Improving deep-learning segmentation performance in 3D neuroimaging with minimal manual annotations

Lorenzo Venturini
New College
University of Oxford

Submitted in partial fulfilment of the requirements of the degree of

Doctor of Philosophy

Michaelmas 2021

Acknowledgments

Personal

First and most importantly, I'd like to thank my supervisor, Ana I.L. Namburete, for the hard work and long hours she took guiding me through this DPhil project. She has been the best mentor I could have asked for, providing me with all the guidance I needed while also giving me the space to make this project my own. She has always been there to listen to me and provide feedback on my ideas and work, and I am incredibly grateful to her.

I'd also like to thank my co-supervisors, Profs. Aris T. Papageorghiou and J. Alison Noble, for helping my research with their expertise, experience, and perspectives. Prof. Noble, in particular, volunteered to take over as my main supervisor for much of 2020 and 2021, when Dr. Namburete was on leave. She was incredibly helpful in her supervision, and they have both been instrumental in shaping this project.

I would also like to thank my fellow students at the Oxford Machine Learning in Neuroimaging Group for all the help they have given me through this project. All the time we spent brainstorming together, proofreading each other's work, and discussing papers with each other has made a profound mark on this thesis.

This DPhil was done as part of the Oxford-Nottingham Biomedical Imaging (ONBI) programme, whose leaders have always been incredibly supportive of me through this DPhil. My thanks especially go to Prof. Peter Jezzard, who has several times provided me with one-on-one help, even when it wasn't at all expected of him.

I would also like to thank all of the friends I made in my time at Oxford, especially those in the New College MCR and the New College Boat Club, for all of the good times we had together, and the experiences we had. I was lucky to have such a strong support network as the friends I made here.

My partner, Dr. Emily Davenport, has been the strongest supporter I've had through all of this. She has always been free to discuss ideas with me, make plans, help me practice presentations, and proofread my writing. Meeting her during this DPhil has been the best thing that has happened to me at Oxford.

Finally, I would like to thank my parents Marco and Shahrzad, and my sister Parissa, for all of the support and encouragement they have given me through this journey. It was thanks to everything they did for me that I was able to come here in the first place, and they have helped me immensely through the years.

Institutional

This work is supported by funding from the Engineering and Physical Sciences Research Council (EPSRC) and Medical Research Council (MRC) [grant number EP/L016052/1]. I would also like to thank New College and the ONBI CDT for their generous grants and

support.We thank the INTERGROWTH-21st Consortium for permission to use 3D ultrasound volumes of the fetal brain.

Abstract

One of the current challenges in applying machine learning to medical images is the difficulty in obtaining labelled training data. While medical images themselves are often available, generating high-quality training labels for them is time-consuming and often requires a trained clinician. This problem is particularly acute in 3D segmentation, which generally requires detailed voxel-by-voxel segmentation maps.

This thesis proposes a series of image analysis methods to leverage additional available information to improve image segmentation quality when there are limited manual labels available. Using a dataset of 3D fetal ultrasound scans with only a small number of initial segmentation labels, we demonstrate a benefit from different information sources. The approaches used in this thesis are:

- Generating segmentation labels automatically from a publicly available spatio-temporal atlas, making use of prior anatomical information. We use anatomical keypoints to guide registration of atlases across different imaging modalities. We then propagate these segmentations to individual ultrasound volumes, automatically generating a set of multi-label segmentations with an average Dice coefficient of 0.818 across the chosen anatomical structures in the range of 20-25 gestational weeks. This set of labels is used to train a segmentation CNN. We find that this produces high-quality segmentations, and that training a single network with the full multi-label training set improves segmentation quality (average Dice coefficient of 0.726 across structures for a multi-task network, versus 0.659 for individual single-task networks).

- Leveraging additional, unlabelled data to improve performance for a cerebellar segmentation task. Unlabelled data does not have a ground-truth label by which segmentation quality may be compared, but measures of output uncertainty can be derived using test-time augmentation and dropout measures. We use these estimates of uncertainty to inform an iterative omni-supervised framework, using the highest-quality segmentations from the unlabelled dataset as additional training data for a segmentation CNN. We show that the use of these labels as part of a training dataset improves segmentation performance, and that this is improved further when measures of uncertainty are used in data selection (an overall improvement in Dice coefficient for a cerebellar segmentation task from 0.673 to 0.727).

- Passing additional, scalar information to a neural network's input. We propose an original method by which scalar data such as gestational age, readily available as part of clinical data acquisition, may be appended to the input of a fully-convolutional neural network. We show that including this information as part of the training dataset can lead to significant improvements in segmentation quality on the same cerebellar segmentation task, in the reported experiments improving test Dice coefficient from

0.673 to 0.723. We find that this improvement is not sensitive to imperfect information passed at test time.

The computational methods developed in this thesis may reduce the reliance on large manually-labelled datasets for medical image analysis and save time in generating expert annotations.

Contents

List of Figures

List of Tables

List of abbreviations

2D,3D	Two- or three- dimensional
AD	Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
BPD	Biparietal diameter
CN	Cognitively normal
CNN	Convolutional neural network
CSF	Cerebrospinal fluid
CSP	Cavum septi pellucidi
DSC	Dice similarity coefficient
ED	Euclidean distance
FSL	FMRIB Software Library
GA,GW	Gestational age, gestational weeks
GT	Ground truth
GPU	Graphics processing unit
HARP	HARmonized Protocol
HD	Hausdorff distance
INTERGROWTH	International Fetal and Newborn Growth Consortium
IOU	Intersection over union
LMP	Last menstrual period
MCI	Mild cognitive impairment (often progresses to AD)
MRI	Magnetic resonance imaging
MSP	Mid-sagittal plane

QuickNAT	Quick segmentation of NeuroAnaTomy
SF	Sylvian fissure (or temporal sulcus)
T1	Longitudinal magnetic relaxation time
T2	Transverse magnetic relaxation time
TCD	Transcerebellar diameter
TSI-Net	Transformed Scalar Inputs-Net
t-SNE	t-stochastic neighbour embedding
TTA	Test-time augmentation
TTD	Test-time dropout
WM	White matter
US	Ultrasound

Chapter 1

Introduction

Chapter layout

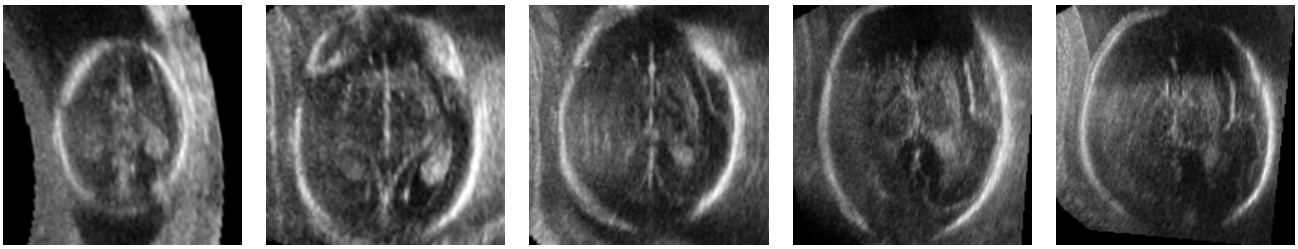
This chapter introduces the challenges tackled by this thesis. It also provides an overview of the work done in this thesis and the structure of the rest of the document.

I begin by motivating for the work done in this thesis in Section 1.1. Section 1.2 describes the structure of this thesis document, providing an outline of the content of each chapter, and listing the contributions made in each chapter. Section 1.3 provides an overview of who contributed to the work presented in each chapter. Finally, Section 1.4 lists the publications, poster presentations, and manuscripts in preparation arising from this thesis.

1.1 Motivation

Over the course of a human pregnancy, extremely complex anatomical structures like the brain arise from a neural tube and rapidly develop to their familiar forms. The development of the fetal brain has long been observed using imaging methods such as ultrasound and magnetic-resonance imaging (MRI): the brain itself is visible by the end of the first trimester of pregnancy, and its initially smooth surface develops the characteristic grooves (sulci and gyri) of human brains throughout the second and third trimesters [5]. During this process abnormalities may also occur and be seen through these imaging methods: conditions such as

ventriculomegaly [6], lissencephaly [7], and congenital heart disease [8] have been reported to be visible and diagnosed through altered development of the fetal cortex throughout the second and third trimesters.



14 weeks 18 weeks 22 weeks 26 weeks 30 weeks

Figure 1.1: Axial slices from ultrasound acquisitions of the fetal brain at different gestational ages. Note that anatomical details in the right (distal) hemisphere are clearly visible, while anatomy in the left hemisphere is indistinct.

These conditions have generally been clinically assessed using subjective measures of development, such as the visibility [9] or depth [10] of specific sulci and gyri. Advances in imaging quality and computational power, together with the development of new analytical techniques, make objective, computational analysis of medical images increasingly feasible and powerful. Automated localisation and segmentation of brain structures in medical acquisitions, in particular, is a task that has been attempted in both ultrasound and MRI, allowing both clearer visualisation of structures for clinicians and a starting point for further automated analysis. Other work has tackled development directly, generating estimates of gestational age (or the extent to which the pregnancy has progressed) automatically based on the development of brain structures [11].

Machine-learning methods, especially neural networks such as convolutional neural networks (CNNs), are one approach by which these measures may be automatically extracted. CNNs can automate classification, regression, and segmentation tasks on image data. CNNs have been used broadly in a range of medical imaging tasks. They have also been used to localise structures in 3D ultrasound data [12], to regress fetal age [11], and to align 3D volumes to a standard orientation [13].

Neural networks, however, typically require large amounts of labelled data to serve as a training set, which provides supervision during training. They usually require far more

data to reach optimal performance than humans do, when trained at the same task [14]. The ImageNet database of natural images for classification, for instance, contains over 14 million natural images [15], each of which is human-annotated.

The manual annotation step in creating labelled training datasets is often limiting, considering the large number of labels required. Humans' ability, and their variation when generating labels[1], provides the upper bound for the potential performance of a network trained using manual labels. Large-scale datasets of natural images typically distribute labelling across many users, using tools such as Amazon Mechanical Turk [16] or Google reCaptcha [17].

This is only suitable for images that can be labelled adequately by non-expert humans, like natural images. Medical image data, on the other hand, often requires trained domain experts, usually clinicians, to annotate. This can be a time-consuming process, particularly for segmentation labels, which makes it difficult to generate labelled datasets on a large scale[2]. One of the major challenges in applying machine-learning techniques to medical image data, therefore, is in how to maximise their performance in the presence of only a limited amount of training data.

In this thesis, we propose several methods to do so, using known information to guide learning when traditional labels are limited. The contributions proposed in this thesis explore different sources of information, such as anatomical prior knowledge, additional unlabelled image data, and additional non-image data.

1.2 Structure and contributions of this thesis

Chapters 1, 2, and 3 provide background on the challenges confronted in this thesis, review the literature on existing approaches, and describe the data used. Chapters 4, 5, and 6 present the original technical work done in this thesis. Each chapter begins with an Outline section,

[1] Known as inter-observer variability when multiple humans are asked to label the same image, and intra-observer variability when the same human is asked to label the same image twice independently.

[2] This is particularly acute when 3D labels are needed, which can be much more time-consuming than 2D labels of the same structure. Chapter 5 provides an estimate of the time required to generate segmentation labels manually.

setting out the structure of the chapter.

Chapter 2 introduces the clinical motivation for fetal brain segmentation and reviews current clinical practice and the available imaging technology. It also reviews the state of the art in medical image segmentation and different techniques by which it is performed, including machine learning-based methods. It also provides an overview of some methods used in this thesis, such as omni-supervised learning and uncertainty estimation using test-time dropout and test-time augmentation.

Chapter 3 is a short chapter which introduces the datasets and annotations used throughout this thesis. Their clinical characteristics, as well as technical details such as acquisition protocol, annotation methods, and number of volumes and labels available from each data source is described here.

Each of Chapters 4, 5, and 6 examines a different method to improve segmentation quality with only limited manual segmentation data, using different information sources to provide additional supervision. Each chapter begins with an **Introduction** section outlining the problem, followed by a **Proposed approach** section setting out the methods proposed in the chapter. The specific experiments are motivated and explained in the following sections, and the **Results** section presents their primary outcomes. These are then discussed in the **Discussion** section, and the high-level message of the chapter is distilled in the **Conclusion** section.

Chapter 4 generates labels using an atlas-based method, which are then used to train a multi-task segmentation CNN. It introduces two spatiotemporal atlases of the developing fetal brain, one assembled from ultrasound data and the other from MRI images. Manual annotation of keypoints in each of those atlases at different time points is used to estimate a set of affine transformations to register the two atlases, and the structural labels in one atlas are then transformed to the other atlas space. These are then propagated to individual volumes within the ultrasound atlas, creating a large set of individual labels.

As this method can only be used for volumes with a known correspondence to the atlas space, these labels are then used to train a multi-task CNN, to segment new volumes. We use

a modified Dice loss function to segment multiple brain structures in 3D simultaneously, and propose a modification of a standard segmentation CNN architecture. This is compared to slicewise 2D approaches as well as to single-task 3D approaches. The multi-task approach was found to outperform traditional segmentation methods.

Chapter 5 proposes a method to use additional unlabelled image data to improve output quality, using uncertainty estimation and omni-supervised learning. It begins by introducing a cerebellar segmentation task in a fetal ultrasound dataset and describing the manual annotations obtained for this dataset. These annotations are used to train a CNN to segment the cerebellum in a larger unlabelled dataset drawn from the same source, as well as to obtain quantitative estimates of aleatoric and epistemic uncertainty for each of these segmentations. These uncertainty measures are found to aid volume selection for a second round of omni-supervised training: selecting the volumes with the lowest aleatoric uncertainty outperforms baseline, while selecting the volumes with the lowest epistemic uncertainty does not. The reasons for this discrepancy are discussed, as well as possible sources of noise within the uncertainty estimates which may affect the results. These methods are validated in a separate MRI dataset, which shows a similar performance boost.

Chapter 6 introduces TSI-Net, a method by which scalar data can be included in a fully-convolutional architecture. Scalar features typically cannot be processed by a fully-convolutional CNN, as convolutional filters can only operate on grid-like data. TSI-Net is a method to transform these scalar features into an image-like format, which can then be passed as an additional input to the network. We use this to pass fetal gestational age to a segmentation CNN alongside image data, measuring a significant performance improvement. We also experimented to find the most appropriate format to which this data could be transformed and found how modifications to the baseline CNN to incorporate scalar data lead to changes in performance, as well as in training time and memory usage.

Finally, Chapter 7 summarises the contributions described in this thesis and considers the limitations of the work presented here. It also presents possible future directions, from which the contributions made in this thesis may lead to further technical and clinical innovations.

1.3 Originality

I declare that I am the sole author of this thesis document, and I produced all the tables and figures unless otherwise specified. Critique and comments were provided by my supervisor Dr. Ana I.L. Namburete and my co-supervisor Prof. J. Alison Noble.

This thesis uses two datasets to demonstrate the methods developed here. One is the INTERGROWTH-21st dataset of fetal ultrasound data [18], from which I used 3D scans of the fetal brain to demonstrate all the contributions developed here. I independently manually generated segmentations for part of this dataset. The other dataset used is the MRI ADNI dataset [19] and the HARP subset of this dataset with manual annotations of the hippocampus [20]. These were used with permission: Prof. Aris Papageorghiou of the INTERGROWTH Consortium gave me permission to use data from the INTERGROWTH-21st dataset, and I obtained permission from the ADNI Data committee to use data from the ADNI dataset.

Chapter 4 This chapter uses two atlases of the developing fetal brain: one is the publicly available MRI atlas[3] developed by Gholipour et al [2], the other is the ultrasound atlas developed for the INTERGROWTH-21st dataset by Dr. Ana Namburete [3]. Dr. Namburete selected the volumes which the atlas was constructed from, performed the registration, and gave me the registration parameters which could be used to propagate the atlas to individual volumes. I independently annotated anatomical keypoints in both of those atlases and developed the Python code to find an affine transformation matrix between those keypoints. I also wrote code to implement that affine transformation and propagate selected labels to the ultrasound atlas. I adapted Dr. Namburete's MATLAB code, originally developed to construct the fetal ultrasound atlas, to propagate segmentations from the atlas to the individual volumes. I also made the design decisions in this process, such as what structural labels to propagate, or the type of interpolation to use, with the help of Dr. Namburete.

I also independently wrote code defining a multi-task 3D segmentation CNN, modifying the standard U-Net architecture to use 3D convolutional layers, and modifying the Dice coefficient

[3]Available at `http://crl.med.harvard.edu/research/fetal_brain_atlas/` and used with permission.

loss function to use multiple labels. I trained this network and comparable architectures, producing the results discussed in that chapter.

Chapter 5 I independently manually segmented the cerebellum in a subset of the INTERGROWTH-21st dataset, creating a training dataset for this. This work was supervised by Dr. Ana Namburete and I incorporated her feedback when adjusting my manual delineations. I wrote the code to modify the standard U-Net to use Dropout and used those segmentations to train this network. I also used this trained network to generate segmentations of unlabelled data. I wrote code to extract aleatoric and epistemic uncertainty measures, as well as rank volumes in those datasets by those measures. I retrained the network using selected segmentations.

I also performed the work necessary to validate this work on the ADNI/HARP dataset: I obtained access to the datasets, selected the appropriate subsets to download, and preprocessed them to obtain consistent image features. Some preprocessing work for MRI image data was done using the FMRIB Software Library (FSL), with the assistance of my colleague Nicola Dinsdale. I adapted the code used to perform omni-supervised training on the ultrasound dataset to the MRI dataset, to provide validation on a different dataset.

Chapter 6 In this chapter, I independently proposed TSI-Net, a method to integrate scalar data into fully-convolutional CNNs. The additional scalar data used in this chapter was available as part of the datasets used throughout this thesis. I wrote the python code to implement the transforms, as well as the code to modify the CNN architecture to accommodate additional inputs. I performed all the experiments to evaluate the most appropriate transformation (and the parameters of the transformation) and reported the results. The image data and segmentation labels used for this chapter were identical to those used in Chapter 5.

I also performed all the work necessary to validate these results on the separate ADNI/HARP dataset. The dataset and labels used for this task were also identical to those used in Chapter 5, to produce comparable results.

1.4 List of Publications

Peer-reviewed conference proceedings

Lorenzo Venturini, Aris T. Papageorghiou, J. Alison Noble, and Ana I. L. Namburete, "Multi-task CNN for Structural Semantic Segmentation in 3D Fetal Brain Ultrasound," in *Medical Image Understanding and Analysis (MIUA),* July 2019 | Podium presentation. [21]

Lorenzo Venturini, Aris T. Papageorghiou, J. Alison Noble, and Ana I. L. Namburete, "Uncertainty estimates as data selection criteria to boost omni-supervised learning," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, July 2020. [22]

Poster presentations

Lorenzo Venturini, Aris T. Papageorghiou, J. Alison Noble, and Ana I. L. Namburete, "Custom layers improve CNN segmentation of fetal ultrasound", at the *Biomedical Imaging CDT Summer School*, Nottingham, UK, 2018

Lorenzo Venturini, Aris T. Papageorghiou, J. Alison Noble, and Ana I. L. Namburete, "Custom layers improve CNN segmentation of fetal ultrasound", at the *12^{th} Oxford Biomedical Imaging Festival*, Oxford, UK, 2018

Lorenzo Venturini, Aris T. Papageorghiou, J. Alison Noble, and Ana I. L. Namburete, "CNN for segmentation of 3D fetal ultrasound", at the *3^{rd} Medical Imaging Summer School*, Favignana, Italy, 2018

Chapter 2

Literature review

Chapter layout

This chapter reviews the state of current research in fetal brain imaging and medical image segmentation, as well as approaches to improve segmentation quality when limited numbers of labels are available. This lays out the groundwork for this DPhil thesis' research contributions. This chapter does not review specific datasets that are used in this thesis: they are reviewed later, in Chapter 3.

I begin with a general review of the medical tasks that are addressed in this thesis, and their particular characteristics. This chapter begins by reviewing the current understanding of fetal brain development and in Section 2.1. Imaging of the fetal brain is reviewed in 2.1.1, as well as the acquisition and visual characteristics of different imaging methods. This chapter also provides a brief review of MRI imaging of Alzheimer's disease in section 2.2, as some of the work in this thesis is also demonstrated on an MRI dataset of aging adults.

Section 2.3 then reviews the field of medical image segmentation, and identifies the key challenges presented by medical data, such as the difficulty of obtaining labelled training data. Section 2.3.1 reviews segmentation methods, such as atlas-based segmentation and machine-learning based methods, and Section 2.3.3 explores ways to evaluate their performance. Section 2.4 considers the challenge of limited training data, and briefly reviews approaches that have been attempted to tackle it. It then reviews one such method, omni-supervised

learning learning, in more detail. Section 2.4.1 gives an overview of uncertainty quantification in image segmentation, which is often difficult to measure from a neural network's output. Finally, Section 2.5 reviews existing work on the integration of image and scalar data in single machine learning models. The background investigated in this chapter is referred to in the rest of this thesis.

2.1 Fetal brain development

A healthy human pregnancy takes place over a period of 40 weeks on average, as measured from the start of the woman's last menstrual cycle [23] (see Figure 2.1). The stage of pregnancy is commonly measured in *gestational age* (or GA) from this same starting point.

In healthy fetuses, the central nervous system begins to develop around 5 gestational weeks (GW) [24]. The first brain structures begin to be visible to ultrasound scans around 9 GW, and initially consist of the high-contrast brain ventricles and choroid plexi [25]. By 15 GW, all large-scale structures and features (including the cortical plate, cerebellum, thalami, and mid-sagittal plane (MSP)) are visible in ultrasound scans. One notable phenomenon during the first and second trimesters of fetal brain development is *neuronal migration*, in which new neurons are initially produced near the centre of the brain and then move to their final positions. This can result in the central regions of the brain (such as the brainstem and thalami) increasing in volume more quickly than peripheral regions like the cortical plate, in the first trimester [26]. In the second and third trimesters, this process is reversed, with peripheral regions of the brain (such as the cortical plate) growing more rapidly than areas near the ventricles [27]. This process continues beyond birth and into infancy.

The fetal brain is still smooth by the end of the first trimester and only gradually develop the characteristic grooves (cortical sulci and gyri) of adult brains' structure. The first such groove, known as the Sylvian fissure, can first be seen in ultrasound at around 17 GW, and the parieto-occipital fissure can be seen around 19 GW [28]. More such structures gradually appear, and by 33 GW, all major sulci and gyri are visible [29]. A similar trajectory can be observed with magnetic resonance imaging (MRI) [30]. These features are visible earlier in

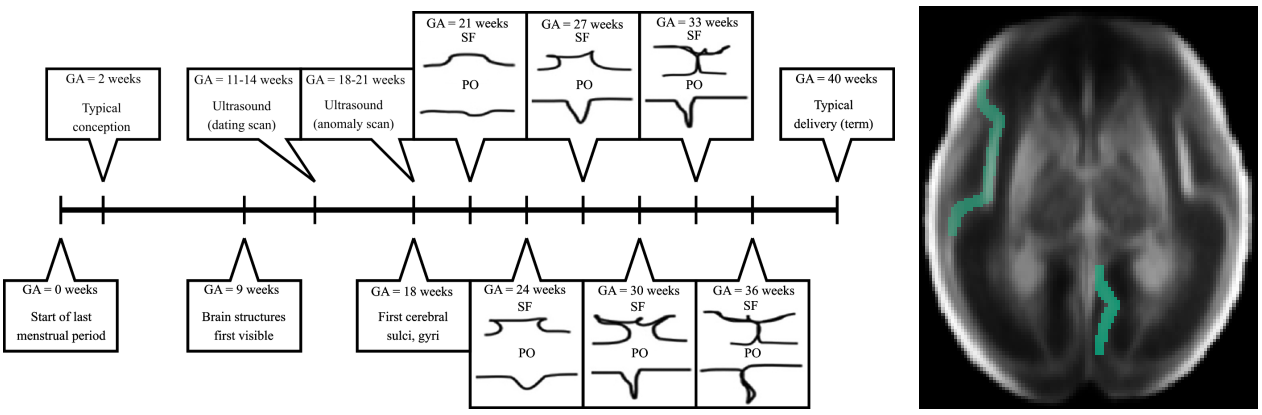


Figure 2.1: (a) timeline of a typical pregnancy, with the key milestones for neuroimaging purposes marked. The typical appearance of the Sylvian fissure (SF) and the parieto-occipital fissure (PO) are shown where visible. Appearance of the sulci reproduced from [1]. (b) Axial view of a fetal brain at 24GW, with the SF (left) and PO (centre) highlighted.

development (by an average of two weeks) in postmortem anatomical samples at the same gestational age [5], likely due to limitations in the contrast and resolution of ultrasound (US) and MRI. This development is not symmetric across cerebral hemispheres, with sulci and gyri on the right hemisphere generally developing a few days earlier than those on the left hemisphere [31].

At first appearance, fissures and gyri are visible just as a small indentation, or dot, in the smooth cortical plate. Throughout gestation, however, these structures become more complex and branched, gradually taking on the more familiar morphology of deep folds in the cortical surface. A timeline of how these structures develop as they appear in ultrasound images, as well as the typical progression of a healthy pregnancy, is shown in Figure 2.1. However, even at term brain structures are not fully developed, and brain gyrification and morphological development continues into infancy [24].

2.1.1 Imaging of the fetal brain

Most commonly, the fetal brain is imaged using ultrasound. Ultrasound is a standard clinical examination, and throughout the developed world most pregnant women receive routine ultrasound scanning as a screening measure [32]. Ultrasound imaging is typically conducted at a frequency of 3-7.5 MHz [33], which results in a spatial resolution typically in the range of

0.3-1mm (depending on the frequency) [34]. Ultrasound can be collected in real-time, with a frame rate often in the range of 10-15 frames per second. While most transducers image in a 2D fan-shaped area, 3D imaging is also common and can be acquired in real time [34].

All ultrasound imaging is characterised by *speckle*: anatomical features smaller than the resolution of the signal result in a persistent granular inhomogeneity in the signal intensity [35]. This adds ambiguity to the signal and makes accurate delineation of structures (and boundaries) more difficult even for an expert.

Typically, only one cerebral hemisphere can be visualised in detail in a sonographic examination. The cause of this is not very well understood, but it is believed to be related to interference with the ultrasound signal reflected by the distal hemisphere of the skull [11]. An example of the appearance of this artifact can be seen in Figure 2.3b. Features in the proximal hemisphere appear more blurred and indistinct with increasing distance from the MSP. The cerebellum, however, straddles the MSP, making it possible to visualise the entire structure regardless of the acquisition angle.

Ultrasound is also susceptible to shadowing artifacts when a reflective structure occludes other tissue behind it in the acquisition pathway. The petrous ridges of the temporal bone can cause strong shadows over the cerebellum [36]. The petrous ridge of the temporal bone is a thick bony ridge in the skull, found behind the ear. This is thicker than the rest of the skull, even *in utero*, so it acts as a strong ultrasound reflector. The direction in which the shadow is cast depends on the angle of the probe, but within standard acquisition protocols it is common for the shadow to extend over the cerebellum. This makes it very difficult to visualise the boundaries of the structure. An example of a shadow cast by the petrous ridge is shown in Figure 2.3(c).

Furthermore, image contrast and grey levels can vary significantly between subjects and even within a single acquisition. Between subjects, this is due to individual characteristics, such as the thickness of (reflective) abdominal adipose tissue of the mother [37]. Within acquisitions, this can be explained by shadowing artifacts closer to the probe as well as acquisition settings such as time-gain compensation [38].

All of these artifacts complicate semantic segmentation: boundaries are often ambiguous and ill-defined, and even a trained expert will not consistently produce identical segmentations. This is explored more quantitatively in section 5.3.1.1.

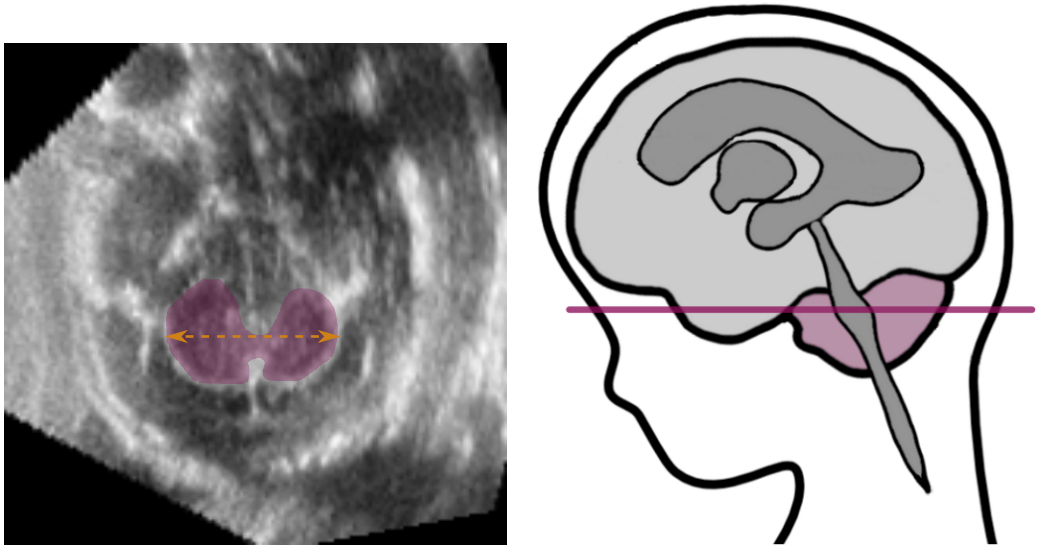
MRI has also been used to image the fetal brain, although it is not a routine examination [39]. Obtaining 3D volumes of fetal brains from MR imaging is a greater challenge, due to the greater time required to acquire a 3D stack of slices. A full isotropic 3D stack can have an acquisition time of up to several minutes [40]. Fetal motion in that period cannot be constrained, which introduces motion artifacts in 3D acquisitions [39]. Earlier MRI studies, therefore, have either used 2D slices only or 3D acquisitions from postmortem [41] or neonatal acquisitions [42, 43]. Improvements in motion correction have since made 3D MRI acquisitions possible in healthy fetuses [44], but remaining challenges in motion correction and the lower spatial resolution of MRI still make it less commonly used at lower gestational ages, and many studies that use 3D fetal MRI are limited to the third trimester [39]. Tissues in the developing brain also exhibit different MR characteristics than in the adult brain: the lower myelination of the fetal brain leads to shorter T1 and T2 relaxation times in the grey matter [45], making the white matter appear hypointense relative to grey matter in fetal acquisitions. This is not fixed but changes over time, and only reverses in infancy [46]. T1-weighted acquisitions offer poor contrast between tissue types at this stage of development, so T2-weighted acquisitions are primarily used [40].

.

2.1.2 Development and imaging of the fetal cerebellum

Development of the cerebellum is an important process d[ref]uring pregnancy. In adulthood, the cerebellum contains over half of the brain's neurons [47]. The cerebellum is divided into three primary anatomical parts: the vermis, located centrally along the MSP separating the two hemispheres, and the two cerebellar hemispheres which can be found on either side of it. Cerebellar growth is, to a close approximation, linear [48] throughout pregnancy, making cerebellar measurement a reliable way to estimate brain development.

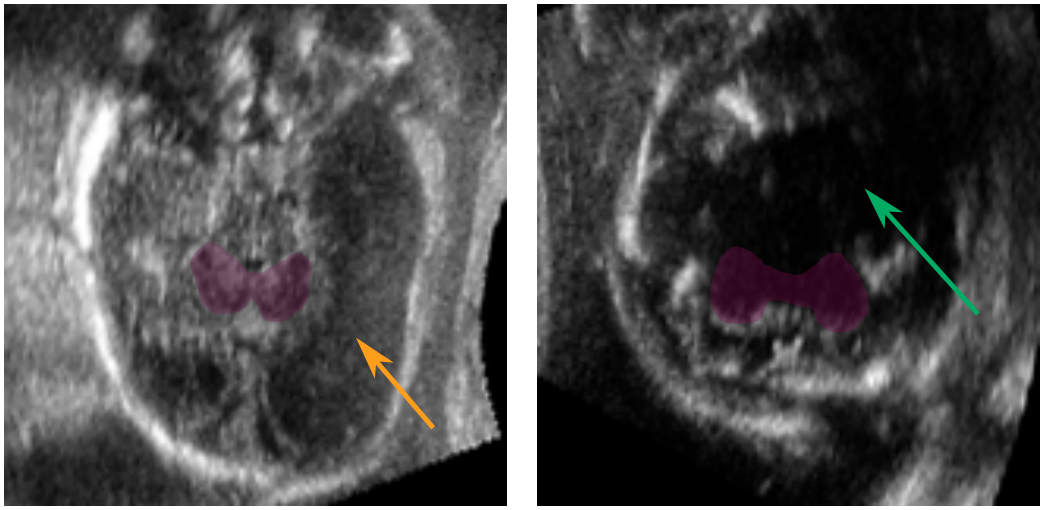
As such, the cerebellum is assessed in routine prenatal examinations [49]. The cerebellum presents in fetal ultrasound as a hypoechoic structure, divided into two hemispheres on either side of the MSP and joined in the middle by the hyperechoic cerebellar vermis. At the fetal anomaly scan, typically performed at 18-21 gestational weeks in the UK, one of the measures of development obtained by clinicians is the transcerebellar diameter (TCD), together with biometric measures of the head such as the occipitofrontal diameter (OFD) and the biparietal diameter (BPD). The TCD is a simple linear measure of the size of the cerebellum, measured as a line perpendicular to the MSP connecting the two furthest points of the cerebellar lobes. Figure 2.2 shows the appearance of the cerebellum in ultrasound, an estimate of the TCD, and the position of the cerebellum in the fetal brain.



(a) (b)

Figure 2.2: a) An ultrasound image showing the cerebellum (in purple), and the TCD in yellow. b) the plane of that slice in the fetal brain. Image credit to Felipe Moser.

The TCD is the most reliable known biomarker for gestational age (GA) estimation, used in combination with the OFD and BPD for that purpose. The TCD is considered more accurate, especially in cases where the fetus is abnormally small or large for its gestational age, due to conditions such as malnutrition or maternal diabetes [50, 51]. This is due to the "brain-sparing effect" [52]: in cases of decreased metabolic conditions such as malnutrition,



(a) (b)

Figure 2.3: a) A volume where the proximal hemisphere (right) has much less visible detail than the distal hemisphere, but the cerebellum remains visible. (b) A volume where much of the cerebellum is obscured by a shadow artifact (resulting from the petrous ridges, on the left of the image).

nutrient supply to the brain is preserved relative to the rest of the body. Cerebral biomarkers of gestational age such as the TCD are therefore more predictive of age than others [53].

2.2 Dementia and hippocampal imaging

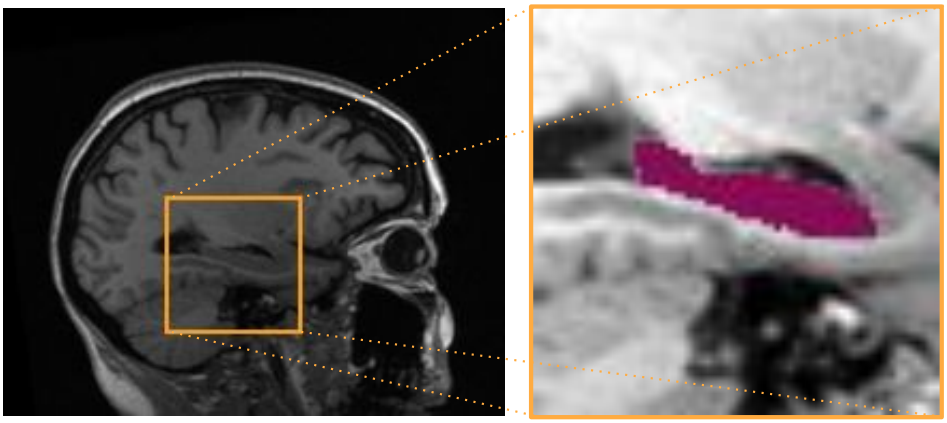


Figure 2.4: One hippocampus, seen in the sagittal view, of a 76 year old healthy female. The image and segmentation label are drawn from the ADNI/HARP dataset.

The hippocampus is a structure found within the brains of all mammals. Humans have two hippocampi, one in each temporal lobe of the brain [54]. It is believed to play a major role in memory acquisition and retrieval, and damage to the hippocampus has been associated with

memory impairment [55]. Imaging the hippocampus can yield insights on how morphological changes to this structure can lead to cognitive decline. Normal ageing is associated with reductions in hippocampal volume, which correlate with age-related cognitive decline [56].

The hippocampus is also affected by neurodegenerative conditions such as Alzheimer's disease (AD) [57]. Alzheimer's disease is a common ($> 20\%$ for those over 75 years old [58]) progressive neurodegenerative condition and one of the main causes of death in the developed world[1] [58]. It occurs primarily in the elderly, with an increasing incidence with increasing age. Despite its prevalence, it is in many ways poorly understood: it remains difficult to predict an individual's risk of developing AD, diagnose AD in patients with symptoms, or devise possible treatments for it. Dementia caused by Alzheimer's disease is typically preceded by a milder form known as mild cognitive impairment (MCI), but cognitive symptoms are often predated by abnormalities that can be seen in medical images [60].

The classical course of Alzheimer's disease is from CN to MCI to AD, with progression from MCI to AD at a rate of about 10-15% per year [61]. However, there is a wide range of outcomes: some patients with MCI progress to AD very quickly, while others display much slower progression. Other patients with MCI do not appear to noticeably progress in their cognitive decline, while others still progress to non-AD dementias. There is a very wide range in imaging findings due to this; some cognitively normal subjects also exhibit imaging signs of AD, which can sometimes predict future cognitive decline [62].

Alzheimer's disease is associated with an accumulation of plaques of β-amyloid protein in the brain, which appear to precede the onset of cognitive symptoms [63]. This accumulation, however, is difficult to image directly using MRI, due to their small physical size (individual plaques are under 100μm in diameter) - they can only be visualised clearly from post-mortem histological slices. Medical imaging methods such as MRI, however, can identify morphological changes in brain structures, such as atrophy, associated with AD.

A prominent symptom of Alzheimer's disease is amnesia, and particularly short-term memory loss. The hippocampus, which is associated with short-term memory acquisition and

[1] In 2019, Alzheimer's disease was the leading cause of death registered in the UK, accounting for 16% of mortality. [59]

retrieval, can therefore be expected to undergo visible atrophy and morphological changes through the progression of AD. Segmentation in hippocampal MRI can be used to assess AD progression, and is hippocampal atrophy can be predictive of cognitive decline [64]. There is interest, therefore, in a way to automatically quantify hippocampal volume from MRI, and the changes in hippocampal volume over time.

2.3 Medical image segmentation

One of the main objectives in the field of image analysis is reliable, automated image[2] *segmentation*: that is, the labelling of sets of pixels within the image to correspond with meaningful regions of interest. In the context of image analysis in brain imaging, this involves identifying and labelling specific anatomical structures (such as the cortex, thalami, brainstem, or the cerebellum) and tissue types (such as grey matter/white matter/CSF) within an image or volume. This can be done manually, but that is time-consuming, suffers from intra- and inter-observer variability, and is difficult to scale to larger datasets. Therefore, an active area of research is automatic image segmentation, where the objective is to reach human levels of accuracy without the need for a trained expert.

Early work in ultrasound and fetal brain image segmentation relied on traditional hand-crafted features and image-processing pipelines, such as edge detection and shape models. Anquez et al [65] used pixel values and a deformable model to segment fetal bodies in 3D ultrasound images, which is a similar approach to that taken by Becker et al [66] to segment the cerebellum in ultrasound. More recent work has also built on these methods: Rueda et al [67] use shape constraints to obtain good quality segmentations. These methods have, however, become less widely used as others take advantage of the increased computing capacity available to produce improved results. Handcrafted features are still widely used to initialise or guide other segmentation methods, but rarely are used alone to perform segmentation themselves.

[2]For simplicity, this section will often refer solely to image segmentation; however, all of the methods described here are applicable to 3D volumes as well as 2D images.

In recent years, the performance of machine-learning methods for image segmentation has improved significantly. This encompasses a wide range of techniques, but all are based on the premise that segmentation should not rely on human selection of hand-crafted features, but rather be performed via features derived automatically through the machine's learning process. The most notable class of methods used is *convolutional neural networks* (CNNs), of which some examples are reviewed here.

2.3.1 Registration and atlas-based methods

One segmentation scheme is *atlas-based segmentation*. This relies on a hand-segmented reference image (or group of images) that all images in the dataset are registered to, from which they inherit its segmentation labels. The power of this segmentation method comes from the possibility to use it to segment large datasets from a small number of labelled images.

Registration is a key step in this process, and it is simply a method to align and transform an image into the same coordinate system as another, to facilitate comparison of images. A number of different schemes have been developed to achieve this, and it is worth going into some detail into the methods underlying them. All image registration algorithms are either *feature-based,* where the registration is guided by the correspondence of specific features in the images, or *intensity-based*, where registration is guided by the pixel values directly. Intensity-based measures require less initialisation, as they work on raw images; however, it is difficult to use intensity-based methods directly on ultrasound images, as grey levels are not consistent within or between different acquisitions (as explained in Section 2.1.1). Feature-based registration scheme range from automatic methods where features are extracted through hand-crafted descriptors such as edge detectors and filters (see Section 2.3.2) to semi-supervised methods where the features of interest are labelled by hand to guide the registration [68].

There are many registration algorithms which may be used to establish correspondence between an image and an atlas [69]. The simplest, global methods are *rigid* and *affine* registration. Rigid registration reaches correspondence using only translation and rotation of

the images, while affine methods build on that by adding scaling and shear transformations. These often can only achieve a rough correspondence if the features of interest are not morphologically identical between each other. Therefore, modern atlas-based segmentation schemes rely on *non-rigid* registration methods. These rely on local deformations of only certain regions of the image to achieve a better correspondence where there are differences in the anatomy or the imaging method used. Popular methods to do so include radial basis functions such as B-splines [70], viscous fluid models [71] or diffeomorphisms [72]. Not all of these methods are invertible: where a transformation maps two points in the input to a single point in the output, the registration will generally not be invertible[3]. All registration methods are guided by an objective function to be minimised, often drawn from the measures listed in Section 2.3.3 [73]. For non-rigid registration schemes, this is often still an ill-posed problem as there are many parameters and local minima, and therefore the objective function is augmented by regularisation terms and heuristics [68].

Since atlas-based segmentation relies on registering every image to an atlas, the construction of an atlas is important to obtaining a good segmentation quality. Early work used a single manually labelled image as an atlas, typically the one considered "best" or "most representative" of the population within the dataset [73]. However, this is a subjective determination, and does not generally account for anatomic and imaging variability within the study population. In 2006, Heckemann et al [74] introduced multi-atlas segmentation in adult MR brain images, where instead of a single image a set of labelled images is used as the atlas and every new image is registered to each of them. The class labels are then decided by a consensus of the predicted labels for each atlas.

Another approach is to construct the reference template from the entire dataset, so that it corresponds not to an individual volume or set of volumes but to an implicit reference, corresponding to aggregate an average of the dataset along some dimension [75]. This typically involves minimising a measure of visual distance between each image and the atlas, such as the mean-squared pixel value [76] or the pixelwise probability distribution [77]. This results in

[3]This is important as atlas-based registration often involves registering the input image to the atlas, and then applying the reverse segmentation to the atlas's segmentation.

an artificial consensus template, which can be considered more representative of the dataset than any individual volume.

A number of atlases have been constructed for fetal and neonatal brains in MRI. Habas et al were the first to construct a spatiotemporal atlas of 2D slices [78] and reconstructed 3D volumes [79] of fetal brains (for $GA = 22 - 24$ weeks), and to use it to perform multi-atlas segmentation of tissue types. Kuklisova-Murgasova et al [43] generated a publicly available 4D probabilistic atlas over a wider range of gestational ages ($GA = 29 - 44$ weeks) that could be used to segment specific structures within the brain; however, this atlas was constructed using neonatal brains born preterm, and is therefore anatomically distinct from fetal brains. Most recently, Gholipour et al [2] proposed a 4D spatiotemporal atlas of the fetal brain spanning $GA = 19 - 39$ weeks, using 3D MRI scans of fetuses and producing atlas labels of tissue type and structure. This atlas can achieve segmentation quality comparable to human experts based on Dice coefficient (see Section 2.3.3).

MR atlases of the developing brain have also been used to characterise the underlying anatomy and clinical conditions. Habas et al [80] used atlas-based segmentation of 2D fetal MRI slices to quantify early cortical folding and hemispheric asymmetry. Gholipour et al [81] used a 4D atlas to segment fetal brain volumes with ventriculomegaly and obtain estimates of the ventricles' segmentation, and subsequently their volume and radius.

Work in ultrasound has been relatively more limited and is at an earlier stage, due to the higher difficulty of registering ultrasound volumes, the lower contrast, and other difficulties in interpreting the image (as mentioned in Section 2.1.1). Kuklisova-Murgasova et al [43, 82] performed what is to our knowledge the only known work in atlas-based segmentation in fetal brain ultrasound, by computationally generating a "pseudo-US" volume from an MR-based atlas and registering that to real ultrasound acquisitions, which is able to segment different brain structures (including many not visible in ultrasound acquisitions). Qiu et al [83] have also constructed an atlas of preterm neonatal brains to conduct segmentation of the cerebral ventricles and therefore measure their volume for clinical monitoring. Namburete et al have also proposed an ultrasound template based on the INTERGROWTH-21st dataset [3], which

is reviewed in more detail in Chapter 3.

2.3.2 Machine-learning based segmentation methods

Another class of methods to perform segmentation on fetal brain structures in MRI and ultrasound is machine learning [39]. This encompasses a wide range of techniques, but all are based on the premise that segmentation should not rely on human selection of hand-crafted features, but rather be performed via features derived organically through the machine's learning process. They rely on the use of relatively large quantities of labelled training data to derive classification rules independently.

One machine learning scheme for medical image analysis is the use of decision trees and random forests. A *decision tree* is a simple classifier, where the training dataset is split at several nodes using binary tests such that the output classes are separated from each other at the output with a high degree of confidence. Several (usually hand-crafted) features are typically derived from each data point to allow the classifier to discriminate more accurately in a higher-dimensional space. Decision trees, however, tend to suffer from *overfit*: can perform virtually flawless segmentations on the training data without generating good segmentations for any images from outside of the training dataset. To combat this poor generalisation they are typically implemented as *random forests.* A random forest is a number of decision trees, each trained on just a subset of the training data. This means that individual trees are statistically independent of each other, and the final classification of each pixel is obtained by a consensus of the predictions for all trees, which significantly reduces the impact of overfitting.

Much work has been done with segmentation of fetal brain images using decision trees and random forests. Kainz et al [84] proposed a random forest-based classifier to localise and segment fetal brain pixels in MRI images, using Gabor filters to provide features to guide segmentation. Yaqub et al [85] attempted a similar approach in 3D fetal ultrasound at $GA = 18 - 26$ weeks, using random forests to detect, and attempt to segment, several brain structures. Namburete et al [86] segment the fetal cranium in 2D ultrasound images at

$GA = 28 - 34$ weeks using a random forest framework. Cuingnet et al [87] similarly segment the fetal cranium for $GA = 19 - 24$ weeks using random forests to align acquisition to a common reference space.

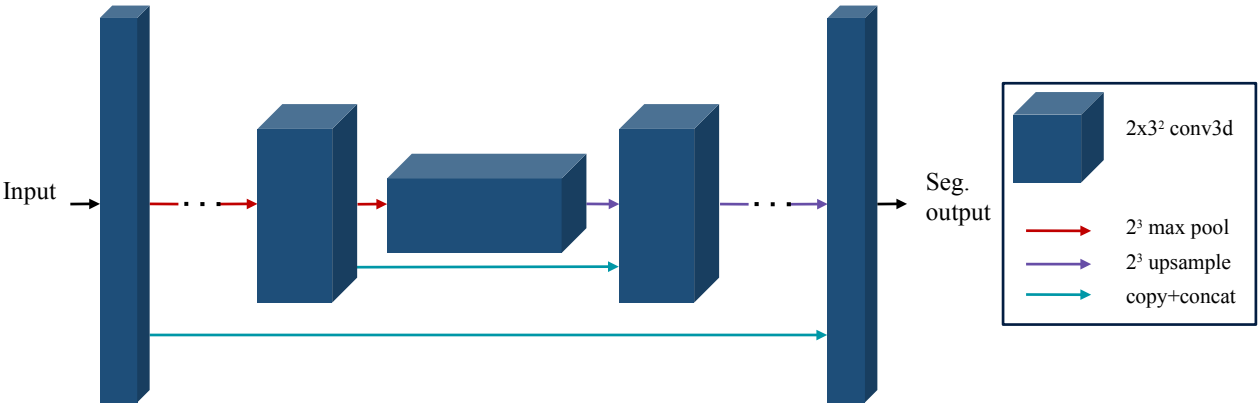


Figure 2.5: A 3D U-Net network structure. For clarity, only the first and bottom two convolutional layers in the downsampling/upsampling bath have been maintained.

One other family of methods that is still being explored is the use of convolutional neural networks (CNNs) to perform image segmentation. CNNs, like other neural networks, adjust weights between different nodes in "hidden" layers to produce an output most similar to the expected data label. They learn convolutional kernels to combine information from neighbouring pixels into feature maps, which can then be adjusted by passing them through a non-linear activation function and then (if necessary) on to further convolutional layers [88]. This allows the network to learn any feature descriptor within a certain scale, and therefore in principle it is able to learn how to reliably extract segmentations from arbitrary (if correctly labelled) data. Each pixel in the image can be classified according to the class that the network predicts for it. The learning, or training, process is guided by an objective function such as the ones listed in Section 2.3.1.

CNNs rely on trainable convolutional kernels. These have a certain visual field, typically 3×3px in 2D networks [89], and produce a single output per pixel which is a linear combination of the values of each pixel in their visual field. Nonlinearities are then added by an *activation function* that processes the output of the kernel, such as a sigmoid function or rectified linear unit (ReLU) [90]. A convolutional *layer* is a set of convolutional kernels, which each filter

the image and produce a different output. A neural network is a set of such layers, often chained sequentially such that the output of one layer serves as the input to the next layer. The output of convolutional layers may be modified by other functions before being passed onto the next layer, such as *max-pool* and *Dropout*. Max-pool is one way to subsample the output of a convolutional layer, to allow the next convolutional layer to process structures at a different scale. The image is split into smaller segments (often 2×2 px) and only the highest pixel value within each pool is kept, effectively lowering the pixel resolution of the output. Dropout is a different way to reduce the output, by randomly changing some outputs of a layer to 0 [91]. The subset of outputs dropped out changes in different batches, making it impossible for the network to rely on a single output or small set of outputs in intermediate layers. Dropout is typically used for regularisation during training, but can also be used at test-time in omni-supervised learning frameworks (see Section 2.4).

In practice, neural networks can be very large and contain several million tunable parameters to be sufficiently powerful to produce good segmentations [92]. This makes naive implementations of neural networks very susceptible to overfitting the training data, especially when the training dataset is limited [93]. This shallow learning of the features of the training set without identifying the underlying patterns can however lead to overfitting on the training dataset. This can be mitigated by making the training dataset larger, either by acquiring more manually labelled images or by performing *data augmentation*: artificially increasing the size of the labelled training dataset by performing random rotations, reflections, and small elastic deformations on the data [94]. It can also be addressed by changing the network structure itself: adding layers that reduce the number of parameters contributing to the output layer, such as dropout or max-pooling layers (where only the maximum output of a number of layers is passed on to the next layer). The images themselves can also be subsampled or split into smaller image patches.

The choice of objective function is also significant. The objective function is usually a measure of the difference between the desired output (the ground truth) and the generated output of the network, and must be differentiable for any small change in the network output.

A simple measure used in early work is the mean-square difference between the ground truth values and the generated classes, which is straightforward to compute. Other methods (such as the ones listed in Section 2.3.3) have been used, with the most popular for medical image segmentation being cross-entropy and Dice coefficient [95]. These generally seem to have similar performance across a wide variety of metrics [95].

A number of different network structures have been proposed for medical image segmentation. Earlier work with neural networks focused on the use of Hopfield neural networks [96], while later work relies on "deep" neural networks (involving more hidden layers between input and outputs) such as the fully convolutional network [97]. In 2015, Ronneberger et al proposed the U-net [94], a CNN structure developed specifically for the segmentation of biomedical images. The basic structure of the U-net is shown in Figure 2.5. It is inspired by the structure of the fully convolutional network, an earlier design [98], and it is able to extract features within images at a number of scales due to the use of several max-pool and up-sampling layers. The max-pooling layers allow for lower-frequency (and larger-scale) features within the image to be extracted, while combining the output from up-sampling layers to output of the previous layer of this scale allows a fusion of these layers. The U-net was originally a designed to process 2D data, but 3D variants have been proposed [99].

CNNs have found applications in brain image segmentation. Moeskops et al [100] propose a CNN to segment tissue types in brain MRI of preterm infants, and Rajchl et al [101] propose an automatic method to segment brain and other anatomical structures in fetal images from a weak "bounding-box" segmentation. Salehi et al [102] perform brain extraction in fetal MR volumes using a network structure based on the U-net. The presence of large motion artifacts in MR images, make analysis a challenging problem for a straightforward implementation of CNNs. CNNs have also started to be used in ultrasound image segmentation and interpretation in more recent publications, with promising results. Li et al [103] use a CNN based on the VGG architecture to reliably segment fetal volumes from the surrounding amniotic fluid in 3D ultrasound. Schmidt-Richberg et al [104] use a fully convolutional network to segment the fetal abdomen. Yu et al [105] also use a custom multiscale CNN to segment the fetal heart

in 4-D sequences. Namburete et al [3] use an FCN network structure to segment the entire fetal brain and eye and use their relative locations to align fetal brains to a common reference space.

Machine learning methods such as CNNs typically rely on large amounts of labelled data to achieve optimal performance However, these are difficult to create, as obtaining a reliable manually labelled dataset of medical images requires a great deal of time investment from a trained clinician. However, large unlabelled datasets do exist, as a number of large-scale studies or even medical databases are available (see Chapter 3). Several approaches have been proposed to obtain improved segmentations from this set of circumstances.

One approach to this problem is "weak supervision", which involves manually labelling the entire dataset with very simple manual annotations, like bounding boxes [106] or points [107] Rajchl et al [101] propose a CNN structure that can accurately segment MR images from unreliable manual segmentations from a crowd of non-expert raters, and a different neural network structure to generate segmentations from bounding boxes [108]. Zhang et al propose a method to use unlabelled datasets to improve the performance of networks with relatively little labelled training data [109]. Yang et al [110] propose a "suggestive annotation" framework where the unlabelled images which would most improve the performance of the network are identified and therefore the network can achieve competitive performance with only a fraction of the entire dataset. Radosavovic et al implement "omni-supervised learning", a method where many different models are used to enhance a small labelled dataset with a large unlabelled dataset and generate new training annotations [111].

Another category of algorithms, *unsupervised* methods, also exist, and do not require any labelled data. These are generally unsuitable for cerebellar or hippocampal segmentation: they are structures whose boundaries are defined by humans, and any attempt to segment them must therefore rely on examples annotated by humans. Therefore, these methods are not reviewed in this thesis.

2.3.3 Assessment of segmentation quality

There are several methods for quantifying the quality of a segmentation and compare proposed segmentation methods. These generally involve comparing the automated segmentation to a manually segmented "ground truth" image. A simple measure is pixel accuracy, where the class label of every pixel is compared to the ground truth and the percentage of correctly classified pixels is then calculated. This is very simple to implement and calculate, but may be misleading, especially in images with unbalanced classes (where more pixels belong to one class than others) where the pixel accuracy can be misleadingly high. Another method that is often used is the Dice coefficient [112]. The Dice coefficient DSC is defined as

$$DSC = \frac{2\left(L_{gt} \cap L_{seg}\right)}{L_{gt} + L_{seg}}$$

where L_{gt} and L_{seg} correspond to the label of interest in the ground truth image(s) and in the automatic segmentation, respectively. This more set-theoretic method is still easy to compute and can give class-specific accuracy. However, it can still be misleading for larger regions where the edges are segmented poorly, since the large central region still results in a high Dice coefficient. One measure that can overcome this is the Hausdorff distance, which measures the worst-case segmentation error [113]. This is a measure of the greatest distance $d(x, y)$ between the boundaries of the two labels, given by

$$d_H\left(L_{gt}, L_{seg}\right) = \max\left\{d\left(L_{gt}, L_{seg}\right)\right\}.$$

However, this is non-differentiable across segmentations which limits its usefulness as an objective function (more on this below and in Section 2.3.2). One final method that will be reviewed here is cross-entropy [114]. This is a method that is most useful when assessing a probabilistic segmentation method, which produces a "soft" segmentation with confidence scores of each segmentation class per pixel. This is a measure of the difference between the automatic and ground-truth segmentations in terms of the difference in information content

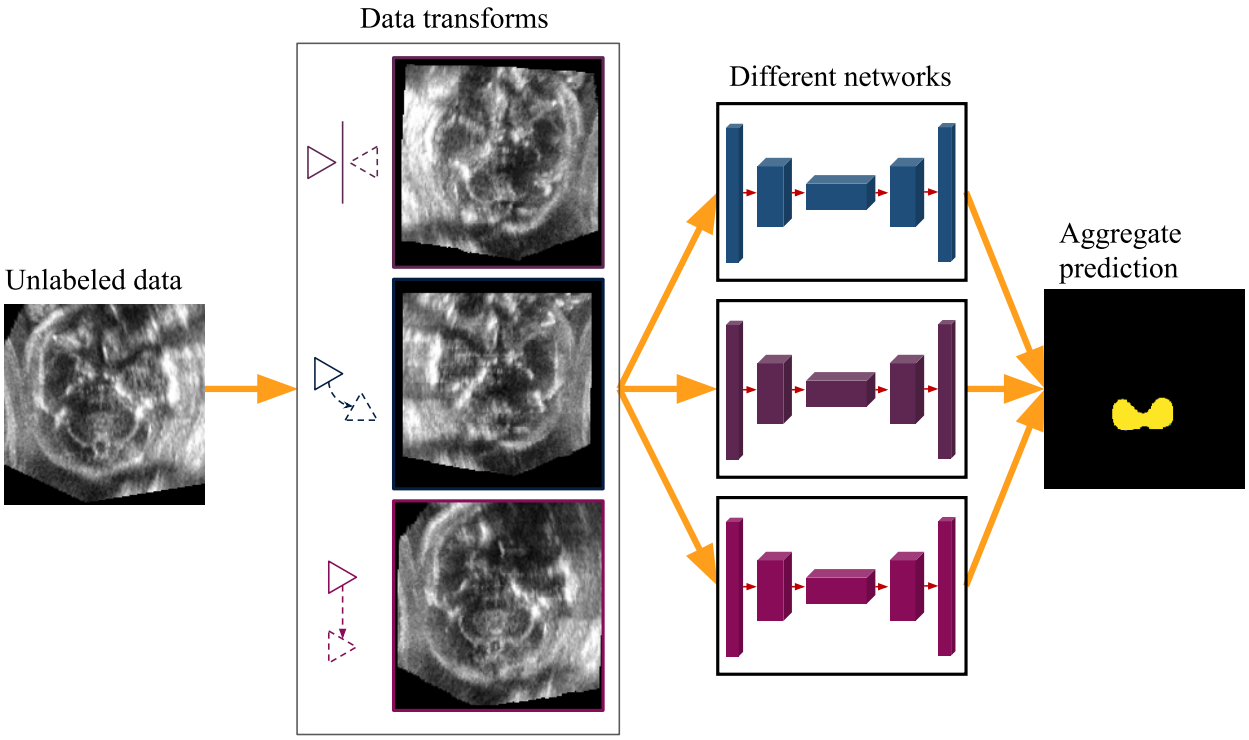


Figure 2.6: Data diversity (shown at the top, as multiple transforms of the same data) and model diversity (shown as multiple networks) can be used to obtain multiple candidate segmentations of a single image.

between them. It is measured as

$$H(p,q) = -\sum_{x} p(x) \log q(x)$$

where $p(x)$ and $q(x)$ are the per-pixel probability distributions of the two image labels. This is computationally efficient and differentiable, which makes it desirable for machine-learning based methods (as seen in Section 2.3.2).

2.4 Boosting omni-supervised learning

Omni-supervised learning is another proposed approach to use the unlabelled data to guide the learning process and improve segmentation performance relative to a network trained using only the small labelled dataset [111]. It does so by using automatically-generated segmentations of the unlabelled dataset as a guide to train the labelled data.

In general, training a neural network using predictions generated by itself does not improve its accuracy [106], and may often make it worse. However, the principle of *boosting* is well established: aggregating the predictions generated by multiple independent learners leads to a prediction superior to any of them [115, 116], and that aggregating predictions generated from different transformations (such as reflections and rotations) of the data can also lead to superior performance [93, 111]. In other words, in both cases, "a set of weak learners can create a strong learner" [117]. These two principles, of *model diversity* and *data diversity*, are what drive the concept of omni-supervised learning. Figure 2.6 shows how model diversity and data diversity can be used.

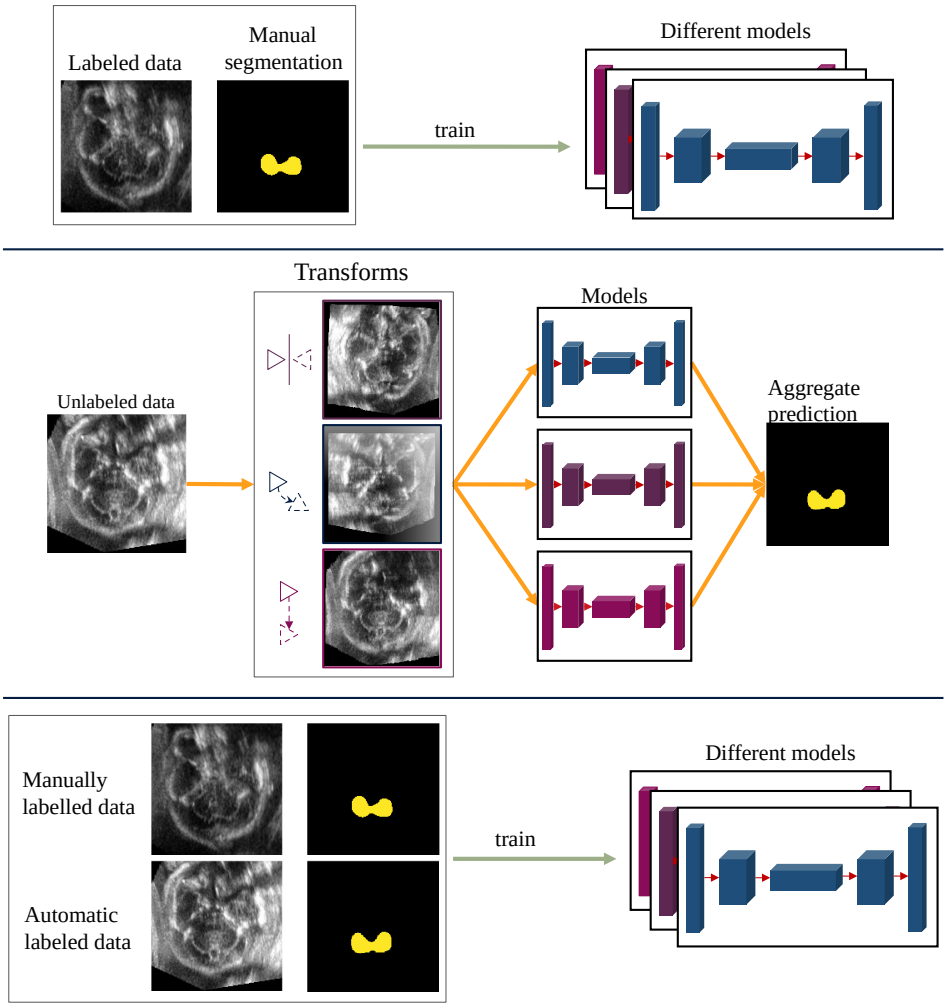


Figure 2.7: The omni-supervised learning algorithm. 1. A neural network is trained on the labelled dataset. 2. The network is used (with model and data diversity, see Figure 2.6) to obtain, and aggregate, multiple segmentations of the unlabelled data. 3. A subset of the unlabelled data is used with the labelled data to retrain the neural network.

Omni-supervised learning uses model diversity and data diversity to obtain multiple predictions on the same data. The outputs are then combined to form a consensus, which should be more robust than any one of the methods, as shown in Figure 2.7.

A subset of the unlabelled dataset is segmented in this manner, and the resulting segmentation is then used to retrain a new iteration of the network. This improves the resulting segmentations produced by the network, and this network is used to segment a new subset of the unlabelled data, and aggregate this segmentation to others as before to create new labels. This process is repeated until the entire dataset has labels, and is claimed to improve the quality of segmentations.

Only a subset of the predictions generated on the unlabelled data are used to seed the next iteration of training. Selecting an appropriate subset of predictions, therefore, is an important task, though often neglected in the literature. Previous work using deep NNs has either selected a subset of the unlabelled dataset at random [118] or used weak heuristics [111], such as ensuring the number of labelled pixels is similar to the average in the training data.

2.4.1 Quantification of uncertainty

A neural network that performs binary segmentation classifies each pixel as belonging to the target class (foreground) or background. This is typically done using a sigmoid activation function after the final layer of the network [94], so each pixel is given a "soft" classification ranging from 0 (confident background) to 1 (confident foreground). This allows the network to express some uncertainty in its segmentation, but neural networks also often confidently make wrong predictions [119], especially when they are presented with noisy or ambiguous data, or when the data presented to them differs in appearance from data they have been trained on. The latter case, in particular, leads to overconfident predictions (often confident and wrong).

A distinction can be made between the two sources of error: (1) error caused by image noise, ambiguity, and inconsistent labelling in the training set, and (2) error caused by the

classifier being shown data that appears different to that in the training set. The first type, known as *aleatoric uncertainty* [120], is due to ambiguity and noise in the data itself. Since this is a measure of genuine uncertainty in the data, this type of uncertainty is unlikely to reduce with increased training data. On the other hand, the second source of error, known as *epistemic uncertainty* [120], is due to the model parameters, due to overfitting or out-of-sample effects. This decreases with increasing training data, as the model's predictions become more consistent. The amount of epistemic uncertainty can therefore give a measure of how different any given example is from the training data, and the level of confidence in the network's predictions. A comparison of these two types of uncertainty is given in Table 2.1.

Gal and Ghahramani [121] have proposed a method to estimate the uncertainty present in the model, using *dropout uncertainty.* Dropout is typically used in neural networks as a form of implicit regularization [122] [91], to make training more robust to overfitting. This is done by randomly zeroing (or "dropping out") a given proportion of the weights in certain layers at training time, and changing which weights are dropped out at each iteration. This prevents the network from putting excessive emphasis on a small number of weights, which can lead to overfitting. In traditional implementations of dropout, this is disabled at test time to use the full learned network architecture to obtain the best prediction.

Gal and Ghahramani propose using dropout at test time, and generating multiple predictions for the same data with the same network only differing in which weights are dropped out. The differences in output across different runs can give a measure of the model's uncertainty. Large differences in output across runs suggest that the model is very uncertain of its output. Gal and Ghahramani originally proposed this method as applied to a classification task, but it can be applied to any network architecture that includes dropout. Kendall et al [123] were the first to propose a network to use dropout uncertainty in semantic segmentation, known as Bayesian SegNet. Kohl et al have also formulated a version of the U-Net segmentation network, known as Dropout U-Net [124], which includes dropout to estimate dropout uncertainty.

Wang et al [125] propose using test-time augmentation to improve estimates of aleatoric uncertainty. The principle behind these differing approaches comes from the different sources

Aleatoric uncertainty	Epistemic uncertainty
Input-based	Model-based
Caused by noise in data and inconsistent labels	Caused by overfitting, and data appearing different from the training set
Does not improve with more training data	Improves with more training data
Measured by varying the input	Measured by varying the model

Table 2.1: A comparison of the two different types of uncertainty in machine learning, aleatoric and epistemic uncertainty.

of the uncertainty. Epistemic uncertainty is introduced by the model's parameters, so making changes to the model is of use in estimating it. Aleatoric uncertainty is introduced by the data, so it follows that varying the data is what can be used to measure the uncertainty in it.

Omni-supervised learning requires data diversity and model diversity, which are also methods to measure aleatoric and epistemic uncertainty, respectively. I believe that these two measures can inform the selection of data for omni-supervised models, without adding complexity.

2.5 Usage of auxiliary features

Machine learning is an approach to optimisation problems in a high-dimensional feature space. Many machine learning methods explicitly incorporate a large number of features for each sample, and let the algorithm learn the relative contribution of each one. Predictive models based on decision trees provide an early example of this: decision trees rely on large numbers of different features to provide a high-quality classification. Similarly, artificial neural networks combine multiple separate input features to produce an often low-dimensional output. Hidden layers combine the input features in a nonlinear way to achieve outputs as close as possible

to that desired. A 3D CNN performing binary segmentation on an $N \times N \times N$ volume can therefore be thought of as classifying each pixel based on N^3 different features.

Convolutional neural networks, however, generate outputs from image data provided at input. This is due to the nature of convolutional layers, which rely on an image-like grid of spatial features and output a similar image-like grid. While there has been some work extending CNN architectures to arbitrary graphs [126], it remains impossible to input scalar data directly into a convolutional layer.

There have been attempts to combine image data inputs with additional scalar inputs. Yap et al train a classifier for skin lesions using both photographic image data and patient metadata (such as age and clinical history) [127]. Their network architecture processes the image data through several convolutional layers, and then into a fully-connected layer to which the scalar features are concatenated. This combined feature set is then processed through further fully-connected layers and then combined into a single categorical output. Ahsan et al use a similar approach to diagnose COVID-19 from X-ray data and patient metadata [128]. They use two separate machine learning models to process the different data formats (a multi-layer perceptron for numerical data, and a CNN for image data) and then use their outputs as inputs for a third model that produces an overall classification.

These methods rely on a combination of convolutional layers with fully-connected layers, to produce a numerical output. The state of the art in medical image segmentation, however, is fully-convolutional networks such as the U-Net [94], where only convolutional layers are employed to retain spatial information. Fully-connected layers cannot be used at any point in the processing pathway of these networks, as that would entail the loss of important spatial context for segmentation and degrade performance. Therefore, there has to my knowledge been no previous work on integrating image and non-image data in a fully-convolutional architecture.

While image data alone encodes much information, there is evidence that humans generating the training data often rely on other, non-image features. Human sonographers, who generally are aware of their subject's gestational age, exhibit expected-value bias while

measuring biomarkers in routine fetal ultrasound scans [129]: their estimates of clinical measurements are closer to the average reported for that gestational age than when they are blind to the subject's gestational age. Similarly, sonographers who look at US scans suspicious of a given diagnosis (such as scans referred to them for suspected hypoxia) are more likely to estimate measurements closer the expected range for that diagnosis, relative to sonographers without access to that information [130].

This is also likely to apply in the generation of training data: clinicians have access to other information (such as age and previous medical history) which informs how they generate an image segmentation used to train a model. This can be interpreted as a bias (as it is in the papers cited in this section), or as a way to improve the quality of their annotations by integrating non-image data. In either case, this is data which fully-convolutional CNNs do not typically have access to, which is used in training data generation. Finding a way to include this information in a fully convolutional CNN would give those networks the same access to information as humans.

There has only been limited work trying to include non-image features in a fully-convolutional CNN. One such effort is CoordConv [131], which concatenates an additional image encoding location coordinates as an input. The authors add a channel to each convolutional layer to encode spatial coordinates in the network's processing, to aid a simple localisation task. This is a way for the network to incorporate additional, non-image data, which is nonetheless presented in a format that can be processed by a convolutional network.

Chapter 3

Datasets and annotations

Chapter overview

This chapter provides a brief overview of the data that is used throughout this thesis. The same datasets are often used in different chapters, so this chapter provides a reference description of the acquisition and properties of certain datasets. Section 3.2 describes the medical imaging datasets used in this thesis, and Section 3.3 covers the sources for the annotations used for training and testing.

3.1 Introduction

Machine learning methods, particularly neural networks, require large labelled datasets to use as training and testing data. This can be a challenge for medical applications: it can be difficult to obtain large sets of medical data, and more difficult to obtain high-quality annotations for experts. Larger medical image research datasets are often collected from multiple centres, as part of large collaborations. These are collected for specific studies (such as the INTERGROWTH-21st dataset discussed in this chapter) and later made available for other research, or occasionally collected explicitly to make large datasets available to researchers (such as the UK Biobank [132]). These datasets often come without manual annotations, such as manual segmentations, which are not a clinical standard.

These constraints mean there are only a small number of medical image datasets that are sufficiently large and well-labelled to use to train machine learning methods.

3.2 Thesis datasets

This section covers the datasets used in this DPhil project, while Section 4.3 covers manual annotations. The techniques covered in later analysis chapters lend themselves better to different subsets of these datasets, or to different starting annotation. The choice of particular data subsets or annotation schemes, and the justification for them, is covered in more detail in each individual chapter. Most of the methods in this thesis were developed for the INTERGROWTH-21st 3D ultrasound dataset (described in Section 3.2.1), while the ADNI MRI dataset described in Section 3.2.2 was used to validate these results.

3.2.1 3D ultrasound dataset

The 3D ultrasound INTERGROWTH-21st dataset is used throughout this thesis. The INTERGROWTH-21st study was a longitudinal and multi-centre study that acquired multiple ultrasound acquisitions from 4321 optimally healthy pregnancies [18]. To ensure optimal health, strict eligibility criteria were used. Eligibility criteria for mothers included nulliparity[1], regular 26-30 day menstrual cycles prior to pregnancy, no use of hormonal contraception in the months prior to pregnancy, and spontaneous conception [133].

Three-dimensional (3D) ultrasound volumes of the fetal body were acquired [134] at centres in eight countries using a consistent acquisition protocol. This study set international standards for fetal growth, and the very large quantities of data acquired for this purpose have been used for image analysis studies [11]. Each volume was labelled with the gestational age of the fetus, as measured in days from the last menstrual period and confirmed by ultrasound measurements of the crown-rump length. Participants were invited for scans every 5 weeks from $GA = 14$ weeks to delivery. These were imaged according to a standard protocol: transabdominal ultrasound was acquired in a lateral recumbent position, with sufficient

[1]Participants needed to have had no previous pregnancies.

magnification to ensure that the structure of interest is fully present and spans at least 30% of the image [135]. At each visit, three 3D ultrasound volumes were acquired of the head, abdomen, and femur. Only the acquisitions of the fetal head were used in this thesis.

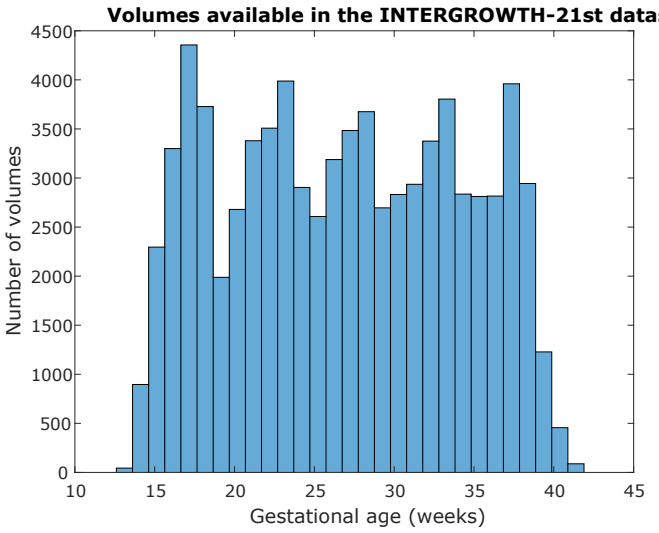


Figure 3.1: The distribution of gestational ages of the 3D ultrasound volumes available in the INTERGROWTH-21st dataset.

Figure 3.1 shows the age distribution of 3D ultrasound volumes available in the dataset, and shows a large and fairly even number of volumes in the dataset across all gestational ages. Not all volumes are usable for the purposes of this study: some do not contain the entire volume of the brain, while others do not show enough detail or exhibit large shadowing artifacts. Due to interference from the fetal skull, in each volume only one hemisphere is visible with a good level of detail: this is explained in Section 2.3. Some sessions also did not include a 3D acquisition of the fetal brain, due to consistently unfavourable positioning of the fetal body. The imaging protocol required fetuses to be positioned on one side but did not specify which during acquisition, so both left and right hemispheres are represented in the data. Approximately two-thirds of fetuses lie to their left at pregnancy term, though the left-right positioning of fetuses earlier in pregnancy is not well-studied [136].

The raw volumes include much of the fetus and some surrounding maternal tissue (one example slice of this can be seen in Figure 3.2). To reduce the large differences in head and probe positioning and fetal size and orientation between acquisitions, all volumes were brain-extracted and the skulls were registered to each other following the method laid out

Figure 3.2: An example of a slice from a raw volume in the dataset. Most of the image space is given to maternal tissue surrounding the fetal head.

by Namburete et al [137]. The volumes were aligned to a common reference frame by using parametric curves to find consistent points in the skull, which are clearly visible in ultrasound. These points are then used to guide a similarity transform which scales and rotates the individual acquisitions to the same reference space. The resulting volumes, after preprocessing, had a consistent image size of $160 \times 160 \times 160$. Since there is significant variation in fetal head size over the period considered in this thesis, this leads to varying physical dimensions for each voxel: though the initial acquisition has a resolution of 0.6mm [134], after scaling the physical space taken by each voxel ranges between approximately 0.5 and 1.1mm.

Heavier preprocessing was used for certain parts of this project. As only one hemisphere is visible in any given volume, earlier parts of the project mirrored volumes across the mid-sagittal plane to simulate a consistently visible brain. It can also be useful, and involve less preprocessing, to treat volumes with a visible left hemisphere separately from volumes with a visible right hemisphere.

3.2.2 Adult MR dataset

In Chapters 5 and 6, an additional MRI dataset was also used to validate the methods developed in a different imaging modality. The Alzheimer's Disease Neuroimaging Initiative (ADNI) [138] is a collaboration to collect data from a large number of older adults. 63 sites participated in acquisition, using different scanners and acquisition methods. [139]

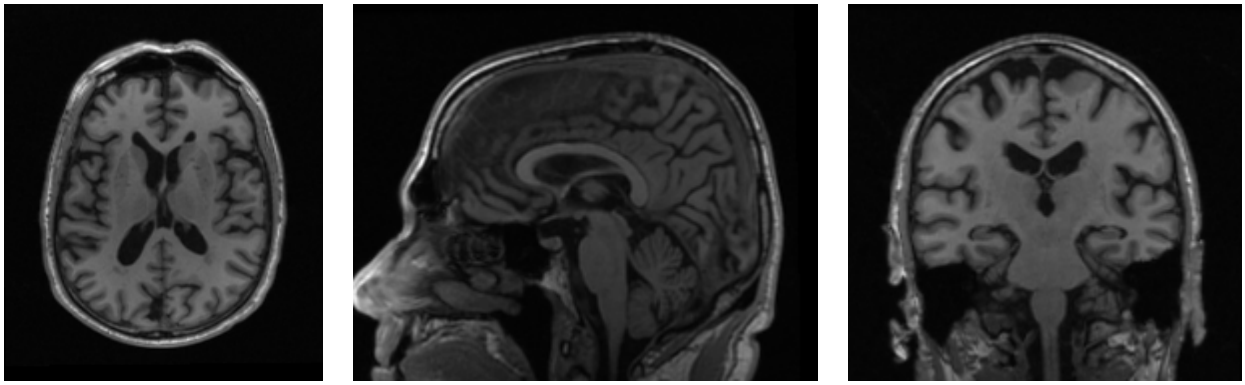


Figure 3.3: Three canonical views of an MRI acquisition of the brain of an 85 year old cognitively normal male, from the ADNI dataset.

The ADNI dataset was collected to identify biomarkers for the detection and tracking of Alzheimer's disease (AD). Alzheimer's disease is a common ($> 20\%$ for those over 75 years old [58]) progressive neurodegenerative condition and one of the main causes of death in the developed world [58], but it is in many ways poorly understood: it remains difficult to predict an individual's risk of developing AD, diagnose AD in patients with symptoms, or devise possible treatments for it. Dementia caused by Alzheimer's disease is typically preceded by a milder form known as mild cognitive impairment (MCI), but cognitive symptoms are often predated by abnormalities that can be seen in medical images [60].

The ADNI dataset is comprised of many different types of clinical, genetic, imaging, and diagnostic data, collected from many different sites using different equipment and different protocols [139]. The imaging data available includes MRI data collected from different scanners, with different sequence weightings and acquisition settings. Acquisition characteristics can introduce significant bias, so this must be taken into consideration for any proposed machine-learning method . Even when selecting volumes with similar acquisition characteristics on paper, it is possible that different postprocessing methods may have been employed at different sites.

The subset of ADNI considered in this thesis comprised only T1-weighted 3D acquisitions of the brain, imaged using a scanner field strength of 1.5T and with a resolution of 1mm along each axis. Figure 3.3 shows three canonical views of one such acquisition, with no additional postprocessing.

None of the volumes in the initial ADNI dataset have segmentation labels, though they are labelled with metadata such as scanner settings, age, and cognitive status.

3.3 Atlases and annotations

The datasets described in Section 3.2 are collections of medical images, labelled with some metadata (such as age and acquisition method) but without segmentation labels. Since this thesis is largely focused on medical image segmentation, availability of annotations for segmentation is very important.

While it is possible to label volumes manually, this is very time-consuming at the level of the datasets discussed above: the exact time needed to generate labels manually varies depending on image modality and structure of interest, but obtaining consistent segmentation labels for an entire dataset on the scale described is a project on the scale of several months. If any labels for these datasets already exist, this may reduce or eliminate this time investment: this section examines what labels were available for the DPhil research.

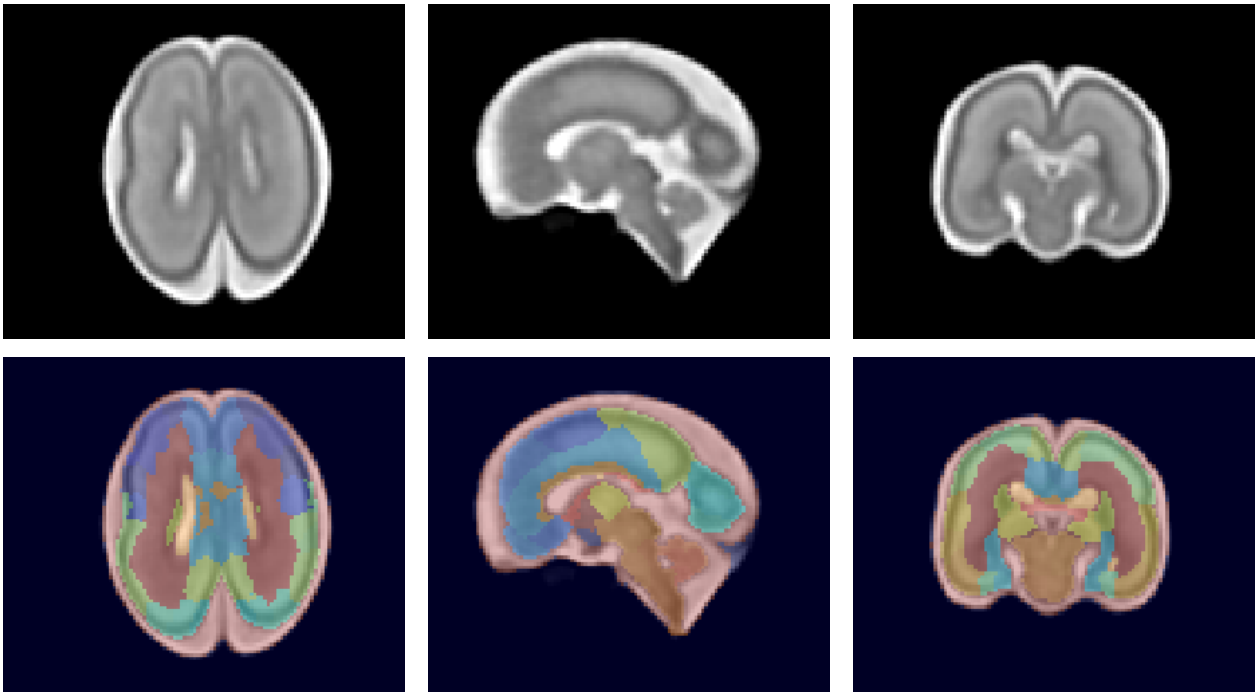
In this thesis, I use *template* to describe a common reference space that volumes can be registered to, such as the MNI152 template [140]. I use the word *atlas* to describe a segmentation for a template. The usage of these words differs in different parts of the literature (see e.g. [140] and [2]), so it is important to use a consistent standard throughout this thesis.

3.3.1 Gholipour fetal MR atlas [21-37 GW]

Gholipour et al [2] constructed an MRI-based spatiotemporal atlas of the fetal brain[2], with structural segmentation labels at a range of gestational ages between 21-37 GW.

This spatiotemporal atlas was constructed using 81 fetal brain MRI volumes in this age range, acquired from healthy fetuses at 1.5T and 3T with a slice thickness of 2mm. These acquisitions are repeated T2-weighted half-Fourier acquisition single shot fast spin echo (T2wSSFSE) scans As explained in Chapter 2, fetal motion is a major challenge in acquisition of fetal MR images: all MRI acquisitions were processed with retrospective motion correction

[2]Available at http://crl.med.harvard.edu/research/fetal_brain_atlas/.



(a) (b) (c)

Figure 3.4: On the top row, the MRI atlas at 23 weeks, and on the bottom row, individual structural labels, as seen in the (a) axial, (b) sagittal, (c) coronal views.

[141]. Following this, they were brain-extracted and normalised, and then registered to a common reference space. This common space was constructed as a 4D space (3D + time) using diffeomorphic registration, where each subject contributed to the atlas at every age point in proportion to the difference between its own age and that point. The atlas was then sampled at one point per gestational week, to provide individual volumes for each age point.

Each age point was then manually labelled over a large number of tissue-type and structural labels - a total of 135 different labels. Labelling was done automatically using a neonatal template [42] at higher gestational ages (>35 weeks), and manually at lower gestational ages.

The small number of volumes used in the construction of this atlas (only 81 across a range of 17 gestational weeks) makes the atlas susceptible to individual variation in anatomy, and potentially unrepresentative of a larger population of developing fetuses.

3.3.2 INTERGROWTH-21st 3D fetal US template [14-30 GW]

Namburete et al [3] have generated a template using the INTERGROWTH-21st dataset, providing a common reference space across a gestational age range of 14-30 GW. This is a lower age range than that of the Gholipour atlas, reflecting the lower ages at which ultrasound imaging can be applied.

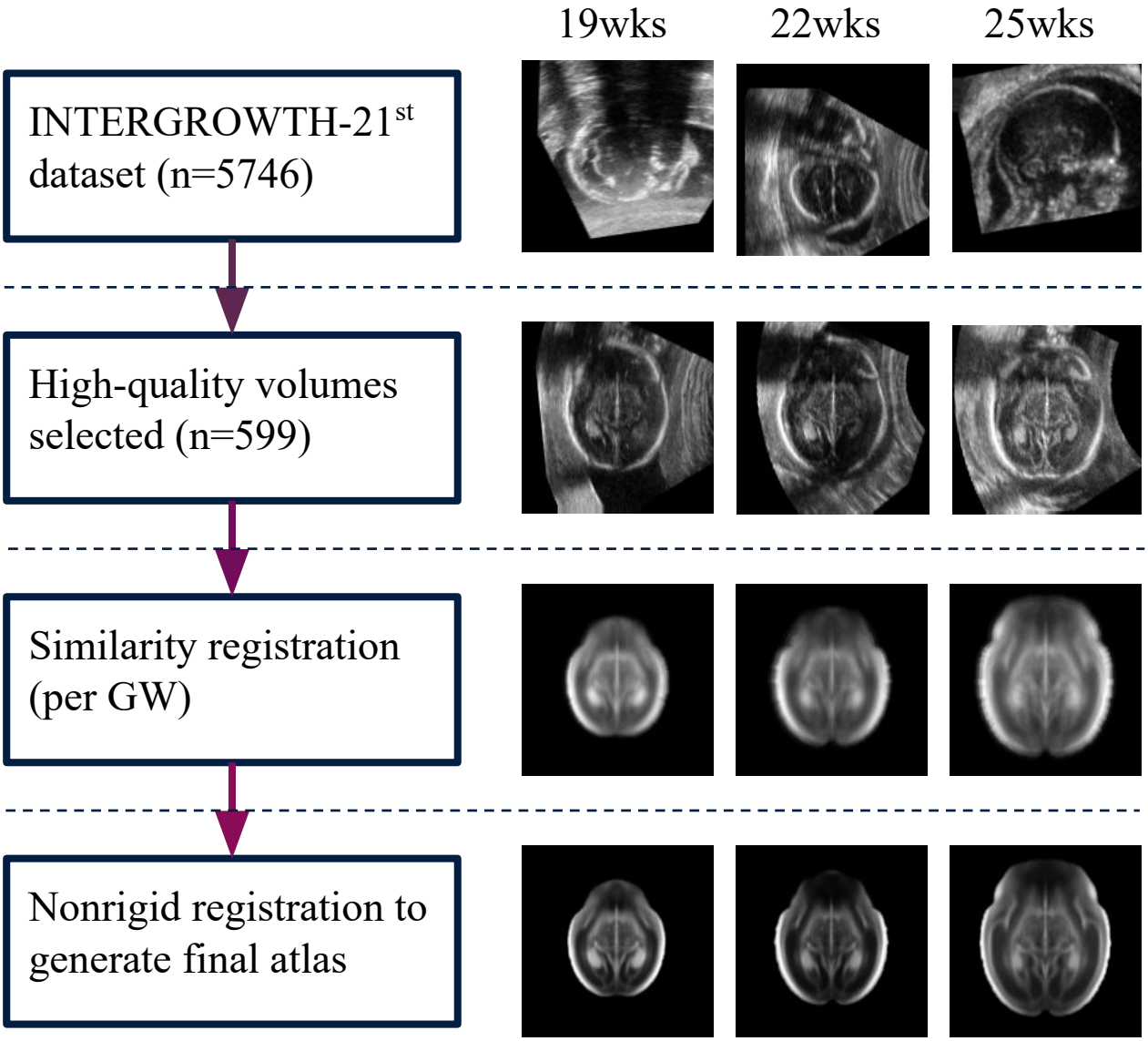


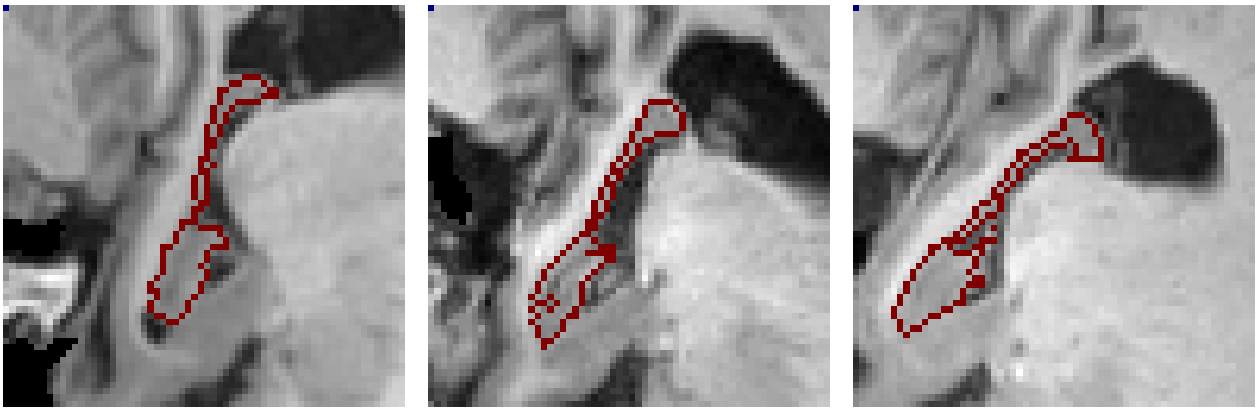
Figure 3.5: Flowchart showing the process by which the INTERGROWTH-21st US template was constructed.

The flowchart in Figure 3.5 shows how this template was constructed. At first a subset of volumes in the INTERGROWTH-21st dataset was selected (599 volumes across the gestational age range of in the template), which had good image quality and few artifacts. Since only one

hemisphere can be consistently visualised in fetal ultrasound [11], these were separated into left hemisphere-visible and right hemisphere-visible volumes. Within each age window and for each hemisphere, an initial 3D image registration was done using a similarity transform, to estimate a linear transformation for all volumes to a standard space. Then a groupwise diffeomorphic nonrigid Demons registration [142, 75] was used to obtain a visually sharp, consensus volume for each hemisphere. Finally, the templates for each hemisphere were combined into a single, complete template. The resulting template is asymmetrical, similarly to real brains.

This template does not have any structural labels, but provides a common reference space for volumes at each gestational age. The transforms performed on each individual volume used to construct this dataset are reversible and their parameters are stored: if labels are generated within this common reference space, they can then be propagated to all the individual volumes that are used to construct the template. It is possible, therefore, to use a single segmentation of the template to generate labels for each volume used to construct the template.

3.3.3 HARP



(a) (b) (c)

Figure 3.6: Example segmentations from the EADC/HARP dataset, showing subjects labelled as (a) cognitively normal, (b) MCI, (c) AD.

The EADC-ADNI HARmonised Protocol (HARP) dataset [20] is a subset of the ADNI dataset. The dataset consists of 135 T1-weighted 3D MRI volumes acquired with 1.5T field strength and $1 \times 1 \times 1$mm voxel dimensions. These were split between cognitively normal

(CN), mild cognitive impairment (MCI), and probable Alzheimer's disease (AD) volumes. 100 volumes are pre-designated as "training" and 35 as "test" volumes.

The hippocampi within each volume were manually labelled by five clinical experts following a harmonised protocol [20]. Each hippocampus was labelled slicewise in 2D, along the coronal axis. Where different raters disagreed, the majority annotation for each voxel was used. Figure 3.6 shows an example of segmentations from three raters, in the sagittal view.

There is a hippocampus in each hemisphere of the human brain, so each acquisition has two hippocampi. Therefore, the HARP dataset contains 270 volumetric labels of individual hippocampi across 135 volumes. In this thesis, the region around individual hippocampi is cropped, so each volume passed to the neural networks used here contains only one hippocampus.

The labels in the HARP dataset cover only a fraction of the ADNI dataset, which is much larger. Chapter 5 of this thesis explores the use of additional unlabelled volumes from ADNI alongside the labelled HARP volumes.

3.4 Discussion and conclusion

The size of the INTERGROWTH-21st 3D ultrasound dataset, and the fact that it was collected using a standard protocol, make it a suitable dataset to train a machine learning method. The lack of high-quality segmentation labels remains a constraint to what can be done using this dataset: Chapter 5 is dedicated to harnessing the large quantities of unlabelled data present in this dataset.

The ADNI dataset covers a different imaging modality (MRI), with different imaging characteristics. The HARP dataset provides high-quality manual segmentation labels for a small part of the ADNI dataset. The larger ADNI dataset contains many more MRI acquisitions with the same acquisition parameters, which remain unlabelled. Chapter 5 explores ways in which segmentation performance can be enhanced with use of this large unlabelled dataset.

This short chapter has introduced the datasets and annotation schemes used throughout

the rest of this thesis. The INTERGROWTH-21st dataset is used in every chapter of this thesis, with different annotations and in different subsets: this is detailed in each chapter. The ADNI/HARP dataset is used in chapters 5 and 6 together with the INTERGROWTH-21st dataset. The annotations and atlases considered in this chapter are also used in later chapters: Chapter 4 covers atlas-based methods using the Gholipour atlas and the Namburete template, while Chapter 5 uses manual annotations and annotations in the HARP dataset.

Chapter 4

Atlas-based label generation and multi-label segmentation

Chapter overview

This chapter uses atlas-based methods to segment a large number of volumes and uses CNNs to extend the number of segmentations to volumes not used in the original construction of the atlas. The main challenge addressed in this chapter is how to obtain high-quality automated 3D segmentations in a large ultrasound dataset without manually-labelled training data.

I begin by introducing the task and motivating the need for a solution, in Sections 4.1 and 4.2. I use affine registration to align two atlases of the developing fetal brain, one obtained from ultrasound and the other from MRI data, in Section 4.3, and propagate the structures labelled in the MRI atlas to the ultrasound atlas. I then propagate those segmentations in turn to the individual ultrasound volumes used to construct the ultrasound atlas in Section 4.3.3, and quantify the reliability of that segmentation (see Section 4.3.4). Since this method can only be used for volumes with known correspondences with the atlas, it cannot generalise to new data, so Section 4.4 then uses those volumes to train a CNN. I consider the design choices that lead to the best CNN-based segmentation. I present the results in Section 4.5 and discuss them in Section 4.6.

A preliminary version of the work presented in this chapter was accepted for presentation at

MIUA 2019 [21]. A further publication based on the full work presented here is in preparation.

4.1 Introduction

Development of the fetal brain can be imaged and tracked using ultrasound (US) and MR imaging. Ultrasound is a more traditional technique: fetal US has been a standard screening test for pregnant women for decades in much of the developed world. In the UK, women are offered at least 2 US scans during pregnancy, one at 8-14 GW (known as the dating scan) and another at $18-21$ GW (known as the anomaly scan) for standard screening purposes [143][1].

These scans are used as a screening tool: the measures used for such screening are either qualitative and subjective (such as an observation of specific anatomy), or measured manually by the clinician (such as the head circumference, or the transcerebellar diameter [143, 33, 144, 145, 32]). In both cases, these rely on the clinician's judgement and experience, and may show expected-value bias [129]. Since they need to be measured manually, the time available for each scanning session limits how much information may be collected from each scan: manual measurements obtained by the clinician are usually simple lengths and circumferences, based on labelling of a small number of points.

More detailed measurements, such as delineation of individual anatomical structures on each scan, may retrieve more usable information, but are time-prohibitive for clinicians to conduct routinely and are thus not part of clinical guidelines. If such measurements could be obtained automatically, however, they could be of clinical value without posing additional burdens on sonographers.

This DPhil project is an attempt to automatically extract such measurements from US scans, enabling these measures to be extracted automatically. This chapter shows one approach to obtain high-quality 3D segmentations without relying on large quantities of manually-labelled data.

[1]Additional US scans in the third trimester are offered for women with additional risk factors, such as high BMI or smoking.

4.2 Proposed approach

This chapter considers multi-label segmentation in 3D US volumes. It tackles two problems: the lack of high-quality manual segmentation labels to construct a training set for a neural network, and the choice of an appropriate architecture for a CNN to perform supervised multi-label anatomical segmentation. The INTERGROWTH-21st dataset, used in this chapter, does not include segmentation labels, so the first task is to generate ground-truth labels to use for training. Manual segmentation is very time-consuming for even individual volumes[2], so it is infeasible to manually segment a large dataset to initialise model training. The first challenge this chapter considers is how to automatically generate sufficient ground-truth labels efficiently and of high enough quality to serve as a training set for a machine-learning method.

Firstly, I explore atlas-based methods to generate high-quality segmentations. I examine the available atlases for the fetal brain, including which segmentation labels are available for each atlas. I consider different ways to register those atlases and to propagate segmentations to each individual volume. I discuss the quality of ground-truth segmentations obtained by the atlas-based methods chosen and find a way to evaluate them quantitatively. Finally, I use the final choice to generate individual segmentations for a large sample of volumes within the INTERGROWTH-21st dataset.

After generating ground-truth segmentations, I experiment with different network architectures for segmentation. The atlas-based method cannot scale to data not used in initial atlas construction, so a CNN is used to do so instead. This can be posed as a 3D multi-task segmentation problem, and it is worth exploring a range of available architectures to find the best-performing type. Section 4.4.3 discusses possibilities: the U-Net architecture [94], an encoder-decoder architecture discussed in Section 2.3.2, has proven successful for medical image segmentation problems, but there are several possible implementations of U-Net-based architectures for this type of task. The network architectures evaluated are:

- A native multi-task 3D network architecture

[2]Even segmentations of a single structure, performed in Chapter 5, were estimated to take 30 minutes per volume.

- A series of single-task 3D network architectures
- A QuickNAT-like [146] series of 2D network architectures, with results aggregated to obtain 3D segmentations.

These architectures have different computer memory requirements and training times, and the possible advantages of each are considered in Section 4.4.3. I then compare the quality of these segmentations, as well as the practical usability of the architectures selected. Finally, I examine particular failure modes in detail.

4.3 Label generation

Given the size of datasets necessary to train a deep-learning method, subject-specific anatomical characteristics, and artifacts inherent in US imaging[3], it is challenging and time-consuming for human experts to manually segment a sufficiently large subset of the dataset used for this chapter.

Manual segmentation of 3D structures in the INTERGROWTH-21st dataset takes an average of 2 hours per volume [118]. Segmenting a training set of 500 volumes would therefore involve a time commitment of 1000 hours, or 25 weeks (6 months) of full-time work (40 hours/week)[4]. This is not a productive use of time in the context of a single DPhil project, so alternative methods were explored.

I used an atlas-based method to generate a large number of weak labels to compensate for this. Atlas-based segmentation relies on a segmentation (the *atlas*) of a single volume (the *template*) to which individual volumes have been registered. Each individual volume has a known transformation to align it to the common reference space of the template. This transformation can then be applied in reverse to the atlas, generating a segmentation for each volume. The advantage of this approach is that no manual segmentations are required, beyond the initial atlas: given a transformation from each volume to the template, it should

[3]Such as speckle and shadowing artifacts, which make it difficult to accurately trace anatomical boundaries.

[4]Other large-scale image datasets, such as ImageNet [15], crowdsource their annotations using services like Mechanical Turk. This is not feasible for medical image datasets, where specialist medical knowledge is required to make accurate segmentations.

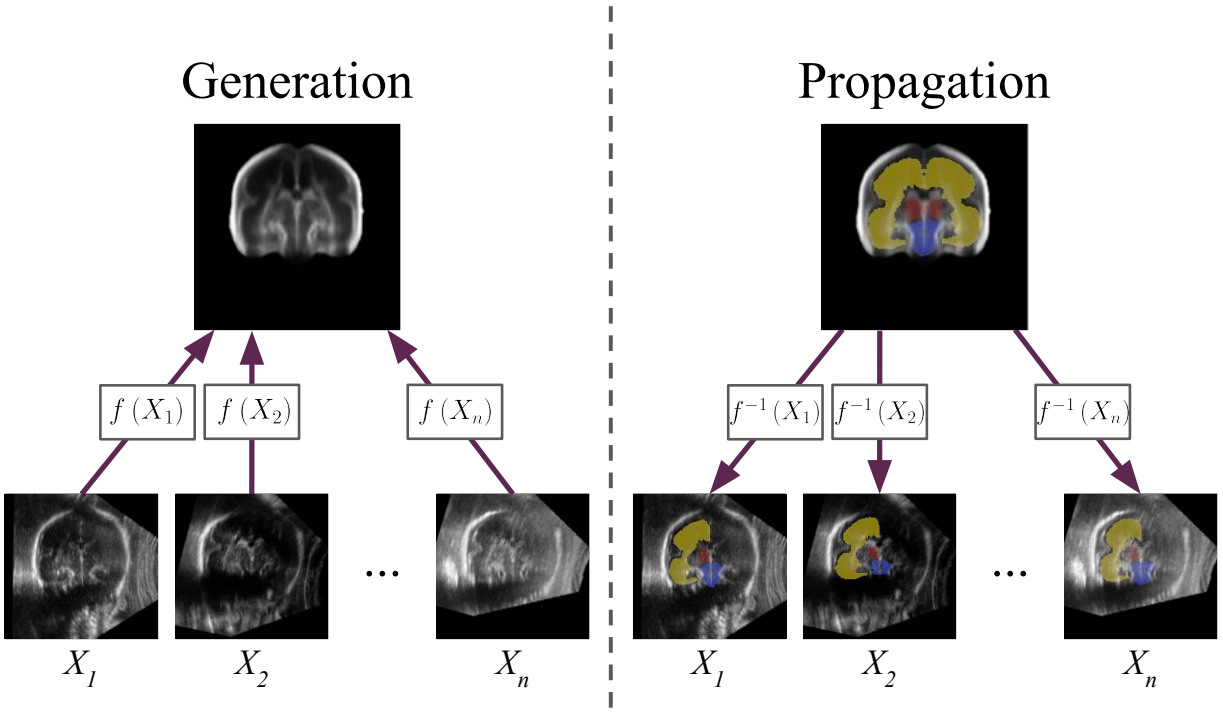


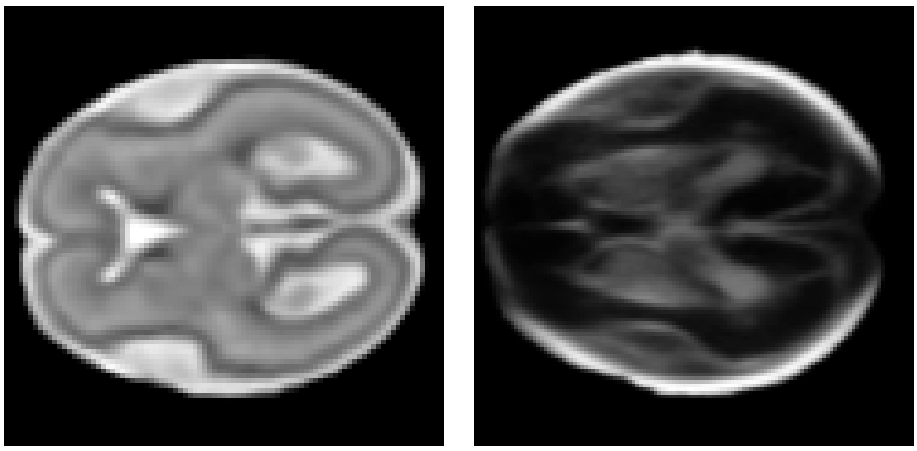
Figure 4.1: A flowchart showing the principle of atlas-based segmentation. A set of volumes $X_1, ..., X_n$ are aligned to a common reference space with a set of transforms $f(X_1, ..., X_n)$. A segmentation in the reference space can then be propagated to each initial volume by inverting the transforms, $f^{-1}(X_1, ..., X_n)$.

be straightforward to transform the atlas labels back to each individual volume. Figure 4.1 illustrates the principle of this method, and Section 2.3.1 reviews atlas-based segmentation in greater detail.

4.3.1 Spatiotemporal atlases

As discussed in Chapter 3, Gholipour et al [2] recently proposed a 4D spatiotemporal atlas of the fetal brain spanning $21 - 37$ gestational weeks (GW), using 3D MRI scans of fetuses and producing atlas labels of tissue type and structure. This atlas has been shown to achieve segmentation quality comparable to human experts based on the Dice coefficient [2]. The labels from this atlas can be used as a starting point for an atlas-based segmentation approach, similar to what was done by Guha Roy et al [147] for the segmentation of brain structures in MRI with limited annotations[5].

[5]They used the FreeSurfer tool [148] to automatically generate training segmentations, with only small manual corrections on a subset of volumes.



(a) (b)

Figure 4.2: A comparison of two spatiotemporal fetal brain atlases at 22GW. (a) the Gholipour MRI atlas [2], and (b) the Namburete US atlas [3]. The different imaging modalities create different contrast and emphasize different structural features.

Chapter 2.4.1 also introduced an US-based template has also recently been generated by Namburete et al [3] using the INTERGROWTH-21st dataset. This spans gestational ages $14-30$ GW, using selected scans from the INTERGROWTH-21st dataset. The age range is earlier in pregnancy than that of the MRI atlas (at $21-37$ GW), and roughly corresponds to the second trimester ($14-26$ GW). This is because of differences between the capabilities of the two imaging modalities: in MRI it is difficult to generate a 3D volume from 2D slices at lower gestational ages due to motion artifacts [44], and MRI scans are rarely acquired at lower ages in the clinic, due to tissue resolution and motion artifacts [39]. On the other hand, 3D US scans are acquired at higher speeds and have fewer problems with motion artifacts, and are routinely acquired at lower gestational ages [33]. On the upper end of the gestational age range, skull ossification in the third trimester can cause shadow artifacts and reduce the signal-to-background ratio of a US scan.

Figure 4.2 shows the difference in visual appearance of the two atlases at the same gestational age. The different modalities in which they are collected have different methods of generating contrast, which emphasise distinct anatomical features. In the MRI atlas, the value of each voxel is determined by its T2 relaxation time, which is tissue-dependent: each voxel with the same type of tissue should appear identical. In US, contrast is given by ultrasound reflectors, often at boundaries between tissue types, and by speckle, which is determined by

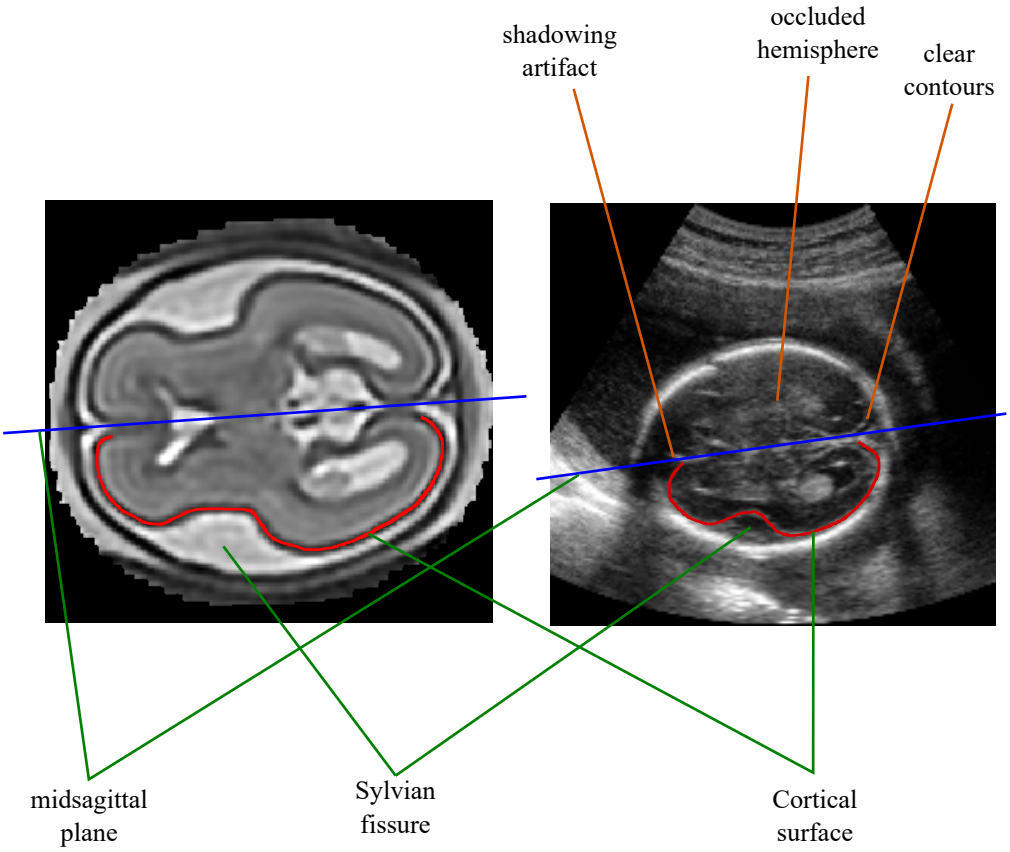


Figure 4.3: Comparison of a T1-weighted MRI fetal brain slice at 23GW and a fetal brain ultrasound image (axial), at 21GW. Notable anatomical features are pointed out in both slices as are some image characteristics in US.

the density of sub-resolution reflectors in each tissue type. Nonetheless, anatomical boundaries of the many structures (such as the cortical plate) are clearly visible in both atlases.

A selection of volumes was registered to a common reference space at each gestational week. Since only one hemisphere can be consistently visualised in fetal ultrasound [11], these were separated into left hemisphere-visible and right hemisphere-visible volumes. Within each age window and for each hemisphere, an initial 3D image registration was done using a similarity transform, to estimate a linear transformation for all volumes to a standard space. Then a groupwise diffeomorphic nonrigid registration [142] was used to obtain a visually sharp, consensus volume for each hemisphere. Finally, the templates for each hemisphere were combined into a single, complete template. Chapter 3 provides more detail on how the template was generated.

Figure 4.4 shows the gestational age (in weeks) of volumes used to generate the template,

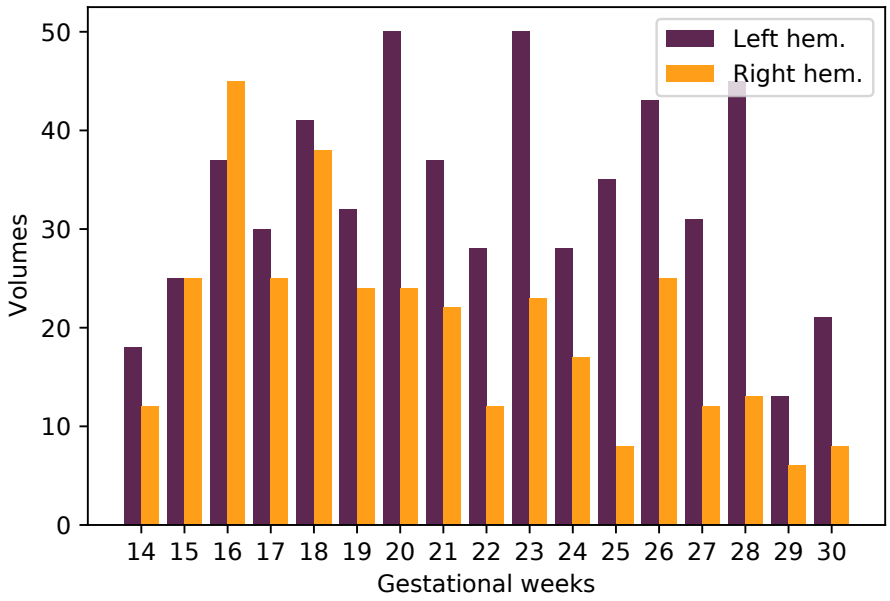
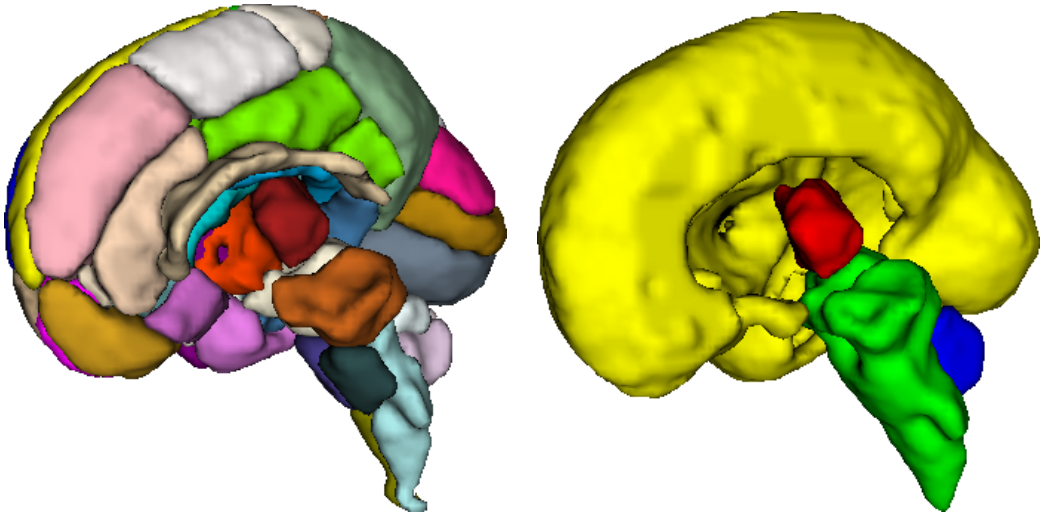


Figure 4.4: Distribution of the US volumes used to form the Namburete atlas, by gestational age and hemisphere. There are significantly more volumes in earlier weeks, due to ossification and progressively lower image quality later in pregnancy.

for each hemisphere. More volumes were used to form the atlas for gestational weeks near the middle of the age range, with fewer nearer the younger and older ends of the distribution. This is consistent with the discussion above: at lower gestational ages, the smaller dataset size makes it difficult to resolve anatomy; at higher gestational ages, ossification reduces image quality. The greatest number of volumes used to construct this atlas was in the range $20 - 25$ GW, around the middle of the age range.

The Gholipour atlas also has structural labels for every gestational age. It has a total of 127 labels [2], showing different structures in each hemisphere as well as different types of tissue. There is some overlap between some of these labels, as they are classified into "tissue" labels and "regional" labels. Figure 4.5a shows a 3D visualisation of some of those structures in the "regional" atlas, within a single hemisphere. Some, but not all, of these structures are visible on a US scan, since the two modalities have different contrast. Figure 4.2 shows a comparison between the Gholipour template and the Namburete template at 22 weeks, where the difference in visibility of different structures can clearly be seen. Some examples of these features are shown in Figure 4.3.

The US template, on the other hand, does not have any structural labels. One way to



(a) (b)

Figure 4.5: A 3D representation of the labels in one hemisphere in the MRI atlas at 23 weeks. (a) the original labels (in the "regional" atlas), and (b) the four anatomical labels chosen. (Red: thalamus, green: brainstem, blue: cerebellum, yellow: white matter)

generate structural labels for US is to find a correspondence between the MRI atlas and the US atlas, and propagate the labels from there. However, since not all structures in the MRI atlas can be seen in US, only some of the structural labels can be selected.

For the purposes of this chapter, the four structural labels selected for propagation to the US atlas were the thalamus, brainstem, cerebellum, and white matter, which are visible on US images. Table 4.1 shows the correspondence between the labels in the Gholipour atlas and the four structural labels selected for use in the US atlas. The labels for "thalamus", "brainstem", and "cerebellum" are all derived from the "regional" atlas, while the labels for "white matter" are derived from the "tissue" atlas: this results in a small amount of overlap between the labels. Table 4.2 shows the overlap of the "white matter" label with the other structural labels. Overlap only exists between the "white matter" and the "brainstem" label, accounting for 0.4% of the volume of the labels on average. In this work, this was resolved (to maintain a simple mapping between voxels and individual labels) by assigning all overlapping voxels to the "brainstem" structure. The small number of voxels affected (averaging 31 voxels per volume out of 4×10^6) make this a small adjustment.

Index	Label (MRI)	Label (US)
77	Thalamus_L	Thalamus
78	Thalamus_R	
92	Lateral_Vent_L	Brainstem
93	Lateral_Vent_R	
94	Midbrain_L	
95	Midbrain_R	
96	Pons_L	
97	Pons_R	
98	Medulla_L	
99	Medulla_R	
100	Cerebellum_L	Cereb.
101	Cerebellum_R	
102	Vermis_Ant_L	
103	Vermis_Ant_R	Cereb.
104	Vermis_Post_L	
105	Vermis_Post_R	
106	Vermis_Cent_L	
107	Vermis_Cent_R	
112	Cortical_Plate_L	WM
113	Cortical_Plate_R	
114	Subplate_L	
115	Subplate_R	
116	Inter_Zone_L	
117	Inter_Zone_R	
118	Vent_Zone_L	
119	Vent_Zone_R	

Table 4.1: The correspondence between the labels used in the MRI atlas and the structural labels propagated to the US atlas. As the MRI atlas had separate labels for each hemisphere, it was straightforward to export only one hemisphere to each part of the US atlas.

Structure overlap	Thalamus	Brainstem	Cerebellum
Overlap (% of structure)	0	0.4 ± 0.2	0

Table 4.2: Overlap of the "white matter" label with other structural labels.

4.3.2 Atlas alignment

The initial approach I took, published in MIUA 2019 [21], aligned the MRI atlas to each US volume by aligning their skulls. A simple skull mask was automatically fitted to each US volume and aligned with the anatomical boundaries of the MRI atlas. A similarity transform was used to find the best alignment using the open-source open3d library [149], as shown in Figure 4.6. A similarity transform was chosen because other transforms gave subjectively worse outputs: a rigid transform does not include a scaling factor (which is important since the fetal brain varies considerably in scale with development), and the additional degrees of freedom given by affine transforms were unnecessary, leading to slower convergence and

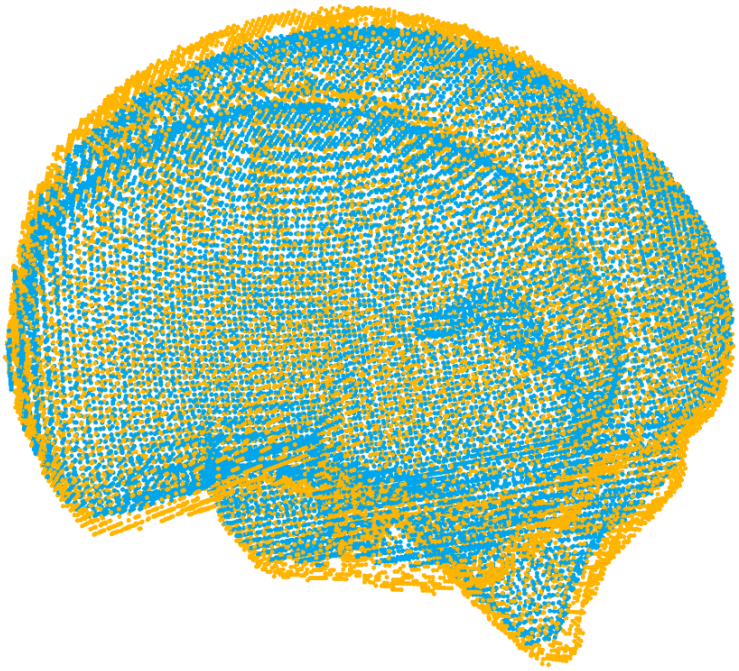


Figure 4.6: The final output of the skull-based alignment method: blue is the MRI atlas and yellow is the US template.

distorted outputs. The transformation matrix produced was then used to propagate the selected structural labels to each volume.

However, the MRI atlas (as can be seen in Figure 4.2) does not include the skull itself, which led to significant misalignment, especially for peripheral structures such as the medulla, in the brainstem. This alignment also does not consider internal anatomy and individual variation: a similarity transform has few degrees of freedom[6]. As a result, all volumes within the same gestational week were initially given very similar segmentations, without accounting for individual differences. This led to inaccurate labels for individual volumes, and in this thesis I propose a different alignment scheme.

A transformation can still be estimated between the US atlas and the MRI atlas by labelling corresponding known points on both volumes and estimating a transformation between them[7]. An affine transform, which requires a minimum of 12 correspondences[8], was

[6]7 degrees of freedom, corresponding to translation (along each axis), rotation (along each axis), and a global scale factor.

[7]In principle, this does not need to be between atlases, but it would be time-prohibitive to manually label corresponding points in all the volumes used.

[8]Corresponding to 12 degrees of freedom: 3 each for translation, rotation, scaling, and shear (along each axis).

Keypoints (US/MRI)	Numbering	Bilateral
Cerebellar lobe	0-1	Yes
Midbrain (substantia nigra)	2-3	Yes
Crown	4-5	Yes
Sylvian fissure (crown end)	6-7	Yes
Sylvian fissure (rump end)	8-9	Yes
CSP (anterior end)	10	No
CSP (posterior end)	11	No
Occiput	12-13	Yes

Table 4.3: The keypoints used to drive affine registration. The correspondences marked "bilateral" were made in both hemispheres.

estimated using a least-squares approximation to find the best correspondence between the two sets of points. Table 4.3 shows the keypoints that were labelled in both the MRI atlas and the US atlas, and Figure 4.7 shows an example slice of the atlases with labelled keypoints.

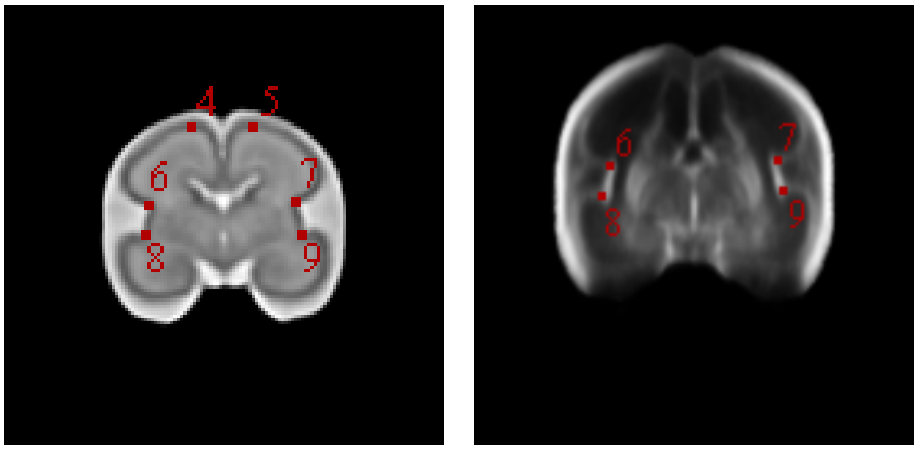
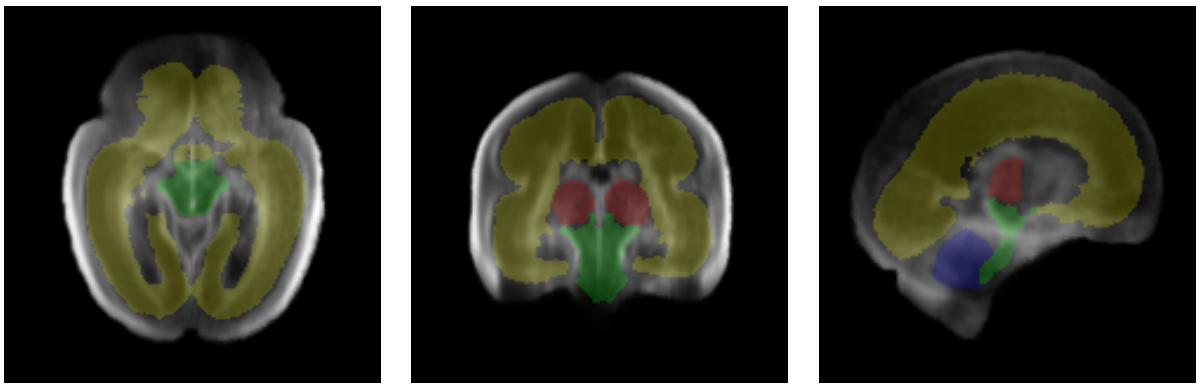


Figure 4.7: Keypoints labelled in a coronal slice in both atlases, at 24 weeks. The numbers correspond to structures as detailed in Table 4.3.

This was only done within the range of $20-25$ gestational weeks. This is the age range in which there are the most US samples available (see Figure 4.4) and in which the MRI atlas has structural labels. The Gholipour atlas' lower end is 21 weeks - to extend it to 20 weeks, I simply used the 21-week label[9].

[9]This reduces the anatomical accuracy of the segmentation labels at 20 weeks, but subjectively the difference is minor enough to be acceptable.



(a) (b) (c)

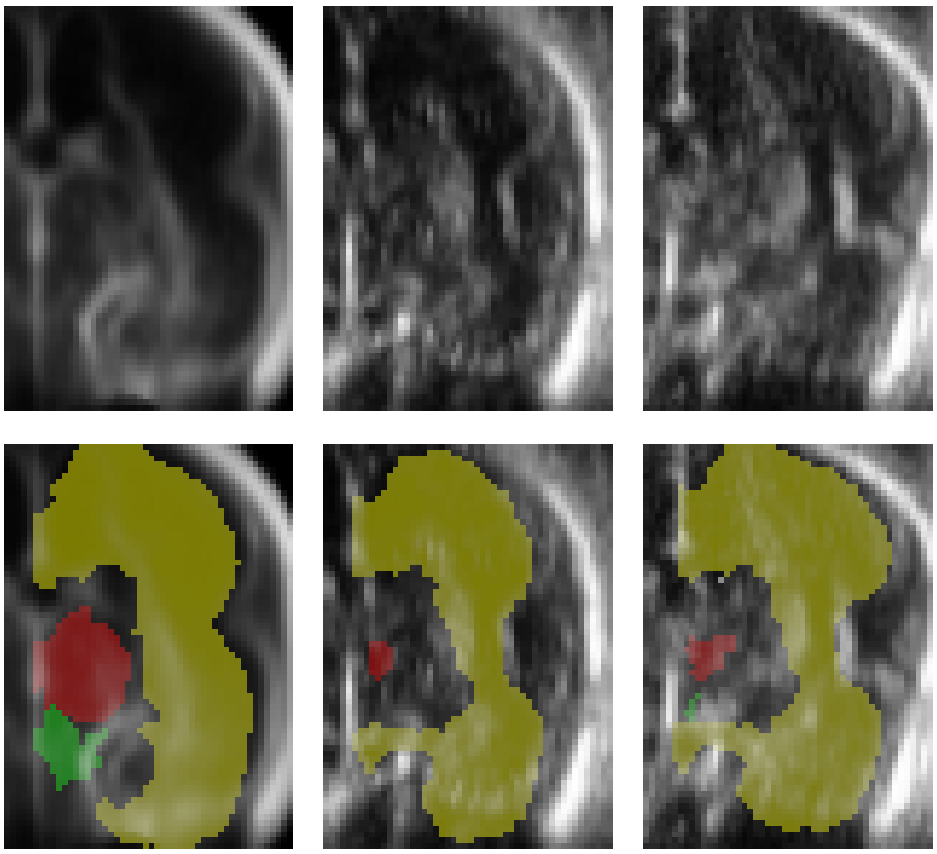
Figure 4.8: The structures propagated from the MRI atlas to the US atlas, at 24 weeks.

Figure 4.8 shows the alignment of the structures to the US atlas at 24 weeks. Subjectively, the alignment seems to be of good quality, with structures closely following visible anatomical boundaries. However, there are still some notable misalignments, with some clearly visible anatomical boundaries being slightly offset from the propagated label. In particular, the edges of the segmentation appear jagged, likely due to interpolation artifacts: this is discussed in more detail in Section 4.3.3.2. A quantitative analysis of the interpolation is presented in Section 4.3.4.

4.3.3 Label propagation

Once the atlas labels are generated, they still need to be propagated to individual volumes. The US atlas was generated using nonrigid diffeomorphic registration, as described in Chapter 3.3. This means that the transformations can be reversed to return to the original reference space. The 3D displacement map from the original volumes to the atlas space was saved when generating the atlas (see Figure **??** in Chapter 2). The reverse transformation (simply using the negative displacement map) can then be used to propagate segmentations from the atlas back to the original volumes.

Figure 4.9 shows the detail of a structure and its label in the US atlas, and how it changes when it is propagated to individual US volumes. The label is deformed to closely follow the anatomical boundary in the individual volumes, as expected.



(a) (b) (c)

Figure 4.9: A detail of (a) the US atlas, and (b - c) two US volumes. The lower row shows how the labelled structures change in response to the displacement map.

4.3.3.1 Separate hemispheres

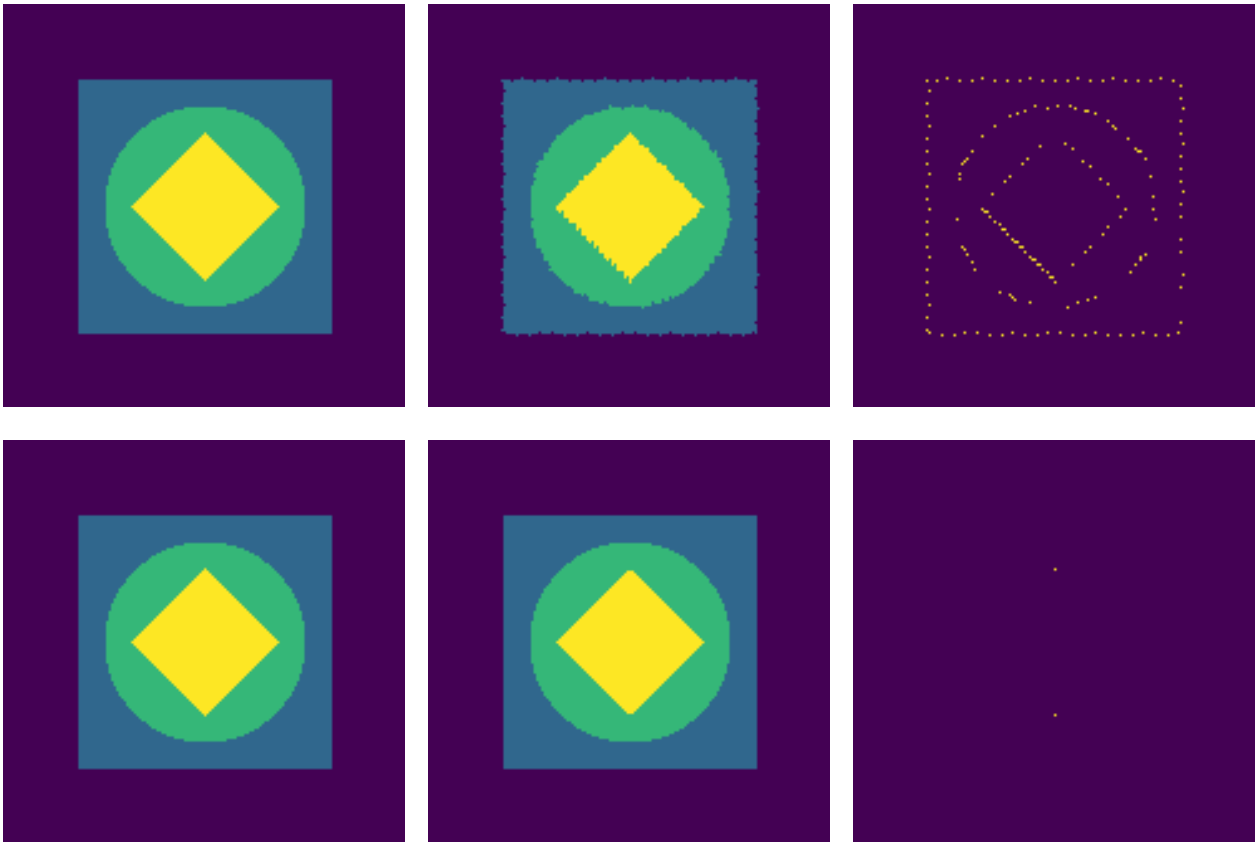
Since only the hemisphere distal to the ultrasound probe can be seen in any detail, each volume was only used to construct part of the atlas, corresponding to the distal hemisphere (see Figure 4.4). Consequently, for each volume only one hemisphere was registered to the atlas[10], and so only structures from that hemisphere could be propagated from the atlas back to the volume. Figure **??** shows, in each sub-figure, the cut-off point: only one hemisphere was kept for the atlas, together with a margin of 20 voxels across the mid-sagittal plane (MSP) (empirically chosen as a point beyond which anatomy is indistinct).

As a result, only structures from one hemisphere of the atlas are propagated to individual volumes. The cerebellum and the brainstem straddle the MSP, with no noticeable difference

[10]To be more precise, a $160 \times 100 \times 160$ section of the $160 \times 160 \times 160$ volume. The extra margin of 20vx across the mid-sagittal plane allows displacement map from that region.

in quality between hemispheres. However, the cerebellum segmentation always crosses into the non-propagated part of the hemisphere, leading to some imperfections in the segmentation of individual volumes. This only affects a small part of the segmentation of the cerebellum in each volume in the distal hemisphere.

4.3.3.2 Interpolation



(a) (b) (c)

Figure 4.10: A synthetic image (a) has a random transform and its inverse applied to it (b). The top row shows the result using nearest-neighbour interpolation, and the bottom row linear interpolation. (c) shows the misclassified pixels.

To generate the ground-truth structural labels for every volume, I used two transformations from the initial MRI labels: an affine transform to the US atlas, followed by the deformation field from the atlas to the individual volumes. Each of these steps requires interpolation.

Since atlas labels are discrete, a naive application of common interpolation methods (such as trilinear interpolation or spline interpolation) would produce a continuous output,

Interpolation	error %	Time (s)
Nearest-neighbour	0.241	1.83
Bilinear	0.023	2.26
Spline (3rd order)	0.040	5.17

Table 4.4: Accuracy of different thresholded interpolation techniques, on the synthetic multi-class image shown in Figure 4.10. These values were estimated from random transformation of the image 1000 times.

leading to some voxels at the boundaries having ambiguous classification. Nearest-neighbour interpolation does not have this feature and outputs a segmentation with discrete outcomes just like the inputs. However, this can lead to outputs with jagged or noisy edges, in contrast with the smooth contours of the original shapes. Figure 4.10 shows an example of this in 2D (a synthetic image was used as a basis for these comparisons and quantitative analysis in this section), and the same interpolation methods should be expected to yield similar results in 3D.

The alternative is to use a continuous interpolation method, such as linear interpolation, and then threshold the result to obtain a discrete map. This needs to be done carefully: the binary segmentation of each class needs to be transformed and interpolated separately, to avoid misclassifications during thresholding. The output in this case can be seen in Figure 4.10. The result is notably smoother and has less of the quantisation noise that characterises the nearest-neighbour approach. The need to transform and interpolate each structure separately, however, can lead to some inconsistency at anatomical boundaries, where (by quantisation noise) a single voxel can be labelled as belonging to multiple structures. To resolve this inconsistency and to minimise jaggedness and uncertainty, the interpolation methods compared here chose the class with higher value, effectively using an argmax function to select a single classification.

Table 4.4 shows the speed and error rate of different interpolation methods. Nearest-neighbour interpolation is the fastest method, as expected, but leads to jagged edges and a higher objective error rate. Other interpolation methods show far fewer misclassifications, at the expense of increased computation time. The results here show that linear interpolation

(thresholded after the final transform to generate a hard map) is almost as rapid as nearest-neighbour interpolation and has the fewest misclassified voxels of the methods investigated. Therefore, I chose thresholded linear interpolation to generate ground-truth labels for the rest of this chapter.

4.3.4 Segmentation quality

DSC	21/22 weeks	22/23 weeks	23/24 weeks	24/25 weeks	Average
Thala.	0.818 ± 0.020	0.473 ± 0.027	0.608 ± 0.017	0.327 ± 0.042	0.595 ± 0.163
BStem	0.659 ± 0.008	0.768 ± 0.007	0.726 ± 0.014	0.588 ± 0.008	0.700 ± 0.059
Cereb.	0.468 ± 0.020	0.784 ± 0.013	0.806 ± 0.014	0.770 ± 0.010	0.705 ± 0.145
WM	0.831 ± 0.011	0.804 ± 0.009	0.847 ± 0.006	0.739 ± 0.011	0.818 ± 0.035

Table 4.5: Dice coefficients for labels generated from adjacent weeks of the atlas, for volumes near the borderline between different weeks. Thala. is the thalamus, BStem is the brainstem, Cereb. is the cerebellum, and WM is the white matter.

It is difficult to measure how reliable the segmentation labels generated by the atlas are, but one way to estimate it is using volumes that lie between two temporal timepoints. Each volume in the dataset is labelled with its gestational age (in days), while the atlas is divided into different gestational ages by week. Therefore, each week's atlas corresponds to a range of 7 days. The volumes nearest the borderline between two weeks can be thought of as being approximately in the middle, in terms of appearance of structures, between the atlas for each week[11]. Therefore, the reliability of segmentation labels generated using the atlas can be estimated by comparing segmentations of these edge cases created using different atlases. Table 4.5 shows the Dice coefficients obtained for each structure based on the volumes near the temporal borderlines.

The Dice coefficients shown in Table 4.5 can be considered a reasonable upper bound for the performance of neural networks trained on this data, as they show the difference between different plausible labels generated by the same method. There are three possible sources of

[11]For the purposes of this chapter, only volumes within 1 day of the cutoff between weeks are considered 'borderline'.

error each contributing to the limited Dice coefficients: imperfect correspondence between the MRI and the US atlas used to generate labels; interpolation errors during atlas propagation; and the limited temporal resolution of the atlases used (1 week each), which can quantise structural features.

This method propagates segmentations from one fetal brain atlas (in MRI) to another (in ultrasound). These are constructed in different modalities using different volumes, so the boundaries of structures visible in the atlases are different. This creates some inaccuracies in the segmentation of the US atlas, which is then propagated to the individual volumes used to construct the atlas. This is consistent among volumes in the same week and (in part) across different ages: some differences in segmentation are due to individual variation among the acquisitions in the atlas, but others are due to the consistent difference in appearance.

Furthermore, the correspondence between atlases relies on an affine transform estimated by manually labelling corresponding anatomical features. The correspondences are noisy and can be inaccurate, and the transform is estimated by least-squares, which attempts to smooth some of the noise in the correspondences. That is another source of systematic error within each week of the atlas. Since volumes within each gestational week uss a different set of correspondences, the magnitude of this error varies across different weeks, and is not consistent.

The structural labels are affinely transformed from the original MRI atlas to the US atlas, and then transformed again (with a nonrigid transform) from the US atlas to individual volumes. Some quantisation noise is introduced by the interpolation step. To reduce this, trilinear interpolation is used in both of these steps, and the labels are only quantised at the end. Nevertheless, this adds some unavoidable noise to the labels.

Finally, the temporal resolution of the atlases is limited. While volumes are labelled with their age by day, the atlases used in this chapter only have one reference volume per week of gestational age. Brain structures change continuously throughout pregnancy, and any atlas that uses discrete temporal bins necessarily introduces some inaccuracies in segmentation, especially for volumes near the edges of their age bins.

4.4 Segmentation

The generated atlas-based labels can only be used for volumes that have a known set of correspondences with the atlas: therefore, they can only be generated with the volumes used to construct the atlas in the first place. While this is sufficient to generate a training set, it cannot generalise to unseen data, and cannot scale to large datasets. For this purpose, it is necessary to use a different method. While there are several possible approaches to this (see Chapter 2), the availability of a labelled set makes supervised machine learning attractive. CNNs are an appropriate tool for multi-task segmentation in these circumstances.

In this section, I investigate the appropriate network architecture for multi-task segmentation in fetal brain US volumes. Encoder-decoder architectures such as the U-Net [94] have proven successful for medical image segmentation tasks [124, 150, 99], even with limited training data. The original implementation of the U-Net is for a binary classification task, where every pixel in the image is labelled as belonging to the structure of interest or background. Furthermore, original implementations of the U-Net were for 2D segmentation, though 3D variants have been proposed [99, 151]. This chapter explores different approaches to extend the U-Net to 3D multi-task segmentation.

The simplest way to extend the U-Net architecture to segment multiple structures is to train an ensemble of single-task U-Nets in parallel for different structures. This is very straightforward but requires significant memory and training time to train the required networks. It also requires some postprocessing to fuse the output of each network in the ensemble, as independent networks may label the same voxel as belonging to different structures.

Alternatively, a multi-task architecture can be devised by simply replacing the final layer: instead of predicting a binary score for each voxel, the network can use a softmax function to predict a one-hot feature vector for each voxel. However, this requires more memory for each training run and may require different loss functions.

2D networks can also be used to generate 3D segmentations. A 3D volume can be sliced into 2D slices, and each slice can then be individually segmented, but this leads to a loss of

context: 2D convolutional layers incorporate context along the X and Y axes, but not the Z axis of the stack. One possible way to alleviate this is the use of three different networks to segment different slices made across the three possible axes. The predictions from each of these networks can then be aggregated, similar to the strategy used by Guha Roy et al in QuickNAT [146] . This has the advantage of the lower number of parameters and memory usage of 2D networks, while maintaining context in all directions as in a 3D network.

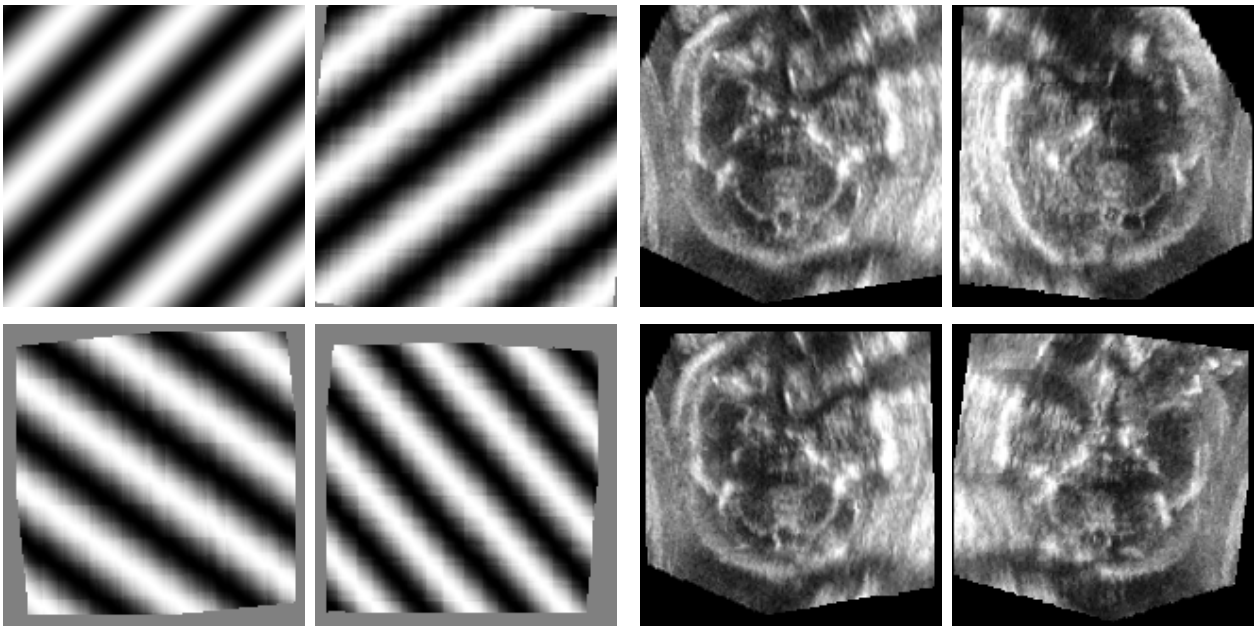
4.4.1 Data preprocessing

All volumes used to construct the US atlas were used for this task. The ultrasound acquisition protocol was agnostic to the direction faced by the fetus during imaging: the dataset is therefore a mixture of volumes acquired from different orientations, some of which have a more visible left hemisphere and others of which have a more visible right hemisphere. The distribution of age and hemisphere of these volumes is shown in Figure 4.4. There are a total of 402 US volumes available, of which 271 are of the left hemisphere and 131 are of the right hemisphere. This is a large imbalance, with more than twice as many volumes from one hemisphere than the other, and cannot be explained by chance alone ($p < 10^{-6}$). This is likely to arise naturally: approximately two-thirds of fetuses lie to their left at pregnancy term [136]. All volumes were in the range of 20-25 gestational weeks.

As discussed above, different volumes were used to construct the atlas for different hemispheres, and as a result only one hemisphere's structural segmentations were propagated from the atlas to individual volumes.

To align all volumes to the same reference space, all volumes whose right hemisphere was segmented were reflected across the MSP, so that all structures to be segmented appear to be in approximately the same location. Each brain was centred and resampled to a $160 \times 160 \times 160$ volume, with the mean voxel sampled at $0.6 \times 0.6 \times 0.6$ mm. Each volume has four structural labels (thalamus, brainstem, cerebellum, white matter): these were transformed into categorical labels, so each voxel is associated with a one-hot vector of length 5 (the four structures of interest and background) corresponding to the ground-truth label.

4.4.2 Augmentation



(a) (b)

Figure 4.11: Three augmentations applied to (a) a synthetic example, and (b) a real ultrasound volume in the dataset (of which one axial slice is shown here). These augmentations were randomly sampled with parameters within the range in Table 4.6. The original in both cases on the top-left.

Augmentation methods
Reflection across MSP
Random translation (±5vx)
Random rotation (±10°)
Random scaling (±10%)

Table 4.6: The transformations used for data augmentation.

Data augmentation was used during training. The augmentation used consists of simple similarity transforms, shown in Table 4.6. Since the images are aligned to a common reference space, the only axis they can be reflected across to maintain alignment is the mid-sagittal plane (MSP), the plane that runs between the brain's hemispheres and between the eyes. As the data alignment is not exact, small random translations (up to ±5 voxels along each axis), small random rotations (up to ±10° around each axis), and small amounts of scaling (up to ±10% linear zoom) were also used. Nearest-neighbour interpolation was used for

computational efficiency[12]. The volumes were reflected with 50% probability, while the degree of all other augmentations was sampled from a uniform distribution. Some examples of slices from the same volume augmented several times are shown in Figure 4.11.

To minimise the use of interpolation, a single transformation matrix was made for augmentation combining all random rotations and scaling, and then applied to the volume. Since this was repeated at every training iteration, computational efficiency was important. Nearest-neighbour interpolation was used, to maximise efficiency.

4.4.3 Network design

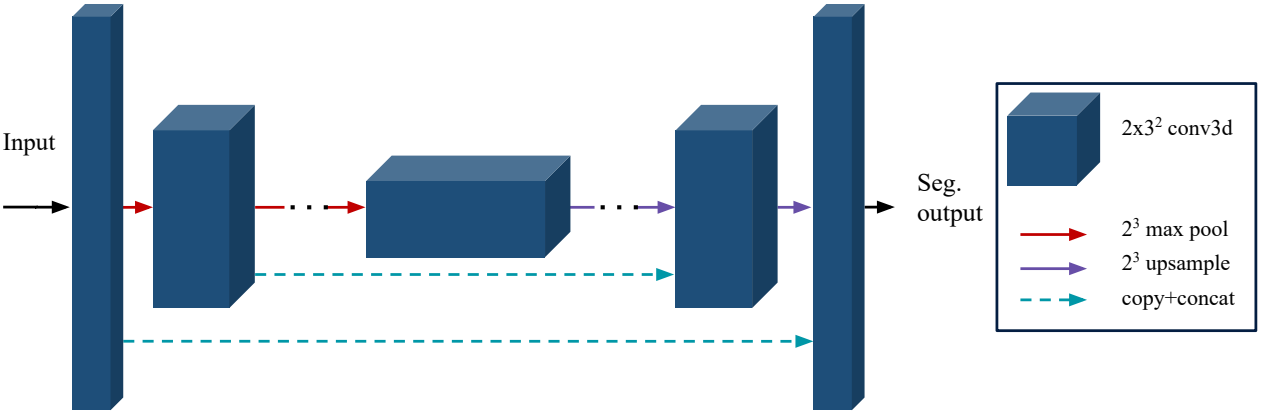


Figure 4.12: The multi-task 3D U-Net architecture used in this chapter. The single-task and 2D architectures are identical, except in the number of dimensions in each convolutional layer and the activation function used in the final layer.

This section compares different network architectures.

Network	No. parameters	Memory use (GB)	Batch size
3D multi-task	8643685	3.36	1
4x 3D single task	4x 8643617	3.11	1
2D multi-task	2850821	2.10	64

Table 4.7: A comparison of the sizes (in terms of number of trainable parameters, and memory usage) of the CNNs implemented. The 4x 3D single task networks and the 2D multi-task network require multiple independently trained networks, which multiplies their memory usage.

[12]During training, the computational efficiency advantage of nearest-neighbour interpolation outweighs the disadvantage of aliasing near the boundaries.

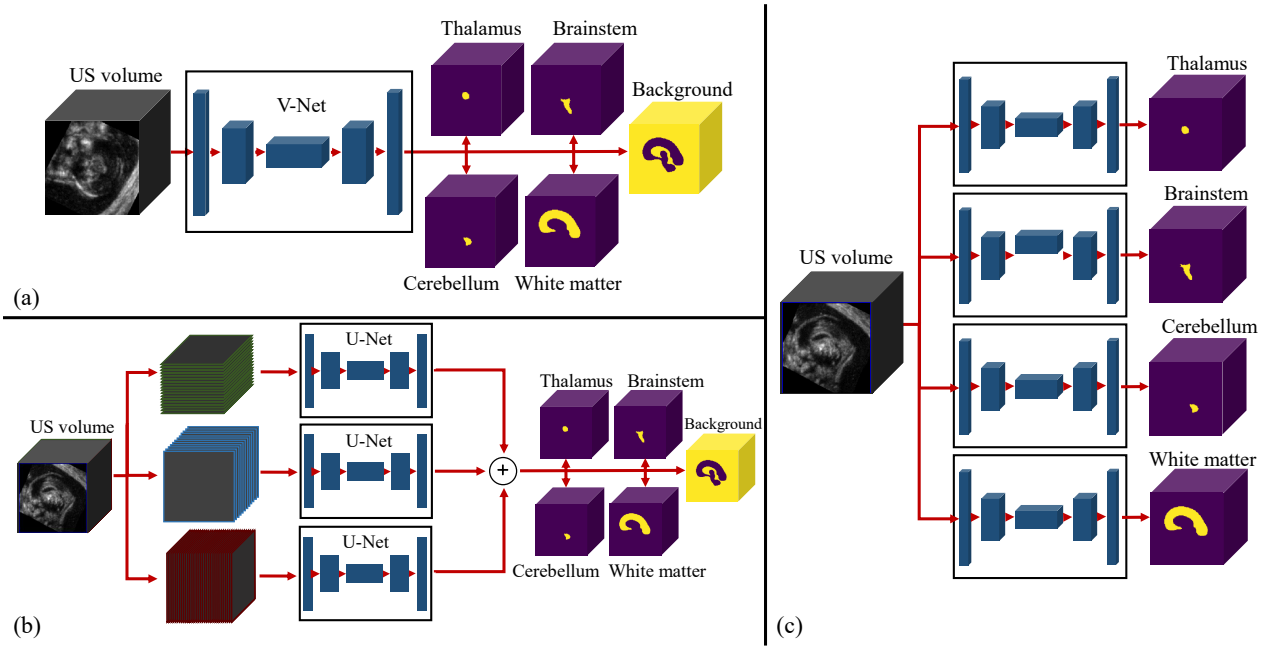


Figure 4.13: The proposed different network pipelines. (a) The proposed 3D multi-task architecture, based on V-net. (b) A 2D multi-task framework, based on U-net, with QuickNAT-style merging of the different views. (c) A 3D single-task architecture, where a different network is trained per structure.

A 3D encoder-decoder network architecture based on the U-net architecture was used to perform multi-task segmentation. This is a large network for 3D volume processing with 3D convolutional kernels, so memory constraints are a real limitation. To remain within the GPU memory available[13], the top-level layer learned 16 3×3×3 feature maps. To satisfy memory constraints, the V-net architecture [151] was used. A softmax activation function was used at the output of the final convolutional layer to classify each voxel. The output was a five-class segmentation $\mathbb{Y} \in \mathbb{R}^{n \times N_x \times N_y \times N_z \times 5}$where n is the number of volumes, and all volumes have dimensions $N_x \times N_y \times N_z$. Segmentation maps for the thalamus, white matter, brainstem, cerebellum, and background were generated[14]. A multi-label Dice coefficient, the sum of the Dice coefficients of all classes, was used as the loss function, as this led to what visually appeared to be the best results. Multi-label Dice is given by

$$DSC_{ml} = \sum_i \frac{2\left(GT_i \cap Seg_i\right)}{GT_i + Seg_i}$$

[13]I used Nvidia GeForce GTX 1080 GPUs for training, with a memory of 10GB.
[14]Due to the size of this network, it used a batch size of 1 volume for training.

where GT and Seg are mappings of voxels corresponding to the ground-truth and generated segmentation, respectively, across all structures of interest i. The other hyperparameters for the network's training were replicated from Milletari et al's V-net study [151] : a depth of 5 layers to the bottleneck, and a learning rate $\alpha = 10^{-3}$. ReLU activation functions were used instead of PReLU for simplicity. Figure 4.12 shows a schematic of the architecture used.

A similar, single-task version of this network was also implemented for comparison. The architecture was identical, but the final layer was given a sigmoid activation function to constrain each output class to the range $[0, 1]$. Four different networks also had to be trained, each to segment a different structure of interest.

A classical 2D U-net architecture was also implemented for comparison. This network took 2D slices as input, and output a segmentation map for each slice. The segmentations of each slice were then stacked to obtain a full 3D semantic segmentation. To incorporate contextual information from other views, the data was sliced in 3 different ways, corresponding to the 3 orthogonal views, in a strategy similar to QuickNAT [146]. Each 2D network outputs "soft" segmentation masks for each structure, with each voxel given a value between 0 and 1 for each structure corresponding to the network's confidence. Combining the output of each network could exploit 3D information for segmentation, and therefore lead to a better accuracy than networks trained on individual views. Section 4.4.4 considers approaches by which labels may be fused to obtain an aggregate segmentation. All three architectures are shown in Figure 4.13.

There are significant differences in model size and memory requirements across different architectures, as shown in Table 4.7. The 3D single-task and multi-task architectures have similar numbers of parameters: the choice of a 3D architecture involves significant memory usage, which limits the batch size that can fit in the memory of one GPU. These networks were trained using a batch size of 1. The 2D network architecture involves much lower memory usage, which allows a higher batch size. The 2D network therefore was trained with a batch size of 64.

All of these network architectures were implemented using Keras 2.3 [152] and Tensorflow

1.14 [153] and trained on an Nvidia GeForce GTX 1080 GPU. Other parameters for the network's training, such as the use of $3 \times 3 \times 3$ convolutional layers, ReLU activation functions for each convolutional layer and the Adam optimisation algorithm, were replicated from Ronneberger et al's study [94].

4.4.4 Postprocessing

Each of the network architectures described above requires postprocessing steps.

A network that generates multi-task segmentation outputs a feature vector for every voxel in the volume, with the implied probability p_i of the voxel belonging to each of 5 structural classes (the 4 structures of interest + background). The use of a softmax activation function at the output layer ensures that the probabilities p_i across every class i in a voxel satisfy $\sum_i p_i = 1$. To turn this into a segmentation map, each vector needs to be reduced to a single label. A simple argmax function can be used to select the class i with the highest probability for each voxel and produce a hard segmentation map,

$$i(n) = \arg_i \max p_i(n). \tag{4.1}$$

Using separate networks to perform binary (structure of interest / background) segmentation for each structure of interest leads to other problems. Each network produces a single output $p_{s,i}$ ranging from 0 (background) to 1 (structure of interest i). Each segmentation can be thresholded to produce a hard segmentation map for individual structures, but there is no guarantee that the implied probability for each voxel sums to 1. In principle, it is possible for a voxel to be segmented as belonging to multiple structures simultaneously by different networks. The most principled way to solve this is to generate a combined prediction $p_{c,i}$, using the predicted probability from each network and the minimum probability that it was

classified as background across the networks,

$$p_{c,i}(n) = \begin{cases} p_{s,i}(n) & i \neq 0 \\ 1 - \max_i \left(p_{s,i}(n)\right) & i = 0 \end{cases}$$

After that, an argmax function can be used as shown in Equation 4.1.

The 2.5D network, shown in Figure 4.13, requires the most postprocessing to obtain a single segmentation. It consists of three 2D networks, each trained on slices of the volumes sliced along different axes (the axial, coronal, and sagittal views). Using a 2D network architecture requires considerably less GPU memory and allows faster training, at the expense of contextual information from the missing axis. Combining predictions generated by networks trained on different views is one way of incorporating that information.

Since there are three networks generating predictions for each voxel, each voxel n has three independent sets of classification probabilities $p_{i,x}(n), p_{i,y}(n), p_{i,z}(n)$ over all classes i. For most voxels, the predictions of individual 2D networks agree and lead to the same segmentation outcome, but in some cases different networks may make different predictions. I considered two possible approaches to combining the predictions from different slices: voting and summation.

Voting involves thresholding each of the three segmentations to produce a single prediction $i_{\mathrm{ax}}(n)$ for each 2D network and then comparing them directly to each other. Since there are three independent predictions per voxel, it is very likely that at least two views for each voxel will have the same class prediction. In those cases, the classification predicted by the majority of networks can be taken as the consensus classification $i_c(n)$,

$$\begin{aligned} i_{\mathrm{ax}}(n) &= \arg_i \max p_{i,\mathrm{ax}}(n) \\ i_c(n) &= \mathrm{mode}\left(i_x(n), i_y(n), i_z(n)\right). \end{aligned} \tag{4.2}$$

It is theoretically possible (though unlikely in practice) that a single voxel would be

classified as belonging to three different structures: in this case, a tie-breaker is needed. Ties can be broken by classifying the voxel as the class with the highest output probability from any of the networks,

$$i_c(n) = \arg_{i,\text{ax}} \max \left(p_{i,\text{ax}}(n) \right)$$

Summation is simpler: to sum the probability values for the feature vectors for each voxel, and then use an argmax function to obtain an overall segmentation combining outputs from the three individual networks.

$$i_c(n) = \arg_i \max \left(\sum_{\text{ax}} p_{i,\text{ax}}(n) \right) \tag{4.3}$$

This produces a different outcome from the voting system in cases where two networks predict a classification with low confidence and the third predicts another one with high confidence. Table 4.11 shows the performance of these two methods. While both ways of aggregating the three networks' predictions outperform any single component network, method 4.3 appears to lead to slightly greater performance. Therefore, I used **summation** for all the results presented below.

4.4.5 Data splits

	Gestational age (weeks)					
Hemisphere	20	21	22	23	24	25
Left	50	37	26	50	28	35
Right	24	22	12	23	17	8

Table 4.8: The number of labelled volumes, by age and hemisphere.

The volumes are split between 6 gestational weeks and 2 hemispheres (see Figure 4.4 and Table 4.8), and this split should be maintained in the validation data. Each of the validation folds in this chapter has equal proportions of volumes.

80% of the data (479 volumes) was used for training, with the remaining 20% (120 volumes)

used for validation. Five-fold cross-validation was performed for all the experiments described here: this means that every volume was in the training set in four of the validation folds, and in the validation set for one of them. All reported results in this chapter use this cross-validation scheme.

4.5 Results

Network	Converg. epochs	Converg. time (min)	Segment time / vol (s)
3D multi-task	20	67	0.34
4x 3D single-task	20	208	1.21
3x 2D multi-task	50	108	1.44

Table 4.9: The number of epochs taken to reach convergence, average time required to reach convergence, and the time to segment all structures on an individual volume

The multi-task network appeared to converge[15] within 20 epochs, after which no further improvement was noticed. This took about an hour on the hardware used in this chapter (see Section 4.4.3). The single-task network also converged after 20 epochs: however, since four networks were trained, this still led to a substantially longer training time required of over three hours. The 2D multi-task network needed 50 epochs to reach convergence, possibly due to the more limited amount of data in each iteration (a batch of 64 2D slices, instead of a 3D volume). Table 4.9 shows that the multi-task network is the fastest to train, largely because a single network needs to be trained. If trained in parallel on multiple GPUs, the reverse would be true: the 3D multi-task network would be the slowest to converge. Training time would remain unchanged at 67 min, compared to 52 min for individual 3D single-task networks and 36 min for 2D multi-task networks.

Table 4.10 shows the segmentation performance of each network across the structures of interest. Across all network architectures, the best performance was obtained on segmentation of the white matter, with a Dice coefficient of 0.831 for the 3D multi-task architecture (and

[15]Defined as the point where the minimum loss was reached on the validation set: beyond this point, overfitting led to decreasing performance on the validation set.

Network	DSC	ED (mm)	HD (mm)
Thalamus			
3D multi-task	**0.649 ± 0.067**	3.17 ± 1.35	**7.05 ± 2.17**
3D single-task	0.616 ± 0.043	2.82± 1.65	6.28 ± 1.94
2D	0.639 ± 0.036	**2.36 ± 1.27**	4.41 ± 1.91
Brainstem			
3D multi-task	**0.713 ± 0.045**	3.09 ± 1.26	**6.87 ± 2.95**
3D single-task	0.606 ± 0.076	4.96 ± 2.26	10.95 ± 4.31
2D	0.656 ± 0.074	**2.48 ± 1.85**	7.52 ± 2.86
Cerebellum			
3D multi-task	**0.711 ± 0.037**	2.42 ± 1.32	4.24 ± 2.06
3D single-task	0.547 ± 0.044	4.20 ± 2.72	9.25 ± 1.47
2D	0.642 ± 0.070	**2.25 ± 1.93**	**3.51 ± 1.80**
White matter			
3D multi-task	0.831 ± 0.028	3.27 ± 1.46	7.53 ± 2.28
3D single-task	0.867 ± 0.005	3.32 ± 1.72	6.90 ± 2.04
2D	**0.856 ± 0.002**	**2.90 ± 1.55**	**6.67 ± 0.33**

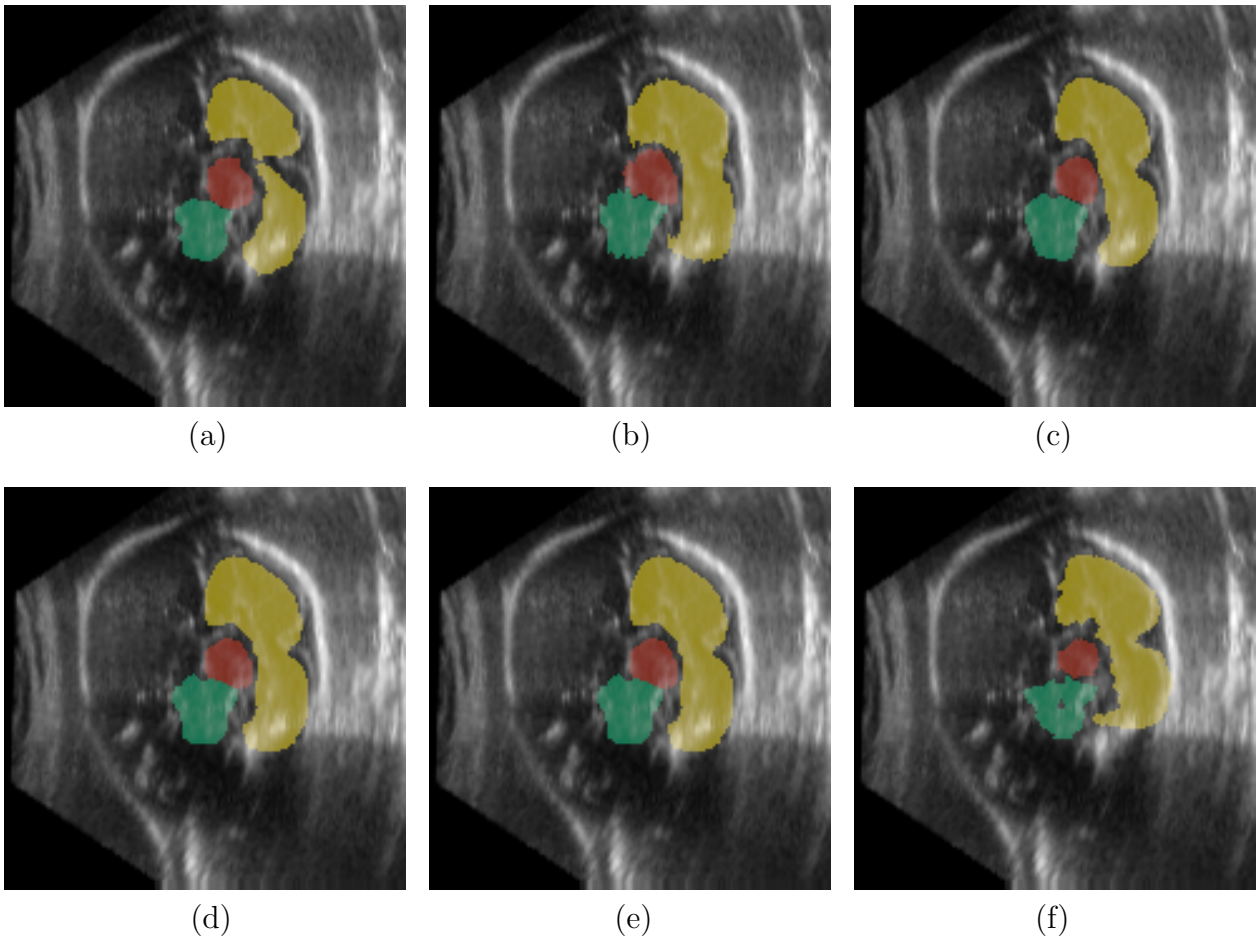
Table 4.10: Segmentation performance of single-task and multi-task segmentation architectures, as measured by Dice coefficient (DSC), Euclidean distance of the centres of mass (ED) and Hausdorff distance (HD). Across measures and brain structures, the multi-task architecture outperforms the single-task network.

higher for alternatives). The thalamus and the cerebellum, on the other hand, had the worst performance, with Dice coefficients of 0.649 and 0.711, respectively, in the 3D multi-task architecture.

The 3D multi-task architecture itself gives the best performance, across different structures and architectures. This performance boost appears larger in Hausdorff distance than in Dice coefficient, suggesting that this architecture is best at generating more anatomically plausible segmentations that do not significantly differ from the structures of interest. The notable exception is the segmentation of the white matter: the single-task network and the 2D networks both outperform the 3D multi-task network in this instance. Section 4.6.2 discusses

the reasons for the difference in performance across different structures and architectures.

4.5.1 Postprocessing



(a) (b) (c)

(d) (e) (f)

Figure 4.14: (a,b,c) are the segmentations (resliced to a coronal slice) of a volume at 24 weeks, generated by 2D networks trained on the axial, sagittal, and coronal views, respectively. The lower row shows the output of aggregating these segmentations by (d) voting, (e) summation, (f) the ground-truth segmentation.

Table 4.11 shows the performance of each individual 2D network in the first three rows, for one of the structures of interest. There is significant variability in prediction quality, with the network trained on the sagittal view performing considerably worse than the network trained on the axial and coronal views. This may reflect the labelling process: the protocol used to label structures in the MRI atlas is that described by Gousias et al [42], in which structures were labelled by reference to 2D axial and coronal views. The sagittal view may therefore be less consistently labelled.

Postprocessing	Dice coefficient	Hausdorff distance (mm)
Axial only	0.839 ± 0.004	10.53 ± 1.85
Sagittal only	0.767 ± 0.021	13.29 ± 2.49
Coronal only	0.847 ± 0.008	9.33 ± 2.49
Voting	0.854 ± 0.003	6.55 ± 0.27
Summation	0.856 ± 0.003	6.89 ± 0.47
Summation + morphology	0.858 ± 0.002	6.67 ± 0.33

Table 4.11: A comparison of the performance achieved by individual components of the 2.5D architecture, and different ways to combine their predictions. Results are given just for the "white matter" segmentation.

Section 4.4.4 considers two approaches to aggregate different 2D segmentations from the network, **voting** and **summation**. Table 4.11 shows the performance of these two methods. While both ways of aggregating the three networks' predictions outperform any single component network, method 4.3 appears to lead to slightly greater performance. Therefore, I used **summation** for all the results presented in this chapter.

One other postprocessing step which was found to improve performance for this network was the use of morphological operations. A 2.5D network operating on individual slices can produce significant discontinuities across slices, over- or under- segmenting individual slices. This can lower the worst-case performance of the network especially, as seen in the Hausdorff distance in Table 4.11. Aggregating its predictions with those of networks trained on different 2D views smooths some discontinuities (as seen in the Hausdorff distance measures) but some still remain. A simple approach to smooth discontinuities further is the use of morphological operations. Another technique which was found to slightly improve performance further was the application of a simple morphological operation (a 3×3×3 morphological closing[16] followed by an opening[17] with the same kernel size) to the resulting segmentation. This operation removes any small gaps from the segmentation and weakly enforces smoothness near the edges of the segmentation. Figure 4.14 shows the qualitative difference in output. The

[16]A binary dilation, followed by an erosion.
[17]A binary erosion, followed by a dilation.

results reported for the 2.5D network in this chapter are for the case where the predictions are averaged, and morphological operations are confirmed.

4.5.2 Analysis of failure cases

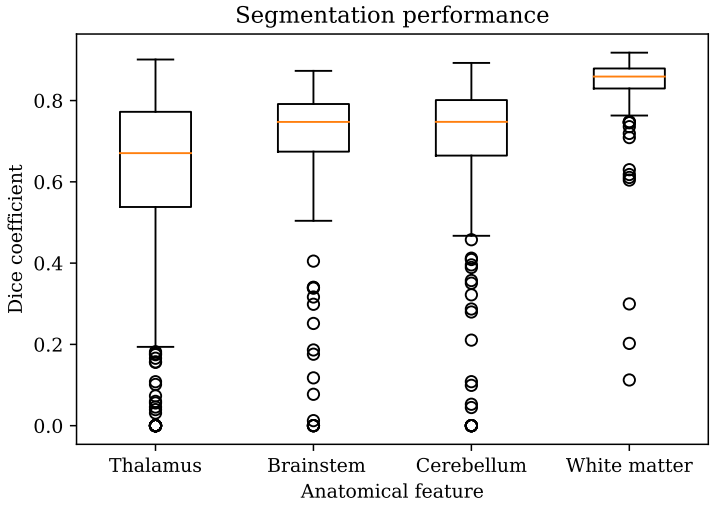


Figure 4.15: Box plot of Dice coefficients of the segmentations obtained by the 3D multi-task architecture.

Figure 4.15 shows a box-plot of the Dice coefficients in the segmentation of different structures, as obtained by the 3D multi-task network. Smaller features tend to have higher variability in Dice coefficient, perhaps as a result of their smaller physical size (see Section 4.6.2 for more discussion of this). On the other hand, segmentation of the white matter is very consistent, with the large majority of volumes having a Dice coefficient within a narrow band.

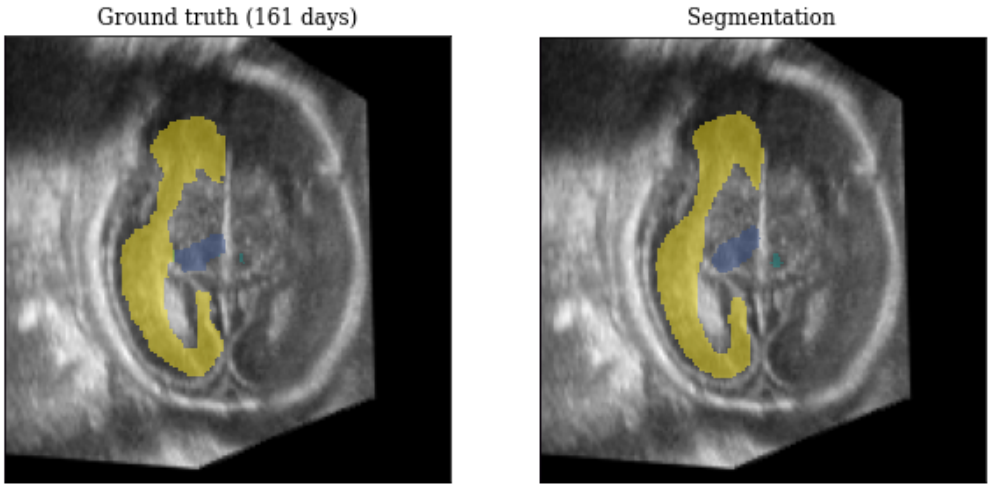


Figure 4.16: An axial slice of a successful segmentation (white matter Dice = 0.91). Yellow corresponds to "white matter", blue is "thalamus", green is "brainstem".

Figure 4.16 shows a slice of one successful segmentation, using the 3D multi-task network. The segmentation generated is very similar to the ground-truth segmentation (if a bit smoother). It is also notable, however, that the ground truth itself is imperfect: some of its boundaries are different from the true visible anatomical boundaries, which the CNN only partially corrects (such as by more faithfully following the borders of the MSP).

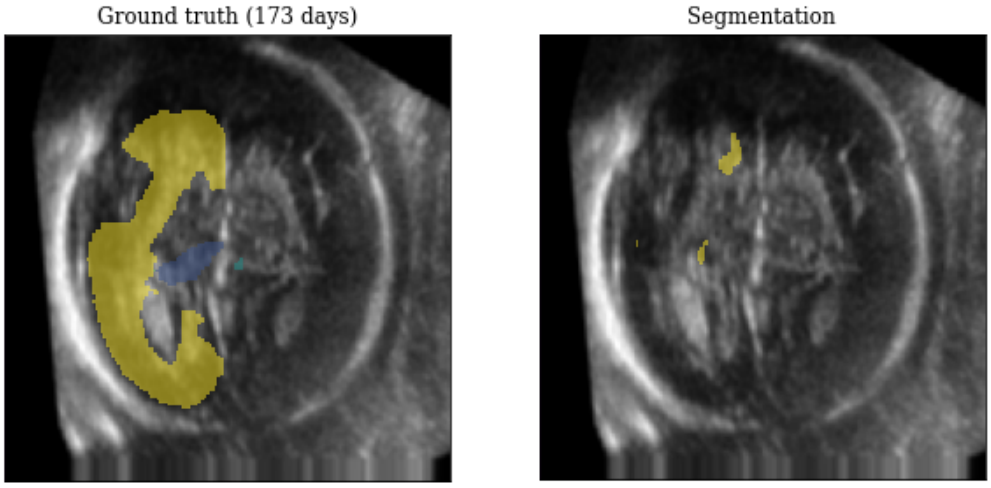


Figure 4.17: An axial slice of a failed segmentation (white matter Dice = 0.11).

There were only a few volumes (3 in this dataset) that resulted in complete failure. Figure 4.17 shows the worst such case across the dataset, which is (incorrectly) largely segmented as background across all structures. The two other failure cases resemble this one, with failures

caused by undersegmentation. This does appear to happen only at higher gestational ages. For volumes in this age range fewer atlas-based training volumes are available (see Figure 4.4), so the training dataset may cover only a limited part of that feature space.

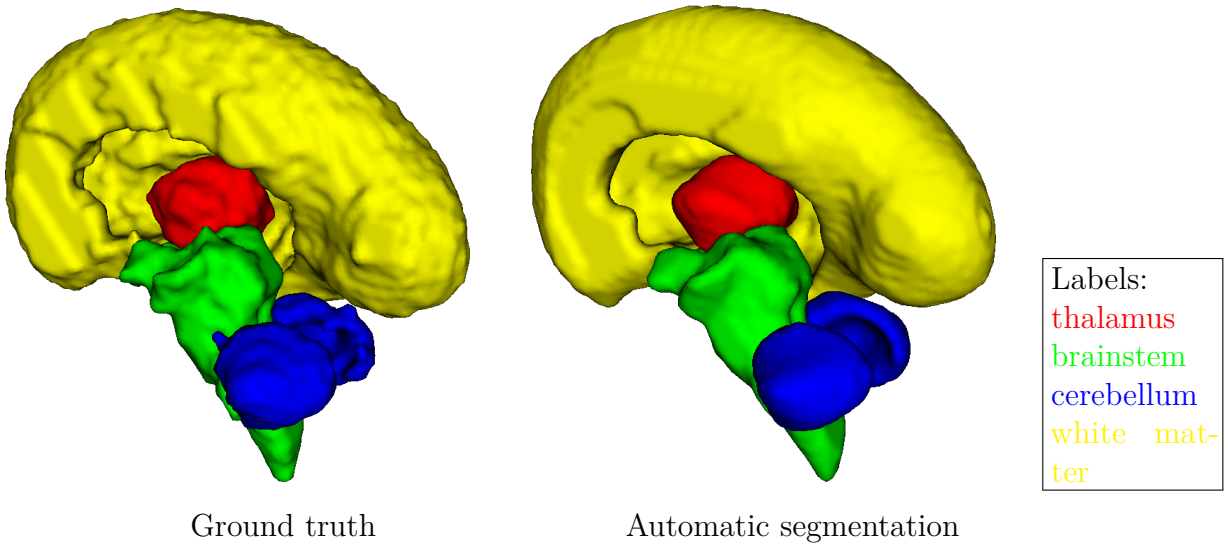


Figure 4.18: Comparison of the visual appearance in 3D of the atlas-based ground-truth labels and the automatic segmentation for a volume.

More generally, the CNN's output seems to be significantly smoother than the atlas-based ground truth labels used for training, as seen in Figure 4.18. This is likely due to the jaggedness of the original atlas-based segmentation: since nearest-neighbour interpolation is necessary[18], aliasing artifacts are likely to be introduced into the volumetric image. The resulting learned volumes, while smoother, do also appear to lose some of the detail available: surface features like sulci and gyri appear to be smoothed.

[18]Other forms of interpolation, such as linear interpolation, are impossible when using categorical data.

4.5.3 Comparison to clinical metrics

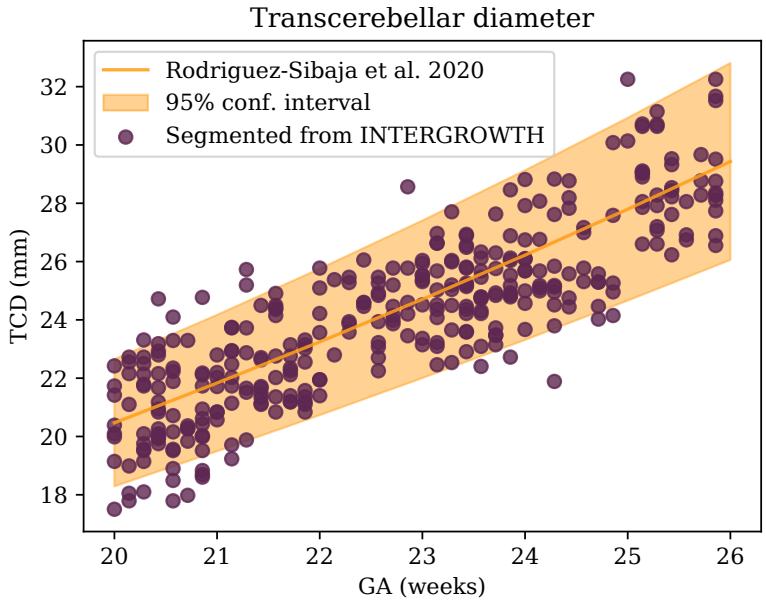


Figure 4.19: Estimates of the transcerebellar diameter (TCD) derived from my method are in general agreement with the literature [4].

Having segmented a volume, clinical measurements can be taken and compared to previous results in the literature. Volumetric segmentations, as have been performed in this chapter, are not routinely performed or well-characterised as a clinical metric. However, from the segmentations obtained in this chapter it is possible to derive other metrics that are clinically validated. One such metric is the transcerebellar diameter (TCD) [143], measured as the distance between the lateral boundaries of the cerebellar lobes. It is straightforward to obtain this from a full cerebellar segmentation: the greatest Euclidean distance between two voxels labelled as part of the cerebellum[19].

Figure 4.19 shows the transcerebellar diameter (TCD), a clinical biomarker often measured in scans [154, 4]. Rodriguez-Sibaja et al [4] propose an empirical equation to estimate the TCD from gestational age, drawing form observations on the INTERGROWTH-21st dataset. The results shown here accord well with this equation, as well as previous studies.

[19]In clinical practice, the TCD is measured as the furthest distance between two points in the cerebellum as seen in the transcerebellar plane: this yields very similar results.

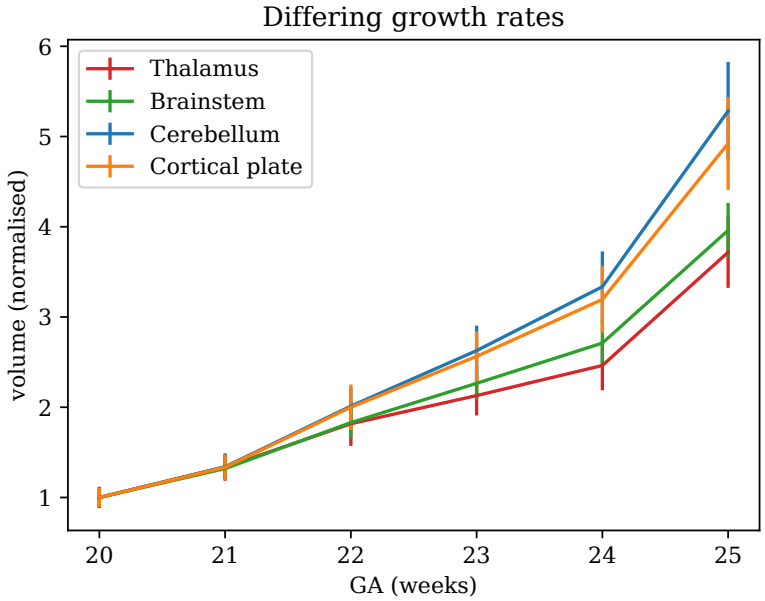


Figure 4.20: Mean and standard deviation of volume of different brain structures across the INTERGROWTH-21st dataset, as measured by the segmentations produced. To emphasize the growth rate, the average volume was normalised to be 1 at 20 weeks.

The raw volumetric segmentations themselves can provide useful insights. Figure 4.20 compares the volume of the four structures of interest across the dataset, from a (normalised) baseline at 20 weeks. In this period the cerebellum and the cortical plate grow notably more quickly than the thalamus and the brainstem, especially near the end of the second trimester. This is consistent with [155], which measured a sequence in the structures that exhibited fast growth in the second trimester: first the diencephalon (including the thalamus), then the white matter, and then other cerebral white matter. It is also consistent with the theory of neuronal migration [26], according to which structures in more central parts of the fetal brain (such as the thalamus) see greater development earlier in gestation, and more peripheral structures (such as the cortical plate) see faster growth in later stages. Chapter 6 goes into more detail on volumetric growth of specific structures.

4.6 Discussion

4.6.1 Label generation

The structural labels generated from the US atlas are much more closely aligned to individual anatomy than the labels generated from the MRI atlas alone registered to skulls of individual volumes. The labels generated using skull registration were registered using only a similarity transform, so other than the parameters of the transform (translation / rotation / scale), the labels themselves were simple copies of the MRI atlas labels. This means that they could not adapt to individual anatomical variation, and all volumes at the same gestational age had labels corresponding to the atlas.

On the other hand, the US atlas was generated with nonrigid registrations, making the shape of each volume's labels specific to the individual volume. The use of displacement maps ensures that individual variation is kept into account, allowing for a much more diverse dataset and labels that are a closer reflection of the underlying anatomy. The number of volumes to which this technique can be applied is limited to those originally selected to construct the US atlas, but that is still a large enough number to train a CNN.

Interpolation can be a significant source of error, especially since the chosen method involves two registration steps which may introduce quantisation noise. This chapter carefully considers possible design choices and selects an approach that minimises this source of error.

Nonetheless, even the US atlas-based labels rely on an initial affine transformation between the MRI atlas and the US atlas, which does not take differences between the modalities into account. This leads to a consistent systematic error in segmentations: the segmentation of the US atlas is slightly inaccurate, and these inaccuracies propagate to all the volumes used to construct the atlas. Another systematic error is caused by the simple fact that the atlases were constructed with different volumes in different modalities: the boundaries of the structures of interest are different regardless of registration.

The temporal resolution of the Gholipour atlas as well as the US atlas (1 week each) is also fairly low and can cause inaccurate segmentations. The combination of interpolation

errors, errors in affine transformation estimation and limited temporal resolution limits the accuracy of the segmentations generated with this method. Table 4.5 shows the consistency of segmentations on volumes near the border between two age bins, which can be thought of as a bound on how consistent the segmentations produced by a neural network can be.

4.6.2 CNN-based network segmentation

Atlas-based segmentation is necessarily limited: it only can only be used on volumes that are included in an atlas, and for which a set of correspondences has been clearly laid out. This makes it difficult to scale up to a clinical dataset and to new volumes, and a different method is needed for unseen volumes.

This chapter set out to compare the performance of three different network architectures, all based on the U-Net for segmentation, with different design philosophies. A 3D multi-task network, a 2D multi-task network, and an ensemble of 3D single-task networks were proposed, and trained on the same data. Due to the design choices made, training times only differed slightly, though inference times did vary between architectures, as in some designs multiple networks were needed.

The 3D multi-task network performed best overall, especially on smaller structures like the thalamus and the cerebellum. The single-task network performed the best segmentations on the large white matter label, but did worse on smaller structure labels. One possible explanation for this difference is the amount of labelled data available: physically smaller structures of interest result in fewer voxels labelled as the structure, and therefore less data available to train the network and a greater degree of overfitting. This is somewhat counteracted in a multi-task architecture, as the other labels are included and can act as a form of regularisation for the network. This is less of a concern for larger structures, such as the white matter, so the performance benefit of a multi-task architecture is diminished in those cases. On the other hand, constraints on GPU memory limit the number of trainable parameters, which may lead to worse segmentation performance as seen for the white matter in the 3D multi-task architecture.

Across architectures, segmentation of smaller structures, such as the thalamus, results in a significantly lower Dice coefficient than segmentation of the white matter on the same network. This can be explained by their differing physical characteristics: in the dataset used, the white matter typically has a volume 15 times greater than the thalamus at 20 weeks, and 20 times greater at 24 weeks. The Dice coefficient is therefore biased by the larger number of interior voxels that can be predicted with high confidence, compared to voxels near the surface for which classification is more difficult. On the other hand, measures such as the Hausdorff distance are lower on smaller structures.

On the other hand, segmentation performance by the CNN appears to be superior to the atlas' consistency between weeks (as seen in Table 4.5). This may be surprising, considering that the consistency metric (comparing segmentations from different weeks near temporal boundaries) can be thought of as a possible upper bound to the performance of a learning algorithm. This can be explained by the fact that the measured atlas consistency is just an estimate: it was after all only measured in volumes near gestational age boundaries: one may expect that these labels are more self-consistent and predictable when the volumes to segment are in the middle of their age bins, and that therefore the accuracy measured is an underestimate.

4.6.3 Clinical implications

Clinical measures that can be estimated from this data, such as the transcerebellar diameter, are broadly in line with other published data. Volumetric changes throughout gestation can also be straightforwardly measured with this data: I found that the different structures I segmented show divergent growth rates in the 5-week period chosen here. This is consistent with results in the literature, such as Nakae et al's [155] measurements from postmortem histology. Scott et al [27] also find using MRI volumetry that the cortical plate displays a faster growth rate than other structures in the second trimester. Clouchoux et al [156], examining a later gestational age range, find that cerebellar growth is fastest of all brain structures they examined. Our results are also consistent with the broader understanding of

fetal brain development.

This analysis is limited by the lack of manually labelled data. Since the network is trained on labels generated by the atlas, the reliability of those labels is necessarily limited. There may be some systematic error introduced by the fact that the original segmentations were performed in MRI, as well as the registration to MRI and interpolation steps. Later chapters investigate clinical applications in more detail.

4.7 Summary

In this chapter, I explored different ways to obtain large quantities of labelled data for 3D US segmentation. I used an atlas-based method to generate hundreds of 3D labels for a US segmentation task, starting from an atlas generated in a different imaging modality. This chapter showed that this is a viable method to generate a training set for a neural network, and a CNN trained on this training set can generalise to unseen data with performance comparable to the initial atlas transform.

I used labels from a publicly available 3D atlas of the developing fetal brain to generate this training set. Since this atlas was constructed from MRI volumes, the first step was to register the labels to an atlas generated from the INTERGROWTH-21st dataset itself. From here, it was possible to propagate the atlas labels to individual volumes used to construct the atlas in the first place, simply by reversing the displacement map. This generated a large dataset of volumes with individual segmentations that account for individual variations without a need for manual segmentation.

Still, this dataset could only cover the volumes used to create the INTERGROWTH-21st US atlas in the first place. A different method was needed to generate volumetric labels for new data. The atlas-based labels obtained in the first part of this chapter provided a sufficient dataset to train a CNN with: I considered several architectures and design choices to find the one that performed best. The final choice obtained segmentations similar in quality to the original atlas-based labels, and with lower inference time. Comparisons with clinical data (where such comparisons can be drawn) show that this method makes anatomically plausible

predictions that seem broadly in line with the literature.

One major challenge to the results in this chapter is that the atlas-based ground-truth generation, while sufficient to train a CNN, is based on imperfect estimates of correspondences and segmentations originally obtained in a different modality. In other words, the initial segmentations are not as good as a fully manually-segmented dataset, and imperfect labels fed in as training data to a CNN necessarily lead to imperfect outputs. Improving the quality of the training labels in this dataset would improve the segmentation performance of this network.

However, it remains impractical to manually segment a training dataset of similar size to obtain good performance on this segmentation task. Different possible approaches to address this are explored in subsequent chapters.

Chapter 5

Uncertainty as a selection tool for omni-supervised segmentation of the fetal cerebellum

Chapter layout

This chapter discusses measures of uncertainty, estimated as prediction variability, in the CNN-based segmentation of the fetal cerebellum, and how those measures can be used to select data for boosting and further training, and enhance segmentation quality. The core challenge addressed in this chapter is leveraging the unlabelled data present in a large dataset to enhance the quality of segmentation obtained when only a small fraction of the data is manually labelled. This is a direct response to the challenge presented in the previous chapter, of the difficulty of obtaining accurate training labels for medical image segmentation tasks. This chapter shows how uncertainty estimates can be used to select volumes to add to the training set in omni-supervised methods.

I begin by providing an overview of the task attempted in this chapter in Sections 5.1 and 5.2. I then describe the datasets and the annotations used, and the validation performed on those annotations, in Section 5.3. This section also describes an external dataset used to validate the results in this chapter. A network and a training protocol are specified to

quantify uncertainty of the unlabelled dataset in Sections 5.4. These results are presented in Section 5.5 and I discuss them in Section 5.6.

5.1 Introduction

A major challenge in medical image segmentation is the scarcity of appropriately labelled data [39]. While imaging is often used in routine clinical applications and it is sometimes possible for researchers to gain access to large datasets containing image data from many subjects, it is much harder to obtain accurate segmentation labels for the data. Accurate manual labels, especially for segmentation tasks in 3D, are time-consuming to produce and often are not part of a standard clinical protocol.

Given the specialisation of medical imaging, only trained clinicians can produce reliable labels, and they often do not have the time required to produce 3D labels for a large dataset. As a result, researchers often face the situation of having labels available for only a subset of medical image datasets, and a larger unlabelled set of data.

The question I approach in this chapter is how to leverage the presence of a large unlabelled dataset to improve the segmentation quality obtained by automated methods from a limited number of labelled images. I use the established method of *omni-supervised learning*, which uses high-quality automatically generated labels to improve predictions with a small initial labelled set. I propose a novel data selection method, based on estimates of segmentation uncertainty, to improve the performance gain for omni-supervised methods. I explore the effects of different data selection approaches on the performance of omni-supervised segmentation methods on my dataset, consisting of a large dataset of 3D fetal brain ultrasounds, of which only a subset is labelled with segmentations of the cerebellum.

I then generate segmentation labels for all volumes in this dataset. This generates a large segmented dataset, which leads to more data-driven analysis to be explored in the next chapter.

5.2 Proposed approach

This chapter aims to maximise performance on a segmentation task in the case where only a fraction of the dataset has manual training labels. I propose to do this by refining the established technique of omni-supervised learning and then applying it to this segmentation task. More specifically, I plan to focus on the use of aleatoric and epistemic uncertainty measures as a possible selection tool for omni-supervised learning.

The central insight leading to this chapter is that the principles of model diversity and data diversity are very similar to the methods used to estimate epistemic and aleatoric uncertainty. Model diversity, like epistemic uncertainty estimation, involves making changes to the segmentation model to obtain different predictions: the difference in predictions with model variation is also used to measure epistemic uncertainty. Similarly, data diversity involves test-time data augmentation, which is also used to measure aleatoric uncertainty. Therefore, it is possible to obtain estimates of the segmentation's uncertainty with no additional training or testing time.

I hypothesise that these uncertainty measures may be a useful guide to selecting initially-unlabelled data to add to the training dataset for subsequent rounds of omni-supervised learning. I propose an experiment to verify this. A CNN is trained on the labelled data used in this chapter and is then used to generate segmentations and uncertainty estimates for all the unlabelled data. The network is then retrained using additional labels from the unlabelled set in an omni-supervised manner, and the additional labels are selected based on different measures of uncertainty, or randomly (as a control). The segmentation quality is then compared between identical volumes segmented by these differently trained networks.

Additional experiments are done to validate the method and select of the necessary hyperparameters. The experiments conducted are:

- Measuring the effectiveness of omni-supervised methods with different starting amounts of labelled data

- Selecting an appropriate amount of dropout

- Measuring and compensating for any artifacts introduced by augmentation methods

These are important for establishing the usefulness the methods used here.

I then analyse the cases where this method fails and attempt to explain its modes of failure. While it is difficult to quantify performance in unlabelled data, I analyse the cases where the entire volume is segmented as background and attempt to determine the cause of the error. This is done on two datasets, the ultrasound INTERGROWTH-21$^{\text{st}}$ dataset and the MRI EADC-ADNI/HARP dataset.

To the best of my knowledge, this is the first application of omni-supervised learning to use estimates of data and model uncertainty as a selection tool.

5.3 Datasets and labelling

5.3.1 INTERGROWTH-21$^{\text{st}}$ dataset

The data used in this chapter comes from the INTERGROWTH-21$^{\text{st}}$ study, a longitudinal and multi-centre study that acquired multiple ultrasound acquisitions from 4321 optimally healthy pregnancies [18]. Stringent exclusion criteria were used to ensure that only the healthiest pregnancies were included: only nulliparous women with low risk factors were included, and they were scanned every 5 weeks from $GA = 14$ weeks to delivery. 3D ultrasound volumes of the fetal body were acquired using a consistent protocol across eight centres in eight countries. This study aimed to set international standards for fetal growth, but the very large quantities of data acquired have made it possible to use this dataset for other image-analysis studies [11].

There are 948 3D ultrasound acquisitions in the dataset. The study was conducted to obtain standard clinical measures based on skull size[1] with no regard to internal brain structures, which are weaker ultrasound reflectors and more prone to imaging artifacts (see Section 2.4). As a result, brain structures in many volumes in the dataset could not subjectively

[1]Specifically, the measures obtained from the head acquisitions were of fetal head circumference, biparietal diameter, and occipitofrontal diameter.

be visualised. After visual inspection, a subset of 908 head acquisitions was selected in which brain structures could be seen and had not been corrupted by artifacts. The selected volumes were all in the range of $GA = 14 - 25$ weeks[2]. Of this dataset, 186 volumes (20%) were selected for manual segmentation. Figure 5.1 shows the distribution of age ranges in the dataset used. I ensured that the volumes in the manually segmented dataset would follow the same age distribution as the overall dataset, by randomly selecting 20% of the data within each gestational week.

The imaging protocol used required acquisitions to be collected perpendicular to the mid-sagittal plane (MSP). This causes only one hemisphere to be visible, as discussed in Section 2.1.1. It also can cause shadow artifacts over the cerebellum from the temporal bone. The acquisition protocol was agnostic as to the hemisphere from which the ultrasound acquisitions were made, so some acquisitions were taken from the right (with the left hemisphere as the more visible distal hemisphere), while others were taken from the left.

The volumes from the ultrasound acquisitions consist of much of the fetus and some surrounding maternal tissue. The large differences in head and probe positioning and fetal size and orientation between acquisitions make it challenging to analyse directly. Therefore, all volumes were brain-extracted and the skulls were registered to each other following the method laid out by Namburete et al [137]. They were also scaled to the same dimensions of $160 \times 160 \times 160$ (using tricubic interpolation where needed).

5.3.1.1 Manual segmentation

It is time-prohibitive to generate manual segmentations of individual brain structures in the entire dataset of 948 3D volumes, but it is possible to do so for a subset of this dataset. Therefore, I attempted to generate 3D manual segmentations of the 186 (20%) volumes in the training set. However, many volumes (especially at higher gestational ages) displayed an ultrasound shadow cast by the petrous ridge that obscured part of the cerebellum and made it impossible to segment some of its boundaries (see Figure 2.3, in Chapter 2). Volumes which

[2]After 30 weeks, ossification of the temporal bone leads to high rates of occlusion of the cerebellum in the imaging protocol followed.

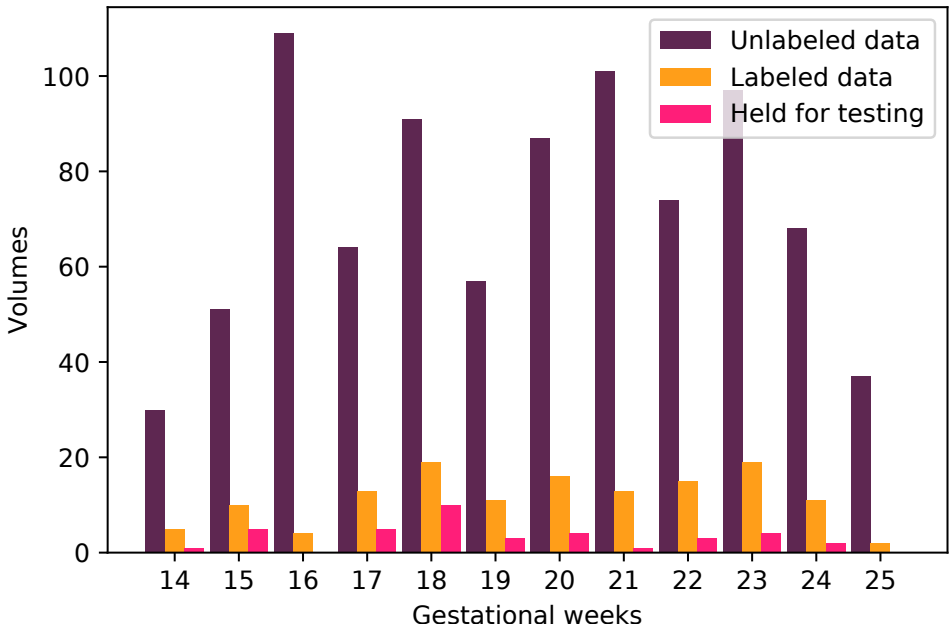


Figure 5.1: Distribution of gestational age in the dataset used for the experiments here. 25% of the volumes in each gestational week were selected for manual segmentation. Of these, 40 were held back for testing, leaving 106 volumes for training.

contained a strong shadow which obstructed view of the cerebellum needed to be discarded, as no manual segmentation could be performed with any confidence. This was only noticed in some volumes at the higher end of gestational age.

A total of 146 (16%) manual segmentations were performed on the ultrasound volumes, after exclusions. Each manual segmentation took an average of 20 minutes to perform, and a further 10 minutes to refine. Segmentations were made using MITK Workbench [157] on a Wacom touchscreen with a stylus. Some examples of the segmentations produced are shown in Figure 5.2. All segmentations were performed by the same individual. Even without access to a clinician to verify the segmentations performed, this ensures that the segmentations are consistent, and delineating the same anatomy. An example can be seen in Figure 5.2. Figure 5.1 shows how the dataset is split between labelled and unlabelled data, and how much was reserved for testing.

5.3.1.2 Intra-observer variability

The low signal-to-background ratio and presence of artifacts in ultrasound acquisitions make it difficult to find precise anatomical boundaries within an ultrasound volume, leading inevitably

to some uncertainty even in the manual segmentation process. This would lead to some variation in the boundaries of the manual segmentations drawn, even if performed by the same segmenter.

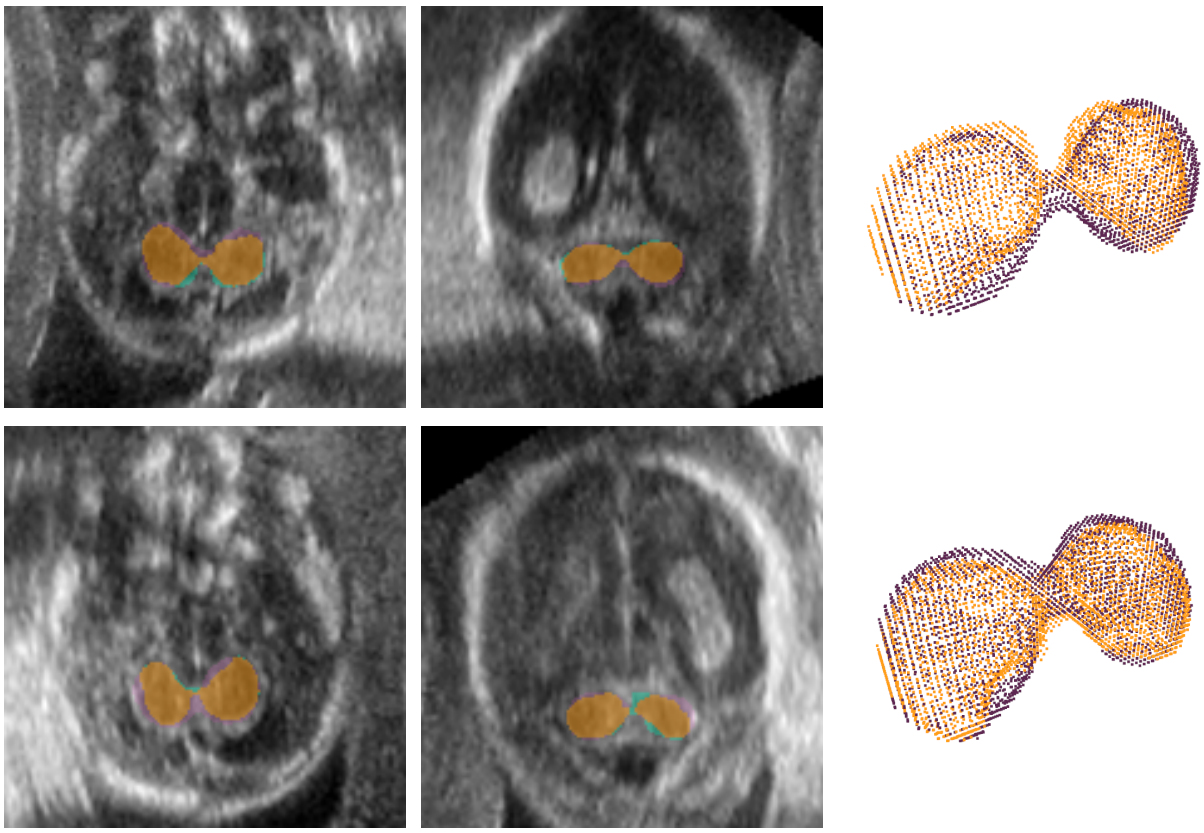
To quantify this variability, I performed a second, independent manual segmentation of a randomly selected subset of 30 volumes. These were all done by the same segmenter who had performed the first manual segmentation, and without reference to the first segmentation. Since all manual segmentations were performed by the same individual originally, the best machine-learning algorithm trained on that dataset can learn to segment as well as the segmenter - thus measuring this individual's variability is a true, and fair, measure of the upper bound of performance by any programmatic method[3]. Some examples of the different segmentations are shown in Figure 5.2.

Metric	Score
Dice coefficient	0.764 ± 0.060
Jaccard coefficient	0.622 ± 0.078
Hausdorff distance (mm)	3.96 ± 0.99
Euclidean distance (mm)	1.21 ± 0.68

Table 5.1: Measures of intra-observer variability for a sample of 30 3D volumes.

Visually, they appear similar, but with noticeable differences around the edges and the centre. Anatomical boundaries are often not very well defined in ultrasound, and it is difficult to delineate them exactly. Table 5.1 shows a quantitative comparison of the two methods, showing the Dice coefficient [112], Jaccard coefficient (or IoU) [158], Hausdorff distance [113] and Euclidean distances of segmentations. The Dice and Jaccard coefficients appear low, which may be due to the small anatomical size of the structure and the difficulty of consistent segmentation. I expect, therefore, that a machine learning method trained on the data used here and corresponding segmentations is likely to display more uncertainty in the regions where the manual annotator also showed more variation: the edges and the connection between

[3]Measuring *inter-observer* variability, using a different segmenter, is likely to lead to worse agreement. I believe this is less relevant to this task, as all of the labels were made by the same segmenter.



(a) (b) (c)

Figure 5.2: On each row, axial and coronal slice of two independent cerebellar segmentations of the same volumes (18 weeks and 19 weeks), by the same segmenter. Voxels that were segmented by both are shown in yellow, while voxels segmented by only one are shown in purple or green. (c), a 3D rendering overlaying both segmentations.

the two hemispheres. These results also provide an upper bound to the segmentation quality of any automated methods.

5.3.2 ADNI-HARP dataset

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is an ongoing multisite longitudinal study that acquires structural MRI brain volumes from cognitively normal ageing adults (CN), as well as those with mild cognitive impairment (MCI) and Alzheimer's disease (AD) [138]. 63 sites participate in acquisition, using different scanners and acquisition methods.

The EADC/HARP dataset consists of a subset of 135 volumes selected from the ADNI dataset, with expert 3D manual segmentations of the hippocampus. All the volumes in EADC/HARP are 1.5T T1-weighted acquisitions with 1mm resolution along all axes, and

brain-extracted and registered to the same image space. Of these, 44 were healthy, 46 had MCI and 45 had AD: the segmentations of these varied substantially as the hippocampus is one of the areas of the brain most affected by Alzheimer's disease [159]. Figure 3.6 shows sample segmentations in the HARP dataset. Of the 135 labelled volumes, 80 were used for training, 20 for validation, and 35 for testing.

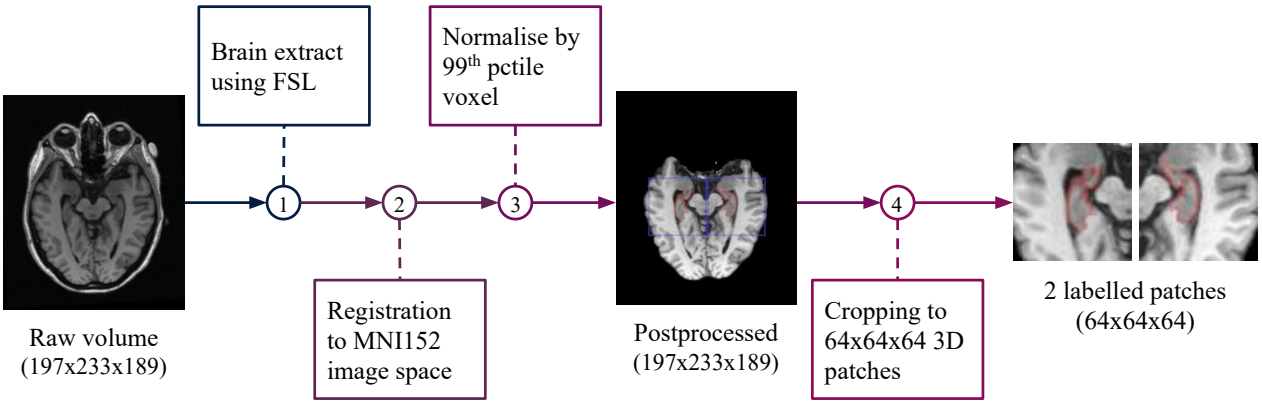


Figure 5.3: The pipeline used to register unlabelled ADNI volumes to a common image space.

680 additional unlabelled volumes from the ADNI dataset were also used, each from different subjects, removing any volumes already in the HARP dataset. These volumes were selected to match the acquisition parameters used (1.5T T1 with 1mm resolution). All volumes were brain-extracted, linearly registered to MNI152 image space [140, 20] and normalized by dividing each voxel by the 99^{th} percentile value. $64 \times 64 \times 64$ patches were extracted around the hippocampus in each hemisphere, leaving 270 labelled patches and 1360 unlabelled patches. FSL [160] was used to perform this preprocessing. Figure 5.3 shows this pipeline.

5.4 Uncertainty quantification in network design

5.4.1 Network design

To measure the aleatoric and epistemic uncertainty in the predictions generated by the model (see Section 2.4.1), I can find the differences in the predictions generated using test-time dropout and test-time augmentation.

As described in Section 2.4.1, epistemic uncertainty arises from the model, where training

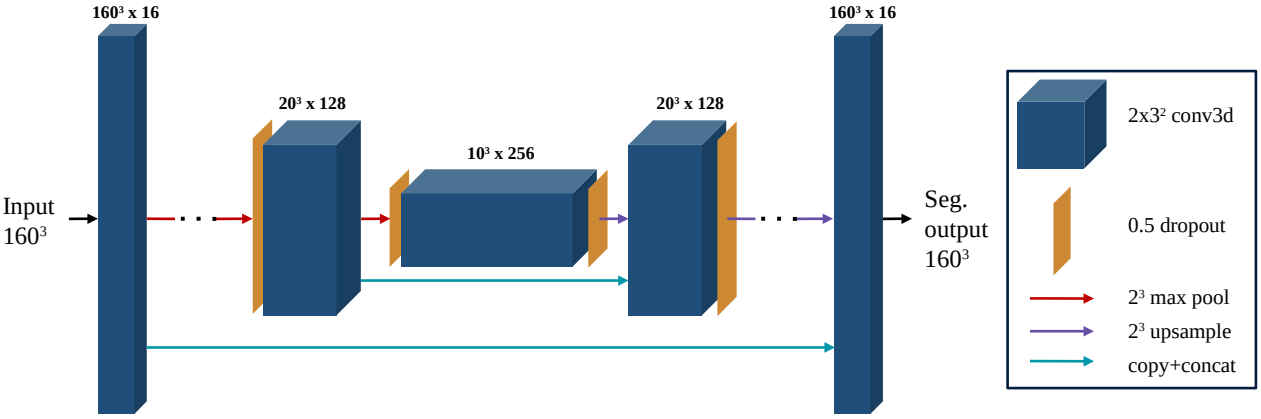


Figure 5.4: The Dropout U-Net architecture I implemented for this work. For clarity, the middle layers (at scale 80^3 and 40^3) are omitted from this diagram.

is insufficient to reach a confident accurate segmentation at test time. This could be because the data does not resemble the training set. The output of the network in these cases will tend to change when varying the structure of the network: epistemic uncertainty therefore can be measured with test-time dropout [121]. This varies the network at test time, producing different predictions even when the same data is presented to the network for prediction.

To this effect, I implement the Dropout U-net proposed by Kohl et al [124][4]. This is a minor variation on the conventional U-Net network architecture that uses Dropout to obtain uncertainty estimates, inspired by Kendall et al's Bayesian Segnet [123]. After each max-pool layer or before each upsampling layer, half of the previous layer's weights are dropped out (or set to 0). This enforces sparsity in training [122] and is a sufficient change in the model to estimate the model's uncertainty. The architecture of the network is shown in Figure 5.4.

The major difference between the network architecture proposed by Ronneberger and the one proposed here is the use of 3D volumes (and 3D convolutions), while the original network architecture only segmented 2D images. This is very similar to the 3D U-Net network [151]. The large-scale network architecture was the same between 2D and 3D versions, but all 2D convolutional layers are replaced by $3 \times 3 \times 3$ 3D convolutional layers and 2D max-pooling layers by $2 \times 2 \times 2$ 3D max-pooling layers. Moving to 3D also led to a large increase in the number of trainable parameters, so to fit the network into the memory of a single GPU the

[4]The U-Net is a widely used architecture for segmentation tasks, making it a sensible choice for this chapter: the focus in this chapter is on uncertainty measures and data selection methods that should be architecture-agnostic, so there is no need to extensively experiment to find the best architecture for this task.

number of filters had to be halved (to 16 in the first layer) and reduce the batch size to 1. All of the models used here were written in Python 3.7 using Tensorflow 1.14 and Keras 2.4. They were trained on Nvidia GTX 1080 Ti GPUs.

5.4.2 Test-time dropout

To compute epistemic uncertainty in my model's predictions, I use the method developed by Kendall and Gal [120]. Using test-time dropout, I generate 40 independent predictions for each unlabelled volume. The uncertainty in the segmentation of each voxel is then given by the variance in the voxel's predicted value across those different runs[5]. I can then obtain a single number corresponding to a global measure of the epistemic uncertainty in a volume, simply by summing the uncertainty of all voxels

$$\text{uncertainty} = \sum_i \text{Var}_n\left(x_i\right)$$

where $\sigma_n^2\left(x_i\right)$ is the variance of the value of each voxel x_i over n samples.

5.4.2.1 Choosing the number of samples

Applying a random dropout to the network's weights leads to slightly different predictions. Each prediction $x_{i,n}$ can be modelled as an independent sample from an (unknown) distribution of all possible predictions obtainable from the different data, with different weights dropped out. An estimate of the epistemic uncertainty of the model can then be found for each voxel by calculating the variance of that voxel's predicted value across N samples:

$$\text{Var}_n\left(x_i\right) = \sum_{n=1}^{N} x_{i,n}^2 - \left(\frac{\sum_{n=1}^{N} x_{i,n}}{N}\right)^2$$

This estimate, however, is based on a finite number of observations, so it is itself prone to uncertainty, and differs from the true variance of the distribution. Cochran's theorem [161] proves that the variance of a finite sample of a random variable is itself a random variable

[5]As most voxels are consistently segmented as cerebellum or background, the uncertainty of most voxels is 0.

that follows a chi-squared distribution[6]. Its mean is the true variance of the function, and its own variance is given by:

$$\mathrm{Var}\left(\mathrm{Var}_n\left(x_i\right)\right) = \frac{2}{N-1}\mathrm{Var}_n\left(x_i\right)^2$$

The variance of the sample estimate, therefore, decreases inversely with the number of samples (or equivalently, its standard deviation follows the inverse square root of the number of samples.) The standard deviation of this is equivalent to the expected error between the true variance of the voxel and the measured variance from the sample. This only shows a slow decline with increasing number of samples.

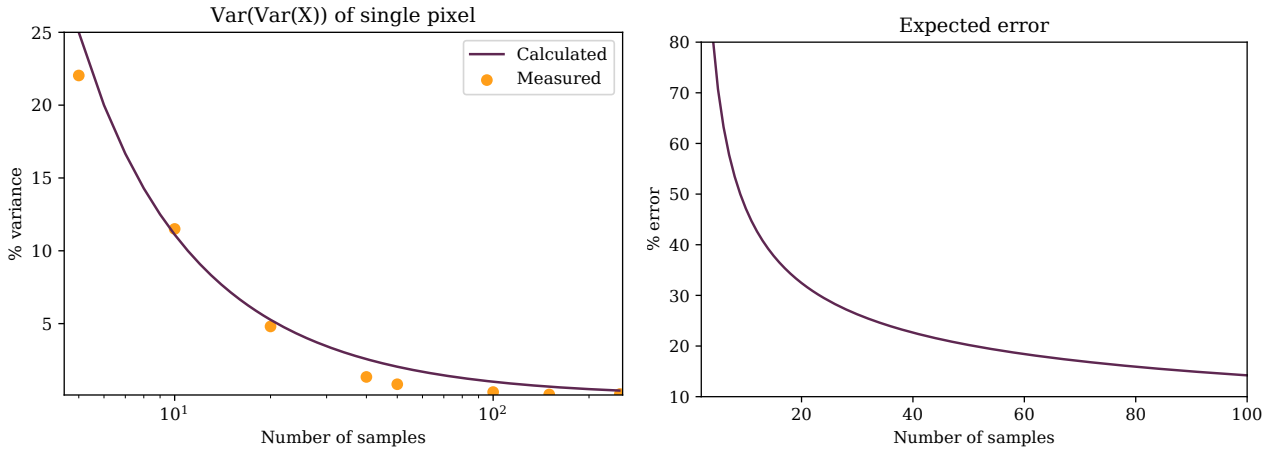


Figure 5.5: (a) The variance of variance (sum of absolute differences) measurements of a voxel with test-time dropout, changing with the number of samples. (b) the expected percentage error between the measured variance of a sample and the sample size.

Figure 5.5b shows how the expected error changes with increasing sample size. My selection of 40 leads to an expected error on the measurement of variance of 22.6%. The choice of 40 was driven by practical considerations: a much larger number would have improved variance estimates but would have led to a much slower segmentation process lasting several days.

This analysis is limited to the variance of an individual voxel. The overall uncertainty of the model is estimated from the sum of variance over all voxels, so in practice I expect that the global error will be lower. Voxel classifications in a single segmentation are not

[6]Subject to the assumption that the possible values for a given voxel are normally distributed.

independent of each other, so it is difficult to estimate how different this is in practice.

Figure 5.5a shows the difference between the expected variance of the variance measurement when it is measured from independent test-time dropout samples. It seems clear that the experimental measurement largely follows the prediction made in the equation above, though it seems to decay somewhat faster with increasing number of samples. The discrepancy seems likely to be due to the assumptions made above: the calculations made here assume Gaussianity, which seems unlikely to match the true distribution. The faster decay means that the true error in variance estimates is likely to be lower than the calculation in Figure 5.5b.

5.4.2.2 Dropout level

Dropout level	Dice coefficient	Mean variance (sum)
0	0.64	N/A
0.1	0.64	245.4
0.2	0.64	431.5
0.3	0.63	593.0
0.5	0.62	917.1
0.8	0.34	1192.2
0.9	0.01	2509.7

Table 5.2: The Dice coefficients and average variance (sum over all voxels) in predictions of networks trained on the same labelled data. The computed Dice coefficient decreases monotonically, though this is a very small effect with low levels of dropout. Example outputs are shown in Fig. 5.6.

An experiment was performed to determine the choice of dropout level. The network implemented by Kohl et al [124] in their "Dropout U-Net" implementation drops out half of the connections per dropout layer. It is worth exploring how varying the dropout level changes the uncertainty measures obtained. There seems to be an implicit trade-off when selecting dropout level: low dropout leads to very low variation in the model and little variation in the predictions generated even when the model is uncertain, while very high dropout leads to poor segmentation performance. The experiment performed involved using different levels of

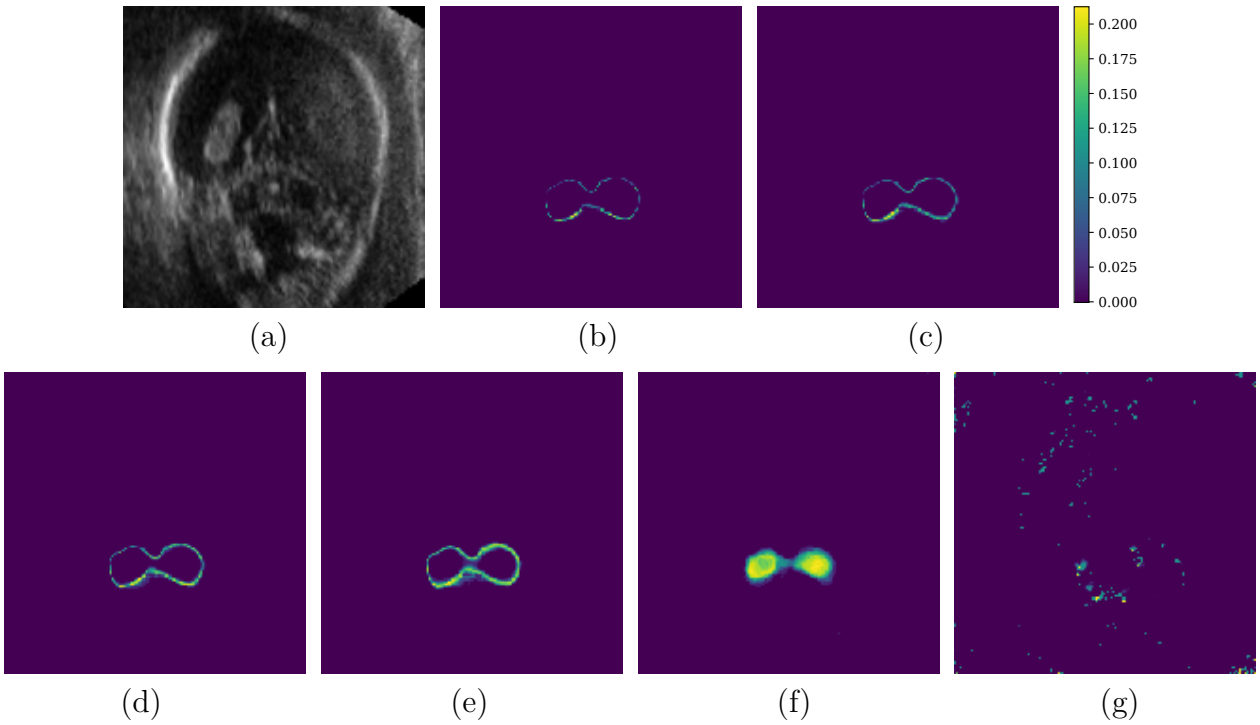


Figure 5.6: Examples of dropout uncertainty in volumes trained with (b) 0.1, (c) 0.2, (d) 0.3, (e) 0.5, (f) 0.8, (g) 0.9 dropout. (a) The original volume.

dropout at test-time, using the same network architecture with the same weights.

Table 5.2 shows the trade-off described above on the data in the validation set. The variance of predictions increases steadily with increasing dropout, while the Dice coefficient (obtained from the median of 40 prediction) slowly declines. This can also be seen in Figure 5.6: up to 0.5 dropout the variance in predictions increases, mostly near the edges of the segmentation, while the shape of the resulting segmentation is largely unchanged. At 0.8 dropout there is considerable uncertainty even in voxels that seem clearly to be in the cerebellum, and the segmentation algorithm completely fails with 0.9 dropout.

The dropout method implemented was the same as proposed by Hinton et al [122], and was applied to every layer of the network. Some experiments to vary this were performed, such as only applying dropout to one half of the network (either the downsampling path or the upsampling path) but the results did not appear to differ much. Therefore, the original Dropout U-Net architecture was maintained.

5.4.3 Test-time augmentation

Aleatoric uncertainty is uncertainty inherent in the data, often due to noise in the data, low contrast, or inconsistent labelling in the training data. It is the second major source of uncertainty in a segmentation task.

Estimating aleatoric uncertainty is conceptually similar to estimating epistemic uncertainty. While epistemic uncertainty arises from the model and can be measured by varying the model at test time, aleatoric uncertainty arises from the data and can be measured by varying the data at test time. I vary the data at test-time using augmentation.

The augmentation used consists of simple similarity transforms, identical to the scheme used in Chapter 4. Data augmentation was employed and consisted of: small translations (up to ± 5 voxels along every axis), small rotations (up to $\pm 10°$), and scaling (up to $\pm 15\%$). Reflections were also used across the MSP. Once a prediction had been obtained on an augmented volume, the transforms for the augmentation were simply reversed to return to the common reference space. Aleatoric uncertainty was estimated using the variance of the resulting predictions, after test-time augmentations. The variance was measured across 40 predictions per volume.

The same augmentation scheme was used in the ADNI/HARP segmentation task. All volumes in the ADNI/HARP task were registered to MNI152 image space as a preprocessing step. Augmentation is thus less physically realistic and would not be expected to improve segmentation performance. Nonetheless, test-time augmentation is used in my scheme to estimate aleatoric uncertainty, so augmentation must be used in training. Empirically, the segmentation performance of a network on this dataset is unaffected by the use of augmentation and the segmentation performance obtained here is similar to other published results on this dataset [162].

Most of the augmentation methods used here, including rotation, translation, and scaling, have the potential to cause a loss of data near the edges of the image (as the transforms may lead to that data being moved out of frame). This is not a concern for ultrasound segmentation: the voxels near the boundary of the image are always outside of the skull and

contain no useful information. In the MRI dataset, the augmentations were applied to larger patches ($74 \times 74 \times 74$) and then cropped to remove these effects.

5.4.3.1 Quantisation errors

(a) (b) (c) (d)

Figure 5.7: (a) A slice of a cerebellar volume, and (b) its segmentation. (c) The same segmentation, after a random augmentation has been applied and reversed. (d) The voxelwise variance of this slice across 40 such transformations.

Segmentation is a voxel-wise classification task, which makes it fundamentally discrete in nature. This complicates the approach described here, because some of the test-time transforms described here, the rotation and scaling operations, are continuous. When the image is transformed, in general voxels from an image lie between the new voxel values, making interpolation necessary. This is a lossy process, and it is one that performed with all segmentations obtained when doing test-time augmentation.

A simple model can give an idea of the magnitude of this effect. One may assume that the amount of shift Δx in a voxel's position along one axis given by nearest-neighbour interpolation following one of the transformations made here is a sample from a uniform distribution

$$P(\Delta x) = \begin{cases} 1 & |\Delta x| \leq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

Since the transformations made here are applied twice (once to augment the volume, and then again to return the segmentation to a common reference space), the shift in the position Δx

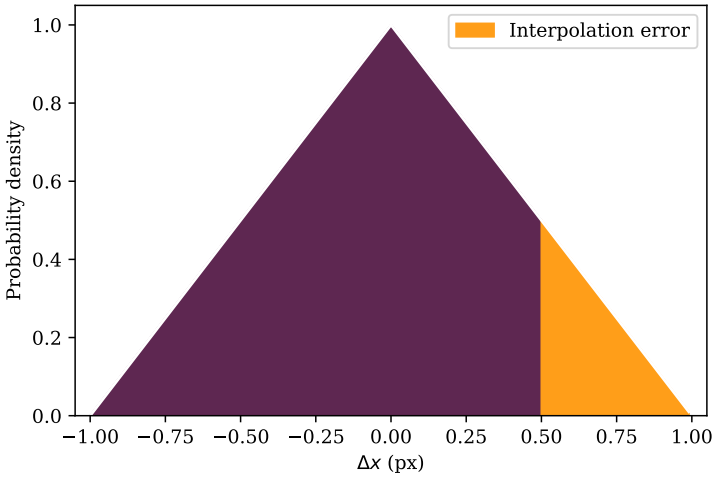


Figure 5.8: Pdf of the possible computed shift after two transformations with nearest-neighbour interpolation. With nearest-neighbour interpolation, a shift $\Delta x \geq 0.5$ moves the voxel by 1, leading to interpolation errors.

of one voxel along one axis following two transformations is given by the sum of two samples from the above pdf:

$$P(\Delta x) = \begin{cases} 1 - |\Delta x| & |\Delta x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

This pdf can be seen in Figure 5.8. When the position shift is greater than 0.5 voxels, nearest-neighbour interpolation shifts the voxel by 1 relative to its initial position. If this occurs in a voxel at a boundary and the shift is towards the boundary, that can lead to the voxel across the boundary being misclassified. The probability of this happening is equal to the area under the yellow area of the curve in Figure 5.8, or $\frac{1}{8}$ under these assumptions. This leads to an expected variance of 0.109 for voxels near a boundary in one dimension. We can assume that the shift along each dimension is independent, and that leads to the calculation that for voxels near boundaries along two axes, the expected variance is 0.179, and for voxels at boundaries along all three axes, the expected variance is 0.221. Considering each boundary has two voxels (one on either side) and the split between the number of boundaries per voxel in our segmentations, we can estimate an additional variance of about $0.33SA$, where SA is the number of voxels on one side of a boundary.

The result is that interpolation artifacts from different augmentations artificially increase

the variance of segmentations near their boundaries. An example of this can be seen in Figure 5.7: a slice from a segmentation volume has been randomly transformed and then the transform has been reversed, using nearest-neighbour interpolation both times. Figure 5.7(c) shows the variance of voxel values across this slice after 40 such examples. It is highest near the edges of the structure, biasing uncertainty estimates in that area. This is inevitable and likely to add some baseline variance to my estimates in later experiments. The example I used represents the worst-case scenario: applying a transform and reversing it means applying nearest-neighbour interpolation twice on the segmentation. In testing, on the other hand, the image is transformed and then the trained model is used to segment it. The inverse transform is then applied to the segmentation and nearest-neighbour interpolation applied - so the segmentation is interpolated only once, leading to milder artifacts.

5.4.4 Omni-supervised segmentation

The volumes with manual segmentations of the cerebellum still account for under 20% of the dataset available. Omni-supervised methods (described in Section 2.4) are one way to exploit large amounts of unlabelled data to improve the quality of the segmentations generated by CNNs.

A CNN is trained on the manually labelled data available and used to generate segmentations of the unlabelled data. Other segmentations of the same data are also obtained, using model diversity and data diversity - changing the model or applying transformations on the data, respectively. An aggregate consensus segmentation is then obtained, which should be more robust than any one segmentation.

The segmentations are then combined to form a consensus, which should be more robust than any one of the methods. The aggregation can be done in several ways: in this work, I used the median prediction for each voxel as a simple method.

A subset of the unlabelled dataset is segmented in this manner and the resulting segmentation is then used to retrain a new iteration of the network. This improves the segmentations produced by the retrained network, which in turn is used to segment a new subset of the

unlabelled data, and aggregate this segmentation to others as before to create new labels. This process is repeated until the entire dataset is labelled and is claimed to improve the quality of segmentations. Listing **??** in Chapter 2 shows an example implementation of this algorithm in pseudocode.

This framework has some variables that should be tuned to achieve the best segmentation quality: the way that different segmentation proposals are aggregated, and the number of iterations. I propose to experiment with these to find the best framework for automatic segmentation of the cerebellum (detailed below). Another experiment evaluates the impact of the quantity of manually segmented training data used, to evaluate whether the manual segmentation was sufficient. This is done by using the same framework with different amounts of labelled data as a starting point for the omni-supervised setup. This experiment also allows me to verify whether the amount of labelled data is sufficient for a high-quality segmentation output, and measure the benefit of using unlabelled data at different initial training set sizes.

5.4.5 Data selection

The central insight of omni-supervised learning is that using data with automatically generated data labels alongside data with manually generated labels can improve overall model performance, provided that the automatically generated labels are sufficiently high quality to aid network training. These labels are generated by aggregating a diverse set of predictions generated by different models and from different transformations of the data, leading to a consensus prediction. Even so, as long as the prediction is not perfect (obviating the need for multiple iterations) there will be some variance in the quality of the automatic segmentations. It would be best to select only those volumes that have the highest-quality automatic segmentations for use in training the next iteration of the network.

Without manual segmentations, it is impossible to directly estimate the accuracy of each segmentation of an unlabelled volume. However, I hypothesise that measures of aleatoric and epistemic uncertainty may provide useful indications of segmentation quality.

The epistemic uncertainty [120] provided by dropout uncertainty merits some more

consideration. As discussed in Section 2.4.1, the uncertainty generated by test-time dropout is uncertainty inherent in the model itself, and the variance in a segmentation under dropout can be conceptualised as a measure of the distance between the volume analysed and the training data. While higher uncertainty, in general, appears to point to a lower-quality segmentation, it is also important to increase the diversity of the training dataset and reduce the differences between the training data and the unlabelled data. I hypothesise, therefore, that selecting volumes segmented with a higher level of dropout uncertainty actually leads to an improvement in segmentation performance in the next round of training.

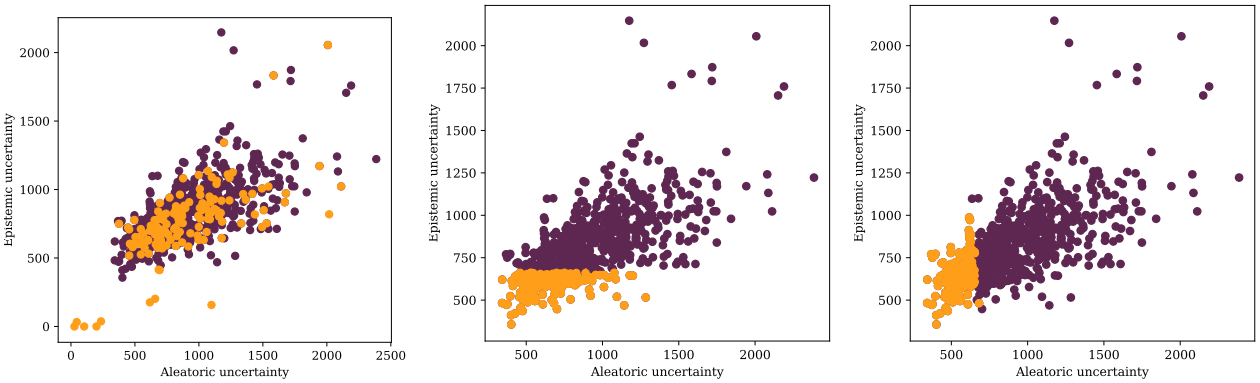


Figure 5.9: The volumes selected by each of the three proposed approaches for second-stage training, by the uncertainty estimates for them after the first stage. (a) Random selection. (b) Selecting the volumes with the lowest epistemic uncertainty. (c) Selecting the volumes with the lowest aleatoric uncertainty.

The same is not true of *aleatoric* uncertainty, the uncertainty inherently present in the data. This does not improve with larger training dataset and is linked to the data, rather than the model: a volume segmented with higher aleatoric uncertainty is likely to have a lower signal-to-background ratio. Therefore, I expect that selecting volumes with lower aleatoric uncertainty as shown in my module for iteration will lead to better results.

5.4.5.1 Proposed experiment

I propose an experiment to test these hypotheses and find the best method for iteration. I trained a model using the 106 labelled volumes in the training set and used the model to generate automated segmentations. The aleatoric and epistemic uncertainty of each

segmentation is estimated using test-time augmentation and test-time dropout as described in Section 5.4, and a single overall value of each measure of uncertainty is obtained. An aggregate overall segmentation is produced for each volume by taking the median value for each voxel across all predictions made of the volume. 150 volumes are then selected for the second stage of omni-supervised training: these 150 volumes are added to the training set (alongside the manual segmentations used) and the model is retrained using this expanded training set. I experiment with different selection criteria for those 150 automatically segmented volumes:

1. Randomly select the volumes, as a control (and similar to current methods)

2. Select the volumes with the lowest epistemic uncertainty

3. Select the volumes with the lowest aleatoric uncertainty.

Figure 5.9 shows what volumes were selected, based on their estimated aleatoric and epistemic uncertainties, for each of these experiments[7]. The same number of $n = 150$ volumes was selected for each method. I then measure the quality of each segmentation method on the 40 labelled volumes in the test dataset.

5.5 Results

5.5.1 Ultrasound dataset

The first experiment performed here is to determine the number of training volumes needed to initialise an omni-supervised method. This is essential to understand the role of omni-supervised learning, as the purpose of this chapter is to improve segmentation performance when only a fraction of the data available has manual labels. To evaluate the change in performance with increased training data and the role of omni-supervised learning, the network described in Section 5.4.1 was trained on a part of the labelled data available. After that, additional unlabelled volumes were selected (equal in number to the training volumes) to

[7]A detailed discussion of the uncertainty estimates in this scatter plot can be seen in Section 5.5.1.1.

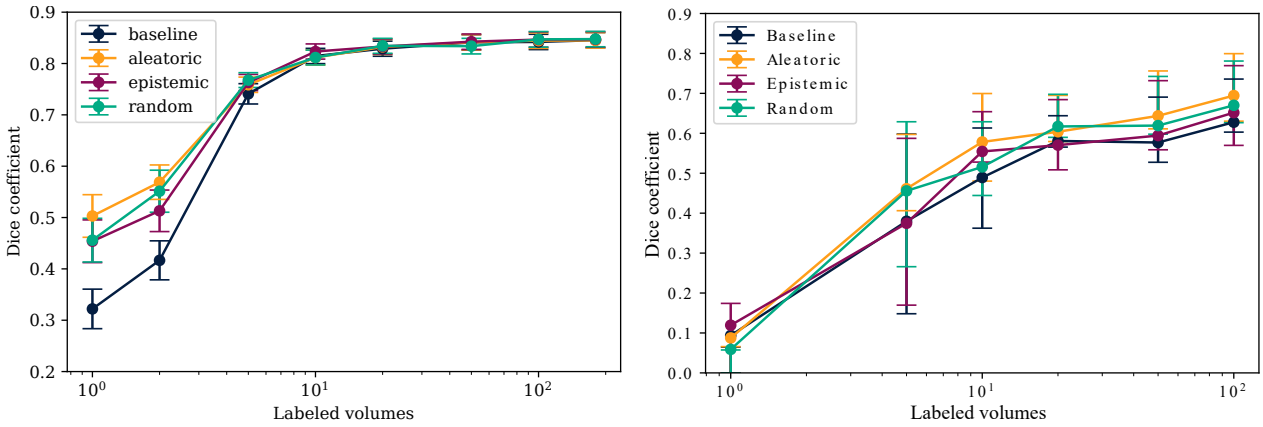


Figure 5.10: Performance of a model trained with different levels of training data, and the performance of the same model after adding 100 volumes labelled in an omni-supervised way to its training set in (a) the MRI dataset and (b) the ultrasound dataset. The different coloured lines represent different selection methods for omni-supervised learning, and the error bars represent the Dice coefficients at the 25$^{\text{th}}$ and 75$^{\text{th}}$ percentile.

serve as additional training data for a second round of learning, to evaluate the effect of omni-supervised learning at different initial data sizes[8].

The results of this can be seen in Figure 5.10. As expected, the performance of the network improves when more training data is provided. In the MRI dataset, diminishing returns are seen with more data, with performance stabilising beyond 20 labelled volumes, while performance improves steadily in the ultrasound dataset. More interestingly, omni-supervised learning seems to improve segmentation quality (especially in ultrasound) regardless of the amount of training data used: even with a single labelled training example, using omni-supervised learning to seed training of the next round leads to significant increases in Dice coefficient ($p \sim 10^{-8}$ when using aleatoric selection, $p \sim 10^{-5}$ when using random selection). The effect size always appears present in the range explored here.

5.5.1.1 Uncertainty estimation

Using the methods outlined above, I obtained estimates of the aleatoric and epistemic uncertainties for each volume in the unlabelled datasets. The global estimates for those volumes are shown in Figure 5.11. The number for aleatoric uncertainty includes the interpolation

[8]The data was selected by aleatoric uncertainty, as described above.

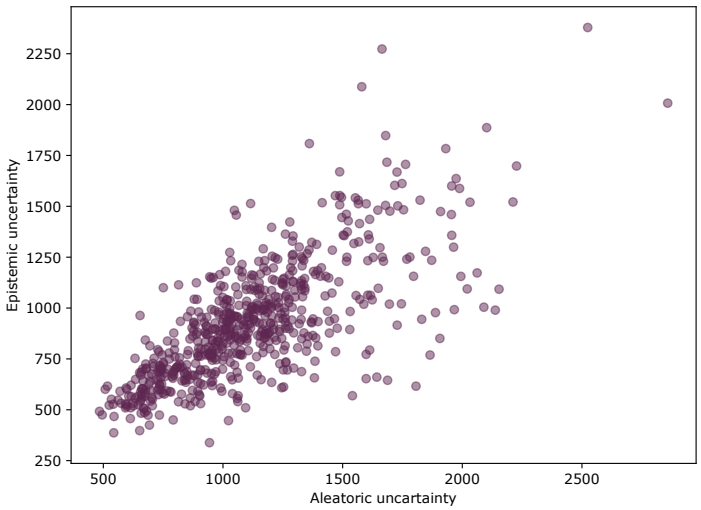


Figure 5.11: Scatter plot showing the estimated aleatoric and epistemic uncertainties of every unlabelled volume after the first round of training. Note the large variation between volumes and the correlation between these estimates.

bias I estimated in Section 5.4.3.1, which causes a large addition to the estimates of aleatoric uncertainty.

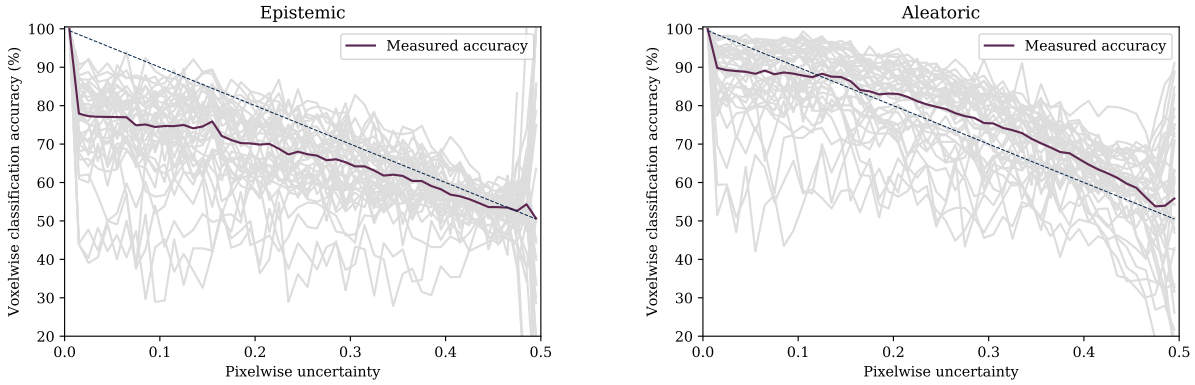
Not included in this plot are a few volumes ($N = 25$, 3.2%) where the segmentation method failed, and no voxels were classified as belonging to the cerebellum. In those cases, estimates of aleatoric and epistemic uncertainty were often very close to 0. I make a detailed analysis of failure cases in Section 5.6.1.

Metric	Measure	Aleatoric	Volume	Age
r	Epistemic	0.695	0.167	0.200
	Age	0.435	0.447	
	Volume	0.431		
p-value	Epistemic	4×10^{-102}	9×10^{-6}	1×10^{-7}
	Age	9×10^{-34}	1×10^{-35}	
	Volume	4×10^{-33}		

Table 5.3: Correlation coefficients between different factors in each segmentation.

As is clear from the plot, there is a strong correlation between the estimates of aleatoric and epistemic uncertainty ($r = 0.69$). They are also both correlated positively with segmentation volume and gestational age. I hypothesise that this is due to acquisition properties: at higher

gestational ages, the skull becomes increasingly calcified and thus a stronger ultrasound reflector, lowering the signal-to-background ratio within the cranial cavity. Furthermore, the morphology of the cerebellum changes throughout pregnancy, becoming more complicated and increasing the surface area. Since errors tend to be near the cerebellar surface, it is plausible that surface area alone is a significant driver of this correlation: this is explored below.



(a) (b)

Figure 5.12: The measured voxelwise classification accuracy in the test set shown as a function of (a) epistemic uncertainty, and (b) aleatoric uncertainty. The dashed line shows the theoretical perfect calibration.

Figure 5.12 shows how the voxelwise accuracy depends on the measured uncertainty of a given voxel. Across both measures, the voxelwise classification accuracy is extremely high when there is no measured uncertainty, and drops consistently with increasing uncertainty. The measure of epistemic uncertainty used seems generally to underestimate the likelihood of a classification error, while the measure of aleatoric uncertainty seems to overestimate its likelihood. One possible explanation is that this is driven in part by the additional interpolation errors that appear in this measure of aleatoric uncertainty (see Section 5.4.3.1), which artificially increases the measured variation.

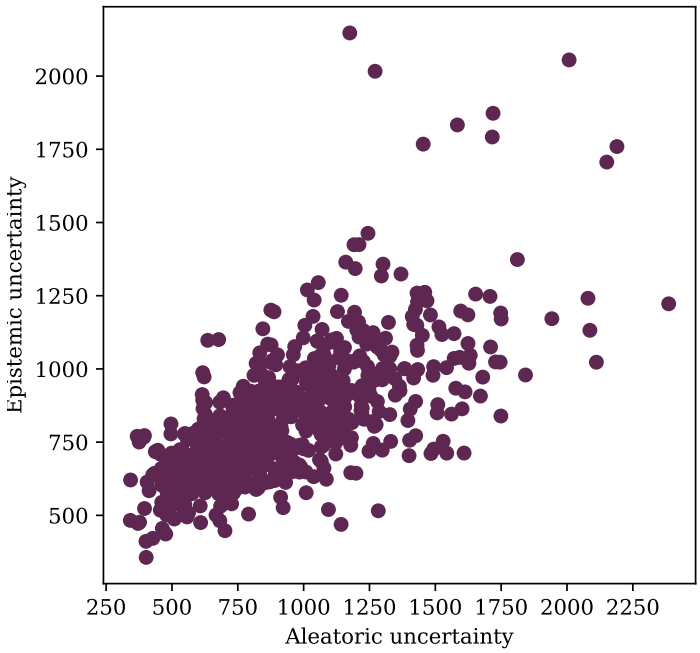


Figure 5.13: Correlation between aleatoric and epistemic uncertainty after accounting for interpolation errors.

5.5.1.2 Accounting for interpolation errors

I showed in Section 5.4.3.1 that variance is introduced in test-time augmentation due to interpolation artifacts, and I quantify this as roughly equal to 0.3 times the surface area[9]. It is difficult to directly estimate what the aleatoric uncertainty would be without this bias, but simply subtracting that estimate leads to estimates of aleatoric uncertainty which are significantly more weakly correlated than in Table 5.3. This can be seen in Table 5.4.

Note that the correlation between aleatoric uncertainty and all other measures tracked here drops. Most notable is the drop in the correlation between aleatoric uncertainty and cerebellar volume, which becomes statistically insignificant. The correlation of uncertainty with age and between aleatoric and epistemic uncertainty measures also drops substantially, though it remains highly significant. This means that these correlations are driven by something beyond just the surface area of the cerebellum: fetuses at higher gestational ages have greater skull calcification and lower signal strength within the brain, for example, which can lead to higher aleatoric uncertainty independent of surface area.

[9]The surface area of a segmentation can simply be estimated by counting the voxels on the surface, as measured using a simple XOR, $\text{seg} \oplus \text{erode}(\text{seg})$.

Metric	Measure	Raw	Decorrelated
r	Epistemic	0.695	0.677
	Age	0.435	0.373
	Volume	0.431	-0.011
p-value	Epistemic	4×10^{-102}	9×10^{-95}
	Age	9×10^{-34}	2×10^{-24}
	Volume	4×10^{-33}	0.769

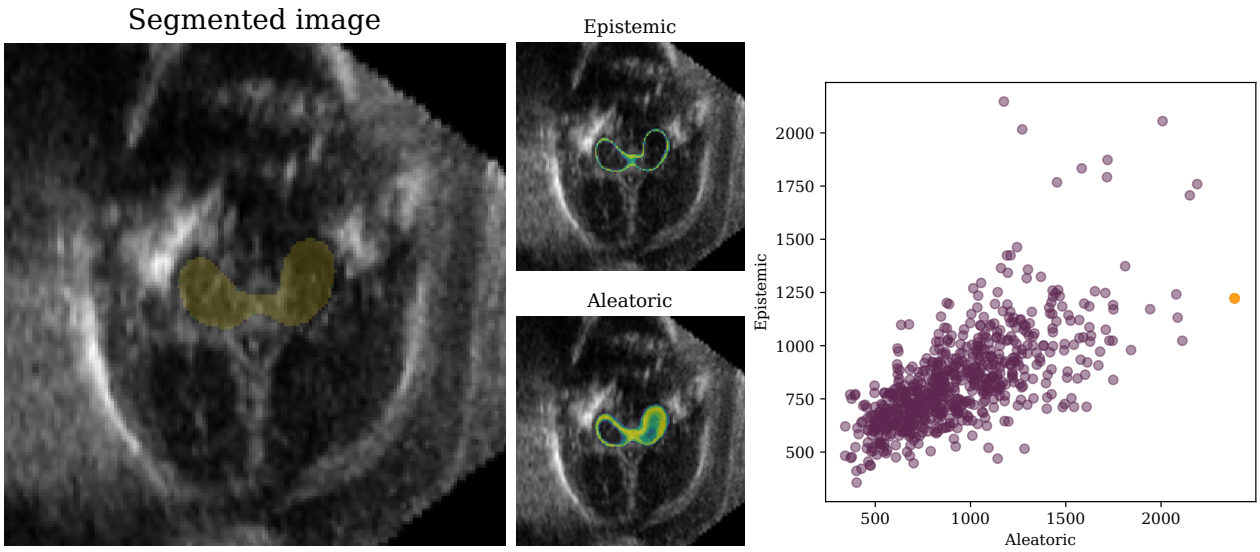
Table 5.4: The effect on correlations with other measures of subtracting a fraction of the surface area from the estimates of aleatoric uncertainty.

5.5.1.3 Analysis of individual cases

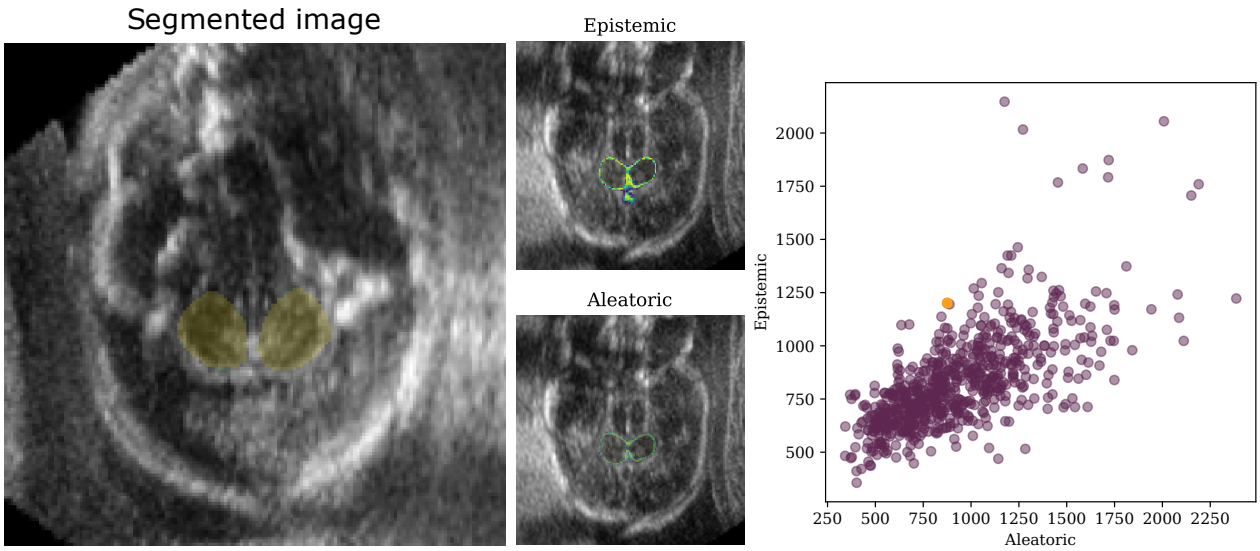
This subsection examines three volumes from the unlabelled dataset as examples of the output of the uncertainty measures explored here. For each volume, one slice is shown (part (a) of each figure), the regions of higher aleatoric and epistemic uncertainty (part (b)) and the global uncertainty measures of that volume in the context of the entire unlabelled set (part (c)).

Figure 5.14a shows a volume with high aleatoric uncertainty and low epistemic uncertainty. The segmentation output by the network is very asymmetrical across the MSP, but it is consistent across test-time dropout runs. Test-time augmentation displays significantly more uncertainty about one of the segmented lobes, corresponding to the proximal hemisphere. This lobe is partially obscured by a shadow, and it would be difficult even for a human to confidently identify a boundary.

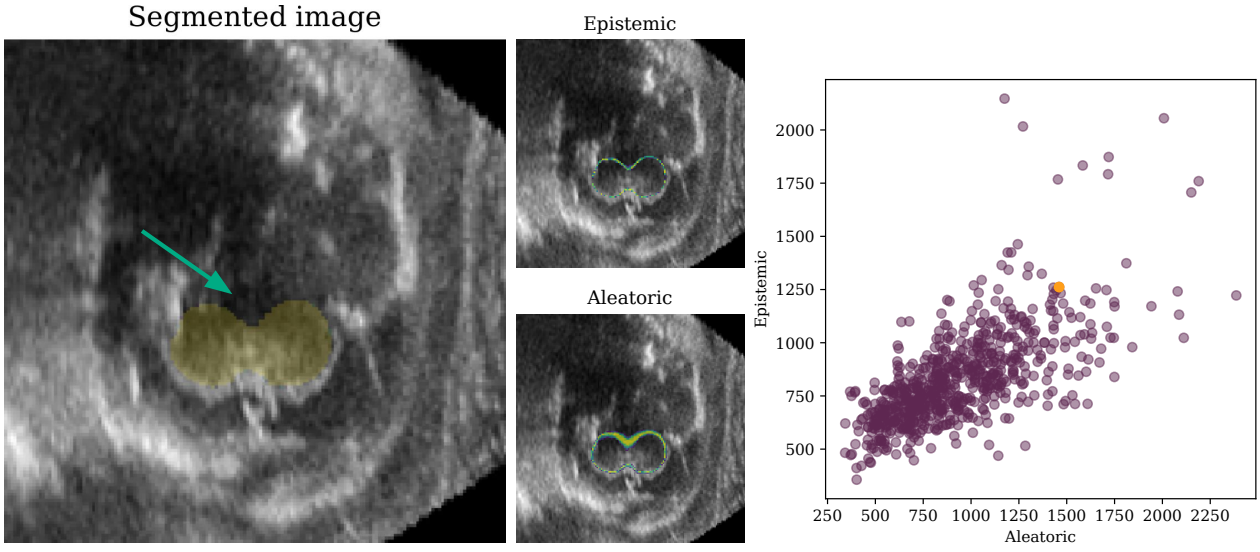
Figure 5.14b shows a volume with high epistemic uncertainty and low aleatoric uncertainty. The segmentation seems anatomically plausible in this plane, though with test-time dropout the network shows considerable uncertainty about the segmentation of part of the vermis. The vermis is often not visualised, especially at lower gestational ages, and it is possible there are few examples of the vermis in the training data. This supports the characterisation of the difference between aleatoric and epistemic uncertainty: the image is not ambiguous in that region, and any inconsistency in the segmentation is just due to the lack of similar training



(a) A volume with higher aleatoric and relatively lower epistemic uncertainty (21 weeks).



(b) A volume with higher epistemic and relatively lower aleatoric uncertainty (22 weeks).



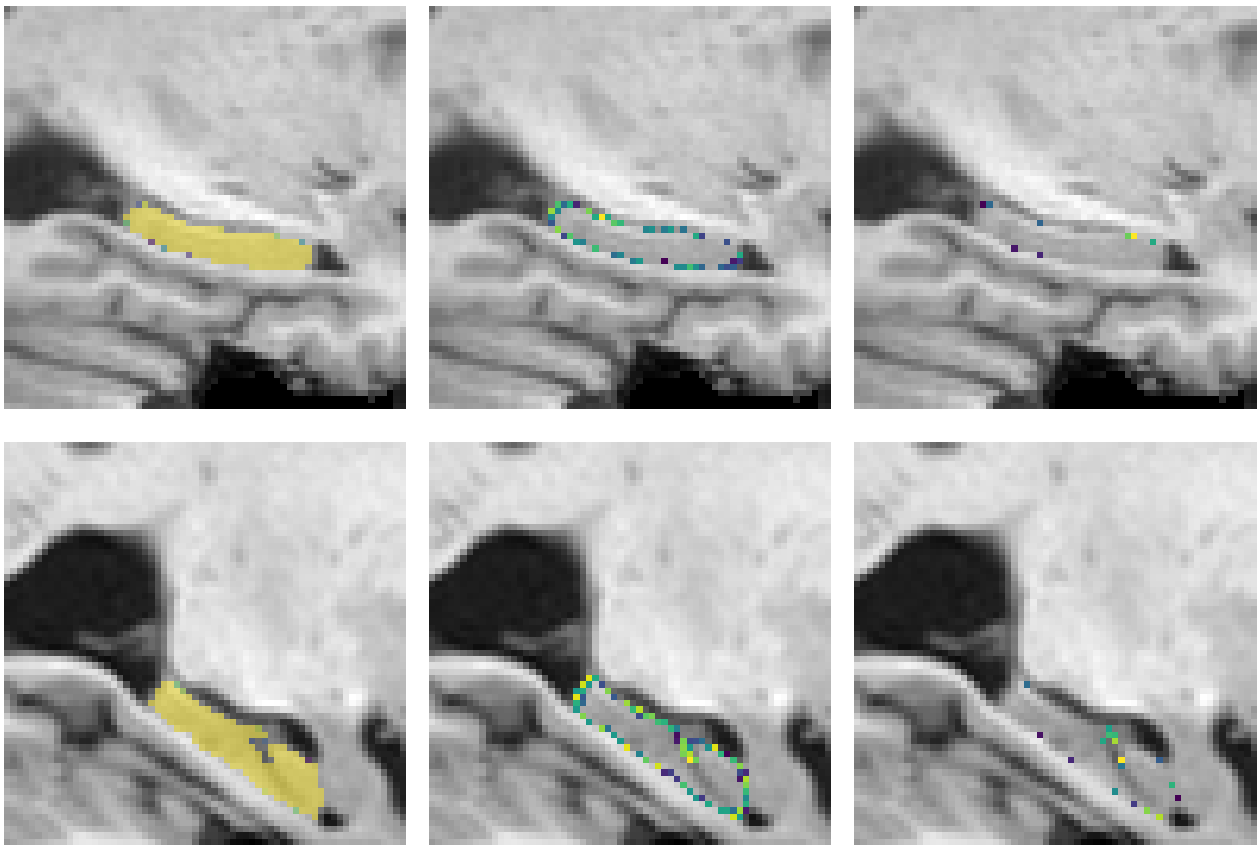
(c) A volume with high epistemic and aleatoric uncertainty (23 weeks).

Figure 5.14: These volumes are drawn from the unlabelled set, so there is no "ground truth" segmentation.

examples.

Figure 5.14c shows a volume with high levels of uncertainty in both metrics. This volume is notable for the very strong shadow artifact obscuring the boundaries on one side of the cerebellum and making it a challenge to place those boundaries accurately. Test-time augmentation does show more uncertainty than test-time dropout here, showing uncertainty present in the image itself. Test-time dropout also leads to high estimates of epistemic uncertainty: I believe this is because any volumes with significant shadow artifacts were excluded from the training set, making this volume appear different from the training set.

5.5.2 MRI dataset



(a) (b) (c)

Figure 5.15: Example segmentations (a) with aleatoric (b) and epistemic (c) uncertainty for two unlabelled examples in the MRI dataset.

The MRI dataset shows generally lower aleatoric uncertainty, with lower variability. This is probably due to more well-defined edges in MRI and lower variability in image quality

which make structural boundaries more ambiguous. MRI does not include speckle or shadow artifacts which are significant contributors to aleatoric uncertainty. The correlation between aleatoric and epistemic uncertainty is also somewhat smaller ($r = 0.69$). Figure 5.15 shows some examples of the uncertainty measured in the MRI dataset. It does seem clear that the uncertainty is largely at the boundary, but it appears generally smaller.

5.5.3 Selection of volumes for iteration

	MRI		Ultrasound	
Volume selection	Dice coeff.	Hausdorff (mm)	Dice coeff.	Hausdorff (mm)
Fully supervised	0.848 ± 0.015	3.97 ± 0.25	0.673 ± 0.046	12.25 ± 3.57
Random	0.849 ± 0.015	3.73 ± 0.18	0.700 ± 0.023	11.44 ± 1.75
Epistemic	0.847 ± 0.015	4.15 ± 0.32	0.689 ± 0.040	10.01 ± 1.48
Aleatoric	$\mathbf{0.851 \pm 0.015}$	$\mathbf{3.69 \pm 0.25}$	$\mathbf{0.727 \pm 0.014}$	$\mathbf{8.54 \pm 1.60}$
Aleat. + epist.	0.848 ± 0.015	4.20 ± 0.35	0.697 ± 0.015	11.10 ± 1.51
IOV	N/A	N/A	0.764 ± 0.060	3.96 ± 0.99

Table 5.5: The segmentation performance of retraining a 3D CNN using different selection methods for the additional labelled data. The "manual segmentation" row indicates the consistency of manual segmentations obtained by the same individual.

150 volumes from the unlabelled set were selected, automatically segmented, and added to the 106 labelled volumes to act as training data. Section 5.4.5 describes different ways by which these volumes could be selected: randomly, for the lowest aleatoric uncertainty or for the lowest epistemic uncertainty. Table 5.5 shows the performance of the different training sets, in terms of Dice coefficient and Hausdorff distance. The consistency of segmentations by the same individual is also shown as a form of upper bound: no network can learn to segment better than the consistency of its training segmentations.

Random volume selection leads to an improved Dice coefficient ($p < 0.01$ in a 1-tailed t-test) in the validation set over using the original labelled set only, but does not lead to a significant improvement in Hausdorff distance. Selecting the volumes with the lowest epistemic

uncertainty leads to no significant improvement ($p = 0.82$) in Dice coefficient, and actually seems to lead to worse performance in Hausdorff distance. Selecting volumes based on the lowest aleatoric uncertainty, on the other hand, led to significant improvements in both Dice coefficient and Hausdorff distance for the segmentations.

5.5.4 Initial number of labels

One possible explanation for the small effect size in the MRI dataset is that the labelled data alone is sufficient to generalize to the test set, and the network is already near the maximum achievable performance on this dataset. Figure 5.10a shows that reducing the number of labelled training examples to 10 or 20 (from the original 180) seems to have only a limited impact on segmentation performance. With fewer initial labelled examples, the performance boost from omni-supervised learning is clearer.

The same is not true in the ultrasound dataset, where segmentation performance increases steadily with increasing dataset size, and the effect of an omni-supervised scheme increasing the amount of data available is therefore larger. I conclude that with less initial training data, the effect size of the schemes discussed in this chapter is greater, and that segmentation performance on the MRI dataset is already near its ceiling for this network structure.

5.6 Discussion

The results shown in Table 5.5 can be explained in terms of the difference in the type of uncertainty measured by test-time dropout and test-time augmentation.

Selecting the volumes with the lowest epistemic uncertainty, as measured by test-time dropout, seems to lead to worse performance over random selection, in both the ultrasound and the MRI dataset. Epistemic uncertainty is believed to be a measure of the model's extrapolation over new data (see Section 2.4.1): the overall epistemic uncertainty present in a volume can be usefully thought of as a measure of how different the volume is from the training data that the model has seen. The volumes with the lowest epistemic uncertainty

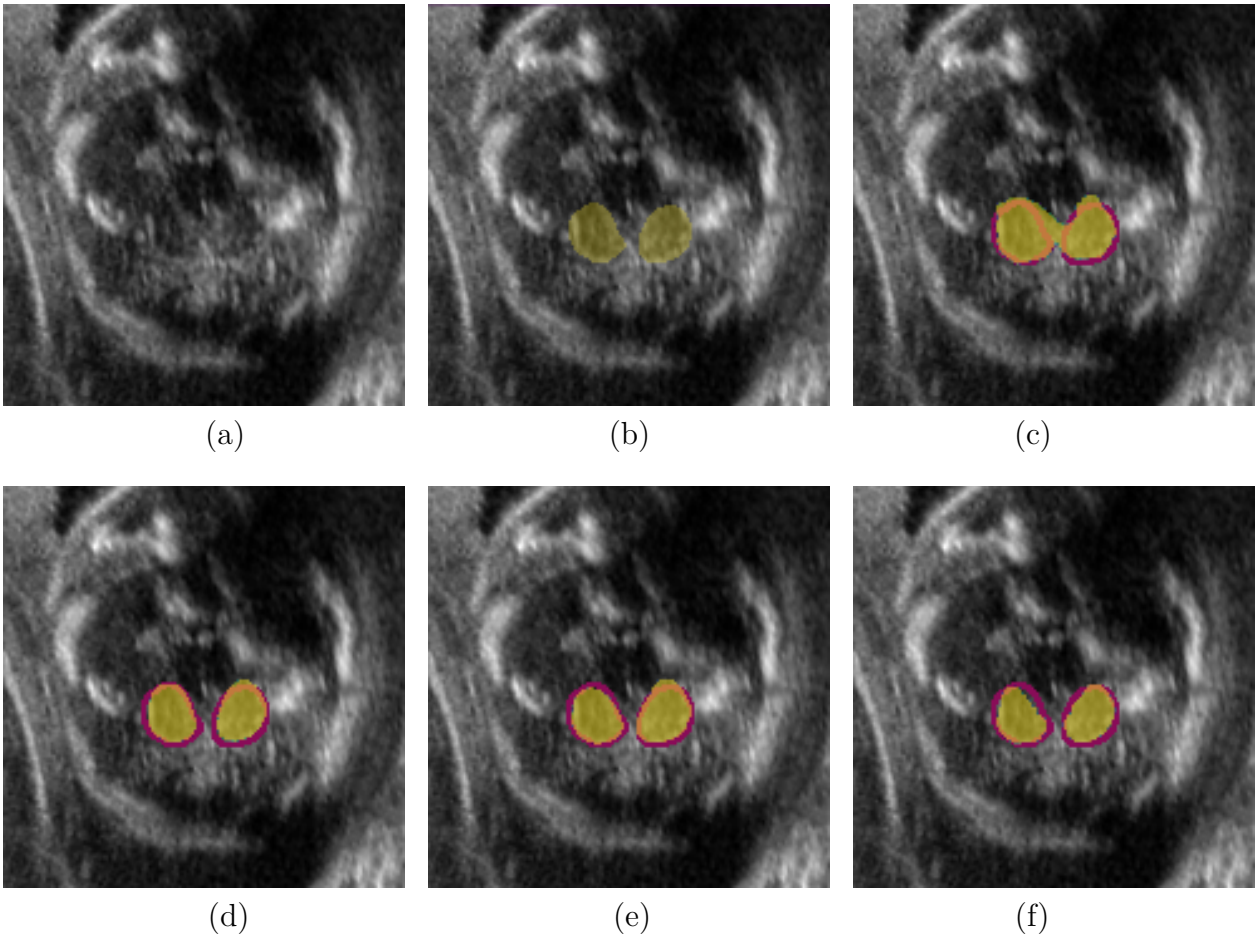


Figure 5.16: An example of the effect of data selection for omni-supervised learning. (a) A slice of a volume in the validation dataset (at 21 weeks). (b) The ground truth segmentation of the cerebellum, (c) the segmentation obtained after training a CNN on labelled data only. Bottom row: the segmentations obtained by re-training the network in an omni-supervised fashion selecting additional volumes (d) randomly, (e) by aleatoric uncertainty, and (f) by epistemic uncertainty.

can therefore be thought of as being the volumes that most closely resemble the training data. Adding these volumes to the training data, therefore, does not improve the generalisation of the network when it is presented with new data. Instead, it introduces a bias in the training data towards the segmentations it has generated in the previous round. This explains the lack of improvement in Dice coefficient and the worsening in Hausdorff distance (11.37mm vs 9.05mm in the ultrasound dataset). The effect size is larger in the ultrasound dataset than in the MRI dataset: I believe this is due to the larger variability of the ultrasound dataset, as evidenced by the larger range (and generally higher level) of aleatoric uncertainty estimates in that dataset.

On the other hand, using aleatoric uncertainty leads to a significant improvement in

performance over a random selection. Aleatoric uncertainty is used as an estimate of the amount of uncertainty inherent in the data itself, such as noise, artifacts, or inconsistent labelling in the training data. In principle, it is unrelated to how different the data appears from training and only a measure of the quality of the data itself. Lower aleatoric uncertainty, therefore, can be thought of as correlating to clearer data and possibly better segmentation performance, without compromising the generalisability of the segmentation method. The effect size is larger in the ultrasound dataset than in the MRI dataset, but it holds in both datasets. Figure 5.16 shows an example of the effects of this method, by significantly improving the segmentation of a sample slice.

It is worth noting that random selection does still lead to an improvement over using no selection at all: this is reassuring evidence that omni-supervised learning methods work. Using aleatoric uncertainty-based selection still seems to roughly double the performance gains of omni-supervised learning relative to random selection in ultrasound (an improvement in Dice coefficient of 0.08 relative to fully-supervised segmentation only, compared to 0.04 for omni-supervised segmentation).

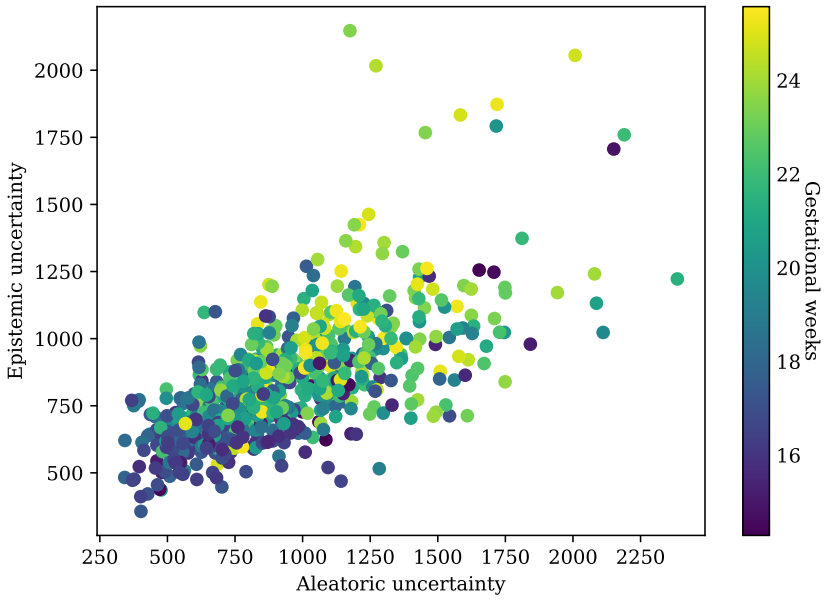
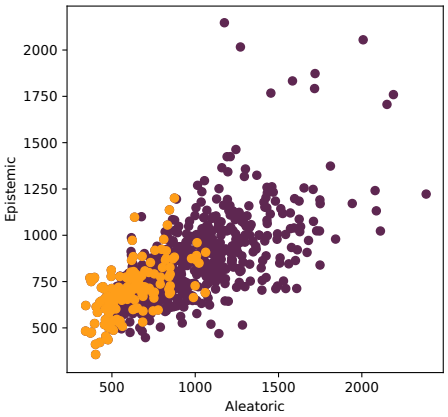


Figure 5.17: A scatter plot of aleatoric and epistemic uncertainties in the data, coloured by gestational age.

I still note a strong correlation between measured aleatoric and epistemic uncertainty, and other measures. This can potentially lead to biases in data selection and could prevent

further improvements in performance. For example, we note a strong correlation between gestational age and both aleatoric ($p < 10^{-33}$) and epistemic ($p < 10^{-24}$) uncertainty measures. A straightforward selection of the volumes with the lowest aleatoric uncertainty is therefore likely to be biased in favour of lower gestational ages (see Figure 5.17). One can attempt to control for the age distribution of the volumes we select, and only base the selection on the uncertainty measures obtained within each gestational week.



(a)

Volume selection	Dice coeff.	Hausdorff (mm)
Aleatoric	0.695 ± 0.111	8.18 ± 5.39
Aleatoric (age-adj.)	0.697 ± 0.089	8.21 ± 5.28

(b)

Figure 5.18: (a) the data selected for omni-supervised learning after controlling aleatoric uncertainty selection for age. (b) the performance of training with this data selection compared to simple aleatoric selection.

Performing such a selection leads to the results in Fig. 5.18. There is no statistically significant change in segmentation quality when controlling for age rather than picking only the volumes with the lowest aleatoric uncertainty.

5.6.1 Analysis of failure cases

After the first round of segmentation, 25 volumes in ultrasound (3.2% of the dataset) were entirely segmented as background, with no voxels classified as cerebellum at all. The unlabelled data these segmentations were generated from had no ground-truth segmentations to allow direct evaluations of performance, but all the fetuses recruited for the INTERGROWTH-21st study have a cerebellum, making these clear examples of failure cases.

Figure 5.19 shows the gestational ages of those failure cases. It appears to be a bimodal

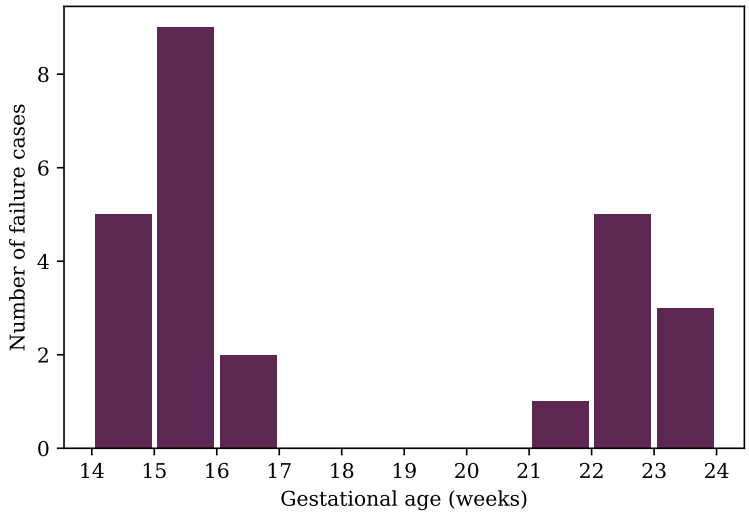
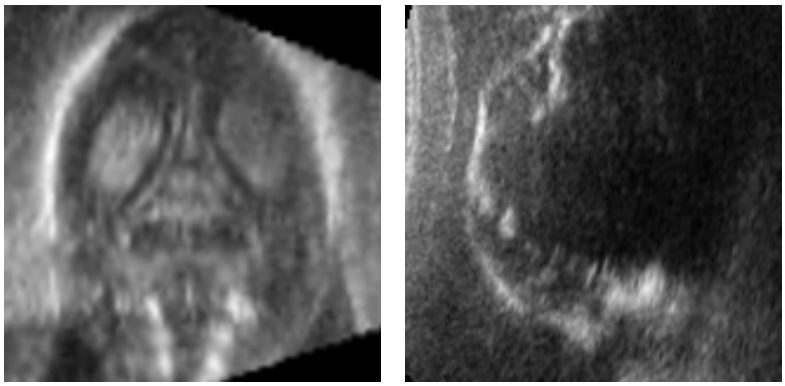


Figure 5.19: The gestational age of volumes in the unlabelled dataset which were entirely segmented as background.

distribution: 16 of the 25 cases (64%) had a gestational between 14 and 16 weeks, while the remaining 9 (36%) had a gestational age of 22 weeks or above. None of the 343 volumes in the range 17-21 weeks were entirely segmented as background.



(a) (b)

Figure 5.20: (a) An example of a failure case at a gestational age of 17 weeks. (b) An example of a failure case at a gestational age of 24 weeks.

Figure 5.20 shows examples of failure cases at either end of the age range. It seems clear that there are two different failure modes for this network. At higher gestational ages, shadows are cast from the petrous ridges of the temporal bone and obscure the cerebellum. In these circumstances, it is impossible to make any segmentation with any confidence. I visualised all failure cases and all the cases at higher gestational ages displayed dark shadow artifacts

obscuring the cerebellum. This is a problem that is more common at higher gestational ages, as the skull ossifies and becomes a stronger ultrasound reflector.

At lower gestational ages, the failure mode is less obvious. The contrast appears low, and while subjectively the cerebellum can be visualised, it is difficult to draw clear boundaries. One may imagine that the reduced size of the cerebellum leads to greater uncertainty in the predictions, leading to cases where dropped-out predictions fail to agree. This is not borne out by the uncertainty estimates for the data: with test-time dropout and test-time augmentation, these volumes are consistently segmented as background.

This analysis seems to justify the age limits selected for this task: at lower gestational ages, it is difficult to visualise the cerebellum, while at higher gestational ages, shadows are more likely to obscure the structures of interest.

Segmentation performance was more consistent in the MRI dataset: the imaging modality and structures visualised are seen regularly in all patches. Therefore, there were no total failure cases in this modality.

5.7 Conclusion

This chapter has proposed a new data-selection method to improve the performance of omni-supervised learning on segmentation of the fetal cerebellum, with results validated on an external MRI dataset. The basic problem I faced was that of little available labelled data: even though the INTERGROWTH-21st dataset contains many fetal ultrasound acquisitions, no reliable segmentation labels are provided with it. I manually segmented a subset of this dataset to provide a starting point, then used omni-supervised learning to generate high-quality automated segmentations. Working with uncertainty estimates to quantify the reliability of segmentations led to the intuition of using these measures to help train omni-supervised methods, which is implemented in the data-selection method outlined in this chapter. I found that selecting data for further omni-supervised training based on aleatoric uncertainty improves the performance of my CNN on this dataset.

These results were replicated on an external, public MRI dataset. A fraction of the

MRI dataset had been manually annotated by experts, and I found similar improvements using omni-supervised learning to boost performance segmentation. Although the effect size was smaller, I found similar results: selection based on aleatoric uncertainty resulted in significantly improved segmentation performance, while epistemic or random selection led to smaller effects.

I then use this data-selection method to train an omni-supervised method to generate segmentations of the cerebellum throughout the whole dataset.

Chapter 6

Adding information from scalar features at network input

Chapter layout

This chapter explores methods to combine image and scalar data at the input of a CNN, providing the network with the same information that a human annotator has access to. Human annotators look at and are influenced by metadata as well as image data, and the segmentations they produce are influenced by the metadata available. CNNs, on the other hand, are usually limited to the image data itself, stripped of relevant metadata. This chapter measures the extent of that information disparity and investigates one method by which information parity between human and machine can be reached. I propose TSI-Net, a network architecture that can combine transformed scalar inputs with image inputs in a segmentation framework.

Section 6.1 introduces and justifies the addition of scalar features as a plausible way to improve segmentation quality. Section 6.2 looks at specific features that are available in the INTERGROWTH-21st and ADNI/HARP datasets and sets out the approach used to input them to a CNN in this chapter. Section 6.30.0 considers various design decisions, such as ways to encode the scalar values described in Section 6.2, and how to modify a CNN architecture to include them in Section 6.3.3. This section proposes experiments to evaluate these design

decisions. The results of these experiments are presented in Section 6.4, where the TSI-Net architecture is finalised. Section 6.5 discusses what these results imply for the usefulness of this method and considers failure cases. Finally, Section 6.6 summarises the outcomes of this chapter and considers their implications.

6.1 Introduction

While obtaining segmentation labels for medical image data is difficult, there are several non-image annotations that are collected routinely in the clinic. For instance, gestational age (measured as days elapsed from the last menstrual period) is often known to sonographers during a prenatal ultrasound scan and is recorded. For clinical imaging, the patient's demographics and clinical diagnosis are also generally recorded and kept with the imaging data.

Human annotators are influenced by these scalar values, as described in Chapter 2. Sonographers told a fetus' gestational age produce clinical measures that are closer to the expected value for that age [129], and cognitive anchoring effects (e.g. when a diagnosis is already present) are well-known across medical image analysis [130]. Inter-rater agreement is also often measured when human annotators have access to this additional non-image data: this is used as a gold standard of segmentation performance, but it may not even be reached by humans who do not have this information available.

CNNs typically are trained on image data alone, without the metadata which humans have access to. This additional information is not fully contained within the image data. It is possible to extract measures such as the transcerebellar diameter (TCD) from the image data, which is used as a proxy for gestational age, and is used by sonographers to estimate GA when more reliable measures are unavailable [53]. However, this estimate still is prone to some error due to individual variation. Rodriguez-Sibaja et al [4] manually measured the TCD of a large subset of the INTERGROWTH-21st dataset, and fit a polynomial function estimate age from TCD. However, this polynomial estimate still has an error of over 7 days for about 10% of volumes in the INTERGROWTH-21st dataset, making it quite noisy (shown

in Figure 6.3). Section 6.4.1 measures how well different methods can estimate these features in the dataset.

This disparity can lower the potential performance of these methods, meaning that human annotation performance cannot be treated as an upper bound on segmentation performance.

6.2 Proposed approach

This chapter proposes TSI-Net, a method to include Transformed Scalar Inputs to a fully-convolutional CNN and integrate image and non-image data. In this chapter, the aim of this is to aid segmentation of the cerebellum in the INTERGROWTH-21$^{\text{st}}$ dataset, to give machines the same information that humans use to perform this labelling.

Segmentation is generally performed with a fully-convolutional network architecture such as the U-Net [94], a variant of which is used throughout this thesis. This is necessary to maintain spatial consistency in the features being processed, but it also restricts all inputs at any layer of the network to grid-like features that can be processed by a convolutional kernel. Scalar features cannot be processed by convolutional kernels: a 3D convolutional kernel with size 3^3 requires a minimum of 9 samples, arranged in a $3 \times 3 \times 3$ grid, to produce a valid output. Therefore, to add non-image features to the input of a fully-convolutional CNN, they must first be transformed to image-like data.

This chapter considers different ways to transform non-image features to image data, and add this as an additional input to a CNN. It proposes TSI-Net as a method to perform this transformation with only minor modifications to a traditional CNN architecture. After numerical values have been transformed, this information can be simply concatenated to the image input at the same layer. This is not the only possible choice: as a separate input, it can be introduced at a different layer of the network, or even be processed through a different pathway at first. This chapter performs a set of experiments to determine the best way to modify a CNN architecture to accommodate this additional input, balancing segmentation performance, training time, and memory usage.

The experiments conducted are:

- Considering and evaluating different function to encode scalar features into a CNN input and their correct parameters
- Finding the appropriate layer at which to modify the U-Net architecture to include additional features
- Measuring the effect of combining multiple scalar features into a single input
- Adding noise to scalar values at training and test time

The methods described in this chapter are then tested on segmentations of anatomical structures from the INTERGROWTH-21st dataset and the ADNI/HARP dataset, with the addition of known non-image features from each of those datasets. For comparability, the segmentation tasks considered are the same as in Chapter 5: segmentation of the cerebellum in the ultrasound dataset, and segmentation of the hippocampus in the MRI dataset.

6.2.1 Scalar features known

The subset of the INTERGROWTH-21st dataset examined in this chapter is the same as that considered in Chapter 5: it only includes 3D brain volumes taken according to a strict protocol, within a gestational age range of 14 to 25 weeks. 14 weeks is the lower bound of the age range for which images were collected during the study, while above 25 weeks shadows from the temporal bone are frequently present and obscure the cerebellum.

The main scalar feature relevant to this segmentation task is gestational age. As part of the selection protocol for data collection in this study, all subjects needed a precisely-known gestational age, defined as the number of days elapsed from the start of the mother's last menstrual period (LMP) [134]. The LMP predates a typical conception by approximately two weeks [23], and is the clinical standard for measuring gestational age outside of ultrasound-based measurements [163]. LMP itself is not perfectly accurate: women's estimates of LMP date exhibits digit preference [164] and inaccurate recall, suggesting some random error in the estimation of the date. This error has not been quantified rigorously.

The ADNI/HARP MRI dataset comes with a set of image and non-image features. As described in Chapter 3, the dataset was collected to study Alzheimer's disease, which has a particular impact on the hippocampus. The subjects in this dataset are older adults, of whom some are cognitively normal (CN), while others suffer from mild cognitive impairment (MCI) or probable Alzheimer's disease (AD). In this chapter, three features are considered as additional inputs: the age of the test subject (with a range of 60-90 years), the sex (M/F), and their cognitive status (CN/MCI/AD). These are distinct, but not fully independent from each other: older subjects are more likely to develop AD [165], and women have a higher prevalence of AD than men at the same age [165]. Nonetheless, each of these features does provide additional information which is not fully contained in the others.

6.3 Methods

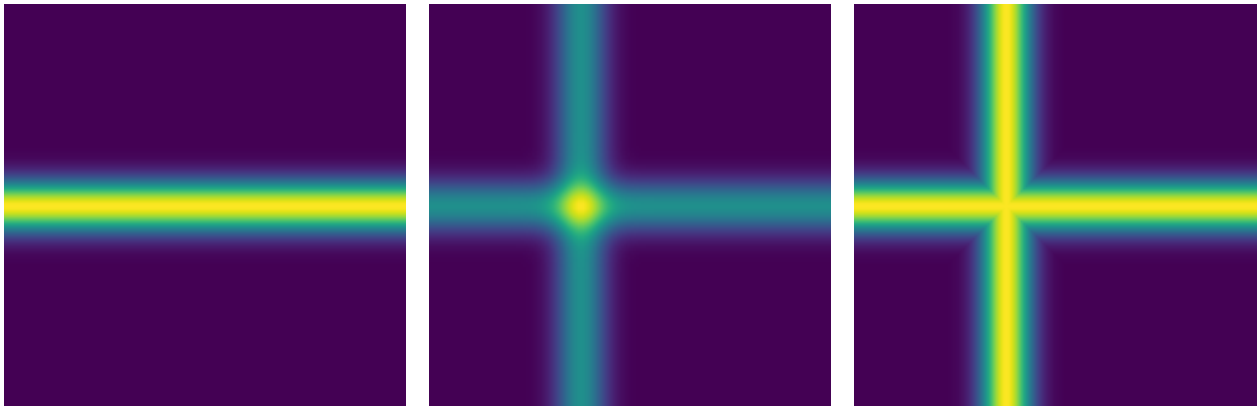
The methods proposed here are applied to the same segmentation tasks performed in Chapter 5. This allows direct comparison with the results compared in that chapter. The same training and test datasets are used (with no need to use the additional unlabeled data included in that chapter).

Both tasks attempted in Chapter 5 are performed in this chapter: segmentation of the fetal cerebellum from 3D ultrasound in the INTERGROWTH-21st dataset, and hippocampal segmentation task in the ADNI/HARP MRI dataset.

6.3.1 Encoding scalar values in a matrix

First, scalar parameters need to be transformed to image-like features. One way to do this is to create synthetic images, whose contents depend on the value of the scalars encoded within them. These images need to match the dimensionality of the image input: a 3D medical acquisition processed by 3D convolutional layers must be accompanied by a 3D synthetic image, to be processed by the same filters.

We propose TSI-Net, a method to use **Transformed Scalar Inputs** in a fully-convolutional



(a) (b) (c)

Figure 6.1: (a) A slice of a 3D volume, with a single scalar value encoded within. (b-c) A slice of a 3D volume with two different scalar values encoded by (b) summation, (c) maximum function.

network architecture. TSI-Net is based on the transformation of scalar features into a grid-like shape, which are then passed as a separate input to the network. This allows them to be processed by a conventional network architecture, with only minor changes to the network design.

In this chapter, this data is encoded as a line in a 3D matrix, creating an image-like output which can take any dimensions that are desired. This is shown in Figure 6.1(a). The position of the line encodes the value of the scalar feature of interest. This position is proportional to the value of the feature, and to use the space available in the image most effectively, the range of positions is normalised: the lowest value in the dataset corresponds to a line near one end of the image, and the highest value corresponds to a line near the opposite end of the image.

Since a volume is three-dimensional, it is possible to include up to three one-dimensional scalar values in the same volume in this scheme, one encoded along each axis. Figure 6.1(b-c) shows a slice of a 3D volume with two scalar values encoded, one along each visible axis. The simplest way to generate such an image C_t is to create images encoding a single value, and then combine them into a single image. Three methods are explored to do so. Starting from a set of image channels $C_1, C_2 ... C_n$, these can be expressed as

$$C_t = \sum_{i=1}^{n} C_i \tag{6.1}$$

$$C_t = \max\left(C_1, C_2 ... C_n\right) \tag{6.2}$$

Method 6.1 is simpler to implement and evaluates more quickly during training (mean time of 28.9ms using two exponentials, compared to a mean time of 64.3ms for method 6.2), but visually appears to emphasise the intersection between the two planes (which is arbitrary, as they represent distinct values). Method 6.2 does not emphasise the intersection between planes, but evaluates more slowly and therefore slows down training. These methods can also be compared to a simpler, "null" method: passing each feature separately, as a different map, and simply concatenating them at the input level. This bypasses the need for any of these methods, but can result in a larger input, larger network size, and longer training time. These methods are all compared by experiment.

6.3.2 Choice of function

The choice of function by which scalar features are transformed into a grid is an important design parameter of TSI-Net. While it is possible to encode the scalar of interest into a single straight line 1 voxel wide[1], this only uses a very small fraction of the visual space in the image, potentially leading to slower training and a higher propensity to overfit.

It may be better to encode this data in a way that uses a higher proportion of the image, by using a function that peaks at the correct scalar value and decreases monotonically away from it. This ensures a higher proportion of the network's convolutional layers are activated for a given value of the scalar feature, improving robustness.

Four functions were considered for this task. Their formulae are described in Table 6.1, in terms of the position x, scalar feature value s, and a tunable width parameter σ (see below). The first function considered is a simple rectangular function, taking value 1 within the chosen width σ from the parameter value and 0 elsewhere. This is the simplest choice, but it also

[1]This is equivalent to encoding it using a rectangular function of width 1.

Function	Formula	Half-width half-max	Time per vol. (ms)
Rectangular	$y = \begin{cases} 1 & \lvert x-s\rvert \leq \sigma \\ 0 & \text{otherwise} \end{cases}$	σ	6.1
Triangular	$y = \begin{cases} 1 - \frac{\lvert x-s\rvert}{\sigma} & \lvert x-s\rvert \leq \sigma \\ 0 & \text{otherwise} \end{cases}$	$\frac{\sigma}{2}$	11.9
Quadratic	$y = \begin{cases} 1 - \left(\frac{\lvert x-s\rvert}{\sigma}\right)^2 & \lvert x-s\rvert \leq \sigma \\ 0 & \text{otherwise} \end{cases}$	$\frac{\sigma}{\sqrt{2}}$	11.0
Gaussian	$y = e^{\frac{-(x-s)^2}{\sigma}}$	$\sigma \ln 2$	12.0

Table 6.1: The formulae of the functions considered to transform scalar data to a grid-like form, where y is their value as a function of voxel value x along one axis. The table also shows their half-width half-maximum (in terms of a parameter σ) and the time taken to generate a 160^3 map using each.

makes it difficult for a convolutional network to find the precise value of the scalar feature, as the function is flat around the peak parameter value s. The second choice considered is a triangular function, which maintains the simplicity but has a well-defined peak. This function decreases linearly from 1 at the peak to 0 at a chosen width σ from the peak. The third function (a quadratic) is similar, but exhibits a different descent profile. This follows from the intuition that larger deviations from the peak value of the chosen parameter (such as age) should lead to supralinearly larger changes in visual features than smaller changes, and therefore the function of choice should have a broader, flatter peak and a faster decline. The final function chosen is a Gaussian, which has a similar shape to the quadratic function but is differentiable everywhere and isn't set to 0 outside of a specified width parameter.

Figure 6.2 shows the 1D shape of the functions considered, and a 2D slice of the final 3D volume input into the network. To obtain the final 3D volume, the 1D shape of the function can be simply tiled along the $y-$ and $z-$ axes until a volume of the desired dimensions is obtained.

The width of these functions is described in Table 6.1, in terms of a parameter σ. The first three functions evaluate to zero outside of the interval $[s-\sigma, s+\sigma]$, while the Gaussian function only tends to zero but never reaches it (maintaining monotonicity and differentiability throughout the image space). A higher value of σ leads to the function covering a greater

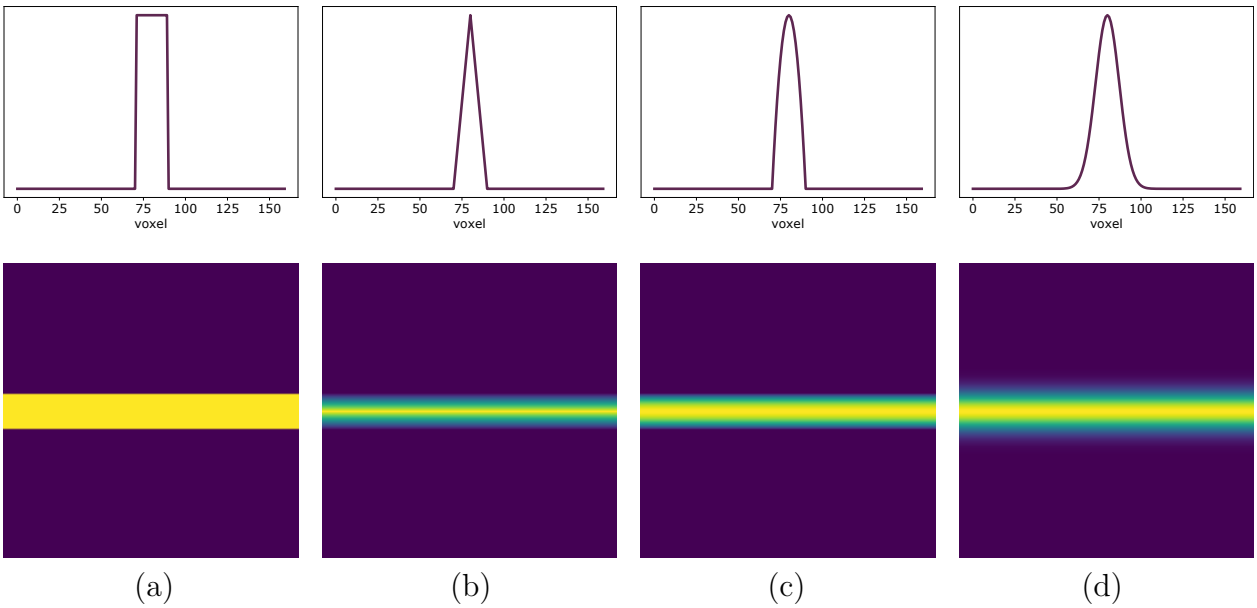


Figure 6.2: The four different functions considered. (a) rectangular function, (b) triangular function, (c) quadratic function, (d) Gaussian function. The top row shows the 1D shape of the function, and the bottom row shows a 2D slice of the 3D volume input in the network.

proportion of the available space, but flattens the function in the peak region and makes the precise parameter value less well-defined, while a lower value of σ results in more image space being set to 0 and not being used for training or inference. Therefore, an experiment is conducted in this chapter to find the optimal value of σ for the tasks described here.

Table 6.1 also shows the time taken to generate volumes of dimensions 160^3 using these functions. As expected, the rectangular function is the simplest one and evaluates most quickly, in only 6ms. The other functions take approximately twice as long to evaluate, but there is no significant difference in evaluation time between them.

6.3.3 Network design

The network architecture used for TSI-Net is the same 3D implementation of the U-Net architecture used in previous chapters. The network's hyperparameters are also kept the same as in previous chapters: the purpose of this chapter is to experiment with varying inputs to the network, not the architecture itself. The network architecture chosen is shown in Figure 6.3.

The network must, however, be modified to accept an additional input representing the

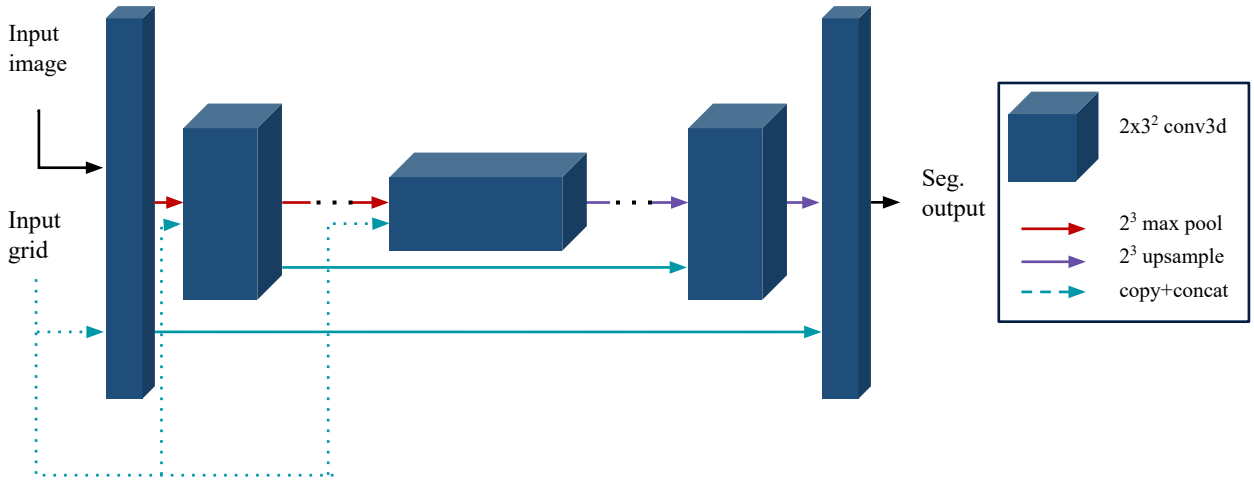


Figure 6.3: The 3D U-Net architecture used for TSI-Net. Different options for where to input the additional grid (shown as a dotted line in the figure) are considered.

additional scalar data, as well as the usual image input. There are several possible approaches to include this data.

The most straightforward option is to concatenate this second input with the image. This allows all levels of the network to use the additional information provided by the scalar feature, but it also means that the additional input needs the same dimensions as the image (160^3 in the the INTERGROWTH-21st dataset, 64^3 in the ADNI/HARP dataset). Where the image input has large dimensions, this seems wasteful and increases memory usage and time needed to generate the feature maps.

An alternative is to provide the additional input part or all of the way down the encoder pathway, as shown in Figure 6.4. Since this is after the processed image data has been max-pooled and downsampled, the additional input can also be smaller and concatenate with the image data. This also reduces the size of the overall network, as fewer convolutional kernels are needed initially in the encoder pathway. It also reduces overall training time and memory requirement, as smaller feature maps need to be made. However, this does mean that all the information provided isn't available to the network until some processing has already been conducted on the image data, which could lead to a lower performance advantage.

It would also be possible to concatenate the input at similar levels in the decoder path, instead of the encoder path, of the CNN. I did not do this in this chapter, as the benefit of

attempting this is unclear: this does not reduce input size or the memory requirements of the CNN relative to input in the encoder path, but reduces the processing this input undergoes and the information available to most of the network.

Input layer	Input size	Net. size (GB)	Input gen. time (ms)	Train time / sample (s)
1	160^3	3.56	66	0.945
2	80^3	3.54	6.1	0.886
3	40^3	3.52	0.39	0.861
separate	80^3	3.56	6.1	0.931
Baseline	N/A	3.50	0	0.852

Table 6.2: Inputting additional scalar data at different layers of the CNN has negligible effect on the overall network size, but reduces the time to generate the synthetic image and therefore the overall training time per epoch.

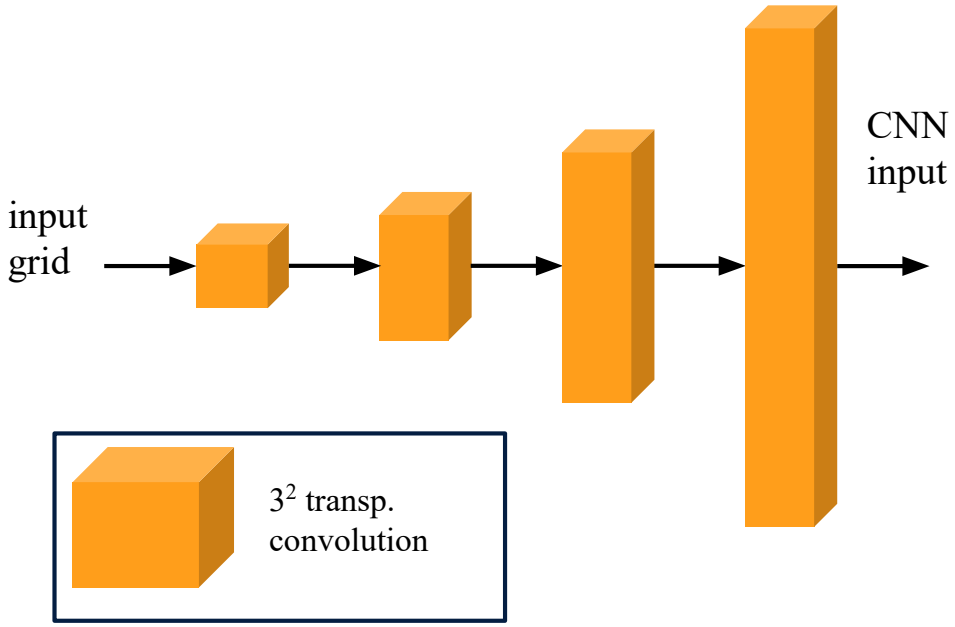


Figure 6.4: One possible way to maintain a small additional input size while processing the additional feature grid by all convolutional layers is to pass it through successive transposed-convolution layers until it reaches the same volumetric dimensions as the input.

Another option which combines smaller input size and processing by all layers in the network is shown in Figure 6.4. The scalar input is passed as a smaller volume to a dedicated deconvolution (or strided-convolution) layer, which then concatenates its output with the

image input. This allows a smaller additional input to be passed to the network encoding the scalar parameters without skipping any processing layers. However, this does not lead to a concurrent reduction in the size of the CNN: the overall number of parameters is slightly greater than an architecture which concatenates two inputs of same size.

Table 6.2 shows the difference in network size, input generation times, and actual training time per iteration of each architecture considered, for a volumetric input size of 160^3. Changing the position of the additional input has a negligible (on the order of 1%) impact on the overall size of the network, which only requires additional convolutional filters in the layer at which the volume is input. The overall size of the network is dominated by all other convolutional layers, so the choice of input location does not have a significant effect on memory requirements.

On the other hand, the time taken to generate the additional input does vary substantially based on the layer at which it is introduced, reducing by approximately an order of magnitude for each successive layer on the encoder path measured. This is significant but it represents only one factor of CNN training time: the majority of training time comes not from generating the input, but by the processing of the convolutional layers and backpropagation.

6.3.4 Quantifying feature contributions

A baseline CNN can be trained, with no additional inputs, for comparison. The lack of an additional input, however, reduces the number of convolutional layers in the network slightly (see Table 6.2). This is a potential source of bias: to address this, I also trained a second baseline CNN to compare to TSI-Net, with an additional input consisting of a uniform array of zeros. This matches the number of trainable parameters between architectures but does not provide any additional non-image information to the network.

The ADNI/HARP dataset has multiple scalar features that can be encoded separately or together in this scheme. Age, sex, and clinical diagnosis (CN/MCI/AD) can all be encoded by the scheme above, either individually or together. It is possible, therefore, to quantify the performance boost given by each of these measures individually, as well as by their

combination.

6.3.5 Training and choice of hyperparameters

The same training dataset was used as in Chapter 5. For the INTERGROWTH-21st dataset, this consists of 106 manual 3D segmentations of the fetal cerebellum with a gestational age range of $14 - 25$ GW, chosen to match the age distribution of the overall dataset. For the ADNI/HARP dataset, this consists of 200 manual 3D hippocampal segmentations, obtained from 100 volumes. The same preprocessing as in previous chapters was also applied: each volume was brain-extracted, linearly registered to MNI152 image space [20] and normalized by dividing each voxel by the 99th percentile value. $64 \times 64 \times 64$ patches were extracted around the hippocampus in each hemisphere.

Data augmentation was used during training, similar to previous chapters. The ultrasound volumes were randomly reflected across the mid-sagittal plane and subjected to small random translations (up to ± 5 voxels along each axis), small random rotations (up to $\pm 10°$ around each axis), and small amounts of scaling (up to $\pm 10\%$ linear zoom). Nearest-neighbour interpolation was used for computational efficiency. The volumes were reflected with 50% probability, while the degree of all other augmentations was sampled from a uniform distribution. Unlike in Chapter 5, augmentation was only applied during training: since aleatoric uncertainty is not measured in this chapter, there is no need for test-time augmentation.

All of the models used here were written in Python 3.7 using Tensorflow 1.14 and Keras 2.4. They were trained on Nvidia GTX 1080 Ti GPUs.

6.4 Results

6.4.1 Scalar value

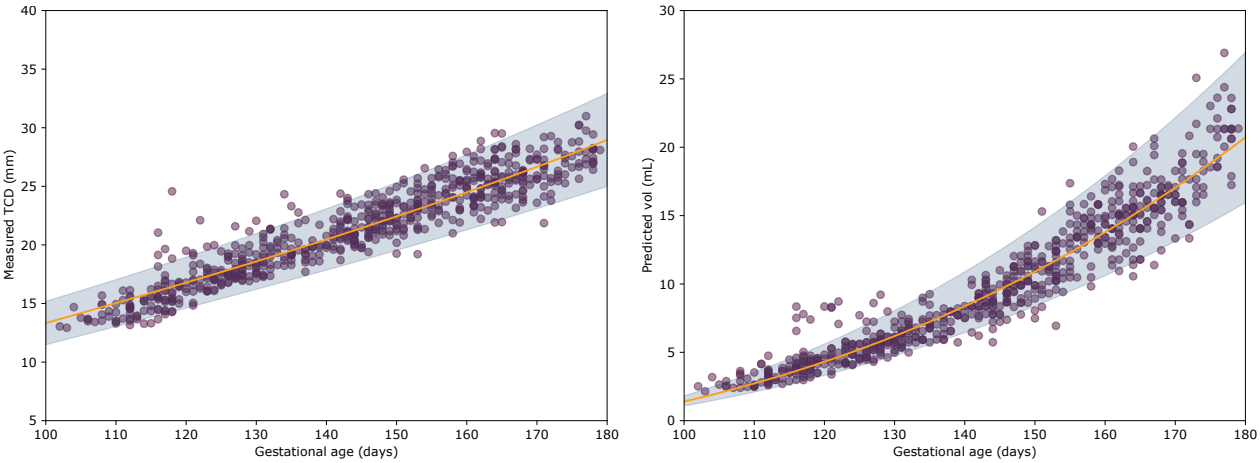
It is important to check to what extent the information being appended to the network in later experiments is already included in the image data itself. Since, for the US dataset, the scalar value added is gestational age, this section considers how well gestational age can be

Source	Method	MSE (days)	MAE	% vols ± 7 days
Rodriguez-Sibaja [4]	Manual TCD measure	7.79	5.52	89.7%
Our method (TCD)	Auto. TCD measure	7.91	6.05	87.1%
Our method (vol.)	Auto. volume measure	5.69	4.21	95.3%
Namburete [11]	Direct pred. on image	6.10	N/A	N/A

Table 6.3: Predictive power of different methods which estimate age from features present in the imaging data.

predicted from the image.

There are several biomarkers visible in fetal ultrasound that are predictive of gestational age, such as the biparietal diameter (BPD), crown-rump length, or femur length [145]. One gestational age predictor that is visible in the 3D neuroimaging volumes considered in this thesis is the transcerebellar diameter (TCD). It is considered robust to changes in fetal environment that otherwise lead to fetuses that are smaller or larger than expected for their gestational age, due to the brain-sparing hypothesis discussed in Chapter 2.



(a) (b)

Figure 6.5: (a) estimated TCD and (b) estimated cerebellar volume drawn from segmentations with a baseline CNN in the unlabelled ultrasound dataset. Curves of best fit drawn from (a) Rodriguez-Sibaja et al [4] and (b) regressed from this dataset.

Rodriguez-Sibaja et al manually measured the TCD of a large sample of the INTERGROWTH-21st dataset [4] and fitted a polynomial function to predict gestational age from TCD measurement. They use a cubic polynomial of the form

$$GA = ad^3 + bd + c$$

in terms of the TCD d and constant terms a, b, c which they found empirically [2]. The curve is shown in Figure 6.5a: the cubic component is very small in the gestational age range considered for this study, and the curve appears roughly linear.

The network used in Chapter 5 was used to segment the cerebellum of the large unlabelled dataset used in that chapter. From a cerebellar segmentation, it is straightforward to measure the TCD: the pair of segmented points furthest from each other can be found, and the distance between them is the estimated TCD. Figure 6.5a shows that this method yields a good estimate of TCD for the dataset used, comparable to that reported by Rodriguez-Sibaja et al on the same dataset.

Another way to estimate gestational age from a cerebellar segmentation is to directly measure the cerebellar volume. Rodriguez-Sibaja et al [4] use a cubic function to estimate the gestational age from TCD, which is approximately linear in the gestational age region of interest. TCD is a 1-dimensional length measure, while volume overall is a 3-dimensional measure. If one assumes that cerebellar development occurs at a similar pace along all three spatial axes, it's reasonable to consider the use of a cubic function to predict fetal cerebellar volume from gestational age. A curve to find gestational age t was found empirically, regressing the constant terms a, c in the cubic equation

$$TCD = at^3 + c.$$

This yielded estimates of $a = 4.0 \times 10^{-6}, c = -2.6$. [3]

Finally, the most straightforward way to estimate gestational age from an ultrasound image is to train a method directly to estimate gestational age. Namburete et al [11] do so, using random forests, and achieve an average MSE of 6.10 days across a wider gestational age

[2] Empirically, they find $a = 7.6187 \times 10^{-5}, b = 0.8154074, c = -3.957113$.

[3] For comparison, I also performed an exponential regression in the form $TCD = ae^{bx}$, to find if an exponential function would be a better fit for the data. Despite an equal number of tunable parameters, this led to significantly worse fit, yielding an MSE of 6.41 days.

range.

Table 6.3 shows a comparison of different approaches (both ours and other published work) that estimate fetal GA from image data alone. These methods all find that the image data can predict gestational age to within 4-7 days, but that nevertheless this leads to a significant minority of volumes for which gestational age is estimated incorrectly (with > 1 week between estimated GA and recorded GA). These comparisons are made from volumes in the same dataset, and using similar age ranges, but across studies with different inclusion criteria for individual volumes: caution must be used when comparing their results directly.

Measuring TCD on the segmentations produced automatically by our network and applying Rodriguez-Sibaja et al's formula to estimate age yields a reasonable estimate of GA, with no noticeable systematic error. Our method to estimate gestational age from volume finds the smallest average error along all metrics considered. While overfitting may be a concern (the parameters of the cubic equation chosen to fit gestational age were determined by the data available), this was minimised by reducing the number of tunable parameters to two: furthermore, the other estimates presented in this table also use parameters derived from the dataset against which they measure their errors. Section 6.5.1 further discusses these results.

6.4.2 Segmentation performance

Three-fold cross-validation was used to generate the segmentation results. The results of adding features, for both the ultrasound and the MRI datasets, are shown in Table 6.4.

In the ultrasound dataset, adding age as an additional layer led to significant increases in segmentation performance, measured both by Dice coefficient ($p < 10^{-11}$) and Hausdorff distance. This effect cannot be accounted for by the additional convolutional layers from adding another input: adding a non-informative input of the same dimensions leads to performance that is not significantly different from baseline.

The effect size is smaller in the MRI dataset, but still statistically significant ($p < 0.01$). Different features appear to have different effects when passed into TSI-Net: age and diagnosis increase segmentation performance, but sex has no significant effect. Section 6.4.3 explores

Fetal ultrasound dataset		
Added features	Dice coefficient	Hausdorff distance (mm)
Baseline	0.673 ± 0.046	12.25 ± 3.57
Zero layer	0.678 ± 0.037	9.81 ± 3.82
Age	0.723 ± 0.023	8.89 ± 3.35
ADNI/HARP MRI dataset		
Added features	Dice coefficient	Hausdorff distance (mm)
Baseline	0.848 ± 0.015	3.97 ± 0.25
Age	0.850 ± 0.014	3.64 ± 0.22
Sex	0.848 ± 0.015	4.08 ± 0.30
Diagnosis	0.849 ± 0.011	4.79 ± 0.48

Table 6.4: Performance of baseline CNNs and CNNs with additional features for both the ultrasound dataset and the MRI dataset.

the effect of combining features on TSI-Net's output.

6.4.2.1 Additional input format

Layer introduced	Training time (ms)	Validation Dice
Baseline (none)	852	0.673 ± 0.023
conv1 zeroed (160^3)	929	0.678 ± 0.037
conv1 ($\mathbf{160^3}$)	**945**	$\mathbf{0.723 \pm 0.023}$
conv2 (80^3)	886	0.697 ± 0.022
conv3 (40^3)	861	0.675 ± 0.021
Transposed (80^3)	931	0.700 ± 0.035

Table 6.5: Training time per sample and validation Dice coefficient for different locations for the additional input layer. Two baseline networks are considered: one with no additional features at all, and one with an all-zero additional input.

Table 6.5 shows the change in training time and segmentation performance from adding the scalar input at different levels of the TSI-Net encoder path. As confirmed by Table 6.2, passing additional features further down the encoder pathway reduces training time per sample. This

reduction in input generation time only accounts for approximately 10% of training time per sample, so it is a minor factor. Validation Dice coefficient decreases monotonically too as the additional input is passed further down the encoder pathway. This is expected, as an input passed part of the way through a CNN processing pathway is processed by fewer filters, and earlier layers lack some informative context. However, this quickly declines to baseline.

Using a smaller additional input, but using transposed convolution to increase its size until it can be concatenated with the image input at the top of the encoder pathway (see Figure 6.4]) does not appear to share the same reduction in training time as passing this input further down the encoder pathway. The results obtained by this method are insignificantly lower than the results obtained by a simple concatenation of inputs, both in terms of training time and in final performance. The additional memory requirements of this architecture are small, so there is little practical difference between these methods. The addition of this transposed-convolution pathway is therefore unnecessary.

Shape	Validation Dice
Square	0.699 ± 0.023
Triangular	0.680 ± 0.018
Cubic	0.651 ± 0.031
Gaussian	$\mathbf{0.723 \pm 0.023}$

(a)

Line width σ	Validation Dice
1	0.702 ± 0.023
5	0.717 ± 0.023
10	$\mathbf{0.723 \pm 0.023}$
50	0.680 ± 0.023

(b)

Table 6.6: (a) the Dice coefficient obtained when using different shapes to encode the scalar value. (b) the Dice coefficient obtained using different values of the width parameter σ to train a CNN with a Gaussian shape.

Table 6.6 shows how the network's ultrasound segmentation performance is sensitive to the form of the function used to encode scalar features, and its width parameter σ (see Section 6.3.2). This table confirms that a Gaussian function produces the greatest boost in performance. This may be because the other functions take values of 0 outside the range $[p - \sigma, p + \sigma]$, where p is the peak value of the function, while the Gaussian function monotonically increases to the peak through the range of the volume (see Figure 6.1).

A very high width parameter σ obscures the precise scalar value of the feature passed

to the network, by drawing a function with a broad and flat peak. A very low value of σ, on the other hand, produces a very narrow function which reduces the training signal for convolutional layers: the optimal value appears to be a balance of these considerations. For the ultrasound dataset (with input size 160^3) the optimum value of σ in this sweep was found to be $\sigma = 10$, so this value was used for all other experiments in this dataset. A similar sweep was not conducted for the ADNI/HARP dataset (with input size 64^3): instead, the value of the parameter σ was simply scaled to the input size for this dataset, resulting in a value of $\sigma = 4$ used for all experiments in this dataset for this chapter.

Feature combination method	Training time (ms)	Validation Dice
Concatenate	450	$\mathbf{0.852 \pm 0.015}$
Add	**447**	0.850 ± 0.015
Maximum	493	0.851 ± 0.016

Table 6.7: A comparison of the training times per batch and test Dice coefficients obtained with the different methods considered to combine different features in the ADNI/HARP dataset. All three available features were used in this experiment.

Finally, on the ADNI/HARP dataset, experiments were performed to combine different features in a single feature map. The three methods compared were concatenating the features as separate maps, adding them into a single map, and combining them using a maximum function (see Figure 6.1).

Table 6.7 shows the training time (per batch) and test Dice coefficients obtained using those methods. These results confirm that simply adding feature maps together is the fastest method, while using a maximum function is the slowest, leading to a roughly 10% increase in training time per batch. Both methods, however, are outperformed by a simple concatenation of feature maps encoding separate features. While this results in a marginal ($\sim 1\%$) increase in the memory usage of TSI-Net, it does not significantly increase training time: it appears to be the best choice for the tasks in this chapter.

Feature added	Test Dice	Hausdorff (mm)
None (baseline)	0.848 ± 0.015	3.97 ± 0.25
Age	0.850 ± 0.014	3.64 ± 0.22
Sex	0.848 ± 0.015	4.08 ± 0.30
Diagnosis	0.849 ± 0.011	4.79 ± 0.48
Age+sex	0.850 ± 0.014	5.40 ± 0.74
Age+diagnosis	$\mathbf{0.852 \pm 0.016}$	$\mathbf{3.64 \pm 0.44}$
Sex+diagnosis	0.850 ± 0.016	5.32 ± 0.70
All features	0.852 ± 0.015	3.90 ± 0.39

Table 6.8: Impact of passing individual scalar features, both by themselves and in combination with other scalar features, on test Dice coefficient and Hausdorff distance on the ADNI/HARP dataset.

6.4.3 Contributions of individual features

The overall effect size of adding any feature in the MRI dataset is small, so caution is necessary when estimating contributions of each feature. Age and diagnosis feature each appear to improve the segmentation Dice coefficient of the network relative to baseline, both when passed individually and when passed together. On the other hand, passing the subject's sex to TSI-Net does not appear to have any predictive value for hippocampal segmentation.

When multiple features are passed together at the input of the network, this appears to have an additive effect on performance: passing age+diagnosis results in higher test Dice coefficient and lower test Hausdorff distance than passing either measure alone ($p = 0.01$). As before, this effect is not observed for sex.

6.4.4 Varying input features

It is also possible to vary the input features during training and/or test, to measure how sensitive the recorded performance boost is to the precision of the features used. An error can be added to the scalar value (passing an additional input corresponding to an age of $t \pm \delta$, where δ is a known error magnitude).

Figure 6.6 shows how TSI-Net's segmentation performance at test time varies when an

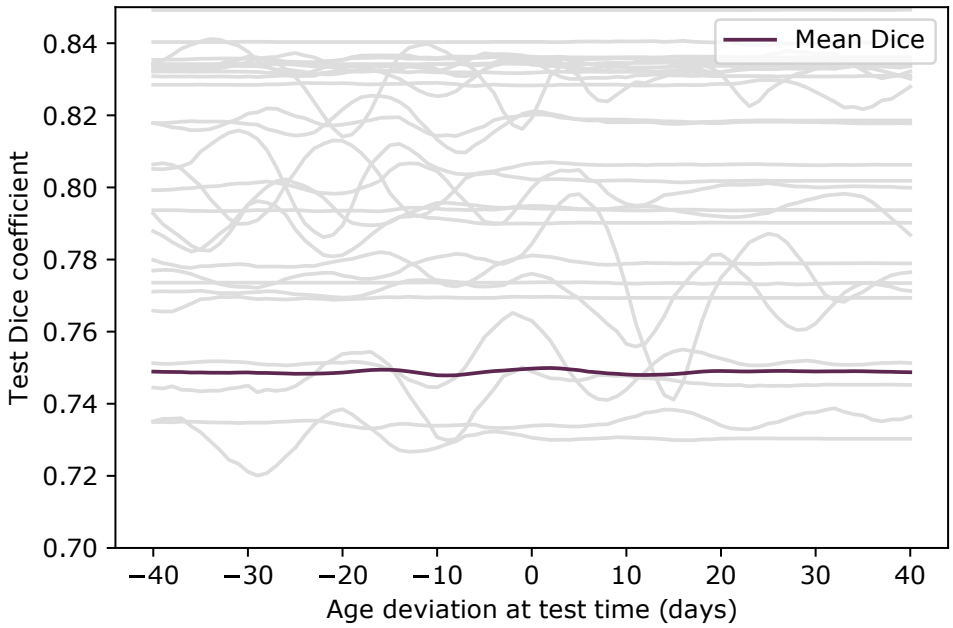


Figure 6.6: When an incorrect age is input into the network at test time, the segmentation Dice coefficient only varies slightly. In grey is the Dice coefficient for each volume in the test set.

incorrect gestational age is input at test time. While individual volumes appear to display some fluctuation in segmentation performance, this is not correlated to how close to the correct gestational age the input is. To verify this effect, an additional experiment was run, passing no age at all (an array of zeros as the additional input) at test time: this also showed no degradation in performance ($p = 0.3$).

Overall, there is no significant effect from passing the network an incorrect feature at test time. This is surprising: TSI-Net shows a significant performance boost from passing this feature during training, and one would expect it to use the information at test time too. I discuss this further in Section 6.5.3.

Figure 6.7 shows the effect of adding uniform noise during training to the numerical gestational age used to generate the additional feature input. This effect is in the expected direction: increasing noise amplitude leads to lower performance, until performance is indistinguishable from baseline (with no added age input) when noise is sufficiently high. The additional benefit of TSI-Net therefore appears to be entirely accrued during training.

Due to the skip-connections present in this network architecture, it is not possible to investigate further how TSI-Net is representing this data internally, and how this representation

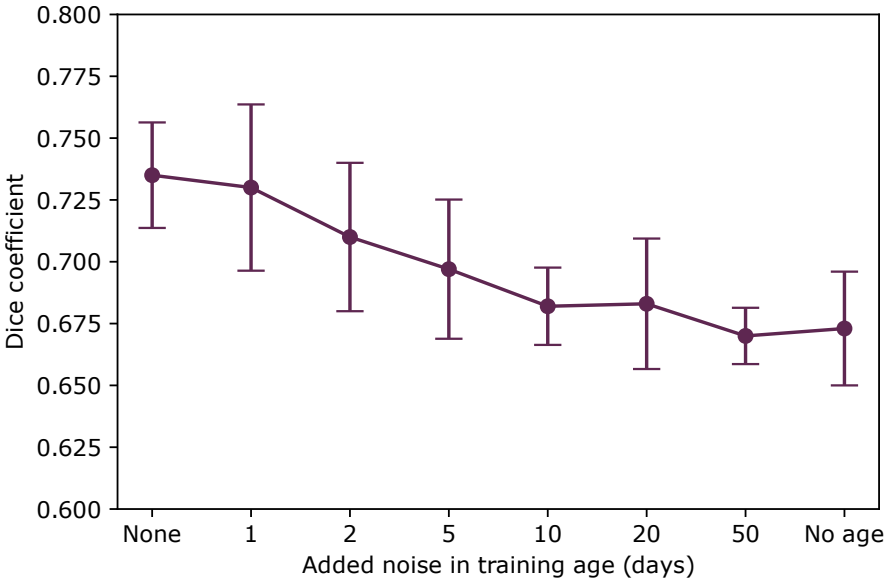


Figure 6.7: Adding uniformly distributed noise to the gestational-age measure in training leads to gradual degradation in segmentation performance.

differs from baseline. Methods such as t-SNE are only suitable for linear architectures such as autoencoders [166]: while the encoder-decoder architecture of the U-Net superficially resembles an autoencoder, the skip connections make it impossible for this method to capture internal representation, as this representation is spread across layers.

6.5 Discussion

6.5.1 Feature prediction

Section 6.4.1 shows that gestational age information is not fully contained within the image data itself. While various methods are proposed to estimate gestational age from data present in the image, all present some error, and for all a minority of volumes exhibit a large (>1 week) prediction error in gestational age. These methods depend on the target for the segmentation network: both TCD and cerebellar volume are in this section measured from a volumetric segmentation of the cerebellum. Volumes for which the segmentation algorithm performs poorly therefore also present higher errors in estimated gestational age: passing an additional scalar input may reduce the segmentation error in these cases.

The target against which these measures of gestational age are compared is the number of

days since the start of the last menstrual period (LMP) before pregnancy. This measurement itself presents some error: self-reported LMP date exhibits digit preference and is recalled poorly [164], which introduces some random error in LMP dating too. The error introduced by this has not been rigorously estimated, but it does present a bound for the accuracy of a CNN method.

Interestingly, total segmented cerebellar volume appears to be a better predictor of gestational age, within the dataset explored here, than TCD. This makes intuitive sense: random variation in anatomy and measurement along one dimension is likely to be greater than that of a volume across three dimensions: using a 3D measure should decrease this type of random error by a factor of $\frac{1}{\sqrt{3}} \approx 0.58$. Other sources of error (such as in the measurement of LMP) are independent of the number of dimensions from which these estimates are derived, so the real-world decline in error is lower.

These results suggest a novel biomarker that may be of clinical use: cerebellar volume, if it can be estimated quickly and reliably from a segmentation method, can be a useful additional tool for clinicians as a gestational-age estimators. Measurements such as TCD are simple linear measures that only require a clinician to draw a straight line in a single ultrasound plane, while manually obtaining an accurate estimate of 3D cerebellar volume is impractical and time-consuming in a clinical setting. An automatic method to segment the fetal cerebellum and measure its volume is therefore necessary for this biomarker to be used in the clinic.

6.5.2 Performance boost

Figure 6.8 shows an example slice of an ultrasound volume, segmented both by the baseline network architecture and by TSI-Net. The headline improvement in Dice coefficient from using TSI-Net is larger than average (0.752 to 0.833), even though this volume has a higher-than-average Dice score for segmentation for both networks. The Hausdorff distance shows an even more pronounced improvement, from 5.66mm to 3.50mm.

While the centre of both cerebellar lobes is well-segmented by both architectures, the

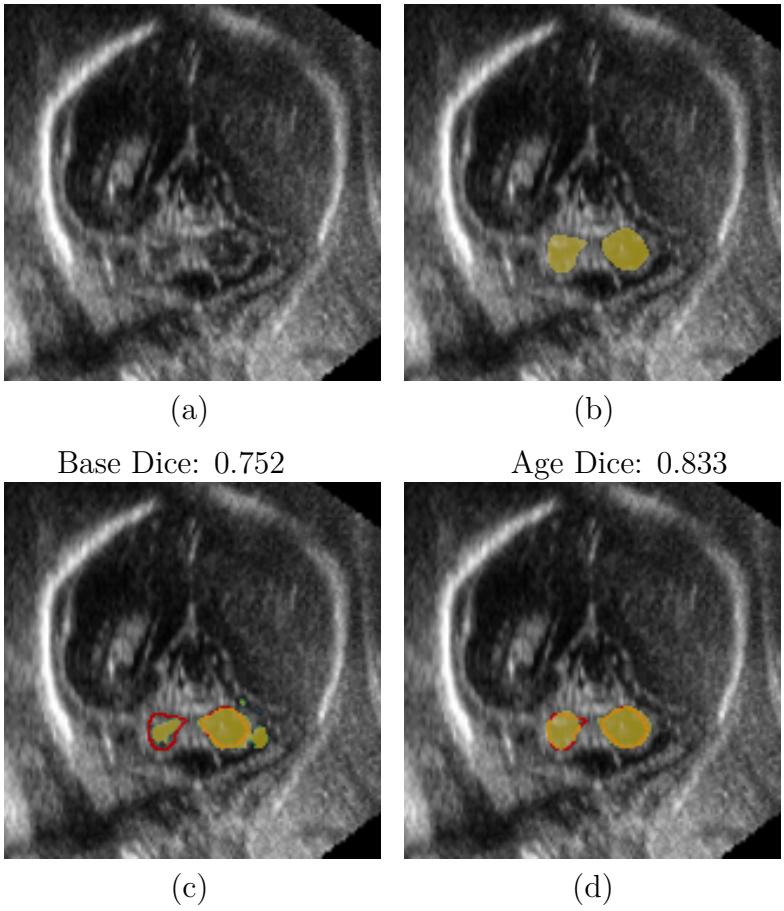


Figure 6.8: An example of the performance boost effect of adding age to the input layer. (a) A slice of a volume in the validation dataset (at 23 weeks). (b) The ground truth segmentation of the cerebellum, (c) the segmentation obtained after training a baseline CNN on labelled data. (d) the segmentation obtained by training the same CNN with an additional age input.

boundaries show considerable uncertainty and misclassification in the baseline architecture, which are corrected by the addition of age at the input. This is a plausible effect: the morphology of the cerebellum changes significantly during the second trimester [47], and anatomical boundaries are often difficult to discern in ultrasound. Passing age as an additional feature therefore seems to provide TSI-Net with a stronger anatomical prior on the segmentation, reducing uncertainty near the boundaries.

Figure 6.9 shows the CDF of the Dice coefficient, on the ultrasound test dataset, for the baseline CNN and the age-augmented CNN. The performance advantage from adding gestational age appears to apply across the dataset, improving segmentation performance both for volumes that are already segmented well by the baseline architecture and for those that are segmented poorly. This effect also holds across the gestational age range in the

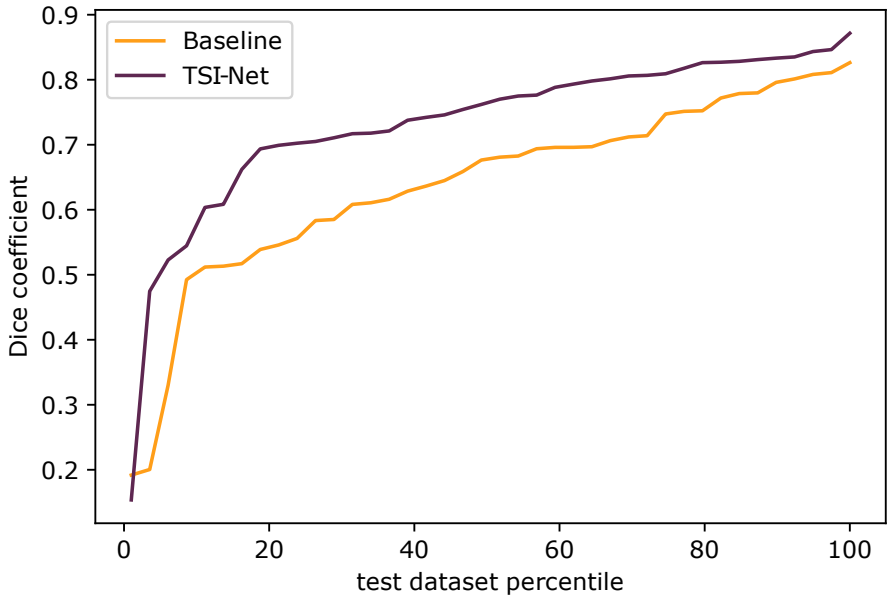
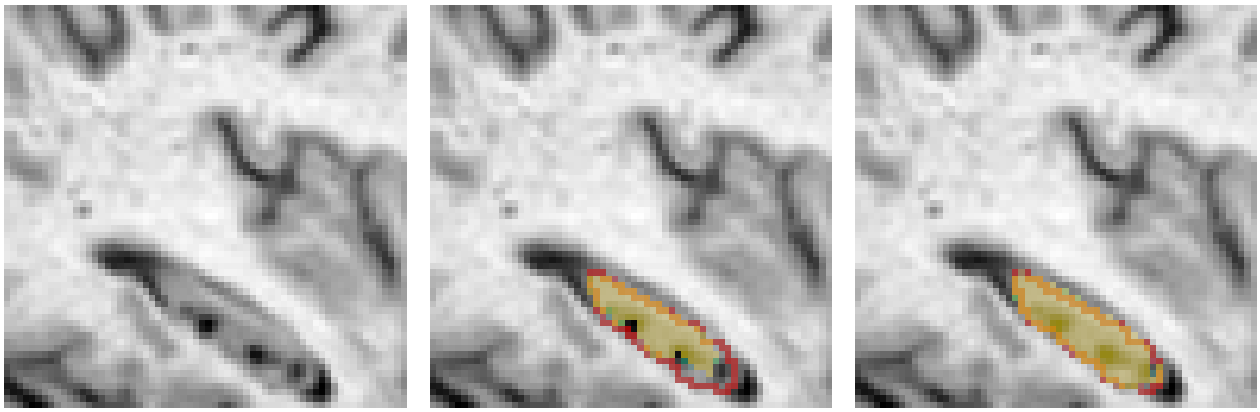


Figure 6.9: CDF of the Dice coefficients of the test set, comparing performance of the baseline network and the network trained with additional gestational age. The CNN with added age outperforms baseline across the dataset ($p < 10^{-11}$).

dataset, with no significant difference in performance change by GA.



(a) (b) (c)

Figure 6.10: (a) Example sagittal slice showing the hippocampus of a 70 year old male with probable Alzheimer's disease. (b) The baseline segmentation of that slice, with ground truth outlined in red, and (c) the segmentation obtained when passing all scalar features to TSI-Net. Overall Dice coefficient increases from 0.883 to 0.913.

The impact of passing scalar features to the network for the hippocampal segmentation task was smaller, but still statistically significant. Age and clinical diagnosis appeared to improve the performance of the CNN, both when passed individually and passed together at the network input, while patient sex did not have a notable impact. Ageing is associated with morphological changes in the hippocampus [167], and regardless of age, Alzheimer's

disease-related memory loss appears to be preceded by noticeable hippocampal atrophy [64, 159]. This supports the hypothesis that these features have independent predictive power on the morphology of the hippocampus and explains why passing these features together is better than passing them individually.

A tangible example of the benefit of passing characteristics such as clinical diagnosis to TSI-Net is shown in Figure 6.10. Figure 6.10 shows an axial slice of an MRI scan of a 70 year old male with probable Alzheimer's disease. This slice displays prominent perihippocampal fissures (PHFs) filled with cerebrospinal fluid, which appear as darker spots inside the hippocampus. These are a biomarker of accelerated hippocampal atrophy [168], which is known to be exhibited in Alzheimer's disease. As these are contained within the hippocampus proper, the segmentation protocol calls for them to be included as part of the segmentation, as they are in the manual ground-truth label [20], but the baseline network architecture segments these lesions as "background". Using TSI-Net, these voxels are segmented correctly, thanks to the additional information the network has available.

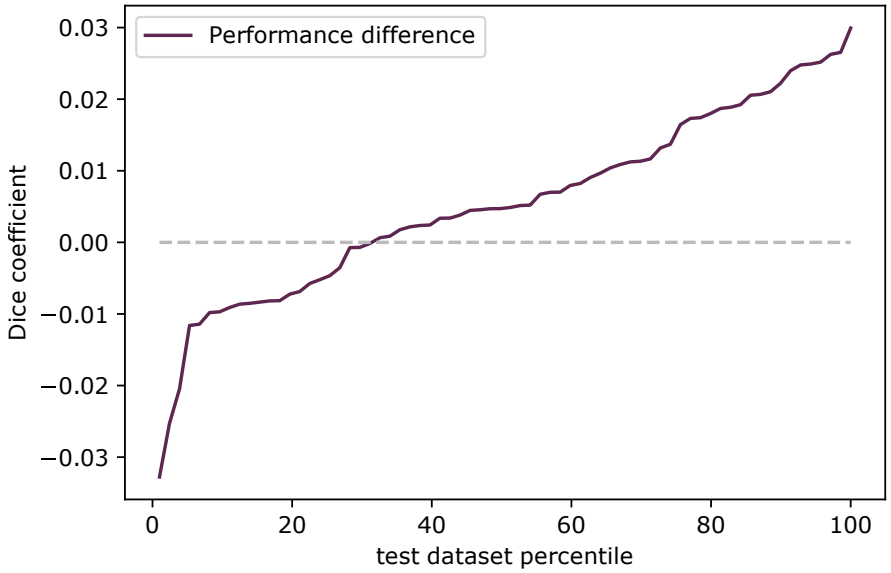


Figure 6.11: CDF of the performance difference for each volume in the ADNI/HARP dataset, relative to baseline. 48/70 (68.5%, $p = 0.001$) volumes show a performance improvement.

Figure 6.11 shows the change in Dice coefficient for each volume in the test set between the baseline network architecture and TSI-Net when all features were passed at input. The overall effect size is small, with an average boost in Dice coefficient of 0.04. However, the

majority of volumes (68.5%, $p = 0.001$) still shows a performance improvement when using TSI-Net, demonstrating the validity of this technique on a different neuroimaging dataset.

Similar to Chapter 5, the performance boost appears larger for the fetal ultrasound task than for the MRI task. This is likely due to the same reason as that found in that chapter: the ultrasound segmentation task is more challenging, with more variable data which makes segmentation more ambiguous. Access to more data, therefore, improves segmentation performance more on this dataset than in the MRI dataset, where structures are more well-delineated and imaging features are more consistent across scans.

6.5.3 Input format and network architecture

Section 6.4.2 finds that passing the additional input further at levels further down the encoder pathway, reducing the size of the input required, decreases training time but also reduces the performance benefit gained from the additional input. Segmentation performance quickly reverts to baseline as features are introduced at deeper layers in the pathway.

Interestingly, Section 6.4.4 finds that passing incorrect scalar features at test time, or no scalar features at all, still yields the same performance improvement: the increase in segmentation quality only comes from passing that information during training. This suggests that the additional features passed to TSI-Net are providing a form of regularisation: the additional information passed to the network appears to change the internal representation of image data and reduce overfitting.

6.5.4 Analysis of failure cases

The CNN trained on the ultrasound dataset was used to segment the same large unlabelled dataset used in Chapter 5. 11 volumes in ultrasound (1.4% of the dataset) were entirely segmented as background, with no voxels classified as cerebellum at all. The unlabelled data these segmentations were generated from had no ground-truth segmentations to allow direct evaluations of performance, but these were selected as clear failure cases (all fetuses in this dataset have a cerebellum).

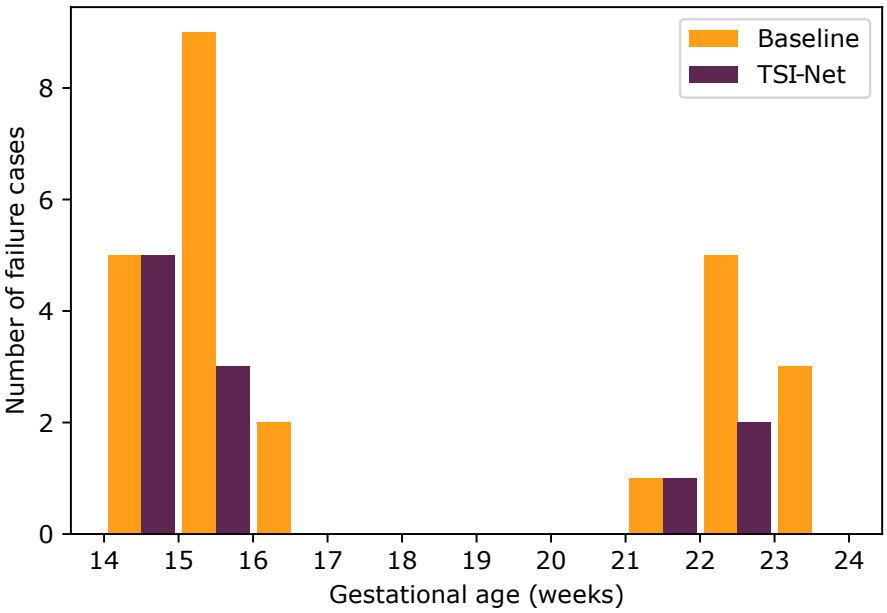
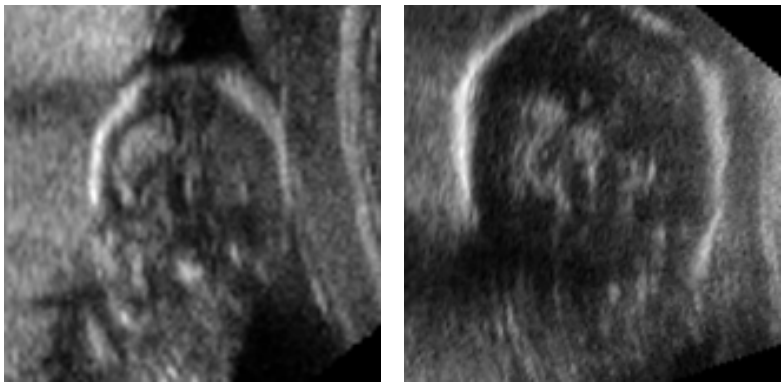


Figure 6.12: The gestational age of volumes in the unlabelled dataset which were entirely segmented as background, for both the baseline model and TSI-Net.

The failure rate is a marked decrease from the rate observed in the baseline network and reported in Chapter 5: a reduction from 25 null-segmented volumes to 11, a decrease of 56%. Figure 6.12 compares the gestational ages of failure cases with this model to those resulting from segmentation with a baseline CNN. The overall age distribution of the failure cases appears similar between the baseline architecture and ours: a bimodal distribution with failures concentrated at the extremes of the age range for the dataset, with no failures in the middle section of the age distribution.

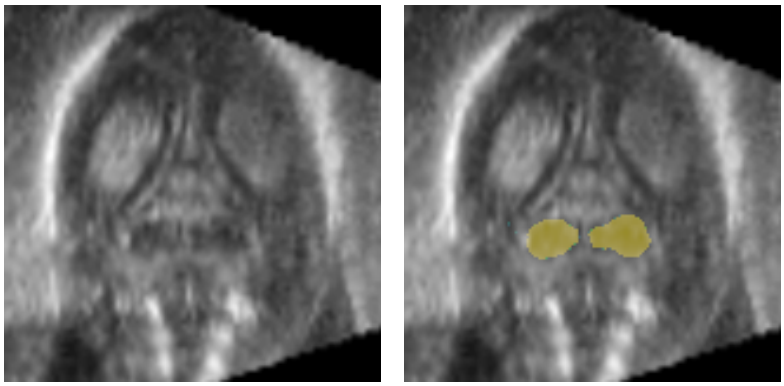


(a) (b)

Figure 6.13: (a) An example of a failure case at a gestational age of 14 weeks. (b) An example of a failure case at a gestational age of 23 weeks.

Figure 6.13 shows examples of failure cases at both ends of the age range. As with the

baseline architecture, two distinct failure modes are visible. At the lower end of the age distribution, the CNN does not appear to have learned to segment the cerebellum, even when it is clearly visible. This is likely due to a lack of training data in this age range: the dataset has fewer volumes in this age range (see Chapter 5), and the training data does not cover the variability seen in this age range.



(a) (b)

Figure 6.14: (a) A volume classed as a failure by the baseline CNN architecture, at 17 weeks. (b) The segmentation produced by TSI-Net.

On the higher end of the age distribution, some volumes display prominent shadow artifacts from the temporal bone which obscure the cerebellum. It is unclear whether these volumes can be truly called failures.

Every individual volume for which the age-based method failed to produce a segmentation was also classed as a failure by the baseline CNN. In this dataset, passing age therefore decreases the number of total failures of segmentation without creating new ones. Figure 6.14 shows an example of a volume at 17 weeks, for which the baseline CNN failed to produce a segmentation at all, while TSI-Net produces a full volumetric segmentation output. This volume (like others in this age range) has fewer shadows and lower contrast between tissue types than older fetuses, which changes the appearance of anatomical boundaries: passing the age at the input can help the CNN interpret these features.

6.6 Conclusion

This chapter has proposed a novel method to combine image and scalar data in fully-convolutional network architectures for medical image segmentation. The challenge presented here is inherent in the nature of convolutional filters: they can only operate on grid-like data, making it impossible to straightforwardly process scalar features using them. To resolve this, I propose TSI-Net, a scheme to encode scalar features into a grid-like volume of appropriate dimensions, that I then pass as an additional input to the CNN.

Using TSI-Net for a cerebellar segmentation task in ultrasound leads to a significant performance improvement, which allows more accurate segmentations without requiring additional image data or segmentation labels. I replicate these results with a hippocampal segmentation task which is also significantly improved with the addition of age and clinical characteristics, and show that adding these features together results in a better output than adding them individually. While the effect size is small for this task and it is difficult to measure accurately the contribution of each feature, passing the subject's sex to TSI-Net did not appear to improve segmentation performance.

This chapter also compares different methods of estimating gestational age from features present in the medical image. It shows that volumetric measures of cerebellar size can outperform traditional measures such as TCD at that task. This is a very simple measure that can be estimated from a 3D segmentation, but is not generally measured in the clinic. This result suggests one possible value of reliable machine-learning segmentation methods: new biomarkers can be found from these measurements to supplement the measures already available.

Chapter 7

Conclusion

Chapter layout

This thesis has presented several approaches to improve 3D medical image segmentation when traditional training labels are limited, a common problem in medical image analysis. It has shown that, with appropriate registration, atlas labels generated in MRI acquisitions can serve as training labels for ultrasound data. It has also shown that uncertainty estimates can boost an omni-supervised framework to include unlabelled data in a network training set. Finally, it showed that passing additional scalar inputs to a network input can improve segmentation performance, and proposed one framework to do so.

This chapter summarises the main contributions presented in this thesis in Section 7.1. Section 7.2 discusses the limitations of this work and explores opportunities for future research directions.

7.1 Overview of contributions

The main contributions of this thesis have been:

Atlas-based label generation and multi-label segmentation

We used prior anatomical knowledge to guide atlas-based registration, reducing the manual annotation time needed to generate a training dataset. We manually labelled a set of known anatomical keypoints in two atlases of the developing fetal brain acquired in different modalities: ultrasound and MRI. This was done in a gestational age range of $20 - 25$ weeks. The anatomical correspondences created by this made it possible to estimate a set of affine transforms to align the two atlases to each other, and therefore to apply the anatomical labels found in the MRI atlas to ultrasound. These labels were then transformed from the common reference space of the atlas to the individual ultrasound volumes, inverting the nonlinear transformation used to construct the atlas. This created a set of high-quality multi-label segmentations for 402 ultrasound volumes, covering the thalamus, brainstem, cerebellum, and white matter.

This method could only be used for those volumes for which a transformation to the atlas reference space was known. To extend this segmentation method to new volumes, we trained a 3D multi-task CNN to segment four key structures in the fetal brain. We found that this architecture produces good results, roughly in line with the literature, and that a multi-task 3D architecture outperforms 2D or single-task 3D architecture. We hypothesise that this is due to increased context from the use of 3D convolutional filters, and increased supervision from additional training labels from more structures (which often share anatomical boundaries).

Uncertainty as a selection tool for omni-supervised segmentation of the fetal cerebellum

This contribution explored the use of additional unlabelled data to enhance training with a limited dataset, as well as the use of estimates of uncertainty to inform data selection in an omni-supervised framework. Omni-supervised learning is based on the principle that "a set of weak learners can create a strong learner" [117]: that combining the predictions of multiple learners produces an aggregate prediction superior to any of them. Combining a set

of segmentations produced by different networks produces aggregate segmentations of higher quality than the outputs of individual networks, and omni-supervised methods then use these segmentations as training labels for a new set of learners.

This contribution proposed a novel method to select additional volumes for further rounds of omni-supervised learning. The volumes with the lowest aleatoric uncertainty in their segmentations, corresponding to less ambiguous image data, are segmented better by a CNN. They therefore lead to a larger performance improvement when added to the training set for further rounds of omni-supervised training. This is shown in two independent datasets from different modalities.

Adding information from scalar features at network input

We proposed TSI-Net, a method to incorporate scalar features at the input of a segmentation CNN architecture. Scalar features, such as age or clinical characteristics, are commonly and routinely collected alongside image data, and used by clinicians in their assessments of image data. However, they cannot be processed by convolutional layers in a fully-convolutional CNN.

TSI-Net transforms this scalar data to an image-like format, allowing a fully-convolutional CNN to process this additional information with minimal changes to its architecture. The best way to transform this data and pass it as an additional input to the network was found experimentally.

Including this information at the network input improves segmentation quality produced by a CNN, in both an ultrasound and an MRI dataset. Intriguingly, this performance boost appears to come entirely during training, and no significant decrease in performance is noted if the additional information is omitted at test time. This suggests that the additional features passed into TSI-Net provide a form of regularisation, improving the internal representation of the data.

7.2 Limitations and future work

7.2.1 Gestational age range

Most of this work focused on ultrasound volumes in the age range of 14-25 weeks. A typical pregnancy is 40 weeks long, so this work does not consider image analysis in the third trimester. In the INTERGROWTH-21st dataset, 3D ultrasound volumes were collected every 4 weeks, from 14 weeks' gestation to delivery [134]. This covers the second and third trimesters of a healthy pregnancy, when individual brain structures are visible.

The third trimester was excluded because the imaging protocol used cast acoustic shadows from the petrous ridges of the temporal bone onto the cerebellum. These shadows became more pronounced at higher gestational ages, as the skull ossifies and becomes more reflective to ultrasound. Above 25 GW, a very high proportion of volumes in this dataset show acoustic shadows that completely obscure the cerebellum, and therefore the entire age range of 25 GW - 40 GW were excluded from this thesis.

This limits the strength of the conclusions to be drawn from the results presented here, on the ultrasound INTERGROWTH-21st dataset. There is no reason to believe that the approaches explored in this thesis would not aid segmentation at higher gestational ages, but this has not been shown experimentally here. A CNN could be trained on additional datasets acquired with a different protocol which does not obscure the cerebellum at higher gestational ages, or segmentation could be attempted on different anatomical structures.

However, it is possible that with a sufficiently wide range of gestational ages, enough variability is present in the morphology that CNN segmentation performance declines with an increasing age range. This was not observed in the age range considered in this thesis, but may occur with a wider age range. In that case, separate networks (or if sufficient memory is available, wider networks) may be needed.

7.2.2 Derivation of fetal biomarkers from segmented structures

Chapter 6 found that fetal cerebellar volume can be a notably more accurate biomarker of fetal gestational age than TCD. In the same dataset, volume predicted GA with a mean-squared error of 5.69 days, while TCD predicted GA with an MSE of 7.91 days.

Most biomarkers measured at routine ultrasound scans, such as TCD, fetal head circumference, biparietal diameter, crown-rump length, and femur length [144], are simple and generally linear measures. They can be easily measured in real time by a clinician, by manually labelling key points (such as the furthest points in the cerebellar lobes from the transcerebellar view, in the case of TCD). Biomarkers such as fetal cerebellar volume cannot be measured as simply in real-time: they require segmentation of the 3D structure, which is time-prohibitive if performed manually (see Section 5.3.1.1).

Automatic segmentation of fetal brain structures, as outlined in this thesis, makes the routine collection of volumetric measures feasible if it can be performed at high quality. From a volumetric segmentation, it is trivial to estimate cerebellar volume, and other volumetric measures can similarly be estimated, such as surface area or gyrification index [169]. These further measures have not been considered in this thesis, but volume itself showed a significant improvement on other measures considered here.

Further validation of fetal volumetry, therefore, would be required before this measure can be used in the clinic, but the results shown in this thesis provide an interesting proof of concept.

7.2.3 Using scalar features as supervision targets

In Chapter 6, scalar features were used as an additional source of information when passed into the network. It may also be possible to use this information as an additional output for the network, rather than an input, and train a CNN to predict the scalar features associated with those volumes.

Chapter 4 also found that a multi-task network architecture outperforms single-task architectures at each of the tasks at which it is trained. In that chapter, I hypothesise that

this is because every additional task also provides additional training labels, and increases the supervision signal passed to the network. This concept can extend beyond segmentation alone: a multi-task network architecture could also be used to simultaneously perform a segmentation task and a scalar regression task.

Amyar et al [170] modify the U-Net architecture to simultaneously classify patients' COVID-19 status and segment lung lesions in a single multi-task architecture, and report that adding a classification task to the CNN architecture also improves segmentation performance. A similar modification could be made to the 3D U-Net architecture presented in this thesis.

Every volume in the INTERGROWTH-21st dataset has been labelled with the gestational age of the fetus, as measured by LMP. Similarly, every volume in the ADNI has a set of scalar values associated with it, as described in Chapter 3. These are available even in volumes which do not have a manual segmentation label. It may be possible, then, to use these volumes for training, using only the age label as a supervision signal.

A multi-task architecture could also be of additional clinical utility: a CNN that can automatically estimate gestational age, as well as provide clinical volumetry, would aid clinical practice.

7.2.4 Evaluating uncertainty as a way to identify abnormalities and suggest features for follow-up

Chapter 5 examined different ways to estimate uncertainty in CNN-based segmentations. Epistemic uncertainty, as measured by test-time dropout, can be thought of as a measure of training-set coverage: high epistemic uncertainty generally corresponds to features that are not well-represented in the training dataset.

This could suggest an active-learning framework: the acquisitions which show the highest epistemic uncertainty can be suggested to a human annotator for manual annotation and added to the training dataset, improving training-set coverage of the image space. di Scandalea et al [171] use such a framework to aid segmentation of nerve cells in histology data, and Gorriz et al [172] use a similar scheme for melanoma lesions. This could be extended to 3D

segmentation of fetal brain structures, increasing coverage of the feature space.

Another option to use epistemic uncertainty for abnormality detection. The dataset available in this thesis includes data from optimally healthy fetuses only: as a result, abnormalities are not represented within it. A CNN trained using the INTERGROWTH-21st dataset would therefore be likely to show high uncertainty when it encounters samples that contain abnormalities, and these could be flagged for a clinician to review. Seebock et al [173] use this method to detect diabetic retinopathy in fundus images from a network trained only using healthy images, reporting 70% sensitivity. This could be used to address the common problem in medical image analysis that abnormalities are generally uncommon, and under-represented in training datasets.

7.3 Summary

The research presented in this thesis explored different ways to include additional knowledge to improve medical image segmentation by machine-learning techniques, especially when there are few traditional manual labels available. One way to do so is to leverage prior knowledge about anatomy: by using a small number of manually annotated keypoints to generate a set of correspondences, it was possible to generate a large set of training labels for individual volumes, which was then used to train a multi-task segmentation network. Another method explored in this thesis is the use of additional unlabelled image data. Using aleatoric uncertainty to select volumes in an omni-supervised framework was found to significantly improve segmentation quality compared to random selection or not using unlabelled data at all, in two datasets. Finally, the third additional information source used in this thesis was additional non-image data, which is generally collected with the image. This thesis proposes a method to transform this information to an image-like format, and shows that using this transformation to pass additional scalar information to a CNN also leads to significant improvement in segmentation performance, as measured across two different tasks.

Fetal volumetry as developed in this thesis may also present an opportunity to find new biomarkers of fetal health, which could be brought to the clinic. Preliminary results presented

in this thesis suggest that this may outperform traditional methods of estimating gestational age: with further validation, this could be a useful clinical contribution.

Bibliography

[1] X. Chen, S. L. Li, G. Y. Luo, E. R. Norwitz, S. Y. Ouyang, H. X. Wen, Y. Yuan, X. X. Tian, and J. M. He, "Ultrasonographic characteristics of cortical sulcus development in the human fetus between 18 and 41 Weeks of gestation," *Chinese Medical Journal*, vol. 130, no. 8, pp. 920–928, 2017. [Online]. Available: http://www.cmj.org/article.asp?issn=0366-6999

[2] A. Gholipour, C. K. Rollins, C. Velasco-Annis, A. Ouaalam, A. Akhondi-Asl, O. Afacan, C. M. Ortinau, S. Clancy, C. Limperopoulos, E. Yang, J. A. Estroff, and S. K. Warfield, "A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth," *Scientific Reports*, vol. 7, no. 1, 2017.

[3] A. I. Namburete, R. van Kampen, A. T. Papageorghiou, and B. W. Papież, "Multi-channel groupwise registration to construct an ultrasound-specific fetal brain atlas," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11076 LNCS, 2018, pp. 76–86.

[4] M. J. Rodriguez-Sibaja, J. Villar, E. O. Ohuma, R. Napolitano, S. Heyl, M. Carvalho, Y. A. Jaffer, J. A. Noble, M. Oberto, M. Purwar, R. Pang, L. Cheikh Ismail, A. Lambert, M. G. Gravett, L. J. Salomon, L. Drukker, F. C. Barros, S. H. Kennedy, Z. A. Bhutta, and A. T. Papageorghiou, "Fetal cerebellar growth and Sylvian fissure maturation: international standards from Fetal Growth Longitudinal Study of INTERGROWTH-21st Project," *Ultrasound in Obstetrics and Gynecology*, vol. 57, no. 4, pp. 614–623, 2021.

[5] A. Monteagudo and I. E. Timor-Tritsch, "Development of fetal gyri, sulci and fissures: A transvaginal sonographic study," *Ultrasound in Obstetrics and Gynecology*, vol. 9, no. 4, pp. 222–228, apr 1997. [Online]. Available: http://doi.wiley.com/10.1046/j.1469-0705.1997.09040222.x

[6] Y. Li, J. A. Estroff, T. S. Mehta, R. L. Robertson, C. D. Robson, T. Y. Poussaint, H. A. Feldman, J. Ware, and D. Levine, "Ultrasound and MRI of fetuses with ventriculomegaly: Can cortical development be used to predict postnatal outcome?" *American Journal of Roentgenology*, vol. 196, no. 6, pp. 1457–1467, 2011.

[7] S. Gha, K. W. Fong, A. Toi, D. Chitayat, S. Pantazi, and S. Blaser, "Prenatal US and MR imaging findings of lissencephaly: Review of fetal cerebral sulcal development," *Radiographics*, vol. 26, no. 2, pp. 389–405, mar 2006. [Online]. Available: https://doi.org/10.1148/rg.262055059

[8] C. Clouchoux, A. J. du Plessis, M. Bouyssi-Kobar, W. Tworetzky, D. B. McElhinney, D. W. Brown, A. Gholipour, D. Kudelski, S. K. Warfield, R. J. McCarter, R. L.

Robertson, A. C. Evans, J. W. Newburger, and C. Limperopoulos, "Delayed cortical development in fetuses with complex congenital heart disease." *Cerebral cortex (New York, N.Y. : 1991)*, vol. 23, no. 12, pp. 2932–2943, 2013.

[9] J. F. De Rijk-Van Andel, W. M. Arts, A. Hofman, A. Staal, and M. F. Niermeijer, "Epidemiology of lissencephaly type I," *Neuroepidemiology*, vol. 10, no. 4, pp. 200–204, 1991. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/1745330

[10] I. V. Koning, A. W. van Graafeiland, I. A. Groenenberg, S. C. Husen, A. T. Go, J. Dudink, S. P. Willemsen, J. M. Cornette, and R. P. Steegers-Theunissen, "Prenatal influence of congenital heart defects on trajectories of cortical folding of the fetal brain using three-dimensional ultrasound," *Prenatal Diagnosis*, vol. 37, no. 10, pp. 1008–1016, oct 2017. [Online]. Available: http://doi.wiley.com/10.1002/pd.5135

[11] A. I. Namburete, R. V. Stebbing, B. Kemp, M. Yaqub, A. T. Papageorghiou, and J. Alison Noble, "Learning-based prediction of gestational age from ultrasound images of the fetal brain," *Medical Image Analysis*, vol. 21, no. 1, pp. 72–86, 2015.

[12] R. Huang, W. Xie, and J. Alison Noble, "VP-Nets: Efficient automatic localization of key brain structures in 3D fetal neurosonography," *Medical Image Analysis*, vol. 47, pp. 127–139, jul 2018.

[13] F. Moser, R. Huang, A. T. Papageorghiou, B. W. Papież, and A. I. Namburete, "Automated Fetal Brain Extraction from Clinical Ultrasound Volumes Using 3D Convolutional Neural Networks," in *Communications in Computer and Information Science*, vol. 1065 CCIS, nov 2020, pp. 151–163. [Online]. Available: http://arxiv.org/abs/1911.07566

[14] C. A. Goodfellow lan, Bengio Yoshua, *Deep Learning Book*, 2016.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[16] A. Sorokin and D. Forsyth, "Utility data annotation with Amazon Mechanical Turk," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*, 2008.

[17] Google, "reCAPTCHA," 2021. [Online]. Available: https://www.google.com/recaptcha/about/?hl=el

[18] A. T. Papageorghiou, E. O. Ohuma, D. G. Altman, T. Todros, L. C. Ismail, A. Lambert, Y. A. Jaffer, E. Bertino, M. G. Gravett, M. Purwar, J. A. Noble, R. Pang, C. G. Victora, F. C. Barros, M. Carvalho, L. J. Salomon, Z. A. Bhutta, S. H. Kennedy, and J. Villar, "International standards for fetal growth based on serial ultrasound measurements: The Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project," *The Lancet*, vol. 384, no. 9946, pp. 869–879, 2014. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0140673614614902

[19] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI)," pp. 55–66, 2005.

[20] M. Bocchetta, M. Boccardi, R. Ganzola, L. G. Apostolova, G. Preboske, D. Wolf, C. Ferrari, P. Pasqualetti, N. Robitaille, S. Duchesne, C. R. Jack, G. B. Frisoni, G. Bartzokis, C. Decarli, L. Detoledo-Morrell, A. Fellgiebel, M. Firbank, L. Gerritsen, W. Henneman, R. J. Killiany, N. Malykhin, J. C. Pruessner, H. Soininen, and L. Wang, "Harmonized benchmark labels of the hippocampus on magnetic resonance: The EADC-ADNI project," *Alzheimer's and Dementia*, vol. 11, no. 2, pp. 151–160.e5, 2015.

[21] L. Venturini, A. T. Papageorghiou, J. A. Noble, and A. I. Namburete, "Multi-task CNN for Structural Semantic Segmentation in 3D Fetal Brain Ultrasound," *Communications in Computer and Information Science*, vol. 1065 CCIS, pp. 164–173, jul 2020. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-39343-4_14

[22] L. Venturini, A. T. Papageorghiou, J. A. Noble, and A. I. Namburete, "Uncertainty estimates as data selection criteria to boost omni-supervised learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12261 LNCS, pp. 689–698, oct 2020. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-59710-8_67

[23] A. M. Jukic, D. D. Baird, C. R. Weinberg, D. R. Mcconnaughey, and A. J. Wilcox, "Length of human pregnancy and contributors to its natural variation," *Human Reproduction*, vol. 28, no. 10, pp. 2848–2855, oct 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23922246

[24] J. Stiles and T. L. Jernigan, "The basics of brain development," *Neuropsychology Review*, vol. 20, no. 4, pp. 327–348, 2010.

[25] I. E. Timor-Tritsch, A. Monteagudo, and W. B. Warren, "Transvaginal ultrasonographic definition of the central nervous system in the first and early second trimesters," *American Journal of Obstetrics and Gynecology*, vol. 164, no. 2, pp. 497–503, feb 1991.

[26] P. Gressens, "Mechanisms and disturbances of neuronal migration," *Pediatric Research*, vol. 48, no. 6, pp. 725–730, dec 2000. [Online]. Available: http://dx.doi.org/10.1203/00006450-200012000-00004

[27] J. A. Scott, P. A. Habas, K. Kim, V. Rajagopalan, K. S. Hamzelou, J. M. Corbett-Detig, A. J. Barkovich, O. A. Glenn, and C. Studholme, "Growth trajectories of the human fetal brain tissues estimated from 3D reconstructed in utero MRI," *International Journal of Developmental Neuroscience*, vol. 29, no. 5, pp. 529–536, aug 2011.

[28] A. Toi, W. S. Lister, and K. W. Fong, "How early are fetal cerebral sulci visible at prenatal ultrasound and what is the normal pattern of early fetal sulcal development?" *Ultrasound in Obstetrics and Gynecology*, vol. 24, no. 7, pp. 706–715, 2004.

[29] L. R. Pistorius, P. Stoutenbeek, F. Groenendaal, L. De Vries, G. Manten, E. Mulder, and G. Visser, "Grade and symmetry of normal fetal cortical development: A longitudinal

two- and three-dimensional ultrasound study," *Ultrasound in Obstetrics and Gynecology*, vol. 36, no. 6, pp. 700–708, 2010.

[30] C. Garel, E. Chantrel, H. Brisse, M. Elmaleh, D. Luton, J. F. Oury, G. Sebag, and M. Hassan, "Fetal cerebral cortex: Normal gestational landmarks identified using prenatal MR imaging," *American Journal of Neuroradiology*, vol. 22, no. 1, pp. 184–189, 2001.

[31] J. G. Chi, E. C. Dooling, and F. H. Gilles, "Gyral development of the human brain," *Annals of Neurology*, vol. 1, no. 1, pp. 86–93, 1977.

[32] D. Paladini, G. Malinger, A. Monteagudo, G. Pilu, I. Timor-Tritsch, and A. Toi, "Sonographic examination of the fetal central nervous system: Guidelines for performing the 'basic examination' and the 'fetal neurosonogram'," *Ultrasound in Obstetrics and Gynecology*, vol. 29, no. 1, pp. 109–116, 2007.

[33] D. Prayer, G. Malinger, P. C. Brugger, C. Cassady, L. De Catte, B. De Keersmaecker, G. L. Fernandes, P. Glanc, L. F. Gonçalves, G. M. Gruber, S. Laifer-Narin, W. Lee, A. E. Millischer, M. Molho, J. Neelavalli, L. Platt, D. Pugash, P. Ramaekers, L. J. Salomon, M. Sanz, I. E. Timor-Tritsch, B. Tutschek, D. Twickler, M. Weber, R. Ximenes, and N. Raine-Fenning, "ISUOG Practice Guidelines: performance of fetal magnetic resonance imaging," *Ultrasound in Obstetrics and Gynecology*, vol. 49, no. 5, pp. 671–680, 2017. [Online]. Available: https://www.isuog.org/uploads/assets/uploaded/76547b04-fbdb-42fc-b0c6c01118299348.pdf

[34] Q. Huang and Z. Zeng, "A Review on Real-Time 3D Ultrasound Imaging Technology," *BioMed Research International*, vol. 2017, 2017.

[35] C. B. Burckhardt, "Speckle in Ultrasound B-Mode Scans," *IEEE Transactions on Sonics and Ultrasonics*, vol. 25, no. 1, pp. 1–6, jan 1978. [Online]. Available: http://ieeexplore.ieee.org/document/1539054/

[36] G. Timor-tritsh, Ilan Monteagudo, Ana Pilu, Giunluigi Malinger, *Ultrasonography and the prenatal brain*. McGraw-Hill Professional, 2012, no. 1.

[37] J. S. Dashe, D. D. McIntire, and D. M. Twickler, "Effect of maternal obesity on the ultrasound detection of anomalous fetuses," *Obstetrics and Gynecology*, vol. 113, no. 5, pp. 1001–1007, 2009.

[38] H. E. Melton and D. J. Skorton, "Rational gain compensation for attenuation in cardiac ultrasonography," *Ultrasonic Imaging*, vol. 5, no. 3, pp. 214–228, aug 1983. [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/016173468300500302

[39] A. Makropoulos, S. J. Counsell, and D. Rueckert, "A review on automatic fetal and neonatal brain MRI segmentation," *NeuroImage*, vol. 170, pp. 231–248, 2018.

[40] D. Prayer, P. C. Brugger, and L. Prayer, "Fetal MRI: Techniques and protocols," *Pediatric Radiology*, vol. 34, no. 9, pp. 685–693, jul 2004. [Online]. Available: https://link.springer.com/article/10.1007/s00247-004-1246-0

[41] Z. Zhang, S. Liu, X. Lin, G. Teng, T. Yu, F. Fang, and F. Zang, "Development of fetal brain of 20 weeks gestational age: Assessment with post-mortem Magnetic Resonance Imaging," *European Journal of Radiology*, vol. 80, no. 3, pp. e432–e439, dec 2011.

[42] I. S. Gousias, A. D. Edwards, M. A. Rutherford, S. J. Counsell, J. V. Hajnal, D. Rueckert, and A. Hammers, "Magnetic resonance imaging of the newborn brain: Manual segmentation of labelled atlases in term-born and preterm infants," *NeuroImage*, vol. 62, no. 3, pp. 1499–1509, sep 2012. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22713673/

[43] M. Kuklisova-Murgasova, P. Aljabar, L. Srinivasan, S. J. Counsell, V. Doria, A. Serag, I. S. Gousias, J. P. Boardman, M. A. Rutherford, A. D. Edwards, J. V. Hajnal, and D. Rueckert, "A dynamic 4D probabilistic atlas of the developing brain," *NeuroImage*, vol. 54, no. 4, pp. 2750–2763, 2011.

[44] C. Studholme, "Mapping fetal brain development in utero using magnetic resonance imaging: The big bang of brain mapping," *Annual Review of Biomedical Engineering*, vol. 13, no. 1, pp. 345–368, 2011. [Online]. Available: http://www.annualreviews.org/doi/10.1146/annurev-bioeng-071910-124654

[45] D. Prayer, G. Kasprian, E. Krampl, B. Ulm, L. Witzani, L. Prayer, and P. C. Brugger, "MRI of normal fetal brain development," *European Journal of Radiology*, vol. 57, no. 2, pp. 199–216, feb 2006. [Online]. Available: http://www.ejradiology.com/article/S0720048X05003852/fulltext

[46] J. W. Murakami, E. Weinberger, and D. W. Shaw, "Normal myelination of the pediatric brain imaged with fluid-attenuated inversion-recovery (FLAIR) MR imaging," *American Journal of Neuroradiology*, vol. 20, no. 8, pp. 1406–1411, 1999.

[47] T. Butts, M. J. Green, and R. J. Wingate, "Development of the cerebellum: Simple steps to make a 'little brain'," pp. 4031–4041, 2014.

[48] F. Triulzi, C. Parazzini, and A. Righini, "MRI of fetal and neonatal cerebellar development," pp. 411–420, 2005.

[49] NHS Screening Programmes, "Fetal Anomaly Screening: Programme handbook," pp. 1–38, 2015. [Online]. Available: https://www.gov.uk/government/publications/fetal-anomaly-screening-programme-handbook

[50] R. J. Kehl, M. A. Krew, A. Thomas, and P. M. Catalano, "Fetal growth and body composition in infants of women with diabetes mellitus during pregnancy," *Journal of Maternal-Fetal and Neonatal Medicine*, vol. 5, no. 5, pp. 273–280, 1996.

[51] C. Bamberg and K. D. Kalache, "Prenatal diagnosis of fetal growth restriction," pp. 387–394, 2004.

[52] E. Cohen, W. Baerts, and F. Van Bel, "Brain-Sparing in Intrauterine Growth Restriction: Considerations for the Neonatologist," *Neonatology*, vol. 108, no. 4, pp. 269–276, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26330337

[53] M. R. Chavez, C. V. Ananth, J. C. Smulian, and A. M. Vintzileos, "Fetal transcerebellar diameter measurement for prediction of gestational age at the extremes of fetal growth," *Journal of Ultrasound in Medicine*, vol. 26, no. 9, pp. 1167–1173, sep 2007. [Online]. Available: http://doi.wiley.com/10.7863/jum.2007.26.9.1167

[54] D. Johnston and D. G. Amaral, "Hippocampus," in *The Synaptic Organization of the Brain*. Oxford University Press, jan 2004. [Online]. Available: /record/2004-00210-011

[55] D. G. Mumby, S. Gaskin, M. J. Glenn, T. E. Schramek, and H. Lehmann, "Hippocampal damage and exploratory preferences in rats: Memory for objects, places, and contexts," *Learning and Memory*, vol. 9, no. 2, pp. 49–57, mar 2002. [Online]. Available: http://learnmem.cshlp.org/content/9/2/49

[56] I. Driscoll, D. A. Hamilton, H. Petropoulos, R. A. Yeo, W. M. Brooks, R. N. Baumgartner, and R. J. Sutherland, "The Aging Hippocampus: Cognitive, Biochemical and Structural Findings," *Cerebral Cortex*, vol. 13, no. 12, pp. 1344–1351, dec 2003. [Online]. Available: https://academic.oup.com/cercor/article/13/12/1344/384492

[57] R. C. Petersen, "Aging, mild cognitive impairment, and Alzheimer's disease," *Neurologic Clinics*, vol. 18, no. 4, pp. 789–805, 2000. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11072261/

[58] Alzheimer's Association, "2020 Alzheimer's disease facts and figures," *Alzheimer's and Dementia*, vol. 16, no. 3, pp. 391–460, 2020.

[59] Office for National Statistics, "Deaths registered in England and Wales. 2019 - Office for National Statistics," 2020. [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsregistrationsummarytables/2019#leading-causes-of-death

[60] D. J. Selkoe and J. Hardy, "The amyloid hypothesis of Alzheimer's disease at 25 years," *EMBO Molecular Medicine*, vol. 8, no. 6, pp. 595–608, 2016.

[61] S. Risacher, A. Saykin, J. Wes, L. Shen, H. Firpi, and B. McDonald, "Baseline MRI Predictors of Conversion from MCI to Probable AD in the ADNI Cohort," *Current Alzheimer Research*, vol. 6, no. 4, pp. 347–361, aug 2009.

[62] A. G. Vlassenko, M. A. Mintun, C. Xiong, Y. I. Sheline, A. M. Goate, T. L. Benzinger, and J. C. Morris, "Amyloid-beta plaque growth in cognitively normal adults: Longitudinal [11C]Pittsburgh compound B data," *Annals of Neurology*, vol. 70, no. 5, pp. 857–861, nov 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243969/

[63] J. A. Hardy and G. A. Higgins, "Alzheimer's disease: The amyloid cascade hypothesis," pp. 184–185, apr 1992.

[64] N. C. Fox, E. K. Warrington, P. A. Freeborough, P. Hartikainen, A. M. Kennedy, J. M. Stevens, and M. N. Rossor, "Presymptomatic hippocampal atrophy in Alzheimer's disease A longitudinal MRI study," *Brain*, vol. 119, no. 6, pp. 2001–2007, 1996.

[65] J. Anquez, E. D. Angelini, and I. Bloch, "Segmentation of fetal 3D ultrasound based on statistical prior and deformable model," in *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Proceedings, ISBI.* IEEE, 2008, pp. 17–20.

[66] B. Gutierrez Becker, F. Arambula Cosio, M. E. Guzman Huerta, and J. A. Benavides-Serralde, "Automatic segmentation of the cerebellum of fetuses on 3D ultrasound images, using a 3D Point Distribution Model." in *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference.* IEEE, 2010, pp. 4731–4734.

[67] S. Rueda, C. L. Knight, A. T. Papageorghiou, and J. Alison Noble, "Feature-based fuzzy connectedness segmentation of ultrasound images with an object completion step," *Medical Image Analysis*, vol. 26, no. 1, pp. 30–46, 2015.

[68] L. Tang and G. Hamarneh, "Medical image registration: A review," *Medical Imaging: Technology and Applications*, vol. 17, no. 2, pp. 619–660, 2013.

[69] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, "Medical image registration," *Physics in Medicine and Biology*, vol. 46, no. 3, pp. R1—-R45, mar 2001. [Online]. Available: http://stacks.iop.org/0031-9155/46/i=3/a=201?key=crossref.708a7842277435684a97102338567917

[70] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: A toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, jan 2010.

[71] M. Bro-Nielsen and C. Gramkow, "Fast fluid registration of medical images," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 1131. Springer, 1996, pp. 267–276.

[72] J. Ashburner, "A fast diffeomorphic image registration algorithm," *NeuroImage*, vol. 38, no. 1, pp. 95–113, 2007.

[73] M. Cabezas, A. Oliver, X. Lladó, J. Freixenet, and M. Bach Cuadra, "A review of atlas-based segmentation for magnetic resonance brain images," *Computer Methods and Programs in Biomedicine*, vol. 104, no. 3, pp. e158–e177, dec 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260711002033

[74] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," *NeuroImage*, vol. 33, no. 1, pp. 115–126, oct 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811906006458

[75] X. Geng, G. E. Christensen, H. Gu, T. J. Ross, and Y. Yang, "Implicit reference-based group-wise image registration and its application to structural and functional MRI," *NeuroImage*, vol. 47, no. 4, pp. 1341–1351, oct 2009. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/19371788/

[76] S. Joshi, B. Davis, M. Jomier, and G. Gerig, "Unbiased diffeomorphic atlas construction for computational anatomy," *NeuroImage*, vol. 23, no. SUPPL. 1, pp. S151–S160, jan 2004.

[77] C. Studholme, "Simultaneous population based image alignment for template free spatial normalisation of brain anatomy," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2717, pp. 81–90, 2003. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-39701-4_9

[78] P. A. Habas, K. Kim, F. Rousseau, O. A. Glenn, A. J. Barkovich, and C. Studholme, "Atlas-based segmentation of the germinal matrix from in utero clinical MRI of the fetal brain," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5241 LNCS, no. PART 1. Springer, 2008, pp. 351–358.

[79] P. A. Habas, K. Kim, F. Rousseau, O. A. Glenn, A. J. Barkovich, and C. Studholme, "A spatio-temporal atlas of the human fetal brain with application to tissue segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5761 LNCS, no. PART 1, pp. 289–296, 2009.

[80] P. A. Habas, J. A. Scott, A. Roosta, V. Rajagopalan, K. Kim, F. Rousseau, A. J. Barkovich, O. A. Glenn, and C. Studholme, "Early folding patterns and asymmetries of the normal human brain detected from in utero MRI," *Cerebral Cortex*, vol. 22, no. 1, pp. 13–25, 2012.

[81] A. Gholipour, A. Akhondi-Asl, J. A. Estroff, and S. K. Warfield, "Multi-atlas multi-shape segmentation of fetal brain MRI for volumetric and morphometric analysis of ventriculomegaly," *NeuroImage*, vol. 60, no. 3, pp. 1819–1831, 2012.

[82] M. K. Murgasova, A. Cifor, R. Napolitano, A. Papageorghiou, G. Quaghebeur, J. Alison Noble, and J. A. Schnabel, "Registration of 3D fetal brain US and MRI," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7511 LNCS, no. 8, pp. 667–674, 2012.

[83] W. Qiu, Y. Chen, J. Kishimoto, S. de Ribaupierre, B. Chiu, A. Fenster, and J. Yuan, "Automatic segmentation approach to extracting neonatal cerebral ventricles from 3D ultrasound images," *Medical Image Analysis*, vol. 35, pp. 181–191, apr 2017. [Online]. Available: http://dx.doi.org/10.1016/j.media.2016.06.038

[84] B. Kainz, K. Keraudren, V. Kyriakopoulou, M. Rutherford, J. V. Hajnal, and D. Rueckert, "Fast fully automatic brain detection in fetal MRI using dense rotation invariant image descriptors," in *2014 IEEE 11th International Symposium on Biomedical Imaging, ISBI 2014*, 2014, pp. 1230–1233.

[85] M. Yaqub, R. Cuingnet, R. Napolitano, D. Roundhill, A. Papageorghiou, R. Ardon, and J. A. Noble, "Volumetric segmentation of key fetal brain structures in 3D ultrasound," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial*

Intelligence and Lecture Notes in Bioinformatics), vol. 8184 LNCS. Springer, 2013, pp. 25–32.

[86] A. I. Namburete and J. A. Noble, "Fetal cranial segmentation in 2D ultrasound images using shape properties of pixel clusters," in *Proceedings - International Symposium on Biomedical Imaging*. IEEE, 2013, pp. 720–723.

[87] R. Cuingnet, O. Somphone, B. Mory, R. Prevost, M. Yaqub, R. Napolitano, A. Papageorghiou, D. Roundhill, J. A. Noble, and R. Ardon, "Where is my baby? A fast fetal head auto-alignment in 3D-ultrasound," in *Proceedings - International Symposium on Biomedical Imaging*. IEEE, apr 2013, pp. 768–771. [Online]. Available: http://ieeexplore.ieee.org/document/6556588/

[88] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998.

[89] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, sep 2015. [Online]. Available: https://arxiv.org/abs/1409.1556v6

[90] V. Nair and G. E. Hinton, "Rectified linear units improve Restricted Boltzmann machines," *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, pp. 807–814, 2010.

[91] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[92] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[93] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Communications of the ACM*, vol. 60, no. 6, 2017, pp. 84–90. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks

[94] W. Weng and X. Zhu, "INet: Convolutional Networks for Biomedical Image Segmentation," in *IEEE Access*, vol. 9. Springer, 2021, pp. 16 591–16 603.

[95] C. F. Baumgartner, L. M. Koch, M. Pollefeys, and E. Konukoglu, "An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10663 LNCS, pp. 111–119, 2018.

[96] J. Jiang, P. Trundle, and J. Ren, "Medical image analysis with artificial neural networks," *Computerized Medical Imaging and Graphics*, vol. 34, no. 8, pp. 617–631, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20713305

[97] A. De Brébisson and G. Montana, "Deep neural networks for anatomical brain segmentation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2015-Octob, 2015, pp. 20–28.

[98] E. Guerra, J. de Lara, A. Malizia, and P. Díaz, "Supporting user-oriented analysis for multi-view domain-specific visual languages," *Information and Software Technology*, vol. 51, no. 4, pp. 769–784, 2009.

[99] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9901 LNCS. Springer, 2016, pp. 424–432.

[100] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. De Vries, M. J. Benders, and I. Isgum, "Automatic Segmentation of MR Brain Images with a Convolutional Neural Network," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1252–1261, 2016.

[101] M. Rajchl, M. C. H. Lee, F. Schrans, A. Davidson, J. Passerat-Palmbach, G. Tarroni, A. Alansary, O. Oktay, B. Kainz, and D. Rueckert, "Learning under Distributed Weak Supervision," *arXiv preprint arXiv:1606.01100*, 2016. [Online]. Available: http://arxiv.org/abs/1606.01100

[102] S. S. M. Salehi, S. R. Hashemi, C. Velasco-Annis, A. Ouaalam, J. A. Estroff, D. Erdogmus, S. K. Warfield, and A. Gholipour, "Real-time automatic fetal brain extraction in fetal MRI by deep learning," *Proceedings - International Symposium on Biomedical Imaging*, vol. 2018-April, pp. 720–724, 2018.

[103] Y. Li, R. Xu, J. Ohya, and H. Iwata, "Automatic fetal body and amniotic fluid segmentation from fetal ultrasound images by encoder-decoder network with inner layers," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. IEEE, 2017, pp. 1485–1488.

[104] A. Schmidt-Richberg, T. Brosch, N. Schadewaldt, T. Klinder, A. Cavallaro, I. Salim, D. Roundhill, A. Papageorghiou, and C. Lorenz, "Abdomen segmentation in 3D fetal ultrasound using CNN-powered deformable models," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, 2017, vol. 10554 LNCS, pp. 52–61.

[105] L. Yu, Y. Guo, Y. Wang, J. Yu, and P. Chen, "Segmentation of fetal left ventricle in echocardiographic sequences based on dynamic convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1886–1895, 2017.

[106] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does It: Weakly supervised instance and semantic segmentation," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, 2017, pp. 1665–1674.

[107] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9911 LNCS, 2016, pp. 549–565.

[108] M. Rajchl, M. C. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M. A. Rutherford, J. V. Hajnal, B. Kainz, and D. Rueckert, "DeepCut: Object Segmentation from Bounding Box Annotations Using Convolutional Neural Networks," *IEEE Transactions on Medical Imaging*, vol. 36, no. 2, pp. 674–683, 2017.

[109] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham: Springer International Publishing, 2017, vol. 10435 LNCS, pp. 408–416. [Online]. Available: https://doi.org/10.1007/978-3-319-66179-7_47

[110] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham: Springer International Publishing, 2017, vol. 10435 LNCS, pp. 399–407. [Online]. Available: https://doi.org/10.1007/978-3-319-66179-7_46

[111] I. Radosavovic, P. Dollar, R. Girshick, G. Gkioxari, and K. He, "Data Distillation: Towards Omni-Supervised Learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4119–4128.

[112] L. R. Dice, "Measures of the Amount of Ecologic Association Between Species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[113] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing Images Using the Hausdorff Distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993. [Online]. Available: http://ieeexplore.ieee.org/document/232073/

[114] D. M. Hutton, "The Cross-Entropy Method: A Unified Approach to Combinatorial Optimisation, Monte-Carlo Simulation and Machine Learning," p. 903, 2005.

[115] T. G. Dietterich, "Ensemble methods in machine learning," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 1857 LNCS, 2000, pp. 1–15.

[116] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," mar 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

[117] R. E. Schapire, "The Strength of Weak Learnability," *Machine Learning*, vol. 5, no. 2, pp. 197–227, 1990.

[118] R. Huang, J. A. Noble, and A. I. Namburete, "Omni-supervised learning: Scaling up to large unlabelled medical datasets," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, Cham, sep 2018, vol. 11070 LNCS, pp. 572–580. [Online]. Available: http://link.springer.com/10.1007/978-3-030-00928-1_65

[119] J. S. Denker and Y. LeCun, "Transforming Neural-Net Output Levels to Probability Distributions," *Advances in Neural Information Processing Systems 3*, pp. 853–859, 1991. [Online]. Available: https://proceedings.neurips.cc/paper/1990/file/7eacb532570ff6858afd2723755ff790-Paper.pdf

[120] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems*, vol. 2017-Decem, 2017, pp. 5575–5585.

[121] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *33rd International Conference on Machine Learning, ICML 2016*, vol. 3, 2016, pp. 1651–1660.

[122] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," jul 2012. [Online]. Available: http://arxiv.org/abs/1207.0580

[123] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *British Machine Vision Conference 2017, BMVC 2017*. British Machine Vision Association, 2017. [Online]. Available: http://www.bmva.org/bmvc/2017/papers/paper057/index.html

[124] S. A. A. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. H. Maier-Hein, S. M. A. Eslami, D. Jimenez Rezende, O. Ronneberger, S. M. Ali Eslami, D. J. Rezende, and O. Ronneberger, "A probabilistic U-net for segmentation of ambiguous images," in *Advances in Neural Information Processing Systems*, vol. 2018-Decem, 2018, pp. 6965–6975. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/473447ac58e1cd7e96172575f48dca3b-Paper.pdf

[125] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, apr 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219301961

[126] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph Cnn for learning on point clouds," *ACM Transactions on Graphics*, vol. 38, no. 5, p. 13, jan 2019. [Online]. Available: https://arxiv.org/abs/1801.07829v2

[127] J. Yap, W. Yolland, and P. Tschandl, "Multimodal skin lesion classification using deep learning," *Experimental Dermatology*, vol. 27, no. 11, pp. 1261–1267, nov 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/exd.13777

[128] M. M. Ahsan, T. E. Alam, T. Trafalis, and P. Huebner, "Deep MLP-CNN model using mixed-data to distinguish between COVID-19 and Non-COVID-19 patients," *Symmetry*, vol. 12, no. 9, p. 1526, sep 2020. [Online]. Available: https://www.mdpi.com/2073-8994/12/9/1526

[129] L. Drukker, R. Droste, P. Chatelain, J. A. Noble, and A. T. Papageorghiou, "Expected-value bias in routine third-trimester growth scans," *Ultrasound in Obstetrics*

and Gynecology, vol. 55, no. 3, pp. 375–382, mar 2020. [Online]. Available: https://obgyn.onlinelibrary.wiley.com/doi/10.1002/uog.21929

[130] A. Sotiriadis and A. O. Odibo, "Systematic error and cognitive bias in obstetric ultrasound," pp. 431–435, apr 2019. [Online]. Available: https://obgyn.onlinelibrary.wiley.com/doi/10.1002/uog.20232

[131] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski, "An intriguing failing of convolutional neural networks and the CoordConv solution," in *Advances in Neural Information Processing Systems*, vol. 2018-Decem. Neural information processing systems foundation, jul 2018, pp. 9605–9616. [Online]. Available: https://arxiv.org/abs/1807.03247v2

[132] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, and J. Marchini, "The UK Biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, no. 7726, pp. 203–209, 2018.

[133] L. Cheikh Ismail, H. E. Knight, Z. Bhutta, and W. C. Chumlea, "Anthropometric protocols for the construction of new international fetal and newborn growth standards: The INTERGROWTH-21st Project," *BJOG: An International Journal of Obstetrics and Gynaecology*, vol. 120, no. SUPPL. 2, pp. 42–47, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/1471-0528.12125

[134] A. T. Papageorghiou, I. Sarris, C. Ioannou, T. Todros, M. Carvalho, G. Pilu, and L. J. Salomon, "Ultrasound methodology used to construct the fetal growth standards in the INTERGROWTH-21st Project," *BJOG: An International Journal of Obstetrics and Gynaecology*, vol. 120, no. SUPPL. 2, pp. 27–32, 2013.

[135] I. Sarris, C. Ioannou, E. O. Ohuma, D. G. Altman, L. Hoch, C. Cosgrove, S. Fathima, L. J. Salomon, and A. T. Papageorghiou, "Standardisation and quality control of ultrasound measurements taken in the INTERGROWTH-21st Project," *BJOG: An International Journal of Obstetrics and Gynaecology*, vol. 120, no. SUPPL. 2, pp. 33–37, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/1471-0528.12315

[136] K. Matsuo, K. Shimoya, N. Ushioda, and T. Kimura, "Maternal positioning and fetal positioning in utero," *Journal of Obstetrics and Gynaecology Research*, vol. 33, no. 3, pp. 279–282, jun 2007. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1447-0756.2007.00524.x

[137] A. I. L. Namburete, R. V. Stebbing, and J. A. Noble, "Cranial parametrization of the fetal head for 3D ultrasound image analysis," *Medical Image Understanding and Analysis (MIUA)*, pp. 196–201, 2013. [Online]. Available: files/183/Namburete-CranialParametrizationoftheFetalHeadfor3DU.pdf

[138] C. R. Jack, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward, A. M. Dale, J. P. Felmlee, J. L. Gunter, D. L. Hill, R. Killiany, N. Schuff, S. Fox-Bosetti, C. Lin, C. Studholme, C. S. DeCarli, G. Krueger, H. A. Ward, G. J. Metzger, K. T. Scott, R. Mallozzi, D. Blezek,

J. Levy, J. P. Debbins, A. S. Fleisher, M. Albert, R. Green, G. Bartzokis, G. Glover, J. Mugler, and M. W. Weiner, "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods," pp. 685–691, 2008.

[139] B. T. Wyman, D. J. Harvey, K. Crawford, M. A. Bernstein, O. Carmichael, P. E. Cole, P. K. Crane, C. Decarli, N. C. Fox, J. L. Gunter, D. Hill, R. J. Killiany, C. Pachai, A. J. Schwarz, N. Schuff, M. L. Senjem, J. Suhy, P. M. Thompson, M. Weiner, and C. R. Jack, "Standardization of analysis sets for reporting results from ADNI MRI data," *Alzheimer's and Dementia*, vol. 9, no. 3, pp. 332–337, may 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1016/j.jalz.2012.06.004

[140] J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, C. Holmes, L. Collins, P. Thompson, D. MacDonald, M. Iacoboni, T. Schormann, K. Amunts, N. Palomero-Gallagher, S. Geyer, L. Parsons, K. Narr, N. Kabani, G. Le Goualher, D. Boomsma, T. Cannon, R. Kawashima, and B. Mazoyer, "A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM)," pp. 1293–1322, 2001.

[141] A. Gholipour, J. A. Estroff, and S. K. Warfield, "Robust super-resolution volume reconstruction from slice acquisitions: Application to fetal brain MRI," *IEEE Transactions on Medical Imaging*, vol. 29, no. 10, pp. 1739–1758, 2010.

[142] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic demons: efficient non-parametric image registration." *NeuroImage*, vol. 45, no. 1 Suppl, pp. S61–S72, mar 2009.

[143] National Health Service, "Mid-pregnancy anomaly scan," pp. https://www.nhs.uk/conditions/pregnancy–and–baby/a, 2018. [Online]. Available: https://www.nhs.uk/conditions/pregnancy-and-baby/anomaly-scan-18-19-20-21-weeks-pregnant/

[144] L. J. Salomon, Z. Alfirevic, V. Berghella, C. Bilardo, E. Hernandez-Andrade, S. L. Johnsen, K. Kalache, K. Y. Leung, G. Malinger, H. Munoz, F. Prefumo, A. Toi, and W. Lee, "Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan," *Ultrasound in Obstetrics and Gynecology*, vol. 37, no. 1, pp. 116–126, dec 2011. [Online]. Available: https://doi.org/10.1002/uog.8831

[145] L. J. Salomon, Z. Alfirevic, F. Da Silva Costa, R. L. Deter, F. Figueras, T. Ghi, P. Glanc, A. Khalil, W. Lee, R. Napolitano, A. Papageorghiou, A. Sotiradis, J. Stirnemann, A. Toi, and G. Yeo, "ISUOG Practice Guidelines: ultrasound assessment of fetal biometry and growth," *Ultrasound in Obstetrics and Gynecology*, vol. 53, no. 6, pp. 715–723, jun 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31169958/

[146] A. Guha Roy, S. Conjeti, N. Navab, and C. Wachinger, "QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy," *NeuroImage*, vol. 186, pp. 713–727, feb 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811918321232

[147] A. G. Roy, S. Conjeti, D. Sheet, A. Katouzian, N. Navab, and C. Wachinger, "Error corrective boosting for learning fully convolutional networks with limited

data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10435 LNCS. Springer, Cham, sep 2017, pp. 231–239. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-66179-7_27

[148] B. Fischl, "FreeSurfer," *NeuroImage*, vol. 62, no. 2, pp. 774–781, aug 2012.

[149] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A Modern Library for 3D Data Processing," 2018. [Online]. Available: http://arxiv.org/abs/1801.09847

[150] A. Beers, K. Chang, J. Brown, E. Sartor, C. Mammen, E. Gerstner, B. Rosen, and J. Kalpathy-Cramer, "Sequential 3D U-Nets for Biologically-Informed Brain Tumor Segmentation," *arXiv preprint arXiv:1709.02967*, 2017. [Online]. Available: http://arxiv.org/abs/1709.02967

[151] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, pp. 565–571, jun 2016. [Online]. Available: http://arxiv.org/abs/1606.04797

[152] F. Chollet and Others, "Keras," 2015.

[153] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*, 2016, pp. 265–283.

[154] A. S. Vinkesteijn, P. G. Mulder, and J. W. Wladimiroff, "Fetal transverse cerebellar diameter measurements in normal and reduced fetal growth," *Ultrasound in Obstetrics and Gynecology*, vol. 15, no. 1, pp. 47–51, jan 2000. [Online]. Available: http://doi.wiley.com/10.1046/j.1469-0705.2000.00024.x

[155] Y. Nakae, N. Goto, and T. Nara, "Development of Human Fetus Brain. Volumetric Analysis of the Component Parts," *No To Hattatsu*, vol. 21, no. 5, pp. 434–439, sep 1989. [Online]. Available: https://europepmc.org/article/med/2803794

[156] C. Clouchoux, N. Guizard, A. C. Evans, A. J. Du Plessis, and C. Limperopoulos, "Normative fetal brain growth by quantitative in vivo magnetic resonance imaging," *American Journal of Obstetrics and Gynecology*, vol. 206, no. 2, pp. 173.e1–173.e8, feb 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0002937811012701

[157] I. Wolf, M. Vetter, I. Wegner, T. Böttger, M. Nolden, M. Schöbinger, M. Hastenteufel, T. Kunert, and H. P. Meinzer, "The medical imaging interaction toolkit," *Medical Image Analysis*, vol. 9, no. 6, pp. 594–604, 2005.

[158] P. Jaccard, "Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines," *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 241–272, 1901. [Online]. Available: https://ci.nii.ac.jp/naid/10027880482/

[159] F. Shi, B. Liu, Y. Zhou, C. Yu, and T. Jiang, "Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer's disease: Meta-analyses of MRI studies," pp. 1055–1064, 2009.

[160] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney, R. K. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. De Stefano, J. M. Brady, and P. M. Matthews, "Advances in functional and structural MR image analysis and implementation as FSL," in *NeuroImage*, vol. 23, no. SUPPL. 1, 2004.

[161] W. G. Cochran, "The distribution of quadratic forms in a normal system, with applications to the analysis of covariance," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 30, no. 2, pp. 178–191, 1934.

[162] N. K. Dinsdale, M. Jenkinson, and A. I. Namburete, "Spatial Warping Network for 3D Segmentation of the Hippocampus in MR Images," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11766 LNCS, 2019, pp. 284–291.

[163] R. E. Rosenberg, A. S. U. Ahmed, S. Ahmed, S. K. Saha, M. A. Chowdhury, R. E. Black, M. Santosham, and G. L. Darmstadt, "Determining gestational age in a low-resource setting: Validity of last menstrual period," *Journal of Health, Population and Nutrition*, vol. 27, no. 3, pp. 332–338, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2761790/

[164] R. H. Van Oppenraaij, P. H. Eilers, S. P. Willemsen, F. M. Van Dunné, N. Exalto, and E. A. Steegers, "Determinants of number-specific recall error of last menstrual period: A retrospective cohort study," *BJOG: An International Journal of Obstetrics and Gynaecology*, vol. 122, no. 6, pp. 835–841, may 2015. [Online]. Available: https://obgyn.onlinelibrary.wiley.com/doi/10.1111/1471-0528.12991

[165] W. A. Rocca, A. Hofman, C. Brayne, M. M. Breteler, M. Clarke, J. R. Copeland, J. â. Dartigues, K. Engedal, O. Hagnell, T. J. Heeren, C. Jonker, J. Lindesay, A. Lobo, A. H. Mann, P. K. Mölsä, K. Morgan, D. W. O'Connor, A. d. S. Droux, R. Sulkava, D. W. Kay, and L. Amaducci, "Frequency and distribution of Alzheimer's disease in Europe: A collaborative study of 1980-1990 prevalence findings," *Annals of Neurology*, vol. 30, no. 3, pp. 381–390, sep 1991. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/ana.410300310

[166] G. Hinton and S. Roweis, "Stochastic neighbor embedding," *Advances in Neural Information Processing Systems*, 2003.

[167] J. Golomb, M. J. Leon, A. Kluger, C. Tarshish, S. H. Ferris, and A. E. George, "Hippocampal atrophy in normal aging: An association with recent memory impairment," *Archives of Neurology*, vol. 50, no. 9, pp. 967–973, sep 1993. [Online]. Available: https://jamanetwork.com/journals/jamaneurology/fullarticle/592529

[168] M. J. De Leon, J. Golomb, A. E. George, A. Convit, C. Y. Tarshish, T. McRae, S. De Santi, G. Smith, S. H. Ferris, M. Noz, and H. Rusinek, "The radiologic prediction

of Alzheimer disease: The atrophic hippocampal formation," *American Journal of Neuroradiology*, vol. 14, no. 4, pp. 897–906, 1993.

[169] K. Zilles, E. Armstrong, A. Schleicher, and H. J. Kretschmann, "The human pattern of gyrification in the cerebral cortex," *Anatomy and Embryology*, vol. 179, no. 2, pp. 173–179, 1988.

[170] A. Amyar, R. Modzelewski, H. Li, and S. Ruan, "Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation," *Computers in Biology and Medicine*, vol. 126, p. 104037, nov 2020.

[171] M. L. di Scandalea, C. S. Perone, M. Boudreau, and J. Cohen-Adad, "Deep Active Learning for Axon-Myelin Segmentation on Histology Data," jul 2019. [Online]. Available: http://arxiv.org/abs/1907.05143

[172] M. Gorriz, A. Carlier, E. Faure, and X. Giro-i Nieto, "Cost-Effective Active Learning for Melanoma Segmentation," nov 2017. [Online]. Available: http://arxiv.org/abs/1711.09168

[173] P. Seebock, J. I. Orlando, T. Schlegl, S. M. Waldstein, H. Bogunovic, S. Klimscha, G. Langs, and U. Schmidt-Erfurth, "Exploiting Epistemic Uncertainty of Anatomy Segmentation for Anomaly Detection in Retinal OCT," *IEEE Transactions on Medical Imaging*, vol. 39, no. 1, pp. 87–98, jan 2020.