

Learning to Map 2D Ultrasound Images into 3D Space with Minimal Human Annotation

Pak-Hei Yeung^{a,*}, Moska Aliasi^b, Aris T. Papageorgiou^c, Monique Haak^b,
Weidi Xie^{a,d}, Ana I.L. Namburete^a

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

^b*Division of Fetal Medicine, Department of Obstetrics, Leiden University Medical Center,
2333 ZA Leiden, The Netherlands*

^c*Nuffield Department of Obstetrics and Gynaecology, University of Oxford, Oxford, United
Kingdom*

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

Abstract

In fetal neurosonography, aligning two-dimensional (2D) ultrasound scans to their corresponding plane in the three-dimensional (3D) space remains a challenging task. In this paper, we propose a convolutional neural network that predicts the position of 2D ultrasound fetal brain scans in 3D atlas space. Instead of purely supervised learning that requires heavy annotations for each 2D scan, we train the model by sampling 2D slices from 3D fetal brain volumes, and target the model to predict the inverse of the sampling process, resembling the idea of self-supervised learning.

We propose a model that takes a set of images as input, and learns to compare them in pairs, the pairwise comparison is weighted by the attention module based on its contribution to the prediction, which is learnt implicitly during training. The feature representation for each image is thus computed by incorporating the relative position information to all the other images in the set, and is later used for the final prediction.

We benchmark our model on 2D slices sampled from 3D fetal brain volumes at 18-22 weeks of gestational age. Using three evaluation metrics, namely, Eu-

*Corresponding author

Email address: `pak.yeung@eng.ox.ac.uk` (Pak-Hei Yeung)

clidean distance, plane angles and normalized cross correlation, which account for both the geometrical and appearance discrepancy between the groundtruth and prediction, in all these metrics, our model outperforms a baseline model by as much as 23%, when the number of input images increases. We further demonstrate that our model generalizes to (i) real 2D standard transthalamic plane images, achieving comparable performance as human annotations, as well as (ii) videos of 2D freehand fetal brain scan.

Keywords: Fetal neurosonography, Convolutional neural network, Plane localization, Self-supervised learning

1. Introduction

Two-dimensional (2D) ultrasound, given its cost-effectiveness, safety and real-time acquisition capabilities, is the preferred tool for routine monitoring of fetal growth, and the assessment of fetal anatomy including the fetal central nervous system (CNS). During routine 2D ultrasound fetal brain scanning, one of the major goals is to acquire standard planes for the assessment of structural development, namely, the transventricular (TV) plane, the transcerebellar (TC) plane and the transthalamic (TT) plane (Paladini et al., 2007). Standard biometric measurements of the fetal head, for example head circumference (HC), atrium of the lateral ventricle and transcerebellar diameter (TCD) are derived from those planes (Paladini et al., 2007). These biometric measurements are correlated with fetal brain development and, hence, serve as the essential metrics for fetal growth monitoring (Loughna et al., 2009). Different anatomic structures, such as the lateral ventricles, cavum septum pellucidum (CSP), cerebellum and cisterna magna can also be qualitatively evaluated from the acquired images, for example, the CSP can be identified in the TV plane from as early as 15 weeks of gestation and its absence or enlargement shown in the ultrasound images may indicate abnormal brain development and diseases, such as septo-optic dysplasia, holoprosencephaly and middle interhemispheric variant (Maligner et al., 2005; Falco et al., 2000; Winter et al., 2010). The cerebellum and cisterna

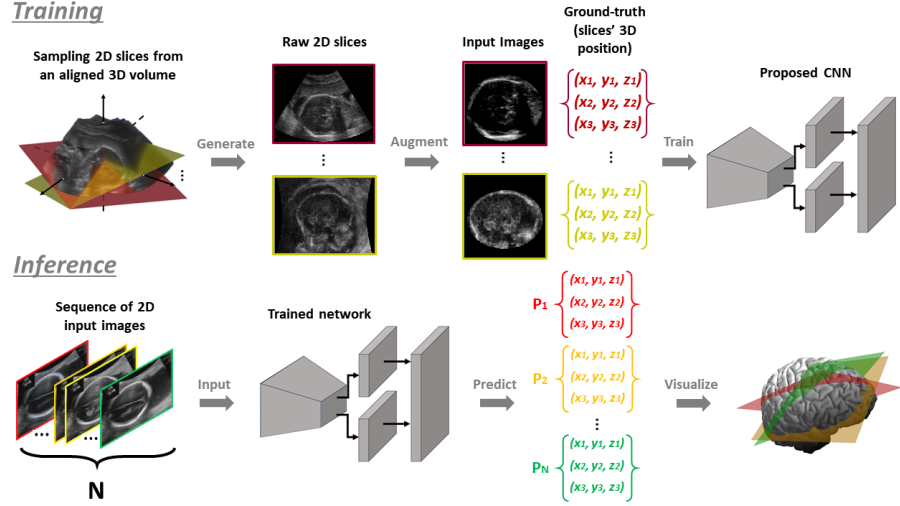


Figure 1: Pipeline of our proposed work. During training, 2D slices sampled from aligned 3D volumes are augmented and used to train our proposed CNN. The trained network can be used to predict the 3D location of arbitrary number of 2D images. (Best viewed in color)

magna can be visualized in the TC plane and CNS abnormalities, such as mega cisterna magna, dandy walker malformation, or spina bifida, may be suspected by the absence or abnormal size or shape of these two structures (Filly et al., 1989). Besides the major assessments and diagnosis from single standard plane, multiplanar approach is sometimes required for a detailed fetal neurosonography, for example assessing the presence of corpus callosum, measuring the depth and position of the cerebellar vermis, and comparing the size of the right and left cerebellar lobes (Paladini et al., 2007; Bethune et al., 2013).

All the aforementioned assessments require accurate identification of different 2D views and matching them with the corresponding planes in the 3D brain. Conventionally, sonographers need to interpret the relationships between the 2D views and the 3D brain anatomy and mentally reconstruct a 3D image given just the 2D information (Gonçalves et al., 2005). This process requires in-depth understanding of fetal anatomy and experience in ultrasound imaging, which requires a significant amount of training and there may be a shortage of adequately skilled personnel in resource-constrained settings (Benacerraf,

2002). Errors in identifying the corresponding 3D location of 2D scans may lead to inaccurate measurements of biometric parameters and misjudgement of fetal brain anatomy. Even though the 2D views are correctly identified, match-
40 ing them with the corresponding planes in the 3D brain involves subjective judgement and, hence, inter-operator variability is inevitable. In this work, we propose a model that predicts the corresponding location of 2D scans in the 3D brain, which may help the clinicians identify and localize different scans, including both standard and non-standard planes, more easily and hence lead to
45 more accurate and objective measurement and analysis. This may be useful for different potential clinical applications: (i) this can be used for training novices because the model may help them visualize the correspondence between 2D scans and 3D space and structures, which are achieved mentally by experienced sonographer (Gonçalves et al., 2005); (ii) mapping 2D ultrasound images to 3D
50 space may facilitate a variety of tasks, such as quality control and guiding the scanning by human-computer interaction.

In this paper, we propose a convolutional neural network (CNN) for predicting the location of 2D ultrasound fetal brain images in a pre-defined 3D reference coordinate system. As such, we present the following contributions:

- 55 (i) We define the localization of 2D ultrasound images of fetal brain in 3D space as a self-supervised learning problem. Using 2D slices sampled from aligned 3D volumes as training data, which are processed by our proposed preprocessing pipeline (*i.e.* Section 3.1), we further demonstrate that our model generalizes to actual 2D ultrasound images and videos (Fig. 1).
- 60 (ii) We propose a new CNN model architecture that takes an arbitrary number of input images as a set, instead of individual images. We demonstrate that this is a better utilization of available information and leads to improved performance. This setting is particularly suitable for 2D freehand ultrasound scanning, where a large but indeterminate number of 2D images are
65 usually available.
- (iii) Inspired by the idea of relation networks (Santoro et al., 2017; Xie et al.,

2018), we apply a pairwise comparison module to evaluate the geometrical relationships between different planes. While fusing feature representations from different planes during prediction, attention mechanisms are applied to dynamically assign importance to the information from other input planes, and we demonstrate that a fully trained model can indeed learn to assign a meaningful attention weight to each input image without extra supervision.

- (iv) We first benchmark our model on a synthetic dataset, where 2D slices are sampled from 3D volumes, and hence the groundtruth location of these slices is known. We show that our proposed model consistently outperforms a strong baseline described in Hou et al. (2017, 2018). In addition, we test our model on real 2D ultrasound images and videos with annotations from two experienced clinicians and medical professionals. Our proposed model also outperforms the strong baseline (Hou et al., 2017, 2018) and achieves comparable performance to human annotation.

2. Related Works

2.1. Standard Planes Detection

In the literature, a number of methods have been proposed for automated standard plane detection for 2D fetal ultrasound. Earlier studies (Zhang et al., 2012; Ni et al., 2013; Yang et al., 2014) proposed to use the Adaboost classifier or support vector machine classifier to detect key anatomical landmarks in a sequence of 2D ultrasound images. Presence and orientation of the detected landmarks were used to identify an image as either a standard or non-standard plane of view.

Recently proposed methods employed convolutional neural networks (CNN) for standard plane detection. Chen et al. (2015b) fine-tuned a pretrained CaffeNet Model (Deng et al., 2009) to detect the standard planes in ultrasound fetal abdominal images. Baumgartner et al. (2017) further trained a CNN model

95 to classify fetal ultrasound images into 14 categories, including different types of standard plane images and background images. Using an attention mechanism, Schlemper et al. (2018) proposed a CNN model that may simultaneously perform standard plane detection and weakly supervised structure localization using only image-level class label for training.

100 Spatio-temporal information of 2D ultrasound videos has also been explored for standard plane detection. Chen et al. (2015a) and Huang et al. (2017) presented different multi-task recurrent neural network models that can utilize the temporal information of consecutive sequences in ultrasound videos to provide extra contextual clues for the detection task. Gao and Noble (2017) used
105 image-level labels to train a two-stream spatio-temporal CNN to recognize fetal heart frames and localize the heart in freehand fetal ultrasound videos.

Despite their effectiveness in detecting standard plane images, all of the above methods can only predict whether the image is acquired at a standard plane, but not the exact location of the image in the corresponding 3D space.
110 Furthermore, a large amount of annotated data is required to train the model. Instead of training a classification model, we aim to learn a regression model that predicts the location of 2D ultrasound images of the fetal brain in a pre-defined 3D atlas space. This is a more general task, which can be easily adapted to standard plane detection by simply identifying the standard planes in the pre-
115 defined 3D atlas space. Our model can further provide information about the relative position between the current plane and any standard or oblique planes of interest. Also, we use 2D images sampled from 3D volumes that are aligned to a common atlas space (Namburete et al., 2018) so that the locations of images are automatically known and no further human annotation is needed, which
120 resembles the idea of self-supervised learning.

2.2. Standard Planes Localization in 3D Volumes

A slightly different task is standard plane localization in 3D volumes, which aims at identifying the standard planes (*i.e.* cross-sectional views) within a given volume. Several studies have suggested different methods for this task.

125 Ryou et al. (2016) proposed to exploit sharp boundary information in the 3D
 ultrasound volume to detect the fetal region-of-interest (ROI) and then classify
 head and body slices within the ROI using a transfer learning CNN. The stan-
 dard head and abdominal planes are automatically selected by incorporating
 prior clinical knowledge about the position of the standard plane within the
 130 two structures. Li et al. (2018) presented a CNN that is able to output the
 transformation required to move the input 2D cross-sectional image of a 3D
 fetal brain ultrasound volume towards the standard plane of view. Such predic-
 tion is computed iteratively during inference. Recent studies (Alansary et al.,
 2018; Dou et al., 2019) proposed different reinforcement frameworks for standard
 135 plane localization in 3D MRI and ultrasound volumes. These RL frameworks
 provide feedback from the environment (*i.e.* the 3D volume) during the search
 for the standard planes, which mimics the navigation performed by experienced
 operators when they are locating the target view planes in the volumes.

The aforementioned methods require 3D volume as an input, either directly
 140 or by having information extracted from the volume as a feedback during the lo-
 calization process. This may limit their application as most of the current stan-
 dard clinical tests rely on only 2D ultrasound, and 3D ultrasound is not always
 available in many settings because of its cost and clinicians’ preference (Pala-
 dini et al., 2007). On the other hand, our proposed method just relies on 2D
 145 ultrasound images and it can be easily used with 2D ultrasound scanning to
 localize any standard or oblique planes of clinicians’ interest.

2.3. Slice-to-Volume Registration

Image registration is the process of aligning two or more images into a shared
 coordinate system and slice-to-volume registration (SVR) is a sub-class of this
 150 problem, where the images to be registered are 2D and the target coordinate
 system is 3D (Ferrante and Paragios, 2017). One of the major applications of
 SVR in medical imaging is motion correction of fetal MRI.

Alansary et al. (2017) summarized a general framework about SVR for mo-
 tion correction, where a few overlapping motion-corrupted volumes are used

155 to reconstruct a clean volume. One volume is normally selected as the initial reference and 2D slices of the remaining motion-corrupted volumes are incrementally aligned to it to update the reference volume by optimizing the similarity between the reference and motion-corrupted volumes. Different objective functions, including mutual information (Rousseau et al., 2006), cross-
160 correlation (Jiang et al., 2007; Kuklisova-Murgasova et al., 2012) and mean square difference (Gholipour et al., 2010), have been explored for SVR. Despite the effectiveness reported in these studies, the proposed approaches are mainly for motion correction of 3D volumes, where the composing slices are roughly aligned and span the whole anatomical structure. These assumptions
165 may not hold for our proposed task because of the relatively long duration and inter-operator variability of typical clinical ultrasound examination. Also, those reviewed methods usually involve complicated pipelines, which are difficult to reproduce and highly specific to the task and data concerned (Ferrante and Paragios, 2017).

170 Recently, Hou et al. (2017, 2018) proposed to train a CNN to predict the rotations and translations of 2D slices sampled from a 3D MRI volume, which is aligned to an atlas coordinate system. They firstly manually corrected a set of motion-corrupted MRI volumes. The volumes were then aligned to an atlas so that they were in the same coordinate system. Slices were randomly
175 sampled from these aligned volumes and a model was trained to predict their position in the atlas coordinate system using geometric loss. With their learned model, slices from a motion-corrupted MRI volume can be registered to an aligned space. Our work is inspired by theirs and we propose a new CNN model architecture. Our proposed model accepts an arbitrary number of input images
180 as a set and takes the relationships between input images into account when predicting the location of each input image, using pairwise comparison and attention mechanism (Santoro et al., 2017; Xie et al., 2018). A comparison of the performance of these two CNN models is included in Section 5.1.1 and 5.1.2.

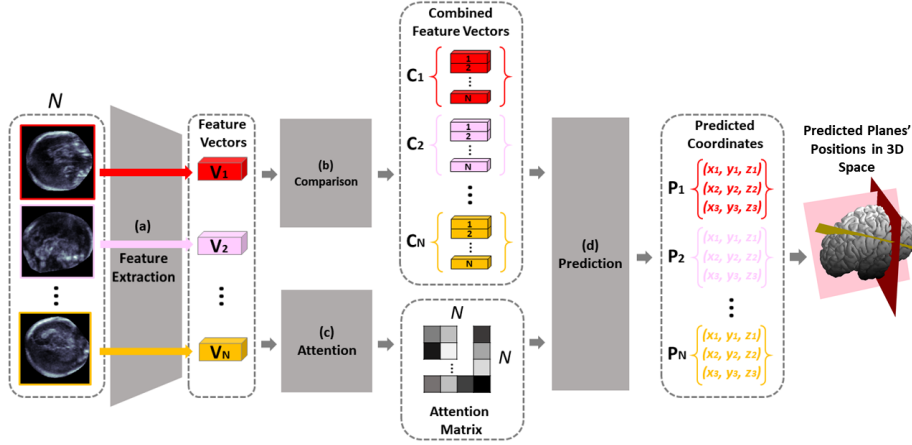


Figure 2: Overview of our proposed network. It consists of 4 sequential modules, namely (a) Feature Extraction, (b) Comparison, (c) Attention and (d) Prediction, which are represented by the grey blocks in the figure. (Best viewed in color)

3. Methods

185 In this section, we describe the proposed regression CNN (Fig. 2) in detail.

3.1. Training Data Generation for Self-Supervised Learning

Supervised learning requires paired training data in the form of $\{x_i, y_i\}$, where x_i is the input data point (*i.e.* 2D ultrasound image of fetal brain) and y_i is the label (*i.e.* the 3D location of the input image). Conventionally, the label is obtained by manual annotation, and the goal is usually to learn a function that maps the input sample x to a corresponding output label y .
 190 However, annotating the location of a random 2D ultrasound image of the fetal brain in the 3D space is very challenging. Therefore, we artificially sample 2D slices from aligned 3D ultrasound volumes of the fetal brain, results in almost infinite number of data pairs $\{x_i, y_i\}$.
 195 Despite the volume alignment is semi-automatic (minimal effort is required from manual correction), the training for our proposed model resembles self-supervised learning, in the sense that training labels can be generated from the data itself. Three main steps are involved in generating the training data, $\{x_i, y_i\}$:

200 (I) The raw 3D volumes are firstly aligned to a common reference atlas space with the method proposed in (Namburete et al., 2018), followed by a manual correction step. For every aligned 3D volume, $V \in \mathcal{R}^{h \times w \times d}$, where h , w and d are the height, width and depth of the 3D volume respectively, there is an associated binary mask of skull, $B \in \{0, 1\}^{h \times w \times d}$.
 205 The masks are generated by the CNN model proposed in (Moser et al., 2019).

(II) Following the sampling scheme adopted by Hou et al. (2017, 2018), 2D images and their corresponding 2D binary masks are sampled from the aligned 3D volumes (V), and 3D binary mask (B). In order to generate 2D images that are evenly distributed in a 3D volume, the surface normal of the sampling planes should be evenly spaced on the surface of a unit sphere (Hou et al., 2017), and this can be achieved by Fibonacci sphere sampling of polar coordinates, $p(\phi, \theta)$, where ϕ and θ are the azimuth and elevation angles respectively. Assuming m surface normals are sampled, $\{\phi_i\}_{i=1}^m$ and $\{\theta_i\}_{i=1}^m$ can be calculated by:

$$\phi_i = \frac{2\pi(i-1)}{(\sqrt{5}+1)/2} \quad (1)$$

$$\theta_i = \cos^{-1} \left(\frac{2(1-i)}{m} \right) \quad (2)$$

By defining the surface normal by Eq. 1 and 2, the coordinate of the centre point of the sampling plane as well as the in-plane rotation (*i.e.* plane rotation about its surface normal), 2D images can be sampled from
 210 3D volumes.

(III) The sampled 2D images are randomly processed by one of the three proposed ways during training, namely (i) masking the 2D images by the convex hull of the associated sampled 2D binary masks to remove *most* of the extracranial contents, (ii) masking the 2D images by 2D circular masks with arbitrary size larger than the associated sampled 2D binary
 215 masks to remove *part* of the extracranial contents or (iii) not masking

the 2D images at all to keep *all* the extracranial contents. While (i) and (ii) prevent the model from making predictions based on the background (*i.e.* extracranial structures) of the images, (iii) tries to minimize the influence of the shape and size of the binary masks, which are normally unavailable during inference, towards the prediction. Also, since 2D images are artificially sampled from 3D fetal brain volumes, they may look differently compared to the actual 2D images, in terms of resolution, intensity and noise. We use extensive data augmentation to make the model more generalizable, including geometrical transformation, scaling, contrast modification, and addition of random noise.

These three pre-processing steps are only required during training, but not for inference when the trained network is being employed to actual 2D images.

3.2. Model Architecture

The input to our proposed network is a set of an arbitrary number of 2D images $\{\mathbf{I}_i\}_{i=1}^N$, $\mathbf{I}_i \in \mathcal{R}^{h \times w}$, where N , h and w are the number of images, height and width of image, respectively. The output is the set of corresponding predicted locations $\{\mathbf{p}_i\}_{i=1}^N$, where $\mathbf{p}_i \in \mathcal{R}^{3 \times 3}$, referring to the 3 Cartesian coordinates (*i.e.* x, y, z) of the 3 landmarks that define the predicted plane. Following the approach proposed in Hou et al. (2017, 2018), we use the centre, the bottom right and left corners of a plane as the landmarks to define the predicted plane. In order to simulate the motion of an actual ultrasound scan, during training, a constraint is imposed to the N input images such that the distance between two consecutively sampled slices (*i.e.* \mathbf{I}_i and \mathbf{I}_{i+1}) should be smaller than a predefined value, which is 20 pixels.

Our proposed network consists of 4 sequential modules: *Feature extraction* (Fig. 2a): a feature encoder (*i.e.* a shared CNN backbone) is used to generate a fixed-length feature vector, \mathbf{v}_i , to represent each input image. *Comparison* (Fig. 2b): the feature vectors for each image are compared pairwise to compute the relationship between every pair of input images, which is further represented

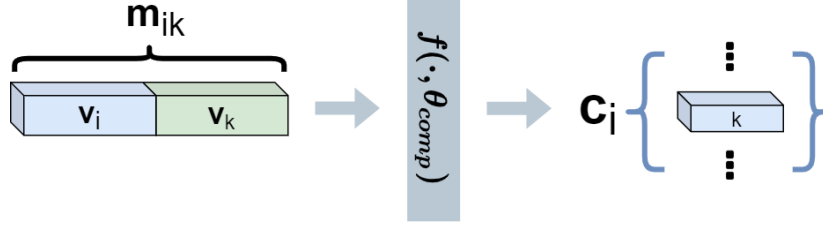


Figure 3: The processing unit of the *Comparison* module (Fig. 2b). A pair of feature vectors, \mathbf{v}_i and \mathbf{v}_k , are concatenated and passed to the comparison network, $f(\cdot, \theta_{comp})$, to output a comparative feature vector, \mathbf{c}_{ik} . (Best viewed in color)

by the set of comparison feature vectors, $\{\mathbf{c}_{ij}\}_{i=1, j=1}^{N, N}$. *Attention (Fig. 2c)*: an attention mechanism is applied on the set of feature vectors to weight the contribution of each pairwise relationship. *Prediction (Fig. 2d)*: while generating a summarization feature vector of every input image for prediction of position in 3D space, the attention matrix, is used to weight the comparison feature vectors. Each module is described in more detail below.

3.2.1. Feature extraction (Fig. 2a)

A feature extractor (*i.e.* a shared CNN backbone) is used to generate a fixed-length feature vector, $\mathbf{v}_i \in \mathcal{R}^{1 \times 512}$, for each input image, \mathbf{I}_i . A common feature encoder (*i.e.* shared weights) is used for all input images, such that the feature extraction is invariant to the permutation and number of input images. This is a desirable property for ultrasound images analysis due to the randomness of freehand image acquisition.

In our case, the feature extractor, $\psi(\cdot, \theta_{feat})$, parameterized by θ_{feat} , is based on the VGG-16 network architecture (Simonyan and Zisserman, 2015). With an arbitrary number of 2D input images, $\{\mathbf{I}_i\}_{i=1}^N, \mathbf{I}_i \in \mathcal{R}^{h \times w}$, the output from this module is:

$$[\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N] = [\psi(\mathbf{I}_1; \theta_{feat}), \psi(\mathbf{I}_2; \theta_{feat}), \dots, \psi(\mathbf{I}_N; \theta_{feat})] \quad (3)$$

3.2.2. Comparison (Fig. 2b)

260 The set of feature vectors, $\{\mathbf{v}_i\}_{i=1}^N$, is compared pairwise in this module. Instead of directly predicting the location of the image from its corresponding feature vector (*i.e.* each image position is predicted independently of all others) (Hou et al., 2017, 2018), we believe that it will be beneficial for each input image to also consider its relative position with respect to other images, as all
265 images are different planes of the brain of the same fetus and, hence, likely to be inter-correlated. This is achieved by combining the feature vector to generate a comparative feature vector, \mathbf{c}_{ij} , of every input image pair, \mathbf{I}_i and \mathbf{I}_j .

This comparison is implemented in two steps, which are summarized by the processing unit as shown in Fig. 3. Firstly, concatenation between vector pairs is computed, which can be formally expressed as:

$$[\mathbf{m}_{11}, \mathbf{m}_{12}, \dots, \mathbf{m}_{NN}] = [(\mathbf{v}_1 \parallel \mathbf{v}_1), (\mathbf{v}_1 \parallel \mathbf{v}_2), \dots, (\mathbf{v}_N \parallel \mathbf{v}_N)] \quad (4)$$

where \parallel is the concatenation operator and $\{\mathbf{m}_{ij}\}_{i=1,j=1}^{N,N}$, $\mathbf{m}_{ij} \in \mathcal{R}^{1 \times 1024}$ is the set of concatenated feature vectors.

Secondly, the set of concatenated feature vectors is passed as input to the comparison network, $f(\cdot, \theta_{comp})$, parameterized by θ_{comp} . The comparison network is a fully connected layer that merges the information of the two feature vectors into a comparative feature vector:

$$[\mathbf{c}_{11}, \mathbf{c}_{12}, \dots, \mathbf{c}_{NN}] = [f(\mathbf{m}_{11}; \theta_{comp}), f(\mathbf{m}_{12}; \theta_{comp}), \dots, f(\mathbf{m}_{NN}; \theta_{comp})] \quad (5)$$

270 where $\{\mathbf{c}_{ij}\}_{i=1,j=1}^{N,N}$, $\mathbf{c}_{ij} \in \mathcal{R}^{1 \times 512}$ is the set of comparative feature vectors.

3.2.3. Attention (Fig. 2c)

Different comparative feature vectors (\mathbf{c}_{ij}), may contribute differently to the final prediction of plane position. We propose to compute the relative contribution of each pairwise comparison by using an attention module (Vaswani
275 et al., 2017). The module will learn to assign more attention (*i.e.* a higher scalar weight) to comparisons with higher relational contribution and vice-versa. Contribution means the extent of any type of relationship, for example the similarity

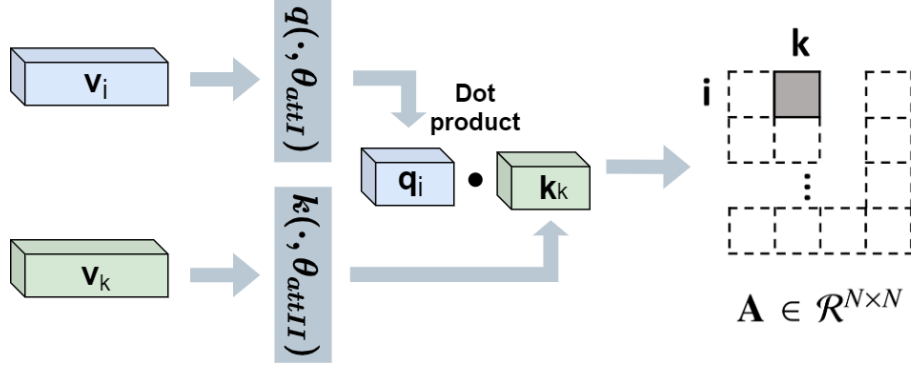


Figure 4: The processing unit of the *Attention* module (Fig. 2c). The dot product between a pair of embedded feature vectors, \mathbf{q}_i and \mathbf{k}_k , gives rise to \mathbf{A}_{ik} . (Best viewed in color)

between a pair of images, which is related to the final prediction and hence can be learned by the model from the loss. The output of this attention module will be an attention matrix, \mathbf{A} .

Fig. 4 displays the processing unit of the *Attention* module. To compute the attention matrix, $\mathbf{A} \in \mathcal{R}^{N \times N}$, we will compute the dot product between pairs of feature vectors, $\{\mathbf{v}_i\}_{i=1}^N$, in an embedding space as follow:

$$\mathbf{A}(i, j) = q(\mathbf{v}_i; \theta_{attI}) k(\mathbf{v}_j; \theta_{attII})^T \quad (6)$$

where $q(\cdot, \theta_{attI})$ and $k(\cdot, \theta_{attII})$ are embedding networks (*i.e.* multilayer perceptrons), parameterized by θ_{attI} and θ_{attII} , respectively, that map the feature vectors into an embedding space, $\mathcal{R}^{1 \times 256}$.

3.2.4. Prediction (Fig. 2d)

Fig. 5 shows the processing unit of the *Prediction* module. To compute the final prediction of each input image, the prediction module uses the attention matrix, \mathbf{A} , to weight the comparative feature vectors, $\{\mathbf{c}_{ij}\}_{i=1, j=1}^{N, N}$, to compute a summarization feature vector, $\mathbf{s}_i \in \mathcal{R}^{1 \times 512}$, for every input image, \mathbf{I}_i . The summarization feature vector, \mathbf{s}_i , gathers information from all images, weighted by the learned contribution towards the prediction of \mathbf{I}_i and is computed as

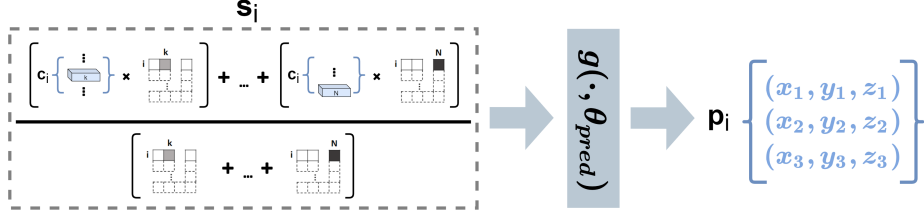


Figure 5: The processing unit of the *Prediction* module (Fig. 2d). The summarization feature vector, \mathbf{s}_i , is computed by $\{\mathbf{A}_{ik}\}_{k=1}^N$ and $\{\mathbf{c}_{ik}\}_{k=1}^N$. It is then processed by the prediction network, $g(\cdot, \theta_{pred})$, to output the set of predicted locations \mathbf{p}_i . (Best viewed in color)

follow:

$$\mathbf{s}_i = \frac{\sum_{j=1}^N \mathbf{A}(i, j) \mathbf{c}_{ij}}{\sum_{j=1}^N \mathbf{A}(i, j)} \quad (7)$$

The set of predicted locations $\{\mathbf{p}_i\}_{i=1}^N$, $\mathbf{p}_i \in \mathcal{R}^{3 \times 3}$, is obtained by passing the set of summarization feature vectors, $\{\mathbf{s}_i\}_{i=1}^N$, to the prediction network, $g(\cdot, \theta_{pred})$, parameterized by θ_{pred} :

$$[\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_N] = [g(\mathbf{s}_1; \theta_{pred}), g(\mathbf{s}_2; \theta_{pred}), \dots, g(\mathbf{s}_N; \theta_{pred})] \quad (8)$$

285 In summary, the predicted location, \mathbf{p}_i , of image, \mathbf{I}_i , is derived from \mathbf{s}_i and hence the weighted sum of \mathbf{c}_{ij} for all j . In other words, when predicting the location of image \mathbf{I}_i , information of all images within the same space, $\{\mathbf{I}_j\}_{j=1}^N$, will be considered. Furthermore, their relative contribution and degree of relationships with \mathbf{I}_i will be taken into account by the attention matrix, \mathbf{A} .

290 3.3. Loss Function

During training, we apply the mean least-square error as the loss function:

$$L_2(\hat{\mathbf{p}}, \mathbf{p}) = \frac{1}{N} \sum_{i=1}^N (\hat{\mathbf{p}}_i - \mathbf{p}_i)^2 \quad (9)$$

where $\hat{\mathbf{p}}$ and \mathbf{p} are the ground-truth and predicted locations, respectively.

4. Experimental Setup

4.1. Dataset

The 3D ultrasound fetal brain volumes ($160 \times 160 \times 160$ voxels at a resolution of $0.6 \times 0.6 \times 0.6$ mm³) were obtained as part of the INTERGROWTH-21st study (Papageorgiou et al., 2014), which were collected using a Philips HD9 curvilinear probe at a 2–5 MHz wave frequency. For each 3D volume, there is at least one associated 2D image taken at the standard TT plane routinely used for biometric and structural assessment. Both the 2D images and 3D volumes were acquired following strict requirements to ensure that the image quality satisfied pre-defined criteria (The INTERBIO-21st Consortium, 2012). For instance, the fetal skull occupied at least 50% of the image, and the image was not affected by fetal or maternal movements. Fetal anomaly ultrasound scan is recommended to be undertaken between 18 to 21 gestational weeks and some flexibilities are allowed for this age range (Public Health England, 2018). In this study, images were selected from fetuses with gestational age ranging from 18 to 22 gestational weeks. Each image was masked and aligned to a coordinate space as described in Section 3.1.

A summary of training and different experiments and their corresponding dataset is presented in Table 4 in the *Supplementary Materials*.

4.2. Training Details

In this study, we re-implement a baseline model (only slight modification based on network architecture proposed by Hou et al. (2017, 2018)) and compared its performance to our proposed model. The exact network architectures of the baseline model and our proposed model are presented in Table 3 in the *Supplementary Materials*. Optimization was achieved using the ADAM algorithm (Kingma and Ba, 2015) with mini-batches of size 32. The initial learning rate was set to 10^{-4} , which was decreased by half when errors plateaued.

Fifty and fifteen 3D volumes acquired at 21 gestational weeks were selected for training and validation, respectively. For each 3D volume in each training

epoch, 50 evenly distributed normals were sampled using the Fibonacci Sphere Sampling method as described in Section 3.1. Along each normal, 15 planes perpendicular to the normal, with average spacing of 2.4 mm were chosen (Fig. 7a). For each plane, four 2D slices (160×160 pixels), with random in-plane rotation were sampled. Therefore, in total, there were 150,000 and 45,000 images for each training and validation epoch, respectively.

Since an infinite number of different 2D slices can be sampled from a 3D volume in principle, we made use of this feature and introduced random variation to the sampling parameters for each training epoch. Therefore, the 150,000 training images were expected to be different for every epoch. This kept the number of training data for each training epoch relatively small as compared to Hou et al. (2017, 2018), while the number of different images used for the whole training was much larger. We regarded this as a type of data augmentation, which may prevent the model from overfitting while having a reasonable amount of varied training data for each epoch.

4.3. Evaluation metrics

Three evaluation metrics were used to evaluate and compare the performance of the models. First, Euclidean distance (ED) between all the coordinates of the predicted and ground-truth planes is computed as follow:

$$ED = \frac{\sum_{i=1, j=1}^{h, w} dist(\hat{\mathbf{p}}_{ij}, \mathbf{p}_{ij})}{h \cdot w} \quad (10)$$

where $\hat{\mathbf{p}}$ and \mathbf{p} are the predicted and ground-truth planes and $dist(\hat{\mathbf{p}}_{ij}, \mathbf{p}_{ij}) = |\hat{\mathbf{p}}_{ij} - \mathbf{p}_{ij}|^2$ computes the Euclidean distance between the two points, $\hat{\mathbf{p}}_{ij}$ and \mathbf{p}_{ij} , where $\hat{\mathbf{p}}_{ij}$ and \mathbf{p}_{ij} are the (x, y, z) coordinates of the pixel ij on the predicted and ground-truth planes.

Secondly, plane angle (PA) between the predicted and ground-truth planes are computed as follow:

$$PA = \cos^{-1}(\hat{\mathbf{n}} \cdot \mathbf{n}) \quad (11)$$

where $\hat{\mathbf{n}}$ and \mathbf{n} are the surface normals of the predicted and ground-truth planes, respectively. Smaller ED and PA suggest that the ground-truth and predicted planes locate closely to each other, which may represent more accurate prediction.

345 Thirdly, normalized cross-correlation (NCC) (Yoo and Han, 2009) between the input image and image sampled from the predicted plane is computed. Larger values may suggest higher similarity between the two images and more accurate prediction of plane position.

4.4. Comparison with Baseline Model

350 Images sampled from 3D volumes were used to quantitatively evaluate the performance of different models. Our proposed model and the baseline model were compared using the evaluation metrics introduced in Section 4.3. In addition, in order to investigate the individual contribution of our newly proposed modules, namely the *Comparison* module (Section 3.2.2) and the *Attention* module (Section 3.2.3), ablation study has been conducted by removing the *At-*
 355 *tention* module of our proposed network and applying equal weighting to every comparative feature vector, \mathbf{c}_{ij} (*i.e.* replacing the attention matrix, \mathbf{A} , with a matrix of ones).

4.4.1. Sensitivity to Input Image Support

360 Since our proposed model makes a prediction for each image by grouping information of all input images, prediction accuracy may be sensitive to the number of input images. Therefore, different numbers, $N \in \{1, 2, 4, 8, 16, 32, 64, 128\}$, of input images were tested to investigate on how changing the number of input images may affect the prediction of our proposed model.

365 4.4.2. Application to Broader Gestational Age Range

Trained on images at 21 gestational weeks, images within a broader gestational age range (*i.e.* 18 to 22 gestational weeks) were tested to evaluate the generalizability of the models to different ages. For different ages, a slight change of brain anatomical structure is expected (Pistorius et al., 2010).

370 As a comparison, images of the whole gestational age range (*i.e.* 18 to 22 gestational weeks) were used to train a different set of models to verify if a single model can be used on a broad gestational age range.

4.5. Relationship between Plane Location and Accuracy of Prediction

375 In Section 4.4, 2D images have been sampled at different locations in each 3D volume. The results of the images sampled from the 15 fetal brain volumes in Section 4.4.1 were further analyzed to investigate how accurate our proposed model is in predicting images in different regions of the 3D brain. Specifically, the accuracy of prediction of images sampled along different directions and at different distances from the centre of the brain were studied. Fig. 7a shows how
380 planes were sampled from the 3D volume along one normal using the Fibonacci Sphere Sampling method as described in Section 3.1.

4.6. Real 2D Image Acquisition of Standard TT Plane

Real 2D images taken at the standard TT plane were tested. These images were acquired with the 15 3D test volumes in Section 4.4.1. For each 2D image,
385 plane location was predicted by our proposed model and annotated by 2 individual experts separately. Using the predicted plane locations, the corresponding 2D images were sampled from the associated 3D volume. Variations, measured by the evaluation metrics introduced in Section 3.1, were estimated between the 3 different sets of predictions and annotations. They were further analyzed by
390 one-way ANOVA.

4.7. Video of Freehand Fetal Brain Scanning

In addition to the single standard plane images as described in Section 4.6, 15 videos of 2D scans acquired from 4 subjects with gestational age between 19 to 21 weeks during fetal exams of the brain were also tested. The videos were
395 acquired by sweeping the ultrasound probe along different directions during scanning. Therefore, the videos were composed of 2D views corresponding to different locations of the fetal brain, which may or may not be a standard

Score	1	3	5
Description	The video frame presents totally different structures from the sampled atlas slice; the location of the plane in 3D atlas space is incorrect	The video frame presents most of the structures as the sampled atlas slice; the location of the plane in 3D atlas space indicates roughly correct and reasonable location	The video frame presents the same structures as the sampled atlas slice; the location of the plane in 3D atlas space indicates the correct location

Table 1: Description of the scoring scale for evaluating the performance by the models on real 2D ultrasound images sampled from scanning videos. Scale of score 1 to 5 is used, where score 1 indicates totally incorrect prediction while score 5 indicates perfect prediction.

plane. Every video was treated as a set of 2D images for testing. Using the predicted plane locations, 2D slices were sampled from the 3D atlas volume.

400 The video frames and the corresponding sampled 2D atlas slices were compared qualitatively, in terms of structures present and image orientation. Also, we selected 50 frames from the videos and obtained the predictions by both our proposed model and the baseline model. Using the scoring scale as described in Table 1, our clinical collaborators scored the predictions from both models for

405 further comparison.

4.8. Impact of Learned Attention

As mentioned in Section 3.2.3, the attention matrix (\mathbf{A}) weights the contribution of each pairwise comparison of the set of input images. To verify that the *Attention* module (Fig. 2c) actually learns to assign meaningful weights, we

410 further analyzed the results of the slices sampled from the 15 fetal brain volumes in Section 4.4.1. Using $N = 4$ input images (for easier comparison and visualization), the normalized attention, $\frac{\sum_{j=1}^4 \mathbf{A}(i,j)}{\sum_{i=1}^4 \sum_{j=1}^4 \mathbf{A}(i,j)}$, associated to each input

image was investigated.

5. Results

415 5.1. Comparison with Baseline Model

The results of the two experimental settings (Section 4.4.1 and 4.4.2) are presented in Fig. 6. For both settings, all three evaluation metrics indicated that the performance of our proposed models surpassed that of the baseline model.

420 For each 3D volume, 3000 2D images were sampled in the same way as described in Section 4.2. Two settings were investigated, namely variation on number of input image (Section 4.4.1) and generalization to a broader gestational age range (Section 4.4.2).

5.1.1. Sensitivity to Input Image Support

425 Fifteen 3D fetal brain volumes with gestational age of 21 gestational weeks were used for evaluation, yielding at total of 45,000 2D test images. Different numbers, $N \in \{1, 2, 4, 8, 16, 32, 64, 128\}$, of input images were tested.

The results of this experiment are presented in Figs. 6a to 6c. Since the number of images would not affect the prediction of the baseline model, results
430 of the baseline model were the same for different number of input images.

For our proposed models, both with and without the *Attention* module, performance increased with the number of input images by as much as 17%, 7% and 5% as indicated by ED, PA and NCC, respectively. This may be reasonable because our proposed models make a prediction for each image by grouping
435 information of all input images. More input images may provide more information for the prediction. Also, the ablation study suggested that the *Comparison* and *Prediction* modules, which are responsible for grouping information of all input images, may primarily lead to improvement when compared to the baseline model by around 19% (ED), 8% (PA) and 15% (NCC). The addition of the
440 *Attention* module, which assigns weights to the grouping of information, contributed to further improvement by an extra 5% (ED), 2% (PA) and 5% (NCC).

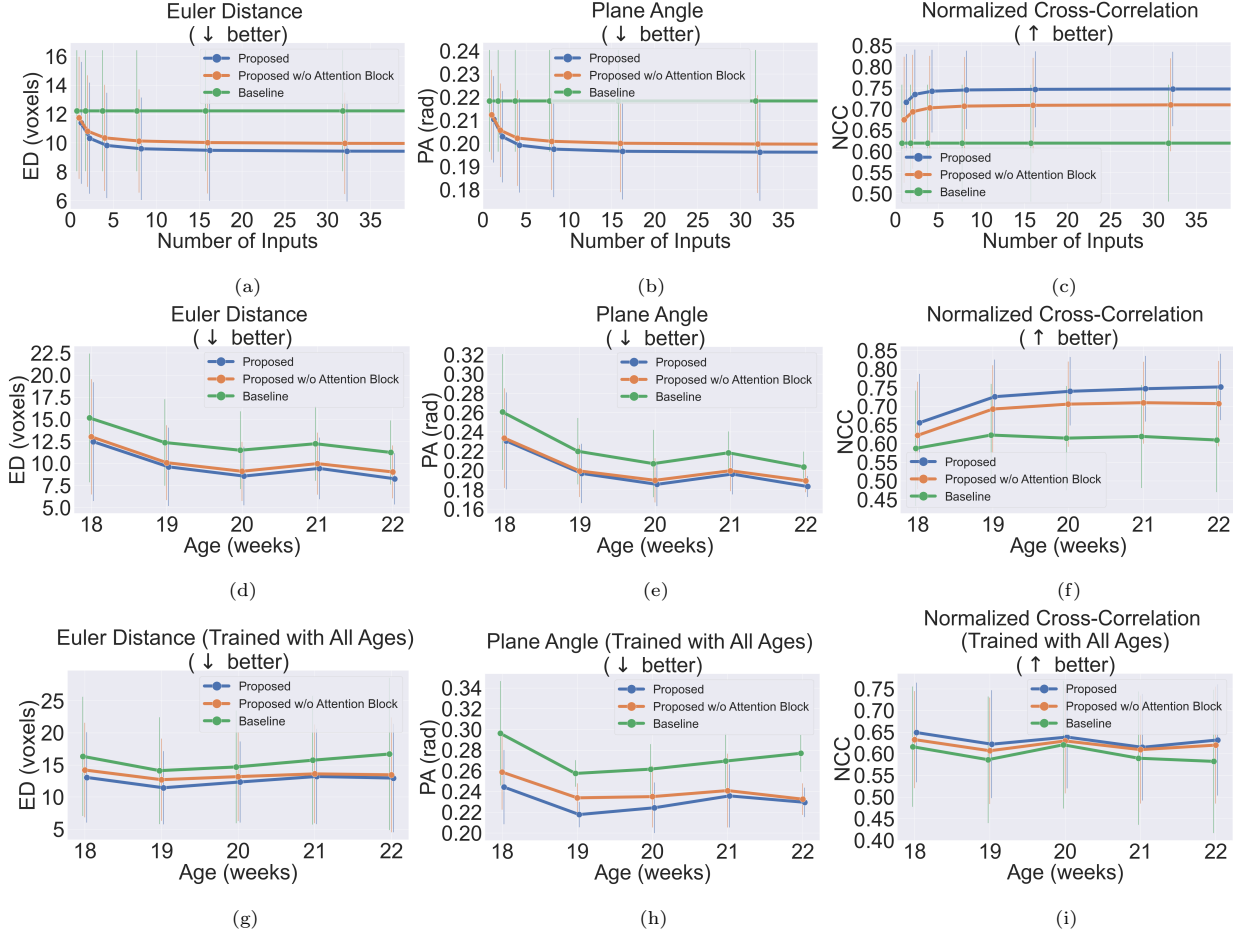


Figure 6: The accuracy of the baseline model (green), our proposed model (blue) and our proposed model without *Attention* module (orange). Upper row shows the mean results (\pm standard deviation) of (a) Euclidean distance, (b) plane angle and (c) normalized cross correlation between groundtruth and prediction for different numbers of input images. Middle row shows the mean results (\pm standard deviation) of (d) Euclidean distance, (e) plane angle and (f) normalized cross correlation between groundtruth and prediction for different gestational ages. Bottom row shows the mean results (\pm standard deviation) of (g) Euclidean distance, (h) plane angle and (i) normalized cross correlation between groundtruth and prediction by models trained with images of all gestational ages for different gestational ages. The dots in the graph are slightly shifted for better visualization of the standard deviation. (Best viewed in color)

Although such further improvement may appear to be marginal as shown in Figs. 6a to 6c, it is statistically significant for every number of input images and evaluation metric concerned ($p < 0.05$, t-test).

445 In addition, all three evaluation metrics showed that the performance of our proposed models surpassed that of the baseline model by as much as 23% (ED), 11% (PA) and 21% (NCC) and when the number of input images increased, the improvement was more significant. We observed that the gain in accuracy nearly saturated when the number of inputs exceeds 32 and therefore in Figs. 6a
450 to 6c, we omitted the results for $N \in \{64, 128\}$ for clearer visualization. The result distribution of $N = 4$ and $N = 64$ is further displayed in Fig. 12, which shows that although increasing the number of input images may not have a significant impact on reducing outliers, it shifted the distribution towards better performance.

455 5.1.2. Application to Broader Gestational Age Range

Fetal brain volumes with gestational age of 18 gestational weeks (50 volumes), 19 gestational weeks (34 volumes), 20 gestational weeks (57 volumes), 21 gestational weeks (15 volumes) and 22 gestational weeks (9 volumes) were used for testing in this experiment.

460 The results of the first part of this experiment are summarized in Figs. 6d to 6f. Using models trained on images with gestational age of 21 weeks, we tested the models on images with gestational age ranging from 18 to 22 weeks.

Two observations can be obtained: firstly, for all ages, predictions made by our proposed models were more accurate than those made by the baseline model
465 by as much as 23% (ED), 11% (PA) and 21% (NCC). Also, predictions made by our proposed model without the *Attention* module were slightly less accurate than the complete version of the proposed model. The slight improvement caused by the incorporation of the *Attention* module is statistically significant for every age and evaluation metric concerned ($p < 0.05$, t-test). Secondly, we
470 observed that in general, predictions on images at younger gestational ages were less accurate by as much as 51% (ED), 26% (PA) and 13% (NCC). A

potential explanation is that fetuses during the second trimester are undergoing rapid neuro-development (Pistorius et al., 2010). Therefore, brain structures of fetuses at younger gestational ages may look quite different from those of
475 fetuses of gestational age of 21 weeks, which are the images that the models were trained on.

The results of the second part of this experiment are summarized in Figs. 6g to 6i, where the models have been trained and tested on images of the whole gestational age range (*i.e.* 18 to 22 gestational weeks). When compared to the
480 results of models trained on images of a single age, two observations can be obtained: firstly, predictions made by our proposed models were more accurate than those made by the baseline model by as much as 21% (ED), 16% (PA) and 9% (NCC). Also, predictions made by our proposed model without the *Attention* module were slightly less accurate than the complete version of the proposed
485 model. The slight improvement caused by the incorporation of the *Attention* module is statistically significant for every age and evaluation metric concerned ($p < 0.05$, t-test). Secondly, predictions made by models trained on images of the whole gestational age range were less accurate when compared to those made by models trained on images of just 21 weeks. This may be reasonable because
490 images of different gestational ages were registered to different atlases as brain structures presented at different gestational ages may look quite different. For a single age, every plane location in the atlas space corresponds to a unique set of 2D image features. However, when a single model is trained with images of different gestational ages, it is equivalent to combining different unique atlas
495 spaces into one and every plane location in this combined atlas space corresponds to multiple sets of 2D image features, each belongs to a specific age and hence they can be quite different to each other. This may be a more difficult and ambiguous learning task when compared to training models on images of a single age. Therefore, one single model trained on images of a broad gestational
500 age range may have poorer performance when compared to models trained on a single age.

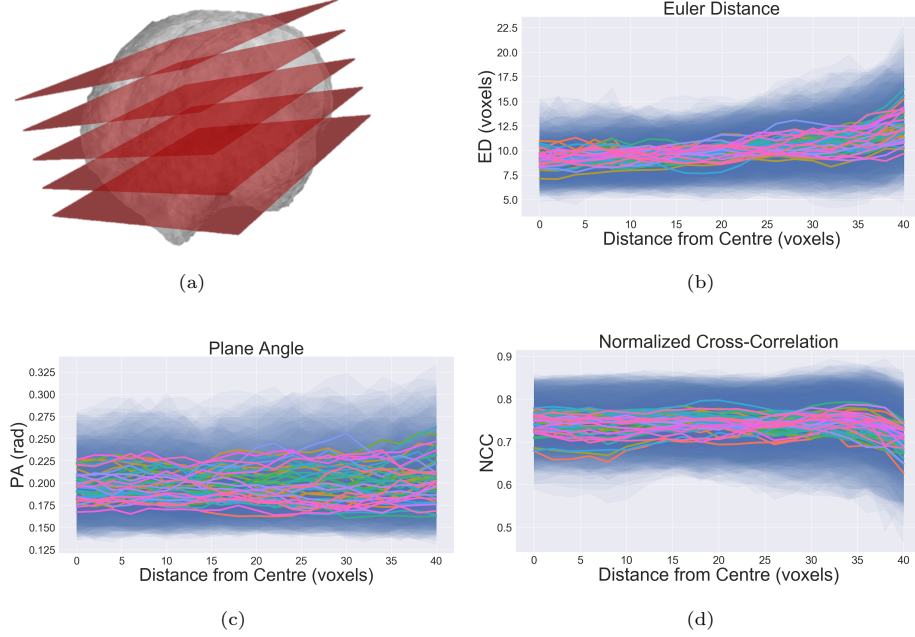


Figure 7: Plane location and accuracy of prediction. (a) shows how planes were sampled from the 3D volume along one normal of the unit sphere. Mean results (\pm standard deviation) of (b) Euclidean distance, (c) plane angle and (d) normalized cross correlation between groundtruth and prediction for images sampled from different locations of the 3D brain volumes are computed. Each curve in the figure indicates the mean results of one normal of the unit sphere and the blue shadow around it is the standard deviation of the results. Slices perpendicular to it and at different distance away from the centre of the 3D brain volumes were sampled and tested. (Best viewed in color)

5.2. Relationship between Plane Location and Accuracy of Prediction

The results of finding the relationship between plane location and accuracy of prediction are presented in Fig. 7. Similar to the sampling procedure as introduced in Section 4.2, for each 3D volume, 50 normals evenly distributed on the unit sphere were chosen and each of them was represented by a colored curve in Fig. 7. In Fig. 7, values on each colored curve indicate the mean results, while the blue shadow around the curve is the standard deviation of the results. Along each normal and at different distance away from the centre of the 3D brain volumes, planes perpendicular to the normal were sampled.

	ED (voxels)	PA (rad)	NCC
M1 <i>v.s.</i> M2	9.12 ± 4.01	0.126 ± 0.055	0.867 ± 0.093
M1 <i>v.s.</i> Model	11.36 ± 3.26	0.179 ± 0.095	0.841 ± 0.096
M2 <i>v.s.</i> Model	11.44 ± 5.02	0.180 ± 0.120	0.837 ± 0.080
P value (one-way ANOVA)	0.257	0.227	0.639

Table 2: Comparison with manual annotation on real 2D images taken at the standard TT plane. Mean results (\pm standard deviation) and one-way ANOVA results between first manual annotation (M1), second manual annotation (M2) and prediction by our proposed model are displayed. P values of the one-way ANOVA suggests the comparable performance by our proposed model and human annotations.

Firstly, suggested by all three evaluation metrics, the performance of our proposed model in predicting images sampled along different directions (*i.e.* different lines in Fig. 7) were similar. The Euler distance, plane angle and normalized cross correlation were around 10 voxels, 0.20 rad and 0.75 respectively, which were similar to the overall result presented in Figs. 6d to 6f. In other words, the performance of our proposed model does not depend on the geometric orientation of the images sampled, which is desirable because during 2D freehand ultrasound scanning, images along different directions may be acquired.

Secondly, as suggested by ED (Fig. 7b) and NCC (Fig. 7d), when the images were farther away from the centre of the 3D brain volumes, the accuracy of the prediction dropped. This is reasonable because in general, images farther away from the centre, especially those near the edge of the brain, contain fewer indicative structures and hence are less informative and it is more difficult to predict their 3D location (Hou et al., 2017, 2018).

5.3. Real 2D Image Acquisition of Standard TT Plane

Real 2D images taken at the standard TT plane were tested. Table 2 summarizes the variations between the plane locations predicted by our proposed model and manually annotated by 2 individual experts. Although the mean val-

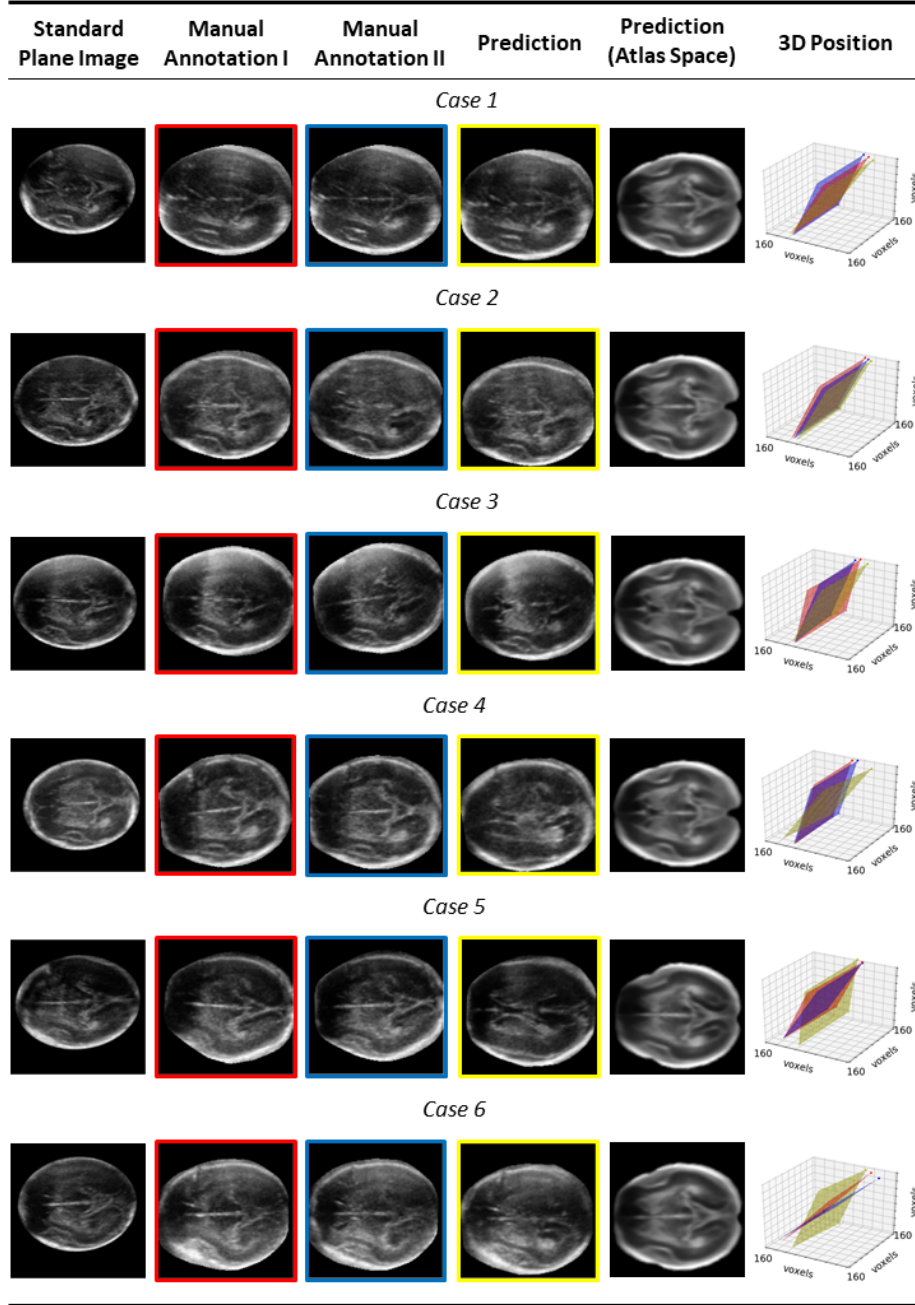


Figure 8: Visualization of manual annotation comparison. Six examples of native and masked 2D scans taken at the standard TT plane (*first column*); slices sampled from the corresponding 3D volume using the first manual annotation (*second column*), second manual annotation (*third column*) and our proposed model's prediction (*fourth column*); slices sampled from the 3D atlas using our proposed model's prediction (*fifth column*) and the position of the aforementioned slices in the 3D atlas space (*sixth column*). Frame color of the images (*second to fifth column*) corresponds to the planes as shown in the 3D atlas space (*sixth column*). (Best viewed in color)

ues of the three evaluation metrics may suggest that the variation between the two sets of manual annotation is smaller than that between the model prediction and the manual annotations, p values of 0.257, 0.227 and 0.639 as calculated by the one-way ANOVA pointed out that we failed to reject the null hypothesis, and there is no difference between the three groups of comparison, suggesting the comparable performance by our proposed model and human annotations.

We understand that the result obtained by the one-way ANOVA may not be convincing enough due to the limited amount of test images. Therefore, we further analyzed the 15 test images independently. While for most cases, the model prediction closely matched both (case 1 and 2 in Fig. 8) or either (case 3 in Fig. 8) set(s) of the manual annotation, we found out that only three cases exhibited significant difference (*i.e.* more than 30% difference) between the model prediction and both sets of the manual annotation. They are presented as cases 4 to 6 in Fig. 8. It is evident that both the appearance (*fourth column*) and 3D location (*sixth column*) of the sampled slices using our proposed model’s prediction differ significantly with those sampled from the manual annotations (*second* and *third column*). However, the slices sampled from the 3D atlas using the prediction by our proposed model (*fifth column*) actually look much more similar to the input standard plane image (*first column*) than the slices sampled from the 3D volume (*fourth column*). In other words, the large variation between the model prediction and the manual annotations in these three cases is mainly due to the misalignment between the three volumes and the atlas. We checked the three volumes again and verified that the poor volume quality makes perfect alignment to the atlas space extremely challenging.

5.4. Video of Freehand Fetal Brain Scanning

Fig. 9 shows the results of four video examples. It can be observed that the video frames and the corresponding slices sampled from the atlas present similar anatomical structures in the same orientation. Also, the predicted plane locations generally match with the motion of the probe when acquiring the videos, which were roughly along the longitudinal axis (videos 1 and 2 of Fig. 9)

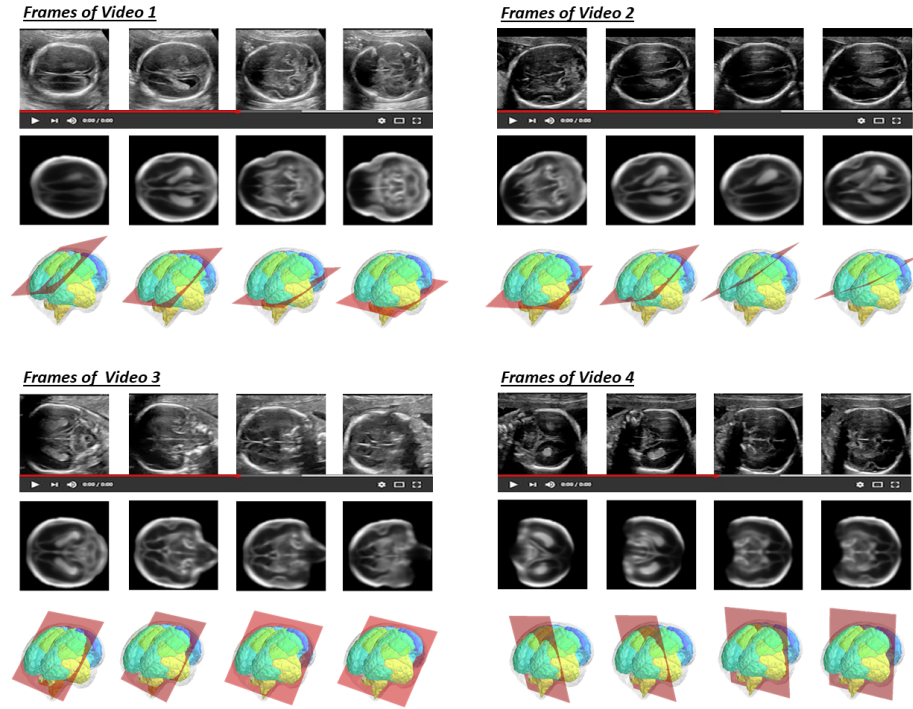


Figure 9: Results of four video examples. For each example, the upper row shows multiple frames of the video, which were input to our proposed model. Using the predicted plane locations, corresponding slices were sampled from the 3D atlas, which were shown in the middle row. The prediction plane location of each input video frame in the 3D atlas space was displayed in the bottom row. (Best viewed in color)

560 and the sagittal axis (videos 3 and 4 of Fig. 9), respectively.

It may be noted that the anatomical structures in the upper hemisphere of some of the video frames (*e.g.* video 2 of Fig. 9) are not clearly discernible. This is due to the interaction between the ultrasound wave and concave fetal skull, which results in the anatomical structures presented in the hemisphere near the ultrasound probe (*i.e.* upper part of the video frames) generally being
565 less visible (Namburete et al., 2017). On the other hand, the atlas represents both hemispheres. Therefore, the upper hemisphere of some of the video frames and that of their corresponding atlas slices may look different.

It was demonstrated in Section 5.2 that the accuracy of prediction decreases
570 when the input 2D image locates farther away from the centre of the brain. Fig. 10 shows two sets of consecutive frames which capture the external areas of the supratentorial region of the fetal brain. It can be observed that the slices sampled from the 3D atlas using the predicted plane location show completely different structures from their corresponding input frames and the predicted
575 locations for consecutive frames do not show a smooth transition, which both further verify that the performance of our proposed model would decline when the input 2D images capture areas farther away from the centre of the brain, which present very limited structural features.

In addition, our clinical collaborator scored the predictions from both our
580 proposed model and the baseline model for 50 selected frames following the scoring system as described in Table 1. The mean (\pm standard deviation) results for our proposed model and the baseline model are 3.12 (± 1.24) and 2.83 (± 0.84) respectively, with full score being 5.0. [Fig. 13 shows the distribution of the scores.](#) The better performance achieved by our proposed model is statistically
585 significant ($p < 0.05$, t-test). This result may further verify that our proposed model not only performs better on images sampled from 3D volumes, but also on real 2D ultrasound images.


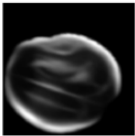
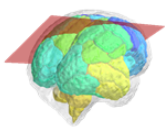

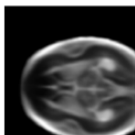
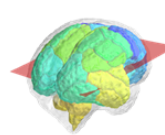


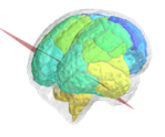

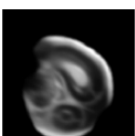
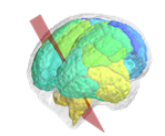
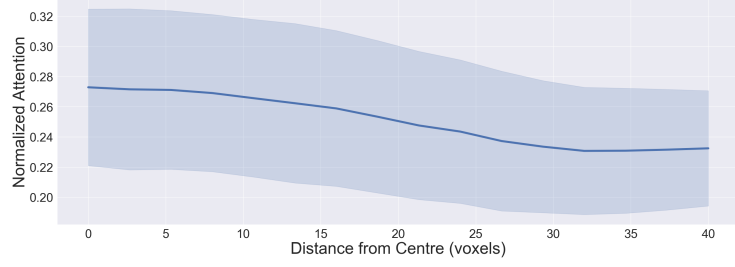
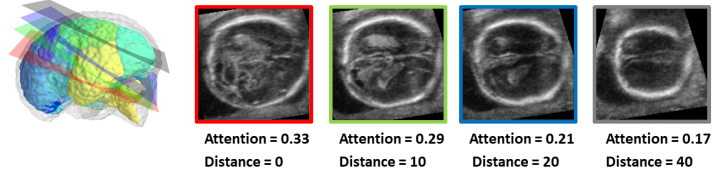
Frame	Input Frame	Prediction (Atlas Space)	3D Position
<i>Case 1</i>			
f			
$f + 1$			
<i>Case 2</i>			
f			
$f + 1$			

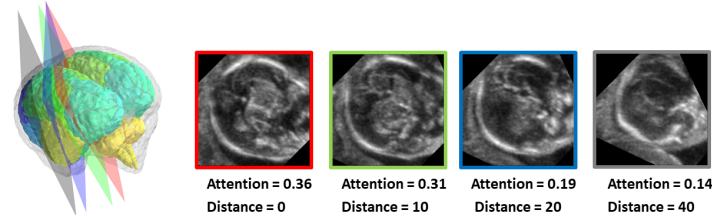
Figure 10: Suboptimal prediction. Predicted locations of consecutive frames, which capture the external areas of the supratentorial region of the fetal brain, are completely different and do not show a smooth transition. (Best viewed in color)



(a)



(b)



(c)

Figure 11: Attention visualization. (a) displays the relationship between the mean normalized learned attention (\pm standard deviation) and the sampled slices' distance from the centre of the fetal brain. (b) and (c) are two sets of attention visualization example. The 3D positions of the images are shown in the 3D fetal brain simulation on the left. (Best viewed in color)

5.5. Impact of Learned Attention

The results of the slices sampled from the 15 fetal brain volumes in Section 5.1.1 were further analyzed to verify that the *Attention* module (Fig. 2c) actually learns to assign meaningful weights. Fig. 11a shows that the learned attention decreases with the increasing sampled slices' distance from the centre of the fetal brain. If the learned attention is interpreted as the weighting of contribution of the pairwise comparison of input images, as mentioned in Section 3.2.3, Fig. 11a may verify that the *Attention* module of our proposed model actually learns to assign meaningful weights, because in general, regions closer to the centre of are more likely to contain richer structural information, and hence more indicative towards the final prediction. This can be visualized in Figs. 11b and 11c, where the images with blue and gray frames (*i.e.* sampled farther away from the centre) present less indicative structural information than the images with red and green frames (*i.e.* sampled closer to the centre). Therefore, the attention weight assigned to the images sampled farther away from the centre of the 3D volume, which quantify their degree of contribution towards the final prediction, is smaller in general.

6. Discussion and Conclusion

The methodology presented in this work was developed for the task of predicting the location of 2D ultrasound fetal brain images in a pre-defined 3D space. This may facilitate better identification and localization of different ultrasound scans clinically, and hence lead to more accurate and objective image acquisition and the analysis of fetal growth and development. In the recent literature, a closely related task is standard plane detection, which can be achieved using deep learning techniques (Chen et al., 2015b; Baumgartner et al., 2017; Chen et al., 2015a; Huang et al., 2017; Gao and Noble, 2017). In contrast, our work attempted to tackle a more general task, where not only standard plane, but also any arbitrary plane of the fetal brain can be detected and located. The idea is inspired by (Hou et al., 2017, 2018), but differs in that an arbitrary

number of images can be packed as a set input. Our proposed model makes use of all input images in the set to predict the location of each image, which leads to more accurate predictions. One may raise the concern that 2D slices sampled perpendicularly to the beam direction of a 3D volume will be of poor resolution, ending up with a domain gap between artificially sampled slices and real ultrasound video scans. However, in this paper, we demonstrate that, based on extensive data augmentation and complementary information from different volumes, *i.e.* 3D training volumes were acquired at different orientations and plane with poor-resolution image at one volume may have a corresponding plane with better-resolution image in another volume, the model trained with 2D slices sampled from 50 aligned 3D volumes can well generalize to real 2D ultrasound acquisitions and videos.

We have shown that the performance of our proposed model surpasses that of the baseline model for images with gestational age ranging from 18 to 22 weeks. It also shows that when the number of input images increases, the improvement is more significant. For freehand 2D ultrasound scanning, it is easy to acquire a large number of images. Therefore, it is practical to utilize our proposed model to predict the 3D location of the acquired images. Nevertheless, when more images are input to the model at the same time, more computing resources are needed but the gain in accuracy is marginal. Therefore, it may be reasonable to keep the number of images around 20 to 40 in practice. In addition, it has been demonstrated that our proposed model that was trained on images with gestational age of 21 weeks may somehow generalize to images of 18 to 22 gestational weeks. Nevertheless, fetuses during the second trimester are undergoing rapid neuro-development and brain structures of fetuses separated by one or two weeks may already look quite different (Pistorius et al., 2010). Also, we have shown that one single model trained on images of a broad gestational age range may have poorer performance. Therefore, different models trained on images with different gestational ages should be used in practice to achieve more accurate prediction. This may not be ideal as the current models are sensitive to gestational ages, which may affect the performance when ab-

normal brain development is present. As future works and clinical application, 3D volumes of different ages should be further registered to a single 3D atlas space to avoid non-deterministic prediction when a single model is trained with
650 images of different gestational ages registered to different atlas spaces.

As shown by the results of both sampled and actual 2D ultrasound data, the performance of our proposed model declined with increasing distance of the input images from the centre of the brain. This may be due to the fact
655 that images farther away from the centre generally contain fewer indicative structures and hence are less informative. Therefore, it is more difficult to predict 3D location of these images (Hou et al., 2017, 2018). Nevertheless, as shown by the results of our tests on the real 2D ultrasound acquisitions and videos, our proposed model can generally predict the location of the planes
660 of view located in the central region of the fetal brain quite successfully. In practice, these predictions and the continuous movement of the probe during freehand scanning may provide clues for identifying the 3D location of images farther away from the central region of the fetal brain.

Computational cost is another important factor when employing our proposed model for prediction. When using single GPU (GeForce GTX 1080 Ti),
665 the total time needed for loading images to the GPU and making prediction is around 0.059 second with 32 images (*i.e.* $N = 32$) being used for attention and comparison calculation. The computation cost increases to around 2.26 seconds if only CPU is available. Therefore, our proposed model can potentially
670 be used for real-time or near-real-time applications if hardware is available and implementation is further optimized.

In summary, we have presented a new CNN that can predict the location of 2D ultrasound fetal brain images in the 3D space. Using sampled 2D images from 3D volumes, we demonstrated that when more images are inputted to
675 our proposed model, prediction is more accurate. Furthermore, the prediction made by our proposed model generalizes to real 2D ultrasound acquisitions and videos, despite the model having only been trained with artificially sampled 2D slices. As future works, we would like to further develop our work as a

training and diagnostic tool that can help clinicians and sonographers when
680 they are acquiring and analyzing 2D fetal brain images in real time, to facilitate
more accurate and objective monitoring of fetal growth and diagnosis of CNS
abnormalities.

Acknowledgment

PH. Yeung is grateful for support from the RC Lee Centenary Scholarship.
685 A. Papageorgiou is supported by the National Institute for Health Research
(NIHR) Oxford Biomedical Research Centre (BRC). W. Xie is supported by
the UK Engineering and Physical Sciences Research Council (EPSRC) Pro-
gramme Grant Seebibyte (EP/M013774/1). A. Namburete is funded by the
UK Royal Academy of Engineering under its Engineering for Development Re-
690 search Fellowship scheme. We thank Lior Drukker for his valuable suggestions
and comments about the work.

References

- Alansary, A., Le Folgoc, L., Vaillant, G., Oktay, O., Li, Y., Bai, W., Passerat-
Palmbach, J., Guerrero, R., Kamnitsas, K., Hou, B., 2018. Automatic view
695 planning with multi-scale deep reinforcement learning agents, in: MICCAI,
Springer. pp. 277–285.
- Alansary, A., Rajchl, M., McDonagh, S.G., Murgasova, M., Damodaram, M.,
Lloyd, D.F., Davidson, A., Rutherford, M., Hajnal, J.V., Rueckert, D., 2017.
PVR: Patch-to-volume reconstruction for large area motion correction of fetal
700 MRI. *IEEE Transactions on Medical Imaging* 36, 2031.
- 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*
Transactions on Medical Imaging 36, 2204–2215.

- 705 Benacerraf, B.R., 2002. Three-dimensional fetal sonography: Use and misuse. *Journal of Ultrasound in Medicine* 21, 1063–1067.
- Bethune, M., Alibrahim, E., Davies, B., Yong, E., 2013. A pictorial guide for the second trimester ultrasound. *Australasian Journal of Ultrasound in Medicine* 16, 98–113.
- 710 Chen, H., Dou, Q., Ni, D., Cheng, J.Z., Qin, J., Li, S., Heng, P.A., 2015a. Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks, in: *MICCAI*, Springer. pp. 507–514.
- Chen, H., Ni, D., Qin, J., Li, S., Yang, X., Wang, T., Heng, P.A., 2015b. Standard plane localization in fetal ultrasound via domain transferred deep
715 neural networks. *IEEE Journal of Biomedical and Health Informatics* 19, 1627–1636.
- Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database, in: *CVPR*, IEEE. pp. 248–255.
- Dou, H., Yang, X., Qian, J., Xue, W., Qin, H., Wang, X., Yu, L., Wang, S.,
720 Xiong, Y., Heng, P.A., Ni, D., 2019. Agent with warm start and active termination for plane localization in 3D ultrasound, in: *MICCAI*, Springer. pp. 290–298.
- Falco, P., Gabrielli, S., Visentin, A., Perolo, A., Pilu, G., Bovicelli, L., 2000. Transabdominal sonography of the cavum septum pellucidum in normal fetuses in the second and third trimesters of pregnancy. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of
725 Ultrasound in Obstetrics and Gynecology* 16, 549–553.
- Ferrante, E., Paragios, N., 2017. Slice-to-volume medical image registration: A survey. *Medical Image Analysis* 39, 101–123.
- 730 Filly, R.A., Cardoza, J.D., Goldstein, R.B., Barkovich, A.J., 1989. Detection of fetal central nervous system anomalies: A practical level of effort for a routine sonogram. *Radiology* 172, 403–408.

- 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: MICCAI, Springer. pp. 305–313.
- 735
- Gholipour, A., Estroff, J.A., Warfield, S.K., 2010. Robust super-resolution volume reconstruction from slice acquisitions: Application to fetal brain MRI. *IEEE Transactions on Medical Imaging* 29, 1739–1758.
- Gonçalves, L.F., Lee, W., Espinoza, J., Romero, R., 2005. Three- and 4-
740 dimensional ultrasound in obstetric practice: Does it help? *Journal of Ultrasound in Medicine* 24, 1599–1624.
- Hou, B., Alansary, A., McDonagh, S., Davidson, A., Rutherford, M., Hajnal, J.V., Rueckert, D., Glocker, B., Kainz, B., 2017. Predicting slice-to-volume transformation in presence of arbitrary subject motion, in: MICCAI, Springer. pp. 296–304.
- 745
- Hou, B., Khanal, B., Alansary, A., McDonagh, S., Davidson, A., Rutherford, M., Hajnal, J.V., Rueckert, D., Glocker, B., Kainz, B., 2018. 3D reconstruction in canonical co-ordinate space from arbitrarily oriented 2D images. *IEEE Transactions on Medical Imaging* 37, 1737–1750.
- 750
- Huang, W., Bridge, C.P., Noble, J.A., Zisserman, A., 2017. Temporal HeartNet: Towards human-level automatic analysis of fetal cardiac screening video, in: MICCAI, Springer. pp. 341–349.
- Jiang, S., Xue, H., Glover, A., Rutherford, M., Rueckert, D., Hajnal, J.V., 2007. MRI of moving subjects using multislice snapshot images with volume
755 reconstruction (SVR): Application to fetal, neonatal, and adult brain studies. *IEEE Transactions on Medical Imaging* 26, 967–980.
- Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization, in: ICLR.

- Kuklisova-Murgasova, M., Quaghebeur, G., Rutherford, M.A., Hajnal, J.V.,
760 Schnabel, J.A., 2012. Reconstruction of fetal brain MRI with intensity match-
ing and complete outlier removal. *Medical Image Analysis* 16, 1550–1564.
- Li, Y., Cerrolaza, J.J., Sinclair, M., Hou, B., Alansary, A., Khanal, B., Matthew,
J., Kainz, B., Rueckert, D., 2018. Standard plane localisation in 3D fetal
ultrasound using network with geometric and image loss, in: *MIDL*.
- 765 Loughna, P., Chitty, L., Evans, T., Chudleigh, T., 2009. Fetal size and dating:
Charts recommended for clinical obstetric practice. *Ultrasound* 17, 160–166.
- Malinger, G., Lev, D., Kidron, D., Heredia, F., HersHKovitz, R., Lerman-Sagie,
T., 2005. Differential diagnosis in fetuses with absent septum pellucidum.
Ultrasound in Obstetrics and Gynecology: The Official Journal of the Inter-
770 *national Society of Ultrasound in Obstetrics and Gynecology* 25, 42–49.
- Moser, F., Huang, R., Papageorgiou, A.T., Papiez, B.W., Namburete, A.I.,
2019. Automated fetal brain extraction from clinical ultrasound volumes
using 3D convolutional neural networks, in: *MIUA*, Springer.
- Namburete, A.I., Xie, W., Noble, J.A., 2017. Robust regression of brain matu-
775 ration from 3d fetal neurosonography using crns, in: *Fetal, Infant and Oph-*
thalmic Medical Image Analysis. Springer, pp. 73–80.
- 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. *Medical Image Analysis* 46, 1–14.
- 780 Ni, D., Li, T., Yang, X., Qin, J., Li, S., Chin, C.T., Ouyang, S., Wang, T.,
Chen, S., 2013. Selective search and sequential detection for standard plane
localization in ultrasound, in: *MICCAI Workshop on Computational and*
Clinical Challenges in Abdominal Imaging, Springer. pp. 203–211.
- Paladini, D., Malinger, G., Monteagudo, A., Pilu, G., Timor-Tritsch, I., Toi, A.,
785 2007. Sonographic examination of the fetal central nervous system: guide-

lines for performing the 'basic examination' and the 'fetal neurosonogram'.
Ultrasound in Obstetrics and Gynecology 29, 109–116.

Papageorgiou, A.T., Ohuma, E.O., Altman, D.G., Todros, T., Ismail, L.C.,
Lambert, A., Jaffer, Y.A., Bertino, E., Gravett, M.G., Purwar, M., 2014.
790 International standards for fetal growth based on serial ultrasound measure-
ments: the fetal growth longitudinal study of the INTERGROWTH-21st
project. The Lancet 384, 869–879.

Pistorius, L., Stoutenbeek, P., Groenendaal, F., De Vries, L., Manten, G., Mul-
der, E., Visser, G., 2010. Grade and symmetry of normal fetal cortical develop-
795 ment: a longitudinal two-and three-dimensional ultrasound study. Ultrasound
in Obstetrics & Gynecology 36, 700–708.

Public Health England, 2018. NHS Fetal Anomaly Screening Programme Hand-
book. Guidance. Public Health England.

Rousseau, F., Glenn, O.A., Iordanova, B., Rodriguez-Carranza, C., Vigneron,
800 D.B., Barkovich, J.A., Studholme, C., 2006. Registration-based approach for
reconstruction of high-resolution in utero fetal MR brain images. Academic
Radiology 13, 1072–1081.

Ryou, H., Yaqub, M., Cavallaro, A., Roseman, F., Papageorgiou, A., Noble,
J.A., 2016. Automated 3D ultrasound biometry planes extraction for first
805 trimester fetal assessment, in: MLMI, Springer. pp. 196–204.

Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia,
P., Lillicrap, T., 2017. A simple neural network module for relational reason-
ing, in: NIPS, pp. 4967–4976.

Schlemper, J., Oktay, O., Chen, L., Matthew, J., Knight, C., Kainz, B., Glocker,
810 B., Rueckert, D., 2018. Attention-gated networks for improving ultrasound
scan plane detection, in: MIDL.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-
scale image recognition, in: ICLR.

- 815 The INTERBIO-21st Consortium, 2012. INTERBIO-21st Study Protocol. Protocol. Oxford.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need, in: NIPS, pp. 5998–6008.
- 820 Winter, T.C., Kennedy, A.M., Byrne, J., Woodward, P.J., 2010. The cavum septi pellucidi: why is it important? Journal of Ultrasound in Medicine 29, 427–444.
- Xie, W., Shen, L., Zisserman, A., 2018. Comparator networks, in: Proc. ECCV, pp. 782–797.
- 825 Yang, X., Ni, D., Qin, J., Li, S., Wang, T., Chen, S., Heng, P.A., 2014. Standard plane localization in ultrasound by radial component, in: ISBI, IEEE. pp. 1180–1183.
- Yoo, J.C., Han, T.H., 2009. Fast normalized cross-correlation. Circuits, Systems and Signal Processing 28, 819.
- 830 Zhang, L., Chen, S., Chin, C.T., Wang, T., Li, S., 2012. Intelligent scanning: Automated standard plane selection and biometric measurement of early gestational sac in routine ultrasound examination. Medical Physics 39, 5015–5027.

7. Supplementary Materials

7.1. Network Architecture

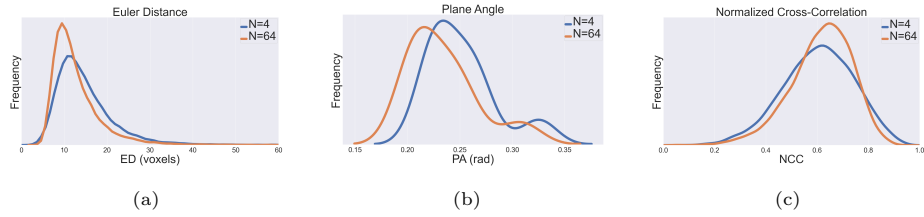
Module	Baseline Model	Proposed Model	Output Size
Feature Extraction	Input Layer		$N \times 160 \times 160 \times 1$
	conv, 3×3 , 64 conv, 3×3 , 64 max pool, 2×2 , stride 2		$N \times 80 \times 80 \times 64$
	conv, 3×3 , 128 conv, 3×3 , 128 max pool, 2×2 , stride 2		$N \times 40 \times 40 \times 128$
	conv, 3×3 , 256 conv, 3×3 , 256 conv, 3×3 , 256 max pool, 2×2 , stride 2		$N \times 20 \times 20 \times 256$
	conv, 3×3 , 512 conv, 3×3 , 512 conv, 3×3 , 512 max pool, 2×2 , stride 2		$N \times 10 \times 10 \times 512$
	conv, 3×3 , 512 conv, 3×3 , 512 conv, 3×3 , 512 max pool, 2×2 , stride 2		$N \times 5 \times 5 \times 512$
	Fully Connected Layer		$N \times 512$ $(\{\mathbf{v}_i\}_{i=1}^N)$
Comparison	-	Pairwise Feature Concatenation	$N \times N \times 1024$ $\{\mathbf{m}_{ij}\}_{i=1,j=1}^{N,N}$
	-	Fully Connected Layer	$N \times N \times 512$ $\{\mathbf{c}_{ij}\}_{i=1,j=1}^{N,N}$
Attention	-	Embedding Networks $\times 2$ (i.e. Fully Connected Layers)	$N \times 256$ $(q(\mathbf{v}_i; \theta_{attI}))$ $N \times 256$ $(k(\mathbf{v}_j; \theta_{attII}))$
	-	Dot Product	$N \times N$ (\mathbf{A})
Prediction	-	Weighted Average	$N \times 512$ $(\{\mathbf{s}_i\}_{i=1}^N)$
	Fully Connected Layer		$N \times 9$ (Resize to $N \times 3 \times 3$, $\{\mathbf{p}_i\}_{i=1}^N$)

Table 3: Network architectures of the baseline model and our proposed model. For the feature extraction module and the final layer of the prediction module, the baseline model and our proposed model have the same architecture, but they do not share weights (i.e. they are trained separately).

Experiment \ Dataset	INTERGROWTH (3D Volumes)	INTERGROWTH (2D TT Plane Images)	Video of Freehand Brain Scanning
Training	✓ (21 weeks & 18-22 weeks)		
Testing			
Sections 4.4.1 and 5.1.1	✓ (21 weeks)		
Sections 4.4.2 and 5.1.2	✓ (18-22 weeks)		
Sections 4.5 and 5.2	✓		
Sections 4.6 and 5.3		✓	
Sections 4.7 and 5.4			✓
Sections 4.8 and 5.5	✓ (21 weeks)		

Table 4: Summary of different experiments and the corresponding dataset.

7.3. Result Distribution

Figure 12: The result distribution of our proposed model. Result distribution of (a) Euclidean distance, (b) plane angle and (c) normalized cross correlation at 21 gestational weeks with $N = 4$ and $N = 64$ are shown.

7.4. Score Distribution



Figure 13: Manual score distribution of prediction on real 2D ultrasound images. The score ranges from 1 to 5, with full score being 5.