

# A Machine Learning Method for Automated Description and Workflow Analysis of First Trimester Ultrasound Scans.

Robail Yasrab, Zeyu Fu, He Zhao, Lok Hin Lee, Harshita Sharma, Lior Drukker, Aris T Papageorgiou, J. Alison Noble.

**Abstract**—Obstetric ultrasound assessment of fetal anatomy in the first trimester of pregnancy is one of the less explored fields in obstetric sonography because of the paucity of guidelines on anatomical screening and availability of data. This paper, for the first time, examines imaging proficiency and practices of first trimester ultrasound scanning through analysis of full-length ultrasound video scans. Findings from this study provide insights to inform the development of more effective user-machine interfaces, of targeted assistive technologies, as well as improvements in workflow protocols for first trimester scanning. Specifically, this paper presents an automated framework to model operator clinical workflow from full-length routine first-trimester fetal ultrasound scan videos. The 2D+t convolutional neural network-based architecture proposed for video annotation incorporates transfer learning and spatio-temporal (2D+t) modelling to automatically partition an ultrasound video into semantically meaningful temporal segments based on the fetal anatomy detected in the video. The model results in a cross-validation A1 accuracy of 96.10%, F1 = 0.95, precision = 0.94 and recall = 0.95. Automated semantic partitioning of unlabelled video scans (n=250) achieves a high correlation with expert annotations ( $\rho = 0.95, p = 0.06$ ). Clinical workflow patterns, operator skill and its variability can be derived from the resulting representation using the detected anatomy labels, order, and distribution. It is shown that nuchal translucency (NT) is the toughest standard plane to acquire and most operators struggle to localize high-quality frames. Furthermore, it is found that newly qualified operators spend 25.56% more time on key biometry tasks than experienced operators.

**Index Terms**—first trimester, ultrasound, spatio-temporal analysis, video classification, clinical workflow.

## I. INTRODUCTION

**T**he first trimester fetal ultrasound (US) scan is an essential part of obstetric care, offered to pregnant women to establish fetal viability, pregnancy dating by measurement of the fetal crown-rump length (CRL) and estimating the likelihood of chromosomal abnormalities through measurement of

the fetal nuchal translucency (NT). The UK Fetal Anomaly Screening Programme (FASP) [1], [2] clinical protocol for the first trimester scan defines these key tasks and measurements, carried out between  $11^{+2}$  and  $14^{+1}$  weeks<sup>+days</sup> of gestation. In recent years, first trimester scans are also increasingly carried out to detect fetal structural anomalies [3]. Detection rates of anomalies (structural and chromosomal) in first trimester ultrasound scans ranges from 32% in low-risk pregnancies to 60% in high-risk pregnancies [4], and recent data also suggest that cardiac defects can be detected at this early stage [5].

Currently, there is no universally accepted standard workflow protocol for the first trimester anatomical screening [4]. The task order for a first trimester scan depends on fetal position, fetal movement, and sonographer preferences. Importantly, sonographer skill and experience play a critical role in first trimester scan acquisition [6]. In addition, structures can be captured in various acquisition planes (axial, coronal and sagittal) using different ultrasound modes (2D, Doppler, and 3D) and methods (Transabdominal and Transvaginal). Hence, it is challenging to automatically interpret and analyze first trimester ultrasound scans due to wide variance in scanning preferences and varied workflows. This study aimed to develop an original spatio-temporal deep learning (DL) architecture trained to provide semantic labels for the first trimester US video. To demonstrate clinical applicability, the trained model is used to semantically partition unlabelled full-length first-trimester US video scans and investigate the clinical workflow. This provides insight into real-world clinical imaging workflow which has not been shown before.

The contributions of the paper are two-fold.

- 1) To automate video annotation: A spatio-temporal deep learning architecture is trained using pre-trained weights from a second trimester video annotation task model to provide semantic labels for a first trimester US video annotation task model. We experimented with various deep neural networks to determine the best performing model for the annotation task. We also investigated the effect of introducing spatio-temporal knowledge during training. The model with the highest performance was assessed for similarity with expert-labelled video scans. The best performing trained model was subsequently applied to annotate the entire dataset of full-length first trimester US videos.

R.Y, Z.F, H.Z, L.H.L, H.S and J.A.N are with the Institute of Biomedical Engineering, University of Oxford, Oxford, OX3-7DQ, UK (corresponding author e-mail: robail.yasrab@eng.ox.ac.uk).

A.T.P, and L.D are with the Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, UK.

L.D is with the Rabin Medical Center, Sackler Faculty of Medicine, Tel-Aviv University, Israel.

- 2) To further our understanding of clinical sonography: For the clinical workflow analysis, a complete anatomical timeline model was built upon partitioned first trimester US scans to investigate the differences and similarities among different scans and sonographer scanning patterns.

The outline of the paper is as follows. Firstly, in Section II, we summarize the related literature on ultrasound video analysis and first trimester US image analysis. Section III outlines the US data acquisition protocol, pre-processing and proposed deep learning model for automated full-length video scan annotation. Automated video annotation is evaluated in Section IV. In section V-A we analyze the annotated video datasets using a subject-specific timeline model to summarize clinical and operator workflow timelines. In section V-C we present the results and discussion.

## II. RELATED WORK

Convolutional Neural Networks (CNNs) have proven to provide a powerful foundation for automated video analysis, by combining the space and time dimensions of an input and performing convolutions in both dimensions [7]. In computer vision, video classification and activity recognition has been extensively studied on several public benchmarks and is an active area of computer vision research [8], [9].

In medical image analysis, a number of studies have explored video classification and analysis methods for second-trimester fetal ultrasound anomaly scan [10]. These include image-based segmentation [11]–[14], frame classification [15], [16], fetal biometry [17], [18] and tracking [19]. Several recent studies have explored obstetric ultrasound standard plane detection [20]–[23], automated biometry [13], [18], [24], [25], activity captioning [26]–[28] and visual attention modelling [20]. Regarding clinical ultrasound workflow analysis, there are a few early studies that have explored and analyzed clinical workflows using machine learning and data science approaches. Blum et al. [29] used Hidden Markov Models (HMM) to generate and visualize surgical workflows for laparoscopic cholecystectomy. Franke et al. [30] proposed an effective strategy for surgical workflow management of lumbar discectomies and brain tumor removals. The proposed approach predicts the remaining intervention time based on a layered model structure of low-level surgical tasks. Holden et al. [31] used Markov models and SVM to automatically segment workflows for tracked needle interventions collected from ultrasound-guided epidural injections and lumbar punctures. In recent years, deep learning has become one of the standard norms for the analysis of surgical workflows [32]. Twinanda et al. [33] explored the use of deep learning for the recognition of surgical workflows in laparoscopic videos. Wang et al. [34], [35] investigated the clinical ultrasound operators' skills using deep learning methods. This study explored the motion of the probe for the purpose of automatic skill assessment for second-trimester fetal ultrasound Sharma et al. [36], [37] recently proposed a spatio-temporal CNN model for second trimester US partitioning and description which is the work most closely related to this paper. Our paper

extends that work by considering a different trimester (and associated differences in fetal appearance and clinical tasks), but also the video annotation method is different (spatio-temporal CNN and use of transfer learning from a second-trimester pre-trained model).

In contrast, only a few first trimester US image analysis studies have been undertaken with a different focus: automated CRL and NT measurements [38], assessment of the maternal placenta [39], classification of fetal brain images [40] and fetal echocardiography [41]. As exemplars of the state of art, Mathewlynn et al. [42] proposed a fully automated placental volume and vascularity measurements method. The proposed DL-based method presented a standardized ultrasound assessment method for 3D volumetric ultrasound. Qi et al. [43] considered automatic localization of placental structural abnormalities to assess placental health. Sobhaninia et al. [44] proposed a multi-task CNN for automatic segmentation and estimation of the head circumference (HC) using 2D ultrasound images. However, no previous publications have considered automated clinical workflow analysis of first trimester ultrasound video. This paper presents a novel, fully-automatic framework to analyse operator clinical workflow from full-length routine first-trimester fetal ultrasound scan videos. In this framework a new DL architecture, which considers spatio-temporal information and transfer learning, is designed to temporally partition ultrasound videos into semantic partitions. Automatic semantic partitioning is the process of segmenting ultrasound videos into semantically meaningful temporal segments. The semantics in our case refer to describing which anatomy is being scanned. The information extracted from labelled scans is employed for large-scale clinical workflow analysis, including knowledge representation, operator clinical workflow analysis, operator skill characterization and variability analysis.

## III. METHOD

### A. Dataset Description

Routine clinical first trimester fetal US scans were recorded as part of the large-scale single site clinical ultrasound study called PULSE [45]. The study was approved by the UK Research Ethics Committee (Reference 18/WS/0051). Scans were performed at the Oxford University Hospitals NHS Foundation Trust. The demographic information of the PULSE first trimester data is given in Table I. The video was acquired by different operators using a commercial Voluson E8 version BT18 (General Electric Healthcare, Zipf, Austria) ultrasound machine. The PULSE protocol proposed the use of two US machines (Voluson E8 and Voluson E10). However, the Voluson E10 has not used for data collection of first-trimester US scans. The sonographers used an LCD monitor with  $1920 \times 1080$  pixels resolution and a refresh rate of 60 Hz. US videos were recorded at 30 frames per second (fps). On average, a recorded first-trimester US video scan takes  $15.7 \pm 4.2$  minutes, with an average of  $28,237 \pm 7,534$  frames per video scan. A dataset of 250 full-length videos (3900 minutes) was acquired from an equal number of pregnant women. We have excluded irrelevant information from the recorded videos and the final 2600 minutes of data that was

used to assess the workflow. For a more detailed description of the full PULSE acquisition protocol the reader is referred to [45].

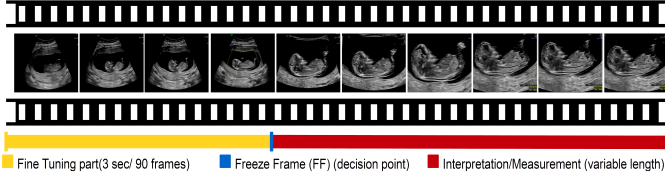


Fig. 1. Illustration of expert annotation process: video frames were annotated as Frozen video segments (blue), measurements (technical annotation) segment (red) and fine-tune segment (yellow). We used the measurements segment (red) for the extraction of the training dataset's technical annotations.

### B. Dataset Annotation

We undertook expert frame-level annotations of fetal anatomy in 95 of the 250 subjects to create a subset of annotated first-trimester video segments. As we had recorded video of the US machine display, all workflow actions were also recorded. As illustrated in Figure 1, a sonographer will scan to identify an anatomy of interest, refine scanning as they approach a standard view, freeze a frame and carry out measurements on the ultrasound video that has been buffered in the machine. A sonographer will then unfreeze the frame and start searching for the next anatomical structure of interest. If a satisfactory standard plane is not seen, the search-freeze-measure procedure is usually repeated. In practice, due to frequent fetal movement and small fetal size in the first trimester, a sonographer may re-visit the same anatomy multiple times in a scan to acquire the best possible view in order to increase the accuracy of anatomical assessment and measurement [46].

The sonographer performs the following actions on the buffered video:

- Diagnostic inspection (including, head, heart, abdomen, optionally other).
- Biometric measurements (NT, CRL, and other anatomical areas such as abdominal or head circumference).
- Doppler or pulse-Doppler based measurements (e.g., heart, maternal uterine artery).
- 3D-mode surface rendering of the entire fetus, fetal face or other anatomy.

A freeze frame (FF) video segment is recorded when a sonographer is satisfied that it is a standard plane. In the continuous recording of the full scan there are significant portions of video unrelated to a fetal standard plane, which in this paper we term "search-time". This is time spent in the searching and refining process (fine-tuning), where a sonographer aims to capture a high-quality view of the relevant fetal anatomy. We added fine-tuning frames to each FF segment (pre-frozen state) to ensure a wide variety of feature maps (Figure 1). We are aiming to train and test the algorithm using a full FF segment as well as pre-frozen video (90 frames) to ensure that it will achieve similar performance on non-FF segments as well. A complete first-trimester video scan can be divided into different FF categories; CRL, NT, Brain, Heart and Abdomen,

which contains fetal biometry measurements and standard plane analysis. FF segments from 95 full-length first trimester

TABLE I

DEMOGRAPHIC INFORMATION OF THE PULSE FIRST TRIMESTER DATA.

Demographic Features	Information
Maternal age (Years)	31.6 $\pm$ 5.4
Smoker at booking	21 (8.5%)
BMI at < 15 weeks (kg/m <sup>2</sup> )	25.3 $\pm$ 5.8
Conception by IVF	2 (0.8%)
Nulliparous	121 (49.2%)
Pregnancy dating by CRL	231 (93.9%)

scans of different subjects were identified by optical character recognition (OCR) and extracted. The average length of each FF segment is 31.8 seconds. We used frames from these sonographers-annotated FFs to build the training and testing dataset (Figure 2) that was used to train the CNNs (section III-E). For spatial-only CNN training, the acquired video was sampled every eighth frame to incorporate a range of anatomical views and spatial diversity for concurrent frames. Figure 3 summarizes the complete dataset used for training, validation, and testing.

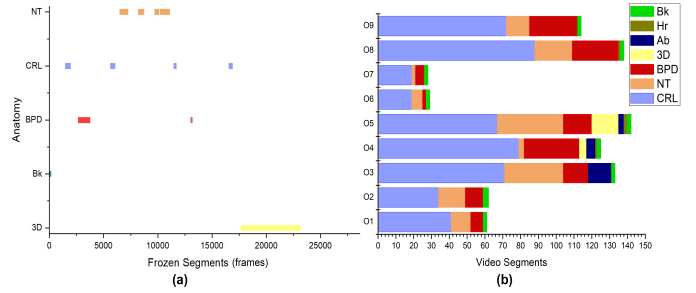


Fig. 2. a. Labeled frozen video segments in a typical sample full-length video US scan. b. Number of expert-annotated short video segments available for each of 9 operators.

There are seven key anatomical categories ["classes"] ("class distribution"): 'Crown Rump Length' [CRL] (48.16%), 'Nuchal Translucency' [NT] (15.51%), 'Biparietal diameter' [BPD] (7.36%), 'Heart' [Hr] (1.63%), 'Abdomen' [Ab] (7.07%), and '3D-mode' (4.19%) and 'Other' [Bk] (16.08%) (Figure 4). 3D-mode is not an anatomical structure however it is an important part of the overall scanning process. Therefore, we wanted to track and evaluate the amount of time spent on scanning in 3D mode. The 'Bk' class includes minority classes (e.g. placenta, etc.). The dataset is divided into the training (77.1%, 73 subjects), validation (17.4%, 16 subjects) and testing sets (5.5%, 6 subjects).

### C. Video Annotation Model

We propose a two-stream CNN architecture, which we call 'PULSE-v', for fetal anatomy annotation in full-length routine first-trimester US scan videos. The proposed network architecture design shown in Figure 5 uses a spatial (2D) supervised transfer learning branch and a spatio-temporal (2D+t) branch.

Full-length First Trimester Ultrasound Scans 95 Unique Videos		
Standard Plane Detection Task 95 Subjects		
Training	Validation	Testing
Subjects=73 (FF=641)	Subjects=16 (FF=145)	Subjects=6 (FF=46)

Fig. 3. Details of dataset used in this study. 95 unique US videos were acquired from 95 subjects.

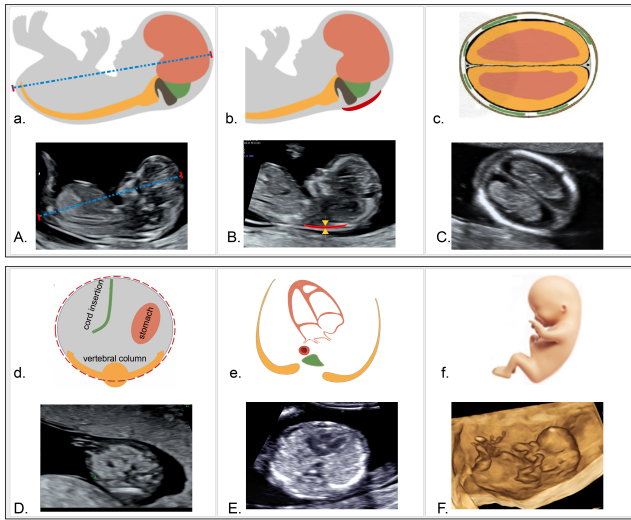


Fig. 4. First-Trimester Fetal Anatomy: A.a Crown Rump Length (CRL), B.b Nuchal Translucency (NT), C.c Brain (BPD), D.d Abdomen (Cord insertion, Ab), E.e Heart (Hr), F.f 3D mode.

The principal design idea for the two branch architecture was to fuse the prior knowledge learned from a second-trimester US feature representation with the spatio-temporal features of a first-trimester representation. To overcome the requirements of large-scale labelled video datasets, we considered transfer learning model fusion with standard plane learning features. Specifically, we investigated using pre-trained weights from a second-trimester model to improve training performance.

**Spatial modelling:** For 2D spatial modelling, we trained and compared a number of CNN architectures. As baselines we used VGG-16, VGG-19 [47], and ResNet-18, ResNet-50 [48] architectures due to their established high benchmark classification performance on public computer vision datasets [49]. We also implemented variants of customized VGG-based models specifically tuned for US images which we refer to as PULSE (architectural configuration shown in Figure 5). We implemented two different configurations of a PULSE model; a PULSE model trained with randomly initialized weights called PULSE(RI); and a PULSE model fine-tuned on second-trimester data weights named PULSE2D(PT). The weights of PULSE2D(PT) were achieved using a VGG16-based network

architecture trained to perform a second-trimester standard plane detection task on 534 second-trimester US scans for 13 standard plane classes: four views of heart, three-vessel and trachea, four-chamber, right ventricular outflow tract, and left ventricular outflow tract, two views of brain, transventricular and transcerebellum, two views of spine, coronal and sagittal, abdomen, kidneys, femur, lips, profile, and the background class.

**Spatio-temporal (2D+t) modelling:** The primary motivation behind designing and implementing a spatio-temporal deep learning method is to utilise the richer spatio-temporal (2D+t) information contained in US video rather than just the 2D standard planes. We consider different approaches to incorporate temporal information and build a spatio-temporal architecture. PULSE spatial CNN was used as a backbone architecture to extract 1D or 2D features from consecutive frames. These features are fed to temporal dependency models: an RNN (PULSE-lstm), a 2D+t CNN (PULSE2Dt) and a multi-stream model (PULSE-v). PULSE-lstm employs long-short-term memory (LSTM) to incorporate temporal information. PULSE2Dt uses 3D convolution kernels for training a 2D+t architecture. For learning temporal dependencies, PULSE-v combines a 2D and a 2D+t branch with a feature fusion unit.

**Feature Fusion:** The fusion of spatio-temporal features is a main challenge in training such models. In order to accomplish our objective, we combined representations constructed from spatial layers (of PULSE2D(PT)) initialized by weights from a large-scale second-trimester dataset and spatio-temporal layers (PULSE2Dt) that were randomly initialized to be trained on the acquired dataset. We investigated the late fusion methodology for the proposed multi-stream framework. The fusion model was composed of a concatenate layer followed by two fully-connected layers in order to reduce the dimensionality of features. Finally, a softmax layer was applied for the final prediction.

#### D. Automated Clinical Workflow Analysis

We follow the analysis approach suggested in [37] to gain understanding of first trimester clinical workflow from the large-scale annotated video dataset.

Specifically we compute:

- **Subject-specific timeline models:** A subject-specific timeline model partitions a full-length US scan into specific scanning events. Each scanning event is dedicated to particular fetal anatomy. The key objective of this analysis is to measure the time spent on each scanning event to determine the accumulative time spent for each anatomical structure by a specific operator.
- **Task frequency:** We use the Apriori algorithm [50] to extract the most frequent anatomical tasks during an US scan. This enables prediction of the most commonly scanned workflows. A Hidden Markov model (HMM) was used to predict the probability of anatomical states and from this we can determine the most frequently visited anatomical structures by a specific operator. A HMM makes use of latent variables in order to deal with



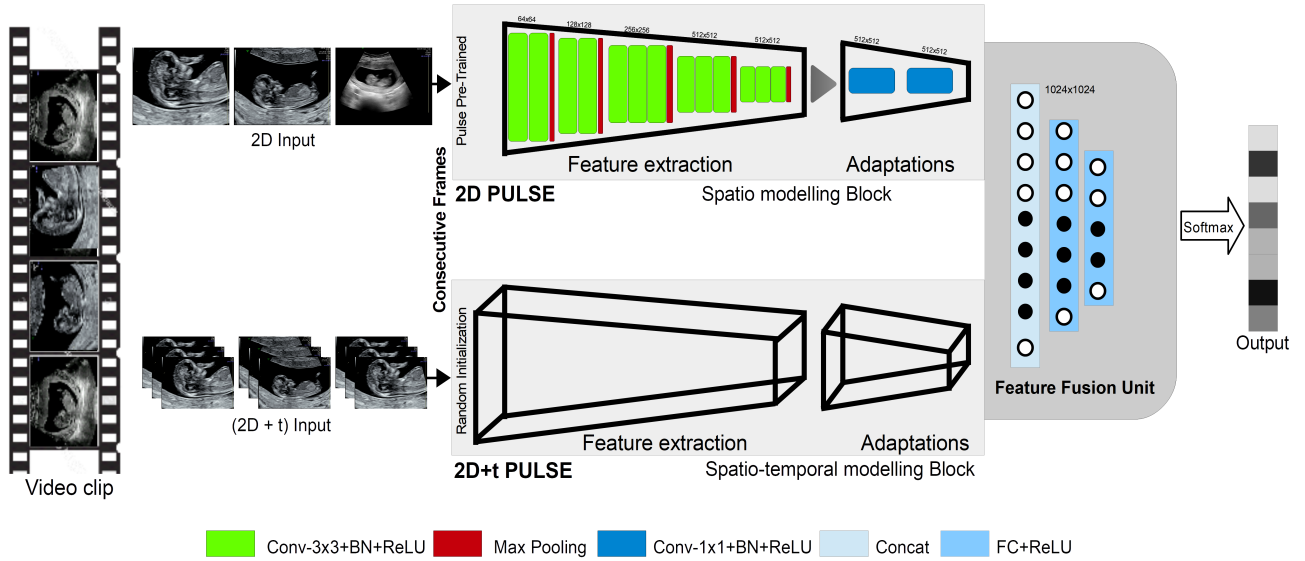


Fig. 5. PULSE-v Model: A spatial-temporal deep learning model with two branches. The 2D (spatial) branch takes a video frame (sampled every 8 frames) and trains using pre-trained weights. The spatial-temporal (2D + t) branch uses a continuous video frames stream (sequence size = 8). The two branches merge at the feature fusion unit to finely classify first-trimester fetal anatomy.

uncertainty and sequential phenomena, making it suitable for addressing a wide variety of biomedical problems [51], [52].

These techniques allow us to describe clinical scanning patterns and preferences for different operators. That can be used to determine skill differences between Experienced (EX) and Newly Qualified (NQ) operators. These representations can also be used to assess the variance between different operators scanning practices as we show later.

### E. Implementation Details

The networks were trained to classify video segments into the 7 classes: CRL, NT, BPD, Hr, Ab, 3D-mode and Bk classes. The CNN architectures were implemented using PyTorch v1.8.0. US video frames were scaled to  $224 \times 224$  pixels. Video frames were extracted and pre-processed to exclude acquisition details like screen commands. Only the area of the screen containing the ultrasound video is used during the training process. Standard data augmentation was used (rotation  $[-30^\circ, 30^\circ]$ , horizontal flip, Gaussian noise, and shear ( $\leq 0.2$ )). Images were normalised to zero-mean and unit variance. The batch size was adjusted according to model size and GPU memory restrictions. We used standard network training configurations for each benchmark (ResNet [48], VGG [47]). In PULSE-net based network configurations, dropout is not used. However, we applied weight decay in order to increase generalizability. The weight decay was set to  $1e-4$ . We used the Stochastic Gradient Descent (SGD) optimizer. All CNN models were trained using a cross-entropy loss function for 200 epochs, constantly reducing the learning rate ( $\times 0.1$  every 20 epochs).

## IV. EXPERIMENTS

### A. Comparison of Video Annotation Models

Recall (R), Precision (P), F1-score (F1), and Top-1 accuracy (A1) were used to assess the performance of the video annotation models. Referring to the quantitative results in Table II, PULSE2D(PT) consistently outperforms the other spatial CNN benchmarks (precision score=0.90). Adding pre-trained weights to PULSE2D(PT) model gives further improvement and the best 2D result (F1-score (3.0%) and A1 (2.84%)) compared to the random initialization (PULSE(RI)). Hence, it was selected as the 2D backbone for the spatial branch of the proposed spatio-temporal CNN architectures.

Table II suggests that the PULSE2Dt has the highest performance (A1 = 95.89%) in spatio-temporal CNN category. With the inclusion of temporal modelling, the spatial-temporal model (PULSE2Dt) performs better (F1-score (5.0%) and A1 (3.84%)) than PULSE2D and also outperforms PULSE-lstm. It appears that using 2D+t data to feed the network is the most natural way of representing the spatio-temporal properties of the US video dataset. Hence, we chose PULSE2Dt as the spatio-temporal branch for the final selected CNN architecture (PULSE-v).

The two-branch model PULSE-v proved to be the best performing end-to-end CNN model for every evaluation metric. The PULSE-v cross-validation evaluation metrics on the test set were  $P = 0.94 \pm 0.05$ ,  $R = 0.95 \pm 0.01$ ,  $F1 = 0.95 \pm 0.03$  and  $A1 = 0.96 \pm 0.01$ . This model translates the spatio-temporal properties of video clips by directly using 2D and 2D+t convolutional and pooling operations simultaneously. The pre-trained weights also boost the performance of the model relative to the RNN-based PULSE+lstm model. We attribute the better performance achieved with PULSE-v due to its ability to learn short-range dynamic features using near consecutive frames (PULSE2Dt) and long-range dynamic features from PULSE2D(PT). Another reason for the efficiency of the proposed model is the less complicated design that combines

TABLE II  
QUANTITATIVE ANALYSIS OF PROPOSED NETWORK.

	Network	Precision ( $\uparrow$ )	Recall ( $\uparrow$ )	F1-score ( $\uparrow$ )	A1( $\uparrow$ )
Spt. Modelling	VGG-16 [47]	0.75 $\pm$ 0.03	0.71 $\pm$ 0.8	0.68 $\pm$ 0.00	72.48 $\pm$ 0.16
	VGG-19 [47]	0.77 $\pm$ 0.23	0.75 $\pm$ 0.01	0.74 $\pm$ 0.41	74.33 $\pm$ 0.07
	ResNet-18 [48]	0.80 $\pm$ 0.03	0.78 $\pm$ 0.11	0.78 $\pm$ 0.49	83.05 $\pm$ 0.01
	ResNet-50 [48]	0.85 $\pm$ 0.35	0.81 $\pm$ 0.10	0.82 $\pm$ 0.01	86.14 $\pm$ 0.05
	PULSE(RI)	0.84 $\pm$ 0.18	0.85 $\pm$ 1.91	0.86 $\pm$ 0.11	89.21 $\pm$ 0.04
	PULSE2D(PT)	<b>0.90<math>\pm</math>0.11</b>	<b>0.90<math>\pm</math>2.60</b>	<b>0.89<math>\pm</math>1.35</b>	<b>92.05 <math>\pm</math>0.11</b>
Spt. Temporal	PULSE+lstm	0.91 $\pm$ 0.22	0.92 $\pm$ 0.18	0.90 $\pm$ 0.15	94.36 $\pm$ 0.05
	PULSE2Dt	0.93 $\pm$ 1.05	0.94 $\pm$ 0.22	0.94 $\pm$ 0.30	95.89 $\pm$ 0.04
	PULSE-v	<b>0.94<math>\pm</math>0.47</b>	<b>0.95<math>\pm</math>0.23</b>	<b>0.95<math>\pm</math>0.18</b>	<b>96.10<math>\pm</math>0.01</b>

a fine-tuned spatial stream with randomly initialised spatio-temporal architecture.

The confusion matrix in Figure 6 depicts the percentage statistical distribution between manually-labelled US scans and automatically-labelled (predicted vs. true label) US scans. Figure 6-b depicts an excellent agreement with manual and automatic semantic labelling. Note that even classes with few samples (e.g. heart, abdomen) are correctly labelled with high accuracy.

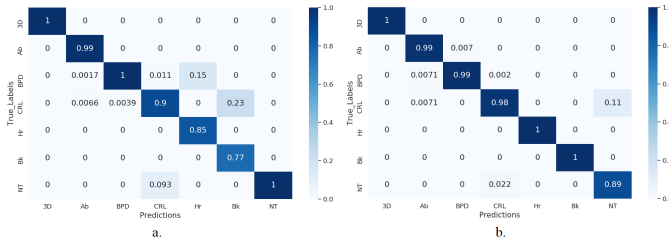


Fig. 6. Confusion matrix for automated semantic annotations vs manual annotations (a) PULSE2D; (b) PULSE-v.

As qualitative analysis, Figure 7 presents the t-distributed stochastic neighbour embedding (t-SNE) visualisation of the penultimate layer of the PULSE2D, PULSE2Dt and PULSE-v models. Observe the confusion between the CRL/NT and Heart/Abdomen classes in Figure 7-a due to their similar appearance and potential co-existence on a standard plane. For the spatio-temporal model (Figure 7-c), feature refinement appears to help to classify anatomical classes into the correct categories. The two-branch model (PULSE-v) with pre-trained weights and spatio-temporal information is a more intuitive solution for US video analysis as it utilises a richer context during training. The automatically-labelled first trimester US scans were validated against manually-labelled test data set and revealed a high correlation Pearson's correlation coefficient ( $\rho = 0.95, p = 0.006$ ). These results established the suitability of the PULSE-v model for the automatic labelling of the rest of the US first-trimester dataset for workflow analysis.

## V. CLINICAL WORKFLOW ANALYSIS

Clinical workflow analysis can provide a simplified US-based workflow model for each subject. In this work a complete US-based workflow is represented as successive temporal events associated with different numerically-coded anatomical structures (class labels), labelled through the validated PULSE-v described earlier in the paper. Statistical tests relevant to each subject, operator and anatomical structure can then be performed.

### A. Anatomical Timeline Model

The automated semantic annotations have been carried out through PULSE-v for full-length US video scans of 250 subjects. We have used temporal regularisation [53] to regularise and smoothen the classification results for each US scan video. The proposed method smooths the results of the classifier by taking into account neighbouring frames. This subject-specific timeline model provides a simplified US scanning workflow timeline, where each video segment is labelled with particular fetal anatomy. This representation is called an Anatomical Timeline Model (ATM) in [37]. An ATM is shown in Figure 8. The transitions between anatomical structures observed in Figure 8 are caused by their existence on similar planes, such as the mid-sagittal view containing NT and CRL. To locate the best plane during the fine-tuning process, an operator may switch opportunistically between these views.

1) *Anatomical Tasks Duration*: The ATMs for the first-trimester dataset provide a relatively low level of abstraction by defining a fine-grained representation of US clinical workflow in terms of distinguishing successive scanning events. This semantic anatomical description of a first trimester ultrasound scan enables assessment of the proportion of time spent performing different anatomical tasks during the scan.

Recall that each US video labelled in our study is a full-length video session acquired through screengrab. This means that it includes screen time when the operator recorded and saved personal details of the subject; and times when the US probe shows no activity during the scanning session which may happen, for instance, when the sonographer is speaking

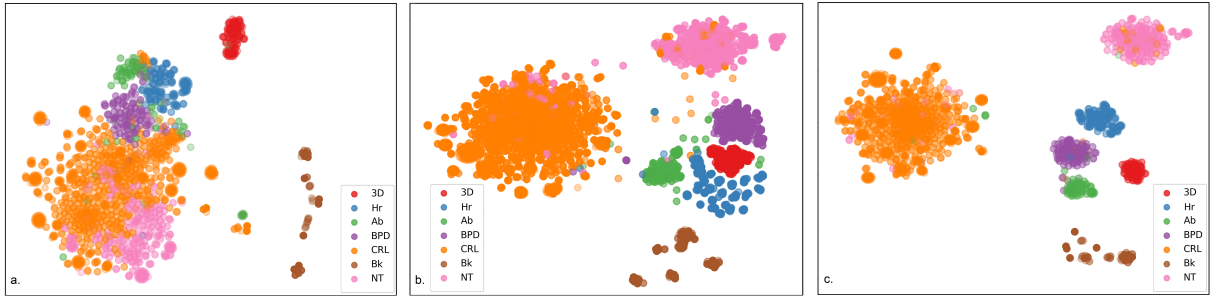


Fig. 7. t-SNE feature visualization of the model penultimate layer of (a) PULSE2D; (b) PULSE2Dt; (c) PULSE-v.

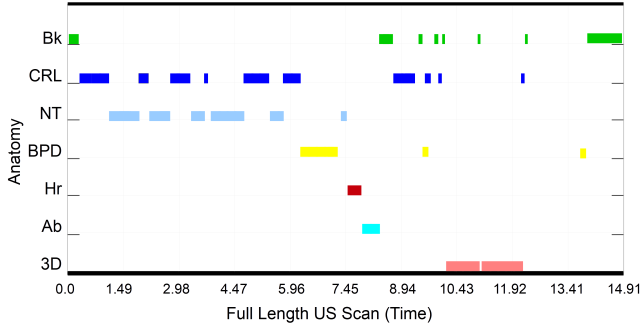


Fig. 8. Anatomical Timeline Model: A sample full length-length US scan labelled using the PULSE-v model.

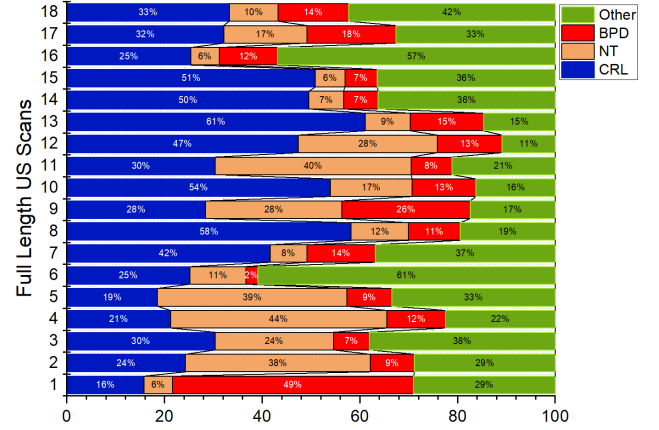


Fig. 9. Anatomical Timeline Model (ATM) for a typical selection of full-length US video scan showing the percentage of time spent on key anatomical tasks (CRL, NT, BPD).

to the subject or colleagues. We excluded these non-anatomical frames from further analysis.

The mean duration of a clinical workflow analysis was found to be 12.4 minutes (interquartile range (IQR) 9.6-19.5 minutes), of which 5.2 (IQR 4.6-9.8) minutes were dedicated to CRL measurement, 2.0 (IQR 1.8-2.0) minutes to NT measurement and 1.5 (IQR 1.9-2.1) minutes to BPD measurement. This shows that approximately one-third of the first-trimester scan duration is dedicated to CRL measurement and the rest to other activities. This finding can be explained by the fact that the mid-sagittal (or more accurately median) view of the fetal profile is a key standard plane viewed during first-trimester US scanning [1], [2]. However, it also represents the fact that several tasks are performed in this plane, including measurement of CRL and NT, and assessing the fetal face, rectangular palate and diencephalon. Figure 9 shows some typical samples of labelled videos, where each video is partitioned into key anatomical tasks carried out during the first-trimester scan.

**2) Operator Specific Tasks Duration:** Our selection criteria included operators with at least 12 full-length video scans. Accordingly, only six operators were selected to measure operator-specific tasks duration. These six operators were divided into two groups based on experience; expert (EX) and newly qualified (NQ). The NQ operators (O1, O3, O4, O6) are qualified sonographers with two or less than two years of experience, and EX (O2 and O5) are operators with more than two years of experience. The choice of two years as the threshold is consistent with [37] and was chosen based on consultation with fetal US specialists ATP and LD.

Scan duration is often thought of as a good surrogate for

skill (i.e. with higher expertise you perform a task quicker). Therefore, we assessed the average time spent by EX and NQ operators on scanning each anatomical structure. The hypothesis was that, as the group EX are more familiar with fetal anatomy than the NQ group, the EX group would have the shorter average scan duration.

The mean duration (interquartile range, IQR) of the EX group was 11.9 (IQR 8.1-17.2) minutes, and the NQ group was 14.7 (IQR 10.3-22.1) minutes, supporting the initial hypothesis. It was also observed that the NQ group took longer to search for and localize different anatomical structures. For example, the average time spent searching and localizing the CRL standard plane for EX operators was 6.9 minutes compared to NQ 9.1 minutes. Another observation is that the NQ operator group spent almost double the amount of time searching for and localizing the NT (Figure 10). The Kruskal-Wallis H test was conducted in order to examine the validity of the hypothesis that there are considerable differences between the NT measurement duration of NQ operators and EX operators. This test results p-value 0.004 that shows a considerable difference among a given set of populations and proves the proposed hypothesis true.

## B. Operator Workflow Analysis

An Operator Workflow Analysis (OWA) maps the anatomical events to model workflow patterns for an individual or group of operators [37]. OWA provides a way to visualise

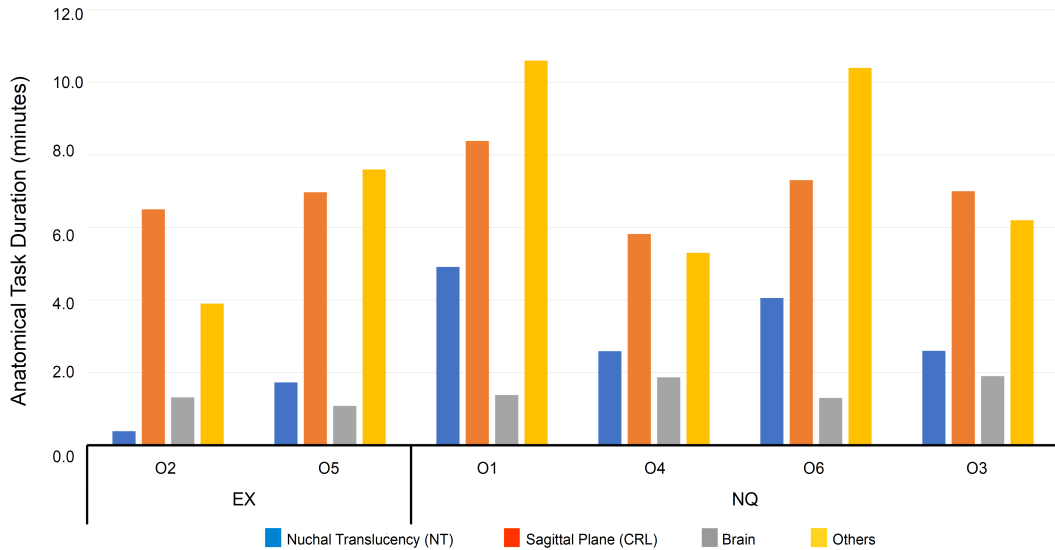


Fig. 10. Average duration of the key anatomical tasks for EX and NQ groups. This shows that the NQ group spends more time locating NT and brain planes. Conversely, the EX group spends less time for the overall scan and identifies NT and brain in a lower average time.

large-scale workflow datasets and, for instance, to extract common scanning practices that may suggest how to standardize future first-trimester US protocols. During first-trimester scans, operators usually scan key anatomical areas involving multiple revisits to essential anatomies (CRL, NT) to be confident in accurate measurement. To map the essential anatomical events, OWAs were designed with six key anatomical classes (NT, CRL, BPD, Ab, Hr, 3D-mode). The Bk class was omitted. It is not a unique anatomical task and accounted for 3.78% of scans. Therefore, excluding the Bk class was not considered to significantly bias the results.

In an OWA initially clinical workflow events are mapped to a connected graph to visualise unique pattern of tasks performed by an operator. We chose a directed relational graph - an Operator Transition Graph (OTG) - showing each sonographer's clinical workflow pattern during the first-trimester scan. To build an OTG we calculate the task transition probability matrix, anatomical task-occurrence probabilities, and anatomical task-start probabilities. These three measures can be calculated from operator-specific ATMs to build a unique operator scanning profile. Assume that there are  $N$  operators (sonographers), and let  $n \in 1, 2, \dots, N$ .

**Task transition probability:** We calculate the task transition probability (TTP) matrix for each scan to measure the probability of transitions between different anatomical tasks, such as  $x$  and  $y$  are two non-identical tasks  $x \neq y$ , and 0 for  $i = j$ . The TTP matrix  $T_n(x, y)$  is stochastic, i.e.  $\sum_{y=1}^m T_n(x, y) = 1$ . A typical transition matrix is shown in Figure 11, illustrating the trend of anatomical scans for O4. We chose O4 due to the diversity of scanning, as the operator acquired a detailed first-trimester scan with all possible anatomical structures.

**Task starting probability:** We were interested to see if an operator always analyses a specific anatomy at the start of a scan or if the sequence of scanning is opportunistic. To study this, the anatomical task-start probability was calculated for each operator and for the overall dataset (all operators)

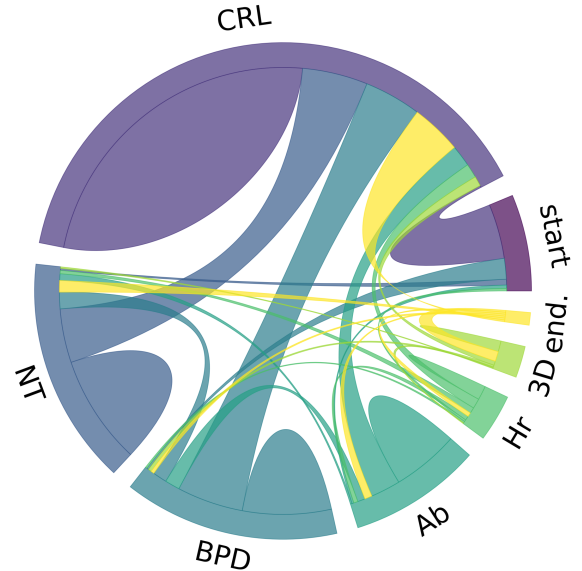


Fig. 11. A visualisation derived from a task transition matrix for operator O4 US workflow. The chord diagram shows a workflow of activities between several first-trimester anatomical structures. Each anatomical structure is represented by a fragment on the outer part of the circular layout. The arcs drawn between these anatomical structures show the workflow patterns of operator O4. The thickness of these arcs is proportional to the frequency of the workflow patterns that O4 follows during US scanning.

by computing the relative occurrence of each task at the beginning of the ATM. The task-start probability for the  $x^{th}$  anatomical task is given as  $P_n(x)$  such that  $\sum_{x=1}^m P_n(x) = 1$ . By analysing the relative occurrence probability of each task in the overall dataset, we selected the three most commonly scanned anatomical tasks (among participating operators) for further analysis: CRL, NT, and BPD.

The Apriori algorithm [50] was used to extract the most frequent starting anatomical tasks; it was also used to obtain the most frequent anatomical task combinations. The Apriori



is a fitting choice to extract the most frequent anatomical task-occurrence probability of each operator's task  $I = \{i_1, i_2, \dots, i_n\}$  and task transition matrix  $T = \{t_1, t_2, \dots, t_m\}$  named as database of anatomical transactions. Here each transaction  $t_x$  in  $T$  has a unique transaction-ID with the subset of item sets in  $I$ . The Apriori rule for any two anatomical activities  $(X, Y)$  stated as  $X \Rightarrow Y$ , where,  $X$  is 'Antecedent' and  $Y$  is 'Consequent' such as  $X, Y \subseteq I$ . Table III shows a different combination of operators' preferences for starting the first-trimester scan. The task starting probabilities with CRL anatomical structure is highest, followed by NT. As expected, it also shows that the anatomical structures close to each other are scanned in succession (i.e. CRL, NT).

TABLE III

THE APRIORI ALGORITHM USED TO EXTRACT THE MOST FREQUENT ANATOMICAL WORKFLOWS ADOPTED BY OPERATORS.

No.	Antecedents	Consequents	Antecedent support (%)
1	(CRL, NT)	(BPD)	96.02
2	(BPD, CRL)	(NT)	86.23
3	(BPD, NT)	(CRL)	58.82
4	(BPD)	(CRL, NT)	88.01
5	(CRL)	(BPD, NT)	75.42
6	(NT)	(BPD, CRL)	19.51

*Task-occurrence probability:* We hypothesized that experienced sonographers have fewer re-visits to key anatomy to re-measure the CRL or NT. Anatomical task occurrence probability can be calculated by calculating the total relative frequency and duration of each task based on the corresponding ATM. The task-occurrence probability for the  $x_{th}$  anatomical task is given as  $O_n(x)$  such that  $\sum_{x=1}^m O_n(x) = 1$ .

Based on computing anatomical task-occurrence probability we found that most operators prefer to view the CRL, the NT, or the brain as the first anatomical tasks. The analysis also showed that these three tasks had the highest task occurrence probabilities (96.02%, (Table III)) which is consistent with clinical expectation as the three anatomies are the most important to assess for any structural and chromosomal abnormalities.

Our modelling shows that most operators prefer to examine Hr, Ab and 3D-mode structures in the latter parts of the first trimester US scan and spend less time looking at them than the other structures. This can be explained by the fact that it is optional and not a protocolised requirement that these anatomical structures are assessed.

**1) Operator Transition Graph (OTG):** Operator Transition Graphs (OTG) are directed graphs  $D_n$  that represent the clinical workflow of each operator. It is composed of nodes (anatomical structures), and edges (a transactional pathway among anatomical structures), such that all nodes are reachable. It also contains self-loops to show the probability of revisiting each node. The definitions (task transition probability, task starting probability, task-occurrence probability) from Section V-B were used to build a database of anatomical transactions represented as an acyclic flow of activities for each operator and scan. Transition probabilities between dis-

crete clinical workflow states could be empirically estimated by the Hidden Markov model (HMM). HMM is a stochastic process that can be parameterized by creating a mathematical representation of underlying ultrasound workflows. The proposed hypothesis is that HMM could predict the most common probabilistic workflow path from a given set of operators US workflow scans. To calculate the most probable path for each sonographer  $n$ , all anatomical scanning pathways  $D_n$  are processed as first-order Markov chains. Hence, the probability of a given path  $x$  from node A to node B is given by  $P(x)$ ,

$$P(x_{t+1} = B \mid x_t = A) \quad (1)$$

Equation 1 describes the probability in the next (time) step  $(t + 1)$  that we transit to state  $B$  from the current state  $A$ . This exemplary construct helps to constitute a typical first-order Markov Model where the next state only depends on the current state. This is the key reason we chose the first-order Markov Model as it is not dependent on the early stages or operators' preferences during the scanning process. From the definition given in Equation 1, we can define the probability of any existing path in the OTG  $D_n$  as,

$$P(path) = P_n(i) \prod_{x,y \in E} T_n(x, y) \quad (2)$$

The final most-probable path for a particular sonographer is calculated from the aforementioned anatomical task-start probabilities, anatomical task-occurrence probabilities, and task transition probability matrix in the following way. The starting node is chosen with the maximum task-start probability. After task initialization, we determine the remaining task ordering. It is hypothesized that OTG will provide the most probable, non-repeating clinical task ordering with the highest probability score. However, multiple paths may have the similar highest path probability value. Thus, to select one of these paths as the most probable path, we calculate the possible pairs of activities and their occurrence probabilities. This provides a map of possible pairs of activities that a particular operator performs in conjunction with. In this way, we set the rest of the activities through this conjunction probability. We use the Apriori algorithm, which builds all possible combinations and task occurrence probability for each operator. This method efficiently discovers complete path sets for each start-node/end-node pair. The path with the highest path probability  $P_{path}$  is determined using Equation 2. Figure 12 shows the most probable paths for five different operators.

Referring to Figure 12 observe that the EX sonographers (O2, O5) follow a systematic scanning approach that likely stems from a mental checklist. It also shows that EX operators tend to scan from CRL to NT that is shown by edges from each node tending to have high probability scores in comparison to the NQ group (O1, O3). The NQ group shows a more opportunistic approach, where their OTG task transaction probabilities are comparably among different anatomical structures.

In addition, operator clinical workflow is also subject to the preferences and priorities of the individual. For example, operator O4 tends to scan comprehensively, taking into account all possible anatomical structures. Based on our analysis,

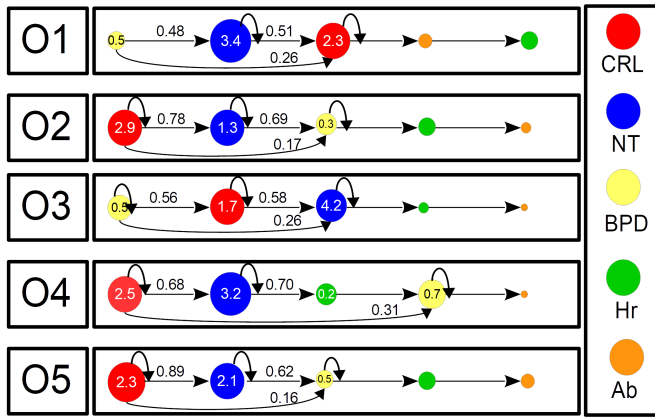


Fig. 12. Operator Transition Graph: The most probable path of each operator, shows the mental map and preference while performing the first-trimester US scan. Also showing each node's anatomical task-occurrence probability (possible revisit to the same anatomical structure to acquire the perfect standard plane).

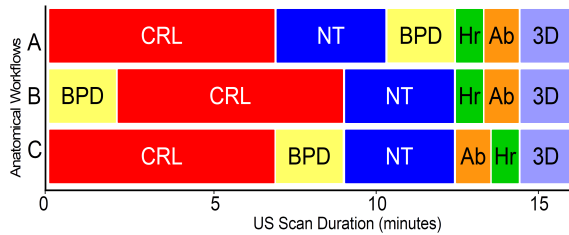


Fig. 13. The most probable general anatomical workflow patterns for 1st-trimester scanning. (A) The anatomical workflow with the highest probability of starting with CRL (Markov chain probability: 61.01%), (B) The second most probabilistic anatomical workflow starting with the fetal brain scan (Markov chain probability: 26.96%). (C) The third most probabilistic anatomical workflow (Markov chain probability: 24.15%).

each operator exhibits a particular pattern and signature for scanning that is repeated in nearly every instance. Based on these statistics, we propose a general anatomical workflow patterns/ mental maps (Figure 13) for first-trimester scanning. This figure illustrates the three most common patterns of scanning (workflows) among US operators, with 'A' being the most commonly practiced workflow followed by 'B' and 'C'.

### C. Clinical Workflow Variability Analysis

**Intra-operator variability for first trimester** In the previous section (Section. V-B) we showed that sonographers scan in different patterns. In this section we analyze operator scanning variability in more detail using anatomical timeline models to assess the variance in task type, order, and time distributions. The 'task type variability' is defined as the standard deviation of the number of unique tasks reported. The 'task order variability' is a measure of variability in the acquisition of anatomical structures, as it is calculated using the standard deviation of the Apriori-based anatomical tasks workflow. The deviation from the mean relative duration of anatomical tasks acquisition is referred to as 'task-time distribution variability' in the first trimester. These variability metrics are reported in the range [0,1] as a normalised deviation from the respective means. Figure 14 reports first trimester US intra-operator variability metrics. These results show the type, order and

time distribution variability of each operator. During the first-trimester US scans, variability among sonographers for type is 23.6%, order is 20.6%, time distribution is 22.2% and overall mean intra-operator variability is 22.13%. Task type and time distribution variability are highest for operators O1 and O3, which may relate to less-developed skills. Specifically, O3 is a newly-qualified operator. It is observed that the average intra-operator variability of the NQ operators is higher than the EX operators.

**Comparison with second trimester** To compare our results with second-trimester operator variance (Figure 14), we chose the same operators as in [37]. Operator 6 did not participate in the second-trimester scanning operations. Therefore, data from that individual cannot be compared in the analysis. The first and second-trimester ultrasound scans differ in length, anatomical classes, and the ultrasound appearance of the same anatomy. For instance, on average, a full-length second-trimester routine US examination usually takes 56.69%; longer than a first-trimester scan. Sharma et al. [37] reported a total of 13 different anatomical classes during the analysis of second-trimester US scans, whereas we reported seven different anatomical structures for the first-trimester US scans.

There was a significant difference in scanning time between NQ and EX groups. As we observed for both trimesters, NQ operators take longer to scan compared to EX operators. The NQ group took 19.88% more time for a full-length second-trimester scan than for a first-trimester scan on average. This could be due to a higher number of anatomical structures being assessed during second-trimester scanning.

Taipale et al. [6] reported a significant 'learning curve' associated with first-trimester anomaly screening. According to Karim et al. [4] greater US scanning sensitivity of fetal anatomy scan could be achieved with the use of a detailed anatomical protocol. Currently, first trimester US scans lack standardised protocols for anatomical screening, and this may cause the higher variability score between first and second trimester US scans. It may also explain the higher intra-operator variability (46.76%) observed during first trimester scans as compared to second trimester scans.

## DISCUSSION AND CONCLUSION

In this study we have investigated clinical workflow during first trimester scanning by large scale data analysis of full-length first trimester US video scans. This work is the first attempt to model clinical workflow on real-world first trimester ultrasound video acquired in a routine fetal screening clinic.

The pre-requisite for such analysis is the semantic temporal partitioning of full-length US scans based on the presence of different anatomies. To this end, we developed a machine learning (ML)-based model (PULSE-v) to automatically partition first trimester full-length video scans, which is subsequently used for operator clinical workflow analysis. PULSE-v is a spatio-temporal CNN architecture transfer-learned to annotate full-length first-trimester US videos. We showed that knowledge transfer from the second-trimester scan to the first-trimester scan improves annotation accuracy. Test set results of automatically versus manually labelled data shows an accuracy

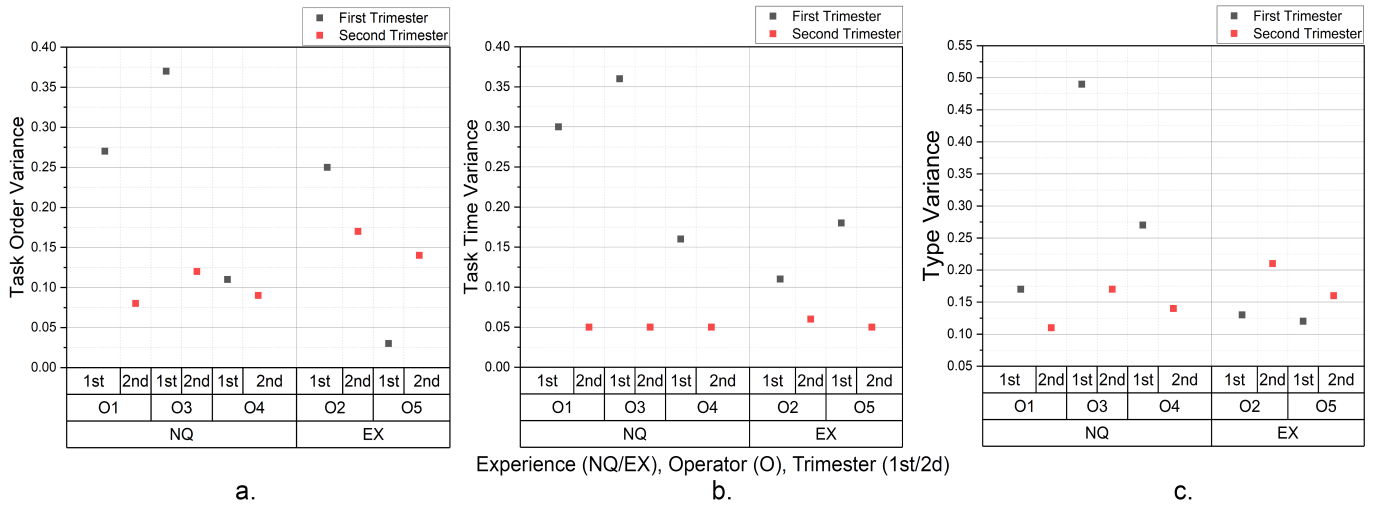


Fig. 14. Intra-operator variability for first and second trimester scanning for 5 operators. a. Task Order Variance, b. Task Time Variance, c. Task Type Variance.

of 96.10% and a correlation of 0.95 ( $p < 0.06$ ). PULSE-v was used as the input to an approach to automated clinical workflow analysis.

We presented subject-specific anatomical timeline models (ATM) that provide a shorthand representation of clinical workflow task duration for a first-trimester US scan and makes observations on large-scale datasets easier and insightful. Our analysis showed that only 50.32% of operators look at the abdomen, heart and 3D-mode (these are optional in guidelines). Similarly, 44% operators did not utilise the 3D/4D US transducer.

Operator workflow analysis (OWA) was used as a quantitative methodology to model and visualise typical patterns of workflow at the individual and group (experienced operator, newly qualified) levels. This showed that EX operators look less into anatomies like Ab, Hr and 3D-mode (which are optional). Some operators (such as O2 and O5) spend less scanning time looking into these three anatomical structures, 11.42% and 25% respectively. On the other hand, NQ operators scan the fetus for longer times, and the majority of NQ operators (95.23%) scanned the Ab standard plane. We observed that detailed imaging of the Heart is the least observed during the first-trimester scan, and this has implications for the detection of such anomalies [5]. Finally, our analysis shows that, on average, NQ operators take 20.24% more time than EX operators to perform a full-length US scan. Most of this time is spent localising the CRL and NT structures. NQ spend 25.56% more time than EX operators on the CRL and NT biometry task.

This study revealed that the time, order, and task variances of the first trimester US scan depend not only on operator experience but also their priorities, and preferences (style) plays a role acquiring US scan. According to Karim et al. [4] the use of standardized anatomical protocol could improve the operator sensitivity for first-trimester US screening. This study provides supporting evidence of this, and tools that could be used to assess future first-trimester clinical workflow standardization protocols.

This study has been limited by the fact that the real-world dataset available for this study is not balanced by anatomy class. The accuracy of clinical workflow analysis could be improved by increasing the number of scans with balanced anatomy acquisitions. However, it should be recognised that we are modeling real-world scans, that are by their nature imbalanced. There is also the possibility that bias in final workflow patterns may result from data acquired from the same institution and using the same imaging device. However, in our case the data is collected following the FASP protocol [1], [2], which is followed by sonographers in the UK NHS. Other international protocols such as the ISUOG [54] guidelines are similar. Thus the scanning protocol would be representative of a typical site.

This study has opened up a number of directions for future possible studies. In the current study, sonographer clinical workflow focused on CRL and NT tasks which are the two primary structures assessed during a first trimester scan. We observed that few operators looked at other anatomical features such as the heart, abdomen, and limbs. It would be interesting to acquire US scans covering a wider range of anatomy and use the methodology reported in this paper to assist in understanding the standardization of first trimester US scans. A second direction of study might utilize the current knowledge obtained from this study to provide the justification for, and subsequent evaluation of assistive tools for first trimester US scanning. For example, it is observed that NQ operators have difficulty localizing complex standard planes (NT). This suggests the need for further research into assistive tools for the navigation and localization of fetal anatomical structures during the first trimester US scan.

#### ACKNOWLEDGMENT

This work is supported by the ERC (ERC-ADG-2015694581, project PULSE), EPSRC (EP/R013853/1 and EP/T028572/1) and the NIHR Oxford Biomedical Research Centre.



## REFERENCES

- [1] "Fetal anomalie screen programme handbook," NHS Screening Programmes, London, UK, Report, 2015.
- [2] S. Alt, A. McHugh, N. Permalloo, and P. Pandya, "Fetal anomaly screening programme," *Obstetrics, Gynaecology & Reproductive Medicine*, vol. 30, no. 12, pp. 395–397, 2020.
- [3] L. J. Salomon, Z. Alfirevic, V. Berghella, C. Bilardo, E. Hernandez-Andrade, S. Johnsen, K. Kalache, K.-Y. Leung, G. Malinger, H. Munoz *et al.*, "Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan," *Ultrasound in Obstetrics & Gynecology (UOG)*, vol. 37, no. 1, pp. 116–126, 2011.
- [4] J. N. Karim, N. W. Roberts, L. J. Salomon, and A. T. Papageorgiou, "Systematic review of first-trimester ultrasound screening for detection of fetal structural anomalies and factors that affect screening performance," *Ultrasound in Obstetrics & Gynecology (UOG)*, vol. 50, no. 4, pp. 429–441, 2017.
- [5] J. Karim, E. Bradburn, N. Roberts, A. T. Papageorgiou, Z. Alfirevic, T. Chudleigh, H. Goodman, C. Ioannou *et al.*, "First-trimester ultrasound detection of fetal heart anomalies: systematic review and meta-analysis," *Ultrasound in Obstetrics & Gynecology (UOG)*, vol. 59, no. 1, pp. 11–25, 2022.
- [6] P. Taipale, M. Ämmälä, R. Salonen, and V. Hiilesmaa, "Learning curve in ultrasonographic screening for selected fetal structural anomalies in early pregnancy," *Obstetrics & Gynecology*, vol. 101, no. 2, pp. 273–278, 2003.
- [7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1725–1732.
- [8] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *ACM Multimedia*, 2015, pp. 461–470.
- [9] Z. Wu, T. Yao, Y. Fu, and Y.-G. Jiang, "Deep learning for video classification and captioning," in *Frontiers of Multimedia Research*, 2017, pp. 3–29.
- [10] T. D. Shipp and B. R. Benacerraf, "Second trimester ultrasound screening for chromosomal abnormalities," *Prenatal Diagnosis: Published in Affiliation With the International Society for Prenatal Diagnosis*, vol. 22, no. 4, pp. 296–307, 2002.
- [11] B. Pu, Y. Lu, J. Chen, S. Li, N. Zhu, W. Wei, and K. Li, "Mobileunet-fpn: A semantic segmentation model for fetal ultrasound four-chamber segmentation in edge computing environments," *IEEE Journal of Biomedical and Health Informatics*, 2022.
- [12] S. Rueda, S. Fathima, C. L. Knight, M. Yaqub, A. T. Papageorgiou, B. Rahmatullah, A. Foi, M. Maggioni, A. Pepe, J. Tohka *et al.*, "Evaluation and comparison of current fetal ultrasound image segmentation methods for biometric measurements: a grand challenge," *IEEE Transactions on Medical Imaging (TMI)*, vol. 33, no. 4, pp. 797–813, 2013.
- [13] J. A. Noble, "Ultrasound image segmentation and tissue characterization," *Journal of Engineering in Medicine*, vol. 224, no. 2, pp. 307–316, 2010.
- [14] J. A. Noble and D. Boukerroui, "Ultrasound image segmentation: a survey," *IEEE Transactions on Medical Imaging (TMI)*, vol. 25, no. 8, pp. 987–1010, 2006.
- [15] S. Gofer, O. Haik, R. Bardin, Y. Gilboa, and S. Perlman, "Machine learning algorithms for classification of first-trimester fetal brain ultrasound images," *Journal of Ultrasound in Medicine*, vol. 41, no. 7, pp. 1773–1779, 2022.
- [16] M. Yaqub, B. Kelly, A. T. Papageorgiou, and J. A. Noble, "Guided random forests for identification of key fetal anatomy and image categorization in ultrasound scans," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 687–694.
- [17] Y. Zeng, P.-H. Tsui, W. Wu, Z. Zhou, and S. Wu, "Fetal ultrasound image segmentation for automatic head circumference biometry using deeply supervised attention-gated v-net," *Journal of Digital Imaging*, vol. 34, no. 1, pp. 134–148, 2021.
- [18] L. H. Lee, E. Bradburn, A. T. Papageorgiou, and J. A. Noble, "Calibrated bayesian neural networks to estimate gestational age and its uncertainty on fetal brain ultrasound images," in *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*. Springer, 2020, pp. 13–22.
- [19] G. I. Sanchez-Ortiz, J. Declerck, M. Mulet-Parada, and J. A. Noble, "Automating 3d echocardiographic image analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2000, pp. 687–696.
- [20] Y. Cai, R. Droste, H. Sharma, P. Chatelain, L. Drukker, A. T. Papageorgiou, and J. A. Noble, "Spatio-temporal visual attention modelling of standard biometry plane-finding navigation," *Medical Image Analysis*, vol. 65, p. 101762, 2020.
- [21] R. Droste, Y. Cai, H. Sharma, P. Chatelain, L. Drukker, A. T. Papageorgiou, and J. A. Noble, "Ultrasound image representation learning by modeling sonographer visual attention," in *International Conference on Information Processing in Medical Imaging*. Springer, 2019, pp. 592–604.
- [22] H. Chen, L. Wu, Q. Dou, J. Qin, S. Li, J.-Z. Cheng, D. Ni, and P.-A. Heng, "Ultrasound standard plane detection using a composite neural network framework," *IEEE Transactions on Cybernetics*, vol. 47, no. 6, pp. 1576–1586, 2017.
- [23] H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang, and P. A. Heng, "Standard plane localization in fetal ultrasound via domain transferred deep neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 5, pp. 1627–1636, 2015.
- [24] L. H. Lee and J. A. Noble, "Automatic determination of the fetal cardiac cycle in ultrasound using spatio-temporal neural networks," in *International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 1937–1940.
- [25] H. P. Kim, S. M. Lee, J.-Y. Kwon, Y. Park, K. C. Kim, and J. K. Seo, "Automatic evaluation of fetal head biometry from ultrasound images using machine learning," *Physiological Measurement*, vol. 40, no. 6, p. 065009, 2019.
- [26] E. Bonmati, Y. Hu, A. Grimwood, G. J. Johnson, G. Goodchild, M. G. Keane, K. Gurusamy, B. Davidson, M. J. Clarkson, S. P. Pereira *et al.*, "Voice-assisted image labeling for endoscopic ultrasound classification using neural networks," *IEEE Transactions on Medical Imaging*, vol. 41, no. 6, pp. 1311–1319, 2021.
- [27] M. Alsharid, R. El-Bouri, H. Sharma, L. Drukker, A. T. Papageorgiou, and J. A. Noble, "A curriculum learning based approach to captioning ultrasound images," in *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*. Springer, 2020, pp. 75–84.
- [28] M. Alsharid, R. ElBouri, H. Sharma, L. Drukker, A. T. Papageorgiou, and J. A. Noble, "A course-focused dual curriculum for image captioning," in *International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 716–720.
- [29] T. Blum, N. Padoy, H. Feußner, and N. Navab, "Workflow mining for visualization and analysis of surgeries," *International Journal of Computer Assisted Radiology and Surgery*, vol. 3, no. 5, pp. 379–386, 2008.
- [30] S. Franke, J. Meixensberger, and T. Neumuth, "Intervention time prediction from surgical low-level tasks," *Journal of Biomedical Informatics*, vol. 46, no. 1, pp. 152–159, 2013.
- [31] M. S. Holden, T. Ungi, D. Sargent, R. C. McGraw, E. C. Chen, S. Ganapathy, T. M. Peters, and G. Fichtinger, "Feasibility of real-time workflow segmentation for tracked needle interventions," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1720–1728, 2014.
- [32] T. Vercauteren, M. Unberath, N. Padoy, and N. Navab, "Cai4cai: the rise of contextual artificial intelligence in computer-assisted interventions," *Proceedings of the IEEE*, vol. 108, no. 1, pp. 198–214, 2019.
- [33] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "Endonet: a deep architecture for recognition tasks on laparoscopic videos," *IEEE Transactions on Medical Imaging*, vol. 36, no. 1, pp. 86–97, 2016.
- [34] Y. Wang, Q. Yang, L. Drukker, A. Papageorgiou, Y. Hu, and J. A. Noble, "Task model-specific operator skill assessment in routine fetal ultrasound scanning," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–8, 2022.
- [35] Y. Wang, R. Droste, J. Jiao, H. Sharma, L. Drukker, A. T. Papageorgiou, and J. A. Noble, "Differentiating operator skill during routine fetal ultrasound scanning using probe motion tracking," in *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*. Springer, 2020, pp. 180–188.
- [36] H. Sharma, R. Droste, P. Chatelain, L. Drukker, A. T. Papageorgiou, and J. A. Noble, "Spatio-temporal partitioning and description of full-length routine fetal anomaly ultrasound scans," in *International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2019, pp. 987–990.
- [37] H. Sharma, L. Drukker, P. Chatelain, R. Droste, A. T. Papageorgiou, and J. A. Noble, "Knowledge representation and learning of operator clinical workflow from full-length routine fetal ultrasound scan videos," *Medical Image Analysis*, vol. 69, p. 101973, 2021.



- [38] R. Singh, M. Mahmud, and L. Yovera, "Classification of first trimester ultrasound images using deep convolutional neural network," in *International Conference on Applied Intelligence and Informatics*. Springer, 2021, pp. 92–105.
- [39] H. Ryou, M. Yaqub, A. Cavallaro, A. T. Papageorgiou, and J. A. Noble, "Automated 3d ultrasound image analysis for first trimester assessment of fetal health," *Physics in Medicine & Biology*, vol. 64, no. 18, p. 185010, 2019.
- [40] S. Gofer, O. Haik, R. Bardin, Y. Gilboa, and S. Perlman, "Machine learning algorithms for classification of first-trimester fetal brain ultrasound images," *Journal of Ultrasound in Medicine*, 2021.
- [41] P. Garcia-Canadilla, S. Sanchez-Martinez, F. Crispi, and B. Bijlens, "Machine learning in fetal cardiology: what to expect," *Fetal diagnosis and therapy*, vol. 47, no. 5, pp. 363–372, 2020.
- [42] S. Mathewlynn and S. Collins, "Volume and vascularity: using ultrasound to unlock the secrets of the first trimester placenta," *Placenta*, vol. 84, pp. 32–36, 2019.
- [43] N. J. Qi H, Collins S, "Automatic lacunae localization in placental ultrasound images via layer aggregation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 921–929.
- [44] Z. Sobhaninia, S. Rafiei, A. Emami, N. Karimi, K. Najarian, S. Samavi, and S. R. Soroushmehr, "Fetal ultrasound image segmentation for measuring biometric parameters using multi-task deep learning," in *Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 6545–6548.
- [45] L. Drukker, H. Sharma, R. Droste, M. Alsharid, P. Chatelain, J. A. Noble, and A. T. Papageorgiou, "Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video," *Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021.
- [46] A. Papageorgiou, S. Kennedy, L. Salomon, E. Ohuma, L. Cheikh Ismail, F. Barros, A. Lambert, M. Carvalho, Y. Jaffer, E. Bertino *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 in Obstetrics & Gynecology (UOG)*, vol. 44, no. 6, pp. 641–648, 2014.
- [47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [48] K. He *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [49] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [50] R. Agrawal, R. Srikant *et al.*, "Fast algorithms for mining association rules," in *International Conference on Very Large Data Bases*, vol. 1215. Citeseer, 1994, pp. 487–499.
- [51] M. L. Bueno, A. Hommersom, P. J. Lucas, and J. Janzing, "A probabilistic framework for predicting disease dynamics: A case study of psychotic depression," *Journal of biomedical informatics*, vol. 95, p. 103232, 2019.
- [52] J. Meier, A. Dietz, A. Boehm, and T. Neumuth, "Predicting treatment process steps from events," *Journal of Biomedical Informatics*, vol. 53, pp. 308–319, 2015.
- [53] M. A. Maraci, C. P. Bridge, R. Napolitano, A. Papageorgiou, and J. A. Noble, "A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat," *Medical image analysis*, vol. 37, pp. 22–36, 2017.
- [54] L. Salomon, Z. Alfrevic, C. Bilardo, G. Chalouhi, T. Ghi, K. Kagan, T. Lau, A. Papageorgiou, N. Raine-Fenning, J. Stirnemann *et al.*, "Isuog practice guidelines: performance of first-trimester fetal ultrasound scan," *Ultrasound in obstetrics & gynecology: the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*, vol. 41, no. 1, pp. 102–113, 2013.