**Appendix S1**

***Data preparation for training***

For the manual labeling, each of the video clips was labeled according to the structure being studied: Head-brain, face axial, face coronal, face sagittal, arms, hands, situs, thorax-heart, abdomen, umbilical cord abdominal insertion, genitalia, bladder, legs, femur, feet, spine, kidneys, full-body fetal sagittal view, placenta-amniotic fluid, maternal anatomy (like uterine artery), 3D/4D face, mixed (two structures or more clearly present in an acquisition), and unidentified. To ensure appropriate representation of each label in the training dataset, we excluded the labels which were poorly represented, defined as labels that appear in less than 2% of the important scan events of the training dataset15. The 12 most commonly used labels were thorax-heart (19.6%), head-brain (11.5%), 3D/4D face (11.2%), mixed (11%), spine (6.1%), abdomen (5.5%), maternal anatomy (5.5%), coronal face (4.4%), placenta-amniotic fluid (4.2%), sagittal face (3.3%), kidneys (2.9%), and femur (2.7%). Labeling was carried out by a clinical obstetric ultrasound expert and seven image analysis scientists with knowledge of obstetric scanning. Overall, the training dataset included 1,158,782 frames. The poorly represented labels were eyes, situs, arms, hands, abdominal cord insertion, bladder, gender, and feet.

A high inter-annotator agreement (76.1%) was found between four expert human annotators. Since our labeling process only allows one label to be assigned, we note that for a few labels, representing structures often seen together, the agreement was relatively low. For example, there was low agreement for the maternal anatomy and placenta as both are often in proximity and therefore visible15.

***Automated video labeling***

A supervised deep learning network architecture for automatic temporal semantic labeling performed the labeling of the 434 previously unseen non-labeled full-length anomaly scans included in the automatic labeling (test) dataset19.

***Labeling testing***

A sample of 28 scans (6.5%) from the automatic labeling dataset was randomly selected and manually labeled. The agreement between manual and automatic labeling was found to be 76.4%. Statistical analysis of the manual and automatically labeled scans showed a high Pearson’s correlation ρ = 0.98 (p < 0.0001). Misclassification was most evident between “mixed” and most other labels15.

***Interpretation***

The high levels of agreement between annotators and similar agreement between manual and automatic labeling provide assurance that the automatic labeling method is appropriate for the current analysis.