defined by the UK NHS Fetal Anomaly Screening Programme (FASP) guideline [2,3]. The global content model ensures that an MS plane contains all key anatomical structures. We combine the QFC predicted probability scores p_i for fetal anatomical structures i = 1, 2, N as a weighted average. Anatomical structures that are more critical to the quality of the MS standard plane are assigned a higher weight, whereas those that are less critical are assigned a lower weight. These weights are referred to as expert weights w_i (Table 1) and were set empirically in consultation with experienced clinical sonographers. We calculate the aggregation probability p as a geometric mean:

$$\hat{p} = \frac{\left[\prod_{i=1}^{N}\left(\frac{p_i}{1-p_i}\right)^{w_i}\right]^{a}}{1+\left[\prod_{i=1}^{N}\left(\frac{p_i}{1-p_i}\right)^{w_i}\right]^{a}} \tag{5}$$

where the parameter a is the systematic bias. In this work we set a = 1. Let the frame confidence status be τ. Its status is defined by

$$\tau = \{\, T_k \quad \text{if } \hat{p} \geq \text{th}_k I_k \text{if } \hat{p} < th_k \tag{6}$$

Here th_k is an empirically determined threshold (defined in our experiments using twofold cross-validation). T_k indicates a confident frame (a high-quality standard plane), and the rest of the frames are called indeterminate frames (I_k), as illustrated in Figure 3a.

Local component model

The local model detects identical frames from the high-quality input (video stream) stream to optimize the algorithm's computational performance. By feeding forward only the best candidates of a set of candidate events, we reduce the overall processing overload, as the next segmentation stage is computationally intensive. The global component model identifies multiple cascading high-quality frames. The structural similarity index (SSIM) [16] is then computed to prune the number of CRL confident frames. In the case of two identical frames with a high SSIM ($m^{\uparrow}$), the algorithm discards the identical frame and proceeds to the next frame. In this local component model, frames at random time stamps are selected to represent different positions and perspectives of the fetus, as illustrated in Figure 3b. To incorporate diverse fetal structures and motion patterns, variable video sequence lengths (10–30 frames) are tested to ensure a high degree of correlation with expert biometry measurements.

Table 1
Expert weights (scores) for anatomical structure significance

Protocol	Score
[CRL] Sagittal section in line with the full length of the body	2
[*Hd*] Crown is clearly defined	3
[*EN*] Nose has an echogenic tip	1
[*P*] Palate has a rectangular shape	1
[Ru] Rump is clearly defined	3

Experienced sonographers rated the importance of each structure present in the standard plane: 3 = highly significant, 1 = least significant.

CRL plane segmentation

Figure 2c illustrates the CRL plane segmentation step. The fetal structure mask and crown and rump landmark locations (hereafter called crown–rump landmarks) were segmented using a training set of 128 manually annotated video clips. As illustrated in Figure 4, CRL plane segmentation is achieved by training a nested encoder–decoder architecture that sandwiches a single stack of an Hourglass CNN [17] between residual blocks [18] and pooling layers. Encoder–decoder CNNs are memory intensive, particularly at points toward the network start and end, where spatial resolution is high. Residual blocks feed the Hourglass encoder–decoder with smaller-sized feature maps, which are less computationally intensive, leading to fewer trainable parameters, thus reducing the computational complexity and increasing performance. At the output, the network is split into two fully convolutional channels that are responsible for separately learning the fetal MS plane segmentation (channel 1) and the crown–rump landmark heatmap regression (channel 2).

For a class such as crown–rump landmarks, there is an extreme imbalance of foreground to background pixels. To address this, we introduce a weighted-loss function, which assigns weights in inverse proportion to the median frequency with which each class appears in the entire training set. Segmentation performance is enhanced by optimizing convergence of the network (adding focus to foreground pixels) without additional trainable parameters.

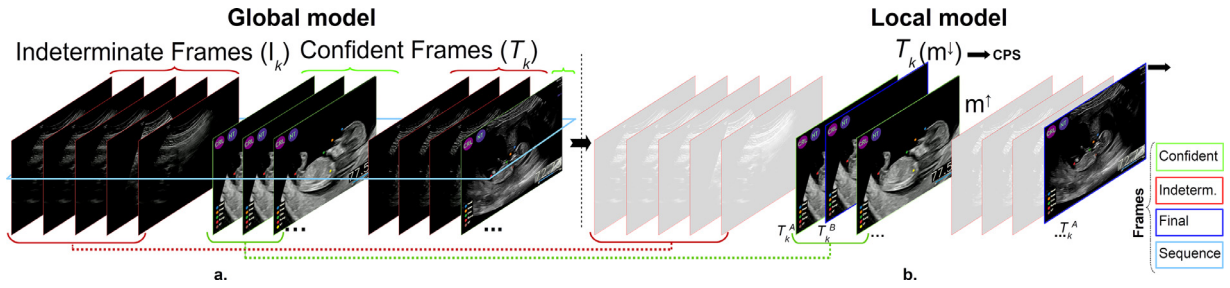

Figure 3. The confident-frame detector (CFD). (a) Global model: detection of confident frames (T_k) and indeterminate frames (I_k) using expert weights and predicted probability scores. (b) Local model: identical frames ($m^{\uparrow}$) are pruned, and confident frames with high dissimilarity score ($m^{\downarrow}$) are selected for further segmentation.

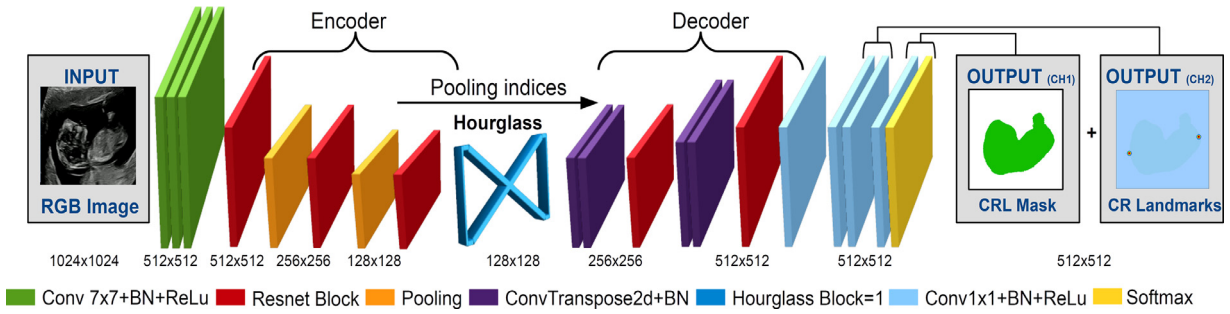

Figure 4. Crown–rump length plane segmentation (CPS) architecture.

The first channel (CH1) segments the MS plane. The l_{CPS_1} loss used in the architecture is

$$l_{\mathrm{CPS}_1} = \frac{\alpha_c}{N}\sum_{n=1}^{N}\sum_{x=1}^{W}\sum_{y=1}^{H}\left[g_{xy}^{n}\log\left(\hat{g}_{xy}^{n}\right) + \left(1-g_{xy}^{n}\right)\log\left(1-\hat{g}_{xy}^{n}\right)\right] \tag{7}$$

where g is the image frame, g_{xy}^{n} is the ground truth and $\hat{g}_{xy}^{n}$ is the predicted class for each of the N features. α_c is the weight of each class, calculated as

$$\alpha_c = \frac{\mathrm{median\_freq}}{\mathrm{freq}(c)} \tag{8}$$

where freq(c) is the frequency of class c, and median_freq is the median of these frequencies over all classes. In addition, a dCRF block [19] was used to improve segmentation results.

The second channel (CH2) predicts the location of crown–rump landmarks. By use of a mean squared error loss, this model predicts likely locations for crown–rump landmarks, represented by a 2-D Gaussian centered at each feature location. The loss l_{CPS_2} used in this architecture is

$$l_{\mathrm{CPS}_2} = \frac{1}{N}\sum_{n=1}^{N}\sum_{xy}\|\hat{P}_n(x,y) - P_n(x,y)\|^2, \tag{9}$$

where $\hat{P}_n(x,y)$ is the predicted confidence at the pixel (x, y) location for the nth part, and $\mathrm{P_n}$(x, y) is the ground truth of the same location. The overall CRL plane segmentation loss is given as $L = l_{\mathrm{CPS}_1} + l_{\mathrm{CPS}_2}$, which balances the objectives of both paths, training the network end to end. In our study, we found that adding additional scaling parameters to the loss of either path, in no circumstances, improved training accuracy.

Feature localization

Feature localization is the next step, as illustrated in Figure 2d. The CRL plane segmentation heatmap regression output consists of 2-D Gaussian distribution maps, with each peak corresponding to a crown–rump landmark location. The discrete location of a crown–rump landmark is computed using a non-maximum suppression algorithm [20] to eliminate predicted pixels that exceed the neighboring pixels (based on an empirical threshold). The algorithm uses non-maximum suppressed landmarks extracted from the predicted mask to avoid landmark overlap.

The non-maximum suppression (NMS) algorithm performs a pixel-wise search and promptly discards any pixel below a pre-defined threshold ($\theta = 0.7$). During this process, each pixel p with coordinates (x, y) is compared with its neighbor pixels in a (3 × 3) window, where the central pixel "c" is non-maximal; if a neighbor pixel with greater or equal intensity is found, the algorithm skips to the next pixel in the scan line. This method executes in a raster scan order, requiring $\lceil(2n+1)^2/2\rceil$ comparisons per pixel for a $(2n+1)\bullet(2n+1)$ neighborhood.

To avoid two locations being returned for a single crown–rump landmark, we identify and suppress adjoining features using an R-Tree-based method. An R-tree [21] is a depth-balanced tree that can store N data rectangles, and the maximum value of its height h is

$$H_{\max} = \lceil \log_m N \rceil - 1 \tag{10}$$

With an allowed number of entries, that is, m for each node, the maximum number of nodes in an R-tree is equal to

$$\sum_{i=1}^{h_{\max}} \lceil N/m_i \rceil = \lceil N/m \rceil + \lceil N/m_2 \rceil + \cdots + 1 \tag{11}$$

In our application, the R-tree takes non-maximum-suppressed candidate crown–rump landmarks and searches for nearby landmarks that have already been added, preventing duplicate landmarks being recorded in close proximity. If a candidate crown–rump landmark location is within an 8 × 8 pixel radius of a recorded landmark, it is regarded as a duplicate of the recorded landmark. As a practical note, non-maximum suppression and the R-Tree landmark reduction method are applied to a downsampled output of the CNN model to lower computational cost.

The FASP guidelines recommend that the fetus must be in a neutral state (not too flexed or extended) to compute an accurate CRL measurement. We performed two sequential checks to determine if the fetus was in a neutral position. The first check is performed for edge detection and finding contours in the edge map. Following this, we detected the outline of the fetus in the edge map and calculated the bounding box surrounding the fetus. In the following steps, the center coordinates (*i, j*) and (*k, l*) of the bounding box are calculated based on the average of the horizontal and vertical directions. Later, the distance between these points was calculated, which gave us the pixel length, allowing us to calculate the absolute distance ratio between the two points. As a result of empirical testing, we determined that if the ratio $d(i,j)/d(k,l) \geq 1.4$, the fetal position will be neutral (Fig. 5a, 5b), where $d(\cdot,\cdot)$ is the Euclidean distance.

The second check is performed by tracing a line between the crown–rump (cr) points and measuring the distance between them. The predicted crown–rump landmarks are required to fall within or on the boundary of the CRL contour points. This check is conducted at four equally positioned points, two in the head area (*yy*', *zz*') and two in the rump area (*ww*', *xx*') (Fig. 5c, 5d). These two checks ensure that at least one of the tests returns a true value to ensure the fetal neutral state.

Fetal biometry estimation

The CRL is measured from the crown to the rump of a fetus (Fig. 2e). In our implementation, CRL contour points are extracted from the segmentation mask, and the Euclidean distance between these contour points is computed. The length in pixels is scaled to millimeters. Several equations for calculating gestational age (GA) from a CRL measurement have been proposed in the clinical literature. In this article, we use guideline [5,22]

$$\mathrm{GA} = 40.9041 + \left(3.21585 \times \alpha_{\mathrm{crl}}^{0.5}\right) + \left(0.348956 \times \alpha_{\mathrm{crl}}\right) \tag{12}$$

where α_{crl} is crown-to-rump distance (mm).

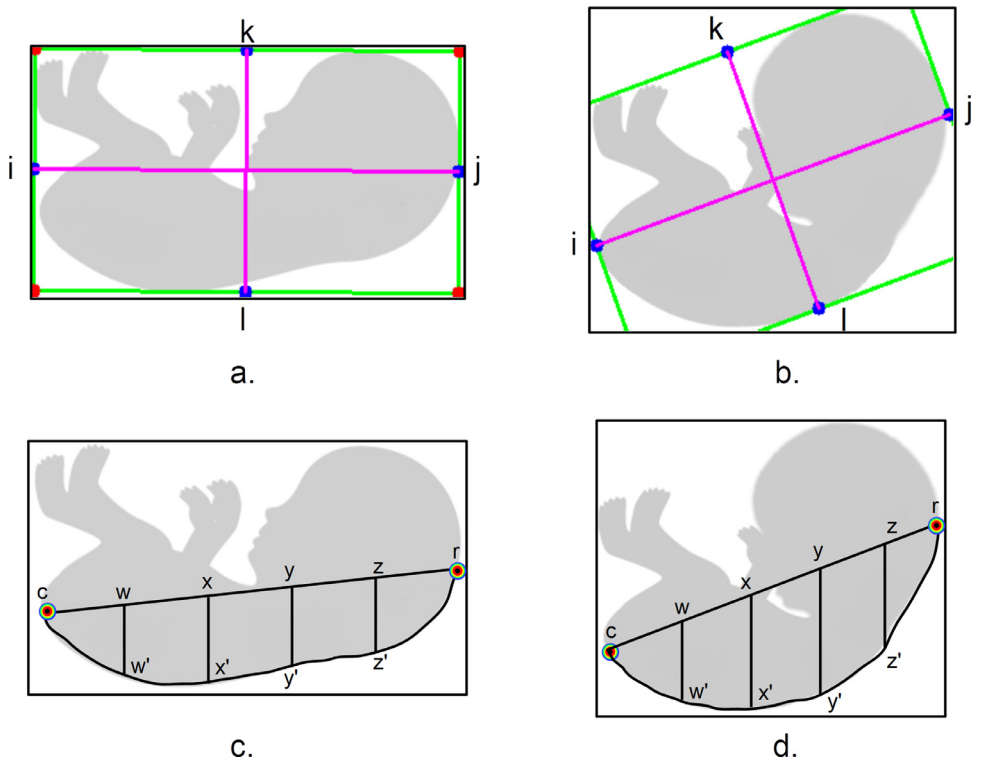

Figure 5. Fetal flex detection. (a) The fetus will be in a neutral state when the absolute distance ratio between points $d(i,j)/d(k,l) \geq 1.4$. (b) Fetus will be in a flexed state when the absolute distance ratio between points $d(i,j)/d(k,l) < 1.4$. (c) Fetus will be in a neutral state when the absolute distance ratio between points $d(x,x')/d(w,w') \leq 1.2$ and $d(y,y')/d(z,z') \leq 1.2$. (d) Fetus will be in a flexed state when the absolute distance ratio between points $d(x,x')/d(w,w') > 1.2$ and $d(y,y')/d(z,z') > 1.2$.

Results

Evaluation of quality frame classifier

The quality frame classifier was trained end-to-end from scratch for 200 epochs with a batch size of 20 via adaptive moment estimation (Adam) and an initial learning rate of $1e^{-3}$ with a decay ($\times$ 0.1) every 30 epochs. Random horizontal flipping with a probability of 50% was used as part of data augmentation. For training, the confidence threshold was set to 0.25, and the non-maximum suppression (NMS) IoU threshold was set to 0.45. For inference, the confidence threshold was set to 0.45, and the NMS IoU threshold was set to 0.60. The quality frame classifier model was evaluated using recall (R), precision (P), F1 score (F1) and Top-1 accuracy (Top-1). The quality frame classifier test set performance based on these metrics is $P = 0.88 \pm 0.05$, $R = 0.85 \pm 0.03$, F1 $= 0.85 \pm 0.10$ and Top-1 $= 0.87 \pm 0.06$.

A confusion matrix is illustrated in Figure 6 for the quality frame classifier. The horizontal sagittal section of the fetus (HS) has a high detection score. On the other hand, the rump (Ru) and translucent diencephalon (TD) are more complicated structures and harder to detect and localize.

In Figure 7, we compare CNN predictions based on the QFC with ground truth. Figure 7a and 7b show two examples where the ground

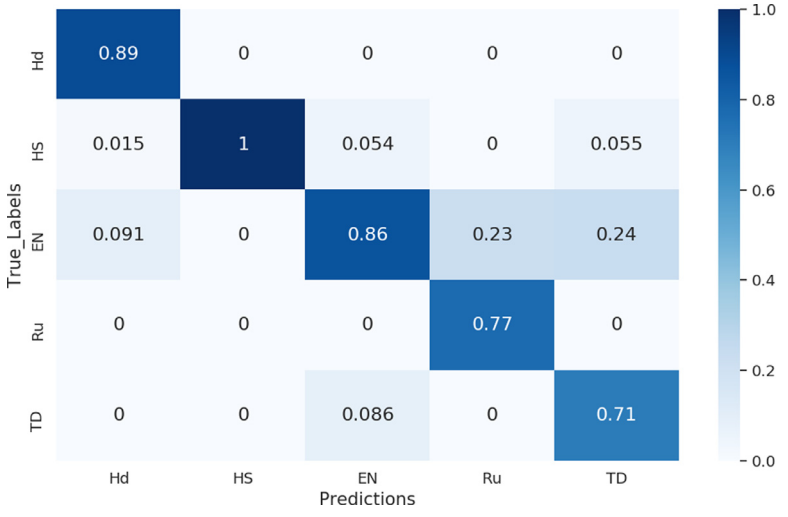

Figure 6. Quality frame classifier test set confusion matrix: predicted versus ground truth.

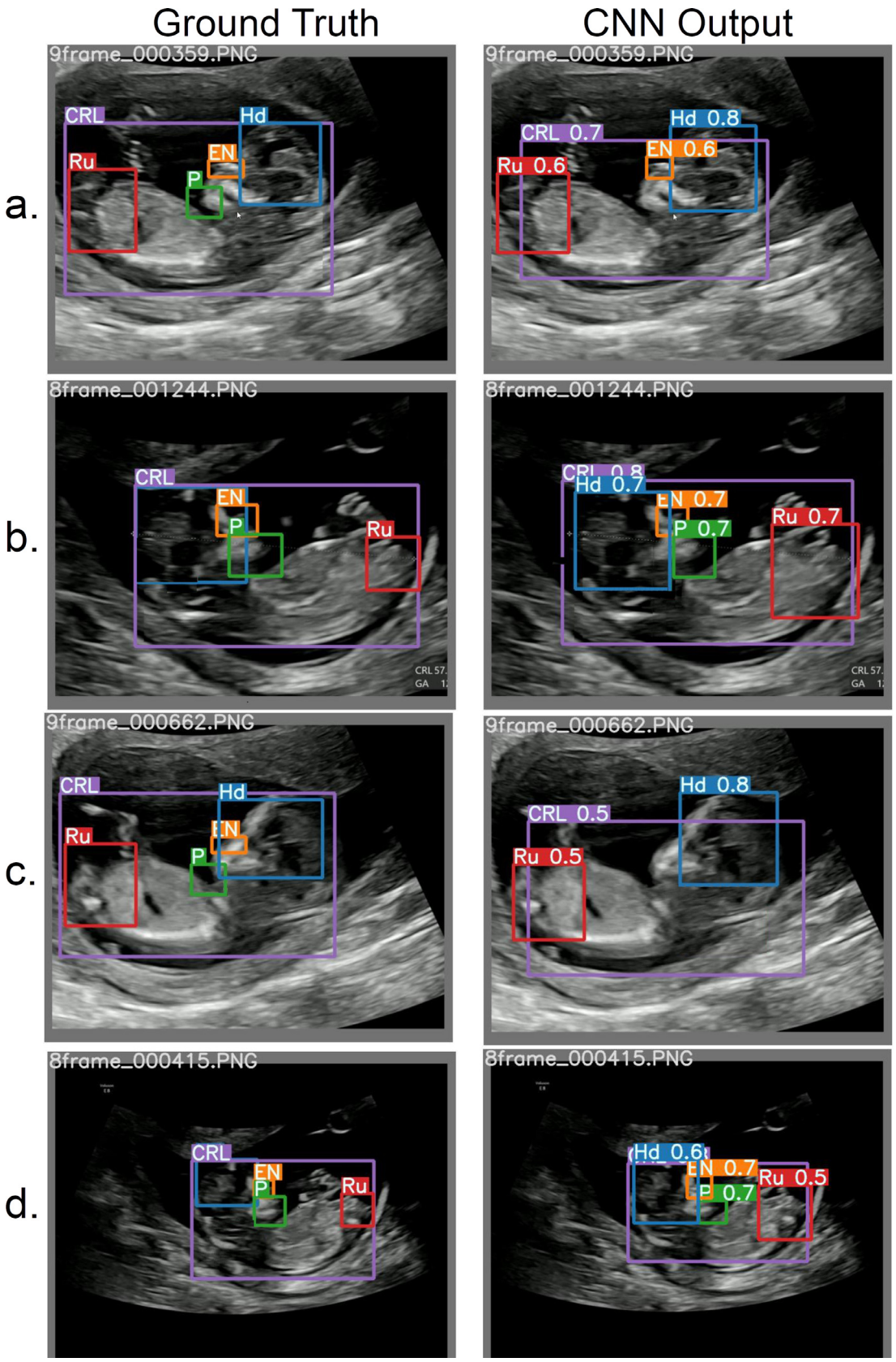

Figure 7. Qualitative analysis of quality frame classifier: comparison of the inference results between the ground truth and CNN predictions. The first column is a list of images that have been labeled by experienced clinicians. The second column presents the automated detection results and confidence scores. CNN, convolutional neural network.

Table 2
Quantitative analysis of segmentation on test data set (meanAsstd [%])

Method	GAA (%) (↑)	MA (%) (↑)	mIoU (%) (↑)
FCN-16 [24]	79.05 ± 0.05	66.23 ± 0.10	54.48 ± 0.20
FCN-32 [24]	81.68 ± 0.01	76.56 ± 0.02	63.87 ± 0.18
U-Net [25]	83.64 ± 0.08	79.80 ± 0.07	67.41 ± 0.25
SegNet [26]	85.08 ± 0.10	83.82 ± 0.10	70.05 ± 0.33
HG (B = 1, S = 2) [17]	89.05 ± 0.09	82.70 ± 0.20	70.83 ± 0.05
CPS (ours)	92.32 ± 0.03	85.01 ± 0.01	74.42 ± 0.04
CPS-Focal-Loss	93.17 ± 0.04	87.33 ± 0.11	76.14 ± 0.01
CPS-Weighted-Loss	94.44 ± 0.03	87.57 ± 0.02	78.69 ± 0.27
CPS-Weighted-Loss + dCRF	94.76 ± 0.01	88.49 ± 0.03	80.02 ± 0.19

CPS, crown–rump length plane segmentation; GAA, global average accuracy; MA, mean accuracy; mIoU, mean intersection over union.

truth and QFC model are in good agreement. Figure 7c illustrates a case where the QFC model finds it difficult to detect some key fetal anatomical landmarks owing to the complexity of first-trimester fetal anatomy and maternal background. Figure 7d indicates the ability to detect fetal anatomy structures at a different imaging scale.

Evaluation of CRL plane segmentation

The CRL plane segmentation (CPS) models were trained from scratch via the RMSprop [23] optimizer using 120 videos. Weighted cross-entropy loss was used for pixelwise segmentation of the fetal mask. MSE loss was used for heatmap regression of crown–rump landmarks for 500K iterations. CNN benchmarks (FCN [24], UNet [25], SegNet [26] and Hourglass [17]) were included for comparison. For each model data augmentation included rotation [−30°, 30°] and horizontal flipping. Models were evaluated using the following metrics: global average accuracy (GAA), mean accuracy (MA) and mean intersection over union (mIoU). All models were implemented in PyTorch 1.8.0.

Table 2 indicates that the CPS model outperforms other CNN benchmark architectures. The CPS model GAA is 5.71% higher than a standard Hourglass (block = 1, stack = 2). The addition of weighted loss to the standard CPS model increases the mIoU by 2.56%. The results suggest that weighted loss is more effective than a class balancing (focal loss [27]) cross-entropy loss. In the post-processing stage, the CPS model improved with the addition of the dCRF block [19], which boosted the mIoU by 1.33%. Figure 8 illustrates some typical qualitative results of CPS models.

Superior segmentation results can be attributed to the nested encoder–decoder design, fewer parameters and choice of a weighted-loss function. Furthermore, a well-balanced diverse training data set significantly contributed to better convergence of the model. We trained the CNN on 12K well-balanced standard plane frames. We added different lie positions and movements of the fetus to add diversity to the training set, which improved the learning efficiency of the model.

Evaluation of automated fetal biometry

Evaluation of automated CRL measurement

With use of the CFD, high-quality MS view standard planes are detected from first-trimester US fetal scan videos. The CRL plane segmentation encoder–decoder architecture was used to extract the fetal segmentation mask, which was later used to calculate the crown–rump length. Figure 9 compares manual versus automated CRL measurement for the 42-participant test data set for different spatiotemporal samplings. Single-shot automatic fetal biometry (Fig. 9a1, 9a2) attained a Pearson correlation coefficient (PCC) of $\rho = 0.85$ and $R^2 = 0.72$. Using additional temporal information from a sequence ($s = 20–30$) with a mean fetal biometry estimated using three frames increased the correlation between manual and automatic fetal biometry (Fig. 9b1, 9b2, 9c1, 9c2). The sequence window size $s = 30$ with a three-frame average achieved a high $p = 0.92$ and $R^2 = 0.84$ (Table 3).

Our results indicate that accuracy of fetal biometry estimation can be further improved by multiframe video analysis rather than single-frame analysis. This approach mimics the clinical experts' practice of taking multiple CRL measurements in different fetal positions. The proposed method selects the best standard plane frames and most optimal fetal positions to determine the final biometry measurement. Figure 10a provides a Bland–Altman plot for single-frame-based manual measurement versus automated fetal biometry analysis, which shows a systematic measurement bias of 2.60 mm. Increasing frame sampling to three frames and incorporation of multiframe video analysis increase the agreement between manual and automated fetal biometry (Fig. 10b). Figure 10c illustrates the proposed approach with a substantial reduction in measurement bias (0.97 mm) that indicates a high correlation between the proposed approach and measurements from experts. In a

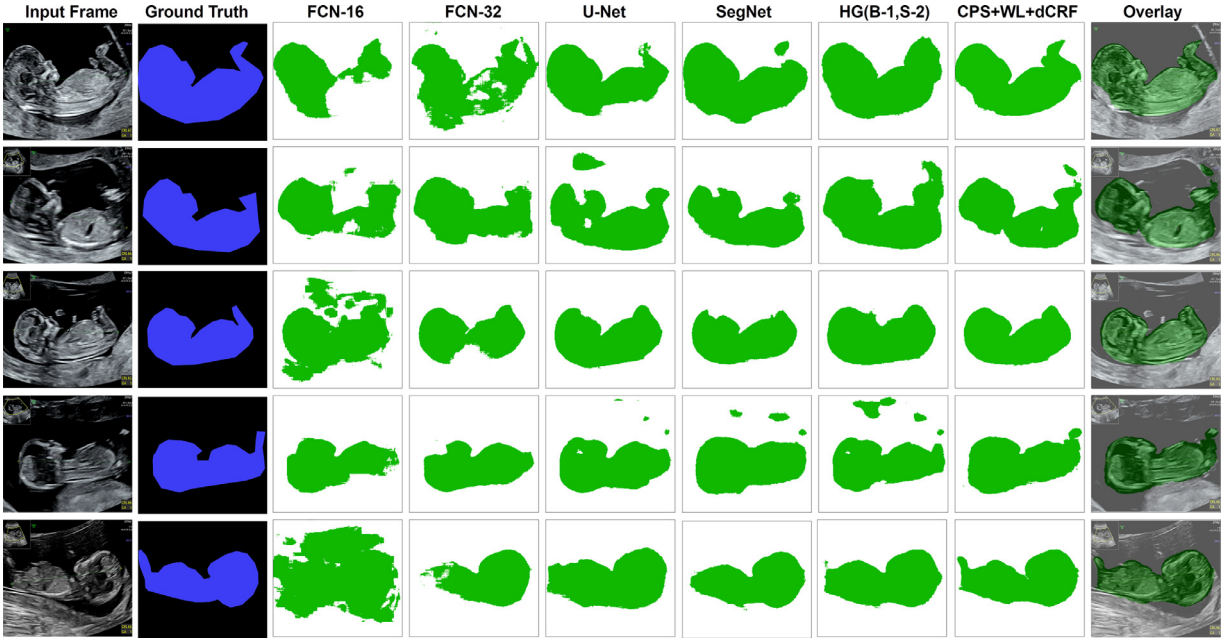

Figure 8. Qualitative analysis of CPS. AI, artificial intelligence; CPS, CRL plane segmentation; CRL, crown–rump length.

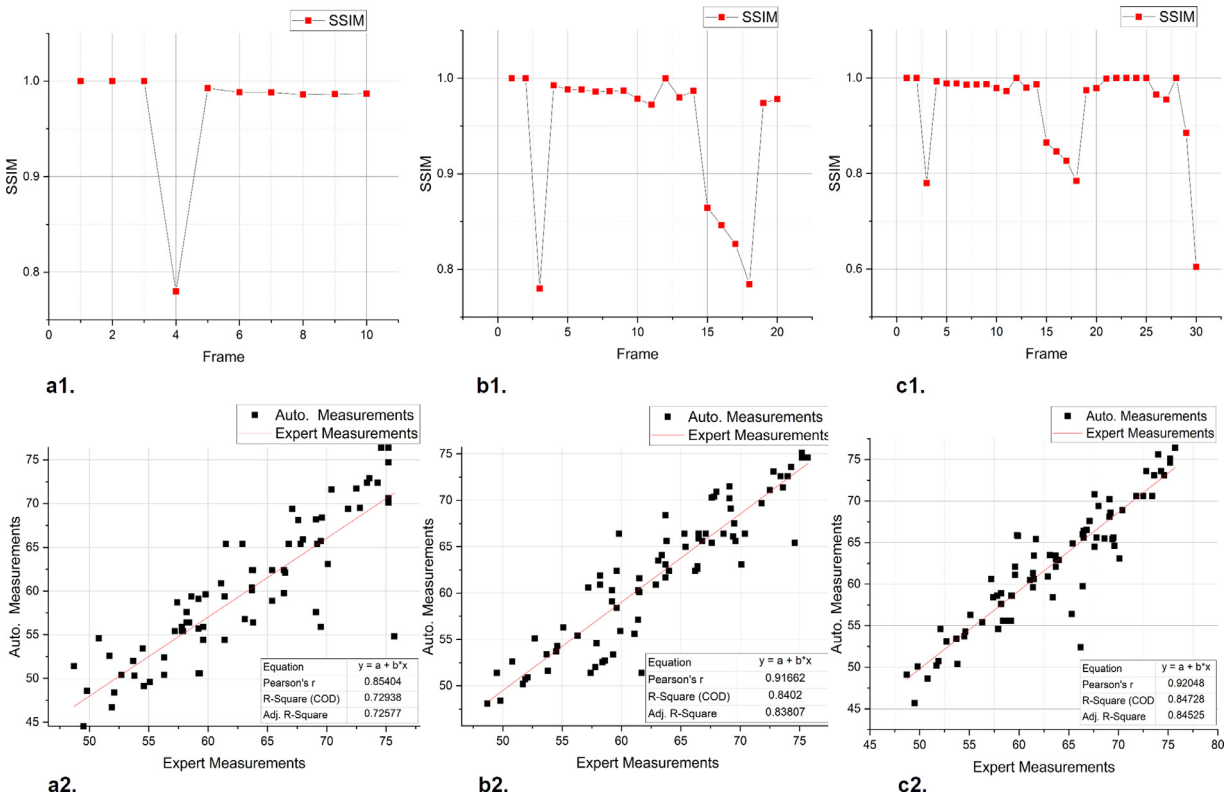

Figure 9. Correlation between manual and automatic CRL fetal biometry (mm). (a1) SSIM index for cascading adjacent frame sequence (s) = 10. (b1) s = 20. (c1) s = 30. (a2) Automatic versus manual fetal biometry measurement from high-quality frames, where frames (f) = 1 and s = 10; (b2) f = 3 and s = 20; and (c2) f = 3 and s = 30. COD, coefficient of determination; CRL, crown–rump length; SSIM, structural similarity index measure.

Table 3
Analysis of confident-frame detector model with respect to fetal biometry

Frame sequence (s)	Candidate frames (f)	Pearson score (ρ) (↑)	R-squared (R^2) (↑)
10	1	0.854	0.729
20	3	0.916	0.840
30	3	0.920	**0.847**
40	4	0.919	0.842

Boldface value indicates to highlight best performance.

standard first-trimester US scan, the operator will take into account the fetal position and movements to make a final CRL measurement. Similarly, the proposed approach has indicated that long frame sequences (s) and accounting for more frames (f) from different time stamps of video scans have improved the accuracy of fetal biometry. The early methods [8,11] typically measured the final fetal biometry using only one frame; however, we accounted for different fetal positions and movements for the final biometry, resulting in a high correlation with clinical expert biometry.

Evaluation of age estimations

We compared traditional single-frame gestational age estimation [8,11] with our multiframe estimation method. A test data set of 42 participants with 82 manual measurements was available. CRL measurements (mm) were converted to gestational age (wk) using the formula in Papageorghiou et al. [5]. The mean absolute error (MAE) and the R^2 of the linear regression between estimated and manual age estimations were computed. As outlined in Table 4, single-frame

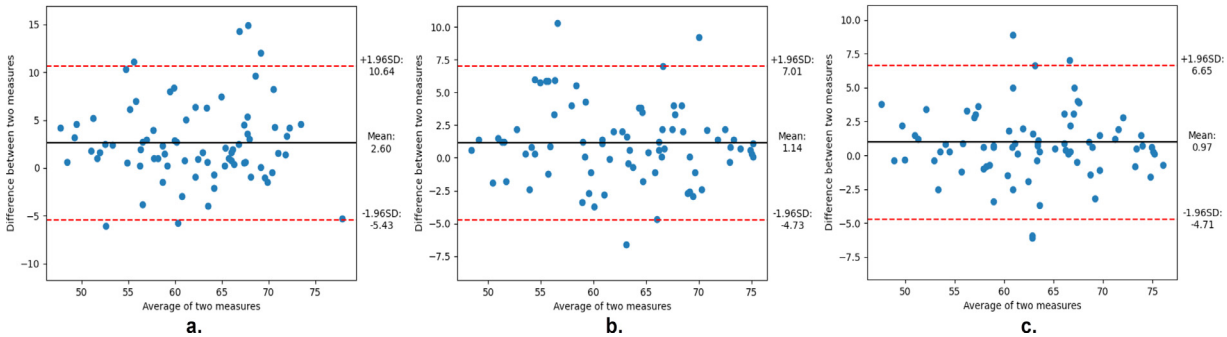

Figure 10. Bland–Altman plots for manual and automated CRL measurement (mm). The *dotted line* represents 2 SD above and below the mean difference. (a) Single-frame (f = 1)-based CRL measurement from sequence of s = 10 frames. (b) Three-frame (f = 3)-based CRL measurement from s = 20 frame sequence. (c) Three-frame (f = 3)-based CRL measurement from s = 30 frame sequence. CRL, crown–rump length; SD, standard deviation.

Table 4
Mean absolute errors for fetal age estimation based on automated crown–rump length estimation

Method	Mean absolute error (wk) (↓)	R^2 (↓)
Single-frame estimation	0.908	0.725
Dual-frame estimations	0.864	0.838
Proposed method	0.834	0.845

gestational age estimation has a higher MAE and a lower R^2 compared with two-frame or multiframe estimation, which indicates that expert versus automatic single frame-based gestational age estimation methods have a high proportion of variance. This suggests that adding more diverse anatomical features in the multiframe strategy leads to a lower MAE (0.834) and higher R^2 (0.845). A high correlation between expert (manual) and automatic (multiframe) gestational age estimations has been indicated.

Complete method: Example results

Figure 11 presents the output stages for three example results. Figure 11a illustrates the single-shot-detector CNN output bounding boxes and their associated probabilities. Figure 11b illustrates how the real-time visual cues provided by CFD algorithms are visualized by the user. There are visual cues such as the outline of the best-quality standard plane (CRL, NT) that can be seen in the top-left corner of Figure 11b. These cues indicate that the current frame may be suitable for biometry. Figure 11c illustrates the segmentation results of the high-quality frames selected in the early stages of the analysis. Figure 11d illustrates the final step of automated fetal biometry. As a whole, the automated pipeline mimics the human action of standard plane detection and fetal biometry measurement.

Discussion

There is a current global shortage of US technicians and a concurrent increase in demand for US-based obstetric care. This is increasing the importance of maximizing the efficiency of scanning time. Automating fetal biometry would aid clinical workflow efficiency as well as standardize measurement. A number of methods automatically estimating fetal biometry have been proposed. In recent years, machine learning-based methods have gained much attention, with success in US image analysis [28–30]. Moreover, these methods have also been applied in fetal biometry to analyze biometric features in US images [31–33]. However, all prior automated fetal biometry measurement methods assume pre-selected fetal standard planes and typically fail to account for frequent fetal movements and fetal positional change over time (particularly notable in the first trimester of pregnancy). Here we argue that to be useful in clinical practice, a fully automated biometry method should mimic the complete human action and not just a subtask. In our case, this means that the method should identify a high-quality standard plane from a real-time video feed and carry out pixelwise semantic segmentation and post-processing for automated CRL estimation.

With further validation and refinement, the method proposed in this article could facilitate and guide newly qualified operators and trainee sonographers who tend to take more time to perform a US scan [34]. From a clinical point of view, successful interpretation of first-trimester clinical US requires a good understanding of fetal anatomy and awareness of frequent fetal movements. Operator experience plays a significant role in acquiring high-quality and reproducible fetal health assessments.

A manual analysis of 250 freehand US first-trimester video recordings revealed that operators typically take, on average, three CRL measurements. Manually measured CRL mean variance was found to be 3.25 mm. The recorded (best) CRL measurement depends on operator experience and fetal position (not too flexed or extended). Automated

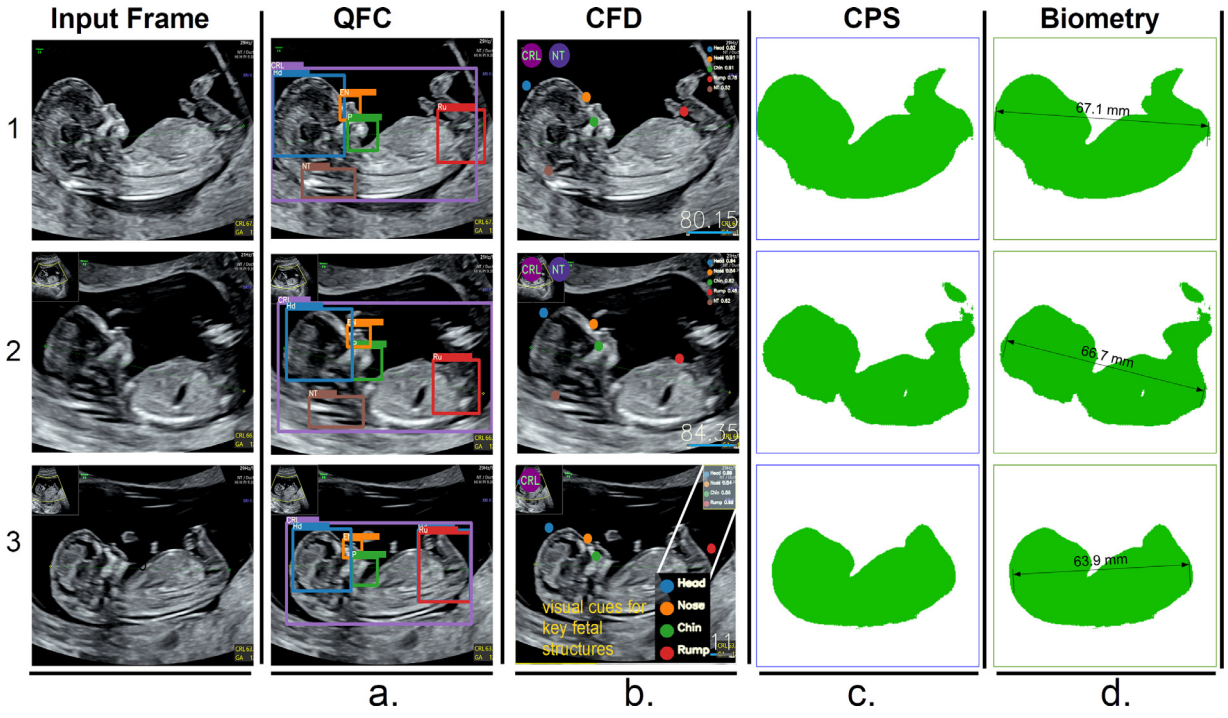

Figure 11. Example results. (a) The QFC detects key anatomical structures and their associated probability scores. (b) The CFD tracks key anatomical structures and provides real-time visual cues (color dots extracted from bounding box predictions) located above the fetal anatomy to identify fetal anatomical structures in addition to a frame quality score (denoted in the bottom-right corner). (c) Segmentation mask of fetal midline sagittal view. (d) Automated biometry. CFD, confident-frame detector; QFC, quality frame classifier.

measurement based on machine learning-based models could offer guidance in locating high-quality standard planes and automating fetal biometry estimation. The automatic spatial–temporal sampling and multishot fetal biometry estimation approach that we have developed resulted in an improvement of 6.13% compared with a single-frame biometry estimation method (Fig. 9) in our evaluation. Additionally, the automatic measurement exhibited a high degree of correlation with clinical manual biometry. The quality frame classifier and CFD process at 26.7 frames/s during the inference stage on an NVIDIA GeForce GTX 1080 (11 GB) GPU.

The proposed system is a sequential pipeline that executes the forthcoming stages if the first stage triggers a positive response. We set the tested quality levels for each stage to ensure a robust method execution as follows: (i) the CFD module ensures frame quality using expert weights. (ii) The CPS module provides frame quality using a fetal flex detection mechanism. (iii) The final feature of the localization module ensures final frame quality using R-Tree to eradicate unnecessary CR points. In the case of later stages, if a module fails, it will move the system to the very first stage to re-initiate the process.

A limitation of this study is that the work was performed on retrospective data. Future work would need to embed the models in a US system and evaluate the system's accuracy and use in a real-world setting.

Conclusion

To the best of our knowledge, this is the first report of an automated first-trimester US fetal biometry estimation method using a multishot spatiotemporal sampling approach. Our automated analysis framework detects and tracks key anatomical structures, selects a high-quality standard plane from a video stream based on the combination of the anatomy predicted probability scores and significance scores provided by sonographers and uses extracted fetal anatomical landmarks from multiple frames for CRL estimation. We evaluated the performance of the framework on real clinical US videos and compared automated fetal biometry measurement with manual measurement by sonography experts. We found that a three-subaction solution can automate the complete human action of performing first-trimester biometry measurement. This may form the basis of a useful automatic real-time capability for fetal biometry estimation suitable for minimally trained operators who do not need to master the intricate details of US guidance and measurement.

Conflict of interest

The authors declare no competing interests.

Acknowledgments

This work is supported by the European Research Council (ERC) (ERC-ADG-2015694581, Project PULSE), the Engineering and Physical Sciences Research Council (EPSRC) (EP/R013853/1 and EP/T028572/1) and the National Institute for Health and Care Research (NIHR) Oxford Biomedical Research Centre.

Data availability statement

Because of the sensitive nature of the data in this study, participants were assured that raw data would remain confidential and would not be shared.

References

[1] Salomon L, Alfirevic Z, Bilardo C, Chalouhi G, Ghi T, Kagan K, et al. ISUOG practice guidelines: performance of first-trimester fetal ultrasound scan. Ultrasound Obstet Gynecol 2013;41:102–13.
[2] Fetal anomaly screening programme handbook 2015.
[3] Alt S, McHugh A, Permalloo N, Pandya P. Fetal anomaly screening programme. Obstet Gynaecol Reprod Med 2020;30:395–7.
[4] Napolitano R, Dhami J, Ohuma EO, Ioannou C, Conde-Agudelo A, Kennedy SH, et al. Pregnancy dating by fetal crown-rump length: a systematic review of charts. BJOG 2014;121:556–65.
[5] Papageorghiou A, Kennedy S, Salomon L, Ohuma E, Cheikh Ismail L, Barros F, et al. International standards for early fetal size and pregnancy dating based on ultrasound measurement of crown–rump length in the first trimester of pregnancy. Ultrasound Obstet Gynecol 2014;44:641–8.
[6] Pu B, Li K, Li S, Zhu N. Automatic fetal ultrasound standard plane recognition based on deep learning and IIoT. IEEE Trans Ind Inf 2021;17:7771–80.
[7] Zeng Y, Tsui PH, Wu W, Zhou Z, Wu S. Fetal ultrasound image segmentation for automatic head circumference biometry using deeply supervised attention-gated V-Net. J Digital Imaging 2021;34:134–48.
[8] Cengiz S, Yaqub M. Automatic fetal gestational age estimation from first trimester scans. International Workshop on Advances in Simplifying Medical Ultrasound. Cham: Springer; 2021. p. 220–7.
[9] Chen H, Dou Q, Ni D, Cheng JZ, Qin J, Li S, Heng PA. Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks. International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer; 2015. p. 507–14.
[10] Chen H, Wu L, Dou Q, Qin J, Li S, Cheng JZ, et al. Ultrasound standard plane detection using a composite neural network framework. IEEE Trans Cybernet 2017;47:1576–86.
[11] Bano S, Dromey B, Vasconcelos F, Napolitano R, David AL, Peebles DM, et al. AutoFB: automating fetal biometry estimation from standard ultrasound planes. International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer; 2021. p. 228–38.
[12] Karim JN, Roberts NW, Salomon LJ, Papageorghiou AT. Systematic review of first-trimester ultrasound screening for detection of fetal structural anomalies and factors that affect screening performance. Ultrasound Obstet Gynecol 2017;50:429–41.
[13] Drukker L, Sharma H, Droste R, Alsharid M, Chatelain P, Noble JA, et al. Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video. Sci Rep 2021;11:1–12.
[14] Jocher G, Chaurasia A, Stoken A, Borovec J, et al. Yolov5, code repository, <https://github.com/ultralytics/yolov5>; 2022 [accessed 19.03.22].
[15] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. New York: IEEE; 2016. p. 779–88.
[16] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004;13:600–12.
[17] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer Vision—ECCV 2016. Lecture Notes in Computer Science Vol. 9912. Cham: Springer; 2016. p. 483–99.
[18] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings, IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. New York: IEEE; 2016. p. 770–8.
[19] Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with Gaussian edge potentials. Adv Neural Inf Process Syst 2011;24:109–17.
[20] Lindeberg T. Edge detection and ridge detection with automatic scale selection. Int J Computer Vision 1998;30:117–56.
[21] Manolopoulos Y, Papadopoulos AN, Papadopoulos, Theodoridis Y. R-Trees: theory and applications. Berlin/Heidelberg: Springer Science & Business Media; 2006.
[22] Kuhn P, de Lourdes Brizot M, Pandya PP, Snijders RJ, Nicolaides KH. Crown–rump length in chromosomally abnormal fetuses at 10 to 13 weeks' gestation. Am J Obstet Gynecol 1995;172:32–5.
[23] Tieleman T, Hinton G. Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks Mach Learn 2012;4:26–31.
[24] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings. *IEEE conference on computer vision and pattern recognition* (CVPR). Boston: IEEE; 2015. p. 3431–40.
[25] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, editors. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. Lecture Notes in Computer Science, Vol. 9351. Cham: Springer; 2015. p. 234–41.
[26] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder–decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39:2481–95.
[27] Lin T, Goyal P, Girschick R, He K, Dollar P. Focal loss for dense object detection. In: Proceedings, 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017. New York: IEEE; 2017. p. 2980–8.
[28] Gao Y, Beriwal S, Craik R, Papageorghiou AT, Noble JA. Label efficient localization of fetal brain biometry planes in ultrasound through metric learning. Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis. Cham: Springer; 2020. p. 126–35.
[29] Dou H, Yang X, Qian J, Xue W, Qin H, Wang X, et al. Agent with warm start and active termination for plane localization in 3D ultrasound. In: Proceedings, International Conference on Medical Image Computing and Computer-Assisted Intervention—MICCAI 2019. Cham: Springer; 2019. p. 290–8.

[30] Droste R, Drukker L, Papageorghiou AT, Noble JA. Automatic probe movement guidance for freehand obstetric ultrasound. In: Proceedings, International Conference on Medical Image Computing and Computer-Assisted Intervention—MICCAI 2020. Cham: Springer; 2020. p. 583–92.

[31] Sobhaninia Z, Rafiei S, Emami A, Karimi N, Najarian K, Samavi S, et al. Fetal ultrasound image segmentation for measuring biometric parameters using multi-task deep learning. In: Proceedings, 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). New York: IEEE; 2019. p. 6545–8.

[32] van den Heuvel TLA, Petros H, Santini S, de Korte CL, van Ginneken B. Automated fetal head detection and circumference estimation from free-hand ultrasound sweeps using deep learning in resource-limited countries. Ultrasound Med Biol 2019;45:773–85.

[33] Płotka S, Włodarczyk T, Klasa A, Lipa M, Sitek A, Trzciński T. Fetalnet: multi-task deep learning framework for fetal ultrasound biometric measurements. Commun Computer Inf Sci 2021;1517:257–65.

[34] Sharma H, Drukker L, Chatelain P, Droste R, Papageorghiou AT, Noble JA. Knowledge representation and learning of operator clinical workflow from full-length routine fetal ultrasound scan videos. Med Image Anal 2021;69:101973.