Subcortical Segmentation of the Fetal Brain in 3D Ultrasound using Deep Learning

Linde S. Hesse^{a,*}, Moska Alias^b, Felipe Moser^a, the INTERGROWTH-21st
Consortium, Monique C. Haak^b, Weidi Xie^c, Mark Jenkinson^{d,e,f}, Ana I.L.
Namburete^g

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

^b*Department of Obstetrics and Fetal Medicine, Leiden University Medical Center, The Netherlands*

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

^d*Wellcome center for Integrative NeuroImaging, FMRIB, University of Oxford, United Kingdom*

^e*Australian Institute for Machine Learning (AIML), Australia*

^f*South Australian Health and Medical Research Institute (SAHMRI), Australia*

^g*Department of Computer Science, University of Oxford, United Kingdom*

Abstract

The quantification of subcortical volume development from 3D fetal ultrasound can provide important diagnostic information during pregnancy monitoring. However, manual segmentation of subcortical structures in ultrasound volumes is time-consuming and challenging due to low soft tissue contrast, speckle and shadowing artifacts. For this reason, we developed a convolutional neural network (CNN) for the automated segmentation of the choroid plexus (CP), lateral posterior ventricle horns (LPVH), cavum septum pellucidum et vergae (CSPV), and cerebellum (CB) from 3D ultrasound. As ground-truth labels are scarce and expensive to obtain, we applied few-shot learning, in which only a small number of manual annotations ($n = 9$) are used to train a CNN. We compared training a CNN with only a few individually annotated volumes versus many weakly labelled volumes obtained from atlas-based segmentations. This showed that

*Corresponding author

Email address: linde.hesse@seh.ox.ac.uk (Linde S. Hesse)

segmentation performance close to intra-observer variability can be obtained with only a handful of manual annotations. Finally, the trained models were applied to a large number ($n = 278$) of ultrasound image volumes of a diverse, healthy population, obtaining novel US-specific growth curves of the respective structures during the second trimester of gestation.

Keywords: Ultrasound, Segmentation, Fetal, Brain, Subcortical, Convolutional Neural Network, Few-Shot Learning, Deep Learning

1. Introduction

During pregnancy, several subcortical structures in the fetal brain are assessed with ultrasound (US) imaging. Especially earlier in gestation, when the fetal skull is not yet fully calcified, the US beam can penetrate the skull and visualize the subcortical structures. The abnormal development of subcortical structures can be a potential sign of a severe neurological condition and, as such, it is important to monitor their development during gestation. For example, an enlarged ventricular cavity can be a sign of a central nervous system (CNS) abnormality [1] whereas the inability to visualize the CSP after 17-20 gestational weeks (GWs) may be indicative of underdevelopment of the corpus callosum, which is a fiber track connecting the two cerebral hemispheres [1–3]. Brain development can be studied in detail with targeted fetal neurosonography [4], however, this is only performed in fetuses at high-risk for CNS abnormalities and is not part of routine obstetric examinations. For this reason, efforts should be made to develop analysis methods that can improve subcortical assessment during routine pregnancy monitoring.

Subcortical structure development is best quantitatively assessed using 3D volumetric information. However, to date, most studies analyzing these structures during gestation only use in-plane measurements obtained from 2D US [5, 6]. These measurements are routinely acquired during a basic examination of the fetal brain [1] but only provide a limited representation of the anatomical development and morphology. For this reason, it is desirable to use 3D US to obtain volumetric measures of subcortical development.

25 As manually annotating several structures is not feasible in routine clinical
26 practice, automated subcortical segmentation methods for 3D US could fa-
27 cilitate this analysis and may provide new insights into in utero subcortical
28 development.

29 Structural segmentation in fetal brain US is a challenging task due to
30 low soft tissue contrast, reverberation artifacts and the characteristic pres-
31 ence of speckle. Consequently, precise structural boundaries can be hard to
32 distinguish, resulting in high inter- and intra-observer variability in manual
33 annotations. Since manual segmentation is not a task usually performed in
34 clinical practice, even trained ultrasonographers can have difficulty in accu-
35 rately segmenting subcortical structures in 3D US volumes. An additional
36 challenge of US volumes acquired with a free-hand scanning protocol, as is
37 typical at the bedside, is the varying position of the fetal brain due to the
38 unpredictable fetal position in the womb as well as movement of the trans-
39 ducer relative to the fetal head. Furthermore, due to interactions of the US
40 beam with the fetal skull, typically only the cerebral hemisphere distal to
41 the US transducer is well visible in the US volumes.

42 Recently, it has been shown that deep learning methods can be success-
43 fully applied to other segmentation tasks in 3D US volumes of the fetal
44 brain [7–9], resulting in higher performance than traditional image analysis
45 methods. However, due to the difficulty of obtaining manual annotations
46 for subcortical structures, a key barrier to applying deep learning methods
47 to this task is obtaining sufficient ground-truth annotations for training.
48 Therefore, to the best of our knowledge, only two previous studies have ap-
49 plied deep learning to subcortical structure segmentation in 3D US, which
50 will be discussed in more detail in Section 2.1.

51 To overcome the need for a large manually annotated dataset, few-shot
52 learning can be used, in which only a small number of manual annota-
53 tions are used to train a convolutional neural network (CNN). Several few-
54 shot learning approaches have been proposed for segmentation tasks in the
55 medical image domain [10–12], such as training a CNN with pseudo-labels
56 obtained from few annotated 2D slices [12] or with additionally generated
57 labels by a generative adversarial network [11]. These studies thus show that

58 good segmentation performance can be obtained using only a very limited
59 amount of voxel-wise manual annotation.

60 In this work, we will use few-shot learning to develop a deep-learning
61 based method for the segmentation of several subcortical structures in 3D
62 US. To the best of our knowledge, few-shot learning has not been applied
63 for this task, and, as such, this will be the first study exploring subcortical
64 segmentation of the fetal brain in a low-data regime. Specifically, we will
65 compare two types of ground-truth labels for training, where both have
66 been obtained using only a few manual annotations: (1) a small number
67 of individually annotated US volumes (referred to as *expert labels*), and
68 (2) a large number of weakly labelled US volumes obtained by propagating
69 annotated template images (referred to as *atlas labels*).

70 Additionally, we investigate the impact of image alignment to a common
71 coordinate system as a pre-processing step. As the anatomical orientation
72 of the fetal brain varies in US image volumes, initial global (affine) registra-
73 tion of the brain is expected to have a positive effect on the segmentation
74 performance. Although the alignment of US volumes is a non-trivial task,
75 methods have been developed for the automated alignment of fetal brain
76 volumes [13, 14]. Furthermore, compared to the voxel-wise annotation of
77 several subcortical structures, manual image alignment requires consider-
78 ably less effort. For these reasons, alignment of the volumes to a common
79 coordinate can be preferred over the additional voxel-wise manual annota-
80 tion, and, as such, it is important to explore global alignment as a requisite
81 preprocessing step.

82 Using few-shot learning, we aim to segment the choroid plexus (CP),
83 lateral posterior ventricle horn (LPVH), cerebellum (CB) and cavum sep-
84 tum pellucidum et vergae (CSPV) during the second trimester of gestation.
85 This trimester is of particular interest since this is when women undergo an
86 US examination as part of routine care to screen for anomalies [15]. During
87 this scan, several linear measurements, such as the transcerebellar diameter
88 (TCD) and the atrial width, are collected and used for diagnostic purposes.
89 Furthermore, in this trimester most subcortical structures have developed
90 enough to be visible on US but are less affected by acoustic shading due to

91 progressing calcification of the skull later in gestation. The specific subcor-
92 tical structures studied in this work were selected based on their importance
93 in this routine anomaly scan.

94 Lastly, to obtain an improved understanding of volumetric subcortical
95 development during the second trimester, we will apply our developed seg-
96 mentation models to a large cohort of healthy fetuses. The resulting model
97 predictions will be validated for the CB, as volumetric growth curves for the
98 CB have been reported in previous studies [16–22], and subsequently used to
99 generate novel US-specific growth curves of subcortical volume development
100 during the second trimester of gestation.

101 2. Related Work

102 Automated subcortical structure segmentation has been widely studied
103 in MRI volumes of the adult brain [23–25]. However, the adult brain is
104 structurally different to the actively developing fetal brain during gestation,
105 which contains structures in transient stages of morphogenesis. For example,
106 certain structures, such as the CSPV (which is in truth a fluid-filled cavity),
107 are only present during gestation and typically disappear after birth [26]. As
108 MRI can also be safely used to visualize the fetal brain, but is clinically only
109 recommended to confirm or complement a diagnosis following an US exam-
110 ination [4], some studies have performed subcortical structure segmentation
111 in fetal MRI volumes [27–31]. Notably, Gholipour et al. [30] published an an-
112 notated atlas of the fetal brain from 21 to 39 weeks of gestation and applied
113 these labels to perform multi-atlas segmentation for new subjects. However,
114 segmentation methods developed for MRI cannot be directly applied to US
115 volumes due to the very different nature of the image acquisitions. While
116 US is based on the interaction of the sound waves with tissue boundaries,
117 resulting in discontinuous boundaries in the resulting image, MRI is based
118 on protons in the tissue that react to magnetic fields. As a result, intensities
119 on an MRI scan are rather uniform, whereas the intensities on an US scan
120 heavily depend on the distance of the boundary to the probe as well as on
121 the boundaries the US beam has already passed through. For example, the

two lateral ventricles will appear very similar on a fetal brain MRI whereas on US typically only one lateral ventricle is well visible due to shadowing of the fetal skull in the proximal hemisphere. As a result of these acquisition differences, there is no direct one-to-one intensity mapping between US and MRI images, which makes it challenging to use fetal MRI atlas labels for US images.

In the remainder of this section we will outline previous work performed for subcortical structure segmentation in US. Furthermore, we will provide a brief overview of studies analyzing volumetric subcortical structure development during gestation.

2.1. Automated subcortical segmentation in US

Due to low soft-tissue contrast, high amount of speckle, and shadowing artifacts, automated anatomical segmentation in US volumes is a challenging task. Since most subcortical structures are small (i.e. the CB covers about 2.5% and the CP 0.8% of the intracranial volume at 22 GWs) and their boundaries are not always easily distinguishable, automated analysis of these structures has not been studied widely. Furthermore, manual segmentation is also challenging due to the aforementioned characteristics of US, and it can therefore be difficult to obtain accurate ground-truth labels. Before summarizing the past work, it is worth noting that the performance reported in different studies is not directly comparable due to the fact that the quality of US volumes can vary tremendously as a result of the transducer, acquisition protocol and scan settings. Additionally, performance measures such as the Dice Similarity Coefficient (DSC), which represent the percentage of overlap between ground-truth and prediction, are biased towards structural size, as for smaller structures the impact of a wrongly predicted voxel is much bigger than for bigger structures. For this reason, these should be carefully interpreted across the gestational age (GA) range, during which most structures grow substantially.

To the best of our knowledge, the first automated subcortical segmentation method for fetal brain US was developed by Gutiérrez-Becker et al. [32]. In that study, a statistical shape model was used to segment the CB

154 in 3D fetal US volumes. The performance was promising, but it is not clear
155 how this method can be easily extended to other subcortical structures with
156 more shape variation than the CB, making it challenging to develop an
157 initial shape model.

158 In another study [33], segmentation of the CP, LPVH, the CSP and
159 the CB was performed in 3D US scans between 18 and 26 weeks of ges-
160 tation. The authors proposed using a random decision forest using both
161 appearance and distance features. Reported performance was good for the
162 CP, LPVH and CSPV, but their method failed to generalize well to the CB.
163 Furthermore, the search region for each structure was limited to the smallest
164 cuboid enclosing all manual ground-truth annotations from the respective
165 structure. More recently, Huang et al. [34] proposed using a region-based
166 based descriptor to segment the CSPV (referred to as corpus callosum in
167 the paper) and CP in 2D US. In contrast to the aforementioned studies, a
168 2D slice of the full fetal brain was used as input for the segmentation model.
169 However, the sampled 2D slices were limited to standard US planes, making
170 the segmentation substantially less challenging than full 3D segmentation.

171 Since the introduction of deep learning for image segmentation, most
172 specifically the U-Net architecture [35], many segmentation tasks in medical
173 imaging have shown an increase in performance. However, only a few studies
174 have applied deep learning methods to subcortical structure segmentation
175 in fetal US, as usually a relatively large dataset with ground-truth labels is
176 needed. Obtaining voxel-wise labels for these structures, especially in 3D, is
177 extremely time-consuming and challenging, due to the structural boundaries
178 that are often hard to distinguish for a human annotator. One method that
179 performs localization by predicting a bounding box, and included here for
180 completeness, is the one developed by Huang et al. [36] that uses three 2D
181 CNNs to localize several subcortical structures in 3D US (LPVH, thalami,
182 CB and cisterna magna). Each of these networks predicted the orthogonal
183 projections of the structures onto one of the three anatomical views (axial,
184 sagittal and coronal), and the resulting 3D bounding box was then con-
185 structed from these 2D projections. However, this method is not suitable
186 to extend to full voxel-wise segmentation in 3D as 2D projections can lack

187 information needed to fully reconstruct an object in 3D.

188 Closely related to the study presented in this paper, Venturini et al. [37]
 189 used MRI atlas labels to overcome the need for manual annotation. Labels
 190 from the respective GW template of the fetal brain MR atlas [30] were reg-
 191 istered to the individual US volumes, resulting in weak labels for the white
 192 matter, brainstem, thalami, and CB. These labels were subsequently used
 193 to train a multi-label 3D CNN to segment the respective subcortical struc-
 194 tures in 3D US volumes. The reported performance was good, however, a
 195 severe downside of this study was that testing was performed on the same
 196 weak MR atlas labels as used in training. As the registered MR atlas labels
 197 can be erroneous, and do not always align well with the structures visible
 198 in the US volumes, this therefore yields a biased assessment of performance
 199 and as such does not analyze the effect of using weak labels. In Hesse and
 200 Namburete [9], active contours were applied to improve atlas-based labels
 201 as a pre-processing step before training a 3D U-Net, showing increased per-
 202 formance for CSP segmentation. However, that method was only validated
 203 for an age range between 20 and 24 GWs, and does not easily translate to
 204 all subcortical structures. Another method for segmenting the CSPV was
 205 presented by Wu et al. [38] and used a modified U-Net to perform segmen-
 206 tation in 2D US. However, as in Huang et al. [34], segmentation was limited
 207 to 2D standard planes.

208 In addition to the previously mentioned methods, some clinical studies
 209 report using multiplanar segmentation or Virtual Organ Computer-Aided
 210 AnaLysis (VOCAL) software (GE Healthcare) to perform their volume mea-
 211 surements in 3D US [17, 39–41]. These methods estimate the structural
 212 volume from 2D contours drawn on parallel (multiplanar) or oblique (VO-
 213 CAL) planes. Although this does speed up analysis compared with fully
 214 annotating the volume, it still requires substantial manual annotation and
 215 is, as such, not considered to be an automated method.

216 2.2. Subcortical structure development

217 During a standard fetal US examination, multiple in-plane measure-
 218 ments, such as the TCD, are performed to assess the development and

quantify the normality of the brain. For this reason, most studies analyzing subcortical growth during gestation use these 2D measurements [1]. Studies that do analyze volumetric development of subcortical structures during gestation most frequently use MRI volumes, either by manually annotating the volumes [19, 20, 22] or by using an atlas-based approach [42]. Theoretically, volumetric measures should be independent of the acquisition type (MRI or US), but in practice these measures can vary due to the different appearance of tissues in each modality. Furthermore, subcortical structures that are easily visible on US, such as the CP, are much harder to discern on fetal MRI. For these reasons, US-specific growth curves are needed.

A small number of studies have analyzed 3D subcortical development using US [17, 18, 21, 39–41, 43, 44], and all have used the previously described multiplanar or VOCAL software. In Benavides-Serralde et al. [40], total intracranial, frontal, thalamic and cerebellar volumes were measured between 28 and 34 weeks of gestation and these measurements were used to estimate the average volume differences between a group of 39 growth-restricted fetuses and a group of 39 age-matched appropriate-for-gestational age fetuses. The same structures were measured between 20 and 36 GWs and used to compare the structural volumes of 73 fetuses with congenital heart disease to a group of 168 normal control fetuses [39].

In order to construct normal ranges for thalamic volume development during gestation, Sotiriadis et al. [41] measured the thalamic volume of 122 normal fetuses between 20 and 34 weeks. Furthermore, it was shown that there were no significant volume differences between the left and right thalamus. Setting out to generate reference values and growth curves for 3D US, Babucci et al. [17] obtained volumetric measures for the thalamus, CB and cortex for 344 fetuses from 22 to 39 weeks. Only a small number of volume measurements were, however, included after 30 weeks, due to shadowing of the fetal skull. Focussing only on cerebellar growth, multiple studies have developed US-specific growth curves for this structure from a healthy population [18, 21, 43, 44].

In summary, even though several studies have proposed subcortical segmentation methods for fetal US, a single method obtaining good perfor-

mance across multiple structures is lacking for 3D US. Furthermore, due to the absence of accurate automated segmentation methods, only thalamic and cerebellar volumetric development during gestation have been studied in previous work. In this study, we address these limitations by developing a segmentation method aiming to obtain competitive segmentation performance across multiple subcortical structures (CB, LPVH, CSPV and CP) and subsequently applying these methods to obtain novel volumetric growth curves for the respective structures during gestation.

3. Methods

In this study, we aim to accurately segment the CB, LPVH, CSPV and CP in 3D US image volumes during the second trimester of gestation using a minimal number of voxel-wise annotations, i.e. few-shot segmentation. Defining n_a as the number of images that is manually annotated, we will consider two approaches, both obtained from an equivalent number of manual annotation: (1) training a model naively with n_a individually annotated 3D images, and (2) training a model with a large number of weak propagated atlas labels obtained from annotating n_a 3D template images.

In this section, we start by introducing some notation, followed by an explanation of the network architecture used for training. Next, the few-shot training labels are explained in more detail, including both the template image generation and the manual annotation protocol. In the experimental section, an outline will be given of the data used in this study as well as the experiments performed and the evaluation metrics. Finally, in the last section we will present how the subcortical growth curves are generated.

3.1. Notation

To facilitate the methods presented in this section, we will introduce some notation. Given a dataset of m image volumes, the set of image volumes in their original acquired orientation will be denoted as $\check{\mathcal{X}} = \{\check{X}_1, \dots, \check{X}_m\}$ with $\check{X} \in \mathbb{R}^3$ and the set of image volumes rigidly aligned to the same reference coordinate system as $\dot{\mathcal{X}} = \{\dot{X}_1, \dots, \dot{X}_m\}$. Referring to either $\check{\mathcal{X}}$ or $\dot{\mathcal{X}}$

will simply be done by using $\mathcal{X}$. Furthermore, we will refer to the original and manually aligned images as *unaligned* and *aligned*, respectively.

Multi-label segmentation masks will be denoted by Y . Referring to the binary mask of a single class will be done by the subscript c , with c being the class (structure), and the type of the label (described in next section) will be marked with a superscript (Y_c^{type}). Model predictions are marked with a hat over them, and follow the same convention to denote the class ($\hat{Y}_c$). As introduced before, n_a refers to the number of manually annotated (template) images used for training in a certain experiment. A full overview of all notation used in this work can be found in Appendix Appendix A.

To avoid ambiguity between *image volumes* and a *volume measurement*, in the remainder of this paper we will refer to our 3D image volumes simply as *images*.

3.2. CNN architecture

We propose a multi-label 3D U-Net [35] with batch normalization [45] for the subcortical segmentation. The network, defined by θ , predicts a multi-channel segmentation for a 3D US input image: $\hat{Y} = \theta(X)$. Following the same convention as for the images, $\check{\theta}$ and $\dot{\theta}$ correspond to networks trained with $\check{\mathcal{X}}$ and $\dot{\mathcal{X}}$ respectively.

The network predicts an output prediction with k channels, in which k is equal to the number of segmented classes. To obtain probabilities for each individual class, a soft-max function is applied across the channel dimension of $\hat{Y}$. Final predictions for a single class, denoted by $\hat{Y}_c$, are subsequently obtained by assigning each voxel to the class with the highest probability. A schematic overview of the segmentation network is shown in Fig. 1.

The U-Net architecture consists of two symmetrical branches, an encoder and a decoder, linked by skip connections. Both the encoder and decoder are composed of several convolutional blocks, each containing two 3D convolutional layers (kernel size of 3x3x3), followed by a ReLU activation function and batch normalization [45]. In the encoder branch, each block is followed by a max-pool layer performing down-sampling, whereas in the decoder branch each block is followed by an up-sampling operation. The

Figure 1: Schematic overview of segmentation network. The segmented structures shown are the CB (pink), LPVH (green), CSPV (yellow) and CP (orange). The LPVH and CP are shown in both hemispheres for visualization purposes, but are only segmented in the visible hemisphere.

skip connections link the output of each block in the encoder to the corresponding block in the decoder branch. To reduce the number of channels in the output to the number of classes, the last block of the decoder branch contains an additional convolutional layer of kernel size 1x1x1.

In this study, we set k to five, resulting in predictions $\hat{Y}_c$ with $c \in \{CP, LPVH, CSPV, CB, background\}$. Based on initial experiments, we designed our network to have a depth of five, defined as the number of blocks in the encoder pathway and empirically set the number of feature maps in the convolutional layers to 16 in the first block of the encoder. For each subsequent block, the number of feature maps was multiplied by two in the encoder branch, and divided by two in the decoder branch. In the decoder pathway, we used nearest-neighbor interpolation for the up-sampling operation. However, although we obtained best performance with this network configuration during initial experiments, we are in this study not particularly interested in the effect of the network architecture on the segmentation task but aim to explore the optimal way of training a CNN with a small number of manual annotations.

3.3. Few-Shot Training Labels

As described before, we consider two types of ground-truth labels (1) a small set of expert annotations, obtained from individually annotated images, and (2) a large set of weak labels obtained from propagated atlas annotations. These will be referred to as *expert* (Y^{exp}) and *atlas* (Y^{atl}) labels respectively. Likewise, models trained with either expert or atlas labels are denoted by θ^{atl} and θ^{exp} . In the next sections, we describe the generation of the atlas labels as well as the manual annotation process in more detail.

3.3.1. Atlas Labels

Atlas-based labels were obtained for the images by manually annotating 3D template images as opposed to annotating individual images (Fig. 2). Template images were created by registering a subset of s rigidly aligned images $\dot{S} \in \dot{\mathcal{X}}$ to each other using a Demons diffeomorphic groupwise registration approach [46], resulting in a set of non-rigid transformations that can transform each image from the rigidly aligned orientation to a reference space R defined by: $\mathcal{T} = \{T_i : \dot{X}_i \rightarrow X_{i,R} | i = 1, \dots, s\}$. The registered images were subsequently averaged to obtain a template image $K = \frac{1}{s} \sum_{i=1}^s X_{i,R}$.

In this template K , the four subcortical structures of interest were manually annotated (see Section 3.4), resulting in a set of labelled template images. For simplicity, the annotated templates images will also be referred to with K . Subsequently, these labels were non-linearly registered back to the individual images using the inverse transform of the registration to the reference space, given by: $\mathcal{T}_{inv} = \{T_i^{-1} : K \rightarrow \dot{Y}_i | i = 1, \dots, s\}$. Due to anatomical variations between individuals and potential errors in the registration, the resulting labels may contain imperfections and are thus considered to be weak labels. To clearly differentiate between the annotated template images (which could be referred to as an atlas) and the propagated atlas-based labels, we will refer to the former only as *annotated template images* and to the latter as *atlas labels*.

Standard Template Construction. In order to generate the template images, rigidly aligned images from a GA range of 1 week were used, resulting in one

template image per GW. To capture possible differences in anatomy between the left and right cerebral hemispheres, template images for each hemisphere were constructed separately using the respective images and fused together at the midsagittal plane by joining the interhemispheric fissure in the two templates (Fig. 2 top row). The set of annotated template images (one per GW) is denoted by $\mathcal{K}$, containing annotations of all segmented structures (LPVH, CB, CP and CSPV).

Cluster-Based Template Construction. During initial experiments with the atlas labels, it was seen that the LPVH shape was not well captured by a single template image per GW. For this reason, our method was extended to generate multiple templates for this structure using a clustering approach (Fig. 2 bottom row). We selected images from an age range of two GWs, grouped them using k-medoids clustering and subsequently constructed four template images of the LPVH using this grouping. As any anatomical variation between left and right can be captured by the clustering, images of the left and right hemisphere were combined by flipping the right side images across the midline. In the resulting clustered template images (denoted by K^{clust}) only the LPVH was manually annotated, resulting in an additional set of atlas labels for the LPVH. Unless explicitly mentioned otherwise, the labels propagated from the cluster-based template images are used for the LPVH in the remainder of this study (i.e. $\mathcal{J}_{LPVH}^{atl}$ is obtained from the cluster-based template annotation).

To perform the k-medoids clustering, all images were cropped to an area around the LPVH (same bounding box dimensions across the whole GA range) and the intensities of the cropped images were normalized using histogram equalization. The size of this crop was empirically set, and was predominantly performed to reduce computation time. Next, all bounding boxes were pairwise registered to each other (using the same Demons registration used for template construction) and the sum-of-squared-differences (SSD) was computed over a region of interest (ROI) around the LPVH. This ROI was manually set for every two-week gestational window. The resulting square matrix of pairwise SSD values was subsequently used as input for the

Figure 2: Schematic overview of atlas label generation. Standard whole brain templates (top row) were constructed for each GW and all structures (CB, CP, CSPV and LPVH) were annotated in these templates. Cluster-based template construction (bottom row) was only performed for the LPVH.

394 k-medoids clustering. As k-medoids clustering can be sensitive to outliers,
 395 we excluded a few outliers based on their visual appearance prior to clus-
 396 tering. After the clustering step, each outlier was assigned to the nearest
 397 cluster based on the distance to the cluster center. Once all images were as-
 398 signed to a cluster, templates were created from the images in each cluster,
 399 including the outliers, as described earlier. The k-medoids clustering was
 400 set to repeat 500 times, and results containing clusters with less than three
 401 images were discarded.

402 3.4. Manual Annotation Process

403 Both the constructed template images, $\mathcal{K}$ for all structures and $\mathcal{K}^{clust}$ for
 404 the LPVH, as well as a subset of individual images, $\mathcal{X}$, had to be manually
 405 annotated. Since subcortical segmentation is not a task usually performed

clinically, and structural boundaries can be very challenging to distinguish in 3D US images, a bespoke segmentation protocol was defined in consultation with experienced ultrasonographers (M. Aliasi and M.C. Haak). For each structure, the 2D views in which the structural boundaries were best visible were identified (see Table 1). Based on this protocol, the manual segmentations (for both template and individual images) were performed by L.S. Hesse and verified by M. Aliasi. To illustrate the challenging nature of subcortical structural segmentation, an example annotation of the CB across the three orthogonal views can be found in the Appendix (Fig. B.12).

As described before, typically only one hemisphere is well visible in the 3D US images. For this reason, annotation of the CP and LPVH in $\mathcal{X}$ was only performed in the most visible hemisphere. However, as the whole brain template images ($\mathcal{K}$) were created for both the left and right hemisphere separately and subsequently fused together, manual annotation of these images involved annotation of the CP and LPVH in both hemispheres, thus resulting in two manual template annotations per GW for these structures. As both the CSPV and CB are structures near the midsagittal plane, these are less affected by acoustic shadowing and are segmented in full. As a result, when setting an equal n_a , defined as the number of manually annotated (template) images, between expert and atlas labels, this results in an equivalent number of annotations for the CB and CSPV, but twice as much annotation of the CP and LPVH for the atlas labels.

The annotation of the LPVH in both left and right hemisphere in the standard template images also means that the annotation effort required to annotate the LPVH in $\mathcal{K}^{clust}$ (four templates per two GWs) was the same as annotating the LPVH in $\mathcal{K}$ (one template per GW, annotating both the left and right LPVH). For consistency among the structures, for the atlas labels n_a will refer to the number of standard template images that were annotated, which is thus equal to one per GW.

All manual annotations were performed on the aligned images, $\mathcal{X}$. Labels for $\mathcal{X}$ images were generated by transforming the labels back to the unaligned space using the inverse of the rigid manual alignment transform.

Table 1: Overview of manual segmentation protocol

Structure	Segmentation Protocol	Example
<p>CP (<i>Choroid Plexus</i>)</p>	<ul style="list-style-type: none"> Identify in transventricular plane as circular echogenic structure inside ventricular cavity. Segment complete CP, which moves from the midline (posteriorly) to a more lateral location (inferior). Use all three planes, predominantly the axial and coronal, to ensure consistent segmentation.
<p>LPVH (<i>Lateral Posterior Ventricle Horn</i>)</p>	<ul style="list-style-type: none"> Ventricular cavity posterior to the CP. Identify in transventricular plane as hypoechoic (dark) cavity. Segment inside echogenic boundary. Use all three planes to ensure consistent segmentation.
<p>CSPV (<i>Cavum Septum Pellucidum et Vergae</i>)</p>	<ul style="list-style-type: none"> Identify in midsagittal plane as fluid-filled cavity. In older fetuses the cavum septum pellucidum (anterior) and cavum vergae (posterior) can be seen as two adjoining cavities but are segmented together. Use axial plane to determine lateral boundaries of CSPV.
<p>CB (<i>Cerebellum</i>)</p>	<ul style="list-style-type: none"> Include bright echogenic boundary*. Identify on axial plane as the two cerebellar hemispheres connected by the vermis. Determine superior and inferior boundaries in sagittal plane. Use all three planes to ensure consistent segmentation.

* Cerebellar boundary included for consistency as transverse cerebellar diameter is clinically measured including this boundary

439 3.5. Experimental Set-Up

440 3.5.1. Dataset and Preprocessing

441 For this study, 3D US images from the INTERGROWTH-21st Fetal
442 Growth Longitudinal Study have been used [47]. The main aim of that
443 project was to study growth, health, nutrition and neurodevelopment in a
444 multi-site ethnically diverse population. Based on a list of strict criteria
445 such as, age, BMI and medical history, healthy women were included from
446 eight carefully selected sites around the world. All images were acquired
447 from the axial plane on a Philips US machine (Philips HD-9, Philips Ul-
448 trasound, USA) with a curvilinear abdominal transducer. We used a total
449 of 537 images of which the fetuses were born without congenital malforma-
450 tions, between the age range of 18 and 26 GWs, inclusive. Furthermore,
451 it was ensured that images from both the left and right hemisphere were
452 included and that only images with sufficient US quality were included in
453 this total. In Fig. 3 the data distribution over the GA range is shown.

454 As an initial pre-processing step, all images were resampled to an isotropic
455 voxel size of 0.6 mm (using trilinear interpolation) and cropped to the same
456 dimensions of $160 \times 160 \times 160$ voxels. Furthermore, all images were man-
457 ually aligned using a rigid transformation to the same reference space, as
458 the anatomical orientation of the brain varies in the original scan due to
459 transducer and fetal position during acquisition.

460 The complete dataset consisted of 537 images, of which 259 were used for
461 model development and the remaining 278 to generate volumetric growth
462 curves of the four subcortical structures (referred to as the *analysis* subset or
463 $\mathcal{X}_{analysis}$). Of the 259 images used for model development, 20 were used for
464 testing ($\mathcal{X}_{test}$) and the remaining for training ($\mathcal{X}_{train}$) and validation ($\mathcal{X}_{val}$).
465 These 20 test set images were evenly distributed across the GA range that
466 we used, with four images for every second GW (at 18, 20, 22, 26, and
467 26 GWs). Furthermore, it was ensured that the training, validation, and
468 analysis subsets were also evenly distributed across the gestational range.
469 All testing images were manually annotated and thus had expert ground-
470 truth labels (as described in 3.4). Ten test images were twice manually
471 annotated by the same observer, in order to obtain intra-observer variability.

Figure 3: Data distribution of the 537 US images used in this study

472 To obtain atlas labels for our dataset $\mathcal{Y}^{atl}$, template images had to be
 473 generated for our dataset (K and K_{clust}). These template images were
 474 created using all images in our dataset ($\mathcal{X}_{train}, \mathcal{X}_{val}, \mathcal{X}_{test}, \mathcal{X}_{analysis}$), as these
 475 were available from earlier work [46]. However, the effect on the appearance
 476 of the template images of using all images versus only the one used for
 477 training is very small, and as such not expected to have a significant effect
 478 on the resulting atlas labels.

479 As the GA of the dataset used in this study ranges between 18 and 26, n_a
 480 for the training data was set to 9. This yielded two overlapping training sets,
 481 one with 239 images containing weak atlas labels, and one with 9 images
 482 containing expert labels. During training of our networks, 10% of the 239
 483 was used as validation. For the 9 expertly labelled individual images, no
 484 validation set was used, but all settings were kept the same as when training
 485 with atlas labels. An overview of how the data were separated into different
 486 subsets, and the number of available labels in each set is summarized in
 487 Table 2.

488 3.5.2. Experiments

489 In order to study the effect of using only a small number of manual
 490 annotations, multiple experiments were performed.

Table 2: Overview of the number of images and available labels for each of the data subsets

	volumes	atlas labels	expert labels
Analysis subset	278	0	0
Model Development subset			
Testing	20	20	20
Training	215	215	9
Validation	24	24	0

491 *Comparison between atlas and expert labels.* Firstly, we aimed to study the
492 effect of using n_a expertly annotated images versus many weakly labelled
493 images, obtained from annotating n_a template images. This first experiment
494 was performed with images from the full GA range in our data, resulting in
495 $n_a = 9$. Defining n_x as the number of training images used during training,
496 two separate networks were trained: one that used all atlas-labelled images
497 (θ^{atl} , $n_x = 215$) and a second that only used the expertly labelled images
498 (θ^{exp} , $n_x = 9$). As the alignment of the images was expected to influence
499 the performance, this experiment was repeated for the aligned images, $\dot{\mathcal{X}}$,
500 as well as for images in their original orientation, $\check{\mathcal{X}}$. Models trained with
501 aligned images were evaluated on aligned testing data, and vice-versa.

502 *Varying the number of manual annotations.* Next, n_a was reduced to study
503 the performance decrease when using even fewer manual annotations for
504 training. For $\mathcal{Y}^{exp}$, a simple subset of n_a images were used whereas for $\mathcal{Y}^{atl}$
505 all images obtained from n_a annotated templates were used. The models
506 trained with these labels are noted as $\theta_{n_a}^{exp}$ and $\theta_{n_a}^{atl}$, and in our experiments
507 n_a ranged from 2 to 9. The images used for training were selected to be
508 uniformly distributed over the whole GA range, e.g. for $n_a = 2$ the images
509 from 20 and 24 GWs were selected. Even though these models were only
510 trained with images from selected GWs, evaluation was performed on all
511 testing images. As for the previous experiment, this was repeated for both
512 $\check{\mathcal{X}}$ and $\dot{\mathcal{X}}$.

513 *Split training by GA.* We then split our training data by GA and trained
514 separate networks for each two-week gestational window. Testing was sub-

sequently only performed on the testing images of the corresponding GWs. These models trained per two-week window are referred to as θ_{2GW}^{exp} and θ_{2GW}^{atl} and are compared to the networks trained on the full GA range, θ^{exp} and θ^{atl} . The aim of this experiment was to study the effect of combining the whole GA range in a single model versus training models for smaller gestational windows. The images in the dataset used span an uneven gestational range of 9 weeks (from 18 to 26 GWs). For this reason, as θ_{2GW}^{exp} is trained using a two-week gestational window, no network was trained with the images of 26 GWs. As the effect of alignment is explored in the first two experiments, we only performed this experiment using $\mathcal{X}$.

Cluster-based atlas labels. During initial testing, we observed that using atlas labels from multiple LPVH templates, K^{clust} , increased performance compared to using the labels from a single annotated template per GW, K . For this reason, all aforementioned experiments using atlas labels were performed with the LPVH annotations from the clustered template images. To quantify the performance increase from these improved template images, we also annotated the LPVH in K , and trained a network with these labels. As this only applies to the atlas labels, no expert training labels were used in this experiment.

3.5.3. Network Training

To train our networks, we used a combination of multi-class DSC ($\mathcal{L}_{MD}$) and Cross-Entropy (CE) loss ($\mathcal{L}_{CE}$), as defined by:

$$\mathcal{L}_{total} = \mathcal{L}_{MD} + \lambda \mathcal{L}_{CE} \quad (1)$$

with λ as a relative weighting parameter between both loss terms. It was shown during initial experiments that the DSC loss term resulted in higher performance for individual structures due to the high unbalance between the structures and background, whereas the addition of the CE loss term ensured convergence for all structures. The λ was set to 1 based on initial experiments, but did not have a strong effect on performance for values within the same order of magnitude.

542 To prevent overfitting, for each sample a combination of the following
 543 geometric augmentations was performed: horizontal flips (across the midline
 544 for aligned images), rotation ($\pm 30^\circ$), translation (± 10 voxels) and scaling.
 545 The scaling range (defined by s_{min} and s_{max}) was set for each GW separately
 546 and given by:

$$\begin{aligned} s_{min}(w) &= \max \left(\sqrt[3]{\frac{0.9 \cdot \overline{V_{brain}}(18)}{\overline{v_{brain}}(w)}}, 1/1.5 \right) \\ s_{max}(w) &= \min \left(\sqrt[3]{\frac{1.1 \cdot \overline{V_{brain}}(26)}{\overline{V_{brain}}(w)}}, 1.5 \right) \end{aligned} \quad (2)$$

547 with w the GA of the fetus in weeks and $\overline{V_{brain}}(w)$ the average whole brain
 548 volume at a certain GW. This ensured that scaling was only performed to
 549 brain sizes within the dataset (i.e. images at 18 GW were predominantly
 550 up-scaled whereas images at 26 GW were mostly down-scaled). We applied
 551 augmentation for training with both $\dot{X}$ and $\check{X}$. For $\dot{X}$, this meant that the
 552 network was trained with images slightly deviating from the initial align-
 553 ment to the same coordinate system, making it more robust to imperfect
 554 alignment.

555 As the size of the training set varies strongly across experiments, all our
 556 models were trained for the same number of iterations (defined as passing
 557 a single batch of data through the network). This number was empirically
 558 set to be equivalent to training with all atlas-labelled images for 100 epochs.
 559 We chose the ADAM optimizer [48] with an initial learning rate of 0.001 for
 560 training, and used a batch size of four. This relatively small batch size was
 561 constrained by the GPU memory due to the large input images (160×160
 562 $\times 160$). The results reported in this study are the performance after the
 563 last training epoch averaged over three training runs.

564 3.5.4. Implementation

565 All models were implemented in Python 3.7 using PyTorch (version
 566 1.7.1). We trained our models on an NVIDIA Tesla V100 with 32 GB of
 567 memory. To generate the template images used to obtain our atlas labels we

568 used Matlab (version 9.8), and all manual annotation was performed using
569 the freely available MITK-Workbench.

570 3.6. Evaluation

571 3.6.1. Postprocessing

572 During visual inspection of the resulting segmentations, we noticed that
573 for some images (about 25% of the test images for $\hat{X}$ and 50% for $\check{X}$),
574 θ^{exp} made predictions that were also in the hemisphere proximal to the
575 transducer (which is partly occluded due to shadowing), whereas our ground-
576 truth only contained manual annotations of the distal hemisphere for both
577 the CP and LPVH. However, these predictions were always smaller in size
578 than the predictions in the distal hemisphere, as only part of the structure
579 was clearly visible. Furthermore, for $\check{X}$ some very small spurious areas
580 were predicted for a few test images. For these reasons, we post-processed
581 our predictions $\hat{Y}_c$ by taking the largest connected component as the final
582 prediction, denoted by $\hat{Y}_{c,post}$. As the evaluation metrics are computed both
583 with and without post-processing, for simplicity the notation $\hat{Y}_c$ is used in
584 the remainder of this section to represent either case.

585 3.6.2. Metrics

586 The predictions for each class were separately evaluated using the fol-
587 lowing metrics, where Y_c^{exp} is a binary manual expert label in the test set
588 for a single class, and $\hat{Y}_c$ the predicted binary mask for this single class:

- The Dice Similarity Coefficient (DSC), defined by:

$$DSC(Y_c^{exp}, \hat{Y}_c) = \frac{2 \cdot \sum_{i=1}^{n_v} Y_{c,i}^{exp} \cdot \hat{Y}_{c,i}}{\sum_{i=1}^{n_v} Y_{c,i}^{exp} + \sum_{i=1}^{n_v} \hat{Y}_{c,i}} \quad (3)$$

589 with n_v the total number of voxels per image. A DSC of 1 indicates
590 perfect overlap between the ground-truth and prediction, whereas a
591 DSC of 0 indicates no overlap.

- The 95th percentile Hausdorff distance (H_{95}), which is the maximum
593 (or 95th percentile in our case) over the set of boundary points of the

594 minimum distances (between a boundary point in one set and any
595 boundary point in the other set). If B_c and $\hat{B}_c$ represent the surface
596 points of Y_c and $\hat{Y}_c$ respectively, the H_{95} is given by [49]:

$$H_{95}(B_c, \hat{B}_c) = \max(h_{95}(B_c, \hat{B}_c), h_{95}(\hat{B}_c, B_c)) \quad (4)$$

where h is the directional Hausdorff distance given by:

$$h_{95}(A, B) = 95^{\text{th}} \min_{\mathbf{a} \in A} d(\mathbf{a}, \mathbf{b}) \quad (5)$$

597 with $d(\mathbf{a}, \mathbf{b})$ the Euclidean distance between points $\mathbf{a}$ and $\mathbf{b}$, and 95^{th}
598 being the 95^{th} percentile. We are using the 95^{th} percentile distance
599 because it is more robust to very small outliers than the standard
600 maximum Hausdorff distance. As the Hausdorff distance represents
601 the distance between prediction and ground-truth, lower distances in-
602 dicate better segmentation performance.

- 603 • The signed (ΔV_{rel}) and unsigned ($|\Delta V_{rel}|$) relative volume differences,
604 defined by:

$$\Delta V_{rel} = \frac{V_{\hat{Y}_c} - V_{Y_c^{exp}}}{V_{Y_c^{exp}}} = \frac{\sum_{i=1}^{n_v} \hat{Y}_{c,i} - \sum_{i=1}^{n_v} Y_{c,i}^{exp}}{\sum_{i=1}^{n_v} Y_{c,i}^{exp}} \quad (6)$$

$$|\Delta V_{rel}| = \frac{|V_{\hat{Y}_c} - V_{Y_c^{exp}}|}{V_{Y_c^{exp}}} \quad (7)$$

605 The signed difference indicates whether the model's prediction is over-
606 or under-segmented. A positive number indicates over-segmentation,
607 whereas a negative number indicates under-segmentation. On the
608 other hand, the unsigned volume difference indicates the error in the
609 volumetric measures predicted by the segmentation network. Given
610 that the segmentation networks were developed in order to extract vol-
611 umetric information of the structures, this measure is important in es-
612 timating the expected error. As the structures segmented in this study

613 vary considerably in size, the relative volume differences (with respect
614 to the ground-truth structural volume) enable comparison across the
615 structures.

616 3.7. Volumetric Subcortical Growth Trajectories

617 In order to generate growth curves for the subcortical structures, we
618 finally applied our models to the US images in our analysis subset. We
619 predicted subcortical volumes ($V_{\hat{Y}_{c,post}}$) for $\mathcal{X}_{analysis}$ using both $\dot{\theta}^{exp}$ and
620 $\dot{\theta}^{atl}$ in order to analyze the differences. Subsequently, a linear or quadratic
621 polynomial was fit to the predicted volumes as function of the GA. The
622 quadratic term was only added if it proved to be statistically significant
623 for both $\dot{\theta}^{exp}$ and $\dot{\theta}^{atl}$ (determined by a two-sided t-test), as the underlying
624 structure growth should be the same between the two networks.

625 In addition to structural volumes, we also computed the relative struc-
626 tural volume with respect to the whole brain volume, defined by:

$$V_{rel_brain} = \frac{V_{\hat{Y}_{c,post}}}{V_{brain}} \quad (8)$$

627 with V_{brain} the whole brain volume of the respective image. These whole
628 brain volumes were computed from whole brain masks derived from an MRI
629 fetal brain atlas [30]. The template images in this atlas, ranging from 21
630 to 31 GWs, were first binarized to obtain whole brain atlas masks and
631 subsequently aligned to the individual US images using an affine transform
632 (rigid + scale). For each US image, the template of the corresponding GW
633 was used, and for fetuses younger than 21 GWs, the template image of 21
634 GWs was used.

635 3.8. Ethics Statement

636 The INTERGROWTH-21st Project was approved by the Oxfordshire
637 Research Ethics Committee “C” (ref: 08/H0606/ 139), the research ethics
638 committee of the individual participating institutions and the corresponding
639 regional health authorities in which the project was implemented. Partici-
640 pants provided written consent to be involved in the project [50].

Table 3: DSC performance of our segmentation networks using aligned images compared to results of previous studies (obtained on different datasets). Numbers in between brackets indicate the standard deviation and the best performance for each structure is shown in bold. For both $\hat{\theta}^{exp}$ and $\hat{\theta}^{atl}$ results of statistical testing with respect to the propagated atlas labels are shown with (**) $p < 0.005$ and (*) $p < 0.05$. Full overview of statistical results can be found in Appendix C.

	CP	LPVH	CSPV	CB
Gutiérrez-Becker et al. [32]				0.80 (0.05)
Yaqub et al. [33] ^a	0.79 (0.09)	0.82 (0.10)	0.74 (0.11)	0.63 (0.15)
Huang et al. [34] ^b	0.76 (0.08)		0.81 (0.06)	
Venturini et al. [37]				0.73 (0.014)
This work				
prop. atlas	0.79 (0.07)	0.68 (0.10)	0.72 (0.10)	0.80 (0.09)
$\hat{\theta}^{atl}$	0.82 (0.05)**	0.77 (0.11)**	0.76 (0.07)	0.86 (0.03)*
$\hat{\theta}^{exp}$	0.85 (0.04)**	0.85** (0.05)	0.78 (0.13)	0.90 (0.02)**
Intra-observer variability	0.85 (0.05)	0.86 (0.03)	0.86 (0.04)	0.91 (0.03)

^a Search region limited to cuboid around ground-truth annotation

^b Segmentation in 2D standard planes

4. Results

All performance values presented in this section were obtained from $\hat{Y}_{c,post}$ (predictions post-processed to only keep the largest connected component). Results without post-processing are given in the Appendix (Figs. B.13 and B.14).

4.1. Weak Atlas Labels

In order to quantify the amount of inaccuracies in our weak atlas training labels, we determined the DSC overlap between the naive atlas propagated labels ($\mathcal{Y}^{atl}$) and the manual ground-truth labels ($\mathcal{Y}^{exp}$) in our test set. These results are shown in Table 3 as *prop. atlas*. For the LPVH, the DSC overlap of the atlas labels obtained from the clustered template images, K^{clust} , is shown. It can be observed that the naive propagated atlas labels

Figure 4: Resulting performance values after post-processing for θ^{exp} and θ^{atl} .

653 obtained relatively low DSC, ranging from 0.68 for the LPVH to 0.80 for
 654 the CB.

655 4.2. Comparison between atlas and expert labels

656 *Effect of image alignment.* The resulting segmentation performance of our
 657 trained networks θ^{atl} ($n_x = 239$) and θ^{exp} ($n_x = 9$) are shown in Fig. 4 and
 658 Table 3. The results shown were based on experiments that were trained
 659 and tested on images with the same alignment, either $\check{\mathcal{X}}$ or $\hat{\mathcal{X}}$. Table 3 shows
 660 the performance for the aligned images whereas in Fig. 4 results for both
 661 the aligned and unaligned settings are presented. From Fig. 4 it can be
 662 seen that higher performance is obtained for the aligned images compared
 663 to the unaligned images, which is most pronounced in the lower H_{95} for
 664 these images. For the aligned images, $\hat{\theta}^{exp}$ outperforms $\hat{\theta}^{atl}$ in terms of DSC
 665 for all structures. This trend is also seen when comparing the H_{95} , except
 666 for the CP. On the other hand, on the unaligned images both $\check{\theta}^{atl}$ and $\check{\theta}^{exp}$

Figure 5: Signed and unsigned relative volume differences after post-processing for θ^{exp} and θ^{atl} .

667 obtain similar performance across the structures.

668 *Volumetric measures.* In Fig. 5 the resulting relative volume differences
669 (ΔV_{rel} and $|\Delta V_{rel}|$) are shown. It can be observed that θ^{atl} shows an average
670 ΔV_{rel} higher than 0 (except for the CSPV in the aligned setting), corre-
671 sponding to over-segmentation. In contrast, θ^{exp} shows, on average, a slight
672 negative ΔV_{rel} , thus under-segmentation, for the unaligned images while
673 the differences are centered around zero for the aligned data. In Fig. 5 also
674 some clear outliers for both the CSPV and LPVH can be observed. However,
675 these correspond to images at 18 or 20 GW where the total ground-truth
676 volume of the respective structures is very small, resulting in large relative
677 volume differences.

678 *Comparison to previous work.* In Table 3, our resulting DSC performance
679 using aligned images is compared to the intra-observer variability and to

previous work performing subcortical structure segmentation. The intra-observability was obtained by annotating ten images from the test set twice. Furthermore, for $\dot{\theta}^{atl}$ and θ^{exp} the significance with respect to the naive propagated atlas labels is shown. The p-values shown were obtained by a repeated measures Analysis of Variance (ANOVA) for each structure, followed by post-hoc testing with a paired t-test. Reported p-values are from the post-hoc testing and underwent Bonferroni correction for the four structures, as well as for the three model comparisons (see Appendix C for full statistical results). For the CP, LPVH and CB, superior performance is obtained compared to previous work. The best performance obtained for these structures is also very close to the intra-observer variability in our test data (DSC of 0.85, 0.85 and 0.90 versus an intra-observability of 0.86, 0.85 and 0.91 for the CP, LPVH and CB, respectively). For the CSPV the DSC is 0.03 lower than previously reported DSC scores, however, it should be noted that that study only performed segmentation in 2D standard planes [34].

Qualitative Results. In Fig. 6, an example prediction is shown for an aligned image at 22 GWs. The manual-ground truth as well as the predictions from both $\dot{\theta}^{exp}$ and $\dot{\theta}^{atl}$ are shown. A qualitative inspection suggests that the predictions of the two networks closely resemble each other as well as the ground-truth. In Fig. 7, slices are shown containing the most prominent error modes in the aligned test set. To also visualize the segmented areas in the proximal hemisphere, the images in this figure are shown before any post-processing. From Fig. 7a-c, it can be observed that $\dot{\theta}^{atl}$ tends to over-segment in some areas, whereas $\dot{\theta}^{exp}$ contains some spuriously segmented areas in the proximal hemisphere. Fig. 7d shows a CSPV segmentation at 18 GW with a very low DSC (0.51).

4.3. Varying the number of manual annotations

For our second experiment, we reduced the amount of manual annotations, n_a , required to generate the training data, thus effectively reducing the number of training images. The results of this experiment are shown

Figure 6: Example prediction of a test set volume at 22 GW (randomly chosen from 5 best performing images, average DSC across structures of 0.85), visualized in the axial (a), sagittal (b) and coronal (c) plane (segmentation was performed in 3D). The structures shown are the LPVH (green), CP (orange), CSPV (yellow), and the CB (pink). Only the outer boundary of the segmentation is shown in color for visualization purposes.

in Fig. 8. It can be seen that for $\theta_{n_a}^{atl}$, only a slight decrease in performance is observed when decreasing n_a , both for the aligned (average DSC changed from 0.80 for $n_a = 9$ to 0.78 for $n_a = 9$) and unaligned images (average DSC changed from 0.78 to 0.76). For $\theta_{n_a}^{exp}$, the performance decreases are stronger with lower numbers of training images, most notably in the unaligned scenario where the average DSC decreases to 0.33 for $n_a = 2$.

4.4. Split training by GA

In Fig. 9, results are shown for networks trained separately for each two-week gestational window, $\dot{\theta}_{2GW}$, compared to having a single model for the

Figure 7: A set of examples showing the most prominent segmentation errors from $\hat{\theta}^{exp}$ and $\hat{\theta}^{atl}$. In contrast to all other results in this section, the segmentations shown are before any post-processing so that false positive segmentation in the proximal hemisphere are visible. (a) axial plane of an image from a subject at 26 GW demonstrating over-segmentation of the LPVH (green) and LPVH segmentation in the distal hemisphere (b) axial plane of an image from a subject at 22 GW demonstrating over-segmentation of the CB (pink) for $\hat{\theta}^{atl}$ and slight under-segmentation for $\hat{\theta}^{exp}$ (d) sagittal plane of an image from a subject at 18 GW, demonstrating over-segmentation for $\hat{\theta}^{atl}$ as well as the difficulty of manually segmenting the CSPV (yellow) at young ages.

720 whole age range, $\hat{\theta}$, for both the atlas and expert labels. For $\hat{\theta}_{2GW}$, each point
721 represents a single network trained using the images of the corresponding
722 GW and the subsequent GW (i.e. 18 + 19), and is tested on the four images
723 of the corresponding GW (i.e. 18), as test images are only available for every
724 second week. For $\hat{\theta}$, each point represents the performance on the four test
725 images of the respective GW.

726 It can be observed that the results for both label types, $\hat{\theta}$ and $\hat{\theta}_{2GW}$,
727 follow a similar trend. Averaged across all structures, $\hat{\theta}$ obtains a DSC of

Figure 8: Varying the number of manual annotations (n_a) used for generating the training labels.

0.81 and 0.85 for the atlas and expert labels, respectively, compared to a DSC of 0.80 and 0.82 for $\dot{\theta}_{2GW}$. For both the atlas and expert labels, $\dot{\theta}$ thus outperforms $\dot{\theta}_{2GW}$. This difference is significant ($p < 0.05$) for the CP, LPVH and CB in $\dot{\theta}^{exp}$, and only for the CP in $\dot{\theta}^{exp}$ (p-values were obtained by a paired t-test with Bonferroni correction for the four structures).

From Fig. 9, the performance of $\dot{\theta}$ across gestation can also be observed. The CP shows a decrease of DSC with advancing GA whereas the CSPV segmentation performance increases. Both the CB and LPVH show a relatively stable performance across the GA range.

4.5. Clustered LPVH Labels

In Table 4 the LPVH segmentation performance of training $\dot{\theta}^{exp}$ with atlas labels either propagated from $\mathcal{K}$ or $\mathcal{K}^{clust}$ is shown. Training a network with the propagated clustered template annotations significantly improved the DSC performance from 0.69 to 0.77 ($p < 0.001$) and H_{95} from 2.2 to 1.6

Figure 9: Performance of a single network ($\hat{\theta}$), versus training separate models for every GA range of 2 weeks ($\hat{\theta}_{2GW}$).

($p < 0.05$). Significance values were calculated using a paired t-test.

4.6. Subcortical Growth Curves

The growth curves obtained from applying our trained networks are presented in Figs. 10 and 11. In Fig. 10 the predicted structural volumes are shown, whereas in Fig. 11 the predicted relative volumes with respect to the whole brain volume (V_{rel_brain}), are shown. The growth curves were fitted with a linear or quadratic fit based on the significance of the quadratic component computed with a two-sided t-test. All parameters obtained through fitting are given in Appendix D.

To illustrate the level of consistency between $\hat{\theta}^{exp}$ and $\hat{\theta}^{atl}$, in Figs. 10 and 11 the samples were colored based on their residual value in the growth curves obtained from $\hat{\theta}^{exp}$. For a certain structure, samples with matching colors between the growth curves of the $\hat{\theta}^{exp}$ and $\hat{\theta}^{atl}$ thus correspond to the

Table 4: LPVH segmentation performance using two different types of atlas-based annotation: LPVH labels obtained from annotating a single template per week ($\mathcal{K}$) or annotating a set of clustered template images ($\mathcal{K}^{clust}$).

Network (template type)	DSC	H_{95} [mm]
$\dot{\theta}^{atl}(\mathcal{K})$	0.69 (0.13)	2.2 (1.40)
$\dot{\theta}^{atl}(\mathcal{K}^{clust})$	0.77 (0.11)	1.6 (1.2)

755 same image. The choice of coloring the samples based on $\dot{\theta}^{exp}$ was made
756 based on its better DSC performance (Table 3), but as it is only used to
757 visualize the matching images, this choice is relatively arbitrary.

758 It can be observed that both of the models, $\dot{\theta}^{exp}$ and $\dot{\theta}^{atl}$, result in
759 a similar growth trend and that the relative sample distribution is largely
760 preserved (i.e. values above or below the growth curve are respectively above
761 or below the curve for the other method, by a similar amount). However,
762 $\dot{\theta}^{atl}$ predicts larger volumes than $\dot{\theta}^{exp}$ and there are also a few outliers that
763 are not preserved between the two models.

764 As volumetric growth curves of the CB have been reported in previous
765 work [16–22], we also compared our cerebellar growth curves to these studies
766 (see Fig. B.15). This showed that our CB growth curves correspond well to
767 previously reported values for cerebellar growth.

768 From the resulting growth curves (Figs. 10 and 11), it can be observed
769 that the CB undergoes rapid growth during the second trimester of gesta-
770 tion, increasing from 1 cm³ at 18 GW to about 4 cm³ at the end of the
771 second trimester. The CSPV shows a linear increase in structural volume
772 during the studied GA range (from about 0.01 cm³ at 18GWs to 0.6 cm³ at
773 26 GWs) but its relative volume increases early in the second trimester and
774 remains constant after approximately 24 GWs. The CP only shows a very
775 small increase in structural volumes during the second trimester. However,
776 the relative volumes show a rapid decline with respect to the total brain
777 volume, ranging from 12% of the total brain volume at 18 GWs to 3% at 27
778 GWs. The LPVH structural volume increases linearly from about 0.1 cm³
779 to 0.3 cm³ in the studied GA range, whereas the relative volume shows a
780 small downward trend.

Figure 10: Estimated structural volumes for subcortical structures as a function of GA. Volumes were fitted with a linear or quadratic fit (black), in which the quadratic term was only added if it was significant for both networks (per structure). The 95% prediction confidence intervals were also computed and are shown with red dashed lines. For each structure, samples were colored based on their residual for $\hat{\theta}^{exp}$, and the same colors (per sample) were used for $\hat{\theta}^{atl}$.

Figure 11: Estimated relative volumes (Rel. Vol.) with respect to the whole brain volumes (V_{rel_brain}) for subcortical structures as a function of GA. Volumes were fitted with a linear or quadratic fit (black), in which the quadratic term was only added if it was significant for both models (per structure). The 95% prediction confidence intervals were also computed and are shown with red dashed lines. For each structure, samples were colored based on their residual for θ^{exp} , and the same colors (per sample) were used for θ^{atl} .

781 5. Discussion

782 In this work, we performed subcortical segmentation in 3D fetal brain
783 US images during the second trimester of gestation in a low-data regime.
784 Our results showed that segmentation performance close to intra-observer
785 variability can be obtained using only nine manually annotated brain images
786 for training. We obtained best DSC performance of 0.85, 0.85, 0.78 and
787 0.90 for the CP, LPVH, CSPV and CB respectively, compared to an intra-
788 observer variability of 0.85, 0.86, 0.86 and 0.91. We also applied our trained
789 networks to generate longitudinal growth curves for subcortical structure
790 development, that are, to the best of our knowledge, the first US-specific
791 volumetric curves for the CP, LPVH and CSPV.

792 In our experiments, we investigated the impact of the number of labelled
793 training samples, particularly assessing whether a network can be trained
794 for structural segmentation in a low-data domain. We compared using a
795 relatively large number (215) of images with weak labels (atlas labels) ver-
796 sus a very small number (9) of images with expert labels for training. For
797 both settings, the number of images that had to be manually annotated to
798 obtain the training labels was kept the same. Furthermore, we repeated our
799 experiments for both *aligned* and *unaligned* images to explore the sensitivity
800 of the model to brain pose. Somewhat surprisingly, in the aligned setting we
801 obtained best performance using only 9 expertly annotated images, as op-
802 posed to a much larger number of weakly annotated images. This indicates
803 that even though segmentation in 3D US by a human annotator is a very
804 challenging task, a network is able to learn the correct features from only a
805 small set of manual annotations, provided that the images are all aligned to
806 the same coordinate space.

807 The network trained with expert labels showed a notably lower perfor-
808 mance when trained with *unaligned* images than with the *aligned* images,
809 which is most pronounced when decreasing the number of manual annota-
810 tions used for training below nine (Fig. 8). On the other hand, this difference
811 was not as pronounced for the network trained with atlas labels. This can
812 be explained by the fact that in the *unaligned* setting the segmentation task

consists of a combination of structure localization and subsequent segmentation. The weak, atlas labels contain imperfections at the boundaries of the segmentations, but are always approximately in the correct location. For this reason, these labels provide more information for the localization task than the few expertly annotated images and, as such, obtain better segmentation performance in the unaligned case. The few expertly annotated images in this setting are not sufficient to provide both localization and segmentation, which can most clearly be deduced from the large H_{95} distances (Fig. 4). These results thus suggest that a large amount of weak labels is better for localization, whereas few expert labels perform better if the task consists of solely delineating a structure in approximately the same location.

For the weak atlas labels, we observed only a small performance decrease when reducing the number of training images. Even by training with labels generated from just 2 GWs (coming from 2 manual annotations) the performance is close to the maximum performance (average DSC of 0.78 versus maximum DSC of 0.80 for the aligned case). The same pattern is observed for both *aligned* and *unaligned* data, confirming that the alignment does not considerably affect segmentation performance for the atlas labels. On the other hand, for the expert labels a performance decrease was observed when reducing the number of training images, showing a stronger decrease for the *unaligned* images than for the *aligned* images. As in this experiment we used as little as two images for training, while evaluating on images from the whole GA range, this drop in performance was in line with our expectations.

We also compared our resulting segmentation performance in the aligned setting to previous studies. However, it should be noted that performance metrics reported in different studies have to be carefully interpreted. Factors such as US image quality, structural definition (i.e. what is included in the segmentation), and GA range can all affect the difficulty of the task. The GA range affects the difficulty of the segmentation tasks as acoustic shadowing increases with advancing GA, due to calcification of the fetal skull. Furthermore, overlap measures, such as the DSC, generally report higher values for larger structures because a single (erroneous) voxel has a smaller

effect on the resulting DSC value for these structures. We also want to stress that small performance increases are very difficult to quantify in 3D US segmentation due to the high intra- and inter-observer variability predominantly resulting from the ambiguity of the exact location of structural boundaries.

For the CP, LPVH, and CB we reported increased performance compared to previous studies (resulting in a DSC of 0.85, 0.85, and 0.90, respectively). An especially large increase was observed for the CB (DSC of 0.90 versus DSC of 0.80, 0.63 and 0.73 in previous studies). However, part of this increase can be attributed to the fact that in this study the CB was manually annotated including the bright echogenic boundaries, thus resulting in a larger segmented region, whereas this boundary was not included in [51] and [33]. This choice was made based on the fact that the transverse cerebellar diameter, which is clinically used to assess growth during a standard fetal examination, also includes this boundary [52]. Furthermore, manual segmentation is more consistent when including this boundary as the outside edge is generally clearly visible in the US images. For the CSPV, slightly higher performance (DSC of 0.81 versus 0.78) was reported in Huang et al. [34], but was obtained in the 2D midsagittal plane. Since the lateral boundaries of the CSPV are hardest to define in 3D images, this naturally results in larger prediction errors than 2D segmentations in this plane. Although these comparisons should thus be interpreted with care, it does show that we obtained competitive segmentation performance for all structures.

As the aim of developing a 3D segmentation model is to extract volumetric information, we also assessed the error of the resulting volume measurements in our test set (Fig. 5). These showed that the network trained with atlas labels generally produced positive volume differences, representing over-segmentation, indicating a tendency to segment a structure larger on the atlas than in the individual images. However, as the visualization of structural boundaries can be subjective in US, a consistent (albeit overestimating) segmentation prediction cannot be considered unacceptable.

We further analyzed our segmentation performance in the *aligned* setting by visualizing the most prominent errors before any post-processing (Fig. 7).

879 These results confirmed that the network trained with atlas labels showed
880 a tendency to over-segment, whereas the network trained with expert labels
881 resulted in some small segmented regions in the proximal hemisphere (as
882 well as occasionally in other areas for the unaligned images). However, these
883 small regions were easily removed by only retaining the largest connected
884 component during post-processing. Furthermore, due to their small size,
885 these regions

886 We thus believe that even though being trained with only a small number
887 of manual annotations, both networks (trained with atlas or expert labels)
888 perform exceptionally well on a task that is challenging for human annota-
889 tors.

890 To observe the effect of training a single model for the entire GA range
891 that was used in this study (18-26 weeks), we also trained separate networks
892 for every biweekly GA window. These results showed that for most struc-
893 tures and GWs, the best performance is obtained by training a single model
894 with data spanning the entire GA range. This confirms our expectation that
895 information from all GWs provides longitudinal structural variability that
896 improves the model's representational power and, thus, segmentation per-
897 formance. However, the differences in performance between both methods
898 were small (DSC of 0.81 and 0.85 for the biweekly model versus a DSC of
899 0.80 and 0.82 for the single model for the atlas and expert labels, respec-
900 tively), suggesting that models for a smaller GA window can also provide
901 good performance.

902 In addition to comparing separately trained networks with a single model,
903 Fig. 9 also shows the DSC performance as a function of the GW. The CSPV
904 performance was low at 18 GWs and increased later in gestation, which
905 matches our expectation that segmentation performance is better for larger
906 structures, as the CSPV is very small (around 0.05 cm^2 at 18 GWs versus
907 0.5 cm^2 at 26 GWs) and therefore difficult to manually segment at 18 GW
908 (Fig. 7d). The CP shows the opposite behavior, which can be attributed to
909 the fact that the shape of the CP is more challenging to segment later in ges-
910 tation but does not increase in size. Similarly, the CB is increasingly difficult
911 to segment during gestation due to acoustic shadowing caused by increasing

912 calcification of the temporal bone. However, as it also grows in size (from an
913 average of 1 cm³ at 18 GW to 3.5 cm³ at 26 GW), the DSC performance is
914 relatively stable across the GA range explored in this study. For the LPVH,
915 both difficulty and size do not change drastically during gestation resulting
916 in a consistent performance across the gestational period.

917 We extended our atlas-based labels with a clustering approach for the
918 LPVH. This approach was only applied for the LPVH, as the other struc-
919 tures did not display as much shape variation within GWs. By annotat-
920 ing the clustered templates instead of a single template per GW, the DSC
921 performance increased from 0.69 to 0.77 ($p < 0.001$), suggesting that the
922 additional variation captured by the clustered templates contributed to an
923 improved segmentation of the LPVH.

924 Finally, we applied our trained networks to extract volumetric infor-
925 mation from a large number of images ($n = 278$), allowing us to generate
926 US-specific growth curves for a healthy population. Previous work on volu-
927 metric subcortical development during gestation from US has largely been
928 limited to cerebellar and thalamic growth [17, 39–41]. Therefore, to the best
929 of our knowledge, this study provides novel US-specific volumetric growth
930 curves of the CSPV, CP and LPVH for a geographically diverse, healthy
931 population. We generated growth curves using the network trained with few
932 expert labels as well as the network trained with atlas labels. This showed
933 that between both networks the sample distribution was largely preserved,
934 confirming that even though the network trained with atlas labels obtained
935 lower performance due to over-segmentation, the predictions are largely con-
936 sistent with the network trained with expert labels. Overall, this shows that
937 both models can potentially be used to predict accurate subcortical vol-
938 umes, however, care should be taken when comparing volumetric measures
939 obtained from different models.

940 To validate the volumetric measures obtained in this study, we compared
941 the growth curves of the structural CB volume to previously reported CB
942 growth curves. This showed that our curves showed excellent agreement
943 with past US studies, especially in the first half of the GA range. After
944 24 GWs the growth rates of the CB (slope of the growth curves) found in

945 this study are slightly lower than in previous work. However, as described
946 previously, the CB becomes more challenging to segment with advancing GA
947 and, as such, this could explain the slightly deviating volume measurements
948 at the end of the second trimester.

949 For the CB a rapid increase in volume was observed, whereas the relative
950 volume, with respect to the total brain volume, remained mostly consistent
951 (between 1.5% and 2.5 %). As the transverse cerebellar diameter is ex-
952 pected to increase linearly during the second trimester [1], this matches the
953 quadratic growth of CB volume. The CP showed a rapid decrease in rela-
954 tive volume during the studied GA range. This aligns with the observation
955 that early in the second trimester the CP almost completely fills the lat-
956 eral ventricles and comprises a large part of the fetal brain, whereas later
957 in gestation it appears as a small structure only partly filling the lateral
958 ventricles. Structural volumes of the LPVH only showed a small increase
959 during the second trimester, which is in agreement with the fact that the
960 atrial width remains stable in healthy fetuses during the second trimester
961 [1]. The structural CSPV volumes increase linearly, whereas the relative
962 volumes show an increase in the first half of the studied GA range. As a de-
963 viating measurement of the CSPV has been related to agenesis of the corpus
964 callosum [3] but is not widely studied, the presented growth curves might
965 facilitate this in future work.

966 6. Conclusion

967 In summary, we showed that only a small number of annotated images
968 are needed to successfully train a network for subcortical segmentation in
969 3D US images. To obtain optimal performance, alignment of the images
970 is required beforehand, however, on unaligned testing images, a high per-
971 formance can still be achieved using several weakly annotated images. By
972 applying our trained networks to a large cohort of fetuses, we were able
973 to generate US-specific growth trajectories of the CP, LPVH, CSPV and
974 CB for a healthy population. This study thus demonstrates the feasibility
975 of subcortical segmentation in 3D US using deep learning, and shows that

volumetric measures obtained from these models can be used to obtain an improved understanding of subcortical growth during gestation. The models developed in this study offer the promise of comparing subcortical development between different fetal cohorts without doing time-consuming manual annotations in future studies. However, more work is needed to extensively validate the segmentation performance for datasets acquired with different US scanners as well as for cohorts containing pathological conditions.

7. Data and code availability statement

The image data is available from the INTERGROWTH-21st Consortium upon reasonable request. All network training code as well as the weights of the trained networks can be requested by emailing the corresponding author. The groupwise registration algorithm used to generate the template image in this work was adopted from [46], and requests for this code can be addressed to the corresponding author of that paper.

8. CRediT authorship contribution statement

Linde Hesse: Conceptualization, Methodology, Investigation, Data Curation, Writing – original draft, Writing - review & editing, Software, Validation. Moska Aliasi: Data Curation, Writing – review & editing. Felipe Moser: Software, Writing - review & editing. Monique Haak: Writing – review & editing. Weidi Xie: Supervision, Writing - review & editing. Mark Jenkinson: Supervision, Writing - review & editing. Ana Namburete: Supervision, Conceptualization, Writing – Review & Editing, Funding Acquisition, Data Curation.

9. Acknowledgements

LH acknowledges the support of the UK Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Award. FM acknowledges the support and funding from the Engineering and Physical Sciences Research Council (EPSRC) and Medical Research Council (MRC)

(EP/L016052/1), as well as the support from University College Oxford and its Oxford-Radcliffe benefaction. WX is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) Programme Grant Seebibyte (EP/M013774/1) and Grant Visual AI (EP/T028572/1). MJ is supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC), and this research was funded by the Wellcome Trust [215573/Z/19/Z]. The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust [203139/Z/16/Z]. AN is grateful for support from the UK Royal Academy of Engineering under the Engineering for Development Research Fellowships scheme, and to St Hilda's College, Oxford.

We would also like to acknowledge the INTERGROWTH-21st Consortium for collecting the data and making it available to us.

References

- [1] G. Malinger, D. Paladini, K. Haratz, A. Monteagudo, G. Pilu, I. Timor-Tritsch, Isuog practice guidelines (updated): sonographic examination of the fetal central nervous system. part 1: performance of screening examination and indications for targeted neurosonography, *Ultrasound in Obstetrics & Gynecology* 56 (2020) 476–484.
- [2] O. Shen, A. Gelot, M. Moutard, J. Jouannic, H. Sela, C. Garel, Abnormal shape of the cavum septi pellucidi: an indirect sign of partial agenesis of the corpus callosum, *Ultrasound in Obstetrics & Gynecology* 46 (2015) 595–599.
- [3] K. Karl, T. Esser, K. Heling, R. Chaoui, Cavum septi pellucidi (csp) ratio: a marker for partial agenesis of the fetal corpus callosum, *Ultrasound in Obstetrics & Gynecology* 50 (2017) 336–341.
- [4] D. Paladini, G. Malinger, R. Birnbaum, A. Monteagudo, G. Pilu, L. Salomon, I. Timor-Tritsch, Isuog practice guidelines (updated): sonographic examination of the fetal central nervous system. part 2: performance of targeted neurosonography., *Ultrasound in Obstetrics &*

- 1034 Gynecology: the Official Journal of the International Society of Ultra-
1035 sound in Obstetrics and Gynecology (2021).
- 1036 [5] S. Shinohara, Y. Okuda, S. Hirata, K. Suzuki, Predictive possibility of
1037 the transverse cerebellar diameter to abdominal circumference ratio for
1038 small-for-gestational-age fetus suspected as a cause of maternal placen-
1039 tal syndromes: a retrospective cohort study, *Hypertension in pregnancy*
1040 39 (2020) 145–151.
- 1041 [6] T. Tongsong, C. Wanapirak, T. Thongpadungroj, Sonographic diagno-
1042 sis of intrauterine growth restriction (iugr) by fetal transverse cerebellar
1043 diameter (tcd)/abdominal circumference (ac) ratio, *International Jour-
1044 nal of Gynecology & Obstetrics* 66 (1999) 1–5.
- 1045 [7] F. Moser, R. Huang, A. T. Papageorgiou, B. W. Papież, A. I. Nambu-
1046 rete, Automated fetal brain extraction from clinical ultrasound volumes
1047 using 3d convolutional neural networks, in: *Annual Conference on Med-
1048 ical Image Understanding and Analysis*, Springer, 2019, pp. 151–163.
- 1049 [8] M. K. Wyburd, M. Jenkinson, A. I. Namburete, Cortical plate seg-
1050 mentation using cnns in 3d fetal ultrasound, in: *Annual Conference on
1051 Medical Image Understanding and Analysis*, Springer, 2020, pp. 56–68.
- 1052 [9] L. S. Hesse, A. I. Namburete, Improving u-net segmentation with ac-
1053 tive contour based label correction, in: *Annual Conference on Medical
1054 Image Understanding and Analysis*, Springer, 2020, pp. 69–81.
- 1055 [10] A. G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, C. Wachinger, ‘squeeze
1056 & excite’guided few-shot segmentation of volumetric images, *Medical
1057 image analysis* 59 (2020) 101587.
- 1058 [11] A. K. Mondal, J. Dolz, C. Desrosiers, Few-shot 3d multi-modal med-
1059 ical image segmentation using generative adversarial learning, *arXiv
1060 preprint arXiv:1810.12241* (2018).
- 1061 [12] D. Al Chanti, V. G. Duque, M. Crouzier, A. Nordez, L. Lacourpaille,
1062 D. Mateus, Ifss-net: Interactive few-shot siamese network for faster

- 1063 muscle segmentation and propagation in volumetric ultrasound, IEEE
1064 Transactions on Medical Imaging (2021).
- 1065 [13] A. I. Namburete, W. Xie, M. Yaqub, A. Zisserman, J. A. Noble, Fully-
1066 automated alignment of 3d fetal brain ultrasound to a canonical refer-
1067 ence space using multi-task learning, Medical image analysis 46 (2018)
1068 1–14.
- 1069 [14] R. Cuingnet, O. Somphone, B. Mory, R. Prevost, M. Yaqub, R. Napoli-
1070 tano, A. Papageorgiou, D. Roundhill, J. A. Noble, R. Ardon, Where is
1071 my baby? a fast fetal head auto-alignment in 3d-ultrasound, in: 2013
1072 IEEE 10th International Symposium on Biomedical Imaging, IEEE,
1073 2013, pp. 768–771.
- 1074 [15] National Institute for Health and Care Excellence, Antenatal care for
1075 uncomplicated pregnancies (NICE Guideline CG62), 2008. Updated
1076 February 2019.
- 1077 [16] V. Kyriakopoulou, D. Vatansever, A. Davidson, P. Patkee, S. Elkom-
1078 mos, A. Chew, M. Martinez-Biarge, B. Hagberg, M. Damodaram,
1079 J. Allsop, et al., Normative biometry of the fetal brain using mag-
1080 netic resonance imaging, Brain Structure and Function 222 (2017)
1081 2295–2307.
- 1082 [17] G. Babucci, K. Rosen, B. Cappuccini, G. Clerici, 3d evaluation of fetal
1083 brain structures: reference values and growth curves, The Journal of
1084 Maternal-Fetal & Neonatal Medicine (2019) 1–6.
- 1085 [18] C.-H. Chang, F.-M. Chang, C.-H. Yu, H.-C. Ko, H.-Y. Chen, As-
1086 sessment of fetal cerebellar volume using three-dimensional ultrasound,
1087 Ultrasound in medicine & biology 26 (2000) 981–988.
- 1088 [19] N. N. Andescavage, A. du Plessis, R. McCarter, A. Serag, I. Evangelou,
1089 G. Vezina, R. Robertson, C. Limperopoulos, Complex trajectories of
1090 brain development in the healthy human fetus, Cerebral Cortex 27
1091 (2017) 5274–5283.

- 1092 [20] J. A. Scott, K. S. Hamzelou, V. Rajagopalan, P. A. Habas, K. Kim,
1093 A. J. Barkovich, O. A. Glenn, C. Studholme, 3d morphometric analysis
1094 of human fetal cerebellar development, *The Cerebellum* 11 (2012) 761–
1095 770.
- 1096 [21] T. Hata, A. Kuno, S.-Y. Dai, E. Inubashiri, U. Hanaoka, K. Kanenishi,
1097 C. Yamashiro, H. Tanaka, T. Yanagihara, Three-dimensional sono-
1098 graphic volume measurement of the fetal cerebellum, *Journal of Medi-
1099 cal Ultrasonics* 34 (2007) 17–21.
- 1100 [22] M. R. Hatab, S. W. Kamourieh, D. M. Twickler, Mr volume of the
1101 fetal cerebellum in relation to growth, *Journal of Magnetic Resonance
1102 Imaging: An Official Journal of the International Society for Magnetic
1103 Resonance in Medicine* 27 (2008) 840–845.
- 1104 [23] B. Fischl, Freesurfer, *Neuroimage* 62 (2012) 774–781.
- 1105 [24] J. Dolz, C. Desrosiers, I. B. Ayed, 3d fully convolutional networks for
1106 subcortical segmentation in mri: A large-scale study, *NeuroImage* 170
1107 (2018) 456–470.
- 1108 [25] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, B. J. Erickson,
1109 Deep learning for brain mri segmentation: state of the art and future
1110 directions, *Journal of digital imaging* 30 (2017) 449–459.
- 1111 [26] M. Dremmen, R. Bouhuis, L. Blanken, R. Muetzel, M. Vernooij,
1112 H. Marroun, V. Jaddoe, F. Verhulst, H. Tiemeier, T. White, Cavum
1113 septum pellucidum in the general pediatric population and its relation
1114 to surrounding brain structure volumes, cognitive function, and emo-
1115 tional or behavioral problems, *American Journal of Neuroradiology* 40
1116 (2019) 340–346.
- 1117 [27] A. Makropoulos, S. J. Counsell, D. Rueckert, A review on automatic
1118 fetal and neonatal brain mri segmentation, *NeuroImage* 170 (2018)
1119 231–248.

- 1120 [28] N. Khalili, N. Lessmann, E. Turk, N. Claessens, R. de Heus, T. Kolk,
1121 M. A. Viergever, M. J. Benders, I. Išgum, Automatic brain tissue seg-
1122 mentation in fetal mri using convolutional neural networks, *Magnetic*
1123 *resonance imaging* 64 (2019) 77–89.
- 1124 [29] G. Sanroma, O. M. Benkarim, G. Piella, K. Lekadir, N. Hahner,
1125 E. Eixarch, M. A. G. Ballester, Learning to combine complementary
1126 segmentation methods for fetal and 6-month infant brain mri segmen-
1127 tation, *Computerized Medical Imaging and Graphics* 69 (2018) 52–59.
- 1128 [30] A. Gholipour, C. K. Rollins, C. Velasco-Annis, A. Ouaalam,
1129 A. Akhondi-Asl, O. Afacan, C. M. Ortinau, S. Clancy, C. Limperopou-
1130 los, E. Yang, et al., A normative spatiotemporal mri atlas of the fetal
1131 brain for automatic segmentation and analysis of early brain growth,
1132 *Scientific reports* 7 (2017) 1–13.
- 1133 [31] K. Payette, P. de Dumast, H. Kebiri, I. Ezhov, J. C. Paetzold, S. Shit,
1134 A. Iqbal, R. Khan, R. Kottke, P. Grehten, et al., An automatic multi-
1135 tissue human fetal brain segmentation benchmark using the fetal tissue
1136 annotation dataset, *Scientific Data* 8 (2021) 1–14.
- 1137 [32] B. Gutiérrez-Becker, F. A. Cosío, M. E. G. Huerta, J. A. Benavides-
1138 Serralde, L. Camargo-Marín, V. M. Bañuelos, Automatic segmentation
1139 of the fetal cerebellum on ultrasound volumes, using a 3d statistical
1140 shape model, *Medical & biological engineering & computing* 51 (2013)
1141 1021–1030.
- 1142 [33] M. Yaqub, R. Cuingnet, R. Napolitano, D. Roundhill, A. Papa-
1143 georgiou, R. Ardon, J. A. Noble, Volumetric segmentation of key
1144 fetal brain structures in 3d ultrasound, in: *International Workshop on*
1145 *Machine Learning in Medical Imaging*, Springer, 2013, pp. 25–32.
- 1146 [34] R. Huang, A. Namburete, A. Noble, Learning to segment key clinical
1147 anatomical structures in fetal neurosonography informed by a region-
1148 based descriptor, *Journal of Medical Imaging* 5 (2018) 014007.

- 1149 [35] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks
1150 for biomedical image segmentation, in: International Conference on
1151 Medical image computing and computer-assisted intervention, Springer,
1152 2015, pp. 234–241.
- 1153 [36] R. Huang, W. Xie, J. A. Noble, Vp-nets: Efficient automatic localiza-
1154 tion of key brain structures in 3d fetal neurosonography, Medical image
1155 analysis 47 (2018) 127–139.
- 1156 [37] L. Venturini, A. T. Papageorgiou, J. A. Noble, A. I. Namburete, Multi-
1157 task cnn for structural semantic segmentation in 3d fetal brain ultra-
1158 sound, in: Annual Conference on Medical Image Understanding and
1159 Analysis, Springer, 2019, pp. 164–173.
- 1160 [38] Y. Wu, K. Shen, Z. Chen, J. Wu, Automatic measurement of fetal
1161 cavum septum pellucidum from ultrasound images using deep attention
1162 network, in: 2020 IEEE International Conference on Image Processing
1163 (ICIP), IEEE, 2020, pp. 2511–2515.
- 1164 [39] S. Zeng, Q. Zhou, J. Zhou, M. Li, C. Long, Q. Peng, Volume of in-
1165 tracranial structures on three-dimensional ultrasound in fetuses with
1166 congenital heart disease, Ultrasound in Obstetrics & Gynecology 46
1167 (2015) 174–181.
- 1168 [40] A. Benavides-Serralde, E. Hernández-Andrade, J. Fernández-Delgado,
1169 W. Plasencia, M. Scheier, F. Crispi, F. Figueras, K. H. Nicolaides,
1170 E. Gratacós, Three-dimensional sonographic calculation of the vol-
1171 ume of intracranial structures in growth-restricted and appropriate-for-
1172 gestational age fetuses, Ultrasound in Obstetrics and Gynecology: The
1173 Official Journal of the International Society of Ultrasound in Obstetrics
1174 and Gynecology 33 (2009) 530–537.
- 1175 [41] A. Sotiriadis, I. Dimitrakopoulos, M. Eleftheriades, T. Agorastos,
1176 G. Makrydimas, Thalamic volume measurement in normal fetuses us-
1177 ing three-dimensional sonography, Journal of Clinical Ultrasound 40
1178 (2012) 207–213.

- 1179 [42] J. A. Scott, P. A. Habas, K. Kim, V. Rajagopalan, K. S. Hamzelou,
1180 J. M. Corbett-Detig, A. J. Barkovich, O. A. Glenn, C. Studholme,
1181 Growth trajectories of the human fetal brain tissues estimated from
1182 3d reconstructed in utero mri, *International Journal of Developmental*
1183 *Neuroscience* 29 (2011) 529–536.
- 1184 [43] M. Rutten, L. Pistorius, E. Mulder, P. Stoutenbeek, L. De Vries,
1185 G. Visser, Fetal cerebellar volume and symmetry on 3-d ultrasound:
1186 volume measurement with multiplanar and vocal techniques, *Ultra-*
1187 *sound in medicine & biology* 35 (2009) 1284–1289.
- 1188 [44] E. A. Júnior, H. A. Guimarães Filho, C. R. Pires, L. M. Nardozza,
1189 A. F. Moron, R. Mattar, Validation of fetal cerebellar volume by three-
1190 dimensional ultrasonography in brazilian population, *Archives of gyne-*
1191 *cology and obstetrics* 275 (2007) 5–11.
- 1192 [45] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network
1193 training by reducing internal covariate shift, in: *International confer-*
1194 *ence on machine learning*, PMLR, 2015, pp. 448–456.
- 1195 [46] A. I. Namburete, R. van Kampen, A. T. Papageorghiou, B. W. Pa-
1196 pież, Multi-channel groupwise registration to construct an ultrasound-
1197 specific fetal brain atlas, in: *Data Driven Treatment Response Assess-*
1198 *ment and Preterm, Perinatal, and Paediatric Image Analysis*, Springer,
1199 2018, pp. 76–86.
- 1200 [47] A. T. Papageorghiou, E. O. Ohuma, D. G. Altman, T. Todros, L. C.
1201 Ismail, A. Lambert, Y. A. Jaffer, E. Bertino, M. G. Gravett, M. Pur-
1202 war, et al., International standards for fetal growth based on serial
1203 ultrasound measurements: the fetal growth longitudinal study of the
1204 INTERGROWTH-21st project, *The Lancet* 384 (2014) 869–879.
- 1205 [48] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization,
1206 arXiv preprint arXiv:1412.6980 (2014).
- 1207 [49] D. P. Huttenlocher, G. A. Klanderman, W. J. Rucklidge, Com-
1208 paring images using the hausdorff distance, *IEEE Transactions*

- 1209 on Pattern Analysis and Machine Intelligence 15 (1993) 850–863.
1210 doi:10.1109/34.232073.
- 1211 [50] J. Villar, D. Altman, M. Purwar, J. Noble, H. Knight, P. Ruyan,
1212 L. Cheikh Ismail, F. Barros, A. Lambert, A. Papageorghiou, et al., The
1213 objectives, design and implementation of the intergrowth-21st project,
1214 BJOG: An International Journal of Obstetrics & Gynaecology 120
1215 (2013) 9–26.
- 1216 [51] L. Venturini, A. T. Papageorghiou, J. A. Noble, A. I. Namburete, Un-
1217 certainty estimates as data selection criteria to boost omni-supervised
1218 learning, in: International Conference on Medical Image Computing
1219 and Computer-Assisted Intervention, Springer, 2020, pp. 689–698.
- 1220 [52] D. Prayer, G. Malingier, P. Brugger, C. Cassady, L. De Catte,
1221 B. De Keersmaecker, G. Fernandes, P. Glanc, L. Gonçalves, G. Gru-
1222 ber, et al., Isuog practice guidelines: performance of fetal magnetic
1223 resonance imaging, Ultrasound in Obstetrics & Gynecology 49 (2017)
1224 671–680.

1225 Appendix A. List of Notations

Table A.5: List of notations used in this work.

Symbol	Description
X	3D US image
$\dot{X}$	3D US image, rigidly aligned to the same coordinate system
$\check{X}$	3D US image in the original acquired orientation
Y	multi-label segmentation mask
$\hat{Y}$	Predicted multi-label segmentation mask
$Y_c, \hat{Y}_c$	Binary segmentation mask of a single class c
Y^{type}	Atlas ($type = atl$) or expert ($type = exp$) labels
$\hat{Y}_{c,post}$	Predicted multi-label segmentation mask of a single class, post-processed with largest connected component
θ^{type}	Network trained with expert or atlas labels
$\dot{\theta}^{type}, \check{\theta}^{type}$	Networks trained with $\dot{X}$ and $\check{X}$, respectively
n_a	Number of annotated (template) images
n_x	Number of images used for training
n_v	Number of voxels per image
k	Number of channels in the output prediction of the network
$\theta_{n_a}^{type}$	Network trained with labels from n_a manually annotated (template) images
θ_{2GW}^{type}	Refers to networks that were trained per two-week gestational window
T	Non-rigid transformation to transform an image to a reference space
K	(Annotated) template image created using the standard template construction
K^{clust}	(Annotated) template image created using a clustering approach
$\mathcal{X}, \mathcal{Y}, \mathcal{K}, \mathcal{T}$	Sets of X , Y , K and T , respectively, following all conventions as described above
$\mathcal{S}$	Subset of $\mathcal{X}$
$\overline{V_{brain}}(w)$	Average whole brain volume (in the dataset used) at gestational week w
V_{brain}	Whole brain volume of an individual image
$B_c, \hat{B}_c$	Surface points of Y_c and $\hat{Y}_c$, respectively.
ΔV_{rel}	Signed relative volume difference between ground-truth and network prediction (with respect to the ground-truth structure size)
$ \Delta V_{rel} $	Unsigned relative volume difference between ground-truth and network prediction (with respect to the ground-truth structure size)
$V_{rel,brain}$	Relative structural volume (with respect to the whole brain volume)

1226 Appendix B. Additional Figures

Figure B.12: Example of a challenging manual annotation for the CB. The three columns each indicate one of the orthogonal 2D views. The top row shows the image without the annotation, and in the bottom row the outer boundaries of the manual labels are shown in pink on top of the image. It can be observed that for this image in both the axial and coronal plane the cerebellar boundaries are very hard to distinguish, whereas on the sagittal view the bean shaped CB is better visible.

Figure B.13: Resulting performance values before the largest connected component post-processing step was applied for θ^{atl} and θ^{atl} .

Figure B.14: Volume-based performance values before the largest connected component post-processing step was applied for both θ^{atl} and θ^{atl} .

Figure B.15: Comparison of growth curves obtained for the CB in this study to previous work. In the left panel the CB growth curves from our study ($\dot{\theta}^{exp}$ (blue) and $\dot{\theta}^{atl}$ (orange)) are compared to other studies reporting cerebellar growth curves from US: Babucci [17] (green), Hata [21] (dark orange), Chang [18] (pink), and the right panel compares our curves to previously reported MRI growth curves: Scott [20] (green), Andescava [19] (dark orange), Hatab [22] (pink), and Kyriakopoulou [16] (light blue).

1227 Appendix C. Statistical Testing

Table C.6: P-values obtained from comparing the DSC segmentation performance of both networks, $\hat{\theta}^{atl}$ and $\hat{\theta}^{exp}$, with each other as well as with the naive propagated atlas masks (*prop. atlas*). For each structure a repeated measures ANOVA was performed followed by post-hoc testing with a paired t-test. Reported p-values are post-hoc tests corrected with Bonferonni correction for the four structures, as well as for the three model comparisons. For the CSPV the ANOVA returned non-significant differences ($p=0.08$), and as such the post-hoc tests are not reported.

	prop. atlas vs $\hat{\theta}^{atl}$	atlas vs $\hat{\theta}^{exp}$	$\hat{\theta}^{atl}$ vs $\hat{\theta}^{exp}$
CP	<0.005	<0.001	<0.001
LPVH	<0.001	<0.001	<0.01
CSPV	-	-	-
CB	<0.05	<0.005	<0.001

1228 Appendix D. Growth Trajectories

Table D.7: Parameters resulting from the fitting of the structural volumes as function of the gestational age ($V = bias + w \cdot x1 + w^2 \cdot x2$, with w the age given in weeks) for the different subcortical structures.

	model	bias	x1	x2	significance x2
CP	expert	0.7008	0.015	-0.0054	$p < 0.001$
	atlas	0.8204	0.0125	0.0024	$p = 0.065$
LPVH	expert	0.2072	0.0241	-0.0004	$p = 0.79$
	atlas	0.2199	0.0354	0.0016	$p = 0.085$
CSPV	expert	0.2857	0.0668	0.0009	$p = 0.38$
	atlas	0.2986	0.0610	0.0610	$p < 0.05$
CB	expert	2.3112	0.3826	0.0197	$p < 0.001$
	atlas	2.6064	0.3924	0.0222	$p < 0.001$

Table D.8: Parameters resulting from the fitting of the relative volumes (with respect to the whole brain volume) as function of the gestational age ($V_{rel,brain} = bias + w \cdot x1 + w^2 \cdot x2$, with w the age given in weeks) for the different subcortical structures.

	model	bias	x1	x2	significance x2
CP	expert	0.071	-0.0010	$4.654 \cdot 10^{-5}$	$p < 0.005$
	atlas	0.084	-0.0013	0.0001	$p < 0.001$
LPVH	expert	0.0018	$-7.785 \cdot 10^{-5}$	$-5.397 \cdot 10^{-6}$	$p = 0.6$
	atlas	0.0019	$-4.979 \cdot 10^{-6}$	$-4.017 \cdot 10^{-6}$	$p = 0.5$
CSPV	expert	0.0021	0.0002	$-4.697 \cdot 10^{-5}$	$p < 0.001$
	atlas	0.0024	0.0001	$-2.916 \cdot 10^{-5}$	$p < 0.001$
CB	expert	0.0914	0.0001	$-3.002 \cdot 10^{-5}$	$p = 0.27$
	atlas	0.0224	-0.0003	$5.121 \cdot 10^{-5}$	$p < 0.05$