Placenta segmentation in ultrasound imaging: Addressing sources of uncertainty and limited field-of-view

Veronika A. Zimmer [a,b,*], Alberto Gomez [a], Emily Skelton [a,c], Robert Wright [a], Gavin Wheeler [a], Shujie Deng [a], Nooshin Ghavami [a], Karen Lloyd [a], Jacqueline Matthew [a], Bernhard Kainz [d,e], Daniel Rueckert [b,d], Joseph V. Hajnal [a], Julia A. Schnabel [a,b,f]

[a] *School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom*
[b] *Faculty of Informatics, Technical University of Munich, Germany*
[c] *School of Health Sciences, City, University of London, London, United Kingdom*
[d] *BioMedIA group, Imperial College London, London, United Kingdom*
[e] *FAU Erlangen-Nürnberg, Germany*
[f] *Helmholtz Center Munich, Germany*

ARTICLE INFO

ABSTRACT

Automatic segmentation of the placenta in fetal ultrasound (US) is challenging due to the (i) high diversity of placenta appearance, (ii) the restricted quality in US resulting in highly variable reference annotations, and (iii) the limited field-of-view of US prohibiting whole placenta assessment at late gestation. In this work, we address these three challenges with a multi-task learning approach that combines the classification of placental location (e.g., anterior, posterior) and semantic placenta segmentation in a single convolutional neural network. Through the classification task the model can learn from larger and more diverse datasets while improving the accuracy of the segmentation task in particular in limited training set conditions. With this approach we investigate the variability in annotations from multiple raters and show that our automatic segmentations (Dice of 0.86 for anterior and 0.83 for posterior placentas) achieve human-level performance as compared to intra- and inter-observer variability. Lastly, our approach can deliver whole placenta segmentation using a multi-view US acquisition pipeline consisting of three stages: multi-probe image acquisition, image fusion and image segmentation. This results in high quality segmentation of larger structures such as the placenta in US with reduced image artifacts which are beyond the field-of-view of single probes.

1. Introduction

Fetal ultrasound (US) is the primary imaging modality to monitor fetal health and development. US is relatively inexpensive and widely available, portable and safe for both mother and fetus. In the UK, all expectant mothers are offered at least two US screening examinations (in the first and second trimester of pregnancy), where the fetus' anatomy and functions are assessed and compared to normal appearances. Mainly 2D US images are acquired, due to their higher resolution, wider availability and ease of acquisition and interpretation compared to 3D US. The rate of anomaly detection in these examinations is highly variable between institutions and sonographers, and significantly below governmental targets for some anomalies and in certain geographical locations (Public Health England, 2020). The main reason for this is that US is a highly operator- and patient-dependent modality (Sarris et al., 2012) and image quality is restricted by the limited field-of-view (FoV) later in gestation, lack of contrast, and view-dependent artifacts.

In recent years, methods from artificial intelligence research, in particular data-driven deep learning approaches, have been successfully investigated to improve fetal screening, for example by automating standard tasks such as detection of standard fetal planes (Baumgartner et al., 2017), estimating fetal biometrics (van den Heuvel et al., 2018; Budd et al., 2019), and investigating the fetal heart (Tan et al., 2020) from 2D US. Further, 3D US (and particularly the combination of multiple 3D views) has been exploited to improve image quality of specific body parts, like the fetal head (Wright et al., 2019) and to extend the field of view (Wachinger et al., 2007; Gomez et al., 2017). The majority of such works focuses on the fetal body, and only few works study the placenta *in utero* (Torrents-Barrena et al., 2019c). Placental assessment during fetal US examination is important for the

identification of pathologies which may be associated with poor fetal and/or maternal outcomes (Fadl et al., 2017). The size, shape and location of the placenta in relation to maternal orientation can be evaluated qualitatively (Salomon et al., 2011), as well as the site and type of cord insertion (Kelley et al., 2020). For example, it has been shown that placental volume in the first (Schwartz et al., 2022) and second (Quant et al., 2016) trimester can be used as a predictor for small-for-gestation (SFG) age birth weight and fetal growth restriction (FGR), as placental growth restriction precedes FGR. However, this does not hold true (especially for first trimester placental volume) for late-onset FGR and preeclampsia pregnancies (Higgins et al., 2016). The authors therefore looked at placentas at late gestation. Pathological conditions such as placenta accreta spectrum (Jauniaux et al., 2018), or lesions including chorioangiomata (Buca et al., 2020), that are likely to require specialist clinical management, may also be visualized.

A full evaluation of the placenta using conventional 2D US, however, is considered to be infeasible beyond the first trimester because of the limited width of the US view sector (Looney et al., 2018; Soongsatitanon and Phupong, 2019; Farina, 2016). As a result, the placenta can only be assessed qualitatively and in segments, which relies on a vigilant operator technique to ensure thorough coverage.

Advances in placental MR imaging including microcirculation assessment (Slator et al., 2018), 3D reconstruction (Torrents-Barrena et al., 2019b) and automatic segmentation (Shahedi et al., 2020) are helping to increase the popularity of fetal MRI as a complementary modality to US in placental evaluation (Prayer et al., 2017). One major advantage of MRI in placental imaging is the larger FoV it affords (Bulas and Egloff, 2013). This enables clinicians to visualize the placenta as a complete structure, facilitating a more coherent and holistic evaluation, and allowing assessment in context to other fetal and maternal structures as well (Miller et al., 2006). Nevertheless, fetal MRI has its own limitations including expense, availability, acoustic noise and sensitivity to maternal and fetal movement, which can degrade image quality (Alansary et al., 2016). Thus, US currently remains the modality of choice for placental assessment during pregnancy. Quantitative assessment of the placenta can be enabled by capturing and segmenting the entire placenta with multiple 3D US images acquired from different views. This is however a difficult task with a number of challenges that need to be addressed.

Automatic segmentations of the placenta are necessary to allow a quantitative assessment throughout the pregnancy. Early works in placenta segmentation in US images have focused on the segmentation of anterior placentas (Stevenson et al., 2015; Oguz et al., 2016). To generalize the segmentation, semi-automatic methods have been proposed in Stevenson et al. (2015) and Oguz et al. (2020). Both methods need a manual initialization to find the position of the placenta in the image. In Oguz et al. (2018), an ensemble of methods is proposed to increase robustness. First, an initial segmentation of the placenta is predicted using a 2D slice, and then a multi-atlas label fusion algorithm is used to provide the full segmentation in 3D.

Convolutional neural networks (CNNs) have recently become the state-of-the-art tools for accurate segmentation (Wang et al., 2020). When a large amount of labeled training data is available, supervised CNN approaches show impressive performance in a variety of medical image segmentation tasks, including good performances for segmenting the placenta in 3D US images (Looney et al., 2018; Yang et al., 2019; Torrents-Barrena et al., 2019a; Zimmer et al., 2019, 2020; Schwartz et al., 2022; Looney et al., 2021). One major drawback is, however, that accurate expert pixel-level annotations are expensive and time-consuming to acquire.

1.1. Remaining challenges in placenta segmentation

Three main challenges have to be overcome: (i) High variability in placental appearance in US; (ii) Intrinsic uncertainty and variability in placenta annotations due to poor US image quality; (iii) limited FoV in

US images, prohibiting whole placenta assessments at late gestation. In the following, we describe these challenges in more detail.

First, we consider variability in appearance. A major factor affecting placenta appearance in US is the location of the placenta. Anterior placentas are located at the front of the uterus towards the mother's abdomen, and posterior placentas at the back of the uterus towards the mother's spine (see Fig. 1, bottom row). Anterior placentas are closer to the US probe, yielding higher contrast between placental and other tissues. On the other hand, the appearance of posterior placentas in US often suffers from shadows (the fetus can lie between the US probe and placenta) and attenuation artifacts. The placenta can be located in any position between the anterior or posterior of the uterine wall with the most common positions being anterior, posterior, lateral and fundal (placentas located at the left or right lateral and top of the uterus, respectively).

Second, we consider variability and uncertainty of segmentations due to poor image quality. US images typically suffer from poor contrast, and view-dependent artifacts, which results in an intrinsic uncertainty for placenta annotation even for clinical experts.

And third, we consider the relatively small FoV of 3D US, which normally cannot capture large structures like the second and third trimester placenta in a single image. Therefore, assessing automatically the whole placenta at late gestation is infeasible with current imaging protocols (Higgins et al., 2016), and it can only be assessed qualitatively in segments.

1.2. Related work

Common strategies in many (medical and non-medical) applications to deal with the lack of large annotated datasets are approaches of transfer, self-supervised and multi-task learning. In transfer learning, information and/or features can be transferred from another image domain, or another task. For the former, one starts with pre-trained models (Shin et al., 2016) (e.g., pre-trained on large natural image datasets such as ImageNet) and then fine-tune the model weights on the new data. The assumption is that the pre-trained weights, even when trained on a different data domain, provide a better initialization for the optimization process during training than random weights, and that fewer data are required to achieve good performance for the final model (Rajpurkar et al., 2020). Another approach is to use self-supervised transfer learning (Shin et al., 2016; Raghu et al., 2019) to adapt the model to a new task. This involves pre-training on the target image domain, but training for a task (the pretext task) which uses different annotations that are already part of the data (or very easy to obtain). In Bai et al. (2019), the prediction of the location of multiple anatomical positions in 2D cardiac MR images was successfully used as a pretext task to boost the accuracy of cardiac segmentation. Here, the transfer learning has been enhanced by a multi-task training strategy, where both the pretext task and the main task are optimized together to achieve the best performance.

In multi-task learning, the idea is to leverage knowledge and information from multiple related tasks to improve performance on all tasks (Zhang and Yang, 2021). The assumption is that related tasks share a common feature representation. This learning strategy is often employed, similar to transfer learning, when the data available for one or all tasks is sparse. Different to transfer learning, the knowledge between all tasks is shared and all tasks are similarly important. In medical imaging, multi-task strategies have been used successfully to detect and correct simultaneously motion-corrupted cardiac MRI sequences during reconstruction (Oksuz et al., 2019), for segmentation and bone suppression in chest X-ray images (Eslami et al., 2020), for the alignment of 3D fetal brain US images and region co-prediction (Namburete et al., 2018), for the segmentation and classification of tumors in breast US (Zhou et al., 2021), and to segment and classify CT images for COVID-19 pneumonia (Amyar et al., 2020), to just name a few.

(a) Single US images   (b) Multi-view US images
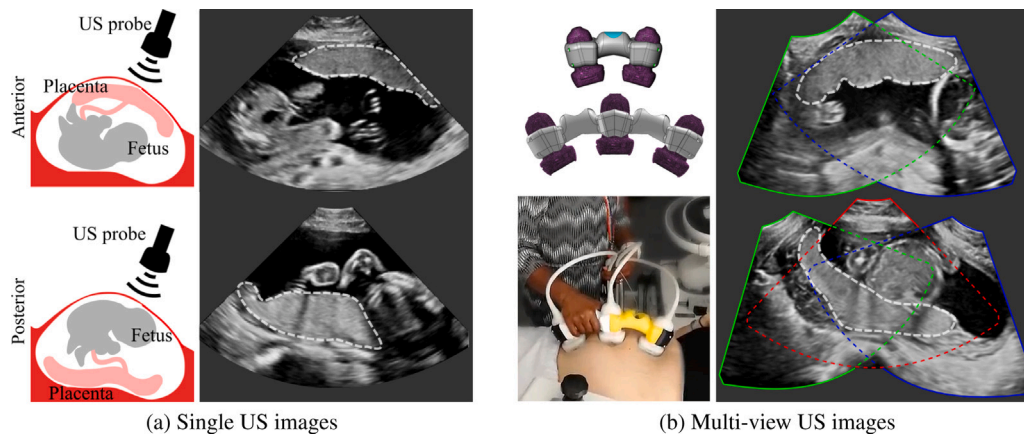
Fig. 1. (a): Examples of anterior (top) and posterior (bottom) placentas in ultrasound (US). (b): Design and implementation of a custom-made multi-probe holder for fetal imaging (left); two- and three-probe multi-view images (right). The placenta is delineated by white dashed lines. (All images are 3D volumes and only central 2D slices are shown.)

To extend the FoV of a single image, multi-view imaging has been previously used. In Wachinger et al. (2007), Ni et al. (2008) and Gomez et al. (2017), registration algorithm and/or tracker information were employed to align the images and provide multi-view US. The resulting image has an extended FoV, and view-dependent artifacts such as shadows can be minimized through the additional signal information from multiple views (Zimmer et al., 2018). In Wright et al. (2019), many different views of the fetal head were registered to a common atlas and fused to provide a detailed, almost tomographic, image of the brain. Aligning US placenta remains however challenging, due to the lack of salient features to drive the registration process, and the high variability in shape, which makes it difficult, if not impossible, to define a placenta atlas space. External tracking, on the other hand, can provide position information of the US probe but is oblivious to maternal and fetal motion.

In general, clinical adoption of segmentation methods requires that clinicians trust the segmentation results. One of the most effective ways to achieve this is by modeling the uncertainty of the estimated segmentations, and communicating this uncertainty to clinicians. Typically, two types of uncertainty are considered: (i) *aleatoric or data/intrinsic uncertainty* and (ii) *epistemic or model/parameter uncertainty* (Kendall and Gal, 2017). The former is caused by the ambiguity and noise inherent in the data itself and is independent of the data used for training. For example, US images are often affected by artifacts and the image quality and contrast can vary greatly. The manual annotation of objects in an image might be therefore ambiguous and rather subjective. Also, the task of manual annotation in 3D images is difficult and their quality is dependent on annotator experience. Previous works have therefore studied the questions *How good is good enough?* or *How good can we actually get?* by looking at inter-rater variability (Joskowicz et al., 2019). Data uncertainty can be incorporated by multiple annotations in the training process as multiple possible labels (Kohl et al., 2018) or as noisy labels (Tanno et al., 2019; Zhang et al., 2020), or estimated using test-time augmentation (Wang et al., 2019). We address the data uncertainty by exploring inter- and intra-rater variability for 3D placenta segmentation in US. The second type of uncertainty, the model/parameter uncertainty, describes the ambiguity in the model parameters, and originates from the data used to train the model. With infinite data, the parameter uncertainty can be neglected. Bayesian approaches have been used to estimate the parameter uncertainty, such as ensemble learning (Kamnitsas et al., 2017; Kohl et al., 2018) and Monte Carlo (MC) dropout as approximation to Bayesian inference (Gal and Ghahramani, 2016).

1.3. Contributions

In this work, we propose a new method to segment 3D US images towards whole placenta segmentation in multi-view images. To achieve this, we address and overcome the three main challenges detailed in Section 1.1.

1. We address the variability in the data by leveraging the information of larger unlabeled data. We propose a transfer and multi-task learning approach combining the classification of placental location and semantic placenta segmentation in a single network to capture data variability in the presence of limited training data;

2. We explore the intra- and inter-rater variability for manual annotation of the placenta in US and study the uncertainty of automatic models. We show that the segmentations obtained by the proposed model lie within the inter-rater variability for manual placenta annotation and that the model shows less uncertainty than baseline models.

3. We describe a multi-view US acquisition pipeline to image larger structures in US as a whole (see Fig. 1(b)). We introduce a new US imaging technique using multiple US probes for the acquisition and fusion of multi-view images. By including the multi-task segmentation model into the multi-view imaging pipeline, we are able to extract whole placentas at late gestation.

In particular, we propose a multi-task approach combining the classification of placental location and semantic placenta segmentation in a single artificial neural network. The location classification as pretext task informs the network about the data variability to improve performance in unfavorable training set conditions for segmentation, which is the clinical downstream task. We discretize the placenta location in three classes: *anterior*, *posterior* and *none*. *Anterior* includes placentas located towards the front uterine wall between the fetus and the US probe, and *posterior* includes placentas located towards the back uterine wall with the fetus and amniotic fluid between the placenta and the tip of the US probe (see Fig. 1(a)). *None* comprises images without placental tissue, independent of the global image label from the corresponding patient.

Since the location of the placenta is typically recorded in fetal screening, training for position classification does not require any additional manual labeling. Hence many more images are available for the pretext task than for the segmentation.

By employing this model in a multi-view US acquisition pipeline, we obtain whole placenta segmentation at late gestation, with significantly better segmentation performances than other UNet-like networks.

This study combines and expands our previous works in Zimmer et al. (2019, 2020). In Zimmer et al. (2019), our multi-view imaging pipeline was described for the first time. Since then, we continued to further improve on the image acquisition process, and we show here new results on a larger data set comprising multi-view images.
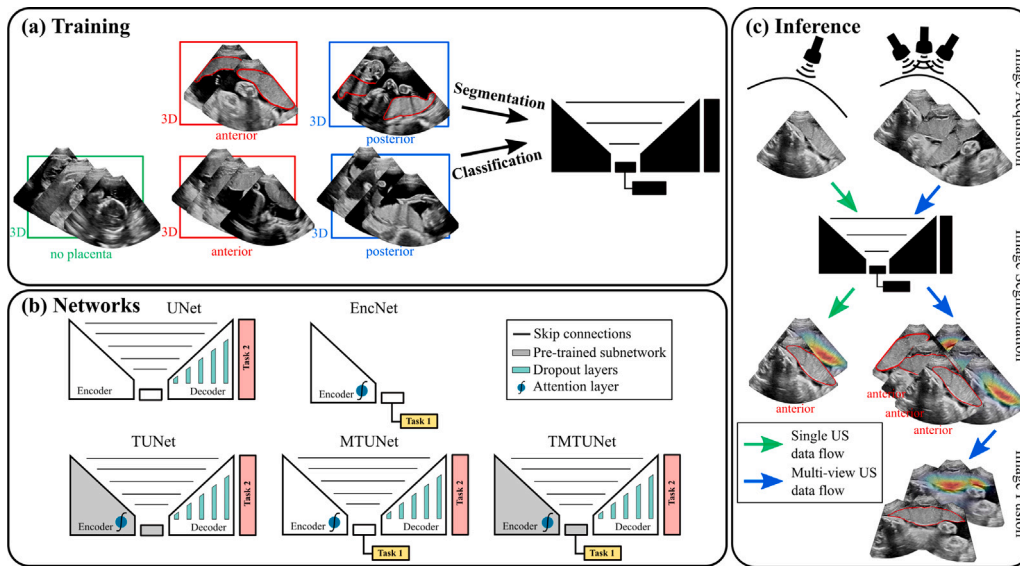
Fig. 2. (a): Training data for multi-task networks using labeled datasets for segmentation and classification (classes anterior, posterior and no placenta); (b): Networks for segmentation (downstream) incorporating information from placental position classification (pretext) in different ways; (c): Inference for single and multi-view ultrasound imaging. (All images are 3D volumes, central 2D slices are shown.)

In Zimmer et al. (2020), we presented a first version of the multi-task model. In this work, we extended the models from 2D to 3D, added Bayesian uncertainty modeling to the UNet architecture, and extended the evaluation and discussion.

2. Methodology

An overview of the entire image segmentation pipeline is shown in Fig. 2. The black box represents any of the models that we compare in this paper, which are illustrated in Fig. 2(b). The pipeline is presented in two parts. First, we describe the multi-task model to segment the placenta using positional information (Section 2.1), and second, we present the multi-view imaging procedure to extract the whole placenta at late gestation (Section 2.3).

2.1. Placenta segmentation and classification

In this section we describe five CNN-based models for segmentation, classification, or both, that are evaluated and compared: *UNet*, *EncNet, TUNet, MTUNet* and *TMTUNet.* These five models are illustrated schematically in Fig. 2(b).

Notation. Let us consider d-dimensional images $I_n : \Omega \subset \mathbb{R}^d \to \mathbb{R}$ and corresponding labels L_n (here class memberships or voxel-wise segmentations). In a fully supervised strategy, the training set $\mathcal{T} = \{(I_n, L_n), n = 1, \ldots, N\}$ contains N pairs of image and reference label, and a CNN model f with parameters Θ is trained to find optimal parameters Θ^* to estimate for an unseen image I its label $\tilde{L} = f(I, \Theta^*)$. During training, a loss function $\mathcal{L}$ is optimized with respect to the parameters Θ. The loss function measures the agreement between reference labels L_n and estimated labels $f(I_n, \Theta)$ over the training set $\mathcal{T}$.

Image Segmentation (*UNet*). We adapt the UNet (Ronneberger et al., 2015; Çiçek et al., 2016) for our segmentation task. UNet has a fully convolutional encoder–decoder structure with convolutional layers, a bottleneck layer in between and skip connections from encoder to decoder. We use a slightly modified version where each layer consists of a residual block with strided convolutions, group normalization and ReLU activations. In the encoder, max pooling is used for downsampling. Dropout is typically used for regularization in CNNs to prevent overfitting. In training, activations of incoming features are randomly removed (with a probability of r). We add dropout with a dropout probability of $r = 0.2$ after each layer of the decoder.

We choose as a loss function $\mathcal{L}_{\text{Seg}}(\tilde{S}, S)$ for training the UNet the sum of the binary cross-entropy loss and Dice loss between the output $\tilde{S}$ of the network and the manual reference segmentations S. This proposed model will be referred to as *UNet* and will form our baseline comparison.

Image Classification (*EncNet*). In image classification, labels L_n are vectors $\mathbf{c}_n \in \mathbb{R}^C$ of class membership for each image I_n. They are defined as $\mathbf{c}_n = \mathbf{e}_c$ if I_n belongs to class c with $c = 1, \ldots, C$ and the cth unit vector $\mathbf{e}_c$.

Our classification (pretext) network has the same structure as the encoder of *UNet* followed by a convolutional block (convolutional layer, layer normalization and ReLU), and a linear block (linear layer, layer normalization and sigmoid activation). We refer to these extra layers as the classification or pretext head. We also incorporate the attention mechanism from Jetley et al. (2018), which not only helps in the interpretation of neural networks by providing visual clues on which image regions are important for the prediction, but also improves final classification accuracy. One attention layer (adapted to 3D volumes) was added after the third layer of the encoder. The trained model $f_{\text{Class}}(I, \Theta_{\text{Class}})$ assigns to an unseen image I two outputs: the class vector $\tilde{\mathbf{c}}$ with predicted class $\tilde{c} = \text{argmax}(\tilde{\mathbf{c}})$, and the attention map $\tilde{A} : \mathbb{R}^d \to \mathbb{R}$, highlighting the region in the image, which most contributed to the predicted classification. We use cross entropy as a loss function for classification, denoted by $\mathcal{L}_{\text{Class}}(\tilde{\mathbf{c}}, \mathbf{c})$. We refer to this model as *EncNet.*

Learning strategies: Transfer and Multi-task Learning (*TUNet, MTUNet* and *TMTUNet*). We explore two different strategies to incorporate the information of unlabeled data or data labeled for a different task in a supervised segmentation network: transfer and multi-task learning.

For transfer learning we use the classification of placental position ($c = 0$: anterior; $c = 1$: no placental tissue in image; $c = 2$: posterior) as a pretext task. Placental position is routinely recorded in each US scanning session and available as meta/clinical data, and does not require any additional expert labeling. Using this strategy, the classification and segmentation tasks are trained sequentially. First, a classification network $f_{\text{Class}}(I, \Theta_{\text{Class}})$ (*EncNet*) is trained on the pretext task. After convergence, the encoder and bottleneck of a UNet $f_{\text{Seg}}(I, \cdot)$

are initialized with the optimized pretrained weights Θ^*_{Class} and further fine-tuned on the downstream task. We refer to this model as *TUNet*.

Another strategy is multi-task learning, where two or multiple tasks are optimized simultaneously. To achieve this for classification and segmentation, we added the classification pretext head after the encoder of the UNet, and added also the attention mechanism to the encoder, as shown in Fig. 2(b). The loss functions $\mathcal{L}_{\text{Seg}}(\tilde{S}, S)$ and $\mathcal{L}_{\text{Class}}(\tilde{\mathbf{c}}, \mathbf{c})$ are combined in a multi-task loss function $\mathcal{L}_{\text{MT}}(\tilde{S}, \tilde{\mathbf{c}}, S, \mathbf{c})$ as

$$\mathcal{L}_{\text{MT}}(\tilde{S}, \tilde{\mathbf{c}}, S, \mathbf{c}) = \mathcal{L}_{\text{Class}}(\tilde{\mathbf{c}}, \mathbf{c}) + \beta \mathcal{L}_{\text{Seg}}(\tilde{S}, S). \tag{1}$$

The parameter $\beta \in \mathbb{R}^+$ is a weighting parameters between classification and segmentation. When $\beta > 1$, it emphasizes the downstream task (placental segmentation) during training.

The multi-task training can be combined with transfer learning by initializing the weights of the encoder and pretext head with the pretrained weights Θ^*_{Class}, and fine-tune the network using both tasks simultaneously using the multi-task loss function in Eq. (1). The multi-task models are referred to as *MTUNet* and *TMTUNet* in the remainder of the paper.

2.2. Variability and uncertainty modeling

We adopt a simple approach towards uncertainty modeling by using dropout at test time. This will allow us to put the uncertainty of the model predictions into context with the inter- and intra-rater variability. In standard dropout, the full activations are used at test time to obtain a single robust prediction. It is also possible (Kendall et al., 2015) to use dropout at test time as an approximation to Bayesian inference. At each test run, activations are removed randomly, yielding multiple possible segmentations for the same image. These can be interpreted as MC samples obtained from the posterior distribution. In the following, we refer to this procedure during test time as MC dropout.

2.3. Multi-view ultrasound imaging

Multi-view placenta imaging with US requires two steps: (i) the image acquisition using multiple probes, and (ii) the multi-view image fusion, see Fig. 1(b) for illustration.

Multi-probe ultrasound imaging. We acquire multiple US images using an in-house US signal multiplexer which allows to connect multiple Philips X6-1 probes to a Philips EPIQ V7 US system. The multiplexer switches rapidly between up to three probes so that images from each probe are acquired in a time-interleaved fashion. The manual movement speeds of the transducer array is within the Nyquist sampling rates. Therefore, for the purpose of data processing, consecutive images are assumed to have been acquired simultaneously over a small time window.

We designed a physical device that fixes the probes in an angle of 30° to each other, which ensures a large overlap between the images (see Fig. 1(b)), and allows easy and comfortable operation. Appendix A.1 with Fig. A.10 describe and show a more detailed illustration of the probe holder design with exact measurements.

Multi-view image fusion. We use a simple, but effective voxel-based weighted fusion strategy to suppress view-dependent artifacts in the images and extend the FoV. First, the images are aligned. This can be achieved via image registration, external tracking information, or fixed multiple probes, as described in the previous section. The weight of a (transformed) data point from each single image is formulated as a function of the depth in the US image with respect to the probe position and the beam angle. In effect, image points with a strong signal (to correct for shadow artifacts) and at a position close to the center of the US frustum (where the quality of the image is typically the best) will receive higher weights. The weighted fusion method is described in detail in Zimmer et al. (2018, 2019). We showed the potential of such acquired and constructed multi-view images for placental volumetry in Skelton et al. (2019).

3. Materials and experiments

3.1. Implementation details

We implemented the models in PyTorch 1.7.1 on a Ubuntu workstation with 48 cores of 3.80 GHz and trained them on a GPU Quadro RTX 8000 48 GB and CUDA 11.1. The code is publicly available.[1]

The hyperparameters and data augmentations for the networks were determined using the validation sets and optimized for *EncNet* (for classification) and *UNet* (for segmentation). We tested different numbers of layers ($\{3, 4, 5\}$) and initial feature maps ($\{4, 16, 32\}$). The best validation performance was achieved using 5 layers with (16, 32, 64, 128, 256) feature maps per layer, both for *EncNet* and *UNet*. For the *EncNet* and the multi-task UNets, we added an attention layer in the third layer of the encoder. A dropout rate of 0.2 is used in the decoder.

The images are resampled to $128 \times 128 \times 128$. We augmented the dataset by flipping the images around the x- and z-axis (an image is not flipped upside down to keep a correct positioning of the frustum), and affine transformations (translation range of 10 voxels, rotations range of 15°, scaling of 10 and shearing of 15 voxels).

All models are optimized using the ADAM optimizer (Kingma and Ba, 2014) and trained until convergence. Convergence was achieved for all folds after 400 epochs (*EncNet*), 100 epochs (*UNet*), 50 epochs (multi-task UNets). For the classification-only *EncNet*, a learning rate of 10^{-5} is employed, for *UNet*, the initial learning rate was 10^{-4} and was reduced by a factor of 0.1 at epochs 30, 70, 90, and for the multi-task and multimodel UNets, the initial learning rate was $5 \cdot 10^{-5}$ and was reduced by a factor of 0.1 at epochs 20, 30, 40.

Since the number of training images for classification differs from the number of training images for segmentation, we follow the training procedure described in Bai et al. (2019). The training alternated between the two different tasks. At each epoch, the task with the higher number of training images, here classification, was optimized for one sub-iteration and the other task, here segmentation, was optimized for β sub-iterations. If $\beta > 1$, a higher weight is assigned to the segmentation task. For our experiments, we empirically chose $\beta = 4$.

The manual reference segmentations for training and evaluation were created using The Medical Imaging Interaction Toolkit (MITK)[2] (Wolf et al., 2005).

3.2. Data

All data were collected as real-time 3D US image streams, on healthy volunteers with a singleton pregnancy (at a gestational age (GA) range of 19–33 weeks). Data were collected under approved institutional ethics (NRES number 14/LO/1806) and all patients were recruited under informed consent. This study was carried out in agreement with the Declaration of Helsinki.

Datasets for classification and segmentation. We collected images from an US examination (duration 30–50 min) of 71 healthy volunteers. A part of the examination were sweeps covering the placenta from different directions. Two expert sonographers (S1 with 10 and S2 with over 15 years of experience) collected the data, and S1 and S3 (with eight years of experience) provided the manual annotations. In 35 patients an anterior placenta is observed and in 32 a posterior placenta. In four patients, only images without placenta visible in the FoV were used. For each patient, 5–30 images were selected, resulting in 1188 images in total, from which 460 show an anterior, 409 a posterior placenta, and 319 show no placental tissue. The images used to train and evaluate segmentation models (see below) were selected from the placental sweeps. Images which are only used in the classification task

---

were collected from different timepoints of the examination and show very different views of the fetus and/or (unavoidable) placental tissue.

We divided the data into two parts. First, the whole dataset $\mathcal{I}^C$ of 1188 images with labels of the classes *anterior*, *posterior* and *none* (no placental tissue in the image), and second, an annotated segmentation dataset $\mathcal{I}^S$ with 292 images and corresponding voxel-wise manual segmentations, manually annotated by S1 from 57 patients. We performed a 5-fold cross-validation where each fold divided the patients into a test, training and validation set. In each fold, approximately 60% of the data $\mathcal{I}^S$ is used for training, and 20% for both validation and testing. Different folds had different amount of images (up to 10%) because of the heterogeneity of the data: each patient had a different number of images, with and without manual segmentations, and with and without placental tissue. However, we made sure that the images from individual patients were not distributed across training/validation/testing sets, the number of training images with segmentations was always the same for posterior and anterior placentas, and that each patient with manual segmentations was exactly once part of a test set. Details about the data distribution in the folds can be found in Table A.5 in the Appendix.

Multi-view data. A subset of the placenta sweeps described above were acquired using the multi-probe acquisition system described in Section 2.3, as follows. On 21 patients, a two-probe and on 32 patients a three-probe holder was used. We selected 1–4 multi-view images per patient which differed in the orientation of the probes with respect to the mother's tummy. This resulted in 32 two-view and 57 three-view images in total. An obstetric sonographer (S1) manually segmented the placenta in all multi-view images (50 images from anterior and 39 images from posterior placentas.)

Datasets for variability and uncertainty. To examine the variability and uncertainty in the segmentations, we created two additional manual reference segmentations for a subset of 53 images from 12 patients by sonographer S1 (around 1 year after the first set), and by sonographer S3, also an expert obstetric sonographer, but without prior experience with MITK. Also the multi-view images from these 12 patients were manually segmented by sonographer S1 twice. In the following, S1.1 and S1.2 denote the two sets of manual segmentations by sonographer S1. On these additional test sets, we investigated the intra- and inter-observer variability.

On a small subset of the multi-view data (17 two- and three-view images), a set of manual segmentations S1.2 is also created. Additionally, we created a third set of annotation (S1.3) of the same subset by fusing the manual segmentations of S1.1 from the single view images to a multi-view segmentation.

3.3. Evaluation metrics

Segmentation and classification. To evaluate the segmentation performance, we use multiple criteria. To compare pairs of segmentations (an automatic and a manual (reference) segmentation), we report both the *Dice* and *IoU* (Intersection over Union) index as overlap measures, and the robust *Hausdorff Distance* (HD) and the *Average Surface Distance* (ASD) as surface metrics. The conventional HD is the maximum distance between two shapes and highly sensitive to outliers. Therefore, we report a robust HD (RHD), by considering the 95 percentile. The classification performance is assessed using the balanced accuracy, precision and F1-score.

Variability in segmentations. To investigate the inter-/intra-expert variability in manual segmentations, and the uncertainty in automatic segmentations, we use the *Generalized Energy Distance* (GED), as described in Kohl et al. (2018). Instead of comparing pairs of segmentations as the measures Dice, IoU, ASD and RHD, the GED compares two distributions of possible segmentations, here a set of possible automatic segmentations obtained with MC dropout and a set of manual segmentations by different annotators. It is based on a distance metric (IoU

in Kohl et al. (2018) and Dice in Zhang et al. (2020)), and leverages pairwise distances. A detailed definition can be found in Kohl et al. (2018).

To test for significance, we performed a paired Wilcoxon signed-rank test between the results of the baseline *UNet* and the proposed models. We report significance at $p < 0.05$ and compute the effect size r as $r = |\frac{z}{\sqrt{N}}|$, where z is the test statistic and N is the number of paired samples. We consider the effect size as small when $r \leq 0.3$, moderate when $0.3 < r < 0.5$ and strong when $r \geq 0.5$ (Cohen, 2013).

3.4. Experiments

We perform two sets of experiments analyzing (i) classification and segmentation performance, and (ii) variability and uncertainty in manual and automatic segmentations.

(i) Placenta classification and segmentation. In the first set of experiments, we compare classification and segmentation performance of different variants of the models described in Section 2, for both individual and multi-view images. We trained all models for segmentation (downstream task) on three different training and validation sets: set A, set P and set AP. In set A, only images with anterior placentas are used for training and validation, in set P only images with posterior placentas, and in set AP both types of images are used. The models are tested on both types of placentas. In the following, we use the term *in-distribution* data (InD) for images whose class was part of the training set (anterior for set A and posterior for set P) and *out-of-distribution* data (OoD) for images whose class was not part of the training set (posterior for set A and anterior for set P).

For classification (pretext task), the baseline *EncNet* is trained on the full classification data $\mathcal{I}^C$. In the multi-task training, we restricted the number of training images for classification to avoid a large difference in numbers between the training data for the pretext and downstream tasks. Next to the 180 images with manual segmentations, we added 90 images without placental tissue and with label *none* for a balanced training set for classification.

The models are tested on the complete test sets both for classification and segmentation and compared for the performance on the individual US images. As described in Section 2 B, the resulting segmentations are then aligned and fused to obtain segmentations of the multi-view images.

(ii) Variability and uncertainty. In a second set of experiments, we investigate the inter- and intra-rater variability of the manual segmentations and compare the variability and uncertainty in automatic segmentations. We measure the variability on a subset of the test data, for which three manual annotations are available, as described in Section 3.2. The intra-rater variability is the agreement between S1.1 and S1.2 and the inter-rater between S1.1 and S3. We compare the automatic segmentation to S1.1 (intra) and S3 (inter). The agreement between pairs of segmentations is measured using Dice, IoU, ASD and RHD.

To assess the general uncertainty for placenta annotation, we compare the distributions of segmentations obtained by manual annotators and by an automatic model using GED_{Dice} and GED_{IoU}. We compare for each training set set A, set P and set AP the baseline *UNet* to the best performing multi-task models. We used MC dropout during test-time to obtain a set of possible segmentations for each image.

(iii) Downstream task: placental volume analysis. As downstream analysis, we extract and compare placental volume from manual and automatic multi-view placenta segmentations. Additionally, we relate the volume extracted from three-probe placenta imaging with reference values throughout gestation obtained from MRI images, as reported by León et al. (2018).
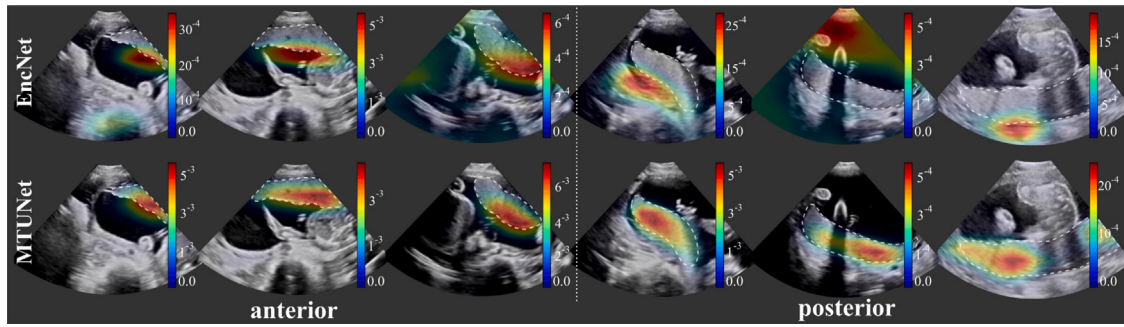
Fig. 3. Attention maps (Jetley et al., 2018) obtained by model *EncNet* (top row) and *MTUNet* (bottom row) trained on set *AP* for both anterior (columns 1–3) and posterior (columns 4–6) placentas. The placenta is delineated by a white dashed line. *EncNet*'s attention lies at the boundary of the placenta and surrounding tissue, *MTUNet*'s on the placenta itself. (All images are 3D volumes, central 2D slices are shown.)
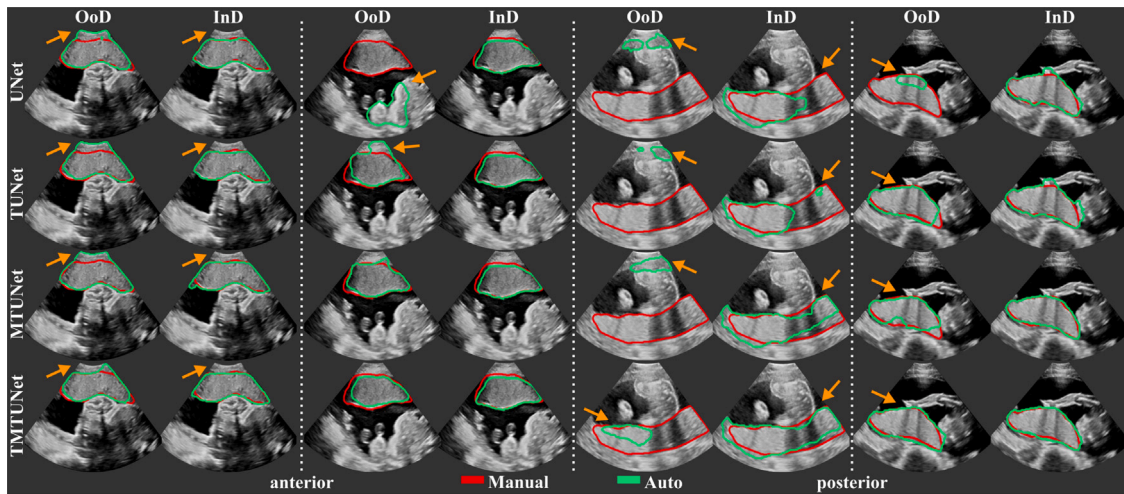

Fig. 4. Examples of automatic placenta segmentations obtained by models *UNet*, *TUNet*, *MTUNet* and *TMTUNet* for pairs of *in-distribution* (InD) and *out-of-distribution* (OoD) test data. The orange arrows indicate areas with segmentation errors and differences between the models. (All images are 3D volumes, central 2D slices are shown.)

4. Results

We first present three types of results: (1) placenta classification and segmentation when using individual images, (2) when using multi-view data, and (3) variability of the annotations and uncertainty of the segmentations.

4.1. Placenta classification and segmentation

4.1.1. Individual images

Classification. The classification results (balanced accuracy, precision, F1-score) obtained by all models are reported in Table 1 and examples of attention maps are shown in Fig. 3. The model *EncNet* trained on the full classification training set of $817 - -840$ images (depending on the fold), is a strong baseline and achieved high performances on all three measures, and in particular a precision of 0.91, 0.90 and 0.88 for the classes anterior, none and posterior, respectively.

Although the training set for classification is 73.41% smaller for the multi-task models, their performance on this task is competitive with the baseline *EncNet* trained on the full training sets. Both multi-task models outperform the baseline for classes anterior and posterior, suggesting that the additional segmentation task has an influence on the performance on the pretext task. This is also confirmed by the better performance of the models trained on the segmentation set *AP*. As an example, the model *TMTUNet* achieved a balanced accuracy of 0.90 for class anterior when trained on set *A*, and 0.94 when trained on set *AP*. The difference between the models is that the latter uses also manual

segmentations of posterior placentas during training, and this increases the performance of the classification of anterior placentas. A final observation is that *EncNet* performs better for class *none* (precision and F1-score) than the multi-task models, which can be explained by the larger number of training images, and that this class is not considered in the downstream task.

We show attention maps obtained by models *EncNet* and *MTUNet* in Fig. 3. In *EncNet*, the model's attention lies rather at the boundary of the placenta and surrounding tissue/space than on the placenta itself. The additional training on segmentation in model *MTUNet*, yields attention maps with good placenta localization.

Segmentation. The segmentation performance of the different models measured by Dice, IoU, ASD and RHD are reported in Table 2 and representative segmentations comparing InD and OoD examples are shown in Fig. 4 with further examples in Fig. B.11. Results using different training and validation sets suggest that anterior and posterior placentas represent two different distributions in the data. The baseline *UNet* trained on set *A* (only anterior) achieves a high Dice score of 0.84 for the InD test set (anterior), but performs poorly on the OoD set (posterior) with a Dice score of 0.26. When trained on set *P* (only posterior), the Dice score for the InD set (posterior) is 0.79, and 0.63 for the OoD set (anterior). The performance on the OoD sets is reduced with a higher standard deviation, indicating that the sets *A* and *P* alone are not representative enough for the segmentation of all types of placenta. These results confirm also that it is easier to segment anterior placentas, which achieve both a higher InD and OoD Dice score. The same trend is observed for the other performance metrics (IoU, ASD, RHD) and models (*TUNet*, *MTUNet*, *TMTUNet*).

Table 1

Classification performance measured by the balanced accuracy, precision and F1-score for classes anterior, none and posterior. The baseline classification model *EncNet* is compared to the multi-task models trained both on classification and segmentation (*MTUNet* and *TMTUNet*). These models are trained on different sets for segmentation: set A (only anterior), set P (only posterior), set AP (both). Bold values indicate best performance on the corresponding class over all models. Gray boxes indicates best performance for each training set.

Train set	Model	Classification performance								
		Balanced accuracy			Precision			F1-score		
		Anterior	None	Posterior	Anterior	None	Posterior	Anterior	None	Posterior
	EncNet	0.91	0.90	**0.93**	0.91	**0.90**	0.88	0.91	**0.90**	0.90
A	*MTUNet*	0.83	0.75	0.74	0.83	0.75	0.74	0.83	0.75	0.74
A	*TMTUNet*	0.90	0.89	0.84	0.91	0.81	0.87	0.90	0.84	0.85
P	*MTUNet*	0.82	0.77	0.86	0.82	0.76	0.86	0.82	0.779	0.86
P	*TMTUNet*	0.92	0.89	0.90	0.92	0.87	0.88	0.92	0.88	0.89
AP	*MTUNet*	0.91	**0.93**	0.90	0.91	0.87	**0.91**	0.91	0.90	0.90
AP	*TMTUNet*	**0.94**	0.89	0.91	**0.94**	0.87	0.90	**0.94**	0.88	**0.91**

Table 2

Segmentation performance for **single-view data** measured by the Dice score, Intersection-over-Union (IoU), Average Surface Distance (ASD) in mm, Robust (95%) Hausdorff distance (RHD) in mm. The baseline (*UNet*) is compared to transfer-based (*TUNet*) and multi-task learning-based (*MTUNet* and *TMTUNet*) models. Showing performance when training on different sets: A (only anterior), P (only posterior), and AP (both). The bold values indicate the best performance of the corresponding class over all models. Gray boxes indicate significance compared to *UNet* (baseline) with a $p < 0.05$ with effect sizes small, moderate (*) and strong (**).

Train set	Model	Dice		IoU		ASD (mm)		RHD (mm)	
		Anterior	Posterior	Anterior	Posterior	Anterior	Posterior	Anterior	Posterior
A	*UNet*	0.84 (0.12)	0.26 (0.29)	0.74 (0.14)	0.19 (0.23)	3.09 (7.26)	33.75 (28.03)	10.99 (14.13)	66.12 (39.86)
A	*TUNet*	0.85 (0.10))	0.41 (0.30)**	0.75 (0.12)	0.30 (0.25)**	2.69 (3.79)	24.03 (29.33)*	10.51 (12.35)	52.54 (39.05)*
A	*MTUNet*	**0.86 (0.09)**	0.27 (0.29)	**0.76 (0.12)**	0.19 (0.23)	**2.23 (1.95)**	34.31 (29.01)	**9.08 (7.76)**	68.24 (39.02)
A	*TMTUNet*	0.85 (0.11)	**0.45 (0.29)**	0.76 (0.12)	**0.34 (0.26)**	2.78 (5.81)	**19.97 (23.41)**	10.33 (11.69)	**48.89 (37.06)***
P	*UNet*	0.63 (0.33)	0.79 (0.10)	0.52 (0.29)	0.67 (0.12)	15.44 (22.74)	4.61 (7.10)	33.68 (32.93)	17.05 (16.75)
P	*TUNet*	0.67 (0.29)	0.80 (0.09)	0.56 (0.26)	0.670 (0.12)	12.25 (19.82)	3.96 (2.46)	28.92 (29.57)	**15.36 (11.35)**
P	*MTUNet*	0.67 (0.29)	**0.81 (0.08)**	0.56 (0.27)	**0.68 (0.11)**	12.66 (20.62)	**3.83 (2.46)**	29.24 (31.10)	15.77 (13.17)
P	*TMTUNet*	**0.74 (0.22)***	0.80 (0.10)	**0.62 (0.21)***	0.68 (0.12)	**7.87 (14.01)***	4.43 (8.26)	**23.08 (23.77)***	15.70 (15.24)
AP	*UNet*	0.864 (0.07)	0.78 (0.12)	0.77 (0.10)	0.65 (0.13)	**2.14 (1.74)**	4.89 (7.26)	**8.57 (7.69)**	18.07 (17.33)
AP	*TUNet*	0.85 (0.12)	0.79 (0.12)	0.76 (0.13)	0.66 (0.13)	3.07 (8.59)	4.88 (9.10)	10.13 (14.49)	17.29 (16.94)
AP	*MTUNet*	**0.87 (0.10)***	**0.80 (0.13)***	**0.77 (0.12)***	**0.68 (0.14)***	2.62 (7.05)	4.73 (8.71)	9.41 (12.06)	**16.50 (16.84)**
AP	*TMTUNet*	0.86 (0.10)	0.79 (0.11)	0.77 (0.12)	0.67 (0.12)	2.67 (6.58)	**4.67 (7.61)**	9.49 (12.26)	17.30 (16.67)

With the incorporation of the classification task with additional training data in models *TUNet*, *MTUNet* and *TMTUNet*, the segmentation performances increase on the OoD data (posterior for set *A* and anterior for set *P*). In particular, it can be observed that with transfer learning on set *A*, i.e., the initialization of the encoder weights with *EncNet*, our method yields a statistically significant (moderate and strong effect size) performance increase from a Dice of 0.258 (baseline *UNet*) to 0.409 (*TUNet*) and 0.450 (*TMTUNet*). The best OoD performance are achieved with model *TMTUNet*. For the InD data, the additional training data for classification, whose information is incorporated in models *TUNet* and *TMTUNet* via weight initialization, is not crucial and the performance increase is not statistically significant. On these data, the best performances are achieved with model *MTUNet*.

When trained on set *AP*, which is representative for both anterior and posterior placentas, good performances are achieved on both classes. The multi-task training improves the segmentation results, and this improvement is statistically significant for the measures Dice, IoU and ASD on all classes with model *MTUNet*, the best performing model.

Notable is that the performance of posterior placentas improve generally more with multi-task learning than the performance of anterior placentas compared to the baseline. As OoD data, posterior placentas improve the Dice score by 74.42%, while anterior only by 17.60% with *TMTUNet*. On the full set *AP*, posterior improve by 2.43% with *MTUNet*, anterior only by 0.35%.

Fig. 4 visualizes examples comparing the segmentation when the images was InD or OoD data. Multi-task models, especially *TMTUNet* (row 4) show a more robust performance with respect to OoD data. For example, *UNet* tries to segment a posterior placenta in OoD of example 2 and an anterior placenta in OoD of example 3. Also, *MTUNet* and *TMTUNet* are more robust to image artifacts, such as shadows, which is shown in InD of example 3. Further examples can be found in Fig. B.11 in the Appendix.

4.1.2. Multi-view images

When the spatial transformation between multiple images is known, e.g., by using a multi-probe system as described in Section 2 for image acquisition, the segmentations in individual images can be combined to obtain the segmentation in the multi-view image. The multi-view segmentation performance is reported in Table 3 and representative results are shown in Fig. 5.

We observe that, in agreement with the results on single views, pre-training significantly improves the performance on OoD data, especially TMTUnet, showing a strong effect size. We would like to emphasize the performance increase on OoD data of TMTUnet trained on set P. Compared to the second best model, TUNet, the ASD is improved by 58.1% (11.81 mm to 4.95 mm) and the RHD by 34.8% (29.22 mm to 19.04 mm).

Interestingly, the performance on OoD data is in general higher on the multi-view data than on single view data. We emphasize here again that the segmentations are obtained from the single view image models and then fused for a multi-view image segmentation. The manual annotations are created on the fused images directly. We surmise that the increased performance measured on multi-view OoD data might be due to the artifact reduction in multi-view US.

For the majority of the performance measures, the multi-task model *MTUNet* performs best on both anterior and posterior placentas on the representative training set *AP*. This is statistically significant for the measures Dice, IoU and ASD with a moderate effect size.

Exemplary multi-view images are shown in Fig. 5 with corresponding placenta segmentations with *MTUNet* and combined attention maps. The placenta is better visualized in the multi-view images with reduced image artifacts and an extended FoV. The multi-task model *MTUNet* provides an accurate segmentation and the combined attention maps localize well the placenta. Further examples of multi-view images with corresponding segmentations can be found in Fig. B.12 in the Appendix.

Table 3

Segmentation performance for **multi-view data** measured by the Dice score, Intersection-over-Union (IoU), Average Surface Distance (ASD) in mm, Robust (95%) Hausdorff distance (RHD) in mm. The baseline (*UNet*) is compared to transfer-based (*TUNet*) and multi-task learning-based (*MTUNet* and *TMTUNet*) models. Showing performance when training on different sets of single-view data: A (only anterior), P (only posterior), AP (both) and subsequently evaluated on the multi-view data. The bold values indicate the best performance of the corresponding class over all models. Gray boxes indicate significance compared to *UNet* (baseline) with a $p < 0.05$ with effect sizes small, moderate (*) and strong (**).

| Train	Model	Dice		IoU		ASD (mm)		RHD (mm)	
set		Anterior	Posterior	Anterior	Posterior	Anterior	Posterior	Anterior	Posterior
A	*UNet*	0.84 (0.09)	0.35 (0.31)	0.74 (0.11)	0.26 (0.25)	2.88 (2.92)	26.00 (23.94)	10.77 (11.32)	57.49 (36.39)
A	*TUNet*	0.84 (0.08)	0.53 (0.23)**	0.73 (0.10)	0.40 (0.21)**	3.12 (2.92)	13.25 (9.38)*	12.23 (13.38)	40.01 (21.72)*
A	*MTUNet*	**0.86 (0.07)***	0.34 (0.30)	**0.75 (0.09)***	0.24 (0.24)	**2.41 (1.38)**	26.96 (23.81)	**9.28 (7.02)**	63.00 (38.37)
A	*TMTUNet*	0.85 (0.08)	**0.57 (0.23)****	0.74 (0.10)	**0.43 (0.23)****	2.82 (2.09)	**11.01 (8.15)****	11.78 (11.41)	**37.08 (22.81)****
P	*UNet*	0.63 (0.30)	0.81 (0.06)	0.52 (0.27)	0.68 (0.09)	14.94 (21.63)	4.52 (2.92)	35.75 (33.49)	18.89 (16.62)
P	*TUNet*	0.68 (0.26)	0.81 (0.06)	0.56 (0.24)	0.69 (0.09)	11.81 (19.91)	4.23 (2.54)	29.22 (31.67)	17.03 (14.16)
P	*MTUNet*	0.64 (0.30)	0.81 (0.06)	0.53 (0.26)	0.69 (0.08)	15.89 (24.20)	4.44 (3.18)	36.63 (35.87)	19.05 (18.40)
P	*TMTUNet*	**0.77 (0.12)****	**0.82 (0.06)***	**0.64 (0.14)****	**0.70 (0.08)***	**4.95 (4.25)****	**3.85 (2.35)***	**19.04 (17.93)****	**15.98 (14.72)**
AP	*UNet*	0.86 (0.05)	0.80 (0.06)	0.75 (0.07)	0.67 (0.07)	2.45 (1.25)	4.76 (2.75)	**9.37 (7.02)**	20.55 (16.76)
AP	*TUNet*	0.85 (0.05)	0.81 (0.08)	0.75 (0.08)	0.68 (0.10)	2.49 (1.30)	4.50 (3.20)	9.79 (7.16)	18.32 (16.09)
AP	*MTUNet*	**0.86 (0.04)***	**0.82 (0.07)***	**0.76 (0.07)***	**0.70 (0.10)***	**2.35 (1.42)***	**4.22 (2.57)***	9.48 (9.46)	**17.70 (15.01)**
AP	*TMTUNet*	0.86 (0.05)	0.80 (0.07)	0.75 (0.08)	0.68 (0.09)	2.64 (1.80)	4.74 (3.33)	10.97 (11.21)	19.66 (17.35)
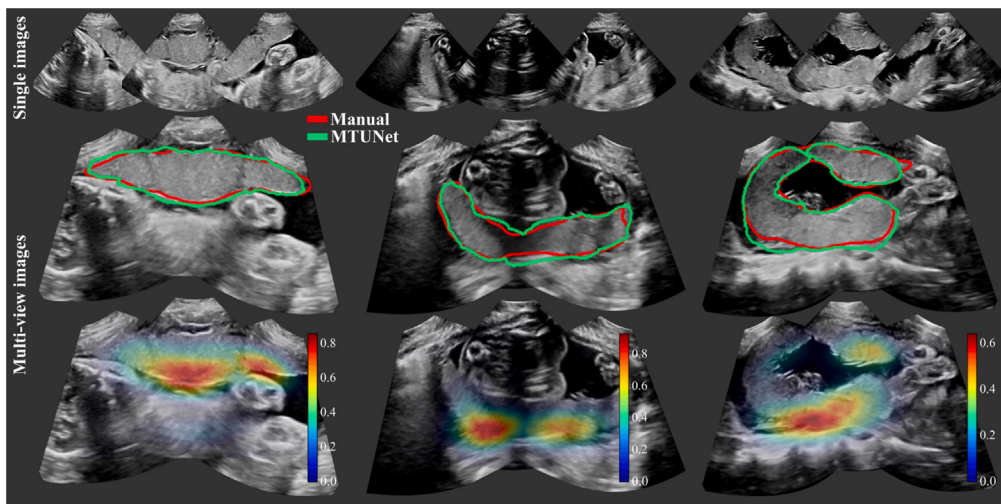
Fig. 5. Three examples of multi-view images, each showing three individual images (top) and fused images with manual (in red) and automatic segmentation (model *MTUNet* in green) (middle) and combined attention maps (bottom). (All images are 3D volumes, central 2D slices are shown.)

4.2. Variability and uncertainty

We investigated the inter- and intra-observer variability for the manual annotation of placental tissue in 3D US. In each fold, we use a subset of the test set, for which three manual annotations are available, as described in Section 3.2. Fig. 6(a) shows the agreement of the segmentations as measured by Dice. We compared the agreement between manual raters S1.1 and S1.2 (intra-variability) and S1.1 and S3 (inter-rater variability), and Fig. 7 shows examples with best and worst intra- and inter-observer agreement. In addition, we assess the agreement between manual and automatic segmentations (*UNet* and *MTUNet*), which are summarized under the term *intra* with reference S1.1 and *inter* with reference S3 in Fig. 6.

Comparing the agreement between manual annotations (plain white bars in Fig. 6), we observe that the intra-observer agreement is higher than the inter-observer agreement for all measures. The difference is statistically significant for anterior placentas with a moderate effect size and for posterior placentas with a strong effect size, denoted by one and two asterisks, respectively, above the bar for inter-rater agreement.

This suggests that the manual annotation of the placenta in US is a subjective task. In all cases and for all measures, the agreement in segmenting posterior placentas is smaller than in anterior placentas, emphasizing that the segmentation of posterior placentas is more ambiguous, possibly due to image artifacts. This is in line with the observation of the previous experiment, that the automatic segmentation models perform worse for posterior than for anterior placentas.

The intra-observer comparison of anterior placentas achieved the best agreement with a Dice of 0.89, an IoU of 0.80, an ASD of 1.70 and a RHD of 12.30. These values can be therefore interpreted as an upper bound and the range between inter- and intra-observer agreement as the desired performance of any automatic segmentation model. For anterior placentas, both the baseline model *UNet* and our best performing model *MTUNet*, as selected in the previous experiment, lie within intra- and inter-rater variability with no significant difference ($p > 0.05$) between the segmentation agreements. For posterior placentas, there is a statistically significant difference (with a moderate effect size) for the baseline model *UNet*, but not for *MTUNet*. The multi-task approach increases the performance and reduces the variance for all measures. The same trend is observed for IoU, ASD and RHD (see Fig. B.13 in the Appendix).

The GED scores for comparing manual and automatic segmentation distributions are shown in Fig. 6(b). For each training set (set A, set P, and set AP) we compare the baseline *UNet* to the best performing model from the first experiment (*TMTUNet* for not representative training dataset A and set P, and *MTUNet* for set AP).

The uncertainty, as measured by GED (based on Dice as a distance measure) of the InD data, both anterior on set A and posterior on set P is small and comparable to the uncertainty obtained with the representative training set AP. There is no statistical significant difference between *UNet* and *TMTUNet* on InD data. On OoD data, the uncertainty and variability increases and is higher for posterior than for anterior
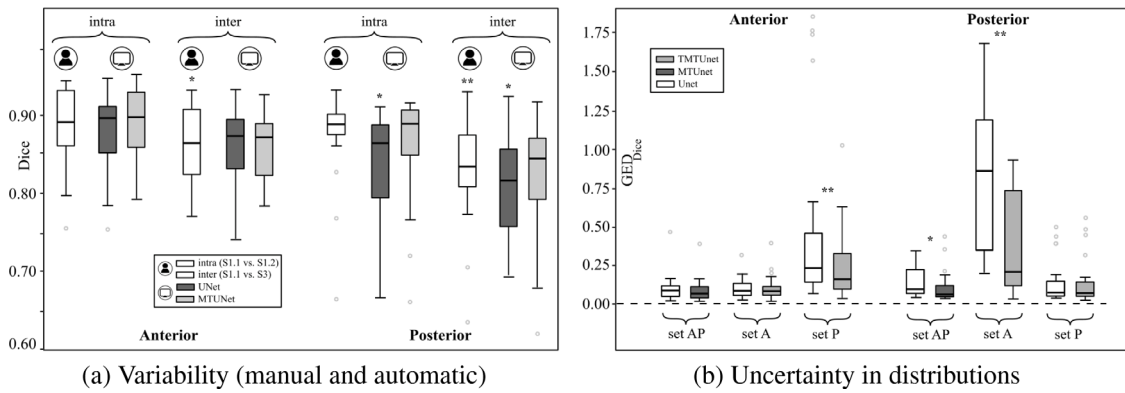
Fig. 6. (a) Variability among manual and automatic segmentations. The agreement of possible segmentation is measures using the Dice score. Manual: S1.1 vs. S1.2 (intra) and S1.1 vs. S3 (inter); *UNet/MTUNet*: S1.1 vs. *UNet/MTUNet* (intra) and S3 vs. *UNet/MTUNet* (inter). (b): The difference in distributions between manual annotations from three raters and automatic segmentations from models *UNet, MTUNet,* and *TMTUNet* with MC dropout is measured by the Generalized Energy Distance using Dice as distance measure. This is compared for models trained on sets *A, P* and *AP* and tested on both anterior and posterior placentas. Statistical significance between *UNet* and *MTUNet/TMTUNet* is indicated by * (moderate effect size) and ** (strong effect size).
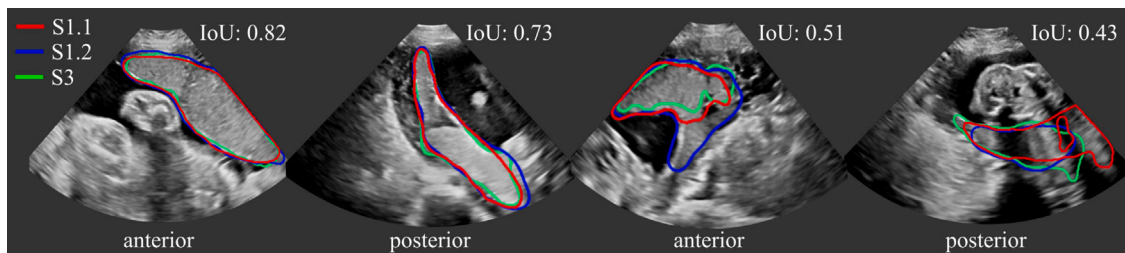
Fig. 7. Manual segmentations S1.1 (red), S1.2 (blue) and S3 (green). All three segmentations agree well in (a) and (b) with an Intersection over Union (IoU) score of 0.82 and 0.73, respectively. Due to strong image artifacts (shadows) and/or low contrast in (c) and (d), the agreement is poorer with an IoU of 0.51 and 0.43. (All images are 3D volumes, central 2D slices are shown.)

placentas. *TMTUNet*, however, obtained significant lower GED scores than *UNet* with a strong effect size both on anterior and posterior placentas. On set *AP*, *MTUNet* shows significantly lower GED scores for posterior placentas compare to *UNet*.

The segmentation performance on this data subset is higher for all measures, classes and models comparing to the performance on the full dataset as reported in Table 2. This suggests that the subset contains images, showing both anterior and posterior placentas, with on average a higher image quality and less artifacts than the full dataset. Thus, we surmise that our observations on variability and uncertainty would be confirmed and even stronger effects could be detected.

4.3. Downstream task: placental volume analysis

Placenta segmentations can be used to extract useful clinical information, such as placental volume. In a last set of experiments, we analyze the volume computed from automatic segmentations obtained with MTUNet when trained on the representative set AP. Fig. 8(a),(b) show Bland–Altman plots placental volume estimates obtained with MTUNet and manual segmentations S1.1 ((a) is color-coded for anterior/posterior and (b) for two-/three-view images). Outliers are mostly posterior placentas, where the image quality is reduced by artifacts. We observe that the majority of two-probe anterior placental volume estimates are relatively small. Anterior placentas are located closer to the probe (where the FoV is very narrow) and tissue is more likely missed even in two-view images.

We compare intra-rater variability with MTUNet in Fig. 8(c)–(e). The intra-rater variability is measured on a subset of the multi-view data, where two manual and one pseudo-manual annotations of rater S1 are available (S1.1, S1.2 and S1.3). For the definition of the pseudo-manual annotation see Section 3.2. We observe from the Bland–Altman

plots, that the differences between MTUNet and S1.1 are comparable to the intra-rater differences (S1.1 vs. S1.2 and S1.1 vs. S1.3).

In addition, we compared the placenta volumes extracted from multi-view images (only acquired with the 3-probe holder) to values of placental volume reported in León et al. (2018) measured in MRI images. The authors found that the equation $f(x) = -0.02x^3 + 1.6x^2 - 13.3x + 8.3$ best describes the volume increase throughout gestation in their cohort. We plot this curve with standard deviations and min/max values reported in León et al. (2018) together with the volumes of our cohort (three-probe holder) in Fig. 9 from (a) the manual annotations S1.1, and (b) the automatic segmentations of MTUNet. We observe a good agreement with the reference volumes S1.1 and the automatic volumes. Overall, there is a good agreement between the volumes from our cohort and the values reported in the literature. However, we observe some outliers (arrows in Fig. 9) of anterior placentas. In these cases, the placenta was close to the probe, where the FoV is very narrow, and the multi-view image does not contain the whole placenta.

5. Discussion

We propose a multi-task approach combining the classification of placental position and semantic placenta segmentation in a single network. Through the classification, the model can learn from larger and more diverse datasets and improve segmentation accuracy, which are comparable to human-level performance. Our results suggest that images of anterior and posterior placentas represent two different distributions in the data. In other words they are OoD data to each other in relation to a placenta segmentation task.

We have shown that multi-task models not only improve significantly the segmentation performance on OoD data, but also the performance when trained on representative data (to a lesser extent). The baseline method, a *UNet* trained on a large dataset including data
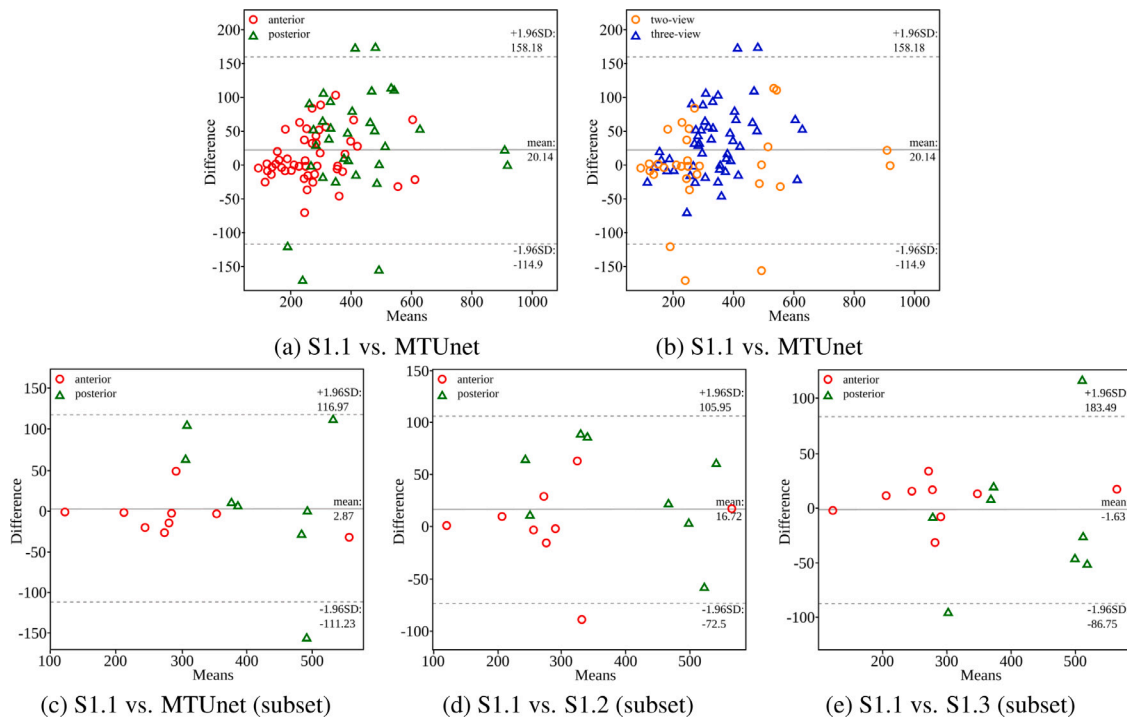
Fig. 8. Bland–Altman plots comparing the placental volume (in mL) extracted from automatic and manual multi-view segmentations. (a)/(b): automatic (MTUNet trained on AP) and manual (S1.1) on the multi-view data color-coded for separating (a) anterior and posterior placentas and (b) two- and three-view images. (c)–(e): Comparison to intra-rater differences on a subset of the multi-view data. (c): automatic (MTUNet trained on AP) and manual (S1.1); (d) intra-rater (S1.1 and S1.2); (e) intra-rater (S1.1 and S1.3). S1.3 is a pseudo-manual segmentation, which is obtained by fusing the manual segmentations from S1.1 for single views.
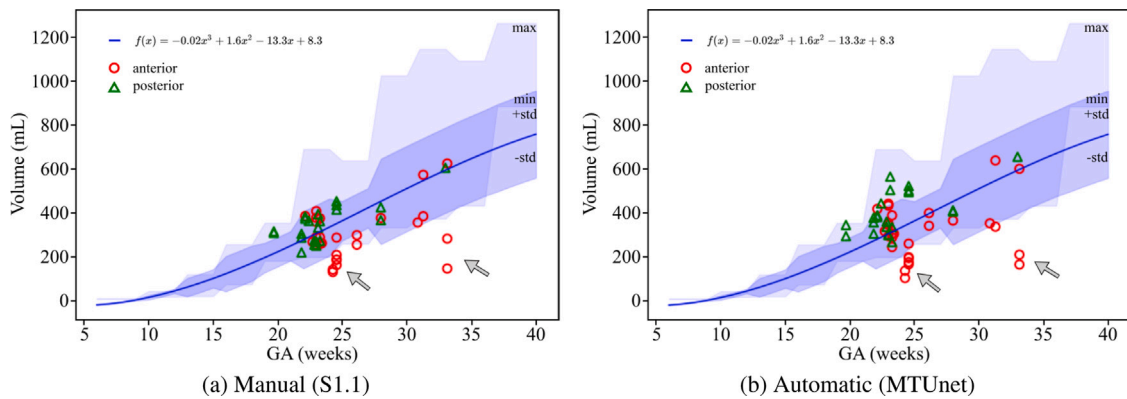
Fig. 9. Comparison of placental volumes (in mL) in multi-view (three-probe holder) images with values reported by León et al. (2018). The curve $f(x) = -0.02x^3 + 1.6x^2 - 13.3x + 8.3$ (blue line) was found to describe best the volume increase throughout gestation in the respective cohort. The shaded area in dark blue indicates the standard deviation and the shaded area in light blue the minimum and maximum placental volumes as reported in Table 2 in León et al. (2018). (a) Manual (S1.1) and (b) automatic placenta segmentations (MTUNet trained on set *AP*) show a good agreement (anterior marked as red circles and posterior as green triangles). There is also a good agreement between the volumes of the cohorts used in León et al. (2018) and in this study. The gray arrows indicate some outliers of anterior placentas, where some tissue is missed by the limited field-of-view close to the ultrasound probe.

from both distributions, can learn reliable segmentations. However, the manual voxel-wise annotation is a difficult, time-consuming and subjective task and therefore availability of such data is not always possible. In unfavorable training set conditions, our multi-task approach achieved up to 70% improvement over the baseline. Overall, the benefits for posterior placenta segmentations were higher, as these are more affected by imaging artifacts. To this end, our multi-task model shares the entire encoder weights for both tasks. This might not be the ideal network structure, as suggested in Guo et al. (2020), where the authors proposed an automated method to learn the best sharing and branching configuration. This would be an interesting avenue for future work.

Our best performing model MTUNet achieves a Dice score of 0.87 ± 0.10 for anterior and 0.80 ± 0.13 for posterior placentas. A direct comparison to performances of other placenta segmentation models reported in the literature is difficult since they are trained and evaluated on different datasets. Table 4 contains a summary of previous approaches with specifications about the training and testing data, the GA of the fetus and the average Dice score achieved for placenta segmentation. The Dice scores vary from 0.64 to 0.92, and the number of data used for training and evaluation from 14 to over 1000. The majority of other works focus on the placenta at the first trimester, and all more recent works (in the last 5 years) employ CNNs. Our segmentation results are comparable to most of these works. Note that only the work

Table 4

Previous work on placenta segmentation in ultrasound with specifications about training and testing data, average performance measured by the Dice score and subjects included in the study. (CNN: convolutional neural network; RNN: recurrent neural network; cGAN: conditional generative adversarial network; CV: average performance obtained in a cross-validation strategy.)

Reference	Method	Dice	Training + Validation	Testing	GA	Subjects
Stevenson et al. (2015)	Random walker, semi-automatic	0.87	–	88	First trimester	3D US, singleton
Oguz et al. (2016)	Multi-atlas label fusion	0.83 ± 0.05	–	14	First trimester	3D US, only anterior
Yang et al. (2019)	Multi-object, 3D CNN + RNN	0.64	50 + 10	44	First trimester (10–14 weeks)	3D US, singleton and twin
Looney et al. (2018)	3D CNN	0.81 ± 0.15	1097 + 100	1196	First trimester (11–14 weeks)	3D US, singleton
Oguz et al. (2018)	2D CNN + 3D Multi-atlas label fusion	0.88 ± 0.05 (anterior) 0.85 ± 0.05 (posterior)	384 slices	73	First trimester	3D US, singleton, 28 anterior 19 posterior
Oguz et al. (2020)	Semi-automatic, Multi-atlas label fusion	0.82 ± 0.06	–	73	First trimester (11–14 weeks)	3D US, singleton, 28 anterior 19 posterior
Schwartz et al. (2022)	2D and 3D CNNs	0.88 ± 0.05	99	25	First trimester (11–14 weeks)	3D US, singleton
Looney et al. (2021)	Single- and Multi-object, 3D CNN	0.85 ± 0.05	1893 + 150	50	First trimester (11–14 weeks)	3D US, singleton
Hu et al. (2019)	2D CNN + shadow detection layer	0.92 ± 0.04	954 + 205	205	First, second and trimester (8–34 weeks)	2D US, singleton and twin
Torrents-Barrena et al. (2019a)	3D cGAN	0.75 ± 0.12	61	61 (CV)	Second and third trimester (15–38 weeks)	3D US, singleton and twin
Ours	3D Multi-task CNN	0.87 ± 0.10 (anterior) 0.80 ± 0.13 (posterior)	1188 (292 with segm.)	292 (CV)	Second and third trimester (19–33 weeks)	3D US, singleton

Two-probe holder design · Three-probe holder design
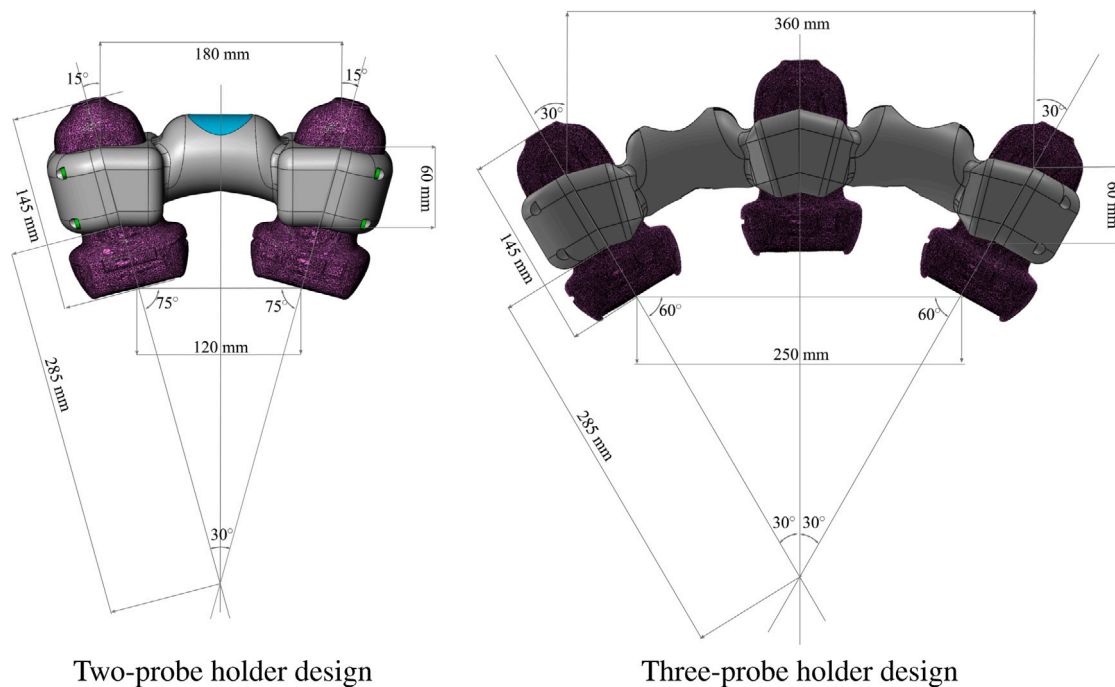
Fig. A.10. Design of a custom-made multi-probe holder for fetal imaging. The probes are fixed in an angle of 30° to each other to ensure a large overlap of the field-of-view. The system is flexible in the sense that it allows the use of two (left) or three (right) probes simultaneously.
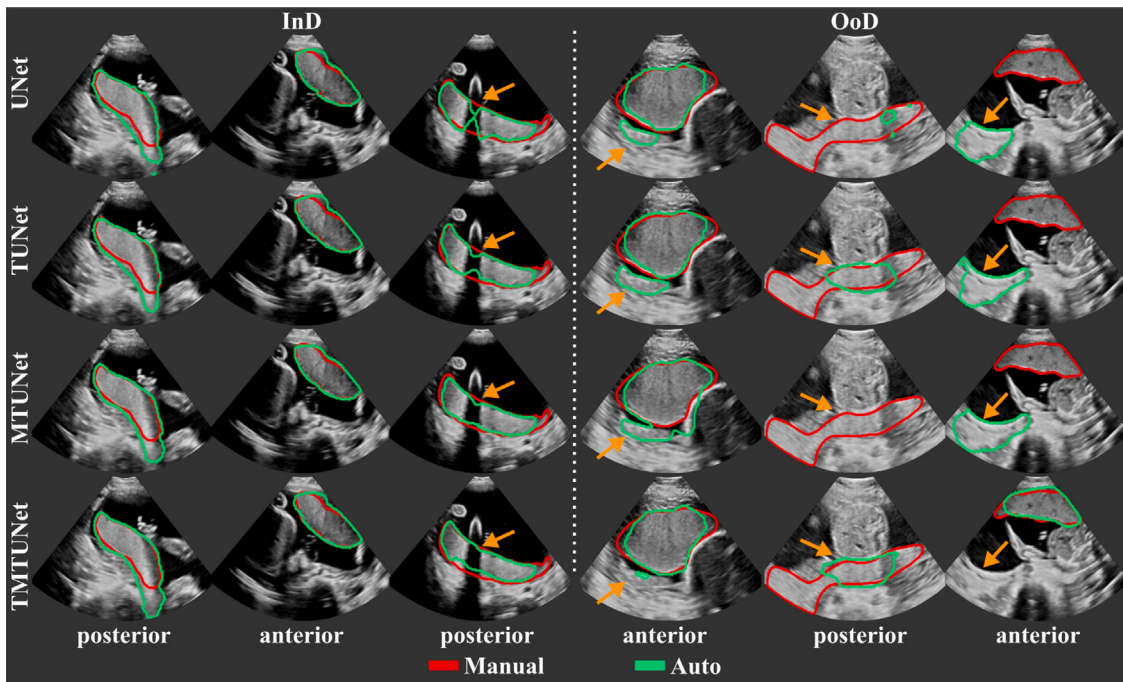
Fig. B.11. Examples of automatic placenta segmentations obtained by models *UNet, TUNet, MTUNet* and *TMTUNet* for *in-distribution* (InD) and *out-of-distribution* (OoD) test data. The orange arrows indicate areas with segmentation errors and differences between the models. (All images are 3D volumes, central 2D slices are shown.)
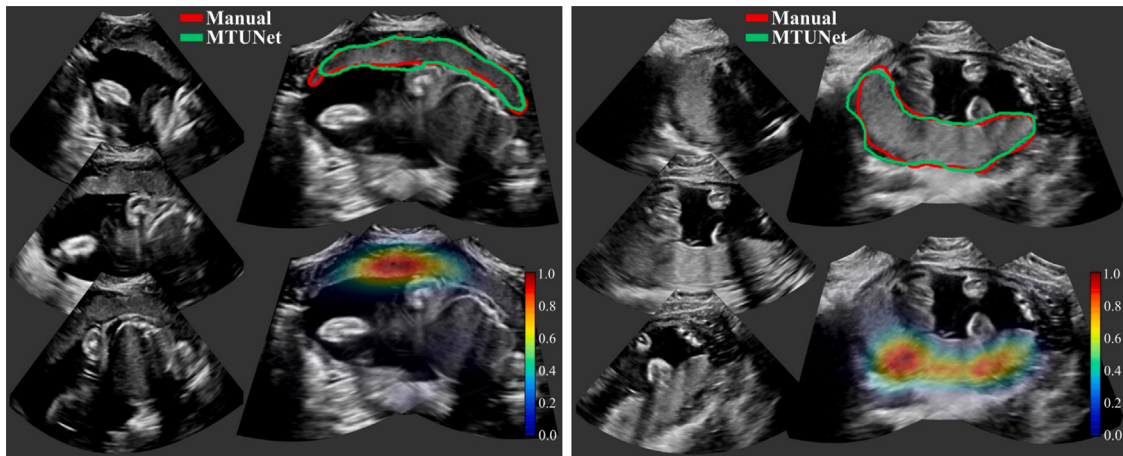

Fig. B.12. Four examples of multi-view images, each showing three individual images (left) and fused images with manual (in red) and automatic segmentation (model *MTUNet* in green) (top right) and combined attention maps (bottom right). (All images are 3D volumes, central 2D slices are shown.)

Table A.5

Data splits for five folds in training, validation and testing sets for the segmentation dataset $\mathcal{I}^S$ and the classification dataset $\mathcal{I}^C$. For $\mathcal{I}^S$, the number of images are given for anterior (ant.) and posterior (post.) placentas. For $\mathcal{I}^C$ additionally the number of images with no placental tissue visible (none) are reported.

| | Segmentation data $\mathcal{I}^S$ | | | | | | Classification data $\mathcal{I}^C$ | | | | | | | | |
| | Training | | Validation | | Testing | | Training | | | Validation | | | Testing | | |
	ant.	post.	ant.	post.	ant.	post.	ant.	post.	none	ant.	post.	none	ant.	post.	none
Fold 1	90	90	34	24	30	22	286	290	241	101	64	49	89	55	29
Fold 2	90	90	33	19	31	27	276	295	267	111	53	34	89	61	18
Fold 3	90	90	34	21	30	25	285	298	240	102	51	48	89	60	31
Fold 4	90	90	33	19	31	27	288	284	261	99	52	11	89	73	47
Fold 5	90	90	32	23	32	23	287	296	267	98	55	19	91	58	33
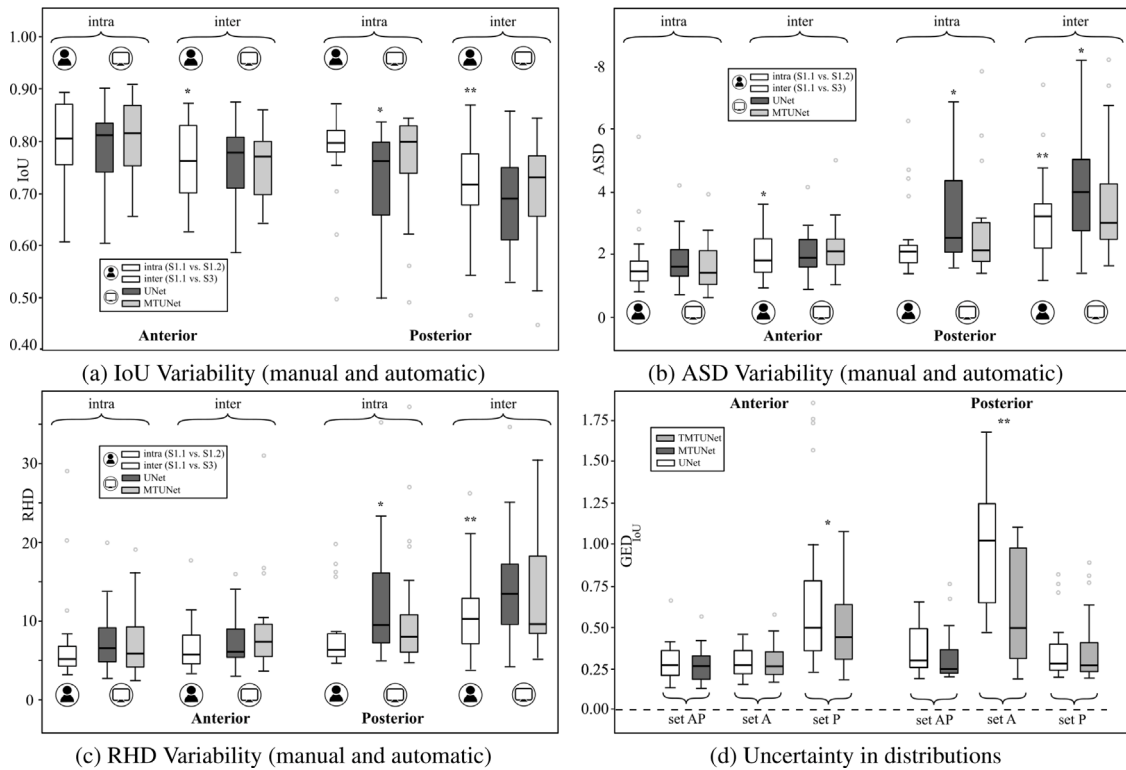
Fig. B.13. (a)–(c) Variability among manual and automatic segmentations. The agreement of possible segmentation is measures using (a) the Intersection-over-Union (IoU), (b) the average surface distance (ASD) and (c) the robust Hausdorff distance (RHD). Manual: S1.1 vs. S1.2 (intra) and S1.1 vs. S3 (inter); *UNet/MTUNet*: S1.1 vs. *UNet/MTUNet* (intra) and S3 vs. *UNet/MTUNet* (inter). (b): The difference in distributions between manual annotations from three raters and automatic segmentations from models *UNet*, *MTUNet*, and *TMTUNet* with MC dropout is measured by the Generalized Energy Distance using IoU as distance measure. This is compared for models trained on sets *A, P* and *AP* and tested on both anterior and posterior placentas. Statistical significance between *UNet* and *MTUNet/TMTUNet* is indicated by * (moderate effect size) and ** (strong effect size).

(Oguz et al., 2018) separates between different positions of the placenta in the evaluation. The works (Hu et al., 2019; Torrents-Barrena et al., 2019a) consider both early and late gestation. The overall best performance is achieved in Hu et al. (2019) with a Dice of 0.92. However, they used 2D US (in contrast to all other methods) which has higher image quality than 3D US. In 3D US, the contrast between placenta and surrounding tissue is low, especially at early but also at late gestation. Shadow artifacts become more apparent at late gestation because of the larger size of the fetus, lying in between the US probe and the placental tissue (posterior). Also, at later gestation, only part of the placental tissue might be visible in the image (especially for our multi-view images, where the middle probe is centered on the placenta and the other two probes only "see" a small part of the placenta, which is visualized with poor contrast (as seen in Figs. 5 and B.12).

Due to poor image quality and shadow artifacts, reproducible manual segmentation is challenging. We studied the intra- and inter-rater variability with two clinical experts. Our results show a higher inter-than intra-rater variability, more pronounced in posterior than in anterior placentas. Our proposed models lie within or very close to the manual rater agreement. When comparing distributions of segmentations, the multi-task approach yields a reduced uncertainty for OoD data than the baseline model. However, the comparison between only two different raters is rather limited and its generalizability should be investigated in the future. This could also be expanded to the fetal anatomy, where accurate segmentations are important.

We do not perform explicit uncertainty modeling or incorporate the knowledge of noisy labels into the model training, as done in Tanno et al. (2019), Zhang et al. (2020), Wang et al. (2019) and Kohl et al. (2018). To this end, we employ an approximation to Bayesian inference by using MC dropout at training and test time and interpret the variability of all possible segmentations for an image as the segmentation uncertainty.

Multi-task models perform statistically significantly better than *UNet*, however, it remains unclear if the improvement, which is rather small for the full dataset, is clinically relevant. The *UNet* is a very strong baseline under ideal training set conditions. However, ideal training set conditions are hard to achieve, due to the variability of the placenta appearance in US and a multi-task approach is favored when only limited annotated data is available.

In addition to a novel segmentation method, we describe a multi-view US acquisition pipeline consisting of three stages: multi-probe image acquisition, image fusion and image segmentation. We designed and printed new accessories for the handling of two or three probes using a standard US system. The obtained images show the anatomy from different view-directions and cover an enlarged FoV, allowing the combined imaging of larger structures in US. Using a simple but effective voxel-based weighted fusion strategy, image artifacts are reduced.

Extracting placental volume is of clinical interest, as it is related to fetal and placental abnormalities (Schwartz et al., 2022; Quant et al., 2016; Higgins et al., 2016). We conducted an analysis of placental volume extracted from manual and automatic segmentation from the multi-view images, and we showed a good agreement of these volumes with reference values extracted from MRI images (León et al., 2018). To this end, we have not used the segmentations/volumes to identify placenta pathologies. While the automatic detection of placental abnormalities would be the overall goal, our study only proposes a first step towards it, which is automatic placenta extraction. Our cohort consists of mainly healthy volunteers without diagnosed placental abnormalities (but blinded to fetal pathologies). A routine clinical workflow typically

does not include a detailed assessment of the placenta. Our study addresses an unmet clinical need and opens up the opportunity to better study placental pathologies throughout gestation. The extension of our work to abnormal cases would be a next logical step.

We only included second and third trimester singleton pregnancies in our study. A next step would be to extend the models and analysis to the whole gestation by including first trimester placentas. This will in addition enable a more concise comparison to previous placenta segmentation methods. Also, it would be important to test our models on twin pregnancies. Twin pregnancies can be monochorionic (shared placenta) or dichorionic (two individual placentas). Individual placentas might pose challenges for models trained only on first trimester singleton pregnancies (when the whole placenta fits in the image). The model might not recognize a second placenta in the image. In our study, however, we use second and third trimester pregnancies. The placenta is rarely completely contained in one image and our models are trained with a variety of different views. Some contain mostly placenta, some only a small part of the placenta. Therefore, we assume that our models would also perform well for twin pregnancies, but this is speculation and has to be confirmed by future studies.

A limitation of this study is that we consider only two separate classes: anterior and posterior placentas (next to the class *none*). The placenta can be located in any position between the anterior or posterior of the uterine wall and it would be interesting to incorporate a finer classification of placentas in our models.

6. Conclusion

In this work we focused on US placenta imaging and address challenges arising due to the high variability of placenta appearance, the poor image quality in US resulting in noisy reference annotations, and the limited FoV of US prohibiting whole placenta assessment at late gestation. We propose a multi-task approach combining the classification of placental position and semantic placenta segmentation in a single network. Through the classification, the model can learn from larger and more diverse datasets and improve segmentation accuracy, which are comparable to human-level performance. Our results suggest that images of anterior and posterior placentas represent two different distributions in the data. In other words they are OoD data to each other in relation to a placenta segmentation task.

We believe that this work presents important contributions for reliable imaging and image analysis in fetal screening using US. Our proposed models show a higher robustness against poor image quality and limited data availability for training. With accurate placenta segmentations together with a pipeline to image the whole placenta at all gestations, we enable clinicians towards a more comprehensive routine examination by considering placental health.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Appendix A. Materials and experiments

A.1. Probe holder design

Fig. A.10 shows the design of the two- and three-probe holder with measurements in mm. The initial design was developed on a fetal phantom in the second trimester (Kyoto Kagu Space-fan CT), and subsequently optimized with regard to comfort and usability in a clinical setting by scanning pregnant volunteers. The result is a flexible system which allows the use of two, three, or even four probes (not used in this study). We fixed the angulation between the probes so that the FoV can be extended with a known spatial alignment of the images. We chose an angle of 30° which empirically showed to angulate the probes sufficiently to maintain contact between the probe's surface and maternal skin. However, other configurations are possible.

A.2. Data

We perform a 5-fold cross-validation and each fold divides the patients in a test, training and validation set. In each fold, approximately 60% of the data I^S is used for training, and 20% for both validation and testing. Different folds had different amount of images for validation and testing (up to 10%) because of the heterogeneity of the data: each patient had a different number of images, with and without manual segmentations, and with and without placental tissue. However, we made sure that the images from individual patients were not distributed across training/validation/testing sets, the number of training images with segmentations is always the same for posterior and anterior placentas, and that each patient with manual segmentations is exactly once part of a test set.

Details about the data distribution in the folds can be found in Table A.5.

Appendix B. Results

B.1. Placenta segmentation - Single images

Fig. B.11 visualizes examples comparing the segmentation when the images was InD or OoD data. Multi-task models, especially *TMTUNet* (row 4) show a more robust performance with respect to OoD data. Only *TMTUNet* is able to localize correctly the placenta in these OoD examples. Also, *MTUNet* and *TMTUNet* are more robust to image artifacts, such as shadows, which is shown in InD, last example.

B.2. Placenta segmentation - Multi-view images

Additional exemplary multi-view images are shown in Fig. B.12 with corresponding placenta segmentations with *MTUNet* and combined attention maps. The placenta is better visualized in the multi-view images with reduced image artifacts and an extended FoV. The multi-task model *MTUNet* provides an accurate segmentation and the combined attention maps localize well the placenta.

B.3. Variability and uncertainty

We investigated the inter- and intra-observer variability for the manual annotation of placental tissue in 3D US. In each fold, we use a subset of the test set, for which three manual annotations are available. Fig. B.13(a)–(c) show the agreement of the segmentations as measured by IoU, ASD and RHD, respectively, and Fig. B.13(d) the difference in manual and automatic distributions (as a measure of uncertainty) measured by the Generalized Energy Distance using the Intersection-over-Union (IoU).

References

Alansary, A., Kamnitsas, K., Davidson, A., Khlebnikov, R., Rajchl, M., Malamateniou, C., Rutherford, M., Hajnal, J.V., Glocker, B., Rueckert, D., et al., 2016. Fast fully automatic segmentation of the human placenta from motion corrupted MRI. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 589–597.

Amyar, A., Modzelewski, R., Li, H., Ruan, S., 2020. Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation. Comput. Biol. Med. 126, 104037.

Bai, W., Chen, C., Tarroni, G., Duan, J., Guitton, F., Petersen, S.E., Guo, Y., Matthews, P.M., Rueckert, D., 2019. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 541–549.

Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D., 2017. SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Trans. Med. Imaging 36 (11), 2204–2215.

Buca, D., Iacovella, C., Khalil, A., Rizzo, G., Sirotkina, M., Makatsariya, A., Liberati, M., Silvi, C., Acharya, G., D'Antonio, F., 2020. Perinatal outcome of pregnancies complicated by placental chorioangioma: systematic review and meta-analysis. Ultrasound Obstet. Gynecol. 55 (4), 441–449.

Budd, S., Sinclair, M., Khanal, B., Matthew, J., Lloyd, D., Gomez, A., Toussaint, N., Robinson, E.C., Kainz, B., 2019. Confident head circumference measurement from ultrasound with real-time feedback for sonographers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 683–691.

Bulas, D., Egloff, A., 2013. Benefits and risks of MRI in pregnancy. In: Seminars in Perinatology, Vol. 37. Elsevier, pp. 301–304.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 424–432.

Cohen, J., 2013. Statistical Power Analysis for the Behavioral Sciences. Academic Press.

Eslami, M., Tabarestani, S., Albarqouni, S., Adeli, E., Navab, N., Adjouadi, M., 2020. Image-to-images translation for multi-task organ segmentation and bone suppression in chest X-Ray radiography. IEEE Trans. Med. Imaging 39 (7), 2553–2565.

Fadl, S., Moshiri, M., Fligner, C.L., Katz, D.S., Dighe, M., 2017. Placental imaging: normal appearance with review of pathologic findings. Radiographics 37 (3), 979–998.

Farina, A., 2016. Systematic review on first trimester three-dimensional placental volumetry predicting small for gestational age infants. Prenat Diagn 36 (2), 135–141.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. PMLR, pp. 1050–1059.

Gomez, A., Bhatia, K., Tharin, S., Housden, J., Toussaint, N., Schnabel, J.A., 2017. Fast registration of 3D fetal ultrasound images using learned corresponding salient points. In: Fetal, Infant and Ophthalmic Medical Image Analysis. Springer, pp. 33–41.

Guo, P., Lee, C.-Y., Ulbricht, D., 2020. Learning to branch for multi-task learning. In: International Conference on Machine Learning. PMLR, pp. 3854–3863.

van den Heuvel, T.L., de Bruijn, D., de Korte, C.L., Ginneken, B.v., 2018. Automated measurement of fetal head circumference using 2D ultrasound images. PLoS One 13 (8), e0200412.

Higgins, L., Simcox, L., Sibley, C., Heazell, A., Johnstone, E., 2016. Third trimester placental volume and biometry measurement: A method-development study. Placenta 42, 51–58.

Hu, R., Singla, R., Yan, R., Mayer, C., Rohling, R.N., 2019. Automated placenta segmentation with a convolutional neural network weighted by acoustic shadow detection. In: International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 6718–6723.

Jauniaux, E., Collins, S., Burton, G.J., 2018. Placenta accreta spectrum: pathophysiology and evidence-based anatomy for prenatal ultrasound imaging. Am. J. Obstet. Gynecol. 218 (1), 75–87.

Jetley, S., Lord, N.A., Lee, N., Torr, P.H., 2018. Learn to pay attention. In: International Conference on Learning Representations.

Joskowicz, L., Cohen, D., Caplan, N., Sosna, J., 2019. Inter-observer variability of manual contour delineation of structures in CT. Eur. Radiol. 29 (3), 1391–1399.

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., et al., 2017. Ensembles of multiple models and architectures for robust brain tumour segmentation. In: International MICCAI Brainlesion Workshop. Springer, pp. 450–462.

Kelley, B.P., Klochko, C.L., Atkinson, S., Hillman, D., Craig, B.M., Sandberg, S.A., Gaba, A.R., Halabi, S.S., 2020. Sonographic diagnosis of velamentous and marginal placental cord insertion. Ultrasound Q. 36 (3), 247–254.

Kendall, A., Badrinarayanan, V., Cipolla, R., 2015. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv preprint arXiv:1511.02680.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? arXiv preprint arXiv:1703.04977.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kohl, S.A., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K.H., Eslami, S., Rezende, D.J., Ronneberger, O., 2018. A probabilistic U-net for segmentation of ambiguous images. In: Advances in Neural Information Processing Systems. pp. 6965–6975.

León, R.L., Li, K.T., Brown, B.P., 2018. A retrospective segmentation analysis of placental volume by magnetic resonance imaging from first trimester to term gestation. Pediatric Radiol. 48 (13), 1936–1944.

Looney, P., Stevenson, G.N., Nicolaides, K.H., Plasencia, W., Molloholli, M., Natsis, S., Collins, S.L., 2018. Fully automated, real-time 3D ultrasound segmentation to estimate first trimester placental volume using deep learning. JCI Insight 3 (11).

Looney, P., Yin, Y., Collins, S.L., Nicolaides, K.H., Plasencia, W., Molloholli, M., Natsis, S., Stevenson, G.N., 2021. Fully automated 3-D ultrasound segmentation of the placenta, amniotic fluid, and fetus for early pregnancy assessment. IEEE Trans. Ultrason. Ferroelectr. Freq. Control 68 (6), 2038–2047.

Miller, E., Ben-Sira, L., Constantini, S., Beni-Adani, L., 2006. Impact of prenatal magnetic resonance imaging on postnatal neurosurgical treatment. J. Neurosurg.: Pediatrics 105 (3), 203–209.

Namburete, A.I., Xie, W., Yaqub, M., Zisserman, A., Noble, J.A., 2018. Fully-automated alignment of 3D fetal brain ultrasound to a canonical reference space using multi-task learning. Med. Image Anal. 46, 1–14.

Ni, D., Qu, Y., Yang, X., Chui, Y.P., Wong, T.-T., Ho, S.S., Heng, P.A., 2008. Volumetric ultrasound panorama based on 3D SIFT. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 52–60.

Oguz, I., Pouch, A., Yushkevich, N., Wang, H., Gee, J., Schwartz, N., Yushkevich, P., 2016. Fully automated placenta segmentation from 3D ultrasound images. In: Perinatal, Preterm and Paediatric Image Anal., PIPPI Workshop, MICCAI. pp. 1–10.

Oguz, B.U., Wang, J., Yushkevich, N., Pouch, A., Gee, J., Yushkevich, P.A., Schwartz, N., Oguz, I., 2018. Combining deep learning and multi-atlas label fusion for automated placenta segmentation from 3D US. In: Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis. Springer, pp. 138–148.

Oguz, I., Yushkevich, N., Pouch, A.M., Oguz, B.U., Wang, J., Parameshwaran, S., Gee, J.C., Yushkevich, P.A., Schwartz, N., 2020. Minimally interactive placenta segmentation from three-dimensional ultrasound images. J. Med. Imaging 7 (1), 014004.

Oksuz, I., Clough, J., Ruijsink, B., Puyol-Antón, E., Bustin, A., Cruz, G., Prieto, C., Rueckert, D., King, A.P., Schnabel, J.A., 2019. Detection and correction of cardiac MRI motion artefacts during reconstruction from k-space. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 695–703.

Prayer, D., Malinger, G., Brugger, P., Cassady, C., De Catte, L., De Keersmaecker, B., Fernandes, G., Glanc, P., Gonçalves, L., Gruber, G., et al., 2017. ISUOG practice guidelines: performance of fetal magnetic resonance imaging. Ultrasound Obstet. Gynecol. 49 (5), 671–680.

Public Health England, 2020. National congenital anomaly and rare disease registration service: Congenital anomaly statistics 2018. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/909405/NCARDRS_Congenital_anomaly_statistics_report_2018.pdf Accessed: 30-4-2021.

Quant, H.S., Sammel, M.D., Parry, S., Schwartz, N., 2016. Second-trimester 3-dimensional placental sonography as a predictor of small-for-gestational-age birth weight. J. Ultrasound Med. 35 (8), 1693–1702.

Raghu, M., Zhang, C., Kleinberg, J., Bengio, S., 2019. Transfusion: Understanding transfer learning for medical imaging. arXiv preprint arXiv:1902.07208.

Rajpurkar, P., Park, A., Irvin, J., Chute, C., Bereket, M., Mastrodicasa, D., Langlotz, C.P., Lungren, M.P., Ng, A.Y., Patel, B.N., 2020. Appendixnet: Deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining. Sci. Rep. 10 (1), 1–7.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Salomon, L., Alfirevic, Z., Berghella, V., Bilardo, C., Hernandez-Andrade, E., Johnsen, S., Kalache, K., Leung, K.-Y., Malinger, G., Munoz, H., et al., 2011. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. Ultrasound Obstet. Gynecol. 37 (1), 116–126.

Sarris, I., Ioannou, C., Chamberlain, P., Ohuma, E., Roseman, F., Hoch, L., Altman, D., Papageorghiou, A., International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st), 2012. Intra-and interobserver variability in fetal ultrasound measurements. Ultrasound Obstet. Gynecol. 39 (3), 266–273.

Schwartz, N., Oguz, I., Wang, J., Pouch, A., Yushkevich, N., Parameshwaran, S., Gee, J., Yushkevich, P., Oguz, B., 2022. Fully automated placental volume quantification from 3D US for prediction of small-for-gestational-age infants. J. Ultrasound Med. 41 (6), 1509–1524.

Shahedi, M., Dormer, J.D., TT, A.D., Do, Q.N., Xi, Y., Lewis, M.A., Madhuranthakam, A.J., Twickler, D.M., Fei, B., 2020. Segmentation of uterus and placenta in MR images using a fully convolutional neural network. In: Medical Imaging 2020: Computer-Aided Diagnosis, Vol. 11314. International Society for Optics and Photonics, 113141R.

Shin, H.-C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M., 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging 35 (5), 1285–1298.

Skelton, E., Matthew, J., Ho, A., Zimmer, V., Roberts, T., Schnabel, J., Hajnal, J., Rutherford, M., 2019. P19. 09: Novel 3D-extended field of view multiprobe ultrasound for placenta volumetry: feasibility and comparison with MRI. Ultrasound Obstet. Gynecol. 54, 218–219.

Slator, P.J., Hutter, J., McCabe, L., Gomes, A.D.S., Price, A.N., Panagiotaki, E., Rutherford, M.A., Hajnal, J.V., Alexander, D.C., 2018. Placenta microstructure and microcirculation imaging with diffusion MRI. Magn. Reson. Med. 80 (2), 756–766.

Soongsatitanon, A., Phupong, V., 2019. First trimester 3D ultrasound placental volume for predicting preeclampsia and/or intrauterine growth restriction. J. Obstetrics Gynecol. 39 (4), 474–479.

Stevenson, G.N., Collins, S.L., Ding, J., Impey, L., Noble, J.A., 2015. 3-D ultrasound segmentation of the placenta using the random walker algorithm: reliability and agreement. Ultrasound Med. Biol. 41 (12), 3182–3193.

Tan, J., Au, A., Meng, Q., FinesilverSmith, S., Simpson, J., Rueckert, D., Razavi, R., Day, T., Lloyd, D., Kainz, B., 2020. Automated detection of congenital heart disease in fetal ultrasound screening. In: Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis. Springer, pp. 243–252.

Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D.C., Silberman, N., 2019. Learning from noisy labels by regularized estimation of annotator confusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11244–11253.

Torrents-Barrena, J., Piella, G., Masoller, N., Gratacós, E., Eixarch, E., Ceresa, M., Ballester, M.A.G., 2019a. Automatic segmentation of the placenta and its peripheral vasculature in volumetric ultrasound for TTTS fetal surgery. In: IEEE International Symposium on Biomedical Imaging). IEEE, pp. 772–775.

Torrents-Barrena, J., Piella, G., Masoller, N., Gratacós, E., Eixarch, E., Ceresa, M., Ballester, M.Á.G., 2019b. Fully automatic 3D reconstruction of the placenta and its peripheral vasculature in intrauterine fetal MRI. Med. Image Anal. 54, 263–279.

Torrents-Barrena, J., Piella, G., Masoller, N., Gratacós, E., Eixarch, E., Ceresa, M., Ballester, M.Á.G., 2019c. Segmentation and classification in MRI and US fetal imaging: Recent trends and future prospects. Med. Image Anal. 51, 61–88.

Wachinger, C., Wein, W., Navab, N., 2007. Three-dimensional ultrasound mosaicing. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 327–335.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing 338, 34–45.

Wang, Z., Zhang, Z., Zheng, J., Huang, B., Voiculescu, I., Yang, G.-Z., 2020. Deep learning in medical ultrasound image segmentation: A review. arXiv preprint arXiv:2002.07703.

Wolf, I., Vetter, M., Wegner, I., Böttger, T., Nolden, M., Schöbinger, M., Hastenteufel, M., Kunert, T., Meinzer, H.-P., 2005. The medical imaging interaction toolkit. Med. Image Anal. 9 (6), 594–604.

Wright, R., Toussaint, N., Gomez, A., Zimmer, V., Khanal, B., Matthew, J., Skelton, E., Kainz, B., Rueckert, D., Hajnal, J.V., et al., 2019. Complete fetal head compounding from multi-view 3D ultrasound. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 384–392.

Yang, X., Yu, L., Li, S., Wen, H., Luo, D., Bian, C., Qin, J., Ni, D., Heng, P.-A., 2019. Towards automated semantic segmentation in prenatal volumetric ultrasound. IEEE Trans. Med. Imaging 38 (1), 180–193.

Zhang, L., Tanno, R., Xu, M.-C., Jin, C., Jacob, J., Ciccarelli, O., Barkhof, F., Alexander, D.C., 2020. Disentangling human error from the ground truth in segmentation of medical images. In: Conference on Neural Information Processing Systems.

Zhang, Y., Yang, Q., 2021. A survey on multi-task learning. IEEE Trans. Knowl. Data Eng..

Zhou, Y., Chen, H., Li, Y., Liu, Q., Xu, X., Wang, S., Yap, P.-T., Shen, D., 2021. Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images. Med. Image Anal. 70, 101918.

Zimmer, V.A., Gomez, A., Noh, Y., Toussaint, N., Khanal, B., Wright, R., Peralta, L., van Poppel, M., Skelton, E., Matthew, J., Schnabel, J.A., 2018. Multi-view image reconstruction: Application to fetal ultrasound compounding. In: Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis. Springer, pp. 107–116.

Zimmer, V.A., Gomez, A., Skelton, E., Ghavami, N., Wright, R., Li, L., Matthew, J., Hajnal, J.V., Schnabel, J.A., 2020. A multi-task approach using positional information for ultrasound placenta segmentation. In: Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis. Springer, pp. 264–273.

Zimmer, V.A., Gomez, A., Skelton, E., Toussaint, N., Zhang, T., Khanal, B., Wright, R., Noh, Y., Ho, A., Matthew, J., et al., 2019. Towards whole placenta segmentation at late gestation using multi-view ultrasound images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 628–636.