



Fully-automated alignment of 3D fetal brain ultrasound to a canonical reference space using multi-task learning

Ana I.L. Namburete^{a,*}, Weidi Xie^{a,b,*}, Mohammad Yaqub^a, Andrew Zisserman^b,
J. Alison Noble^a

^a Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, United Kingdom

^b Visual Geometry Group, Department of Engineering Science, University of Oxford, United Kingdom

ARTICLE INFO

Article history:

Received 22 June 2017

Revised 13 February 2018

Accepted 19 February 2018

Available online 21 February 2018

Keywords:

Ultrasound

Fetal brain

Fully convolutional neural networks

Multi-task learning

Alignment

ABSTRACT

Methods for aligning 3D fetal neurosonography images must be robust to (i) intensity variations, (ii) anatomical and age-specific differences within the fetal population, and (iii) the variations in fetal position. To this end, we propose a multi-task fully convolutional neural network (FCN) architecture to address the problem of 3D fetal brain localization, structural segmentation, and alignment to a referential coordinate system. Instead of treating these tasks as independent problems, we optimize the network by simultaneously learning features shared within the input data pertaining to the correlated tasks, and later branching out into task-specific output streams. Brain alignment is achieved by defining a parametric coordinate system based on skull boundaries, location of the eye sockets, and head pose, as predicted from intracranial structures. This information is used to estimate an affine transformation to align a volumetric image to the skull-based coordinate system. Co-alignment of 140 fetal ultrasound volumes (age range: 26.0 ± 4.4 weeks) was achieved with high brain overlap and low eye localization error, regardless of gestational age or head size. The automatically co-aligned volumes show good structural correspondence between fetal anatomies.

© 2018 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY license. (<http://creativecommons.org/licenses/by/4.0/>)

1. Introduction

Fetal neurosonography has improved significantly in the last few decades. It is emerging as a clinically useful imaging technology for assessing brain development and detecting cerebral abnormalities in the womb, which has applications in settings where expensive magnetic resonance imaging (MRI) is unavailable or not well-suited. Regardless of imaging modality, fetal brain localization and geometric alignment are the primordial steps for neuroimage analysis. This analysis relies on (i) initial localization of the brain, (ii) removal of extracranial and maternal tissues, and (iii) alignment of the region of interest to a referential coordinate system (*reference space*). Establishing a coordinate system over the fetal brain serves as a precursor to the localization of anatomical landmarks (Yaqub et al., 2015; Namburete et al., 2015; Qiu et al., 2017), extraction of standard clinical planes for biometric assessment of fetal growth (Namburete et al., 2014; Yaqub et al., 2016), and ex-

traction of oblique planes to study the evolution of image-based biomarkers from womb to cot (Ball et al., 2012; Yaqub et al., 2016).

Indeed, 3D fetal neurosonography has also been shown to capture neurodevelopmental image signatures from which to predict gestational age and brain maturation, as demonstrated observationally (Pistorius et al., 2010) and, more recently, automatically (Namburete et al., 2015; 2017). However, processing these data remains a challenging task due to interactions between the skull and the ultrasound (US) signal. Specifically, as the density of the skull increases with gestational age, it increasingly reflects the US waves, blocking the signal from penetrating to deep-seated tissues and creating strong acoustic shadows. The concave shape of the skull also creates shadows that limit clear structural visibility to only one of the cerebral hemispheres (located in the lower half of the image). These factors complicate the design of a neurosonography-specific coordinate reference, but these high-intensity structures are also the most salient cues for alignment. Consequently, our solution uses these bony landmarks, coupled with intensity contrast across cerebral soft tissue boundaries, to achieve an alignment of the fetal head to a reference space.

We propose an automated tool which capitalizes solely on sonographic image signatures to achieve an alignment of the fetal

* Corresponding authors.

E-mail addresses: ana.namburete@eng.ox.ac.uk (A.I.L. Namburete), weidi.xie@eng.ox.ac.uk (W. Xie).

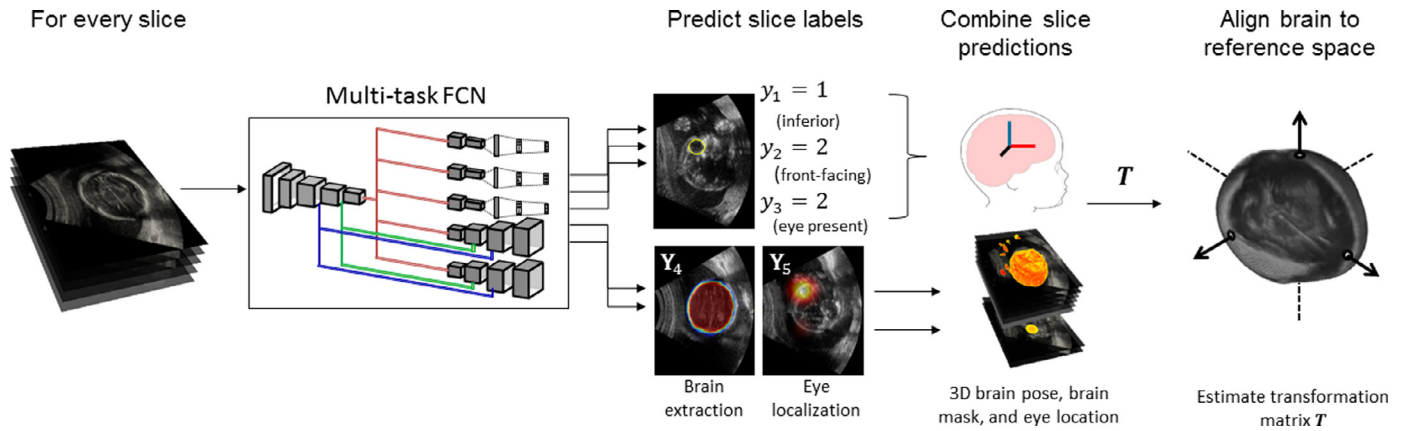


Fig. 1. Pipeline for automated alignment of fetal 3D neurosonography to a canonical coordinate space. Each slice extracted from the volume is classified and segmented by the multi-task network. The predicted labels are combined to estimate a transformation matrix T that aligns the volume to a pre-defined reference space.

brain at any gestational time point. Our pipeline normalizes image volumes to a reference space through estimated affine transformations. As such, this paper presents the following contributions:

- We decompose the complex affine transformation task into several simple ones, which can be readily tackled with CNNs. Using one multi-task fully convolutional neural network (FCN), we derive fetal brain orientation, eye localization, and brain masking (Fig. 1). Our model leverages domain-specific information to train these closely related tasks by sharing low-level features in the early layers, before branching into independent output streams for each task. The task-specific predictions are then combined to produce a transformation matrix as the desired final output.
- We combine the predicted labels from a stack of axial slices to recover 3D brain orientation by respecting their spatial ordering within the volume. A stack of segmentation maps facilitates localization and segmentation of the brain and eye sockets in 3D space. With this information, we ultimately approximate a transformation matrix that maps a given brain into a common 3D reference space.
- We compare our multi-task approach to the performance of predicting each task independently. Our analysis shows that a multi-task network not only saves on training time and memory requirements, but also improves the performance of tasks with unbalanced data labels.
- The model is generalizable to a wide gestational age range, and does not require additional age or fetal size input to achieve head segmentation, eye localization, and prediction of brain orientation. Our method capitalizes on structures that are observed in US images, regardless of gestational age. The network is trained to learn the appearance and structural variability within a large longitudinal dataset.

In Section 2, we review the limited related work in aligning fetal neurosonography. We also review related deep learning-based approaches for segmentation and multi-task learning applications to biomedical images. Section 3 outlines our approach and a description of network design details. Experiments are described in Section 4. The resulting method is evaluated in two ways (Section 5): first by assessing the slice-by-slice classification, brain segmentation, and eye localization; then by comparing the result of brain alignment to a manually initialised parametric surface. We conclude with a discussion in Section 6.

2. Related work

To achieve alignment between 3D fetal neurosonography scans of different ages, a key challenge is to identify longitudinally-

invariant signal responses in the images. Ideally, these may be used to determine brain orientation. Previous alignment approaches have used the fetal skull (Chen et al., 2012; Cuingnet et al., 2013; Namburete et al., 2015), which appears as a prominent, consistently echogenic structure. Its distinct oval shape is recognizable, independent of the position of the probe and fetal age, thus lending itself as a stable structure for fetal brain alignment. Deformable model-based methods (Chen et al., 2012; Namburete et al., 2013; 2015) and shape models (Cuingnet et al., 2013) have been used to determine a *skull-based* frame of reference for alignment. Using 3D snake deformation, Chen et al. (2012) contoured the entire fetal head, face, and neck using a reference surface model pre-constructed from a phantom. Their approach combined information about the intensity, shape, and geometry of the fetal head to align the model to the image. However, that method is heavily reliant on the presence of strong edges as well as visibility of the skull bones and a facial contour in its entirety, which is found in young fetuses (19–22 weeks), but cannot always be expected at later gestation. Similarly, Namburete et al. (2015) used a 3D parametric surface as a surrogate representation of the skull which was aligned to the image and then automatically deformed to the inner skull boundary. That approach is limited by the necessity for the surface to be *manually* initialized to the skull pixels, which is laborious, and thus ineffective for processing large datasets.

Cuingnet et al. (2013) exploited information about the skull, falx, and eye orbits to achieve fully-automated alignment of fetal head US data in early gestation (19–24 weeks). However, that approach presupposes the presence of strong edges corresponding to the objects of interest. Furthermore, intracranial intensity information was overlooked in the alignment procedure, and their approach required the visibility of *both* eye orbits, which cannot be relied upon in older fetuses. Kuklisova-Murgasova et al. (2013) aligned 3D US to MRI data using a complex Bayesian approach coupled with block-matching, demonstrating the anatomical correspondence between the two modalities in a similar gestational window (18–22 weeks). While demonstrating the potential for fusing information from these complementary modalities to enhance neurosonographic anatomies, the method performed poorly on subjects with rotated brain pose, which affects visibility of the necessary registration landmarks. Furthermore, the method required the availability of MR image data to serve as the anatomical reference guiding the alignment.

The recent success of deep convolutional neural networks (DCNNs) originated from Krizhevsky et al. (2012), where the authors showed the effectiveness of DCNNs on a large-scale image classification task. Later, Long et al. (2015) adapted the pre-trained CNN classifier for pixel-wise segmentation in a fully convolutional man-

ner, resulting in networks capable of taking images of arbitrary size as input, and returning heat maps of the same resolution as output. To achieve the desired resolution, up-sampling kernels were trained, and skip layers further applied to fuse features of different levels. In biomedical image analysis, Fully Convolutional Networks (FCNs) were first applied on microscopy images for cell segmentation, detection and counting (Ronneberger et al., 2015; Xie et al., 2015). In the context of fetal US image data, FCNs have been used to classify 2D frames and locate organs in freehand sweeps (Gao and Noble, 2017) and cardiac video sequences (Sundaresan et al., 2017; Baumgartner et al., 2017). Yang et al. (2017) proposed an FCN coupled with a bidirectional long-short term memory recurrent neural network (LSTM-RNN) architecture to segment the placenta, gestational sac, and fetal body in 3D first-trimester volumes. Segmentation of the fetal abdomen has also been achieved using FCNs based on deformable models (Schmidt-Richberg et al., 2017). Indeed, 3D networks have demonstrated success in semantic segmentation of several anatomies in 3D fetal ultrasound volumes, despite the dependence of anatomical appearance on acquisition angle. Realising the tractability of FCNs to make predictions from challenging fetal US data, our work aims to segment structures and classify slices extracted from a head volume to ultimately align the brain to a reference space. One approach is to use multi-task learning.

Multi-task learning aims to boost the generalization performance by *simultaneously* learning multiple related tasks (Caruana, 1997). One recent successful example applied to a medical imaging task is the SpineNet (Jamaludin et al., 2016), where the architecture is an extension of the canonical classification CNN. Building on the shared low-level features, each stream was then trained to predict a radiological score. In this paper, to make use of the information from the related tasks (e.g. where superior-to-inferior brain orientation should inform on relative eye location), our architecture is trained to address multiple tasks, three of which are formulated as classification problems, and two as segmentation tasks. During training, we balance these different tasks by a weighted loss. It is worth highlighting that although our proposed network predicts multiple outputs, these serve only as intermediary results to our desired task. Our interest is in using these to predict the transformation matrix that aligns the fetal US brain to a reference space (Fig. 1), which, to the best of our knowledge, has not previously been attempted.

3. Methods

Fetal neurosonography scans are typically acquired following a pre-defined protocol set by international standards. According to the guidelines set by the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG), the acquisition plane of the US probe should be orthogonal to the midsagittal plane (or falx plane, defining the falx cerebri), and the axial plane traversing the centre of the brain should be located in the middle of the US volume (ISUOG, 2007). Given such a pre-defined standard acquisition protocol, the approximate geometry of image acquisition is known. Specifically, we assume that the US probe is positioned such that the head is imaged axially, and the insonation plane is perpendicular to the midsagittal plane (which separates the cerebral hemispheres, Fig. 2a). It is also expected that the skull appears as the most echogenic oval-shaped structure, occupying at least 50% of the image space. We use this information to perform brain masking, and to guide the prediction of brain orientation.

Although the location of the head can be assumed from the acquisition protocol, the geometric orientation of the brain regions relative to the coordinate axes remains unknown. Due to interactions between the US beam and tissue boundaries, the sonographic appearance of brain structures depends on the insonation plane,

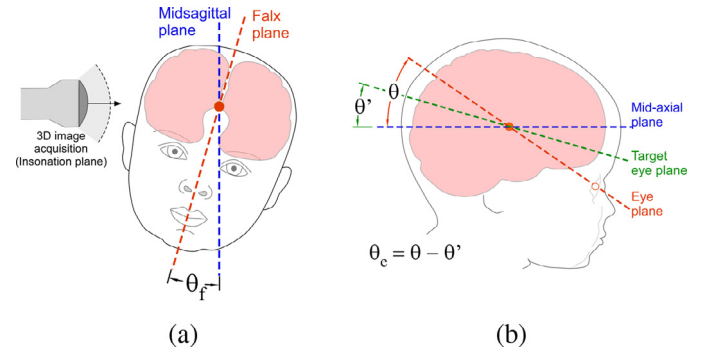


Fig. 2. Data acquisition (a) Schematic of probe position for fetal neurosonography. Insonation plane is expected to be orthogonal to the midsagittal plane (P_s^0) in the sagittal view. The falx plane (P_f) must be rotated by angle θ_f to align to the midsagittal plane. (b) Forward tilt correction is achieved by aligning the fetus' eye plane (P_e) with the target eye plane by rotating by angle θ_e .

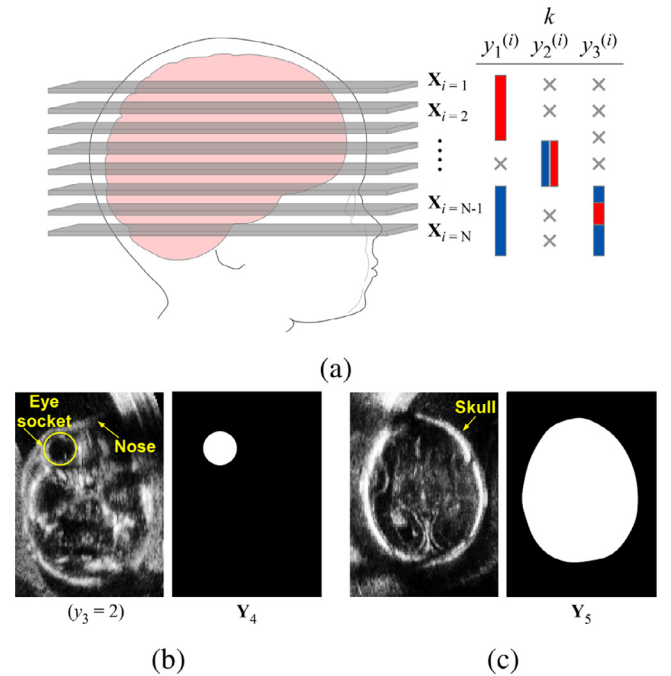


Fig. 3. Annotation description. Schematic of a set of axial slices extracted from a 3D brain volume, spanning the cerebrum. For each task k , respectively, the slices to be considered for classification (red: $y_k = 2$, blue: $y_k = 1$), and the slice regions to ignore ($y_k = 0$) are labelled with a cross. y_1 : Superior (annotated as 2) vs Inferior (annotated as 1); y_2 : Anterior (annotated as 2) vs Posterior (annotated as 1); y_3 : Eye Present (annotated as 2) vs Absent (annotated as 1) (b) Y_4 : Eye localization (binary mask). (c) Y_5 : Skull segmentation (binary mask). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and is thus informative of overall orientation. The appearance of an individual 2D axial slice differs from its adjacent slices as a function of spatial distance (Fig. 3a). That is, the anatomies seen in the slices at the base of the brain (near the neck) have a different spatial configuration from those near the crown. We rely on this appearance-based ordering of slices, the unlikely sudden transition between superior and inferior classes, and the low similarity between them to achieve a prediction of head pose. Furthermore, axial slices collected at the level of the thalami contain structures which are arranged in a consistent *spatial configuration*, which clearly distinguishes anterior from posterior regions¹. Re-

¹ We refer the reader to supplementary material for examples of neurosonographic axial slices and their appearance with respect to spatial configuration.

ardless of gestational age, the structures observable in these slices are generally reliable and consistent in terms of echobrightness.

We treat brain alignment as a transformation which relates the data volume to a 3D common coordinate space. The problem is defined as a 3D affine transformation, \mathbf{T} , composed of the product of four individual sub-transforms,

$$\mathbf{T} = \mathbf{M}_e \mathbf{M}_o \mathbf{M}_s \mathbf{M}_a \quad (1)$$

Here, \mathbf{M}_e encodes the rotation matrix which refines brain alignment by using eye location to correct for forward brain tilt (Fig. 2b). \mathbf{M}_o corresponds to the rotation and translation of the imaged falx plane (\mathcal{P}_f) onto the common coordinate frame, as shown in Fig. 2a. \mathbf{M}_s defines the transformation required to flip (or reflect) the brain such that it is correctly aligned in the superior-to-inferior direction of the coordinate axes, and lastly, \mathbf{M}_a corresponds to the anterior-posterior reflection matrix (e.g. ensuring that all brains are ‘forward-facing’).

3.1. The multi-task model

In order to estimate the geometric transformation that maps the 3D brain volume onto a common coordinate space, we decompose the complex problem into several simple ones, which can be readily tackled with CNNs. We train a deep multi-task network for three classification tasks, and two segmentations tasks that produce maps with binary units at each pixel location (m, n). We use a stack of 2D axial slices sampled from a 3D volume, spanning from the crown to the neck regions of the fetal head. Each 2D image is then passed into the network to predict whether the image is located near the superior (near the crown) or inferior (near the neck) regions of the head (task 1). The model simultaneously determines the anteroposterior direction of the fetal head (task 2), and the presence or absence of the eye socket (task 3). The segmentation tasks are such that task 4 aims to localize the eye socket, and task 5 segments the skull.

To summarize, the proposed network (Fig. 1) takes a 2D axial slice as input and outputs task-specific labels $\{y_1, y_2, y_3, \mathbf{Y}_4, \mathbf{Y}_5\}$, where labels are from a finite set: $y_{\{1, 2, 3\}} \in \{0, 1, 2\}$, and $\mathbf{Y}_{\{4, 5\}} \in \{0, 1\}$. Fig. 3 shows the slice annotations for each task. More specifically, $y_1 \in \{\text{ignore, inferior, superior}\}$, $y_2 \in \{\text{ignore, posterior, anterior}\}$, and $y_3 \in \{\text{ignore, absent, present}\}$. Note that, $y_k \in \mathbb{R}^1$ are 3-way classification labels where $y_k = 0$ corresponds to the ‘ignore’ class. Intuitively, by adding an ‘ignore’ label in all classification tasks, we have formalized the sequential prediction into a multi-task problem. In contrast, $\mathbf{Y}_k \in \mathbb{R}^{m \times n}$ are heat maps, indicating the presence probability of structures. Therefore, the corresponding ground-truth (GT) label for \mathbf{Y}_4 is annotated with binary values indicating the voxels of an eye socket (Fig. 3b). It is, however, worth noting that in standard 3D fetal US data, only the eye socket nearest to the US probe is clearly visible due to reverberation artefacts and fetal position in relation to the probe. As such, the GT eye mask represents the annotation of only one of the fetal eye sockets. For \mathbf{Y}_5 , another GT binary mask is produced for each slice to indicate the image region within the skull boundary (foreground), or extracranial regions such as maternal and uterine tissues (background), as depicted in Fig. 3c. We postpone until Section 3.2 the description of how the output of these five tasks is used to compute the transformation \mathbf{T} , and turn next to the design of the multi-task network.

3.1.1. CNN network design

Convolutional Neural Networks (CNNs) are hierarchical models mainly composed of convolutional layers with non-linear and pooling layers. By varying the depth and breadth of the networks, the capacity of the model can be controlled effectively. In a CNN model, the convolutional layers calculate pixel dependencies locally, and pooling layers are used to cut down the computational

burden by reducing the image resolution, and simultaneously increasing the receptive field and invariance. Then, high-level, task-specific features are gradually built on top of low-level features generated from the previous layers.

As shown in Fig. 4, inspired by the recent residual networks (He et al., 2016a; 2016b), the basic module in our proposed network extensively uses skip layers to fuse feature representations at multiple scales. In order to build a multi-task learning architecture, the model is subdivided into two parts: one comprising low-level features that are shared among all tasks, and the second branches out into task-specific streams. The reason for this choice is that rather than training individual CNNs for each task, we aim to fully exploit the correlation between different tasks, thereby providing more supervision on learning the shared features. Thus, during training, we simultaneously minimize the label prediction loss for all the tasks, which ensures both the invariance of underlying low-level features and discriminative nature of the five classifiers operating on task-specific features.

In our proposed architecture, all the convolutional layers use kernels of fixed size (3, 5 or 7) with sliding step size $\delta = 1$. We employ rectified linear unit (ReLU) activations after each convolutional layer. Max-pooling layers with kernel of size 2×2 ($\delta = 2$) are used to avoid too much spatial information loss. Naturally, this leads to relatively deep networks. In order to compensate for the loss of spatial information after max-pooling, the number of feature maps is gradually increased to and then retained at 256 to keep the number of parameters under control. To avoid overfitting during training, we employ a dropout layer in which 50% of the neurons are randomly set to zero in the last shared convolutional layer. This step also encourages the neurons to operate independently, thus avoiding co-adaptation of neurons (Srivastava et al., 2014). In the end, our network branches out from the convolutional layer to produce outputs for each task-specific stream.

In the eye localization and brain segmentation tasks, to avoid adding too much complexity, we do not predict maps of the same size as the input image (256×192 pixels) for the eye and brain segmentations (tasks 4 and 5). Instead, the outputs are probability maps of size 64×48 pixels, where each pixel position indicates the likelihood (0.0 to 1.0) of lying within the object of interest. Although this choice results in imprecise eye location estimations after up-sampling to original input image size, we found this size yields sufficient accuracy for forward tilt correction, and accelerates the inference speed.

In our experiments, we explore different network architectures by varying the kernel size, and network depth to investigate whether the extraction of more complex features improves predictions (see Section 5).

3.1.2. Training details

During training, each input image is rescaled such that $\mathbf{X} \in \mathbb{R}^{256 \times 192 \times 1}$. Each input \mathbf{X} is mapped to feature maps of size $32 \times 24 \times 256$ or $16 \times 12 \times 256$ (depending on network depth) by the low-level feature extractor $G_s(x, \theta_s)$. These feature maps are further passed through several task-specific streams, and convolved by different sets $G_{y_k}(\cdot, \theta_{y_k})$ for prediction of task k . For tasks $k = 4$ (eye localization) and $k = 5$ (brain segmentation), upsampling kernels are trained to recover the spatial resolution of the predicted segmentation map back to 64×48 pixels (Fig. 4a).

More formally, we consider the energy function:

$$E(\theta_s, \theta_{y_1}, \theta_{y_2}, \theta_{y_3}, \theta_{y_4}, \theta_{y_5}) = \sum_{i=1}^N \left(\sum_{k=1}^3 (\alpha_k \cdot L_k^i(\theta_s, \theta_{y_k})) + \sum_{k=4}^5 \left(\alpha_k \cdot \sum_{mn} L_{k(mn)}^i(\theta_s, \theta_{y_k}) \right) \right) \quad (2)$$

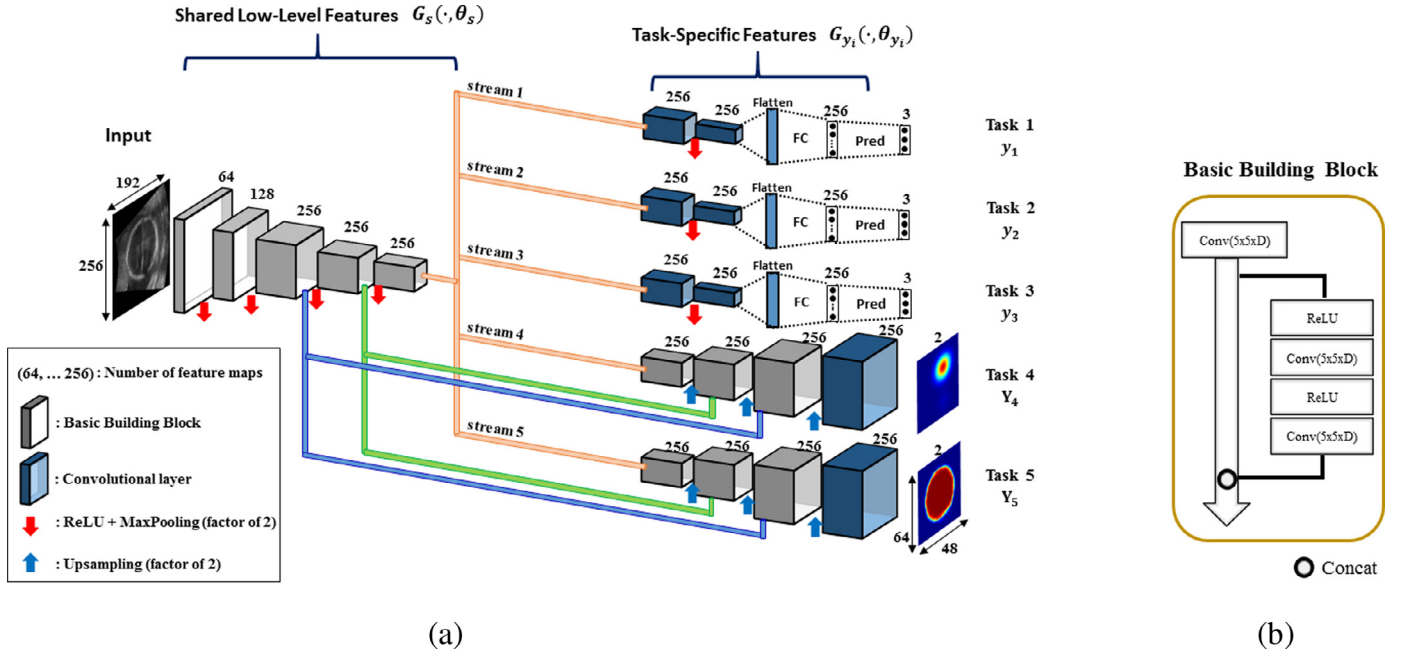


Fig. 4. Proposed architecture: (a) Network architecture (4 pooling layers) for multi-task learning. The architecture contains two parts, namely the shared low-level features, and task-specific features. Every convolution is followed by ReLU as non-linearity functions. Convolutional layers are displayed in blue, and bespoke building blocks are shown as gray blocks. Upsampling and downsampling operations are displayed by red and blue arrows, respectively. (b) A zoomed-in schematic of the basic building block, and its constituent parts. The linear convolutions are followed by ReLU. Concatenation is used for skip layers to fuse multiple-level representations, and to avoid gradient vanishing problems. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Here, $L_y^i(\cdot, \cdot)$ is the loss for label prediction (multinomial for y_1 , y_2 , y_3 , and binary for y_4 and y_5). In this work, we use the negative log-likelihood as a loss function, and α 's refer to the weights for the different tasks ($\alpha_1 = \alpha_2 = \alpha_3 = 3$, $\alpha_4 = \alpha_5 = 1$).

Thus, we seek the parameters $\hat{\theta}_s, \hat{\theta}_{y_1}, \hat{\theta}_{y_2}, \hat{\theta}_{y_3}, \hat{\theta}_{y_4}, \hat{\theta}_{y_5}$ that deliver a point, where:

$$\hat{\theta}_s, \hat{\theta}_{y_1}, \hat{\theta}_{y_2}, \hat{\theta}_{y_3}, \hat{\theta}_{y_4}, \hat{\theta}_{y_5} = \underset{\theta_s, \theta_{y_1}, \theta_{y_2}, \theta_{y_3}, \theta_{y_4}, \theta_{y_5}}{\operatorname{argmin}} E(\theta_s, \theta_{y_1}, \theta_{y_2}, \theta_{y_3}, \theta_{y_4}, \theta_{y_5}) \quad (3)$$

3.1.3. Optimization with back-propagation

When optimizing task-specific parameters, standard stochastic gradient updates for a single sample i are performed as follows:

Classification tasks, $k = \{1, 2, 3\}$:

$$\theta_{y_k} \leftarrow \theta_{y_k} - \lambda_k \frac{\partial L_k^i}{\partial \theta_{y_k}} \quad (4)$$

Segmentation tasks, $k = \{4, 5\}$:

$$\theta_{y_k} \leftarrow \theta_{y_k} - \lambda_k \sum_{mn} \frac{\partial L_{k(mn)}^i}{\partial \theta_{y_k}} \quad (5)$$

where λ 's are the learning rates. Update of the parameters specific to the eye (θ_{y_4}) and brain (θ_{y_5}) segmentation tasks are dependent on all the pixels.

The gradients from all tasks contribute to update the low-level features during back-propagation.

$$\theta_s \leftarrow \theta_s - \left(\underbrace{\sum_{k=1}^3 \lambda_k \alpha_k \frac{\partial L_k^i}{\partial \theta_s}}_{\text{Classification}} + \underbrace{\sum_{k=4}^5 \left(\lambda_k \alpha_k \sum_{mn} \frac{\partial L_{k(mn)}^i}{\partial \theta_s} \right)}_{\text{Segmentation}} \right) \quad (6)$$

We apply the RMS-prop optimization algorithm during training (Tieleman and Hinton, 2012).

3.2. Estimation of transformation (T)

The goal of our pipeline is to estimate a 4×4 geometric transformation matrix T to align 3D brain images to a common co-ordinate space (Eq. (1)). T is estimated by combining predictions from a stack of slices extracted from a given volume. As our object of interest is the fetal brain, only the segmented brain volume is considered for alignment. Thus, the brain extraction mask is first achieved by stacking the skull segmentations (Y_5) in 3D space, and approximating the centre point c and ellipsoidal dimensions $[e_x, e_y, e_z]$ of the skull (Fig. 6b). The remaining matrices are computed as follows:

M_s is the reflection matrix aligning the head in the inferior-to-superior direction. It is approximated by processing the y_1 labels of the axial slices, in their spatial ordering within the volume (i.e. z-position, Fig. 5). The ambiguous slices in the volume (i.e. $y_1 = 0$) are first detected, and the modal labels in the slices above and below these z-positions determine whether the volume must be flipped in order to align to an inferior-to-superior configuration. In Fig. 5, for instance, the ambiguous slice is at $z = 0.71$; below this, the modal label is $y_1 = 1$, and above it is $y_1 = 2$. Thus, the head is in the correct pose and does not require reflection across the xy-plane about the segmented skull centre (i.e. M_s is the identity matrix).

M_a is the reflection matrix which aligns the head in an anterior-to-posterior direction, across the yz-plane about the skull centre. It is approximated by ignoring all ambiguous slices and determining the modal label of all slices voting for a specific y_2 direction, namely forward- ($y_2 = 1$) or backward-facing ($y_2 = 2$). In Fig. 5, the brain is forward-facing.

M_o defines a rotation about the brain centre c correcting for lateral tilt. The brain centre is the centre of mass of the masked brain region. To estimate M_o , we use the algorithm presented by Huang et al. (2015). Briefly, the 2D plane \mathcal{P}_f describing the midsagittal plane is recovered by detecting the membrane of the falx traversing the midline of the masked brain in each ax-

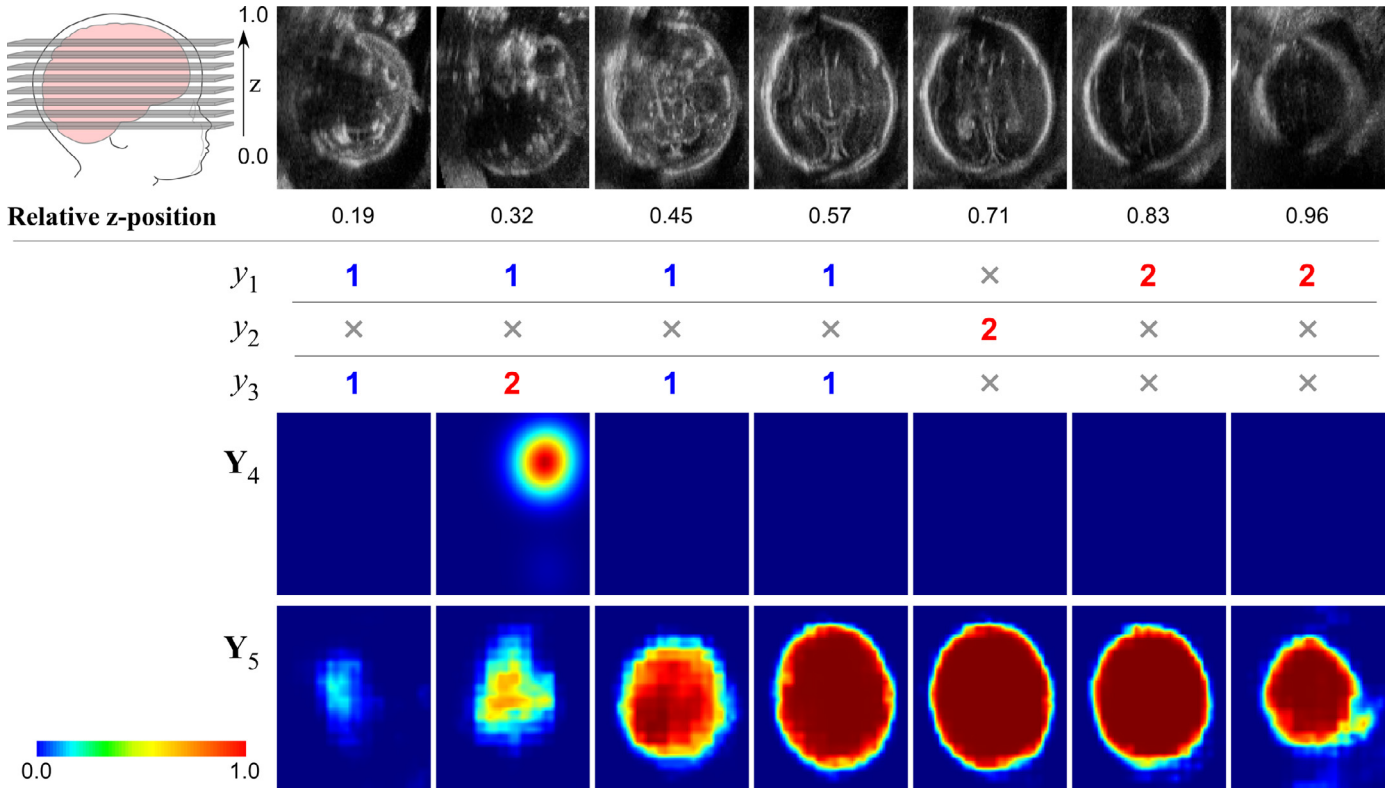


Fig. 5. Network output. Representative example of a typical output from the multi-task network for a single 3D brain volume. Task-specific slice classifications and output segmentations for a stack of axial slices, according to their relative position (z) in the volume, $z = 0.0$: base of the brain, $z = 1.0$: crown.

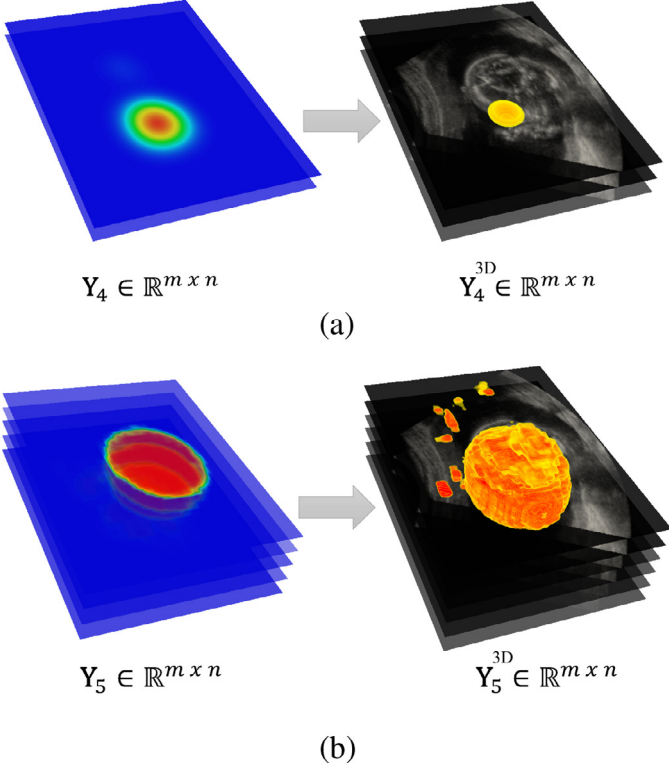


Fig. 6. Inferring 3D volume results from 2D slice predictions. Slice-to-volume segmentation of the (a) eye and (b) brain, respectively. Stacking and thresholding of 2D heat maps to obtain a 3D segmentation of the eye (Y_4^{3D}) and brain (Y_5^{3D}), respectively. The smaller outlying brain regions in the 3D brain map are removed by selecting the largest connected region as the final 3D brain mask.

ial slice. \mathcal{P}_f is approximated by compiling line candidates recovered from the stack of slices, and plane fitting is achieved using the RANSAC method (Torr and Zisserman, 2000). The lateral tilt transform \mathbf{M}_o is then derived from the center-point (\mathbf{c}) and normal vector (\mathbf{n}) of the midsagittal plane \mathcal{P}_f , and its angular deviation (in this case, the dihedral angle) from plane \mathcal{P}_o which is incident to the insonation plane in the sagittal view (s). Ultimately, $\mathbf{M}_o : \mathcal{P}_f \mapsto \mathcal{P}_o^s$.

\mathbf{M}_e If any slices vote for $y_3 = 2$, the eye localization stream is activated (Fig. 3b), and the stack of Y_4 yields an approximate location of the eye in 3D space (Fig. 6a). We consider the eye sockets as anchor points to correct for rotation of the brain in the sagittal direction, about the brain centre. The eyes are annotated in each axial slice such that eye segmentation provides a localization of eye center position, and hence \mathbf{M}_e . As shown in Fig. 2b, the rotation angle is determined by the dihedral angle between the plane intersecting the head center and two eye locations (\mathcal{P}_e), and the axial plane on the coordinate system (\mathcal{P}_o^a).

3.3. Implementation details

The proposed framework was implemented using a parallel computing architecture (CUDA, NVIDIA Corp., Santa Clara, CA), coupled with the Tensorflow library. Our method was developed on an Intel Xeon E5-2630 CPU (2.4 GHz, 16 cores) and a NVIDIA Titan X GPU. Processing took on average 15 ms per 2D slice (3.4s per volume) to output all the classification labels (y_k) and probability maps (Y_k), and an average of 1s per 3D volume.

In our implementation, we initialize the learning rate to $\alpha_{lr} = 10^{-3}$ and divide by 10 every 10 epochs, and the learning momentum is set to $\rho_a = 0.9$. The training took approximately 10 min/epoch and 26 min/epoch for the 3- and 4-layered networks, respectively. All networks were trained for a total of 36 epochs which was sufficient for convergence.

4. Experiments

4.1. Image data

Data used to validate our proposed framework were obtained from the INTERGROWTH-21st and INTERBIO-21st multi-site databases of fetal US images², comprising of healthy and growth-restricted fetuses. Gestational ages ranged from 18 to 34 gestational weeks (GW), spanning a period of active brain maturation and hence rapid developmental changes. Thus, structural appearance and composition varied across images. Images were unregistered and not pre-emptively cropped, and thus encompassed maternal tissues and other soft tissues observed in a typical obstetric scan. Subject inclusion was determined by visibility of internal brain structures in at least one cerebral hemisphere (accounting for shadowing caused by skull bones). The 3D volumes were acquired using a Philips HD9 curvilinear probe (2.5 MHz wave frequency) by different clinicians, adding variability to probe positioning and thus image appearance. Training data comprised of 599 volumes, from which approximately 16 axial slices were extracted from each volume for tuning the multi-task network, yielding a total of 9770 slices for training and validation. Image size varied with each acquisition, ranging from $117 \times 126 \times 126$ to $271 \times 343 \times 337$ voxels, acquired at resolutions ranging from $0.27 \times 0.69 \times 1.08$ to $0.46 \times 0.99 \times 1.83$ mm per voxel³. For processing, all images were standardized by resampling to an isotropic resolution of $0.6 \times 0.6 \times 0.6$ mm, and each 2D axial slice was resized to 256×192 . The independent test set comprised of 140 volumes, also sampled from the INTERGROWTH-21st and INTERBIO-21st databases.

4.2. Preprocessing and data labelling

Manual labelling was performed for all 739 volumes (599 for training, 140 for testing). Acknowledging the skull's shape as not being strictly ellipsoidal, we generated referential ground-truth (GT) segmentations of the skull by deforming ellipsoidal meshes to adhere to the inner cranial contour, using the method described in Namburete et al. (2013). Briefly, a 3D ellipsoidal mesh is manually aligned to the imaged skull, and B-spline regression is applied to deform the mesh vertices to the inner skull boundaries.

Two-dimensional axial slices were extracted from each 3D volume and manually annotated by a single rater with GT labels pertaining to each of the tasks considered during training. For our method, we expect that all sampled slices are extracted from the cerebrum (i.e. above the cerebellum, Fig. 3), to include only structural information from the cerebral hemispheres and the eye sockets. Each axial slice is associated with three classification labels and two 2D binary label maps, $\{y_1, y_2, y_3, Y_4, Y_5\}$.

4.3. Network architectures

As shown in Table 1, six different architectures (A,B,C,D,E,F) were tested. By varying the depth of the shared low-level layers (3 or 4 pooling layers), we explored the effect of the size of the receptive field on performance. For instance, it is expected that accurate skull segmentation is achieved by extracting information from a large region, yet the opposite might be true for eye segmentation. By varying the size of the convolution kernel (3×3 , 5×5 , or 7×7), we aimed to find the balance between the model capacity and data availability. Following the VGGNet (Simonyan and Zisserman, 2014), the selected kernel size remained unchanged in all convolutional layers.

4.4. Evaluation metrics

We performed five-fold cross-validation of randomly selected volumes, splitting the data into 80% for training (7813 ± 11 slices), and 20% for validation (1957 ± 11 slices). After that, the best-performing model was re-trained using the full dataset and applied to an independent left-out test set of age-matched 3D US data. The proposed pipeline was evaluated for slice-wise classification, brain segmentation, and eye localization accuracy. Slice classification was evaluated by calculating the number of correctly classified slices for tasks y_k ($k = \{1, 2, 3\}$). We report accuracy both from three-way confusion matrices, and two-way confusion matrices where ignore labels ($y_k = 0$, $k = \{1, 2, 3\}$) were not taken into account. Eye localization error was measured as the mean distance between the center of the GT eye annotation and the mean position of the detected eye pixels, in 2D and 3D.

Brain segmentation was evaluated using precision and recall measures, to quantify successful extraction of brain pixels whilst minimizing the inclusion of non-brain tissues (Baeza-Yates and Ribeiro-Neto, 1999). These metrics are independent of the background pixels, which is ideal for our dataset of varying head sizes (foreground) in relation to extracranial tissues (background). Recall is the proportion of correctly segmented brain regions:

$$R(Y_g, Y_p) = \frac{\|Y_g \cap Y_p\|}{\|Y_p\|} \quad (7)$$

where Y_g and Y_p are the GT and predicted image maps, respectively. $\|\cdot\|$ represents the voxel count. Precision is the proportion of correctly excluded background voxels:

$$P(Y_g, Y_p) = \frac{\|Y_g \cap Y_p\|}{\|Y_g\|} \quad (8)$$

We further summarize these two measures by computing the Jaccard index, to evaluate overlap between the GT annotation and the predicted segmentation:

$$J(Y_g, Y_p) = \frac{\|Y_g \cap Y_p\|}{\|Y_g \cup Y_p\|} \quad (9)$$

where $\|Y_g \cup Y_p\| = \|Y_g\| + \|Y_p\| - \|Y_g \cap Y_p\|$.

5. Results

In this section, we analyze slice- and volume-wise results for each task. Fig. 5 shows an example obtained by applying the network to a set of slices extracted from a single volume. For each slice, the network yields a prediction for each task, which can be aggregated to infer brain orientation. In the given example, the fetal head is in an inferior-to-superior orientation, forward-facing, with a detectable eye socket. Fig. 6a and b show how slice-wise predictions can be combined to infer 3D eye (Y_4^{3D}) and brain masks (Y_5^{3D}), respectively.

5.1. Slice classification results

In order to determine the stability of the model, we evaluated the accuracy of slice predictions by five-fold cross-validation. Table 1 summarizes the slice prediction results, and the comparison of different networks. For all models, an average accuracy of 87.9% or higher was achieved when considering the three-way classification ($y_k \in \{0, 1, 2\}$) on slice identification tasks ($k = \{1, 2, 3\}$). Exclusion of predictions achieved on the ambiguous slices ($y_k = 0$), however, yielded a performance drop for tasks 2 and 3, regardless of network design. This phenomenon can be explained by the severely unbalanced training sets: task 1 consisted of more relevant than ambiguous slices, whereas the opposite was true for tasks 2 and 3.

² Refer to www.intergrowth21.org.uk and www.interbio21.org.uk for details.

³ Dimensions expressed in the following order: coronal \times axial \times sagittal.

Table 1

Slice classification accuracy. Mean accuracy (\pm standard deviation) of slice-wise classification computed over the five-fold cross-validation sets. Axial slice label prediction accuracy on 2D images for tasks y_1 , y_2 , y_3 . The accuracy for 3-way (where $y_k \in \{0, 1, 2\}$) and 2-way (excluding $y_k = 0$) classification is reported for all six network architectures. Network E (in bold) outperformed the others in all classification tasks.

Network	Kernel size	Classification (3-way)			Classification (2-way)		
		y_1	y_2	y_3	y_1	y_2	y_3
3-pool	A 3×3	90.4 \pm 1.0	90.2 \pm 1.0	87.9 \pm 0.6	94.7 \pm 0.5	63.8 \pm 5.3	83.2 \pm 0.4
	B 5×5	90.6 \pm 1.0	90.7 \pm 0.9	89.5 \pm 0.8	94.9 \pm 0.7	66.7 \pm 3.3	85.4 \pm 1.3
	C 7×7	91.4 \pm 1.4	91.3 \pm 1.3	89.1 \pm 1.1	95.0 \pm 1.0	68.6 \pm 5.3	84.6 \pm 2.1
4-pool	D 3×3	91.0 \pm 0.5	90.8 \pm 0.9	88.5 \pm 1.0	94.4 \pm 0.5	68.9 \pm 2.2	83.5 \pm 1.6
	E 5×5	92.1 \pm 0.7	91.9 \pm 0.7	90.3 \pm 0.6	95.8 \pm 0.4	70.8 \pm 2.0	86.4 \pm 1.5
	F 7×7	91.3 \pm 1.1	91.3 \pm 0.7	89.7 \pm 1.2	95.5 \pm 1.1	66.3 \pm 2.2	85.7 \pm 2.2

From networks A, B, to C, kernel size was increased in all convolutional layers. Two-way classification accuracy steadily improved, particularly for task 2 (from 63.8% to 68.6%). These increases in kernel sizes also resulted in models with 13.7 to 40.0 million parameters. A similar trend was observed from networks D to E with 4 poolings.

When keeping the kernel size fixed and increasing network depth (comparing network B with E), the performance for all tasks was boosted, especially for task 2 (by about 5%). The same behavior was observed when comparing between networks A and D. Thus, network depth solves the problem of unbalanced training samples. Moreover, network E improves on both segmentation and classification tasks, which may be attributed to the importance of the receptive field in slices where the skull occupies a large region.

Following the trend, the kernel size was further increased in network F (55.4 million parameters). Interestingly, the performance dropped for all tasks when compared with network E. One explanation would be that the model started to overfit given the amount of training data in each fold (about 7000 slices).

Ultimately, we selected architecture E as our best performing model in terms of accuracy. We calculated the receiver operator characteristics (ROC) curves for the three classification tasks. The areas under the curve were 0.84 ± 0.02 when distinguishing between superior and inferior brain slices (task 1), 0.84 ± 0.02 for determining the direction in which the brain faces (task 2), and 0.82 ± 0.01 for identifying eye-containing slices (task 3). The small standard deviations show that the model is stable, as designed, for performing the tasks. We therefore retrained network E using all the training data (about 10,000 slices), and report results of applying it to the independent test set in the subsequent sections. For reference, the corresponding ROC curves can be found in the supplement.

Comparing the performance of multi-task and single-task networks (Table 3), we find that each single-task network achieves at least comparable results to its multi-task counterpart. However, the eye localization task is boosted by training concurrently with the other related tasks. The single-task FCN for eye localization did not generalize well to the test set, likely due to the fact that in each volume (16 slices) the eyes only appear in a small proportion of the sampled slices (approx. 2–3 slices), making this a task with an imbalanced set of labels. In the multi-task training scenario, the other tasks would help to train the low-level feature representations, and thus, only a small number of eye slices is needed to tune the task-specific branch.

Multi-task training also proves more advantageous in terms of training time, and memory demands: multi-task weights occupy 29.6 million parameters in memory, as compared to the 112.4 million required for individually training all tasks. This is particularly of note as single-task training does not yield performance gains.

5.2. Volume classification results

An approximation of overall brain orientation can be inferred by combining the predictions from a stack of slices extracted from a given volume. For instance, the brain's superior-to-inferior configuration is revealed by the order in which the axial slices are labelled relative to slice position. Fig. 5 shows a brain volume with 'inferior' labels ($y_1 = 1$) in the lower half of the volume, and 'superior' labels ($y_1 = 2$) closer to the top, consistent with the brain being in an inferior-to-superior configuration ($y_1^{3D} = 1$). Table 4 presents the 3D volume-based prediction, showing that prediction of superior-to-inferior brain orientation was 99.8% accurate for the cross-validation set. The y_1^{3D} task failed on only one of the 140 test volumes, where the brain did not occupy at least 30% of the image space.

The proposed model was capable of detecting an eye socket ($y_3^{3D} = 2$) in 96.4% (135/140) of the test volumes. As with slice-wise classification, the most challenging prediction proved to be the anterior-to-posterior orientation (y_2^{3D}), where average accuracy was 88.6% (124/140), also due to a small number of y_2 voting slices in the stack ($y_2 = 1$ or $y_2 = 2$), relative to the number of ambiguous slices.

5.3. Brain segmentation results

Segmentation of the brain is achieved by classifying the pixels within the skull boundaries on a slice-by-slice basis. In Table 2, we compare the performance of the networks in terms of slice-wise segmentation. Network E yielded the best performance, with a Jaccard score of 0.83 ± 0.18 . Fig. 7a demonstrates that the probabilities of voxels near the brain center are typically above 90%. This value, if taken as a threshold, could inform a brain localization task, despite the presence of a few outlying regions as shown in Fig. 6b. Indeed, we found that by thresholding the probability map at 0.5, the mean error distance between the centroids of the GT and the detected 2D brain regions was 1.19 ± 1.26 mm, only slightly increasing with gestational age, accounting for brain growth ($r = 0.22$, $p < 0.002$).

It is also evident from Fig. 7c that segmentation performance is best in the axial slices extracted near the brain centre (Jaccard > 0.80), whereas segmentation accuracy lowers in the extremal brain slices. These slices either contain the cerebral fontanelles or the infratentorial region, where skull tissues are not fused and so image boundaries are incomplete. This behaviour was observed in all test cases.

The slice-wise predictions can be further combined to obtain a 3D brain mask (Y_3^{3D} , Fig. 6b). Fig. 7a shows a typical segmentation result superimposed on the 3D volume from which the axial slices were extracted. However, as shown from the sagittal view, simply stacking the axial slices to reconstruct 3D volumes may introduce small prediction inaccuracies due to incomplete skull boundaries. Therefore, we run the same model on the coronal slices. Note that,

Table 2

Slice segmentation and object localization accuracy. Mean accuracy (\pm standard deviation) of slice-wise segmentation and object localization computed over the five-fold cross-validation sets. Jaccard index ($J_k = 0.0$: no overlap; $J_k = 1.0$: perfect overlap) and center-point distance (d_k in mm, between GT and predicted segmentations) shown for the eye and skull segmentation tasks, Y_4 and Y_5 , respectively.

Network	Kernel Size	Segmentation		Localization (mm)		No. Params
		Y_4 (eye)	Y_5 (skull)	d_4 (eye)	d_5 (skull)	
3-pool	A 3×3	0.47 ± 0.13	0.71 ± 0.22	2.50 ± 2.96	1.82 ± 1.71	13.7 M
	B 5×5	0.53 ± 0.13	0.79 ± 0.21	2.11 ± 2.01	1.50 ± 1.68	25.1 M
	C 7×7	0.52 ± 0.14	0.80 ± 0.20	2.49 ± 3.02	1.49 ± 1.64	40.0 M
4-pool	D 3×3	0.52 ± 0.15	0.79 ± 0.20	2.49 ± 2.95	1.50 ± 1.68	11.9 M
	E 5×5	0.55 ± 0.12	0.83 ± 0.18	1.92 ± 1.36	1.19 ± 1.26	29.6 M
	F 7×7	0.55 ± 0.13	0.82 ± 0.19	2.14 ± 2.17	1.24 ± 1.25	55.4 M

Table 3

Slice-wise predictions: multi-task versus single-task networks. Mean accuracy (\pm standard deviation) of slice-wise classification and segmentation tasks computed over the five-fold cross-validation sets. Jaccard index ($J_k = 0.0$: no overlap; $J_k = 1.0$: perfect overlap) the eye and skull segmentation tasks. Single-task learning greatly increases memory demands (112.4 compared to 29.6 million parameters), without improving performance.

Network		Classification			Segmentation		No. Params
		y_1	y_2	y_3	Y_4 (eye)	Y_5 (skull)	
Multi-task	E	92.1 ± 0.7	91.9 ± 0.7	90.3 ± 0.6	0.55 ± 0.12	0.83 ± 0.18	29.6 M
Single-task	E	91.5 ± 0.8	–	–	–	–	18.2 M
	–	–	90.0 ± 0.9	–	–	–	18.2 M
	–	–	–	90.0 ± 0.6	–	–	18.2 M
	–	–	–	–	0.34 ± 0.19	–	19.9 M
	–	–	–	–	–	0.83 ± 0.20	19.9 M

Table 4

Volume classification performance. Application of network E with 4 pooling layers and kernel of size 5×5 on an independent test set. Mean accuracy (\pm standard deviation) of volume-wise classification computed for tasks y_k^{3D} . Classification scores indicate the accuracy of combining slice-based classification labels to generate a prediction for overall brain orientation. Eye localization performance is quantified by Euclidean distance, and Jaccard index scores express brain segmentation performance. Target alignment error is quantified as the average of Hausdorff distances between pairs of skull surface landmarks (manual vs predicted alignment). A total of 6144 skull landmarks were used per volume.

Dataset	No. samples	Age range (weeks)	Classification			Localization (mm)	Segmentation	Alignment error (mm)
			y_1^{3D}	y_2^{3D}	y_3^{3D}	Y_4^{3D}	Y_5^{3D}	
Cross-validation set	111 ± 5	26.6 ± 4.3	99.8 ± 0.5	96.1 ± 1.7	93.5 ± 2.5	5.0 ± 5.3	0.82 ± 0.08	9.3 ± 4.1
Test set	140	26.0 ± 4.4	99.3	88.6	96.4	6.9 ± 6.9	0.82 ± 0.07	9.3 ± 4.4

although the networks have only been trained with axial slices, it still shows good generalization for skull segmentation of the coronal slices, and there is only a narrow margin of low probability pixels (meaning uncertainty), near the skull edges (Fig. 7b). Conveniently, by obtaining a segmentation map of a coronal slice (which has clear skull edges), it is possible to exclude the extremal slices by inferring the top and bottom bounds (z) of the axial view. As shown in Fig. 7b, this enables the restriction of axial slices to test so as not to extend beyond the skull bounds.

To recover the 3D brain region, outlying regions were removed from the thresholded probability map by selecting the largest 3D connected component (Fig. 6b). The histogram plots in Fig. 7d summarize the distributions of precision and recall of segmenting the brain region. The probability mask yielded high recall (0.93) and low precision (0.76) when compared to the GT mask. Generally, the probability mask confidently includes tissues inside the fetal skull (high recall), but over-segments by incorrectly including a few non-brain voxels in the extremal axial slices (near the top and bottom of the brain, Fig. 7c).

An ellipsoidal mask was generated by fitting an ellipsoid to the probability mask. Such a mask generally excludes non-brain tissues in the extremal slices, and fully covers all intra-cranial structures. As illustrated in Fig. 6b, this mask yields high average recall (0.94) and good precision (0.84) on account of full brain coverage. However, it fails to adhere to the brain's non-ellipsoidal shape, thereby always overestimating by including some background voxels.

5.4. Eye localization results

The model's eye detection stream outputs a 2D probability map $Y_4 \in \mathbb{R}^{m \times n}$, indicating possible eye locations (Fig. 8). The detections were able to incorporate most foreground eye pixels, which yielded a recall of 0.744. The mean eye localization error on 2D slices was 1.92 ± 1.36 mm (Table 2), and gestational age was found to be of no statistical significance to this behaviour ($r = -0.07$, $p = 0.45$).

Each axial slice typically votes for one eye position, and by averaging the stack of eye locations for a given volume, the eye location can be approximated in 3D space (Fig. 6a). However, there are instances in which the probability map yields two locations. In such cases, both locations detected on each axial slice are kept and a Ball Tree algorithm is used to retain only the position within a 95% confidence interval of the eye locations recovered from adjacent slices (Omohundro, 1989). When the slice-based eye detections were averaged to estimate 3D eye localization for each head volume, the mean error was 6.94 ± 6.87 mm (Table 4).

We consider an eye detection successful if the predicted eye centre lies within a radial search area equivalent to 50% of the age-appropriate eye-to-eye diameter (Jeanty et al., 1982). By this definition, eye detection was successful in 95.5% (127/133) of cases. Of the remaining six cases, eye detection failed when the fetal head did not fill at least 50% of the image (4/6), or in images with incomplete eye socket boundaries caused by fetal motion (1/6) or partial occlusion (3/6). Nevertheless, the model detected structures with similar intensity and boundary characteristics as the eye orbits, but of a larger scale or containing stronger edges. Ultimately,

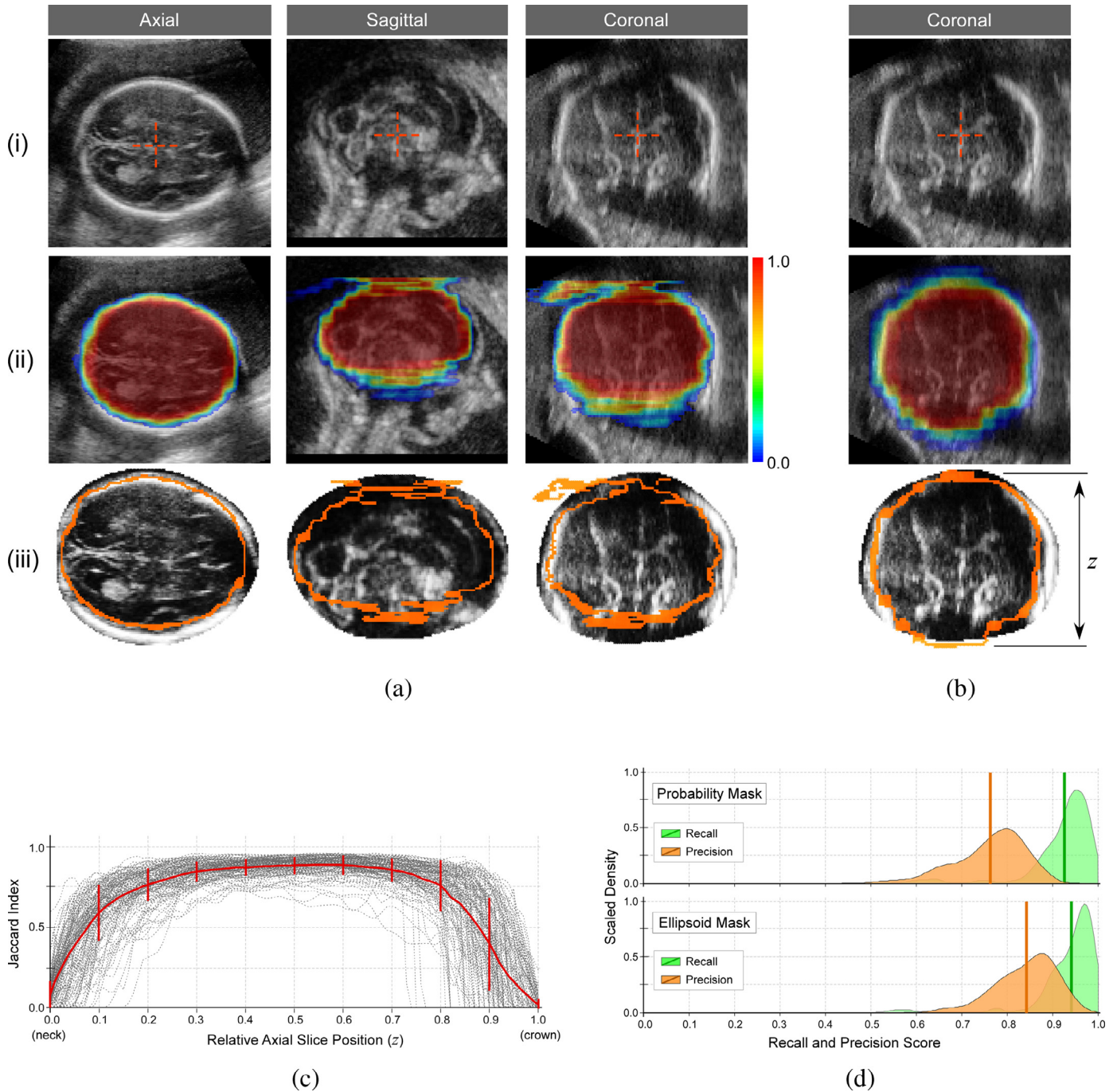


Fig. 7. Brain segmentation performance. (a) Sagittal, coronal, and axial views of an example US image (top row), with the output slice-wise probability maps superimposed (middle row). A 3D brain mask is obtained by preserving the largest connected component of the probability map (bottom row, in orange), which is then compared to the GT cranial segmentation (used here for brain masking). (b) Segmentation result of passing a coronal slice as input to the network. Skull boundaries are clearly detected, and the range of axial slices (z) can be determined. (c) Slice-wise brain segmentation overlap with respect to axial position. Each dotted line represents a single volume. Overlap is highest in the middle slices of the brain, and deteriorates in the extremal slices, near the fetal neck ($z = 0$) or the crown ($z = 1$). (d) Histogram plots showing the precision and recall 3D brain segmentation. When using the probability mask (top), the recall and precision of brain voxels is lower than if an ellipsoid is approximated from the probability mask. The latter tends to exclude incorrectly classified voxels in the extremal slices.

failure to detect the eye sockets can be attributed to improper image acquisition.

5.5. Brain alignment estimation

The quality of brain alignment was scored on the basis of how well skull geometry was approximated by the automatically-recovered transformation T (Eq. (1)). Due to low definition of skull boundaries and longitudinal variations in structural shape

and intensity, we opted for a geometric comparison between a manually-aligned ellipsoidal surface, to a surface aligned by T (Fig. 9a). Surface agreement was measured as the Euclidean distance between the manually- and automatically-aligned surface points, whilst preserving the topology of the surface points on alignment (Fig. 9a). The Hausdorff distances between the manual and automated surfaces are shown in Fig. 9b. The mean distance error was 9.3 ± 4.7 mm for volumes in which the brain alignment

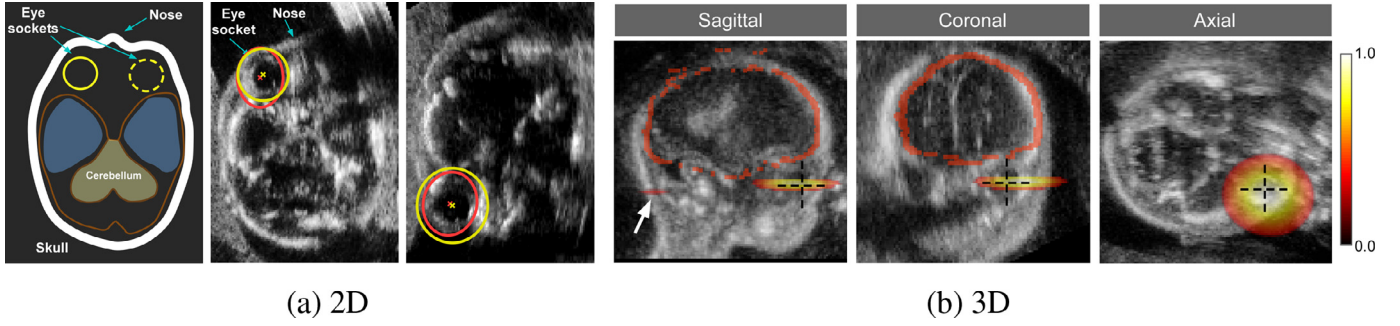


Fig. 8. Representative examples of eye localization. (a) Axial slices collected at the level of the eye sockets, displaying GT eye annotations (yellow), and predicted segmentations (orange). Corresponding eye centers are annotated with crosses. Localisation error (d_c) is the distance between these GT and predicted eye centers. (b) Sagittal, coronal and axial views of eye annotation in 3D space. White arrow indicates second eye hot-spot location detected by the network but with lower probability and in fewer slices. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

was correctly classified for all tasks (124/140). Among the 16 volumes whose orientations (y_k^{3D}) were misclassified, only one had an incorrect superior-inferior prediction (y_1^{3D}). The remaining 15 subjects had misclassified anterior-to-posterior directions. Their surface distance error increased with gestational age, correlating with occipitofrontal diameter ($r = 0.981$): the longest diameter extending from the front to the back of the brain.

From the ellipsoidal surface, a binary mask that encompasses the brain can be extracted, and thus approximates intra-cranial volume. Fig. 9c shows the good correlation between true and predicted cranial volumes ($r = 0.969$). This suggests that our model is also capable of approximating intra-cranial volume, which has been proposed as a proxy for brain growth (Chang et al., 2003; Roelfsema et al., 2004; Hsu et al., 2013). This is true even when the prediction of brain orientation fails.

The quality of brain alignment is further illustrated in Fig. 9d, which shows the average intensity and voxel variance of all correctly classified brains, using the estimated transformations. Intensity variance is a measure of the difference between the intensities of a set of images (Song et al., 2010), defined as follows:

$$\mathbf{X}^{var} = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{X}_i(\mathbf{T}_i) - \bar{\mathbf{X}}) \quad (10)$$

and the intensity average is

$$\bar{\mathbf{X}} = \frac{1}{N} \sum_{i=1}^N \mathbf{X}_i(\mathbf{T}_i) \quad (11)$$

where $\mathbf{X}_i(\mathbf{T}_i)$ is the i th 3D image transformed by matrix \mathbf{T}_i , and N is the number of images in the set. Strong alignment is characterized by the minimum pixel intensity variance among the registered images. It is evident from Fig. 9d that before alignment there is high variance, mostly corresponding to variations in image acquisition. The structures are also indiscernible in the average image. After alignment using our method, variance values were lower and high correspondence was observed in several anatomical regions, such as the Sylvian fissure, thalami, corpus callosum, and choroid plexus. Variance will, in part, reflect structural differences between fetuses at different stages of development, but also accounting for the fact that our approach estimates non-deformable transformations. Strikingly, both the average and variance maps captured hemispheric symmetry, despite the fact that each image contributed structural information from only one hemisphere. This demonstrates that the proposed method successfully aligned the 3D US scans of the fetal brain.

6. Discussion and conclusion

The methodology presented in this work was developed to address the challenge of aligning 3D US images of the fetal brain, to further advance automated analysis of brain maturation (e.g. Namburete et al., 2015; Yaqub et al., 2015). One of the factors limiting the use of 3D neurosonography for such studies is the partial appearance of brain structures, caused by fetal position and interactions between the skull bones and the US signal. Nonetheless, certain structures are consistently observed over the course of pregnancy, regardless of gestational age and fetal pose. In this paper, we present a single multi-task fully convolutional neural network which automatically locates and segments the fetal brain and eye sockets in 2D and 3D images. In the recent literature, 3D deep learning frameworks have proven successful in anatomical segmentation from volumetric ultrasound data (Looney et al., 2017; Schmidt-Richberg et al., 2017; Yang et al., 2017). Our proposed model differs in that it makes simple, computationally inexpensive predictions from 2D slices and is capable of incorporating this information to estimate 3D brain orientation. This results in a tool for fast, age-independent, and consistent co-alignment of 3D fetal neurosonographic images.

Our model was developed from 599 volumes of data ranging from 18 to 34 gestational weeks, and evaluated on a large clinical dataset consisting of 140 healthy and growth-restricted fetuses, acquired from several ethnic and geographical groups. This age range is characteristically marked by dramatic neurodevelopmental changes and increasing cranial ossification, which lead to variations in observable structures. Despite these longitudinal variations, the brain alignment results were consistent and generalizable. Our model does not require any age input, and is applicable to a wide distribution of gestational ages (18–34 weeks). It rather addresses the problem of intensity-based image alignment by segmenting and localizing stable structures observed across different developmental stages. The average of the co-aligned images depicted high correspondence in fetal brain anatomies, revealing the potential of the proposed framework in aligning large datasets for inter-subject comparisons, or longitudinal studies for tracking brain maturation.

We have shown that this fully-automated method successfully aligns 3D fetal neurosonographic data onto a pre-defined coordinate system, which could enable visual interpretation of brain anatomies. In a clinical setting, this may assist in the navigation through the brain volume and establishing coordinate locations of key cerebral landmarks (Monteagudo et al., 2000). This would allow neurosonographers to localize the standard biometry planes from which to extract 2D measurements for fetal growth monitor-

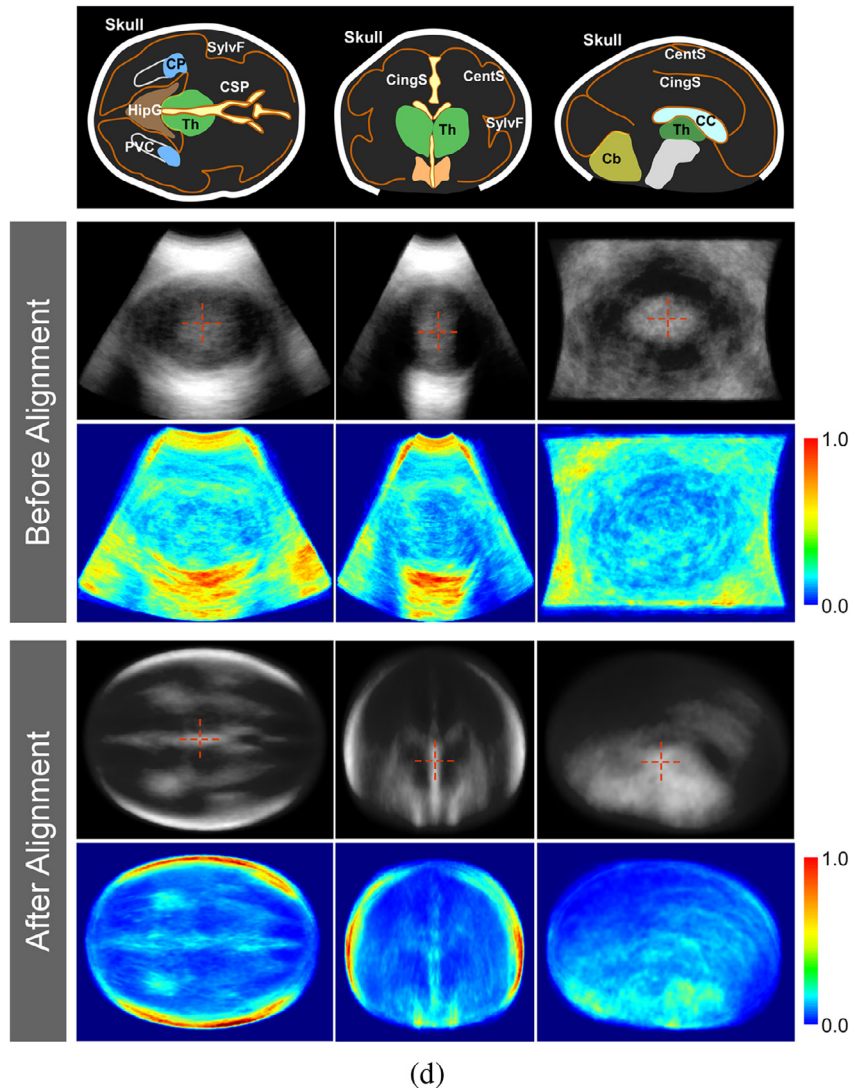
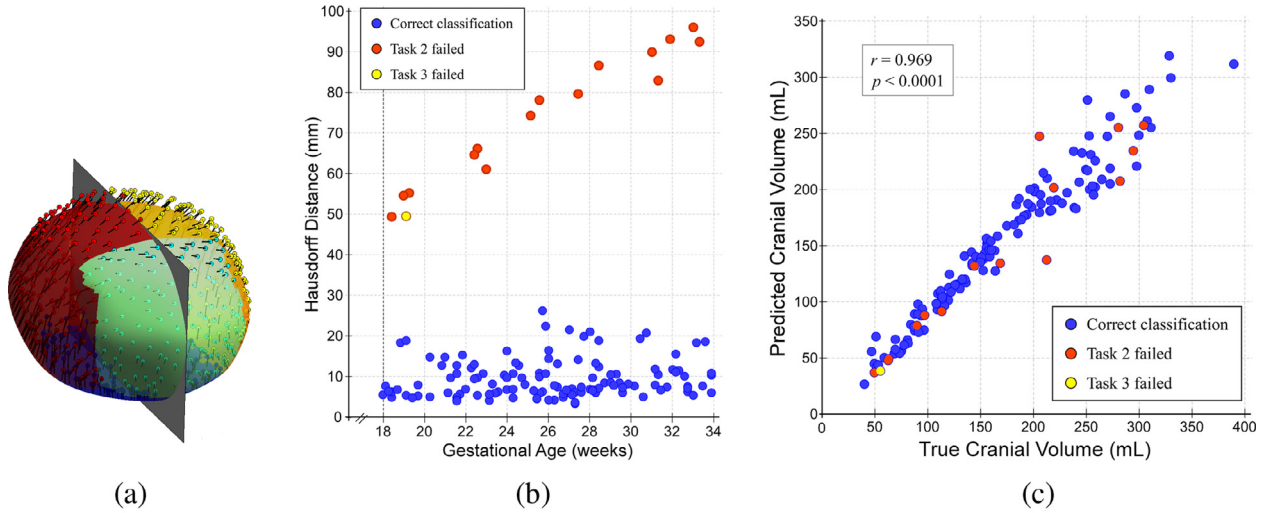


Fig. 9. Brain alignment results. (a) Illustration of the distance between the manually defined cranial surface and the points on the automatically-recovered surface (shown as coloured circles). (b) Hausdorff distance (in mm) plotted against gestational age. Each data point represents a single 3D volume. Data points with correctly classified brain orientation (blue) had a lower distance error than those with incorrect classification of superior-inferior (yellow) or anterior-posterior (orange) orientation. (c) High agreement between model-estimated and true cranial volume. (d) Axial, coronal, and sagittal views of US images and the intensity average (grayscale) and variance (colour) maps and of all 124 successfully classified volumes, before and after alignment. Mean brain maps constructed by aligning the images and averaging their pixel intensities (red: strong structural correspondence; blue: low correspondence). Prior to alignment, brain structures are unintelligible. SylvF: Sylvian fissure; CC: corpus callosum; CSP: cavum septum pellucidum; Th: Thalami; HipG: hippocampal gyrus; CP: choroid plexus; CingS: cingulate sulcus; CentS: central sulcus; PVC: posterior ventricle cavity; Cb: cerebellum. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ing, and to extract oblique planes for neuropathological assessment (ISUOG, 2007; Yaqub et al., 2016).

As the brain develops during the fetal period, some structural changes are independent of head size and their assessment would warrant the need for affine alignment. Our framework can readily be used to estimate an affine transformation matrix that excludes brain size from the co-alignment process. Thus, a size-independent coordinate system can be defined over the brain, which would facilitate feature extraction for further analysis of brain maturation. This, for instance, has clear applications in fully-automating the gestational age prediction framework presented in our previous work (Namburete et al., 2015) by automatically initializing the parametric surface onto the imaged fetal head for the extraction of features which are fed into the age classifier.

In general, we observed that some errors in predicting the overall head orientation can be attributed to the fetal head failing to occupy at least 50% of the image space, as per the protocol recommendations (ISUOG, 2007). Such images would classify as failed acquisitions, and this result indicates the model's sensitivity to image resolution and scaling, and highlights a current requirement for any input data.

In summary, we have presented a deep learning-based approach for predicting the affine transformation which maps a 3D brain image onto a common coordinate space. This fully-automated pipeline for segmentation and co-alignment of 3D fetal neurosonography provides a solution to an important step towards anatomical assessment early in pregnancy. An interesting direction for future work would be to explore the merits of a fully 3D framework for this task. Capitalizing on 3D contextual information for volumetric alignment would require re-designing the current pipeline but may reveal algorithmic and performance improvements.

Acknowledgement

A. Namburete is grateful for support from the **Bill and Melinda Gates Foundation** Grand Challenge Explorations grant, Round 14 (OPP1128941) and the UK Royal Academy of Engineering under its Engineering for Development Research Fellowship scheme. W. Xie is supported by a Google DeepMind Scholarship and the UK Engineering and Physical Sciences Research Council (EPSRC) Programme Grant Seebibyte (EP/M013774/1). M. Yaqub is funded by Innovate UK (Project 101684) and the UK EPSRC (EP/L505316/1). We thank the INTERGROWTH-21st and INTERBIO-21st Consortia for provision of 3D fetal US image data.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at [10.1016/j.media.2018.02.006](https://doi.org/10.1016/j.media.2018.02.006).

References

- Baeza-Yates, R., Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison Wesley Longman Publishing Co. Inc.
- Ball, G., Boardman, J.P., Rueckert, D., Aljabar, P., Arichi, T., Merchant, N., Gousias, I.S., Edwards, A.D., Counsell, S.J., 2012. The effect of preterm birth on thalamic and cortical development. *Cerebr. Cortex* 22 (5), 1016–1024.
- Baumgartner, C.F., Kamnitsas, K., Matthews, 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. Imag.* PP (99), 1–1. doi: [10.1109/TMI.2017.2712367](https://doi.org/10.1109/TMI.2017.2712367).
- Caruana, R., 1997. Multitask learning. *Mach. Learn.* 28 (1), 41–75.
- Chang, C.-H., Yu, C.-H., Chang, F.-M., Ko, H.-C., Chen, H.-Y., 2003. The assessment of normal fetal brain volume by 3-D ultrasound. *Ultrasound Med. Biol.* 29 (9), 1267–1272. [https://doi.org/10.1016/S0301-5629\(03\)00989-X](https://doi.org/10.1016/S0301-5629(03)00989-X).
- Chen, H.-C., Tsai, P.-Y., Huang, H.-H., Shih, H.-H., Wang, Y.-Y., Chang, C.-H., Sun, Y.-N., 2012. Registration-based segmentation of three-dimensional ultrasound images for quantitative measurement of fetal craniofacial structure. *Ultrasound Med. Biol.* 38 (5), 811–823. doi: [10.1016/j.ultrasmedbio.2012.01.025](https://doi.org/10.1016/j.ultrasmedbio.2012.01.025).
- Cuingnet, R., Somphone, O., Mory, B., Prevost, R., Yaqub, M., Napolitano, R., Papageorgiou, A.T., Roundhill, D., Noble, J.A., Ardon, R., 2013. Where is my baby? A fast fetal head auto-alignment in 3D-ultrasound. In: *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 768–771. doi: [10.1109/ISBI.2013.6556588](https://doi.org/10.1109/ISBI.2013.6556588).
- Gao, Y., Noble, J.A., 2017. Detection and Characterization of the Fetal Heartbeat in Free-hand Ultrasound Sweeps With weakly-supervised Two-streams Convolutional networks. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (Eds.), *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer International Publishing, Cham, pp. 305–313. doi: [10.1007/978-3-319-66185-8_35](https://doi.org/10.1007/978-3-319-66185-8_35).
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks. In: *European Conference on Computer Vision*. Springer, pp. 630–645.
- Hsu, J.-C., Wu, Y.-C., Wang, P.-H., Wang, H.-I., Juang, C.-M., Chen, Y.-J., Chang, C.-M., Horng, H.-C., Chen, C.-Y., Yang, M.-J., Yen, M.-S., Chao, K.-C., 2013. Quantitative analysis of normal fetal brain volume and flow by three-dimensional power doppler ultrasound. *J. Chin. Med. Assoc.* 76 (9), 504–509. <https://doi.org/10.1016/j.jcma.2013.05.006>.
- Huang, R., Namburete, A.I.L., Yaqub, M., Noble, J.A., 2015. Automated mid-sagittal plane selection for corpus callosum visualization in 3D ultrasound images. In: *Medical Image Understanding and Analysis (MIUA)*, pp. 46–51.
- ISUOG, 2007. Sonographic examination of the fetal central nervous system: guidelines for performing the 'basic examination' and the 'fetal neurosonogram'. *Ultrasound Obstet. Gynecol.* 29 (1), 109–116.
- Jamaludin, A., Kadir, T., Zisserman, A., 2016. SpineNet: automatically pinpointing classification evidence in spinal MRIs. In: *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 166–175.
- Jeanty, P., Dramaix-Wilmet, M., Gansbeke, D.V., Regemorter, N.V., Rodesch, F., 1982. Fetal ocular biometry by ultrasound. *Radiology* 143 (2), 513–516.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (Eds.), *Advances in Neural Information Processing Systems* 25. Curran Associates, Inc., pp. 1097–1105.
- Kuklisova-Murgasova, M., Cifor, A., Napolitano, R., Papageorgiou, A.T., Quaghebeur, G., Rutherford, M.A., Hajnal, J.V., Noble, J.A., Schnabel, J.A., 2013. Registration of 3D fetal neurosonography and MRI. *Med. Image Anal.* 17 (8), 1137–1150. doi: [10.1016/j.media.2014.12.006](https://doi.org/10.1016/j.media.2014.12.006).
- Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440.
- Looney, P., Stevenson, G.N., Nicolaides, K.H., Plasencia, W., Molloholli, M., Natsis, S., Collins, S.L., 2017. Automatic 3d ultrasound segmentation of the first trimester placenta using deep learning. In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 279–282. doi: [10.1109/ISBI.2017.7950519](https://doi.org/10.1109/ISBI.2017.7950519).
- Monteagudo, A., Timor-Tritsch, I.E., Mayberry, P., 2000. Three-dimensional transvaginal neurosonography of the fetal brain: 'navigating' in the volume scan. *Ultrasound Obstet. Gynecol.* 16 (4), 307–313. doi: [10.1046/j.1469-0705.2000.00264.x](https://doi.org/10.1046/j.1469-0705.2000.00264.x).
- Namburete, A.I.L., Stebbing, R.V., Kemp, B., Yaqub, M., Papageorgiou, A.T., Alison Noble, J., 2015. Learning-based prediction of gestational age from ultrasound images of the fetal brain. *Med. Image Anal.* 21 (1), 72–86. doi: [10.1016/j.media.2014.12.006](https://doi.org/10.1016/j.media.2014.12.006).
- Namburete, A.I.L., Stebbing, R.V., Noble, J.A., 2013. Cranial parametrization of the fetal head for 3D ultrasound image analysis. In: *Medical Image Understanding and Analysis (MIUA)*, pp. 27–32.
- Namburete, A.I.L., Stebbing, R.V., Noble, J.A., 2014. Diagnostic plane extraction from 3D parametric surface of the fetal cranium. In: *Medical Image Understanding and Analysis (MIUA)*, pp. 27–32.
- Namburete, A.I.L., Xie, W., Noble, J.A., 2017. Robust regression of brain maturation from 3d fetal neurosonography using crns. In: Cardoso, M.J., Arbel, T., Melbourne, A., Bogunovic, H., Moeskops, P., Chen, X., Schwartz, E., Garvin, M., Robinson, E., Trucco, E., Ebner, M., Xu, Y., Makropoulos, A., Desjardin, A., Vercauteren, T. (Eds.), *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)– Workshop on Fetal, Infant and Ophthalmic Medical Image Analysis*. Springer International Publishing, Cham, pp. 73–80. doi: [10.1007/978-3-319-67561-9_8](https://doi.org/10.1007/978-3-319-67561-9_8).
- Omohundro, S.M., 1989. Five Balltree Construction Algorithms. Technical Report. International Computer Science Institute, Berkeley.
- Pistorius, L., Stoutenbeek, P., Groenendaal, F., De Vries, L., Manten, G., Mulder, E., Visser, G., 2010. Grade and symmetry of normal fetal cortical development: a longitudinal two-and three-dimensional ultrasound study. *Ultrasound Obstet. Gynecol.* 36 (6), 700–708.
- Qiu, W., Chen, Y., Kishimoto, J., de Ribapierre, S., Chiu, B., Fenster, A., Yuan, J., 2017. Automatic segmentation approach to extracting neonatal cerebral ventricles from 3D ultrasound images. *Med. Image Anal.* 35, 181–191. doi: [10.1016/j.media.2016.06.038](https://doi.org/10.1016/j.media.2016.06.038).
- Roelfsema, N.M., Hop, W.C., Boito, S.M., Wladimiroff, J.W., 2004. Three-dimensional sonographic measurement of normal fetal brain volume during the second half of pregnancy. *Am. J. Obstet. Gynecol.* 190 (1), 275–280. [https://doi.org/10.1016/S0002-9378\(03\)00911-6](https://doi.org/10.1016/S0002-9378(03)00911-6).
- 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 (MICCAI)*. Springer, pp. 234–241.

- Schmidt-Richberg, A., Brosch, T., Schadevaldt, N., Klinder, T., Cavallaro, A., Salim, I., Roundhill, D., Papageorgiou, A., Lorenz, C., 2017. Abdomen segmentation in 3d fetal ultrasound using cnn-powered deformable models. In: Cardoso, M.J., Arbel, T., Melbourne, A., Bogunovic, H., Moeskops, P., Chen, X., Schwartz, E., Garvin, M., Robinson, E., Trucco, E., Ebner, M., Xu, Y., Makropoulos, A., Desjardin, A., Vercauteren, T. (Eds.), International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)– Workshop on Fetal, Infant and Ophthalmic Medical Image Analysis. Springer International Publishing, Cham, pp. 52–61. doi:[10.1007/978-3-319-67561-9_6](https://doi.org/10.1007/978-3-319-67561-9_6).
- Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv:[1409.1556](https://arxiv.org/abs/1409.1556).
- Song, J.H., Christensen, G.E., Hawley, J.A., Wei, Y., Kuhl, J.G., 2010. Evaluating Image Registration Using NIREP. In: Fischer, B., Dawant, B.M., Lorenz, C. (Eds.), Biomedical Image Registration: 4th International Workshop, WBIR 2010, Lübeck, Germany, July 11–13, 2010. Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 140–150. doi:[10.1007/978-3-642-14366-3_13](https://doi.org/10.1007/978-3-642-14366-3_13).
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.
- Sundaresan, V., Bridge, C.P., Ioannou, C., Noble, J.A., 2017. Automated characterization of the fetal heart in ultrasound images using fully convolutional neural networks. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 671–674. doi:[10.1109/ISBI.2017.7950609](https://doi.org/10.1109/ISBI.2017.7950609).
- Tieleman, T., Hinton, G., 2012. RMSprop: Divide the gradient by a running average of its recent magnitude. COURSE 4, 2.
- Torr, P., Zisserman, A., 2000. MLESAC: a new robust estimator with application to estimating image geometry. *Comput. Vis. Image Understand.* 78 (1), 138–156. <https://doi.org/10.1006/cviu.1999.0832>.
- Xie, W., Noble, J.A., Zisserman, A., 2015. Microscopy cell counting with fully convolutional regression networks. In: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)–Workshop on Deep Learning in Medical Image Analysis, pp. 1–10.
- Yang, X., Yu, L., Li, S., Wang, X., Wang, N., Qin, J., Ni, D., Heng, P.-A., 2017. Towards automatic semantic segmentation in volumetric ultrasound. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (Eds.), International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). Springer International Publishing, Cham, pp. 711–719. doi:[10.1007/978-3-319-66182-7_81](https://doi.org/10.1007/978-3-319-66182-7_81).
- Yaqub, M., Kelly, B., Papageorgiou, A.T., Noble, J.A., 2015. Guided random forests for identification of key fetal anatomy and image categorization in ultrasound scans. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer International Publishing, pp. 687–694.
- Yaqub, M., Rueda, S., Kopuri, A., Melo, P., Papageorgiou, A.T., Sullivan, P.B., McCormick, K., Noble, J.A., 2016. Plane localization in 3-D fetal neurosonography for longitudinal analysis of the developing brain. *IEEE J. Biomed. Health Inform.* 20 (1120–1128).