

## Article

# Single-Input Multi-Output U-Net for Automated 2D Foetal Brain Segmentation of MR Images

Andrik Rampun <sup>1</sup>, Deborah Jarvis <sup>1</sup>, Paul D. Griffiths <sup>1</sup>, Reyer Zwiggelaar <sup>2</sup> , Bryan W. Scotney <sup>3</sup>  
and Paul A. Armitage <sup>1,\*</sup>

<sup>1</sup> Academic Unit of Radiology, Department of Infection, Immunity & Cardiovascular Disease, University of Sheffield, Sheffield S10 2RX, UK; y.rampun@sheffield.ac.uk (A.R.); deborah.jarvis@sheffield.ac.uk (D.J.); p.griffiths@sheffield.ac.uk (P.D.G.)

<sup>2</sup> Department of Computer Science, Aberystwyth University, Wales SY23 3DB, UK; rrrz@aber.ac.uk

<sup>3</sup> School of Computing, Ulster University, Jordanstown, County Antrim BT37 0QB, Northern Ireland, UK; bw.scotney@ulster.ac.uk

\* Correspondence: p.armitage@sheffield.ac.uk

**Abstract:** In this work, we develop the Single-Input Multi-Output U-Net (SIMOU-Net), a hybrid network for foetal brain segmentation inspired by the original U-Net fused with the holistically nested edge detection (HED) network. The SIMOU-Net is similar to the original U-Net but it has a deeper architecture and takes account of the features extracted from each side output. It acts similar to an ensemble neural network, however, instead of averaging the outputs from several independently trained models, which is computationally expensive, our approach combines outputs from a single network to reduce the variance of predications and generalization errors. Experimental results using 200 normal foetal brains consisting of over 11,500 2D images produced Dice and Jaccard coefficients of  $94.2 \pm 5.9\%$  and  $88.7 \pm 6.9\%$ , respectively. We further tested the proposed network on 54 abnormal cases (over 3500 images) and achieved Dice and Jaccard coefficients of  $91.2 \pm 6.8\%$  and  $85.7 \pm 6.6\%$ , respectively.

**Keywords:** foetal brain segmentation; MRI; U-Net; HED network; deep learning; convolutional neural network



**Citation:** Rampun, A.; Jarvis, D.; Griffiths, P.D.; Zwiggelaar, R.; Scotney, B.W.; Armitage, P.A. Single-Input Multi-Output U-Net for Automated 2D Foetal Brain Segmentation of MR Images. *J. Imaging* **2021**, *7*, 200. <https://doi.org/10.3390/jimaging7100200>

Academic Editors:  
Raimondo Schettini and Jussi Tohka

Received: 30 June 2021  
Accepted: 26 September 2021  
Published: 1 October 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The study of brain development in utero is an important research field for early diagnosis and intervention during pregnancy. While ultrasound scanning is well-established for routine foetal screening, magnetic resonance imaging (MRI) is performed clinically in cases where a suspicious abnormality is detected during ultrasound scanning and has been shown to improve diagnostic accuracy and confidence [1]. It is now possible to acquire 3D volumetric in utero MR imaging (iuMRI) of the fetal brain [2]. Automated segmentation of the iuMRI data is highly desirable as it can provide accurate and timely measurements of morphological brain structures (e.g., volume) which can be used for monitoring and characterising fetal brain development. According to a recent review conducted in 2018 by Makropoulos et al. [3], limited studies have attempted to extract the foetal brain using deep learning approaches. This might be due to a lack of annotated foetal brain data. In comparison to adult studies, foetal brain segmentation is a challenging task as the boundary of the foetus is less visible, particularly towards the bottom and top slices, which can lead to under-segmentation. On the other hand, in more central slices, the foetal and skull boundaries might be overlapping, resulting in over-segmentation. Furthermore, in foetal MRI, the brain often occupies a relatively small portion of the imaging field of view (FOV) and its location and orientation within the FOV can be highly variable.

In this study, we propose SIMOU-Net, a hybrid network inspired from the U-Net [4] and fused with the holistically nested edge detection (HED) [5] architectures. Single-Input Multi-Output U-Net (SIMOU-Net) inherits the advantages of both U-Net (combines

location information (down-sampling) and contextual information (up-sampling)) and HED (combines multi-scale and multi-level contextual information via a fusion procedure) networks. In this study, multi-scale information refers to the representation of global and local features at different convolutional layers within the network architecture. In contrast, multi-level information refers to the representation of global and local features at different hierarchies within the network architecture. Our motivations are three-fold:

- (1) The U-Net architecture combines both location and contextual information but does not combine information collectively (e.g., making a prediction from different prediction maps). As a result, it relies on a single prediction map derived from the previous layers. In machine learning, many studies [6–8] have shown that ensemble learning reduces the variance of predictions and reduces generalisation error, often producing better results.
- (2) By contrast, the HED network does combine local and global information collectively (e.g., features from different side outputs are fused) but it does not store spatial and contextual information together, which is essential in semantic segmentation. As a result, from our own experience, the HED network is less accurate in cases where the boundary of the foetal brain is less visible.
- (3) The majority of the ensemble-based networks in the literature are (i) multi-stream architectures where several independent networks are placed into a single architecture to produce a single prediction map, or (ii) involve separate training of several independent networks to produce different prediction maps, that are averaged to produce a final prediction. SIMOU-Net bypasses the need for training several different networks independently (to combine the predictions).

Our study has the following contributions:

- (1) We propose SIMOU-Net, an ensemble deep U-Net inspired from the U-Net and HED architectures which combines multi-scale and multi-level local and global information collectively together with spatial information.
- (2) To the best of our knowledge, this is the largest cross-validated foetal brain segmentation study in the literature covering 254 cases (normal and abnormal) with over 15,000 images in total.
- (3) To further evaluate the robustness and generalisability of the proposed method, we train our network on normal cases and test it on abnormal cases, which can be extremely challenging due to differences in shape/geometry information. To the best of our knowledge, it is the first study in the literature to have performed this kind of investigation/evaluation.
- (4) Most of the fetal brain segmentation studies in the literature evaluate their results on a 2D slice-by-slice basis. From a neuro-radiology point of view, volumetric information is one of the key elements in abnormality detection. Thus, we further evaluate our method by comparing the fetal brain volumes generated automatically with those obtained manually both for normal and abnormal cases.

## 2. Related Work

In foetal brain segmentation, the majority of existing studies are based on conventional image processing (e.g., level set, superpixel and region growing) and machine learning techniques (e.g., Support Vector Machine or ensemble classifiers). In an early study, Anquez et al. [9] developed a method that relies on detecting the foetus' eyes. In a similar approach, Taleb et al. [10] proposed a method based on age-dependent template/atlas matching. The method defines a region of interest (ROI) using the intersection of all scans and then registers this ROI to an age-dependent foetal brain template. This is followed by refining the initial segmentation to a final segmentation map using a fusion step. The disadvantage of these early methods is the dependency on prior knowledge which can be unreliable. To overcome the requirement for prior knowledge, Alansary et al. [11] proposed a method using the simple linear iterative clustering (SLIC) [12] technique to extract all superpixels in an image followed by feature extraction in each superpixel. Subsequently, su-

perpixels with similar features were merged and the random forest classifier was employed to differentiate brain and non-brain regions. They reported a Dice score of 73.6% based on 55 cases when compared to expert manual segmentations. A semi-automatic method was proposed by Link et al. [13], where a user has to initialise the method by drawing a circle around the brain. This is followed by a region growing technique to estimate the brain region. Their experimental results reported a correlation value of  $r^2 = 0.91$  for the brain volume when compared with manual segmentation based on 199 fetal brains of normal cases. Attallah et al. [14] presented a method that uses an elaborate combination of adaptive thresholding, morphological and watershed operations in order to segment the fetal brain. The data used appears to be single slice data from three central orthogonal planes rather than 3D volumetric data. They report encouraging Dice scores of 0.975 for 50 of the 227 scans, but do not report data from the remaining 177 scans, nor do they report how the ground truth data for comparison was obtained. This work was primarily concerned with the classification of foetal brain abnormalities, with the segmentation being a step in that process. From our own experience, intensity-based methods such as the region growing and SLIC often fail to segment the foetal brain region toward the top and bottom slices due to obscure boundaries and the brain having a similar intensity to surrounding tissues.

To overcome the disadvantages of intensity-based methods, Ison et al. [15] proposed a texture-based method. Firstly, the random forest (RF) classifier with 3D Haar-like descriptors was used to roughly estimate the location of the foetus' head. The probability map segmentation is refined by taking centroids with high probabilities and optimised using a Markov random field (MRK) appearance model. The method was tested on 43 foetal brains and produced 82% and 89% sensitivity and specificity, respectively. Keraudren et al. [16] proposed a similar pipeline with the following steps but used the scale-invariant feature transform (SIFT) features to represent the brain region and the conditional random field (CRF) was employed to estimate the brain boundary. The authors evaluate the performance of their method on 66 foetal brains and report a Dice score of 93%. In contrast, Kainz et al. [17] developed a method that bypasses the head detection step by directly estimating the fetal brain via an ensemble forests classifier using spherical 3D Gabor features. The resulting probability map was further processed via thresholding and a 2D level-set was employed to perform the segmentation. The authors tested the proposed method on 50 cases and reported a Dice score of 90%.

This summary makes clear that the state-of-the-art in foetal brain extraction methods to date is based on a combination of several traditional image processing techniques such as image registration and level set/graph cuts. Each of these techniques has its own parameters to be set and optimised, which ultimately make foetal brain extraction difficult and semi-automatic or supervised rather than fully automatic. More specifically, we identify the following disadvantages:

- (1) The template-based methods rely heavily on the alignment of the query images to atlases or make a strong assumption about the orientation and geometry information [18]. As a result, when the foetal brain is structurally abnormal, this approach often produces inaccurate results due to differences in geometry/shape information.
- (2) Most of the methods developed require initial detection of the eyes or head in order to aid localisation of the brain. However, in many images, the eyes are not visible and the skull can be obscured particularly towards the bottom and top of the brain. Furthermore, since this initial step is necessary for subsequent steps, it is essential for the first step to achieve 100% localisation accuracy which is often difficult, or time consuming in the case of manual region placement.
- (3) The processing steps that generate initial candidates typically result in many false positives. Therefore, in subsequent steps, it is essential to have a robust post-processing method for false-positive reduction.
- (4) The 2D level-set and region growing methods often perform poorly when the brain boundary is unclear or obscured. Unfortunately, this is a frequent occurrence in foetal MRI, especially in the bottom and top slices. While the methods might work

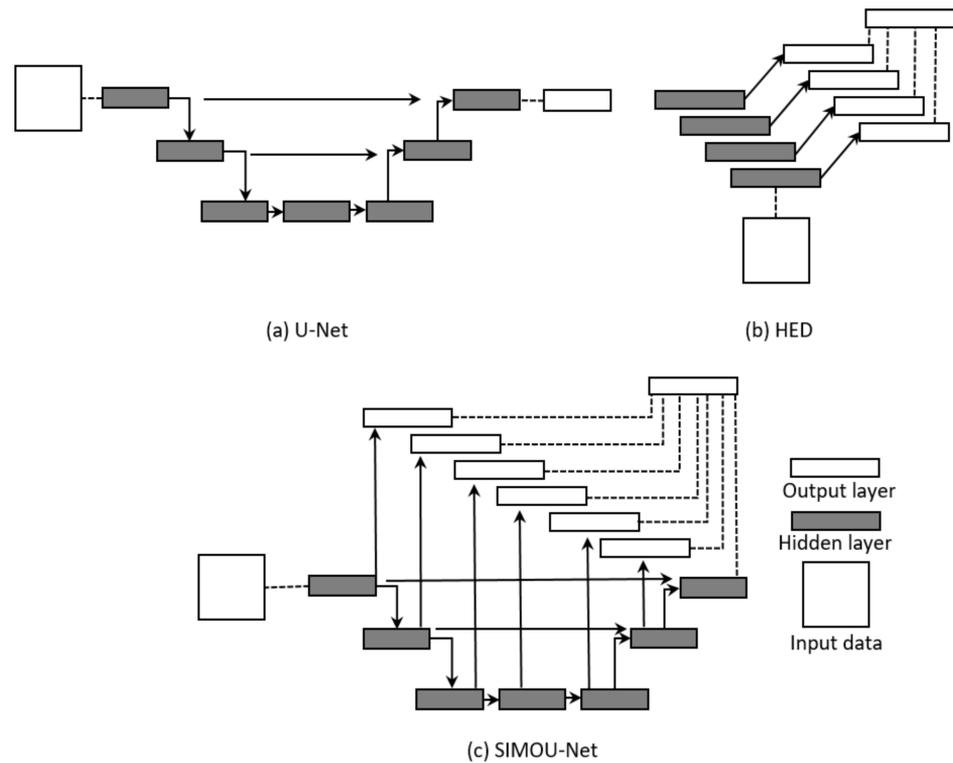
successfully on the central slices, in some cases the brain and the skull boundaries can be overlapping, resulting in over-segmentation. Furthermore, these methods do not generally work well with low contrast images.

The recent promising results in deep learning-based image segmentation motivate further development of learning-based, geometry-independent and registration-free foetal brain segmentation [18]. Deep learning (e.g., convolutional neural network (CNN)) is becoming a methodology of choice in brain MRI segmentation [19–21]. However, in foetal brain segmentation, the number of studies in the literature is rather small. Recently, Salehi et al. [22] proposed a three-pathway automatic foetal brain extraction based on a modified U-Net [4] using auto-context CNN in which intrinsic local and global image features are learned through 2D patches of different window sizes. The authors reported a Dice score of 95.97% based on 75 normal cases. The main disadvantage is that it requires three separate models that need to be run independently to segment the foetal brain (e.g., from coronal, sagittal and axial views). Furthermore, it is essential to obtain manual annotation for each slice from all planes which is highly time consuming even for an expert annotator. In another study, Rajchl et al. [23] extended the GrabCut [24] method to include machine learning by training a neural network classifier from bounding box annotations (which is called DeepCut). A simple variant of the DeepCut method was proposed and compared with a naïve approach to CNN training under weak supervision. The energy function of the densely connected conditional random field (CRF) was used to iteratively update the training targets to achieve semantic segmentation. The experimental evaluation produced Dice scores of 90.3% and 94.1% for the extended GrabCut and DeepCut methods, respectively. Although the results are promising, experimental results suggest that the CRF parameters are inconsistent and need to be tuned manually to maximise their performance. Furthermore, since the training is interrupted after a fixed number of epochs to update class labels, this process makes training very slow [23]. In a more straightforward approach, Khalili et al. [25] trained their network directly (image-to-image prediction) from the whole image. They proposed a modified U-Net architecture using more robust functions to avoid overfitting and reduce the number of features extracted in each convolutional layer. Based on 12 normal cases, they reported a Dice score of 88%.

Ebner et al. [26] proposed a method based on the P-Net network architecture which consists of three stages: (1) a coarse segmentation for brain localisation, (2) fine segmentation with a multi-scale loss function, and (3) brain reconstruction. They reported an average dice score of 93.4% based on two different datasets. Recently, Lou et al. [27] proposed a multi-stage deep supervision U-Net (DS-U-Net). Their method also contained three stages: (1) a coarse segmentation using the original DS U-Net introduced by Dou et al. [28] to define a 3D bounding box to find the location of the brain, (2) using another DS U-Net [28] to focus on the extracted region to produce finer segmentation and (3) refining the final segmentation by using another DS-U-Net but training it using a local image patch strategy. The authors reported a dice score of 91.68% based on 43 cases. The main disadvantage of these methods [26,27] is that they require three training phases which is very time consuming.

### 3. Materials and Methods

The SIMOU-Net architecture comprises a single stream deep network with multiple side outputs which promotes an ensemble learning scheme. Figure 1 shows the semantic architecture of the U-Net, HED and SIMOU-Net architectures.



**Figure 1.** A high-level view of the U-Net, HED and SIMOU-Net architectures.

### 3.1. Network Architecture

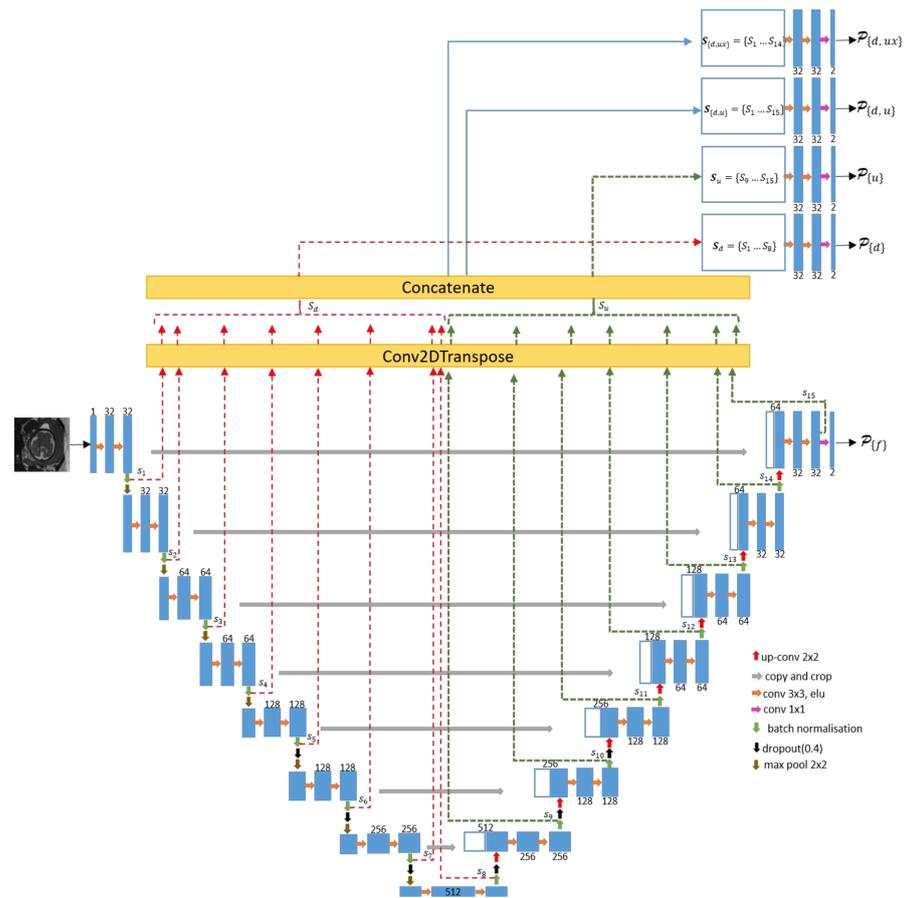
Figure 2 shows details of the SIMOU-Net architecture. The lower part of the network is similar to the original U-Net architecture except it has eight levels of convolutional blocks instead of five. Our deeper architecture enables us to extract finer texture features. The left side of the network consists of repeated application of two  $3 \times 3$  unpadded convolutions followed by an exponential linear unit (‘elu’). Each level of convolutional block is followed by batch normalisation and a  $2 \times 2$  max pooling operation with stride 2 for downsampling.

At each downsampling step, there is one side output ( $s_n$ , where  $n = \{1, 2, \dots, 15\}$ , there are 15 side outputs in total) and we use the following number of feature channels: 32, 32, 64, 64, 128, 128, 256, 512. A dropout operation is applied in the last three layers of the downsampling path and the first three layers of the upsampling path. This is to avoid overfitting due to the number of features in this section of the network being very dense. For the right side (expansive path) of the network, each step consists of upsampling of the feature map from the contracting path followed by a  $2 \times 2$  convolution. To optimise the expansive process, instead of using the conventional upsampling operation (e.g., nearest neighbour interpolation) we use the 2D transpose convolution to fill in details, as well as upsampling the image during the model training process. The concatenation operation with the corresponding cropped feature map from the contracting path and two  $3 \times 3$  convolutions are also applied, each followed by an ‘elu’. A similar process occurs in the contracting path, where each upsampling step has one side output. For the upper part of the network, it consists of a 2D transpose convolution for each side output. In total, the SIMOU-Net contains 15 side outputs ( $s_1, \dots, s_{15}$ );  $s_n$  from the left and right paths of the network are concatenated separately using equation 1. Suppose each side output is an  $M$  (row)  $\times N$  (column) matrix and  $i$  and  $j$  represent each element’s location in  $s_n$ . We concatenate all side outputs (see Figure 2 for more details) from downsampling and upsampling paths ( $\mathbf{S}_{\{d,u\}} = \{s_1, \dots, s_{15}\}$ ), without  $s_{15}$  ( $\mathbf{S}_{\{d,ux\}} = \{s_1, \dots, s_{14}\}$ ), downsampling path only

$(\mathbf{S}_{\{d\}} = \{s_1, \dots, s_8\})$  and upsampling path only  $(\mathbf{S}_{\{u\}} = \{s_9, \dots, s_{15}\})$  by adding each  $s_n$  in the third dimension. Hence, the resulting concatenated matrix can be represented as

$$\mathbf{S}_c = \begin{bmatrix} s_n(i, j, p) & \dots & s_n(i, j + N, p) \\ s_n(i + 1, j, p) & \dots & \vdots \\ \vdots & \ddots & \vdots \\ s_n(i + K, j, p) & \dots & s_n(i + K, j + N, p) \end{bmatrix}, \tag{1}$$

where  $c \in \{(d, u), (d, ux), d, u\}$ . Note the additional dimension of  $p$  of  $\mathbf{S}_c$  leading to  $M \times N \times P$ . In Keras, this can be implemented using the concatenate operation by setting axis = 3. After concatenation, a couple of  $3 \times 3$  convolution operations with ‘elu’ are applied to each  $\mathbf{S}_c$  (e.g.,  $\mathbf{S}_{\{d,u\}}$ ). At the final layer, we use a  $1 \times 1$  convolution to map each 64-component feature vector to two classes (e.g., brain and non-brain) and in total, the SIMOU-Net architecture produces five output prediction maps. Four of them,  $(P_{\{d,ux\}}, P_{\{d,u\}}, P_{\{u\}}, P_{\{d\}})$  are produced from the upper part of the network and one ( $P_{\{f\}}$ ) is produced by the deep U-Net (see Figure 2). The  $P_{\{f\}}$  map is also known as the main output of deep U-Net,  $P_{\{d,u\}}, P_{\{u\}}$  and  $P_{\{d\}}$  are known as side outputs and  $P_{\{d,ux\}}$  is known as the combined output of the side outputs and main output.



**Figure 2.** The SIMOU-Net architecture. Prediction maps for each output are denoted as  $P_{\{i\}}$ . Note that in this architecture (and also in the classical U-Net) the downsampling path extracts global information and the upsampling path extracts local information. The ‘Conv2DTranspose’ is also known as deconvolution, which is similar to ‘UpSampling2D’. However, the main difference is that it both performs the upsample operation and interprets the coarse input data to fill in detail while it is upsampling. The concatenation operation merges together the side outputs  $s_d$  and  $s_u$  using Equation (1), enabling production of the four prediction maps  $P_{\{d\}}, P_{\{u\}}, P_{\{d,u\}}$  and  $P_{\{d,ux\}}$ .

### 3.2. Training, Validation and Testing

We denote our input training dataset as  $D_{tr} = \{(I_q, T_q), q = 1, \dots, Q\}$ , where sample  $I = \{x_r^{(q)}, r = 1, \dots, |I_q|\}$  denotes the raw input image and  $T_q = \{y_r^{(q)}, r = 1, \dots, |I_q|\}$ ,  $y_r^{(n)} \in \{0, 1\}$ , denotes the corresponding ground truth binary map for image  $I_q$ . We use the same formulation for our validation data set  $D_{vd} = \{(I_q, T_q), q = 1, \dots, Q\}$ . For simplicity, we denote the collection of all standard network layer parameters as  $\mathbf{W}$ . In the SIMOU-Net architecture, we have five outputs and each output is associated with a classifier in which the corresponding weights are denoted as  $w = w^{(1)}, \dots, w^{(T)}$ . We consider the following objective function

$$F_{Sp}(W, w) = \sum_{t=1}^T \alpha_t L_{Sp}^{(t)}(W, w^{(t)}), \tag{2}$$

where  $L_{Sp}$  denotes the image-level loss function for the network's outputs and  $\alpha$  is the weight. Following the study in [5], the loss function is computed over all pixels in a training image  $I = \{x_r^{(q)}, r = 1, \dots, |I_q|\}$  and its associated ground truth  $T_q = \{y_r^{(q)}, r = 1, \dots, |I_q|\}$ ,  $y_r^{(n)} \in \{0, 1\}$ . Since the brain region often occupies a relatively small portion of the FOV, we employed the loss function proposed by [29], a combination of binary cross-entropy and Dice coefficient, to offset the imbalance of pixels between foreground and background, which is described as:

$$L_{Sp}(I_b, T_b) = -\frac{1}{B} \sum_{b=1}^B \left( \frac{1}{2} I_b \log T_b + \frac{2I_b T_b}{I_b + T_b} \right), \tag{3}$$

where  $I_b$  and  $T_b$  are the flatten predicted probabilities and the flatten ground truth of the  $b^{th}$  image, respectively, and  $B$  indicates the batch size. We train our network with the root mean square propagation (RMSprop) [30] implementation in Keras with a Tensorflow backend. The initial learning rate ( $lr$ ) and gradient moving average decay factor ( $\rho$ ) are 0.0003 and 0.8, respectively. To minimise training time (e.g., faster to converge) and to have a better estimation of the gradient of the full dataset (hence it converges faster), we favour a maximum batch size ( $bs = 32$ ) that our machine can accommodate. The number of iterations used per epoch ( $E$ ) is based on the number of samples divided by batch size ( $\frac{\#Q}{bs}$ ). We monitor the Dice and Jaccard coefficients, set  $E = 1000$  and employ the 'EarlyStopping' mechanism on the validation set to stop the training automatically when the loss function value does not decrease after 100 epochs.

In our proposed network, with many convolutional layers and different paths within the network, a good initialization of the weights is essential. Otherwise, parts of the network might give excessive activations, while other parts never contribute. For this purpose, we follow the weight initialization implemented by Ronneberger et al. [4] using the Gaussian distribution with a standard deviation of  $\sqrt{2/\eta}$  where  $\eta$  is the number of incoming nodes of one neuron. This strategy was chosen due to the architecture of our network (alternating convolution and ReLU/ELU layers) and was further suggested in the study of [31].

During testing, given an image  $I$ , we obtain five prediction maps which are  $P_{\{d,ux\}}$ ,  $P_{\{d,u\}}$ ,  $P_{\{u\}}$ ,  $P_{\{d\}}$  and  $P_{\{f\}}$ . Subsequently, ensemble outputs can be obtained by further aggregating these prediction maps using the following equation

$$P_{arg} = arg (P_{\{d,u\}}, P_{\{d,ux\}}, P_{\{u\}}, P_{\{d\}}, P_{\{f\}}) \tag{4}$$

where  $arg$  indicates the different operations used, namely 'mean' ( $P_{AVG}$ ), 'median' ( $P_{MED}$ ) and 'maximum' ( $P_{MAX}$ ).

### 3.3. Data Description, Pre-Processing and Experimental Setup

More information on the datasets used in this study can be found in [1,2,32]. Our first dataset consisted of MR images of 200 healthy/normal fetuses with gestational age (GA) ranging between 18 and 37 weeks. The second dataset consisted of MR images of 54 abnormal fetuses with GA between 20 and 34 weeks. All studies were performed on a 1.5T whole-body scanner with images acquired using a 3D FIESTA sequence. Slice thickness varies according to gestational age/size of foetal head (range between 2.0 mm and 2.6 mm). Some of the images contain a degree of motion artefacts that are typical for such acquisitions. For each image, the brain was manually segmented by a radiographer with over 10 years' experience in foetal MRI (who had been trained by an expert neuro-radiologist) using the '3D Slicer' software. The freehand segmentation was performed on axial images with the sagittal and coronal planes used as additional references to improve accuracy. Each case took approximately two to three hours to segment every slice depending on the complexity of the brain and clarity of the image. These manual segmentations were considered to be the 'gold standard' for comparison purposes throughout the rest of this study.

To further enrich the training dataset in this study we applied the following data augmentation, where each input image has a corresponding "augmented" image assigned with a: (i) random rotation range of up to  $180^\circ$ , (ii) zooming in and out with a range of 0.1 to 2.0, and (iii) horizontal and vertical flips. Training and testing were performed on a high-performance computing server with Nvidia DGX-1 graphics cards 8x Tesla P100 GPUs (16GB RAM each) (Santa Clara, CA, USA), dual 20-core Intel Xeon E5-2698 v4 2.2 GHz and 512 GB system RAM (Santa Clara, CA, USA), and all models were implemented in Keras with a Tensorflow backend.

## 4. Results and Discussion

### 4.1. Normal Cases

The following metrics are used to quantitatively evaluate the performance of the method: Jaccard ( $J$ ), Dice ( $D$ ), Sensitivity ( $Sen$ ), Hausdorff distance ( $HD$ ) measured in mm, and mean absolute volume differences ( $MAVD$ ) measured in  $\text{cm}^3$ .  $TP$  (cyan),  $TN$  (black),  $FP$  (red) and  $FN$  (yellow) are true positive, true negative, false positive and false negative, respectively used throughout the paper to illustrate segmentation results.

Table 1 shows the average results using the proposed method based on a stratified 3-fold cross validation. We include in the same group (e.g., training set) all slices/images of the same fetus. We compare our results with the original U-Net [4], wide U-Net (wU-Net) [29], nested U-Net with (U-Net++), without deep supervision (U-Net+++ [29], HED [5], SegNet [33], Att-U-Net [34], DS-U-Net [27], P-Net(S) [26] and a modified U-Net version of [18]. All experiments have the same implementation details, such as the number of epochs, function loss, batch size, etc.

Table 1 shows that the SIMOU-Net produced competitive results across the different metrics. Friedman tests indicated significant differences between methods for all metrics;  $J$  ( $p < 0.001$ ),  $D$  ( $p < 0.001$ ),  $Sen$  ( $p < 0.01$ ),  $MAVD$  ( $p < 0.001$ ) and  $HD$  ( $p = 0.002$ ). The Nemenyi post hoc test was performed and significant differences with respect to the  $P_{AVG}$  prediction map score are indicated in Table 1. It significantly outperforms all other methods based on the  $Sen$  metric and all but the DS-U-Net based on Jaccard and Dice metrics. The  $MAVD$  and  $HD$  metrics were less sensitive to the generally small differences observed between methods, with only the wU-Net ( $p < 0.001$ ) and HED ( $p = 0.029$ ) performing significantly worse than the SIMOU-Net  $P_{AVG}$  prediction map based on  $MAVD$  and only the HED network performed worse with the  $HD$  metric ( $p = 0.001$ ). The  $P_{\{d\}}$  prediction map performed poorly, and this is expected because features are extracted from only the side outputs of the downsampling path (global features) of the network architecture without taking into account local features from the upsampling path. The SIMOU-Net produced the highest  $J = 88.9\%$ ,  $D = 94.2\%$ ,  $Sen = 97.5\%$  and lowest  $MAVD = 4.2 \text{ cm}^3$  and  $HD = 7.5 \text{ mm}$  when all prediction maps are aggregated using the averaging procedure ( $P_{AVG}$ ). Many classification and segmentation deep learning studies [6–8] have shown that

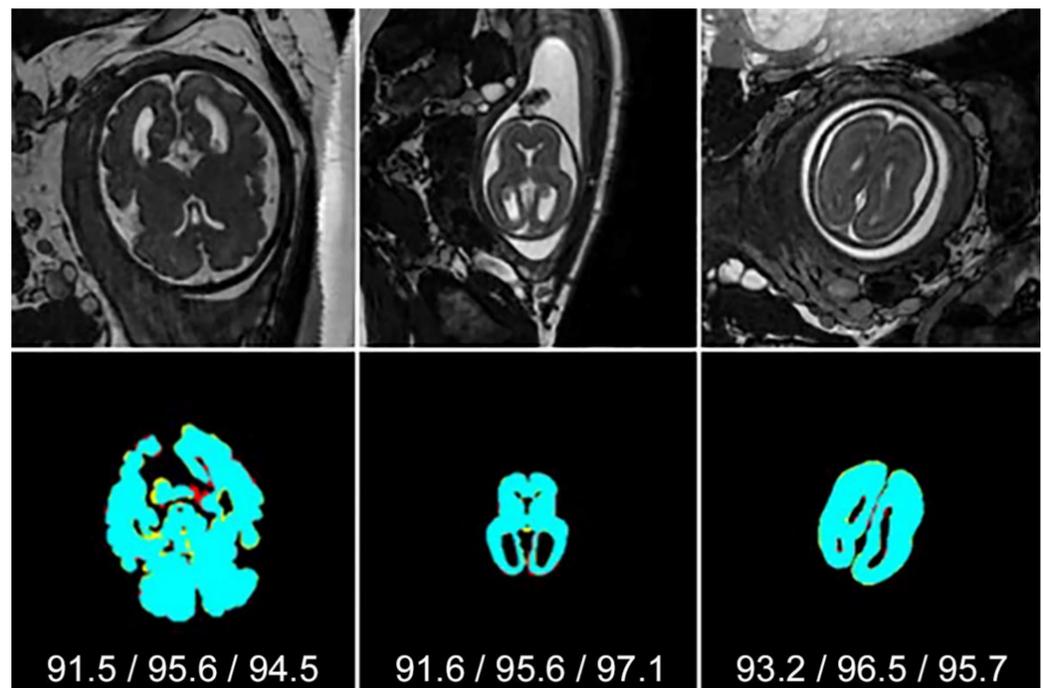
combining models/prediction maps often leads to better classification results because it reduces the variance of the predictions and generalization error. Using the ‘median’ ( $P_{MED}$ ) operation, the SIMOU-Net produced  $J = 87.2\%$ ,  $D = 93.3\%$ ,  $Sen = 95.9\%$ ,  $MAVD = 4.4 \text{ cm}^3$  and  $HD = 7.6 \text{ mm}$  which are close to those from the ‘mean averaging’ operation and only significantly worse for the  $Sen$  metric. When using the ‘maximum’ operation ( $P_{MAX}$ ), we achieved lower accuracy than  $P_{AVG}$ ,  $P_{MED}$ ,  $P_{\{d,ux\}}$ ,  $P_{\{d,u\}}$ ,  $P_{\{u\}}$  and  $P_{\{f\}}$ . This is expected because this operation tends to include outliers in the probability map.

**Table 1.** Average quantitative results for 200 normal cases. Each numerical value represents the mean value  $\pm$  standard deviation from stratified three-fold cross-validation. Bold typeface indicates the best performing method for each metric. \* indicates a significant difference with respect to the  $P_{AVG}$  score after performing a Nemenyi test.

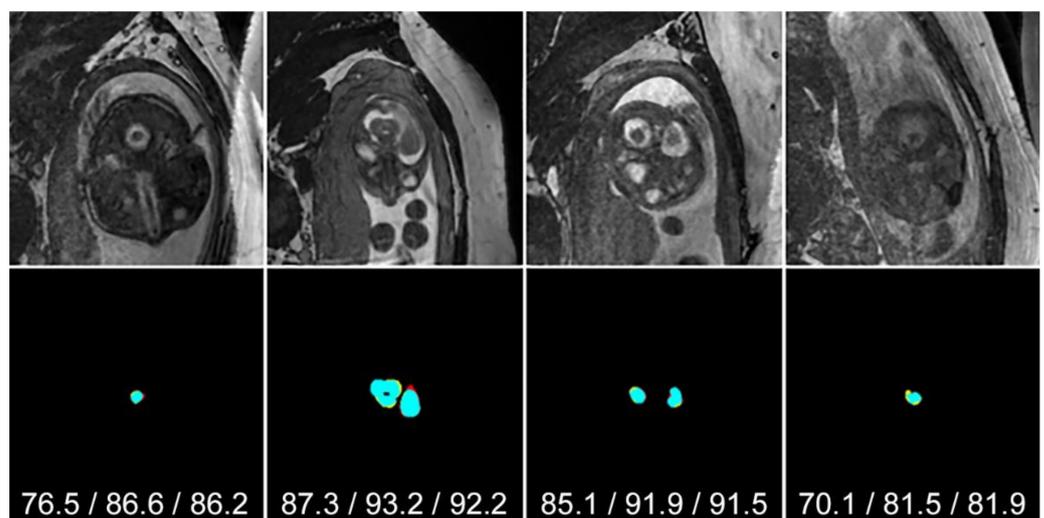
Network	$J$ (%)	$D$ (%)	$Sen$ (%)	$MAVD$ ( $\text{cm}^3$ )	$HD$ (mm)
SIMOU-Net					
$P_{\{d,ux\}}$	$86.9 \pm 7.8$	$93.8 \pm 6.2$	$95.9 \pm 6.5$ *	$4.5 \pm 2.9$	$7.7 \pm 3.7$
$P_{\{d,u\}}$	$86.8 \pm 6.9$	$93.6 \pm 6.3$	$95.7 \pm 6.5$ *	$4.5 \pm 2.9$	$7.7 \pm 3.7$
$P_{\{d\}}$	<b><math>84.3 \pm 8.2</math> *</b>	<b><math>88.1 \pm 7.3</math> *</b>	$92.1 \pm 7.9$ *	$5.1 \pm 3.5$	$7.9 \pm 3.9$
$P_{\{u\}}$	$86.9 \pm 7.8$	$93.0 \pm 6.2$	$95.9 \pm 6.5$ *	$4.5 \pm 2.9$	$7.8 \pm 3.7$
$P_{\{f\}}$	$86.8 \pm 7.4$	$93.2 \pm 6.2$	$95.8 \pm 6.4$ *	$4.5 \pm 2.9$	$7.8 \pm 3.7$
$P_{MAX}$	$86.1 \pm 7.9$ *	$91.2 \pm 6.6$ *	$95.8 \pm 6.2$ *	$4.7 \pm 3.1$	$8.1 \pm 4.1$
$P_{MED}$	$87.2 \pm 6.9$	$93.3 \pm 6.1$	$95.9 \pm 6.2$ *	$4.4 \pm 2.9$	$7.6 \pm 3.7$
$P_{AVG}$	<b><math>88.9 \pm 6.9</math></b>	<b><math>94.2 \pm 5.9</math></b>	<b><math>97.5 \pm 6.2</math></b>	<b><math>4.2 \pm 2.7</math></b>	<b><math>7.5 \pm 3.6</math></b>
U-Net	$79.2 \pm 8.1$ *	$86.8 \pm 7.4$ *	$92.4 \pm 7.9$ *	$6.3 \pm 5.7$	$8.6 \pm 4.8$
wU-Net	$74.9 \pm 10.6$ *	$83.5 \pm 9.3$ *	$92.8 \pm 7.6$ *	$7.6 \pm 6.4$ *	$8.9 \pm 4.8$
U-Net+++	$81.3 \pm 7.5$ *	$90.6 \pm 6.8$ *	$95.9 \pm 6.4$ *	$5.9 \pm 5.6$	$8.1 \pm 4.2$
U-Net++	$80.9 \pm 7.6$ *	$89.7 \pm 6.6$ *	$95.7 \pm 8.3$ *	$6.1 \pm 5.5$	$8.4 \pm 4.3$
HED	$76.9 \pm 10.9$ *	$85.6 \pm 9.9$ *	$84.7 \pm 9.9$ *	$7.2 \pm 6.4$ *	$10.1 \pm 5.6$ *
SegNet	$81.9 \pm 9.2$ *	$87.9 \pm 9.6$ *	$90.7 \pm 7.6$ *	$5.5 \pm 5.5$	$8.4 \pm 4.7$
Att-U-Net	$83.5 \pm 7.8$ *	$90.9 \pm 6.3$ *	$95.1 \pm 6.4$ *	$5.1 \pm 3.5$	$8.1 \pm 4.0$
P-Net(S)	$85.4 \pm 7.9$ *	$90.1 \pm 7.1$ *	$94.2 \pm 8.5$ *	$5.1 \pm 3.5$	$8.2 \pm 4.3$
DS-U-Net	$86.9 \pm 7.8$	$92.2 \pm 6.5$	$96.1 \pm 8.4$ *	$4.5 \pm 2.9$	$7.6 \pm 3.6$
U-Net2	$80.3 \pm 7.9$ *	$88.1 \pm 6.7$ *	$92.9 \pm 9.1$ *	$6.1 \pm 5.5$	$8.4 \pm 4.3$

Combining all side outputs ( $P_{\{d,ux\}}$ ) produced  $J = 86.9\%$ ,  $D = 93.8\%$ ,  $Sen = 95.9\%$ ,  $MAVD = 4.5 \text{ cm}^3$  and  $HD = 7.7 \text{ mm}$ . Including  $S_{15}$  ( $P_{\{d,u\}}$ ) produced very similar results. These experimental results suggest that there is useful information from the side outputs of the network, as  $P_{\{d,ux\}}$  and  $P_{\{d,u\}}$  tended to produce slightly better results than  $P_{\{f\}}$  alone, although the differences did not reach statistical significance. This might be due to the loss of information because of repetitive convolution operations in the contraction and expansion paths, resulting in the slightly worse result in  $P_{\{f\}}$ .

Figures 3 and 4 show examples of segmentation results using  $P_{AVG}$  for middle and bottom/top slices, respectively. Black, cyan, red and yellow represent  $TN$ ,  $TP$ ,  $FP$  and  $FN$ , respectively. In most cases, segmenting the brain from the bottom/top slices is far more difficult than in central slices. This is because the brain region is very small and tends to be more obscure due to low contrast or noise. Figure 3 shows that in all examples the SIMOU-Net produced both  $J > 90\%$  and  $D > 90\%$  whereas for bottom/top slices in Figure 4 the SIMOU-Net produced  $J = 70\%$  to  $J = 90\%$  for very challenging cases.



**Figure 3.** Typical examples of  $P_{AVG}$  segmentation results from middle slices. Black, cyan, red and yellow represent  $TN$ ,  $TP$ ,  $FP$  and  $FN$ , respectively. Jaccard/Dice/Sensitivity scores are shown under each segmentation.



**Figure 4.** Typical examples of  $P_{AVG}$  segmentation results from top and bottom slices. Black, cyan, red and yellow represent  $TN$ ,  $TP$ ,  $FP$  and  $FN$ , respectively. Jaccard/Dice/Sensitivity scores are shown under each segmentation.

#### 4.2. Abnormal Cases

We further evaluated the proposed network architecture by training it on 180 normal cases, validating it on 20 normal cases (in total 200 normal cases from the first dataset) and then testing the model on 54 abnormal cases. The abnormal cases consisted of the following abnormalities: agenesis of corpus callosum (ACC), mild ventriculomegaly (MV) and structurally deformed. From a radiological perspective, ACC manifests as a partial or wholly absent corpus callosum, MV refers to larger than normal cerebral ventricles and structurally deformed covers a wide range of structural brain abnormalities. From a neuro-radiology point of view, estimating the brain volume from these cases can be

difficult due to extreme and obscure brain shape. In fact, it is also difficult for an observer to annotate these cases without substantial experience in dealing with abnormal cases.

Experimental results from the abnormal cases are shown in Table 2. Friedman tests indicated significant differences between methods for  $J$  ( $p < 0.001$ ),  $D$  ( $p < 0.001$ ) and  $Sen$  ( $p < 0.001$ ), but not for  $MAVD$  ( $p = 0.73$ ), nor  $HD$  ( $p = 0.79$ ). The Nemenyi post hoc test was performed for the three metrics showing differences and significance with respect to the  $P_{AVG}$  prediction map score is indicated in Table 2. The SIMOU-Net architecture ( $P_{AVG}$ ) achieved  $J = 85.7\%$ ,  $D = 91.2\%$ ,  $Sen = 93.2\%$ ,  $MAVD = 7.1 \text{ cm}^3$  and  $HD = 8.5 \text{ mm}$ . When using the ‘median’ operation ( $P_{MED}$ ), it produced  $J = 83.7\%$ ,  $D = 89.3\%$ ,  $Sen = 92.4\%$ ,  $MAVD = 7.2 \text{ cm}^3$  and  $HD = 8.7 \text{ mm}$ , which was not significantly different from  $P_{AVG}$  for any metric. These results presented in Table 2 once again suggest that using the ‘mean’ and ‘median’ operations, or taking account of features from side outputs, tends to produce a better result than relying on a single predictive map, although the differences are small and do not generally reach significance with the exception of  $P_{\{d\}}$  and  $P_{\{u\}}$ . Some of these differences may begin to reach significance in a larger dataset, as is indicated by the data presented in Table 1 which is based on 200 cases, as opposed to Table 2 which is derived from 54 cases.

**Table 2.** Average quantitative results (mean  $\pm$  standard deviation) trained and validated on 200 normal cases and tested on 54 abnormal cases. Bold typeface indicates the best performing method for each metric. \* indicates a significant difference with respect to the  $P_{AVG}$  score after performing a Nemenyi test.

Network	$J$ (%)	$D$ (%)	$Sen$ (%)	$MAVD$ ( $\text{cm}^3$ )	$HD$ (mm)
SIMOU-Net					
$P_{\{d,ux\}}$	83.5 $\pm$ 6.8	89.2 $\pm$ 6.4	92.3 $\pm$ 5.9	7.2 $\pm$ 12.1	8.7 $\pm$ 12.4
$P_{\{d,ui\}}$	83.6 $\pm$ 6.8	89.4 $\pm$ 6.4	92.4 $\pm$ 5.9	7.2 $\pm$ 12.1	8.7 $\pm$ 12.4
$P_{\{d\}}$	75.1 $\pm$ 6.9 *	80.1 $\pm$ 6.2 *	85.3 $\pm$ 5.5 *	7.9 $\pm$ 12.7	8.9 $\pm$ 12.6
$P_{\{u\}}$	81.1 $\pm$ 6.4 *	81.2 $\pm$ 6.0 *	88.3 $\pm$ 5.3 *	7.5 $\pm$ 12.3	8.8 $\pm$ 12.5
$P_{\{f\}}$	83.4 $\pm$ 6.1	89.1 $\pm$ 6.4	92.2 $\pm$ 5.9	7.2 $\pm$ 12.1	8.7 $\pm$ 12.4
$P_{MAX}$	77.1 $\pm$ 6.3 *	84.2 $\pm$ 6.6 *	89.2 $\pm$ 6.2	7.7 $\pm$ 12.5	8.8 $\pm$ 12.6
$P_{MED}$	83.7 $\pm$ 6.6	89.3 $\pm$ 6.4	92.4 $\pm$ 7.1	7.2 $\pm$ 12.1	8.7 $\pm$ 12.4
$P_{AVG}$	<b>85.7 <math>\pm</math> 6.6</b>	<b>91.2 <math>\pm</math> 6.8</b>	<b>93.2 <math>\pm</math> 6.2</b>	<b>7.1 <math>\pm</math> 12.1</b>	<b>8.5 <math>\pm</math> 12.4</b>
U-Net	76.2 $\pm$ 8.2 *	82.8 $\pm$ 7.2 *	84.4 $\pm$ 7.1 *	7.8 $\pm$ 12.5	9.0 $\pm$ 12.9
wU-Net	71.1 $\pm$ 10.1 *	79.6 $\pm$ 9.3 *	80.8 $\pm$ 8.6 *	8.1 $\pm$ 12.7	9.3 $\pm$ 13.2
U-Net+++	76.3 $\pm$ 8.5 *	82.9 $\pm$ 7.2 *	84.7 $\pm$ 7.2 *	7.7 $\pm$ 12.5	8.8 $\pm$ 12.6
U-Net++	76.9 $\pm$ 8.4 *	83.2 $\pm$ 6.9 *	85.7 $\pm$ 7.1 *	7.7 $\pm$ 12.5	10.1 $\pm$ 12.8
HED	70.1 $\pm$ 12.7 *	75.2 $\pm$ 10.2 *	79.6 $\pm$ 9.9 *	8.2 $\pm$ 12.9	11.5 $\pm$ 13.9
SegNet	78.2 $\pm$ 10.6 *	84.4 $\pm$ 9.1 *	87.9 $\pm$ 9.6	7.7 $\pm$ 12.5	8.9 $\pm$ 12.7
Att-U-Net	79.5 $\pm$ 7.8 *	85.9 $\pm$ 7.3	87.1 $\pm$ 7.1 *	7.6 $\pm$ 12.3	8.8 $\pm$ 12.7
P-Net(S)	83.9 $\pm$ 6.9	89.7 $\pm$ 6.5	91.3 $\pm$ 6.5	7.2 $\pm$ 12.1	8.8 $\pm$ 12.4
DS-U-Net	85.5 $\pm$ 6.6	90.9 $\pm$ 6.5	92.9 $\pm$ 6.4	7.1 $\pm$ 12.1	8.5 $\pm$ 12.6
U-Net2	78.5 $\pm$ 7.9 *	84.1 $\pm$ 6.9 *	85.9 $\pm$ 7.6 *	7.7 $\pm$ 12.5	9.0 $\pm$ 12.7

In comparison to the other methods, the P-Net(S) and DS-U-Net produced very similar results to our method across different metrics, similar to what was reported in Table 1. The Att-U-Net also showed comparable performance to SIMOU-Net when using the  $D$  metric. All other networks perform significantly worse than SIMOU-Net when using the  $J$ ,  $D$  and  $Sen$  metrics. Similar to Table 1, the  $MAVD$  and  $HD$  show less sensitivity to the performance differences between networks with no significant differences found. We expected these results because the DS-U-Net architecture combines its prediction map (e.g.,  $P_{\{f\}}$ ) with several side outputs from the upsampling path. The Att-U-Net clearly improved the performance of the classical U-Net by at least 2% for the  $J$ ,  $D$  and  $Sen$  metrics.

The SegNet produced similar results to U-Net2, U-Net++ and U-Net+++. The HED produced the worst results with  $J = 70.1\%$ ,  $D = 75.2\%$  and  $Sen = 79.6\%$ ,  $MAVD = 8.2 \text{ cm}^3$  and  $HD = 11.5 \text{ mm}$ . This may be due to the network being less robust at storing spatial information, which is essential in semantic segmentation.

#### 4.3. Fetal Brain Volume Estimation

We also evaluate the robustness of the proposed architecture by comparing the volumes generated automatically with those obtained manually. We compute brain volume by calculating the segmented area for each slice and sum it over the total number of slices, taking account of the pixel dimension and slice thickness. Figure 5a demonstrates that there is an excellent correlation between the automatically generated brain volumes (using the  $P_{AVG}$  segmentations) and those obtained from the expert manual segmentations ( $R^2 = 0.9935$ ,  $p < 0.001$ ) for the 200 normal healthy foetuses. Figure 5b shows the corresponding Bland–Altman plot which indicates that there is good overall agreement with a mean difference between the measurements of  $1.29 \text{ cm}^3$  (95% limits of agreement  $\{-11.22 \text{ cm}^3, 13.80 \text{ cm}^3\}$ ), which suggests that the proposed method tends to slightly overestimate the brain volume. Both graphs demonstrate that the distribution of volumes spreads more as the volume increases. This is due to the more complex appearance of the brain in older foetuses (e.g., above 30 weeks) compared to younger foetuses, which makes the brain more challenging to segment due to increased partial volume effects and there may also be more intrinsic inter-subject variation by this age.

Similarly, Figure 5c demonstrates that there is an excellent correlation between automated and manual volume measurements in the 54 abnormal cases ( $R^2 = 0.9678$ ,  $p < 0.001$ ). The corresponding Bland–Altman plot in Figure 5d once again suggests that the proposed method tends to slightly overestimate brain volumes compared to the expert gold standard (mean difference  $0.31 \text{ cm}^3$ , 95% limits of agreement  $\{-18.17 \text{ cm}^3, 18.79 \text{ cm}^3\}$ ). While the mean bias is slightly lower for the abnormal cases, the limits of agreement indicate a greater amount of variability in the measured volumes when compared to the gold standard expert segmentations. This is expected due to (i) the model being independently trained/validated on normal cases only and (ii) that some of the abnormalities introduce a more complex appearance to the brain, which, like the older foetuses, makes the brain harder to segment reliably. Nevertheless, the automated method produced promising results with most estimated volumes lying very close to the straight line. This gives us confidence that the method will provide a reliable segmentation in a wide range of abnormalities, even when the model is trained on normal healthy cases. This is particularly useful for rare pathologies, where it may be difficult to build up large training sets of abnormal cases.

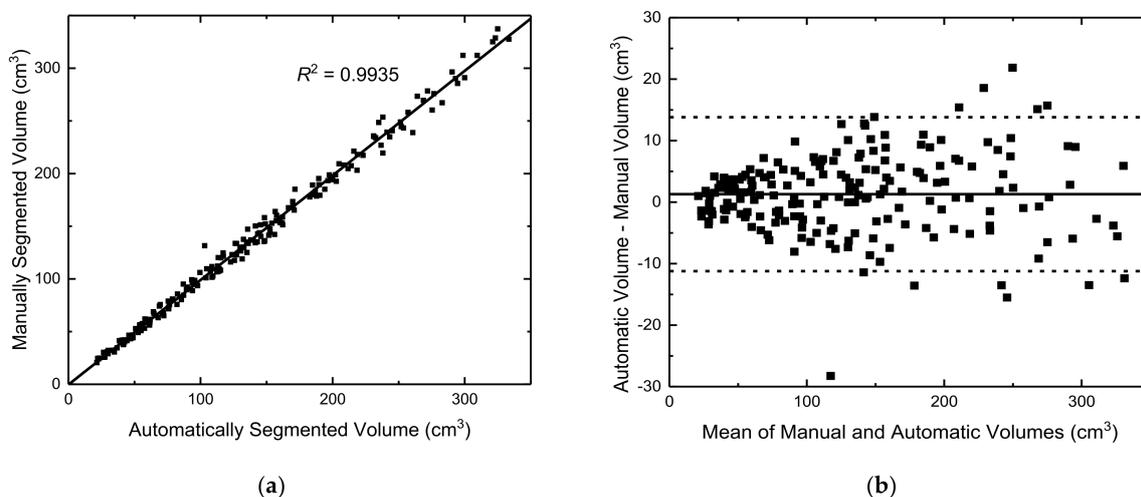
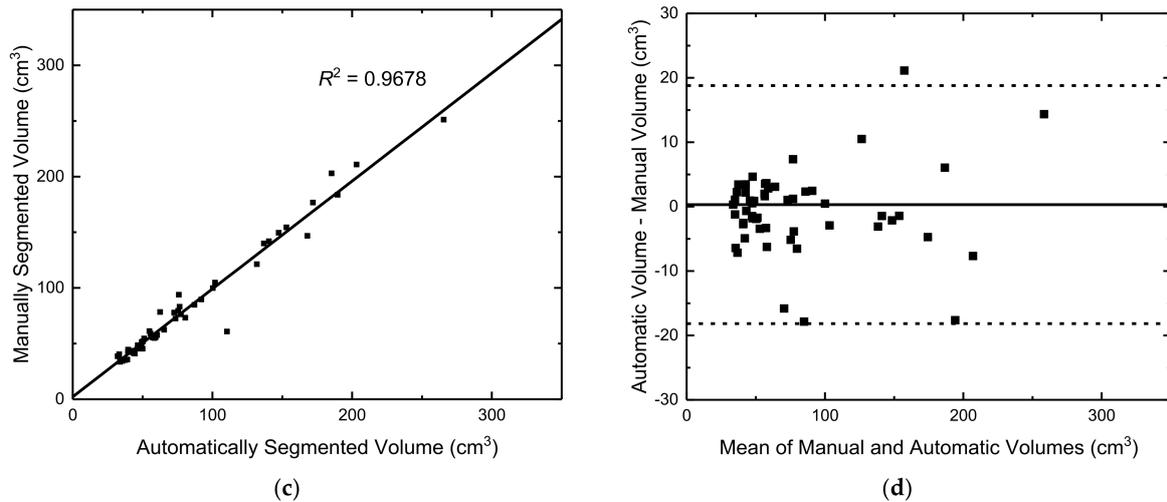


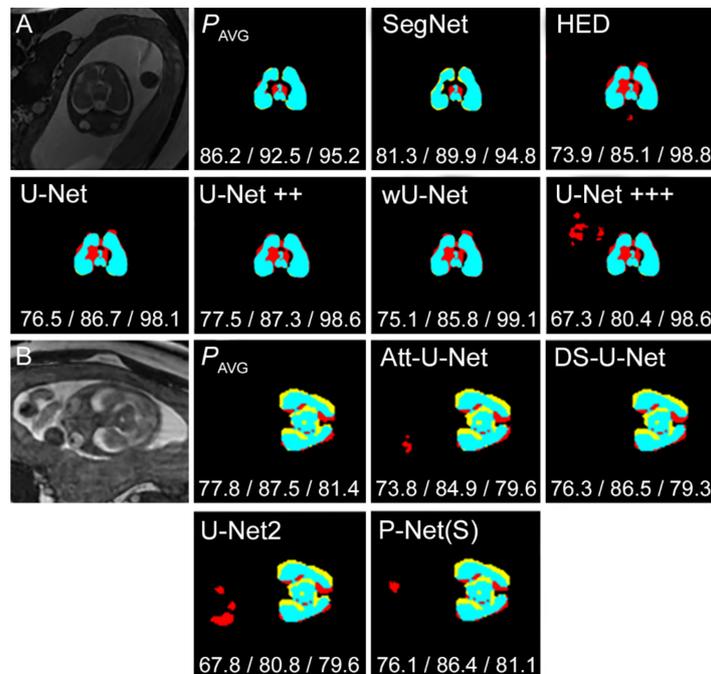
Figure 5. Cont.



**Figure 5.** Correlation between manual and automated ( $P_{AVG}$ ) foetal brain volume measurements for (a) 200 normal cases and (c) 54 abnormal cases. Bland–Altman plots showing the limits of agreement between automatic and manual foetal brain volumes for (b) the 200 normal cases and (d) the 54 abnormal cases.

4.4. Visual Comparison with Other Network Architectures

Figure 6A provides a visual comparison between the segmentation results produced by the SIMOU-Net and the original U-Net (and its variants), HED [5] and SegNet [31]. In example A, the upper part of the brain is overlapping with the skull. All architectures except the SegNet over segmented this region resulting in false positives (red), whereas SegNet under segmented this region resulting in false negatives (yellow). The SIMOU-Net segments this area with small false positives producing the best results of  $J = 86.2\%$  and  $D = 92.5\%$ . The U-Net+++ (with deep supervision) exhibits overfitting, producing the worst results of  $J = 67.3\%$  and  $D = 80.4\%$ .

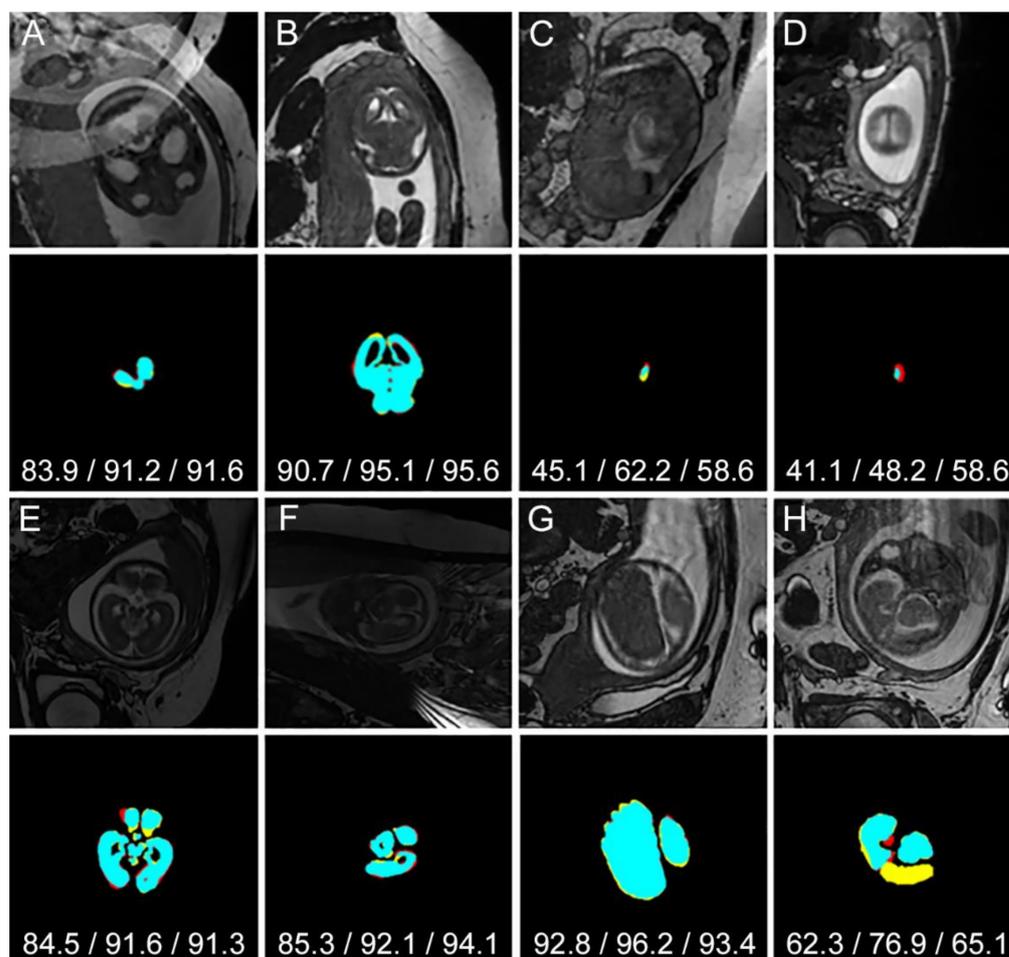


**Figure 6.** Two segmentation examples (A,B) providing a visual comparison between SIMOU-Net ( $P_{AVG}$ ) and (A) SegNet, HED, the original U-Net and its variants and (B) some of the latest networks, Att-U-Net, DS-U-Net, U-Net2 and P-Net(S). Black, cyan, red and yellow represent  $TN$ ,  $TP$ ,  $FP$  and  $FN$ , respectively. Jaccard/Dice/Sensitivity scores are shown under each segmentation.

In Figure 6B, we compare the SIMOU-Net with some of the latest network architectures specifically designed for foetal brain segmentation. Our method outperformed the U-Net2 [12] by at least 8% for  $J$  and  $D$  metrics. However, it only produced marginally better results across different metrics, particularly when compared to the DS-U-Net and P-Net(S).

#### 4.5. Segmentation of Challenging Cases

Figure 7 shows example images and quantitative results for some of the challenging cases. Example A shows the brain region overlapping with a ghosting artefact and the proposed method still achieved a credible  $D = 91.2\%$ . In example B, the boundary of the brain is obscured as it is overlapping with the skull. In this case, the proposed method achieved  $D = 95.1\%$ . In more challenging cases (examples C and D), the region is very small and obscured making it very difficult to detect even for experienced observers. In examples C and D, the SIMOU-Net produced Dice scores of 62.2% and 48.2%, respectively. The bottom half of Figure 7 shows quantitative results for abnormal cases (agenesis of corpus callosum (ACC), mild ventriculomegaly (MV) and structurally deformed). Example E shows an MV case where SIMOU-Net produced  $D = 91.6\%$ . For the ACC case in example F, the image suffers from low contrast, but the proposed method still managed to produce  $D = 92.1\%$ . Structurally deformed cases can be seen in examples G and H.



**Figure 7.** Examples of  $P_{AVG}$  segmentation results for challenging cases (A–D) and for abnormal cases; mild ventriculomegaly (E), agenesis of corpus callosum (F) and structurally deformed (G,H). Black, cyan, red and yellow represent  $TN$ ,  $TP$ ,  $FP$  and  $FN$ , respectively. Jaccard/Dice/Sensitivity scores are shown under each segmentation.

#### 4.6. Ensemble Learning on the Original U-Net and Att-U-Net

We conducted another experiment by employing the SIMO approach (taking into account features from side outputs) on the original U-Net and Att-U-Net. Table 3 indicates that ensemble learning increased the overall results across different metrics for both backbone architectures. For example, the original U-Net ( $P_{\{f\}}$ ) produced  $J = 79.2\%$ ,  $D = 86.8\%$  and  $Sen = 92.4\%$ , but fusing different probability maps using the ‘average’ operation ( $P_{AVG}$ ) achieved  $J = 83.2\%$  (+4%),  $D = 89.2\%$  (+2.9%),  $Sen = 90.1\%$  (+1.3%). When using the ‘median’ operation, the ensemble learning approach improved the results by  $J = +3.5\%$ ,  $D = +1.7\%$ ,  $Sen = +0.4\%$  compared to the original U-Net. The original Att-U-Net ( $P_{\{f\}}$ ) produced  $J = 83.5\%$ ,  $D = 90.9\%$  and  $Sen = 95.1\%$ , but fusing different probability maps using the ‘average’ operation ( $P_{AVG}$ ) achieved  $J = 85.6\%$  (+2.1%),  $D = 92.5\%$  (+1.6%),  $Sen = 95.2\%$  (+0.1%). However, these differences did not generally reach significance with the exception of the U-Net evaluation base on the  $J$  metric.

**Table 3.** Average quantitative results (mean  $\pm$  standard deviation) from 200 normal cases employing the SIMO approach to the U-Net and Att-U-Net with three-fold cross-validation. Bold typeface indicates the best performing method for each metric. \* indicates a significant difference with respect to the  $P_{AVG}$  score after performing a Nemenyi test.

Network	$J$ (%)	$D$ (%)	$Sen$ (%)	MAVD (cm <sup>3</sup> )	HD (mm)
<b>U-Net</b>					
$P_{\{d,ux\}}$	80.2 $\pm$ 7.9 *	87.2 $\pm$ 7.1	92.2 $\pm$ 5.3	6.3 $\pm$ 5.7	8.6 $\pm$ 4.9
$P_{\{d,u\}}$	80.5 $\pm$ 7.5	88.9 $\pm$ 7.1	92.9 $\pm$ 5.3	6.2 $\pm$ 5.7	8.5 $\pm$ 4.7
$P_{\{d\}}$	70.2 $\pm$ 9.4 *	76.2 $\pm$ 9.0 *	79.6 $\pm$ 7.9 *	6.7 $\pm$ 5.9	9.5 $\pm$ 5.2
$P_{\{u\}}$	79.0 $\pm$ 8.1 *	86.1 $\pm$ 7.6 *	91.1 $\pm$ 6.6	6.3 $\pm$ 5.7	8.6 $\pm$ 4.9
$P_{\{f\}}$	79.2 $\pm$ 8.1 *	86.8 $\pm$ 7.4	92.4 $\pm$ 6.4	6.3 $\pm$ 5.7	8.6 $\pm$ 4.8
$P_{MAX}$	73.1 $\pm$ 9.3 *	79.1 $\pm$ 7.9 *	81.1 $\pm$ 7.2	6.5 $\pm$ 5.8	8.8 $\pm$ 5.1
$P_{MED}$	82.7 $\pm$ 7.6	88.5 $\pm$ 6.4	92.8 $\pm$ 5.8	6.0 $\pm$ 5.7	7.8 $\pm$ 4.8
$P_{AVG}$	<b>83.2 <math>\pm</math> 7.6</b>	<b>89.2 <math>\pm</math> 6.2</b>	<b>93.7 <math>\pm</math> 5.7</b>	<b>6.0 <math>\pm</math> 5.2</b>	<b>7.8 <math>\pm</math> 4.5</b>
<b>Att-U-Net</b>					
$P_{\{d,ux\}}$	84.9 $\pm$ 6.9	90.4 $\pm$ 6.2 *	94.1 $\pm$ 5.3	5.6 $\pm$ 3.5	8.1 $\pm$ 4.2
$P_{\{d,u\}}$	84.5 $\pm$ 6.8	90.2 $\pm$ 6.1 *	93.8 $\pm$ 5.4	5.6 $\pm$ 3.5	8.1 $\pm$ 4.2
$P_{\{d\}}$	74.3 $\pm$ 8.8 *	79.9 $\pm$ 8.7 *	82.6 $\pm$ 7.1	6.3 $\pm$ 4.2	9.4 $\pm$ 5.2
$P_{\{u\}}$	81.1 $\pm$ 7.9 *	86.8 $\pm$ 7.3 *	93.3 $\pm$ 6.3	6.1 $\pm$ 3.8	9.2 $\pm$ 4.9
$P_{\{f\}}$	83.5 $\pm$ 7.8	90.9 $\pm$ 6.3	95.1 $\pm$ 6.4	5.5 $\pm$ 3.5	8.1 $\pm$ 4.0
$P_{MAX}$	73.9 $\pm$ 9.3 *	80.6 $\pm$ 7.6 *	84.2 $\pm$ 7.0	6.5 $\pm$ 4.1	10.1 $\pm$ 6.1
$P_{MED}$	84.3 $\pm$ 7.2	90.2 $\pm$ 6.1 *	94.4 $\pm$ 5.3	5.6 $\pm$ 3.5	8.3 $\pm$ 4.4
$P_{AVG}$	85.6 $\pm$ 7.2	92.5 $\pm$ 6.6	95.2 $\pm$ 5.2	4.5 $\pm$ 2.9	7.8 $\pm$ 4.3

#### 4.7. Inter-Observer Evaluation

To further evaluate whether the automated approach is performing at a sufficient level, or not, we compared the automated segmentations to inter-observer measurements. For this purpose, we asked a second observer (who was trained in foetal segmentation by the same expert neuroradiologist as for our ‘gold standard’ observer) to annotate 25 randomly selected cases. Subsequently, we separated the 25 cases from our training sample and used the remaining cases to train and validate the network (150 and 25 cases for training and validation, respectively). Dice scores were calculated between the automated segmentation  $P_{AVG}$  and ‘gold standard’ observers and between the second and ‘gold standard’ observers. Data were checked for normality using a Shapiro–Wilk test and a paired t-test demonstrated that there was no significant difference in Dice scores between the automated segmentation ( $D = 89.88 \pm 2.59$ ) and inter-observer ( $D = 88.58 \pm 3.96$ ) situations ( $p = 0.15$ ). Similarly, a one-way ANOVA indicated that there was no significant difference in the total brain volumes measured between the automated segmentation, second observer and ‘gold standard’ ( $p = 0.61$ ), suggesting that the automated method performs comparably to the two trained observers. In addition, the intraclass correlation coefficient was calculated between the

automated and two observers' volume measurements and gave a value of 0.953 which is also indicative of an excellent agreement.

#### 4.8. Advantages, Disadvantages and Limitations

The main advantage of the proposed network architecture is that it optimises the model by averaging several activation maps, reducing false positives and negatives. Furthermore, this strategy promotes ensemble learning in the network which reduces the variance of predictions and generalization errors. The network can also recover important cues that might disappear due to repetitive convolution operations in the upsampling and downsampling paths. For example, some cues might only be available in the downsampling path (or at the beginning of the upsampling path) of the network and these cues might disappear after many convolutions. In addition, the concept of global and local information in our study is richer because we not only extract this information from the network's main convolutional outputs but also from side outputs.

The main disadvantage is that our network is more complex due to more convolutional layers and additional feature maps extracted from the side outputs. As a result, it has more parameters and takes a longer time to train (approximately 24 h depending on the number of training samples). However, the network is still considered to be 'lighter' compared to U-Net++ and U-Net+++ [28]. The methods of Zhou et al. [29], despite being capable of producing good results, have complex skip connection paths due to the nested strategy which requires maps from downsampling paths to undergo a dense convolution block. The U-Net [4] and SegNet [33] have simpler architectures, but as shown in Tables 1 and 2, their performance is lower compared to the recent methods developed in 2018 and 2019 such as the Att-U-Net [34] and DS-U-Net [27]. Hence, there is still room for improvement.

As for the HED [5] and wU-Net [29] networks, they are much simpler alternatives and took less than 10 h to train. Nevertheless, their performances are among the worst in this study. The Att-U-Net [34] is more complex than the classical U-Net but is considered 'lighter' than our network. The three pathways method proposed by Salehi et al. [22] produced promising results but requires annotations from different views of the brain such as axial, sagittal and coronal. The entire methodology requires training three networks, eventually merging all outputs, which is time consuming and complex. Recent methods that also show excellent results are the P-Net(S) [26] and DS-U-Net [27]. Nevertheless, both methods require several training runs of the same network for localisation, coarse segmentation and fine segmentation purposes.

While SIMOU-Net ( $P_{AVG}$ ) was consistently the best performing network in this study, it should be acknowledged that this network has been developed and optimised for this specific study. This may introduce some positive bias towards its performance, even though we aimed to ensure that the other networks were equally well optimised for the task. This may affect the generalizability of the findings when applied to other similar segmentation problems, although we would expect the relative performances to be similar to that found in this study. Furthermore, while the incorporation of information from the side outputs into the final prediction map consistently improved the performance of the segmentation, the differences were small between  $P_{AVG}$  and  $P_{\{f\}}$  in most instances and rarely reached statistical significance.

## 5. Conclusions

We proposed SIMOU-Net, a hybrid network architecture inspired by the original U-Net and HED architectures. SIMOU-Net not only takes advantage of the original U-Net features, which extract location and contextual information, but it also takes account of local and global information extracted from the side outputs. As a result, instead of relying on a single prediction map, as in the original U-Net network, SIMOU-Net produces several prediction maps and then applies 'averaging' operations on these maps. This approach reduces both the variance of predictions and the generalisation error, which often leads to better results. Furthermore, we have demonstrated that fusing information from multiple

side outputs also improves the performance of other network configurations, suggesting that the approach has more general applicability. Experimental results obtained from foetal brain MRI segmentation are promising and indicate that SIMOU-Net is robust and competitive with similar studies in the literature. For future work, we plan to evaluate SIMOU-Net for multi-class segmentation of the foetal brain and for liver segmentation in the foetus via a transfer learning approach.

**Author Contributions:** Conceptualization, P.A.A., A.R. and P.D.G.; methodology, A.R., D.J., P.D.G., R.Z., B.W.S., P.A.A.; software, A.R.; validation, A.R., P.A.A.; formal analysis, A.R.; investigation, A.R.; resources, A.R., P.A.A.; data curation, A.R., P.A.A., D.J.; writing—original draft preparation, A.R., P.A.A.; writing—review and editing, A.R., D.J., P.D.G., R.Z., B.W.S., P.A.A.; visualization, A.R., P.A.A.; supervision, P.A.A.; project administration, P.A.A.; funding acquisition, P.A.A., P.D.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was part funded by a grant from the National Institute for Health Research (UK), (NIHR HTA 09/06/01).

**Institutional Review Board Statement:** The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the relevant Ethics Board (reference numbers: REC11/YH/0006 and REC10/H1308/2).

**Informed Consent Statement:** Written informed consent was obtained from the pregnant women involved in the study.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author, subject to ethical considerations.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References

1. Griffiths, P.D.; Bradburn, M.; Campbell, M.; Cooper, C.L.; Graham, R.; Jarvis, D.; Kilby, M.D.; Mason, G.; Mooney, C.; Robson, S.C.; et al. Use of MRI in the diagnosis of fetal brain abnormalities in utero (MERIDIAN): A multicentre, prospective cohort study. *Lancet* **2017**, *389*, 538–546. [[CrossRef](#)]
2. Griffiths, P.D.; Jarvis, D.; McQuillan, H.; Williams, F.; Paley, M.; Armitage, P. MRI of the foetal brain using a rapid 3D steady-state sequence. *Br. J. Radiol.* **2013**, *86*, 20130168. [[CrossRef](#)] [[PubMed](#)]
3. Makropoulos, A.; Counsell, S.K.; Rueckert, D. A review on automatic fetal and neonatal brain MRI segmentation. *NeuroImage* **2018**, *170*, 231–248. [[CrossRef](#)] [[PubMed](#)]
4. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*; Springer: Munich, Germany, 2015; Volume 9351, pp. 234–241.
5. Xie, S.; Tu, Z. Holistically-nested edge detection. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 8–10 June 2015; pp. 1395–1403.
6. Rampun, A.; Zheng, L.; Malcolm, P.; Tiddeman, B.; Zwigelaar, R. Computer-aided detection of prostate cancer in T2-weighted MRI within the peripheral zone. *Phys. Med. Biol.* **2016**, *61*, 4796–4825. [[CrossRef](#)] [[PubMed](#)]
7. Avnimelech, R.; Intrator, N. Boosted Mixture of Experts: An Ensemble Learning Scheme. *Neural Comput.* **1999**, *11*, 483–497. [[CrossRef](#)] [[PubMed](#)]
8. Rokach, L. Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. *Comput. Stat. Data An.* **2009**, *53*, 4046–4072. [[CrossRef](#)]
9. Anquez, J.; Angelini, E.D.; Bloch, I. Automatic segmentation of head structures on fetal MRI. In Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Boston, MA, USA, 28 June–1 July 2009; pp. 109–112.
10. Taleb, Y.; Schweitzer, M.; Studholme, C.; Koob, M.; Dietemann, J.L.; Rousseau, F. *Automatic Template-Based Brain Extraction in Fetal MR Images*; HAL: Lyon, France, 2013.
11. Alansary, A.; Lee, M.; Keraudren, K.; Kainz, B.; Malamateniou, C.; Rutherford, M.; Hajnal, J.V.; Glocker, B.; Rueckert, D. Automatic Brain Localization in Fetal MRI Using Superpixel Graphs. In Proceedings of the Machine Learning Meets Medical Imaging, Lille, France, 11 July 2015; pp. 13–22.
12. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Susstrunk, S. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2281. [[CrossRef](#)]
13. Link, D.; Braginsky, M.B.; Joskowicz, L.; Sira, L.B.; Harel, S.; Many, A.; Tarrasch, R.; Malinger, G.; Artzi, M.; Kapoor, C.; et al. Automatic Measurement of Fetal Brain Development from Magnetic Resonance Imaging: New Reference Data. *Fetal Diagn. Ther.* **2018**, *43*, 113–122. [[CrossRef](#)]

14. Attallah, O.; Sharkas, M.A.; Gadelkarim, H. Fetal brain abnormality classification from MRI images of different gestational age. *Brain Sci.* **2019**, *9*, 231. [[CrossRef](#)]
15. Ison, M.; Dittrich, E.; Donner, R.; Kasprian, G.; Prayer, D.; Langs, G. Fully automated brain extraction and orientation in raw fetal MRI, In MICCAI Workshop on Paediatric and Perinatal Imaging 2012 (PaPI 2012). In Proceedings of the MICCAI Workshop on Paediatric and Perinatal Imaging 2012 (PaPI 2012), Nice, France, 1 October 2012; pp. 17–24.
16. Keraudren, K.; Kuklisova-Murgasova, M.; Kyriakopoulou, V.; Mala-mateniou, C.; Rutherford, M.A.; Kainz, B.; Hajnal, J.V.; Rueckert, D. Automated fetal brain segmentation from 2D MRI slices for motion correction. *NeuroImage* **2014**, *101*, 633–643. [[CrossRef](#)]
17. Kainz, B.; Keraudren, K.; Kyriakopoulou, V.; Rutherford, M.; Hajnal, J.V.; Rueckert, D. Fast fully automatic brain detection in fetal MRI using dense rotation invariant image descriptors. In Proceedings of the IEEE 11th International Symposium on Biomedical Imaging (ISBI), Beijing, China, 29 April–2 May 2014; pp. 1230–1233.
18. Salehi, S.S.M.; Hashemi, S.R.; Velasco-Annis, C.; Ouaalam, A.; Estroff, J.A.; Erdogmus, D.; Warfield, S.K.; Gholipour, A. Real-time automatic fetal brain extraction in fetal MRI by deep learning. In Proceedings of the IEEE 15th International Symposium on Biomedical Imaging (ISBI), Washington, DC, USA, 4–7 April 2018; pp. 720–724.
19. Rampun, A.; López-Linares, K.; Morrow, P.J.; Scotney, B.W.; Wang, H.; Ocaña, I.G.; Maclair, G.; Zwiggelaar, R.; Ballester, M.A.G.; Macía, I. Breast pectoral muscle segmentation in mammograms using a modified holistically-nested edge detection network. *Med. Image Anal.* **2019**, *57*, 1–17. [[CrossRef](#)] [[PubMed](#)]
20. Hamidinekoo, A.; Denton, E.; Rampun, A.; Honnor, K.; Zwiggelaar, R. Deep Learning in Mammography and Breast Histology, an Overview and Future Trends. *Med. Image Anal.* **2018**, *76*, 45–67. [[CrossRef](#)] [[PubMed](#)]
21. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [[CrossRef](#)] [[PubMed](#)]
22. Mohseni Salehi, S.S.; Erdogmus, D.; Gholipour, A. Auto-Context Convolutional Neural Network (Auto-Net) for Brain Extraction in Magnetic Resonance Imaging. *IEEE Trans. Med. Imaging* **2017**, *36*, 2319–2330. [[CrossRef](#)] [[PubMed](#)]
23. Rajchl, M.; Lee, M.C.H.; Oktay, O.; Kamnitsas, K.; Passerat-Palmbach, J.; Bai, W.; Damodaram, M.; Rutherford, M.; Hajnal, J.; Kainz, B.; et al. DeepCut: Object Segmentation From Bounding Box Annotations Using Convolutional Neural Networks. *IEEE Trans. Med. Imaging* **2017**, *36*, 674–683. [[CrossRef](#)] [[PubMed](#)]
24. Rother, C.; Kolmogorov, V.; Blake, A. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph. (TOG)* **2004**, *23*, 309–314. [[CrossRef](#)]
25. Khalili, N.; Lessmann, N.; Turk, E.; Claessens, N.; de Heus, R.; Kolk, T.; Viergever, M.A.; Benders, M.J.N.L.; Isgum, I. Automatic brain tissue segmentation in fetal MRI using convolutional neural networks. *Magn. Reson. Imaging* **2019**, *64*, 77–89. [[CrossRef](#)] [[PubMed](#)]
26. Ebner, M.; Wang, G.; Li, W.; Aertsen, M.; Patel, P.A.; Aughwane, R.; Melbourne, A.; Doel, T.; David, A.L.; Deprest, J.; et al. An Automated Localization, Segmentation and Reconstruction Framework for Fetal Brain MRI. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*; Springer: Granada, Spain, 2018; Volume 11070, pp. 313–320.
27. Lou, J.; Li, D.; Bui, T.D.; Zhao, F.; Sun, L.; Li, G.; Shen, D. Automatic Fetal Brain Extraction Using Multi-stage U-Net with Deep Supervision. *Mach. Learn. Med Imaging* **2019**, 592–600. [[CrossRef](#)]
28. Dou, Q.; Chen, H.; Jin, Y.; Yu, L.; Qin, J.; Heng, P.-A. 3D Deeply Supervised Network for Automatic Liver Segmentation from CT Volumes. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*; Springer: Athens, Greece, 2016; Volume 9901, pp. 149–157.
29. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
30. Hinton, G. Neural Networks for Machine Learning—Lecture 6a—Overview of Mini-Batch Gradient Descent. Available online: [https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture\\_slides\\_lec6.pdf](https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf) (accessed on 4 July 2019).
31. Hinton, G.E. Learning to represent visual input. *Philos. Trans. R. Soc. B Biol. Sci.* **2010**, *365*, 177–184. [[CrossRef](#)]
32. Jarvis, D.A.; Finney, C.R.; Griffiths, P.D. Normative volume measurements of the fetal intra-cranial compartments using 3D volume in utero MR imaging. *Eur. Radiol.* **2019**, *29*, 3488–3495. [[CrossRef](#)]
33. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [[CrossRef](#)] [[PubMed](#)]
34. Schlemper, J.; Oktay, O.; Schaap, M.; Heinrich, M.; Kainz, B.; Glocker, B.; Rueckert, D. Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.* **2019**, *53*, 197–207. [[CrossRef](#)] [[PubMed](#)]