Fully-Automated Deep Learning Pipeline for 3D Fetal Brain Ultrasound

Felipe Andres Moser

University College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hilary 2023

Abstract

Three-dimensional ultrasound (3D US) imaging has shown significant potential for in-utero assessment of the development of the fetal brain. However, in spite of the potential benefits of this modality over its two-dimensional (2D) counterpart, its widespread adoption remains largely limited by the difficulty associated with its analysis.

While more established 3D neuroimaging modalities, such as Magnetic Resonance Imaging (MRI), have circumvented similar challenges thanks to reliable, automated neuroimage analysis pipelines, there is currently no comparable pipeline solution for 3D neurosonography.

With the goal of facilitating medical research and encouraging the adoption of 3D US for clinical assessment, the main objective of my doctoral thesis is to design, develop, and validate a set of fundamental automated modules that comprise a fast, robust, fully automated, general-purpose pipeline for the neuroimage analysis of fetal 3D US scans.

For the first module, I propose the fetal Brain Extraction Network (fBEN), a fully-automated, end-to-end 3D Convolutional Neural Network (CNN) with an encoder-decoder architecture. It predicts an accurate binary brain mask for the automated extraction of the fetal brain from standard clinical 3D US scans.

For the second module I propose the fetal Brain Alignment Network (fBAN), a fully-automated, end-to-end regression network with a cascade architecture that accurately predicts the alignment parameters required to rigidly align standard clinical 3D US scans to a canonical reference space.

Finally, for the third module, I propose the fetal Brain Fingerprinting Network (fBFN), a fully-automated, end-to-end network based on a Variational AutoEncoder (VAE) architecture, that encodes the entire structural information of the 3D brain into a relatively small set of parameters in a continuously distributed latent space. It is a general-purpose solution aimed at facilitating the assessment of the 3D US scans by recharacterising the fetal brain into a representation that is easier to analyse.

After exhaustive analysis, each module of this pipeline has proven to achieve state-of-the-art performance that is consistent across a wide gestational range, as well as robust to image quality, while requiring minimal pre-processing. Additionally, this pipeline has been designed to be modular, and easy to modify and expand upon, with the purpose of making it as easy as possible for other researchers to develop new tools and adapt it to their needs. This combination of performance, flexibility, and ease of use may have the potential to help 3D US become the preferred imaging modality for researching and assessing fetal development.

Fully-Automated Deep Learning Pipeline for 3D Fetal Brain Ultrasound

Felipe Andres Moser

University College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2023

To everyone who have lost their brains in the fog. May you all find it one day, as I hope to one day find mine.

Acknowledgements

Personal

First, I would like to thank my supervisors, Prof. Ana Namburete and Prof. Bartłomiej Papież for their support and guidance throughout my DPhil. Without you, the work presented in this thesis would simply no have been possible. I would like to especially thank you both for the patience and understanding you have shown to me while I struggled to juggle the health issues of these last couple of years.

I would also like to thank everybody that has been a part of the OMNI group. From those that were part of it from before the name OMNI existed, like Lorenzo, Nora, Marianne, Izzy, Nicola, and Jielai, to those like Hoda that have just joined: thank you very much for being a part of this journey and helping make it unique in your own weird little ways. Thank you Maddy and Linde for sharing endless hours of distractions when focusing on actual work seemed impossible, and for all those bouldering sessions we shared. Thank you Ruobing for teaching me everything when I had just joined and had no clue what to do. And an especial thank you to Hugo. No only for countless quick chats that inevitably became multi-hour discussions about the absolute nerdiest things, but for the incredible support and encouragement you have given me during the most difficult times of this DPhil. This thesis would just have never been finished without you.

I would also like to thank Prof. Peter Jezzard and the entirety of ONBI. That first year was one of the most fun times of the entire DPhil. Especially you, Zelekha, thank you for becoming such an important part of my life, then and now.

Thank you to Alexandra Brown. Not only was your help essential to get me to focus and get things done, but you were also an invaluable source of comfort and positivity when things were tough.

Finally, thank you Jon for being an immovable part of my life, form before I ever even thought of starting this DPhil. Your friendship was one of the very few constants throughout these crazy few years. I guess fetching all that liquid nitrogen for you was worth it in the end.

Institutional

I have been supported throughout my DPhil thanks to the support and funding from the Engineering and Physical Sciences Research Council (EPSRC) and Medical

Research Council (MRC) (EP/L016052/1), as well as the support from University College Oxford and its Oxford-Radcliffe benefaction.

Abstract

Three-dimensional ultrasound (3D US) imaging has shown significant potential for in-utero assessment of the development of the fetal brain. However, in spite of the potential benefits of this modality over its two-dimensional (2D) counterpart, its widespread adoption remains largely limited by the difficulty associated with its analysis.

While more established 3D neuroimaging modalities, such as Magnetic Resonance Imaging (MRI), have circumvented similar challenges thanks to reliable, automated neuroimage analysis pipelines, there is currently no comparable pipeline solution for 3D neurosonography.

With the goal of facilitating medical research and encouraging the adoption of 3D US for clinical assessment, the main objective of my doctoral thesis is to design, develop, and validate a set of fundamental automated modules that comprise a fast, robust, fully automated, general-purpose pipeline for the neuroimage analysis of fetal 3D US scans.

For the first module, I propose the fetal Brain Extraction Network (fBEN), a fully-automated, end-to-end 3D Convolutional Neural Network (CNN) with an encoder-decoder architecture. It predicts an accurate binary brain mask for the automated extraction of the fetal brain from standard clinical 3D US scans.

For the second module I propose the fetal Brain Alignment Network (fBAN), a fully-automated, end-to-end regression network with a cascade architecture that accurately predicts the alignment parameters required to rigidly align standard clinical 3D US scans to a canonical reference space.

Finally, for the third module, I propose the fetal Brain Fingerprinting Network (fBFN), a fully-automated, end-to-end network based on a Variational AutoEncoder (VAE) architecture, that encodes the entire structural information of the 3D brain into a relatively small set of parameters in a continuously distributed latent space. It is a general-purpose solution aimed at facilitating the assessment of the 3D US scans by recharacterising the fetal brain into a representation that is easier to analyse.

After exhaustive analysis, each module of this pipeline has proven to achieve state-of-the-art performance that is consistent across a wide gestational range, as well as robust to image quality, while requiring minimal pre-processing. Additionally, this pipeline has been designed to be modular, and easy to modify and expand upon,

with the purpose of making it as easy as possible for other researchers to develop new tools and adapt it to their needs. This combination of performance, flexibility, and ease of use may have the potential to help 3D US become the preferred imaging modality for researching and assessing fetal development.

Contents

List of Figures

List of Tables

List of Abbreviations

2D, 3D . . . . Two- or three-dimensional

AC . . . . . Abdominal Circumference

AE . . . . . AutoEncoder

AGA . . . . . Appropriate for Gestational Age

AI . . . . . Artificial Intelligence

BCE . . . . . Binary Cross-Entropy

BPD . . . . . Biparietal Diameter

CD . . . . . Centroid Distance

CER . . . . . Cerebellum

CM . . . . . Cisterna Magna

CNN . . . . . Convolutional Neural Network

CP . . . . . Choroid Plexus

CRL . . . . . Crown-Rump Length

CSP . . . . . Cavum Septum Pellucidum

CV . . . . . Computer Vision

DL . . . . . Deep Learning

DSC . . . . . Dice Similarity Coefficient

EFW . . . . . Estimated Fetal Weight

fBEN . . . . . Fetal Brain Extraction Network

fBAN . . . . . Fetal Brain Alignment Network

fBFN . . . . . Fetal Brain Fingerprinting Network

FH . . . . . Frontal Horn

FL . . . . . Femur Diaphysis Length

FN . . . . . False Negative

FP . . . . . False Positive

FSIM . . . . . Feature Similarity Index

GUI . . . . . Graphical User Interface

GW . . . . . Gestational Weeks

HC . . . . . Head Circumference

HD . . . . . Hausdorff Distance

IAM . . . . . Incidence Angle Map

ISUOG . . . . International Society of Ultrasound in Obstetrics and Gynecology

KLD . . . . . Kullback-Leibler Divergence

LGA . . . . . Large for Gestational Age

LMP . . . . . Last Menstrual Period

LV . . . . . Lateral Ventricle

LSTM . . . . . Long Short-Term Memory

MAE . . . . . Mean Absolute Error

ML . . . . . Machine Learning

MRI . . . . . Magnetic Resonance Imaging

MSE . . . . . Mean Squared Error

OFD . . . . . Occipitofrontal Distance

OH . . . . . Occipital Horn

PCA . . . . . Principal Component Analysis

PSNR . . . . . Peak Signal-to-Noise Ratio

PHE . . . . . Public Health England

RF . . . . . Random Forest

SC . . . . . Symmetry Coefficient

SCM . . . . . Shadow Casting Map

SF . . . . . Sylvian Fissure

SGA . . . . . Small for Gestational Age

SIFT . . . . . Scale-Invariant Feature Transform

SL . . . . . Spatial Landmarks

SSFSE . . . . . Single-Shot Fast Spin Echo

SSIM . . . . . Structural Similarity Index Measure

SVM . . . . . Support Vector Machine

TN . . . . . True Negative

TP . . . . . True Positive

US . . . . . Ultrasound

VAE . . . . . Variational AutoEncoder

VOI . . . . . Volume Of Interest

vxl . . . . . Voxel

1 Introduction

Contents

1.1 Motivation

After an initial embryonic period of 8 gestational weeks (GW) [1], the brain starts its fetal development period as a smooth, i.e. lissencephalic structure. From here, the brain undergoes a folding sequence in a process called gyrification, which results in the formation of ridges and grooves, also known as gyri, and sulci or fissures. These folds occur at predictable points in the gestational period [2] and deviations from this developmental timeline can be related to brain abnormalities such as ventriculomegaly and lissencephaly [2][3]. Therefore, a thorough and accurate clinical assessment of the development of the fetal brain during gestation is crucial.

The primary modality for the assessment of the fetal brain development is two-dimensional ultrasound (2D US), the use of which has been standard clinical practice since the 1970s [4]. This imaging modality allows for the cross-sectional, in-utero view of the fetal brain throughout gestation. In order to ensure a thorough and consistent assessment, standard practice guidelines for sonographic examination of the fetal central nervous system have been proposed by relevant entities such as the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) [3][5] and Public Health England (PHE) [6]. While these guidelines might differ in aspects such as the number of recommended screenings during the pregnancy, the general assessment procedure remains consistent. It comprises of a qualitative and a quantitative assessment of 2D US scans of three standard cross-sectional planes of the fetal brain planes: transthalamic, transventricular, and transcerebellar. The assessment focuses on the head shape and size, the brain texture, and the analysis of four relevant structures as proxy indicators of the integrity of the ventricular system and potential developmental abnormalities: the Lateral Ventricles (LV) (Frontal Horn (FH), Occipital Horn (OH), and Choroid Plexus(CP)), the Cavum Septum Pellucidum (CSP), the Cerebellum (CER), and the Cisterna Magna [3][7][8][9][10][11][12][13][14][15][16][17][18][19].

However, while the 2D US assessment of the fetal brain offers a relatively straightforward and standardised solution for the challenging task of in-utero clinical assessment of fetal brain development, its simplicity is also the source of several critical limitations. Firstly, the reliability of the measurements depends on the accurate acquisition of the standard brain planes by the sonographer, as well as the accurate positioning of the calipers, which is not trivial [20]. This results in a significant amount of inter- and intra-user variability of the measurements [20]. Secondly, the assessment of cerebral structures is achieved by measuring their diameter to represent their size [3], which is a strong oversimplification of their structural characteristics, particularly when considering the continuous state of morphological development the brain undergoes during gestation. Moreover, which biometric measurements are considered abnormal can vary greatly depending on the

reference chart used [21]. Lastly, and perhaps the most significant limitation, is the assessment of the three standard planes as a proxy for the development of the entire brain. This results in the vast majority of the available structural information of the brain being discarded, severely limiting the amount of clinical evidence on which the assessment is performed. Consequently, potential developmental abnormalities that fall outside the three standard planes, such as abnormalities in cerebral growth in infants with congenital heart disease [22][23], are at risk of being overlooked.

In contrast to the 2D US approach, three-dimensional (3D) US allows for the entirety of the fetal brain to be captured in a single scan. If so desired, this can be used to accurately extract the three standard planes after acquisition for the 2D approach just described, greatly reducing the difficulty of the task. However, the main benefit of 3D US is that it allows for the possibility of a complete assessment of the cerebral development to be performed based on the entire structural information available, rather than relying on proxy planes, facilitating a deeper understanding of the developmental process of the fetal brain during gestation. This potential has been highlighted by the recent recommendation of 3D US by ISUOG in their 2021 updated guidelines [5].

In spite of the potential benefits of a 3D-based assessment of in-utero cerebral development, the adoption of this modality over its 2D counterpart has been slow. While slightly higher costs are partly to blame, the main problem is the difficulty of analysing 3D US brain data and the consequent limited perceived benefits [24]. Without robust, automated solutions for the processing and analysis of these scans, the same user-dependent variability issues endure, but the tasks become considerably more difficult. Placing an ellipsoid on the skull to measure the head-circumference becomes extracting the 3D brain (masking of extra-cranial tissue), locating the standard brain planes becomes aligning the fetal brain to a canonical position in 3D space. Since the subsequent neuroimage analysis relies directly on the consistency and accuracy of these two fundamental processing steps, reliable automated solutions for the extraction and alignment of the fetal brain from 3D US scans are crucial.

Furthermore, without automated tools, any subsequent analysis will be limited to a slice-by-slice approach, since a sonographer can only analyse one slice at a time.

More established 3D neuroimaging modalities, such as Magnetic Resonance Imaging (MRI), have circumvented these issues through the use of reliable neuroimage analysis pipelines [25][26] that have been made possible thanks to the development of fast, and robust libraries and toolboxes, such as FSL [27] and BrainVISA [28]. By automating difficult, fundamental tasks such as brain extraction, alignment, and segmentation, these pipelines have facilitated clinical assessment, medical research, and the further development of automated tools [29][30][31]. However, since MRI and US images are acquired through entirely different physical interactions, there are fundamental differences in the structural information rendered, as well as other intensity properties, such as contrast and sharpness. Therefore, the implementation of these pipelines for 3D US is non-trivial and the lack of one-to-one intensity mapping between US and MRI would potentiate the need for major modifications to most intensity-based approaches. Therefore, an equivalent, 3D US-native pipeline would be instrumental in the widespread adoption of 3D US for the analysis of fetal brain development.

While there is a growing number of works aimed at automating 3D US tasks, such as anatomy-based gestational age prediction [32][33][34] or tissue segmentation [35][36][37][38][39][40], there is currently no comparable pipeline solution for 3D neurosonography. This can be attributed, in part, to the intrinsic challenges associated with this imaging modality, as well as the effect that temporal changes of the fetal head have on image quality. Specifically, the growth and morphological changes that the brain undergoes during gestation [1][41] result in a large intrinsic variability in the scans of the fetal cohort. The increasing ossification of the skull during gestation affects the interaction of the tissues with the US beam. As shown in Fig. 1.1, this results in a variation of shadows, occlusions, and reverberation artefacts throughout pregnancy, which vary depending on the relative orientation and position of the US probe in relation to the skull [42]. Furthermore, due to the acoustic cavity created by the increasingly calcified skull, the hemisphere of the

brain that is located closest to the US probe has most of its structural information obscured [43][32], resulting in an intrinsically asymmetric representation of the brain. All these challenges make the development of a reliable pipeline difficult.

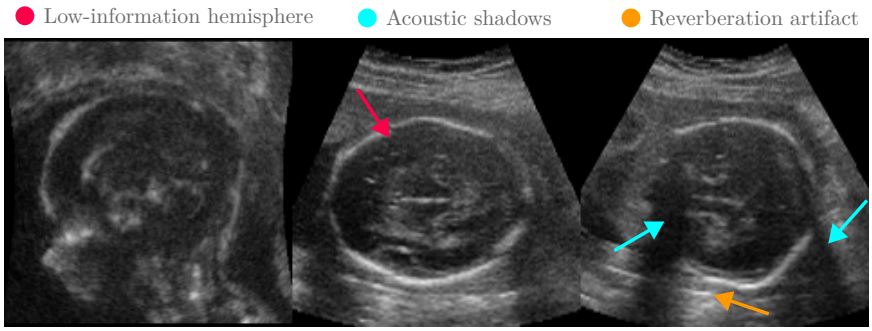

Figure 1.1: Orthogonal midplanes of an example 3D US scan at 26 GW.

Therefore, there is currently a great need for a robust, automated neuroimage analysis pipeline for fetal 3D neurosonograms. Its absence makes medical research of 3D US more difficult, stagnating its progress. The lack of clinical adoption limits available datasets, and makes new discoveries hard for sonographers to implement in their assessment. Simultaneously, without significant medical research to highlight the benefits of 3D neurosonography and a pipeline that can facilitate its analysis, there is limited incentive for clinicians to use this modality as part of the clinical examination. Furthermore, the potential adoption momentum generated by the recent endorsement of this imaging modality by ISUOG [5] could be wasted if the right tools for analysis are not available to make its use compelling. Therefore, the need for such a pipeline is not only great, but urgent.

The main objective of my doctoral research is to design, develop, and validate a set of fundamental automated tools that comprise a fast, robust, fully automated, general-purpose pipeline for the neuroimage analysis of fetal 3D US scans, aiming to facilitate medical research and encouraging the adoption of this modality for clinical assessment. This pipeline is designed to be modular, and easy to modify and expand upon, with the purpose of making it as easy as possible for other researchers to develop new tools and adapt it to their needs. The fundamental tools that comprise the pipeline have been specifically designed to address the

main challenges associated with 3D fetal neurosonography, i.e. the high intrinsic variability of the data and the difficult nature of analysing the whole 3D brain. To address the former, I have first developed fully automated solutions for the tasks of fetal brain extraction and alignment from minimally pre-processed 3D US scans. These tools remove the extra-cerebral data from the scans and align the brain to a canonical reference space, significantly reducing the difficulty of subsequent analysis. As a result, brain extraction and alignment are the fundamental first steps in most neuroimage analysis pipelines. To address the challenging task of analysing the fetal 3D brain I have developed a fully-automated *fingerprinting* solution that encodes the entire structural information of the brain into a relatively small set of parameters in a continuously distributed latent space. This new characterisation of the structural information of the brain aims to provide a general-purpose, condensed structural representation on which subsequent analysis can be performed, facilitating potential novel approaches to study and assess the development of the fetal brain.

However, in order for this pipeline to facilitate the widespread adoption of 3D US, its use must represent a meaningful advantage for the research and clinical community as a whole. After all, the performance of the pipeline is meaningless if nobody uses it. Therefore, to ensure that the pipeline is useable and accessible for as many people as possible, I have chosen a series of criteria that each of its modules must meet:

C1 - Performance: Each module must achieve state-of-the-art performance for their respective task.

C2 - Usability: Each module must be able to fulfil criterion **C1** while requiring a minimal amount of pre-processing to be performed on the 3D US scans before the pipeline.

C3 - Robustness to age: Each module must be able to fulfil criterion **C1** across a wide gestational range. In particular, the performance of the modules must remain consistent across the entire gestational range of 14.0 to 30.9 GW.

C4 - Robustness to misalignment: The performance of each module must remain invariant to the location, and orientation of the 3D brain relative to the original scan, as these can vary depending on the expertise of the sonographer, as well as the varying fetal pose during image acquisition.

C5 - Robustness to quality: Each module must be able to fulfil the previous criteria when analysed against a dataset that offers a realistic representation of the range of image quality (contrast, artefacts, etc.) that is usually expected from 3D US scans.

1.2 Thesis Contribution

The main contribution of this thesis is the development of a fully-automated, Deep Learning (DL) pipeline for 3D fetal brain ultrasound. This pipeline is currently comprised of three modules: a fetal Brain Extraction Network (fBEN) described in Ch. 4, a fetal Brain Alignment Network (fBAN) described in Ch. 5, and a fetal Brain Fingerprinting Network (fBFN) described in Ch. 6. This section outlines each of their major contributions. These modules are summarised in the graphical abstract of this thesis, shown in Fig. 1.2.

1.2.1 Chapter 4: Fetal Brain Extraction Network

For the first contribution, I propose the fetal Brain Extraction Network (fBEN). The fBEN is a fully-automated, end-to-end 3D Convolutional Neural Network (CNN) with an encoder-decoder architecture for the automated extraction of the fetal brain from minimally pre-processed, standard clinical 3D US scans. This network is the first solution that that directly extracts the brain from fetal 3D US scans, without relying on shape approximations or slice-by-slice predictions. It manages to accurately segment and extract the complete brain with a high degree of accuracy and reliability, regardless of the gestational age (ranging from 14.0 to 30.9 weeks), brain and probe positions, and visible brain structures. The fBEN achieved state-of-the-art segmentation performance, significantly outperforming all current

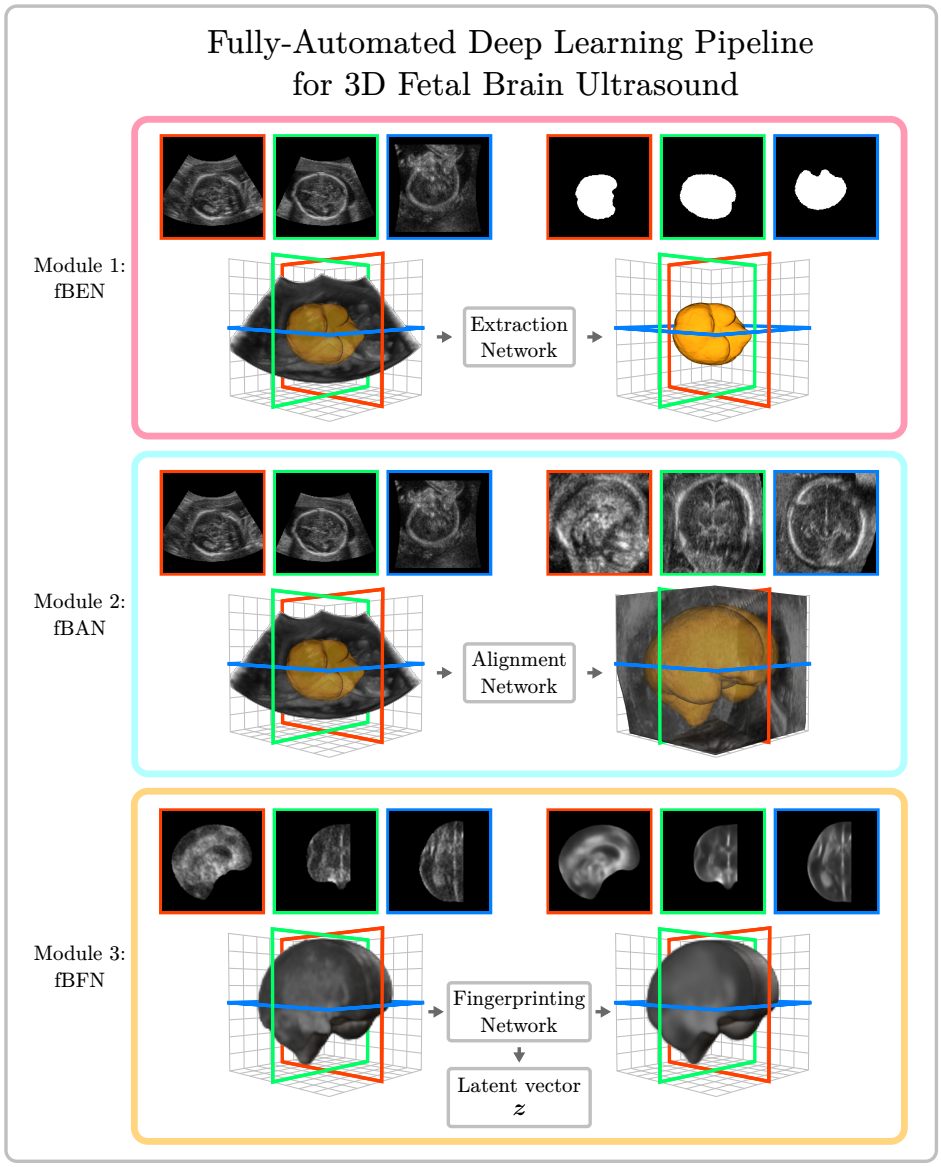

Figure 1.2: Schematics of the three modules that comprise the pipeline presented in this thesis.

alternatives for 3D US and achieving comparable performance to the current state-of-the-art MRI solutions while covering a significantly larger gestational age range.

1.2.2 Chapter 5: Fetal Brain Alignment Network

For the second contribution, I propose the fetal Brain Alignment Network (fBAN). This is a fully-automated, end-to-end regression network with a cascade architecture that accurately predicts the alignment parameters required to rigidly align minimally pre-processed, standard clinical 3D US scans, to a canonical reference space. The

fBAN manages to achieve consistent, state-of-the-art performance across the entire gestational age range of 14.0 to 30.9 weeks, regardless of the initial location and orientation of the brain in the scan. The network achieves this performance thanks to a novel, multi-stage training approach, that includes (i) the use of Transfer Learning from the fBEN (Ch. 4) to ensure that fBAN is making its prediction based on the understanding of the structural information of the brain, hindering it from "cheating", as well as (ii) a novel, highly efficient representation of the alignment task through the use of Spatial Landmarks (SL).

1.2.3 Chapter 6: Fetal Brain Fingerprinting Network

Finally, for the third contribution, I propose the fetal Brain Fingerprinting Network (fBFN). The fBFN is a is a fully-automated, end-to-end network based on a Variational AutoEncoder (VAE) architecture, that encodes the structural information of the 3D brain into a relatively small set of parameters in a continuously distributed latent space. It is a general-purpose solution aimed at facilitating the analysis of the 3D US scans by recharacterising the fetal brain into a representation that is easier to analyse. The fBFN manages to encode the entire structural information into 500 parameters in a continuously distributed latent space, from which it can subsequently generate a structurally-accurate reconstruction. The fBFN manages to predict the gestational age of the input scan as the first of these parameters, achieving state-of-the-art performance. Additionally, the latent space distribution learnt by fBFN allows for the artificial manipulation of the gestational age, with preliminary results showing that the reconstructed scan is structurally consistent with subsequent scans of the same subject. Furthermore, this learnt distribution has the potential to facilitate the study of the normal development of the fetal brain without the need for additional labelling. Finally, the Euclidean distance between latent vectors in latent space can be used to efficiently locate structurally similar scans at a minimal computational cost.

1.3 Thesis Structure

This thesis is comprised of seven chapters. In this first chapter (Chapter 1) I cover the motivation behind the development of the proposed Fully-Automated Deep Learning Pipeline for 3D Fetal Brain Ultrasound, the main contributions involved, and the general structure in which the thesis has been written. Chapter 2 contains a comprehensive literature review of the relevant topics and methods. Chapter 3 covers the data used for the development of the pipeline, its acquisition, and how it was processed and prepared. Chapters 4, 5, and 6 discuss the development and performance of the extraction (fBEN), alignment (fBAN), and fingerprinting (fBFN) modules of the pipeline, respectively. Finally, the current limitations of the proposed pipeline, as well as potential future works are discussed in Chapter 7.

For each contribution chapter, I first briefly summarise the motivation and contributions involved, as well as the data used and its processing. I then cover the technical development of each module, from its initial development to its final refinement, with particular focus on relevant milestones. Finally, I perform an exhaustive, multifaceted assessment of the performance of each module, and I summarise the findings in the context of the proposed pipeline and the current state of the field.

1.4 Statement of originality

I hereby declare that this thesis is my own work, except where specific references are made to the work of others. The contents of the included chapters are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.

1.5 Publications

The works presented in this thesis have been published on the following conferences and journals:

- **Moser, F.**, Huang, R., Papageorghiou, A. T., Papież, B. W., Namburete, A. I. (2019). Automated fetal brain extraction from clinical ultrasound volumes using 3D convolutional neural networks. In Annual Conference on Medical Image Understanding and Analysis (pp. 151-163). Springer, Cham.

- **Moser, F.**, van der Vaart, M., Papageorghiou, A. T., Papież, B., Namburete, A. I. L. (2020). Brain volume from 3D ultrasound for fetal growth assessment using deep convolutional neural networks. In Organization for Human Brain Mapping

- (*poster*) **Moser, F.**, van der Vaart, M., Papageorghiou, A. T., Papież, B., Namburete, A. I. L. (2020). Brain volume from 3D ultrasound for fetal growth assessment using deep convolutional neural networks. In Organization for Human Brain Mapping

- **Moser, F.**, Huang, R., the INTERGROWTH-21st Consortium, Papież, B. W., Namburete, A. I. L. (2022). BEAN: Brain Extraction and Alignment Network for 3D Fetal Neurosonography. NeuroImage, 119341.

Additionally, the works presented in this thesis have been instrumental for the works published on (or submitted to) the following conferences and journals:

- Hesse, L. S., Aliasi, M., **Moser, F.**, the INTERGROWTH-21st Consortium, Haak, M. C., Xie, W., Jenkinson, M., Namburete, A. I. L. (2022). Subcortical segmentation of the fetal brain in 3D ultrasound using deep learning. NeuroImage, 254, 119117.

- (*under review*) Namburete, A. I. L., Papież, B. W., Fernandes, M., Wyburd, M. K., Hesse, L. S., **Moser, F. A.**, Ismail, L. C., Gunier, R. B., Squier, W., Ohuma, E. O., Carvalho, M., Jaffer, Y., Gravett, M., Qingqing, W., Lambert, A., Winsey, A., Restrepo-Méndez, M. C., Bertino, E., Purwar, M., Barros, F. C., Stein, A., Noble, J. A., Molnár, Z., Jenkinson, M., Bhutta, Z. A., Papageorghiou, A., Villar, J., Kennedy, S. H.. Normative spatiotemporal

dynamics of human fetal brain maturation associated with satisfactory growth and neurodevelopment up to 2 years of age. *Submitted to Nature.*

2

Literature review

Contents

2.1 Normal brain development during gestation

In this section, a brief overview of the normal development of the brain during pregnancy is provided. This development can be divided into two main period: an embryonic period that takes place from conception until 8 GW, and a fetal period from 9 GW until birth.

2.1.1 Embryonic brain development

After conception, the zygote, i.e. fertilized egg, begins the process of *blastulation*, repeatedly subdividing until becoming a fluid-filled spherical structure called a *blastula* [44]. This is followed by a process called cell *gastrulation*, in which a slit-like opening called the primitive streak is created on the blastula, reorganising it into a multi-layer structured called the *gastrula* [45]. This streak divides the gastrula into dorsal/ventral and cranial/caudal regions. The rostral region will eventually become the head of the fetus.[1].

At around 3 GW, a process called *neurulation* begins with a thickened area of cells called the *neural plate* is formed at the cranial end of the embryo. Two ridges form along the neural plate of the embryo. They fold inwards and fuse from the middle outwards creating a hollow tube shaped structure called the *neural tube*, which will eventually become the brain and spinal cord. By 4 GW, the anterior end of this tube expands and forms the three primary brain vesicles: the *prosencephalon* which will become the forebrain, the *mesencephalon* which will become the midbrain, and the *rhombencephalon*, which will become the hindbrain. From this point until 8 GW, the prosencephalon will subdivide into the *telencephalon* and the *diencephalon*, while the rhombencephalon will subdivide into the *metencephalon* and the *myelencephalon*, creating the five secondary brain vesicles.

The schematics of these steps can be seen in Figure 2.1 (obtained from [1]), which shows the development of the embryo between 19 days of gestation (E19), to 49 days of gestation (E49). This is considered the primary organization of the central nervous system and the starting point of the fetal period. [46]. At this point, major compartments have differentiated within the diencephalic and midbrain regions [47] [48], and the hindbrain and spinal separation has been specified [49] [50]. While outside the scope of this work, an in-depth description of the cellular and molecular steps responsible for this structural evolution of the embryo can be found in [51] [52] [53] [54] [55] [56] [57].

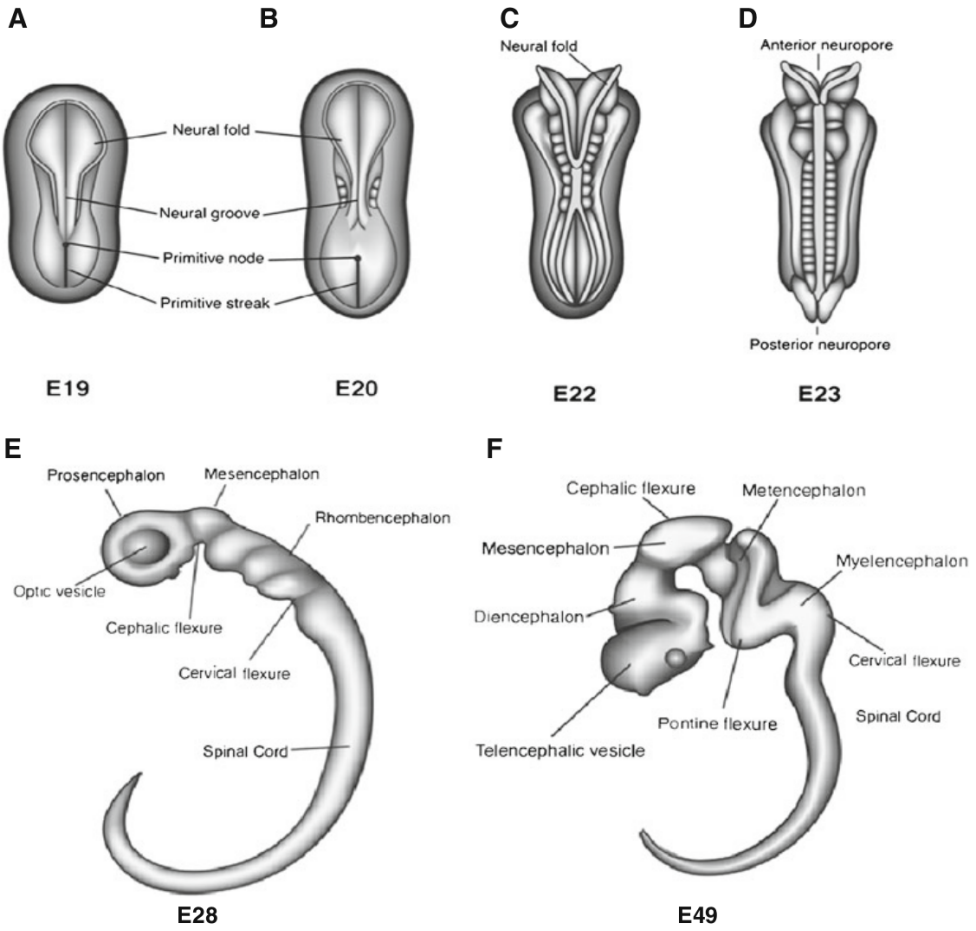

Figure 2.1: Figure obtained from [1] . Steps A through D show the process of the formation of the neural tube. Step E shows the primary vesicles. The emergence of the secondary vesicles can be seen in step F.

2.1.2 Fetal brain development

At the end of the embryonic period around 8 GW, the brain is a smooth structure, also referred to as *lissenphalic*. From the start of the fetal period onwards, it undergoes a sequence of folding, generating ridges, i.e. *gyri*, and grooves, i.e. *sulci* at predictable time points in the gestation period [2]. This process is called *gyrification* and it starts with the creation of the *longitudinal fissure*, which is the sulcus that separates the left and right hemispheres of the brain. This is the earliest gyrification step, starting at around 8 GW [58] and continues developing all the way to 22 GW. The Sylvian sulcus, also known and Sylvian Fissure (SF) develops between 14 and 16 GW and separates the frontal and parietal lobes from the temporal lobe. During this period, the Cingulate and Parieto-Occipital and

Calcarine sulci also develop. Between 20-24 GW, both the Central and Superior Temporal sulci are developed. Finally, the Superior Frontal, Precentral, Inferior Frontal, Postcentral, and Intraparietal sulci form between 25-26 GW. These sulci then constitute the primary sulci and constitute the most important gyrification steps during fetal brain development. Secondary sulci develop during 30-35 GW and tertiary sulci start their development around 36 GW and continue developing after birth [1], at which point the general shape of the brain conforms to what is essentially the same shape of the adult brain. [59] [60] [61]. Figure 2.2 (obtained from [41]) shows a detailed summary of the folds that appear at certain gestational periods, accompanied by a schematics showing the change in appearance.

AGE	SULCI AND FISSURES	GYRI	
8 - 13 wks.	Interhemispheric fissure, sylvian fissure, callosal sulcus		
14 - 17 wks.	Parieto-occipital fissure, olfactory sulcus, cingulate sulcus, calcarine fissure	Gyrus rectus, insula, cingulate gyrus	
18 - 22 wks.	Rolandic sulcus, collateral sulcus, superior temporal sulcus	Parahippocampal gyrus, superior-temporal gyrus	
22 - 25 wks.	Prerolandic sulcus, middle temporal sulcus, postrolandic sulcus, inter-parietal sulcus, superior frontal sulcus, lateral occipital sulcus	Pre and postrolandic gyri, middle temporal gyrus, sup. and inf. parietal lobules, sup. and mid. frontal gyri, sup. and inf. occipital gyri, cuneus, lingual and fusiform gyri.	
26 - 29 wks.	Inferior temporal sulcus, inferior frontal sulcus	Inf. temp. gyrus, triangular gyrus, med. and lat. orbital gyri, calosomarginal gyrus, tran. temp. gyrus, angular gyrus, supramarginal gyrus, external occipito-temp. gyrus	

Figure 2.2: A detailed list of sulci, fissures, and gyri developed at specific gestational ages. Figure obtained from [41].

The description of the gyrification process of the brain has been achieved by using post-mortem anatomical evaluations [58] [62] [59] [63], as well as different imaging techniques such as US [64] [65] [2] [66] and MRI [41] [67] [68] [69] [70] [71] [72]. A comparison of how these folds are seen in each of these modalities

can be seen in Figure 2.3. It is worth noting that the gestational age at which each gyrus and sulcus appear is dependent on the modality used. In general, these structures are visible at earlier points in gestation in post-mortem anatomical methods. Depending on the structure, however, MRI or US may show signs of their development earlier. For example, the interhemispheric fissure can first be seen through anatomical examination around the 10 GW, while MRI does this at 14 GW and US only at the 18 GW. However, US manages to show the first signs of the callosal sulcus appearing around 18 GW, while both MRI and anatomical studies first report its visualization around 22 GW. A more detailed timeline of the gyrification process as observed with US can be seen in Cohen-Sacher et al. [66]. It is important to know the visualization timeline when assessing for abnormalities in the gyrification process, since deviations from the normal timeline could be related to brain anomalies such as ventriculomegaly and lissencephaly [2].

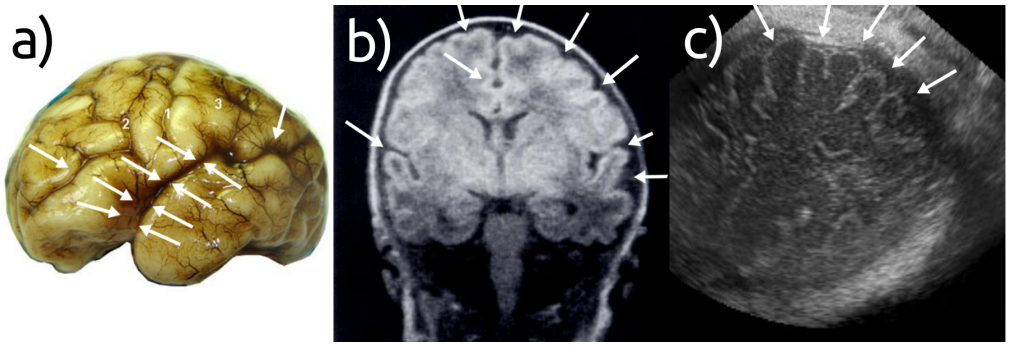

Figure 2.3: Gyri and sulci as seen in different studies. Arrows point to examples of these structures. a)Anatomically [63], b) MRI [41], c) US [66].

2.2 Clinical Assessment of the fetal brain

In order to have a consistent and reliable assessment of the brain development during gestation, the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) published in 2007 a series of guidelines for performing the "basic examination" of the fetal central nervous system.

The basic examination is mainly performed through transabdominal sonography on low risk pregnancies, although in some cases transvaginal and transfundal sonograms

may be used instead. This examination usually takes place around the 20th GW [3] but it can be performed during the late first (0-12 GW) [73], second [74] (13-27 GW), or third trimester of gestation (28 GW to birth) [75]. It is important to take into consideration that the ossification of the fetal skull starts affecting the quality of the scan after the 30th GW due to beam attenuation, hampering the visualization of intracranial structures. However, scanning is still possible after this point if required.

The basic examination can be divided into two main components: qualitative evaluation (Sec. 2.2.1) and quantitative evaluation (Sec. 2.2.2).

2.2.1 Qualitative evaluation

To assess the anatomical integrity of the brain in routine clinical assessment, the transventricular and transcerebellar axial planes are imaged. Schematics of the relative positions of these planes can be seen in Fig. 2.4. Note that a third transcerebellar plane is also usually visualized for the biometric measurements that are discussed in Sec. 2.2.2.

Besides a general assessment of the head shape and the brain texture, there are four key structures that clinicians rely on to determine the health of the fetal brain: LV, CSP, CER, and CM. Deviations in size or shape of these structures have been associated with several developmental anomalies and are therefore closely inspected as part of the basic fetal neurosonogram examination [3].

The LV, divided into an anterior and posterior portion, are clearly visualized in the transventricular plane, shown in Fig. 2.4a. The anterior portion, also called frontal or anterior horns, consists of two dark comma-shaped structures, with well defined walls, separated by the CSP. The posterior portion, also called occipital or posterior horns, actually consists of two parts: the bright CP and the dark occipital horns themselves. They are enveloped by bright medial and lateral walls. Under normal conditions, the CP should almost completely fill atrium of the ventricle, with deviations from this expectation signalling possible developmental problems, such as ventriculomegaly [76] [77] .

Like the LV, the CSP is visualized using the transventricular plane and it is

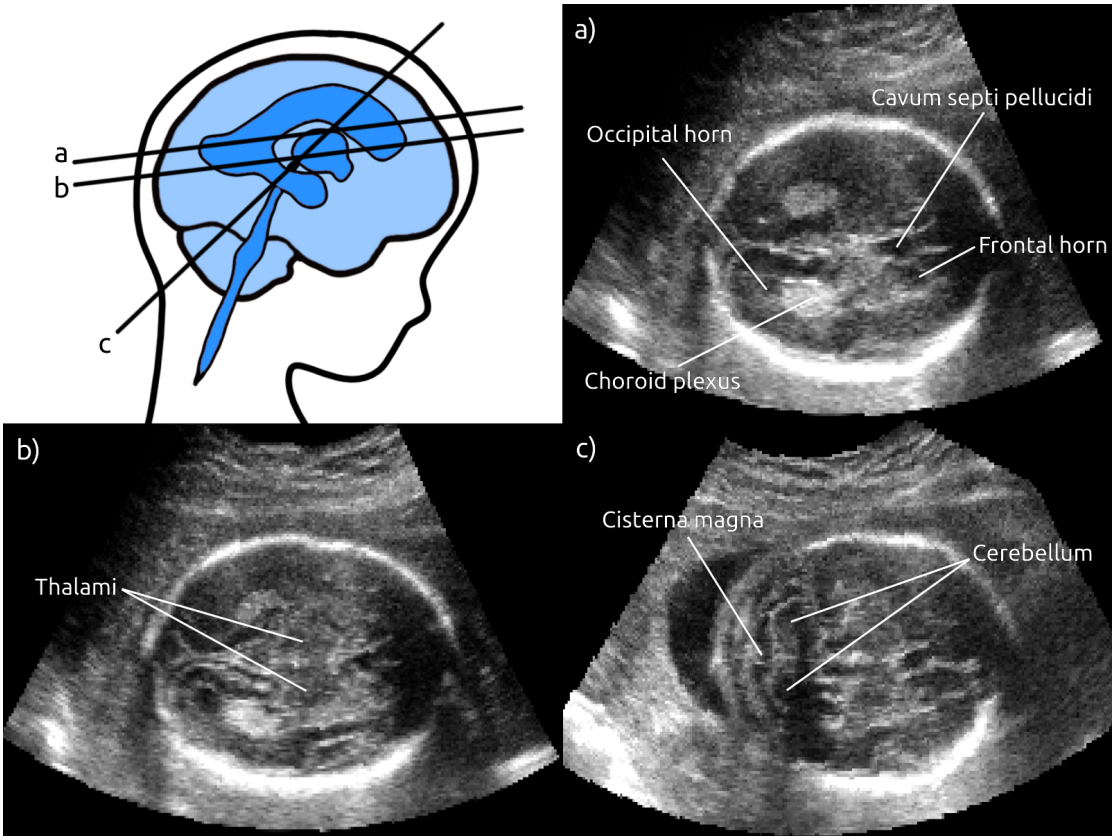

Figure 2.4: Schematics of the relative positions of the three standard planes for clinical assessment at 20 GW: a) transventricular plane, b) transthalamic plane, c) transcerebellar plane.

shown as a fluid filled cavity between two thin membranes. It is debated if the visualization of the CSP provides valuable information for general assessment of the brain development [18] but since it is straightforward to locate and its visual characteristics are affected by several cerebral lesions, it is nevertheless included in the basic evaluation. For example, although the failure to visualize the CSP prior to 16 GW or after 37 GW is normal, it has been associated with several brain anomalies, such as agenesis of the corpus callosum, or schizencephaly [8] [78]. An enlarged CSP has also been associated with other brain anomalies, such as hydrocephalus, chromosomal translocation, and growth retardation [7].

The transcerebellar plane, shown in Fig. 2.4c, cuts through the middle of the CER and is therefore the plane used for its developmental evaluation. The CER is visible as a butterfly shaped structure composed of two round cerebellar hemispheres joined by the cerebellar vermis. The hemispheres of the CER are hypoechoic with echogenic

borders, making its appearance quite distinctive, as can be appreciated in Fig. 2.4c.. The vermis is also hyperechoic and the primary fissure can be constantly observed from 24 GW [79]. Abnormalities of the cerebellar morphology are associated, for example, with chromosomal anomalies [9] [10], as well as spina bifida. [11]
The CM, also called cisterna cerebello-medullaris, is composed by the space between the CER and the occipital edge of the skull. It is filled with fluid and often shows thin separations that, while completely normal, can be confused with abnormalities. However, other structural abnormalities of the CM have been linked to developmental problems such as the Dandy-Walker malformation [12].

2.2.2 Quantitative evaluation

Biometrics are a vital part of the general health assessment of the fetus during gestation. They provide a point of comparison with the developmental range of the rest of the population, facilitating the assessment of whether the fetus is within the normal range of development, also referred to as Appropriate for Gestational Age (AGA), or not. This is of great importance, since disorders such as Small for Gestational Age (SGA) and Large for Gestational Age (LGA) have been associated with adverse maternal and perinatal outcomes.
For fetal biometry, the most common measurements are of the biparietal diameter (BPD), head circumference (HC), abdominal circumference (AC) and femur diaphysis length (FL). These measurements are usually used to calculate the estimated fetal weight (EFW). While focus will be on the fetal head biometry, it is important to mention that before week 14 of gestation, the crown-rump length (CRL) is used as the most reliable method to determine gestational age [80]. After this, however, it is the HC that is used for this purpose [73] [81].
Both head biometric measurements are performed using the transthalamic plane shown in Fig. 2.4b. The BPD is measured by using calipers, while the HC can be measured, depending on the equipment, directly by overlaying an ellipse or indirectly by using the same calipers as for the BPD to measure the occipitofrontal distance (OFD) of the skull and calculating it using the equation $HC = 1.62 \times (BPD + OFD)$

[3]. The ellipse approach is the recommended technique for most accurate results [81]. These values are then compared to charts of reference values to determine the relative position of the fetal biometrics compared to the population, usually by the use of Z scores. Three of the most widely used charts are Snijders et al.[82], Chitty et al.[83], and Kurmanavicius et al.[84].

Besides the measurements of the BPD and HC, the measurement of the internal diameter of the atrium, as well as the transverse cerebellar diameter and the depth of the CM are recommended. The atrium measurement is seen as the most effective method to assess the integrity of the ventricular system and to determine abnormalities such as ventriculomegaly, as mentioned in section 2.2.1. It is expected to have a stable diameter of 6-8 mm in the second and early trimesters [76] [85] [13], and a value of 10 mm or larger is considered suspicious [13] [14]. The diameter of the CER, on the other hand, can be used to assess fetal growth [15] [86] and its biometric deviation is associated with several conditions, such as cerebellar hypoplastia, brainstem hypoplastia, and rhombencephalosynapis [70]. Finally, abnormally large and small measurements of the CM may be benign, such as the case of mega-CM. However, some conditions have been linked to this abnormal dimensions, such as the case of amniocentesis [17] [18] [19] and trisomy 18 [12] (although debated in [87]), but technical limitations make this particular measurement difficult and often the patient will be referred to further evaluation.

While not contained in the current guidelines for clinical evaluation, the grade and symmetry of the fetal cortical development has been proposed as a method to assess the health of the brain development during gestation. Pistorious et al. [64] showed that rapid and accurate grading of the cortical development can be performed with 2D and 3D US. As mentioned in Sec 2.1.2, the gyrification of the brain follows a predictable pattern. Deviation from the norm at a particular gestational age can be used to detect anomalies in the brain development, such as ventriculomegaly [2].

2.2.2.1 Shortcomings of clinical assessment

The general approach of clinical assessment is well thought out and involves a series of qualitative and quantitative steps to ensure a complete and reproducible evaluation of the development of the fetal brain. The structures that have shown an association with certain developmental anomalies are carefully observed and measured. However, there is a large amount of variability intrinsic to the assessment performed by each clinician. Firstly, the qualitative assessment is based on the experience of the person performing the scan, which can vary greatly from sonographer to sonographer [20]. The quantitative measurements also contain a large amount of variability. For example, the accuracy of placing the calipers to measure the BPD, OFD, or fitting the ellipse to the HC, will determine how reliable those measurements are. Using the wrong plane for the measurement can also lead to wrong dimensions being measured. Additionally, the chart used to compare the measurements to the general population can also have a massive impact on the assessment itself. Salomon et al. [21] showed that by simply comparing the biometry to different standard charts, significant deviations could be observed, with the number of biometric measurements considered abnormal varying between 2.6% and 23.6% depending on the chart used. There is a need for automation to reduce the intrinsic variability associated with these assessment. Several methods and approaches to this particular problem have been proposed in the last decade, with a large amount of them harnessing the potential of Artificial Intelligence (AI) through powerful computational techniques such as Computer Vision (CV), Machine Learning (ML), and DL. A review of current automated approaches to different tasks related to fetal brain development can be found in Sec. 2.3.

2.3 Automated methods for 3D fetal neurosonography

2.3.1 3D brain extraction

While there are several publications proposing methods for the task of brain segmentation from 3D US scans, the majority of these works focus on the segmentation of sub-cerebral structures, such as the CER, CP, CSP, or the LV [35] [88] [89] [90] [37] [91] [92] [40] [93] [94] [94] . In contrast, the number of publications that have proposed methods for the segmentation or extraction of the brain itself is significantly more limited.

Chen et al. [95] proposed a method based on a coarse-to-fine strategy for fetal head registration and segmentation. The eye of the fetus was first detected through Gabor Features [96] to identify head-pose, which was then followed by registering a reference model of the fetal head to the imaged head. However, the reliance on a reference model, as well as the presence of strong edges in the image strongly limit its implementation. Also, the method is not fully-automated as the Volume Of Interest (VOI) needs to be first identified manually for each scan.

Cuingnet et al. [97] proposed to first detect the fetal head using a prolate spheroidal shell model which is then fitted to the 3D scan by maximising the response of an anisotropic plate diffusion filter. This shell is subsequently used as a deformable template which is fitted to the actual shape of the skull through global and local non-rigid deformations obtained through maximisation of the gradient flux across its surface, as described in [98].

Namburete el al. [99] proposed a method that relies on a template deformation approach for the parametrisation of the 3D fetal skull. A basic surface model of the 3D skull is first manually aligned to the imaged brain through translation, rotation, and anisotropic scaling. This is finally refined through an iterative deformation process that relies on standard US edge detection techniques.

Cerrolaza et al. [36] proposed using a structured geodesic Random Forest (RF) approach that implements structured labels and semantic features for automated

segmentation of the fetal skull. While this solution showed good performance, it resulted in a significantly lower sensitivity when compared to a classic non-structured RF method, as well as a U-Net [100] based CNN method. In a subsequent study [101], their improved method for 3D skull segmentation used a two-step CNN approach. The first CNN performed an initial segmentation of the skull, which was then used to derive an Incidence Angle Map (IAM) and a Shadow Casting Map (SCM), which provided relevant complementary information of the underlying physics of US image acquisition. The initial segmentation was then passed along with the two derived maps as input for the second CNN, which returned the final segmentation.

Perez Gonzalez et al. [102] proposed a method that relied on texture, intensity and edges to train a Support Vector Machine (SVM), which was then used to segment the fetal skull from the 3D US scan. This was subsequently improved upon by replacing the SVM with a RF classification approach in [103].

Finally, Namburete et al. [104] proposed a multi-task fully CNN that addressed brain localization, segmentation, and alignment to a canonical coordinate system. The alignment task of this work is discussed in Sec. 2.3.2. The brain segmentation task consisted of predicting the fetal brain segmentation of 2D US image slices with a CNN network. These predictions were then stacked together, creating a 3D volume to which an ellipsoid was fitted as an approximation of the fetal brain. However, although this network showed good performance in their testing, two important limitations affect its potential. Firstly, by predicting the brain segmentations from 2D images, the 3D context of the original volume is lost. This context is important information that could be used to improve the segmentation. Secondly, the ellipsoid approximation of the shape of the brain does not represent its actual morphology and will therefore always include background voxels in its brain segmentation.

To the best of my knowledge, the fBEN described in Ch. 4, which has previously been partially published in [105] and [106], is the first and only approach that directly extracts the brain from fetal 3D US scans, without relying on shape approximations, or template registration.

2.3.1.1 Relevant 3D MRI methods

Several solutions have been proposed for the task of fetal brain extraction from MRI scans. Anquez et al. [107] proposed using a template-matching approach to locate the fetal eye, subsequently segmenting the surrounding skull bone using a graph cut approach [108]. However, this method relies on the eye structure to be clearly visible inside the scan, as well as a clear contrast between the skull and the brain. Automated solutions that rely on brain templates have also been proposed [109] [110] [111] [112] [113] [114]. While these works showed good results for fetal brain segmentation of MRI scans, such solutions tend to struggle on scans that show a large deviation from the average, such as structurally abnormal and pathological cases.

ML solutions have also been explored for fetal brain extraction from MRI scans. Kainz et al. [115] proposed a RF classifier trained on 3D Gabor descriptors, with a subsequent segmentation refinement using a 2D level-set. Keraudren et al. [116] used Scale-Invariant Feature Transform (SIFT) features for localization clustered with k-means, and subsequently classified with a SVM classifier and refined using a RF classifier on patches of the 2D slices.

More recently, the high performance of CNNs for the task of fetal brain extraction has been thoroughly demonstrated by the works of Rajchl et al. [117] [118], Salehi et al. [119], Khalili et al. [120] [121], Ebner et al. [122], and Ranzini et al. [123]. These solutions rely on an initial 2D brain segmentations from each slice using CNNs, which can then combined into a 3D volume. Although such an approach results in a more compact network, it greatly limits the contextual information available to the network. However, in the case of fetal brain MRI segmentation, the larger inter-slice spacing and potential motion corruption between neighbouring slices make this a compelling trade-off [122].

2.3.2 3D brain alignment

Kuklisova-Murgasova et al. 2012 [124] [125] proposed an automated method for the registration of 3D fetal neurosonograms with MRI images. This was done by segmenting the MRI volume using a probabilistic atlas. A simulated US volume

is created based on this segmentation, and then the real US volumes are affinely aligned using a block-matching approach in order to deal with intensity artefacts and missing features in the US volumes.

Cuingnet et al. [97] proposed a multi-step approach for automated alignment of the fetal brain from 3D US scans. In addition to the skull segmentation steps described in Sec. 2.3.1, it relied on a weighted Hough transform to detect the mid-sagittal plane, a minimum intensity gradient across the skull surface to detect the neck location, and a RF classifier to detect the eyes. However, this approach is strongly limited by the reliance on a clear view of the neck and eye, limits its implementation in cases where these structures are occluded or affected by acoustic shadows.

Perez Gonzalez et al. [102] and [103] proposed the use of a Coherent Point Drift method to register the point-cloud representation of the skull segmentations described in Sec. 2.3.1. While this approach showed good performance, its focus on skull registration limits its potential as a brain alignment method, for which a secondary registration step would be required.

As mentioned in Sec. 2.3.1, Namburete et al. [104] proposed the use of a multi-task fully CNN that addressed brain alignment to a referential coordinate system. The alignment was achieved by defining a parametric coordinate system based on skull boundaries, location of the eye sockets, and head pose. The network performed 5 tasks for each 2D axial slices of a 3D US scan: 3 classification tasks, namely (i) slice located near crown or neck, (ii) ante-posterior orientation of the head, and (iii)presence or absence of eye in the slice, as well as and 2 segmentation tasks, namely (i) eye segmentation, and (ii) brain segmentation. The transformation matrix for rigid alignment of the brain was then calculated by multiplying four separate matrices based on the combined prediction of the stacked 2D predictions of the 5 tasks. This method was subsequently used as the initial alignment step for the multi-channel groupwise registration presented in [126]. One strong limitation of this approach is the loss of 3D context for the predictions, since they are done

for each 2D slice independently. This also means that there is need for post-processing to fuse the network predictions to estimate the transformation matrix for alignment. The transformation matrix for the alignment is calculated based on these 2D segmentations, causing any discrepancies between slices to impact the alignment result. Similarly to [97], the reliance on a clear view of the eye also limits its implementation.

More recently, Wright et al. [127] proposed using the Long short-term memory (LSTM) spatial co-transformer solution proposed in [128] to co-align images of the fetal head to a canonical pose, showing remarkable results. The network performs an initial alignment, which is then iteratively refined by group-wise registration with a saliency-weighted compounded volume. The predictions are performed independently for each image, allowing for the alignment of single scans. However, analogous to [128], the network relies on multiple views of the same subject for training, which is not always available.

The fBAN described in Ch. 5, which has been previously partially published in [106], is an end-to-end, fully automated DL solution for the task of fetal brain alignment from 3D US scans. It accurately and robustly aligns the 3D fetal brain to a canonical space, without any additional requirements such as the identification of multiple structures, multiple images of the same subject, or subsequent registration steps.

2.3.2.1 Relevant 3D MRI methods

In the case of fetal MRI, research is heavily focused on the reconstruction of the 2D slices into a 3D volume rather than alignment of the 3D brain itself due to the nature of how the MRI scans are generated [122] [129] [130]. While fast imaging methods such as Single-Shot Fast Spin Echo (SSFSE) can minimize the effects of fetal in-plane motion, motion-corruption is still a common issue between slices of the stack. In contrast, while motion artifacts can still occur when acquiring 3D US scans, the shorter acquisition time of this imaging modality strongly minimizes their occurrence. The acquisition can also simply be repeated if such artefacts are

indeed noticed by the sonographer during the scanning. A reconstruction of 2D slices into a 3D volume is therefore not needed for 3D US. The few works that are particularly relevant to this thesis, namely those of Kuklisova-Murgasova et al. [124] [125] and Wright et al. [128] have already been described in Sec. 2.3.2.

2.3.3 3D brain fingerprinting

After successfully addressing the two fundamental tasks of the 3D US fetal neuroimage analysis pipeline, i.e. brain extraction and alignment, the next focus of my doctoral research was to develop a method that would facilitate the challenging task of analysing the 3D fetal neurosonograms. As mentioned in Sec. 1.1, the difficulty of analysing 3D US brain data obfuscates the benefits of this modality over its 2D counterpart, contributing to its slow adoption[24]. Therefore, the goal was to develop a method that could represent the entire structural information of the fetal brain in the 3D US scan in a form that can be more easily analised: a fetal brain *fingerprint*.

The idea of generating a fingerprint that represents the characteristics of an individual brain has been rapidly growing in popularity over the last decade. While the general characteristics of the brain are shared across the population, multiple studies have shown that there is a significant amount of variability between individuals [131], both structurally [132] and functionally [133][134][135]. These individual characteristics have been shown to be largely determined by genetic factors, and some have been linked to neurological disorders such as Parkinson [136], Dandy-Walker Malformation Complex [10], Attention Deficit Hyperactivity Disorder (ADHD) [137], and Autism [138][137]. Therefore, their representation and understanding is crucial for a more precise, personalised medical care [139].

However, to the best of my knowledge, no such fingerprinting solution has been proposed for the characterisation of the structural information of the fetal brain, with the closest works being those of Ciarrusta et al. [140] and Kim et al. [141], which focused on the characterisation of the connectome of the fetal and neonatal brain from functional and diffusion MRI scans.

Therefore, the fBFN described in Ch. 6 is the first and only approach that condenses the structural information of the fetal brain from 3D US scans into a general-purpose fetal brain fingerprint.

2.3.3.1 Relevant 3D MRI methods

Several works for the generation of a brain fingerprint have been proposed for adult brain MRI. Some of these works have been focused on characterising the functional characteristics of the brain [142][143][144], while others focused more on the connectivity of the brain [145][146][147][148]. However, of particular relevance to this thesis are those approaches focused on the structural information of the 3D scan.

Decarli et al. [149] proposed fingerprinting the adult brain based on the volumetric size of cerebral substructures, which can be achieved through the use of robust segmentation libraries, such as FreeSurfer [150]. However, similarly to the biometric measures of the current 2D clinical assessment discussed in 2.2, the volumetric size of the structures is a crude oversimplification of the brain information contained in the scan.

Toews et al. [151] proposed representing the brain as a collage of generic, localized image features. This approach relies on scale-space theory to analyse the features at the characteristic scale of underlying anatomical structures. Through a probabilistic model, the features can be compared across different subject groups, allowing for the detection of distinctive anatomical patterns of subjects with Alzheimer's disease, as well as genetically-related individuals.

Wachinger et al. [152] proposed capturing the morphology of the adult brain from MRI scans by generating a fingerprint of the morphology of the brain. The cortical and subcortical structures are first segmented using FreeSurfer [150], then a descriptor of the morphology of each is generated through the use of Laplace-Beltrami spectra, as proposed by ShapeDNA [153]. The combined descriptors form the fingerprint, on which any subsequent analysis is then performed. However, the morphology of the brain still represents a simplification of the structural information contained in the brain, as it dismisses important information such as the tissue

texture, or their relative intensities. Additionally, similarly to [149], this method relies on a complete and accurate segmentation of the brain structures for capturing the correct morphology. While several works have been published on the topic of fetal brain segmentation from 3D US scans[154][91][37][40][36][38][39], many of which show good performance and reliability, such a complete solution as FreeSurfer is not currently available for 3D US.

To the best of my knowledge, no fingerprinting solution has been proposed for the characterisation of the structural information of the fetal brain from 3D US, making the fBFN proposed in Ch. 6 the first of its kind.

3 Data

Contents

3.1 INTERGROWTH-21st

The 3D US scans used in this thesis have been made available thanks to a collaboration with Professor Aris Papageorghiou and Professor Stephen Kennedy from the University of Oxford. The scans were acquired as part of the INTERGROWTH-21st[81] study, which provided a total of 19702 3D US scans of the fetal brain, ranging from 13 to 42 GW.

The aim of the INTERGROWTH-21st [81] study was to research the normal growth, health, nutrition, and neurodevelopment of the fetus from fewer than 14 GW to 2 years of age. For this, 2D and 3D US scans of normal, healthy fetuses were collected. In order to achieve this, the women that participated in the study had to fulfil several strict criteria. First, candidates had to have no clinically relevant obstetrical, gynaecological, or medical history. They had to be in optimal health,

and meet the criteria of nutrition, education, and socio-economic status, as well as have initiated antenatal care before the 14th GW. The study took part during April 2009, and March 2014, in eight sites that met the necessary criteria in Brazil, Italy, Oman, United Kingdom, United States of America, China, India, and Kenya. During the first visit, the gestational age was determined by the last menstrual period (LMP) and crown-rump length (CRP). After this initial procedure, a scan was performed every 5±1 weeks, including measurements of HC, BPD, OFD, AC, and FL. These were measured three times from separately obtained US images of each structure. A 3D scan of the fetal brain was taken during each visit using a Philips HD-9 ultrasound machine with a V7-3 curvilinear abdominal transducer, which is the data that will be used in this thesis.

A total of 4607 eligible women took part of the study, with 4321 having no complications. In total, 19702 3D US scans of the fetal brain were available from this study, of which 626 are from subjects that were born pre-term [155], 1795 that were born SGA [156], 113 that were born both pre-term and SGA, and 17168 that were born normally. However, this thesis will only rely on 13773 scans spanning the gestational range of 14.0 to 30.9, since the loss of structural information of the brain becomes too significant at later gestational ages due to the ossification of the skull. The gestational age distribution of this dataset is shown in orange in Fig.3.4. The volumes were resampled to an isotropic voxel (vxl) size of 0.6 mm/vxl in each dimension, and subsequently centre-cropped to a size of $160 \times 160 \times 160$ voxels.

3.2 Labels for brain alignment

For the work discussed in this thesis, it was crucial to align the 3D scans to a canonical space. This would generate the necessary ground-truth alignment parameters required for the development of fBAN (Ch.5). Additionally, these alignment parameters were crucial for the creation of the ground-truth brain mask annotations for the development of fBEN (Ch.4), as discussed in Sec.3.3.

Relying on an expert sonographer for the manual alignment of these scans was not feasible, as the time and cost of aligning thousands of scans would be prohibitive.

Therefore, I developed a simple Graphical User Interface (GUI) in MATLAB [157] that would allow for researchers with basic understand of the anatomy of fetal brain and minimal training to accurately align each scan, as shown in Fig. 3.1. The GUI shows the three mid-orthogonal planes of the 3D scans, along with a set of cross-hairs (red) to help locate the centre of the volume, and the outline of a 25 GW fetal brain as an alignment goal. The slide-bars under each view control the translation parameters, the Euler angles [158] for rotation, and the scaling parameter used to align the scan.

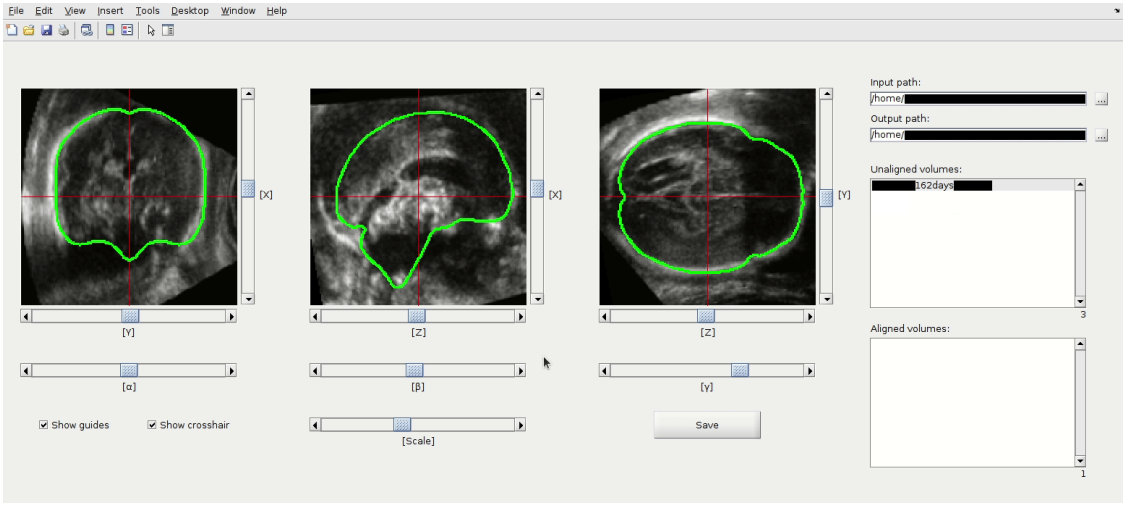

Figure 3.1: Screenshot of the GUI developed by me, used to manually aligned the 3D US scans to a canonical space.

The brain was iteratively translated and rotated until the mid-sagittal, mid-coronal, and mid-axial planes corresponded to the central orthogonal planes of the volume. Finally, each brain was isometrically scaled to match the average brain size at 30 GW in order to facilitate morphological comparisons between different GWs.

Initially, this GUI was used to manually align 1185 that were selected based on their high-quality, based on a subjective assessment of the visibility of the internal structures and the overall contrast. This was then extended to 4290 scans of both high- and low-quality, to better represent the normal quality-variability of the data. Representative examples of high- and low-quality scans are shown in Fig. 3.2.

The manual alignment yielded three translation parameters (p_x, p_y, p_z), three Euler angles (p_α, p_β, p_γ), and a single scaling parameter p_s. These ground-truth

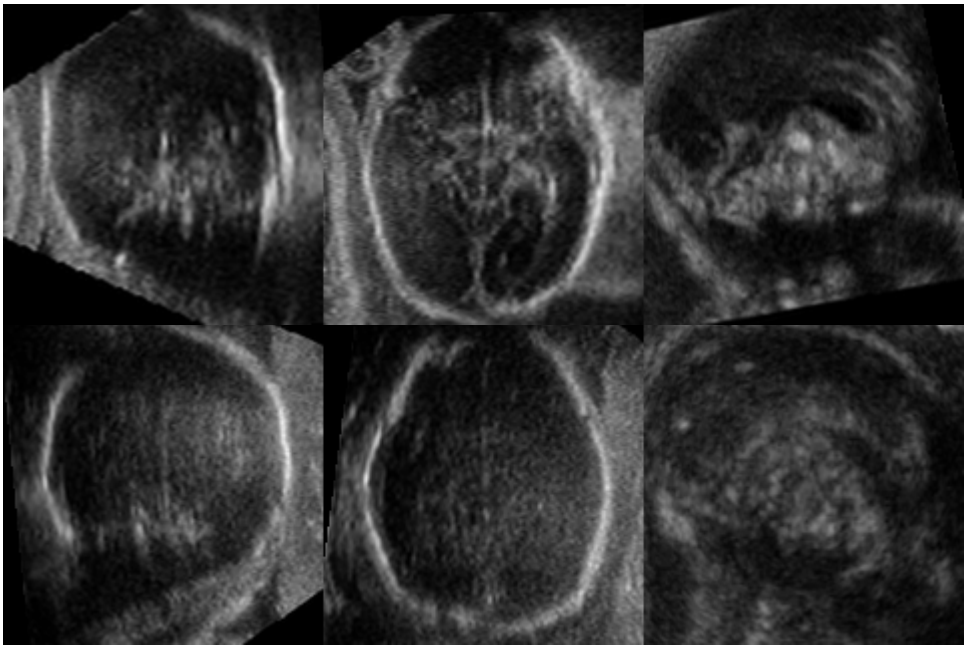

Figure 3.2: Mid-coronal, mid-axial, and mid-sagittal planes of two representative, manually aligned examples of high (top) and low (bottom) quality 3D US scan at 25 GW. Note that the structures are well defined in the high-quality example, while they are difficult to discern in the low-quality example.

alignment parameters $\boldsymbol{p}$ were then used to generate a translation transform $\boldsymbol{T}_T$, a rotation transform $\boldsymbol{T}_R$, and an isotropic scaling transform $\boldsymbol{T}_S$, which combine to generate the similarity transform $\boldsymbol{T}_{scan}$, as shown in Eq. (3.1).

$$\boldsymbol{T}_{scan} = \boldsymbol{T}_S \cdot \boldsymbol{T}_R \cdot \boldsymbol{T}_T \tag{3.1}$$

This transform $\boldsymbol{T}_{scan}$ generated with the alignment parameters $\boldsymbol{p}$ is then used to transform the original scan $\mathbf{S}$ (resampled and centre-cropped as discussed in Sec. 3.1) into the aligned scan $\mathbf{S}^{\boldsymbol{p}}$ in the canonical reference space, as shown in Eq. (3.2). A visual example of this is shown Fig. 3.3.

$$\mathbf{S}_{\boldsymbol{p}} = \boldsymbol{T}_{scan} \cdot \mathbf{S} \tag{3.2}$$

3.3 Labels for brain extraction

Due to the complexity of visually assessing 3D US images, manual labelling of the brain extraction masks by a clinician would be extremely time-consuming and

require a high degree of expertise, making it even more prohibitive than the manual alignment discussed in Sec. 3.2. In order to circumvent this problem, I relied on the normative spatio-temporal atlas of the fetal brain proposed by Gholipour et al. [159]. These atlases were generated using MRI images and represent the average brain shape for each GW in the range of 21 to 38 weeks. Each atlas was manually aligned to the same canonical reference space as the 3D US scans, using the same GUI as the scans, as shown in Fig. 3.3, resulting in the similarity transform $\boldsymbol{T}_{atlas}$. The aligned atlases were then binarised, resulting the aligned brain masks $\mathbf{M}^{p}$. Finally, using the inverse similarity transform $\boldsymbol{T}_{scan}^{-1}$, the aligned brain masks of the corresponding GW were transformed to the original position of the fetal brain, resulting in the brain masks $\mathbf{M}$. For the 14 to 20 GW scans, the spatiotemporal atlas of 21 GW was used instead.

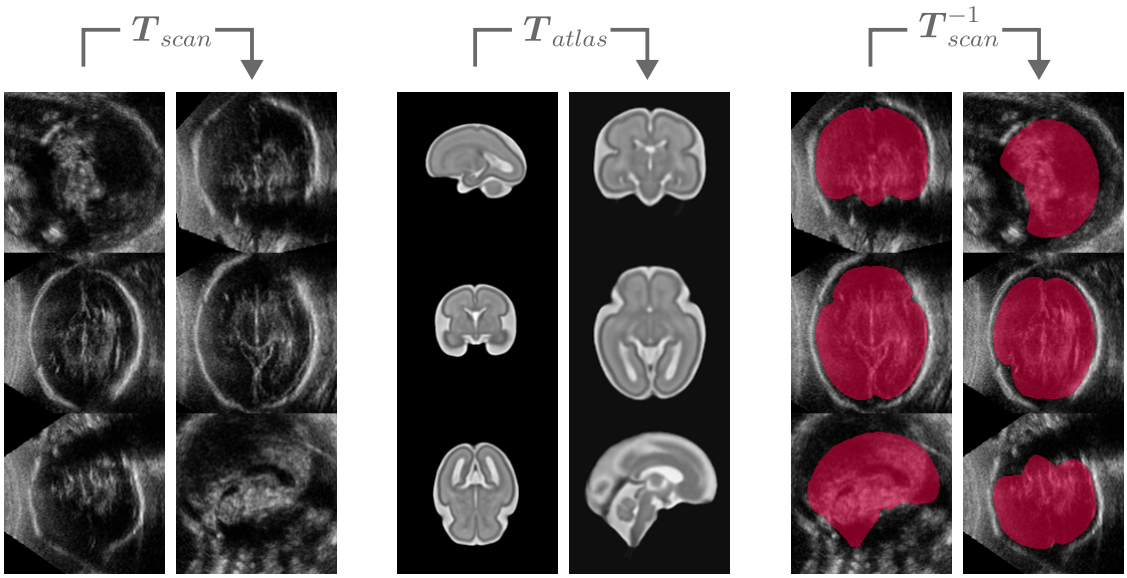

Figure 3.3: Visualisation of the process of data annotation used to generated the ground-truth 3D brain masks $\mathbf{M}$ for an example scan at 25 GW. Left: Each scan $\mathbf{S}$ is first manually aligned to a canonical space using the GUI, resulting in the similarity transform $\boldsymbol{T}_{scan}$ (Sec. 3.2). Centre: The spatiotemporal atlas [159] for each GW is manually aligned to the same canonical space, resulting in the similarity transform $\boldsymbol{T}_{scan}$ (Sec. 3.3). Right: The inverse transform $\boldsymbol{T}_{scan}^{-1}$ aligns the corresponding binarised atlas to the original position of the fetal brain (Sec. 3.3).

3.4 Datasets

Throughout this thesis, the data was subdivided into a number of datasets, namely $\mathfrak{D}_A$, $\mathfrak{D}_B$, $\mathfrak{D}_C$, $\mathfrak{D}_D$, and $\mathfrak{D}_E$, which were used at different stages of development. In this section the characteristics of each one are summarised, with examples shown in Fig. 3.5. A histogram with the gestational age distribution of each dataset is shown in Fig.3.4.

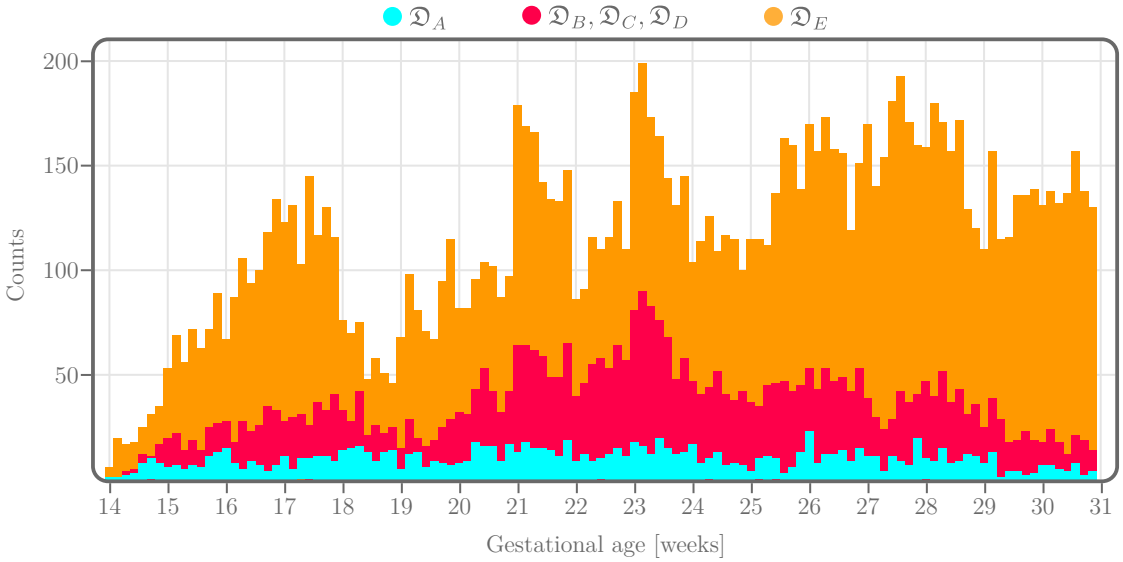

Figure 3.4: Histogram of the gestational age distribution of the different datasets used throughout this thesis. Note that $\mathfrak{D}_B$, $\mathfrak{D}_C$, and $\mathfrak{D}_D$ rely on the same original scans and therefore share the same distribution.

Dataset $\mathfrak{D}_A$: consists of a set $\mathbb{S}_A$ of 1185 high-quality 3D US scans of the fetal brain, a set of their corresponding binary masks $\mathbb{M}_A$, and a set of the corresponding alignment parameters $\mathbb{P}_A$. These parameters align the scan to a canonical space that results in an equidistant spacing between the aligned scan and the edges of the (160,160,160) volume. Dataset $\mathfrak{D}_A$ was split into a 829 set for training and validation, and a 356 hold-out set for testing. This split was done by randomly shuffling the scans based on their gestational age in days, ensuring representative distributions with a Kolmogorov-Smirnov test yielding an $s - value < 0.05$ and a $p - value > 0.95$. This dataset was used for the early development stages of fBEN (Ch. 4) and fBAN (Ch. 5).

Dataset $\mathfrak{D}_B$: consists of a set $\mathbb{S}_B$ of 4290 high- and low-quality scans, and the corresponding binary masks $\mathbb{M}_B$ and alignment parameters $\mathbb{P}_B$. However, unlike $\mathbb{P}_A$, the parameters $\mathbb{P}_B$ align the scans to a canonical space that is consistent with the space of MNI152 [160][161][162], which is a brain atlas derived 152 structural MRI images, averaged together through high-dimensional nonlinear registration. This change in space was done with the goal of facilitating future comparisons between 3D US and 3D MRI. Dataset $\mathfrak{D}_B$ was split into a 3217 set for training and validation, and a 1073 hold-out set for testing. This split was done using a stratified approach, randomly shuffling the scans based on their gestational age in days as well as the structural similarity between scans, based on the Structural Similarity Index Measure (SSIM) [163], thus ensuring that the structural characteristics of the scans are evenly distributed. A Kolmogorov-Smirnov test yielded $s - value < 0.02$ and $p - value > 0.90$ for the age distribution, and $s - value < 0.02$ and a $p - value > 0.94$ for the structural similarity distribution. Note that the split was performed in such a manner, that the splits of $\mathfrak{D}_A$ are subsets of the corresponding split of $\mathfrak{D}_B$. This dataset was used for the final refinement of fBEN (Ch. 4) and fBAN (Ch. 5).

Dataset $\mathfrak{D}_C$: consists of the aligned and masked versions of $\mathfrak{D}_B$, as shown in Fig.3.5. Therefore, the set of scans $\mathbf{S}_C$ and the corresponding set of masks $\mathbf{M}_C$ were split in the same manner as $\mathfrak{D}_B$. This dataset was used for the early development stages of fBFN (Ch. 6).

Dataset $\mathfrak{D}_D$: was generated by mirroring the volumes of $\mathfrak{D}_C$ along the mid-sagittal plane, so that the information-rich distal hemisphere of the brain is always located on the left side. Additionally, the scans were cropped to as size of (78,144,126), removing the right hemisphere. The set of scans $\mathbf{S}_D$ and the corresponding set of masks $\mathbf{M}_D$ were split in the same manner as $\mathfrak{D}_B$. This dataset was used for the intermediate development stages of fBFN (Ch. 6).

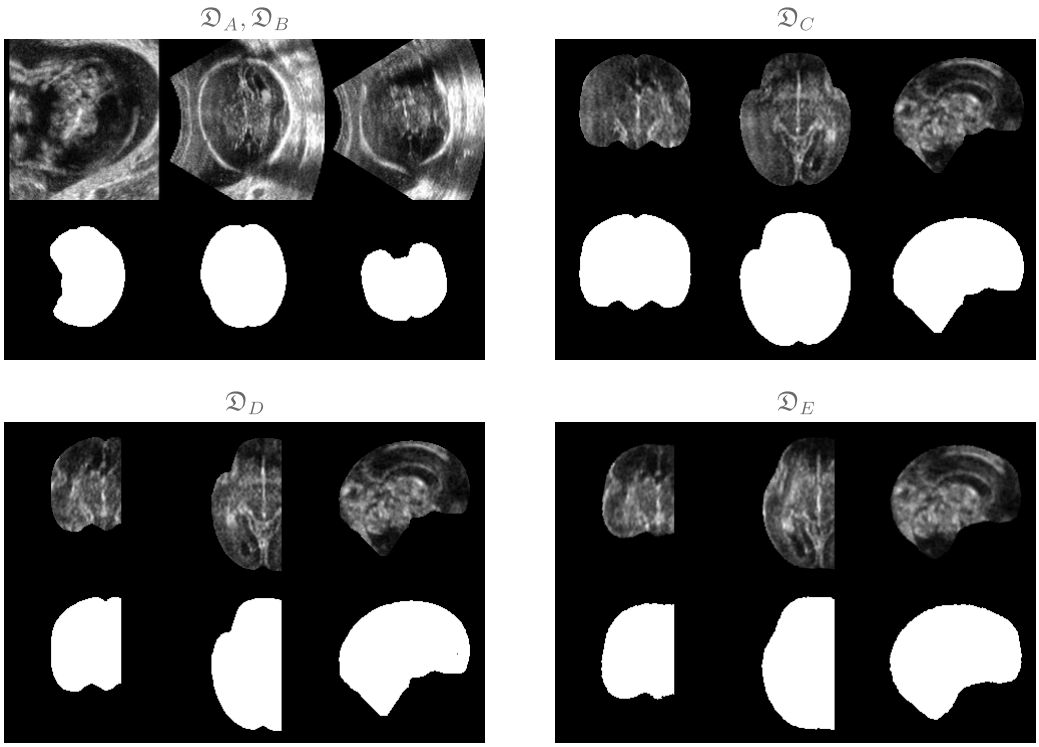

Figure 3.5: Mid-orthogonal planes of an example scan **S** and mask **M** of the different datasets used in this thesis. For each dataset, the same original 25 GW scan was used. Note that while $\mathfrak{D}_A$ and $\mathfrak{D}_B$ were generated using a different number of scans, and a different canonical space, the scans and masks in the original position are nearly identical, and have therefore been grouped together.

Dataset $\mathfrak{D}_E$: was generated in the same manner as $\mathfrak{D}_D$, with the distinction that the sets of scans $\mathbb{S}_E$ and masks $\mathbb{M}_E$ were generated entirely through the automated predictions created by the automated approaches developed in this thesis, namely fBEN (Ch. 4) and fBAN (Ch. 5). This dataset consists of 13779 scans, split into 10334 for training and validation, and a hold-out testing set of 3445. This split was done by randomly shuffling the scans based on their gestational age in days, ensuring representative distributions with a Kolmogorov-Smirnov test yielding an $s{-}value < 0.0015$ and a $p{-}value > 0.99$. Note that the split was performed in such a manner, that the splits of $\mathfrak{D}_A$ and $\mathfrak{D}_B$ (and therefore $\mathfrak{D}_C$ and $\mathfrak{D}_D$) are subsets of the corresponding split of $\mathfrak{D}_E$. This dataset was used for the final refinement of fBFN (Ch. 6).

4

Automated Fetal Brain Extraction

Contents

4.1 Introduction

As discussed in Chapter 1, one of the main factors limiting the widespread adoption of 3D US for medical research and clinical assessment of the fetal brain is the difficulty associated with their analysis. Without robust, automated methods that can facilitate the processing and analysis of these scans, this is unlikely to change.

One of the most important tasks is that of separating the cerebral tissue from the rest of the data in the 3D scan, i.e. brain extraction. In addition to allowing for

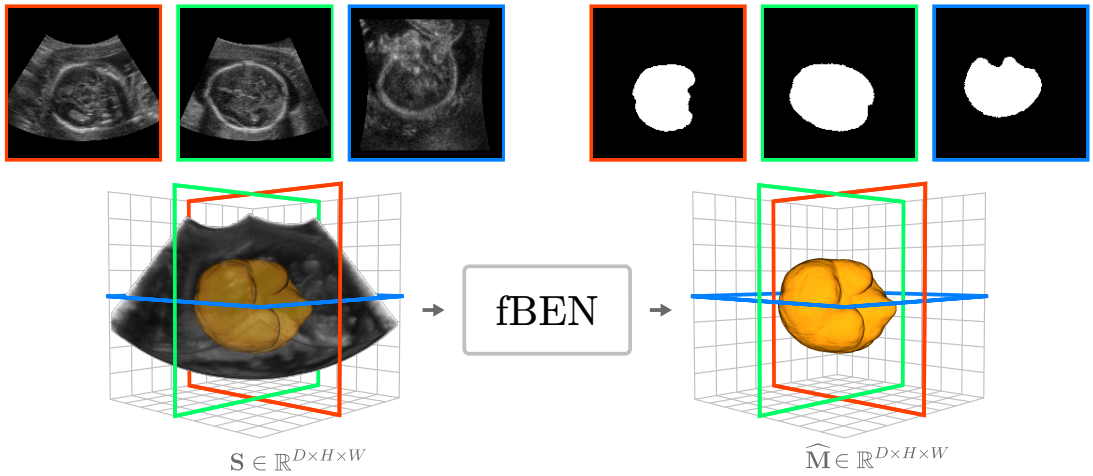

Figure 4.1: Graphical abstract of the fetal Brain Extraction Network (fBEN).

the analysis of the morphology of the brain, this task ensures that extra-cerebral data such as the skull, amniotic fluid, maternal tissues, and background do not affect any subsequent analysis. However, manually annotating each 3D scan is a tedious, time-consuming task that requires a high level of expertise. While this problem has led to the development of a multitude of robust, automated brain extraction solutions for more traditional 3D neuroimaging modalities such as MRI, there is currently no equivalent solution for fetal 3D US.

However, developing a reliable, automated US-specific solution for automated brain extraction is challenging. Conventional image processing methods struggle to perform well with 3D US data, due the low contrast and high intrinsic variability of the data, as well as irregular artefacts such as acoustic shadows, reverberations, and structural asymmetry

Thankfully, the last decade has seen an explosion in the use of AI for medical images, with an extensive number of DL solutions achieving state-of-the-art performance for a broad range of tasks, such as image classification, segmentation, and registration [164]. Specifically, DL solutions have shown great potential for processing US image data, in spite of the aforementioned challenges, making it an ideal candidate for developing a robust, automated fetal brain extraction solution.

In this chapter I propose the first module of the Fully-Automated DL Pipeline for 3D Fetal Brain Ultrasound: the fetal Brain Extraction Network (fBEN). This is

an end-to-end, fully-CNN that extracts the fetal brain from minimally pre-processed, standard clinical 3D US scans, as shown in Fig. 4.1. fBEN is the first of its kind for 3D neurosonography, directly extracting the 3D brain from the image without relying on shape approximations or 2D predictions.

Through exhaustive analysis, I confirm that the network achieves state-of-the-art performance, comparable with the current state-of-the-art solutions for fetal MRI. I demonstrate that fBEN performs consistently for the entire gestational range of 14.0 to 30.9 GW. I also show that the performance of the network is invariant to the location and orientation of the fetal brain in the 3D scan.

The contributions in this chapter are:

- Development of fBEN, an end-to-end, fully-CNN that extracts the fetal brain from minimally pre-processed, standard clinical 3D US scans.
- Development of the first and currently only that directly extracts the fetal brain from 3D US scans without relying on shape approximations or 2D predictions.

4.2 Methods

In this section I describe the iterative process of developing the fBEN, from its initial developmental stages up to its final refinement. I also cover the data used, the measures chosen for a multifaceted assessment of performance, and the implementation details of the method.

4.2.1 Data

The initial development of fBEN, described in Sec. 4.2.4, was performed with dataset $\mathfrak{D}_A$ (Sec. 3.4). $\mathfrak{D}_A$ is made up of n=1185 high-quality 3D US scans of the fetal brain $\mathbb{S}_A$, and the corresponding set of brain masks $\mathbb{M}_A$. This dataset was randomly split (proportionally to the gestational ages) into a 356 hold-out set for testing, and an 829 set for training and validation.

For the final refinement of fBEN, described in Sec. 4.2.5, dataset $\mathfrak{D}_B$ (Sec. 3.4) was used instead. $\mathfrak{D}_B$ contains n=4290 scans $\mathbb{S}_B$ and their corresponding masks $\mathbb{M}_B$. In contrast to $\mathfrak{D}_A$, the scans of $\mathfrak{D}_B$ have been selected to represent the variability of the INTERGROWTH-21$^{\text{st}}$dataset, containing both high and low quality scans. $\mathfrak{D}_B$ was split into a 1073 hold-out set for testing, and a 3217 set for training and validation. However, unlike $\mathfrak{D}_A$, this split was performed using an iterative stratification approach that evenly distributed the data based on the gestational age of the scans, as well as the structural similarity of the brain.

Both datasets span the gestational age range of 14.1 to 30.9 GW (99 to 216 gestational days). The pre-processing of the scans was limited to resampling to an isotropic voxel size of 0.6 mm/vxl, subsequently centre-cropping to a size of (160,160,160), and finally normalising their features, i.e. intensities, to within 0 and 1.

Note that the manual alignment of $\mathfrak{D}_A$, and the manual alignment of $\mathfrak{D}_B$ were performed by different users, and are in different canonical spaces. Consequentially, while the scans of $\mathbb{S}_A$ are contained in $\mathbb{S}_B$, their corresponding masks in $\mathbb{M}_B$ are not identical to those in $\mathbb{M}_A$.

A more detailed description of the data used in this thesis can be found in Chapter 3.

4.2.2 Implementation details

During initial development, the networks were implemented in Python using the Tensorflow [165] and Keras [166] libraries. The networks were developed on an Intel Xeon E-2146G CPU (3.50GHz, 6 cores) and an Nvidia GTX 1080 Ti. For the final refinement, the networks were re-implemented in Python using the Pytorch [167] library, and an Nvidia A10. Unless specified, all fBEN networks were trained with a batch size $bs = 4$, the Adam [168] optimiser, and an empirically selected learning rate $lr = 0.01$. All fBEN networks were trained using a 3-fold cross-validation, and tested against a hold-out dataset.

4.2.3 Evaluation measures

In order to perform a thorough evaluation of the predictions of fBEN, I selected four different measures that reflect different aspects of performance. Combined, they facilitate a quick understanding the strengths and weaknesses of the network.

The first measure used is the Centroid Distance (CD). This measure represents the Euclidean distance between the centroid of the reference mask $\mathbf{M}$ and the centroid of the predicted mask $\widehat{\mathbf{M}}$, and reflects whether the network is localising the brain correctly in the scan $\mathbf{S}$. The volumes are comprised of $m = 160^3$ voxels, with the i-th voxels $\mathbf{M}_i$ and $\widehat{\mathbf{M}}_i$ having the binary value $\mathbf{M}(x_i, y_i, z_i)$ and $\widehat{\mathbf{M}}(x_i, y_i, z_i)$, respectively. The centroid $(\bar{x}_\mathbf{M}, \bar{y}_\mathbf{M}, \bar{z}_\mathbf{M})$ of $\mathbf{M}$ is defined as shown in Eq. (4.1), from which the CD is calculated as shown in Eq. (4.2). This follows analogous for $\widehat{\mathbf{M}}$.

$$(\bar{x}_\mathbf{M}, \bar{y}_\mathbf{M}, \bar{z}_\mathbf{M}) = \frac{1}{\sum_{i=1}^{m} \mathbf{M}_i} \left(\sum_{i=1}^{m} (\mathbf{M}_i \cdot x_i), \sum_{i=1}^{m} (\mathbf{M}_i \cdot y_i), \sum_{i=1}^{m} (\mathbf{M}_i \cdot z_i) \right) \tag{4.1}$$

$$\mathrm{CD}(\mathbf{M}, \widehat{\mathbf{M}}) = \sqrt{(\bar{x}_\mathbf{M} - \bar{x}_{\widehat{\mathbf{M}}})^2 + (\bar{y}_\mathbf{M} - \bar{y}_{\widehat{\mathbf{M}}})^2 + (\bar{z}_\mathbf{M} - \bar{z}_{\widehat{\mathbf{M}}})^2} \tag{4.2}$$

To quantify the overlap between the reference mask $\mathbf{M}$ and the predicted mask $\widehat{\mathbf{M}}$, I rely on the Dice Similarity Coefficient (DSC), also known as the Sørensen–Dice coefficient. This measure determines how similar $\widehat{\mathbf{M}}$ is to $\mathbf{M}$, with a value of 0 for no overlap between the masks, and 1 for identical volumes. For the i-th voxel in $\widehat{\mathbf{M}}$, it is defined as true positive(TP) if $\widehat{\mathbf{M}}_i = 1 = \mathbf{M}_i$, false positive (FP) if $\widehat{\mathbf{M}}_i = 1 \neq \mathbf{M}_i$, true negative (TN) if $\widehat{\mathbf{M}}_i = 0 = \mathbf{M}_i$, and false negative (FN) if $\widehat{\mathbf{M}}_i = 0 \neq \mathbf{M}_i$. With this, the DSC is calculated as shown in Eq. (4.3)

$$\mathrm{DSC}(\mathbf{M}, \widehat{\mathbf{M}}) = \frac{2TP}{2TP + FP + FN} \tag{4.3}$$

One of the limitations of using DSC as a measure of overlap is that it does not give return any information regarding the regional performance. When extracting the brain, it is crucial to know whether the mismatched areas between $\mathbf{M}$ and $\widehat{\mathbf{M}}$ are evenly distributed around the borders, or localised around a specific region of the

brain. To evaluate this, I rely the Hausdorff Distance (HD). Define sets A and B as the locations of positive voxels of the binary masks $\mathbf{M}$ and $\widehat{\mathbf{M}}$, as shown in Eq. (4.4).

$$\begin{aligned} \mathrm{A} &= \{(x,y,z) \in \mathbb{Z}^3 | \mathbf{M}(x,y,z) = 1\} \\ \mathrm{B} &= \{(x,y,z) \in \mathbb{Z}^3 | \widehat{\mathbf{M}}(x,y,z) = 1\} \end{aligned} \tag{4.4}$$

The HD is the maximum minimum distance between every point in set A and every point in set B, as shown in Eq. (4.5). For the same DSC value, a lower HD value indicates that the discrepancies between $\mathbf{M}$ and $\widehat{\mathbf{M}}$ are more evenly distributed, while a higher HD value indicates a more localised discrepancy.

$$\mathrm{HD}(\mathbf{M}, \widehat{\mathbf{M}}) = \max\left(\max_{a\in A}\min_{b\in B}\|a-b\|, \max_{b\in B}\min_{a\in A}\|a-b\|\right) \tag{4.5}$$

Finally, the symmetry of the predicted masks needs to be assessed. As discussed in Sec. 4.1, 3D US have an intrinsic asymmetry regarding the structural information of the fetal brain. To ensure that this asymmetry does not translate into an asymmetric prediction $\widehat{\mathbf{M}}$, I define the Symmetry Coefficient (SC) as the DSC between the predicted hemispheres of the brain. If $\widehat{\mathbf{M}}^{\boldsymbol{p}}$ is the predicted mask $\widehat{\mathbf{M}}$, aligned using the reference alignment parameters $\boldsymbol{p}$, $\widehat{\mathbf{M}}^{\boldsymbol{p}}$ is split along the midsagittal plane, resulting in the predictions for the left-hemisphere $\widehat{\mathbf{M}}^{\boldsymbol{p}}_L$ and the right-hemisphere $\widehat{\mathbf{M}}^{\boldsymbol{p}}_R$. Finally, $\widehat{\mathbf{M}}^{\boldsymbol{p}}_{right}$ is mirrored to match the orientation $\widehat{\mathbf{M}}^{\boldsymbol{p}}_L$, resulting in $\widehat{\mathbf{M}}^{\boldsymbol{p}}_{R^*}$. The SC is then calculated as shown in Eq. (4.6).

$$\mathrm{SC}(\widehat{\mathbf{M}}) = DSC(\widehat{\mathbf{M}}^{\boldsymbol{p}}_L, \widehat{\mathbf{M}}^{\boldsymbol{p}}_{R^*}) \tag{4.6}$$

The combined use of $\mathrm{CD}(\mathbf{M}, \widehat{\mathbf{M}})$, $\mathrm{DSC}(\mathbf{M}, \widehat{\mathbf{M}})$, $\mathrm{HD}(\mathbf{M}, \widehat{\mathbf{M}})$, and $\mathrm{SC}(\widehat{\mathbf{M}})$ results in a thorough, detailed assessment of the performance of fBEN, allowing for the quick and efficient analysis and comparison of the results throughout development, as well as against alternative methods.

To assess the statistical significance when comparing these measures, the normality is first determined with a D'Agostino and Pearson's test [169], followed by a paired Student's t-test [170] or a Wilcoxon signed-rank test [171], for normal and non-normal samples, respectively. A significance threshold of $p < 0.05$ is used for both tests.

4.2.4 Initial development

As mentioned in Sec. 1.1, one of the most difficult aspects of developing a robust fBEN for 3D US is the challenging nature of the data itself. In particular, a tool that can reliably handle the large intrinsic variability of the 3D US scans is need. Therefore, the first step was to find a suitable DL architecture that would work well with the available data. Inspired by the U-Net [100], I decided to use an encoder-decoder architecture with skipped connections, as this general architecture has been shown to handle large data variability well.

I started the development by building a basic network upon which I would then develop fBEN. As shown in Fig. 4.2, it consisted of 4 MaxPooling layers, with convolutional blocks comprised of two subsequent sets of Convolutional layer, a Batch Normalisation layer, and a ReLU activation layer. A final Convolutional layer followed by a Sigmoid activation layer outputs the predicted mask $\widehat{\mathbf{M}}$, which has the same dimensions as the input scan $\mathbf{S}$. I empirically chose an initial kernel size $ks = (3, 3, 3)$ and hidden dimensions $hd = (16, 32, 64, 64, 64, 64, 64, 64)$ of the convolutional blocks, as this showed the best performance in the pre-development phase.

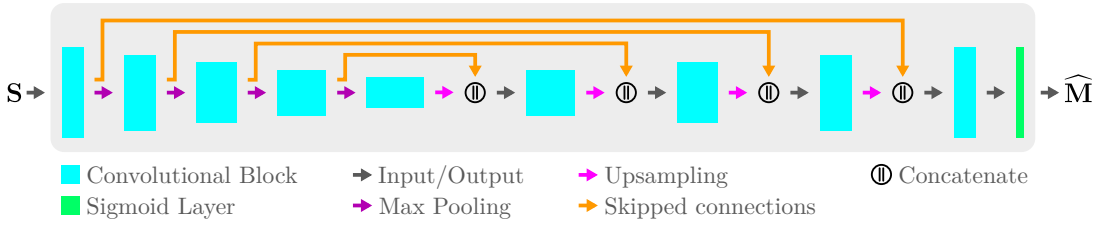

Figure 4.2: Schematics of the initial architecture of fBEN.

To train the network, I chose to use the DSC($\mathbf{M}, \widehat{\mathbf{M}}$) (Eq. (4.3)) measure as a loss function $\mathfrak{L}_{\text{DSC}}$. The main reason for this is its robustness and simplicity, for which is has been widely used for segmentation tasks [172][173]. While there are other possible measures, such as Binary Cross-Entropy (BCE), that have been proven to achieve better performance in some circumstances [174], the noisy nature of the reference labels $\mathbf{M}$, as discussed in Sec 3.3 means that achieving a perfect prediction is unwanted. $\mathfrak{L}_{\text{DSC}}$ is calculated as shown in Eq. (4.7), with smoothing factor $s = 1$. Since thresholding the predicted mask $\widehat{\mathbf{M}}$ is not differentiable,

definition shown in Eq. (4.3) is not implementable. Instead, this approach is known as a soft DSC loss, which directly uses the predicted probabilities, allowing the implementation to remain differentiable. The smoothing factor s is added to stabilise the division by a weak denominator.

$$\mathfrak{L}_{\mathrm{DSC}} = 1 - \mathrm{DSC}(\mathbf{M}, \widehat{\mathbf{M}}) = 1 - \frac{2 \cdot \sum_{i=1}^{m} (\mathbf{M}_i \cdot \widehat{\mathbf{M}}_i) + s}{\sum_{i=1}^{m} \mathbf{M}_i + \sum_{i=1}^{m} \widehat{\mathbf{M}}_i + s} \tag{4.7}$$

During the initial development, I constrained the task of brain extraction by reducing the spatial resolution of the $\mathfrak{D}_A$ dataset from 0.6 mm/vxl to 1.2 mm/vxl. While this results in a coarser prediction, the resulting (80,80,80) volume contains 1/8 of the voxels, allowing for faster development thanks to a greatly reduced computational cost. Additionally, the downsampling results in an effective reduction of speckle and noise, as the voxels are averaged together, which should make the task easier for the network to learn. This makes the initial optimisation of the network significantly faster, after which the full resolution scans can be reintroduced for the final refinement. Nevertheless, for the sake of consistency, unless explicitly stated, all low-resolution predictions have been compared against the reference masks after being binarised with a threshold of 0.5, upsampled to (160,160,160) with a spline interpolation of order 3, and binarised again with a threshold of 0.5.

I began the iterative optimisation of fBEN by determining the optimal kernel size ks for the Convolutional Layers of the network. As shown in Tab. 4.1, I compared three different kernel sizes and evaluated their performance using $\mathrm{CD}(\mathbf{M}, \widehat{\mathbf{M}})$, $\mathrm{HD}(\mathbf{M}, \widehat{\mathbf{M}})$, and $\mathrm{DSC}(\mathbf{M}, \widehat{\mathbf{M}})$ as measures. The network with $ks = (3, 3, 3)$ resulted in the highest performance for $\mathrm{CD}(\mathbf{M}, \widehat{\mathbf{M}})$ and $\mathrm{DSC}(\mathbf{M}, \widehat{\mathbf{M}})$, outperforming the next best by 12.5% and 1%, respectively. While this configuration was outperformed in the $\mathrm{HD}(\mathbf{M}, \widehat{\mathbf{M}})$ measure, only $\mathrm{CD}(\mathbf{M}, \widehat{\mathbf{M}})$ and $\mathrm{DSC}(\mathbf{M}, \widehat{\mathbf{M}})$ resulted in statistically significant differences. Therefore, I continued the development using the convolutional kernel size $ks = (3, 3, 3)$, which has the added benefit of resulting in the lowest number of trainable network parameters and therefore the lowest computational cost.

Table 4.1: Performance comparison of using different kernel sizes *ks* for the Convolutional layers of the network, along with the number of trainable parameters. The best performance for each measure (Centroid Distance CD, Dice Similarity Coefficient DSC, Hausdorff Distance HD) is highlight in bold. The arrows indicate whether a higher (up) or lower (down) value is preferred.

Kernel size *ks*	CD[mm] ↓	DSC ↑	HD[mm] ↓	Param.
(3x3x3)	$\mathbf{1.47 \pm 0.85}$	$\mathbf{0.94 \pm 0.02}$	10.50 ± 5.85	1.6 M
(5x5x5)	1.68 ± 1.20	0.93 ± 0.03	$\mathbf{9.81 \pm 3.99}$	7.4 M
(7x7x7)	1.67 ± 0.85	0.93 ± 0.03	10.05 ± 3.93	20.3 M

I continued the iterative optimisation by focusing on the optimal number of hidden dimensions *hd* of the network, i.e., the output channels of the convolutional blocks. The Convolutional layers of the same convolutional block share the same number of output channels, while the last Convolutional layer always has a single output channel. I compared three configurations that share the same distribution of hidden dimensions throughout the network as the initial configuration. As shown in Tab. 4.2, a configuration with $hd = (8, 16, 32, 32, 32, 32, 32, 32)$ resulted in the best overall performance. While all configurations achieved almost identical CD($\mathbf{M}, \widehat{\mathbf{M}}$) and DSC($\mathbf{M}, \widehat{\mathbf{M}}$) performance, this configuration achieved a statistically significant improvement of 13.8% for HD($\mathbf{M}, \widehat{\mathbf{M}}$), for which it was used for the next steps of development.

Table 4.2: Performance comparison of using different hidden dimensions *hd* for the Convolutional blocks of the network, along with the number of trainable parameters. The best performance for each measure (Centroid Distance CD, Dice Similarity Coefficient DSC, Hausdorff Distance HD) is highlight in bold. The arrows indicate whether a higher (up) or lower (down) value is preferred.

Hidden dimensions *hd*	CD[mm] ↓	DSC ↑	HD[mm] ↓	Param.
4	1.47 ± 0.85	0.94 ± 0.02	10.50 ± 5.85	0.1 M
8	$\mathbf{1.36 \pm 0.72}$	$\mathbf{0.94 \pm 0.02}$	$\mathbf{9.05 \pm 3.56}$	0.4 M
16	1.37 ± 0.70	0.94 ± 0.02	10.90 ± 6.22	1.6 M

For the third and final optimisation, I determined the optimal depth of the network by comparing the performance architectures that use different number of pooling layers *l*. As Tab. 4.3 shows, I compared networks with 2, 3, 4, and 5

pooling layers, with $l = 4$ resulting in the optimal number. The results indicated that shallower networks resulted in significant lower performance across the board, with $l = 4$ significantly outperforming $l = 3$ by 6.6% for CD($\mathbf{M}, \widehat{\mathbf{M}}$), 1% for DSC($\mathbf{M}, \widehat{\mathbf{M}}$), and 24.6% for HD($\mathbf{M}, \widehat{\mathbf{M}}$). In contrast, a deeper network offered no additional benefits, with no statistically significant difference in performance. Therefore, a depth of $l = 4$ was chosen. .

Table 4.3: Performance comparison of using different number of pooling layers l, along with the number of trainable parameters. The best performance for each measure (Centroid Distance CD, Dice Similarity Coefficient DSC, Hausdorff Distance HD) is highlight in bold. The arrows indicate whether a higher (up) or lower (down) value is preferred.

Pooling Layers l	CD[mm] ↓	DSC ↑	HD[mm] ↓	Param.
2	1.98 ± 1.05	0.90 ± 0.04	22.89 ± 8.99	0.1M
3	1.45 ± 0.77	0.93 ± 0.02	12.00 ± 4.72	0.3M
4	$\mathbf{1.36 \pm 0.72}$	$\mathbf{0.94 \pm 0.02}$	$\mathbf{9.05 \pm 3.56}$	0.4M
5	1.38 ± 0.74	0.94 ± 0.02	9.23 ± 3.61	0.5M

At this point of development, I had a well-optimised network, so I focused on determining which threshold t resulted in the best performance. So far, the output of the Sigmoid layer has been binarised using a default threshold $t = 0.5$, i.e. the predicted mask has been comprised of the voxels where fBEN has 50% confidence or more. However, since the optimal performance can vary depending on the confidence of the predictions, I compared the performance of fBEN by thresholding its predictions with four additional values between 0 and 1. At this point in development, I also decided to include the SC($\widehat{\mathbf{M}}$) measure in the assessment, as it provides valuable information regarding the symmetry of the prediction. The results shown in Tab. 4.4 indicate that the original choice of $t = 0.5$ resulted in the best overall performance, sharing the first place with $t = 75$ for CD($\mathbf{M}, \widehat{\mathbf{M}}$), and with $t = 0.25$ for DSC($\mathbf{M}, \widehat{\mathbf{M}}$). With the exception of $t = 0$, all thresholds resulted in a similar HD($\mathbf{M}, \widehat{\mathbf{M}}$) performance, with no statistically significant differences. The best SC($\widehat{\mathbf{M}}$) performance, however, was achieved with a threshold of $t = 1$. Nevertheless, this is due to an overzealous erosion of the outer shell of the predicted mask $\widehat{\mathbf{M}}$, where the confidence is lower, resulting in

a less accurate but more symmetrical mask. Finally, it is worth noting that the consistent performance between $t = 0.25$, $t = 0.5$, and $t = 0.75$ highlights the high confidence that the fBEN has in its predictions. The only substantial difference between these three is a 16.8% drop in the SC($\widehat{\mathbf{M}}$) performance for $t = 0.75$, indicating a slightly asymmetric confidence of $\widehat{\mathbf{M}}$, likely due to the asymmetric structural information of the scanned brain discussed in Sec. 1.1. This configuration of fBEN was originally published in [105], and I will refer to it as fBENv0 from now on. Its schematics are shown in Fig. 4.3a.

Table 4.4: Performance comparison of using different thresholds t to binarise the predicted mask $\widehat{\mathbf{M}}$. The best performance for each measure (Centroid Distance CD, Dice Similarity Coefficient DSC, Hausdorff Distance HD, Symmetry Coefficient SC) is highlight in bold. The arrows indicate whether a higher (up) or lower (down) value is preferred.

Threshold t	CD[mm] ↓	DSC ↑	HD[mm] ↓	SC ↑
0	8.34 ± 3.26	0.27 ± 0.15	62.50 ± 7.91	0.80 ± 0.02
0.25	1.43 ± 0.93	$\mathbf{0.94 \pm 0.02}$	9.24 ± 4.77	0.95 ± 0.02
0.5	$\mathbf{1.36 \pm 0.72}$	$\mathbf{0.94 \pm 0.02}$	9.05 ± 3.56	0.95 ± 0.02
0.75	$\mathbf{1.36 \pm 0.72}$	0.93 ± 0.03	8.97 ± 3.54	0.80 ± 0.02
1	1.42 ± 0.80	0.90 ± 0.03	$\mathbf{8.72 \pm 4.23}$	$\mathbf{0.99 \pm 0.01}$

As a final step in the initial development, I investigated whether thresholding the predicted mask $\widehat{\mathbf{M}}$ after upsampling could result in a more accurate prediction. As stated before, the low-resolution predictions have been thresholded and upsampled before comparing with the high-resolution reference mask $\mathbf{M}$. However, this sequence results in all voxels being equally weighted during the upsampling interpolation. By interpolating before thresholding, the hypothesis was that the prediction confidence of each voxel would act as an effective weighting during interpolation, resulting in more accurate edges. The results shown in Tab. 4.5 confirmed this, with interpolating before thresholding resulting in a 27.9% improvement for CD($\mathbf{M}, \widehat{\mathbf{M}}$), and 13.0% for HD($\mathbf{M}, \widehat{\mathbf{M}}$), both statistically significant. In contrast, no statistically significant changes were observed for DSC($\mathbf{M}, \widehat{\mathbf{M}}$) and SC($\widehat{\mathbf{M}}$), highlighting the importance of relying on multiple measures for a comprehensive performance assessment. While statistically significant, the improvements are relatively small. Nevertheless, they

require no changes to the network architecture, and result in a negligible additional computational cost. This fBEN was published as part of [106] and I will refer to it as fBENv1 from now on. Its schematics can be seen in Fig. 4.3b.

Table 4.5: Performance comparison of upsampling the low-resolution predicted mask before or after thresholding it. The best performance for each measure (Centroid Distance CD, Dice Similarity Coefficient DSC, Hausdorff Distance HD, Symmetry Coefficient SC) is highlight in bold. The arrows indicate whether a higher (up) or lower (down) value is preferred.

Upsampling	CD [mm] ↓	DSC ↑	HD [mm] ↓	SC ↑
Before threshold	$\mathbf{0.98 \pm 0.53}$	$\mathbf{0.94 \pm 0.02}$	$\mathbf{7.87 \pm 3.12}$	$\mathbf{0.95 \pm 0.02}$
After threshold	1.36 ± 0.72	0.94 ± 0.02	9.05 ± 3.56	0.95 ± 0.02

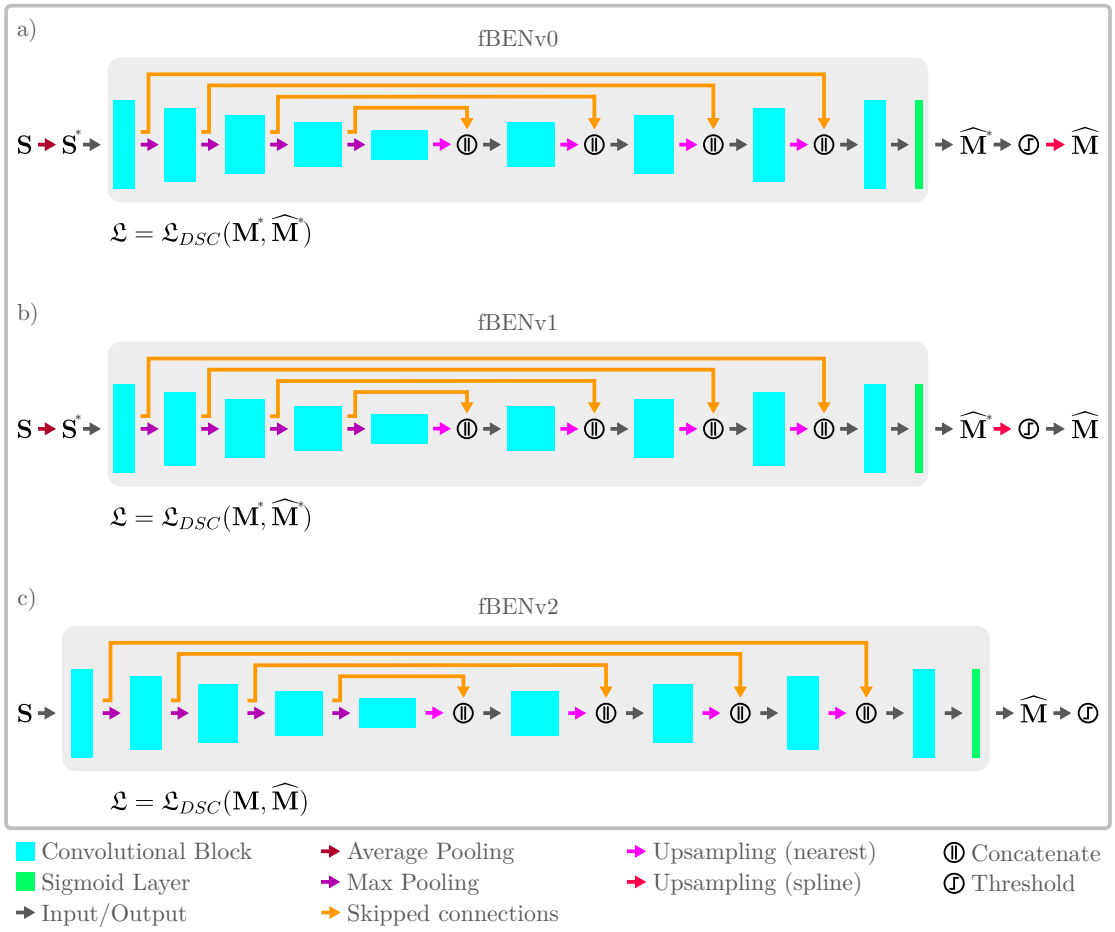

Figure 4.3: Schematics of the architecture of a) fBENv0, b) fBENv1, and c) fBENv2. The notation $\mathbf{S}^*$, $\mathbf{M}^*$, and $\widehat{\mathbf{M}}^*$ represents the low-resolution versions of the reference scan $\mathbf{S}$, the reference mask $\mathbf{M}$, and the predicted mask $\mathbf{M}$. Note that the thresholding is only performed during testing and not during training or validation.

4.2.5 Final refinement

At this stage, I had finished the initial development of fBEN. I had fully optimised the network, which had already achieved state-of-the-art performance for the task of fetal brain extraction from 3D US scans. Additionally, fBEN had also surpassed the performance of the equivalent state-of-the-art solutions for fetal MRI, such as Salehi et al. 2017 [175] and Ebner et al. 2020 [122]. A full analysis of the performance of fBENv0 and fBENv1, as well as alternative methods, can be found in Sec.4.3. However, there were still a couple of issues that I wanted to address.

Firstly, while the 1185 scans contained in dataset $\mathfrak{D}_A$ are significantly more than the number of scans used in other relevant works, covering the gestational range from 14.0 to 30.9 GW, the scans were originally selected specifically avoiding particularly low-quality examples (see Sec. 3.4). However, since one of the criteria of the pipeline (see Sec. 1.1) is robustness, a dataset that better represents the true quality diversity of 3D US would be crucial. Therefore, the final refinement of fBEN was done with dataset $\mathfrak{D}_B$, which is comprised of 4290 scans spanning the same gestational range, but specifically selected to represent the quality range of the INTERGROWTH-21$^{\text{st}}$dataset. Additionally, in order to increase robustness even further, dataset $\mathfrak{D}_B$ has also been augmented by randomising the orientation of the input scans. The 24 possible orientations result in an effective increase from 4290 to 102960 scans. Along with increasing the diversity of the dataset, this augmentation also aims to force the network to perform consistently, regardless of the orientation of the input scan, which may vary from user to user. A detailed description of $\mathfrak{D}_B$ can be found in Sec. 3.4.

Secondly, both fBENv0 and fBENv1 were working under the constrained task of low-resolution (1.2 mm/vxl) input scans. This might be limiting the accuracy of the prediction, especially around the borders of the extraction mask. Therefore, a network working with full-resolution (0.6 mm/vxl) scans might be able to exceed the current performance. However, the higher computational cost of training this network became a limiting factor, as the 11 GB of memory available on the used Nvidia GTX 1080 Ti was not enough to fit the new network at the same configuration

as before. While this could be easily addressed by training the network on better hardware, such equipment is not easily accessible for everybody. At the time of developing this fBEN, an Nvidia GTX 1080 Ti was generally considered to be the best consumer-grade GPU for DL, with 11 GB of memory. Since the memory of the equivalent Nvidia RTX 3080 Ti has only increased to 12 GB at the time of writing of this thesis, the original of 11 GB of memory remains a reasonable limit. Instead, I solved this issue by implementing accumulated gradients [176]. As the name implies, the calculated gradients during training are accumulated for multiple iterations before updating the network parameters. Therefore, rather than training with a batch size of 10, for example, the same can be achieved with a batch size of 1 by accumulating the gradients of 10 iterations. While this approach is slower than simply using a larger batch size, it allows for the GPU memory limitations to be circumvented and for the high-resolution fBEN to be trained with the same configuration as before.

Finally, I performed a final optimisation of the network in the same iterative manner as discussed in Sec. 4.2.4. The final architecture of fBEN, which I will refer to as fBENv2, is shown in Fig. 4.3c. It has a convolutional kernel size of $ks = (3, 3, 3)$, a depth of $l = 4$, and $hd = (16, 32, 64, 128, 256, 128, 64, 32)$ hidden dimensions, and was trained for 100 epochs using the AdamW [177] optimiser with a learning rate of $lr = 0.001$. An exhaustive analysis of the performance of fBENv0, fBENv1, and fBENv2, can be found in Sec. 4.3.

4.3 Results

In this section I perform an exhaustive analysis of the performance of fBENv2, comparing it against alternative methods, as well as fBENv0 and fBENv1. I evaluate the performance by gestational age, brain region, and brain misalignment. Finally, I assess the robustness of the network against input orientation. Since the reference masks $\mathbf{M}$ of dataset $\mathfrak{D}_B$ are slightly different from those in $\mathfrak{D}_A$ (see Sec. 5.2.1), in order for the comparisons to be as accurate as possible, fBENv0 and

fBENv1 have been retrained using the scans from $\mathfrak{D}_A$ and the corresponding reference masks from $\mathfrak{D}_B$.

4.3.1 Mean performance

The mean performance of fBENv0, fBENv1, and fBENv2, on the hold-out testing split of dataset $\mathfrak{D}_B$ is shown in Tab. 4.6. Additionally, the performance of several alternative methods is also evaluated.

Table 4.6: Average performance of fBENv0, fBENv1, and fBENv2, compared with multiple alternative methods. All methods were evaluated using the hold-out testing split of the $\mathfrak{D}_B$ dataset.

Method	CD [mm] ↓	DSC ↑	HD [mm] ↓	SC ↑
SimpleElastix [178]	25.77 ± 9.88	0.51 ± 0.12	58.98 ± 19.42	0.69 ± 0.17
ANTS [179]	20.55 ± 13.15	0.56 ± 0.17	39.49 ± 16.94	0.62 ± 0.26
Salehi [175]	41.25 ± 26.13	0.00 ± 0.00	71.29 ± 21.70	0.03 ± 0.16
Ebner [122]	66.45 ± 34.48	0.00 ± 0.00	103.83 ± 25.75	0.20 ± 0.38
Namburete [104]	6.23 ± 6.84	0.69 ± 0.17	19.50 ± 4.38	0.73 ± 0.07
fBENv0	1.93 ± 0.94	0.75 ± 0.04	11.25 ± 2.78	0.95± 0.02
fBENv1	1.93 ± 0.94	0.77 ± 0.04	10.50 ± 2.78	0.94 ± 0.02
fBENv2	**1.53 ± 0.71**	**0.93 ± 0.02**	**5.73 ± 2.01**	**0.95 ± 0.02**

First, I explored the use of *SimpleElastix* [178] and *ANTS* [179] to register a template to each 3D US scan. The resulting transform could then be used to register a brain mask, similarly to the method used to generate the reference masks, as discussed in Sec. 3.3. For the template, I used the average masked and aligned scan for the corresponding GW. For each of the two methods I performed an exhaustive evaluation of the possible configurations, including several different metrics and transform types. However, neither method was able to register the template to the scans. This is reflected in the results shown in Tab. 4.6, which correspond to the best performing configuration for *SimpleElastix* and *ANTS*. While these methods have been extensively used in the literature, with state-of-the-art performance for multiple imaging modalities, the high intrinsic variability of the 3D US data, and the large positional variability seem to have hinder their performance.

I also explored the use of DL methods for fetal brain extraction from 3D MRI scans. As discussed in Sec. 1.1, the fundamental differences in the acquisition of MRI and US scans make the use of MRI methods for 3D US non-trivial. I highlight this by exploring the use of two of the current state-of-the-art methods: *Salehi* et al. [175] and *Ebner* et al. [122]. As expected, while the *Salehi* and *Ebner* methods managed a mean DSC of 0.92 and 0.94 when extracting the fetal brain from MRI scans, respectively, both failed to repeat their performance for 3D US scan, achieving a DSC of 0. These networks were trained and developed to work on very different data, so these results are not surprising. However, they reinforce the need for a US-specific solution.

The final comparison method was that of *Namburete* et al. [104]. As mentioned in Sec 2.3.1, this method was one of the best performing solutions for the task of automated fetal brain extraction from 3D US scans. As expected, it significantly outperformed all other comparison methods with a $\text{CD}(\mathbf{M}, \widehat{\mathbf{M}})$ of 6.23 mm, a $\text{DSC}(\mathbf{M}, \widehat{\mathbf{M}})$ of 0.69, a $\text{HD}(\mathbf{M}, \widehat{\mathbf{M}})$ of 19.50 mm, and a $\text{SC}(\widehat{\mathbf{M}})$ of 0.73. However, its performance is still significantly lower than the equivalent state-of-the-art methods for fetal MRI. Furthermore, it relies on approximating the fetal brain with an ellipsoid by fitting it onto compounded 2D segmentations. Not only is this a strong approximation of the brain shape, but it requires a precise and accurate fit for this approximation to be sensible. Its $\text{SC}(\widehat{\mathbf{M}})$ score of 0.73 shows that this can be a challenge, since an ellipsoid should have perfect symmetry if aligned correctly. An example of these limitations can be seen in Fig. 4.4. Therefore, this method is not accurate or robust enough for a general purpose 3D US pipeline, reinforcing the motivation for developing fBEN.

The results shown in Tab. 4.6 show that all versions of fBEN significantly outperformed every other method for every measure. However, when comparing the results for fBENv0 and fBENv1 against those shown in Tab. 4.5, a significant drop in performance for both fBENv0 and fBENv1 can be observed for every measure, except for $\text{SC}(\widehat{\mathbf{M}})$. The most significant of these was for $\text{DSC}(\mathbf{M}, \widehat{\mathbf{M}})$, where their performance decreased from 0.94 to 0.75 and 0.77, respectively. This

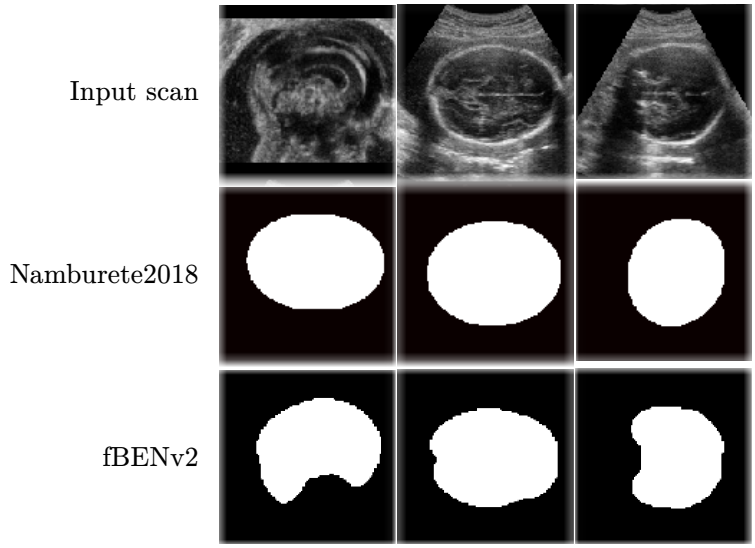

Figure 4.4: Qualitative comparison of the predicted mask generated using the method from Namburete et al. [104], compared to the mask predicted by fBENv2, for an example scan of 23 GWs.

is most likely a result of being trained on dataset $\mathfrak{D}_A$, which lacks the low-quality scans contained in $\mathfrak{D}_B$. The difference between fBENv0 and fBENv1 is also less dramatic, suggesting that thresholding before or after binarising is less significant when performance is overall lower. While fBENv0 and fBENv1 managed have state-of-the-art performance when dealing with good quality scans, they lack the robust performance to deal with more challenging data that is require for proposed pipeline. In contrast, fBENv2 managed to significantly outperform the previous version, achieving state-of-the-art performance for every measure, in spite of the more challenging dataset. In fact, fBENv2 managed to achieve a $\mathrm{HD}(\mathbf{M}, \widehat{\mathbf{M}})$ of 5.73 mm, which is 27.1% better than the previous versions managed on dataset $\mathfrak{D}_A$.

Figure 4.5 shows a comparison of the performance of fBENv0, fBENv1, and fBENv2 for each GW. While all three network manage to perform evenly throughout the gestational range, fBENv0 and fBENv1 show a slight drop in performance for the earlier and later weeks. In contrast, fBENv2 is significantly more consistent, without a visible drop at the edges. It also manages a significant increase in performance for $\mathrm{CD}(\mathbf{M}, \widehat{\mathbf{M}})$, $\mathrm{DSC}(\mathbf{M}, \widehat{\mathbf{M}})$, and $\mathrm{HD}(\mathbf{M}, \widehat{\mathbf{M}})$ at every GW, with the latter two showing the largest improvements, just like the results in Tab. 4.6 show. Additionally, the

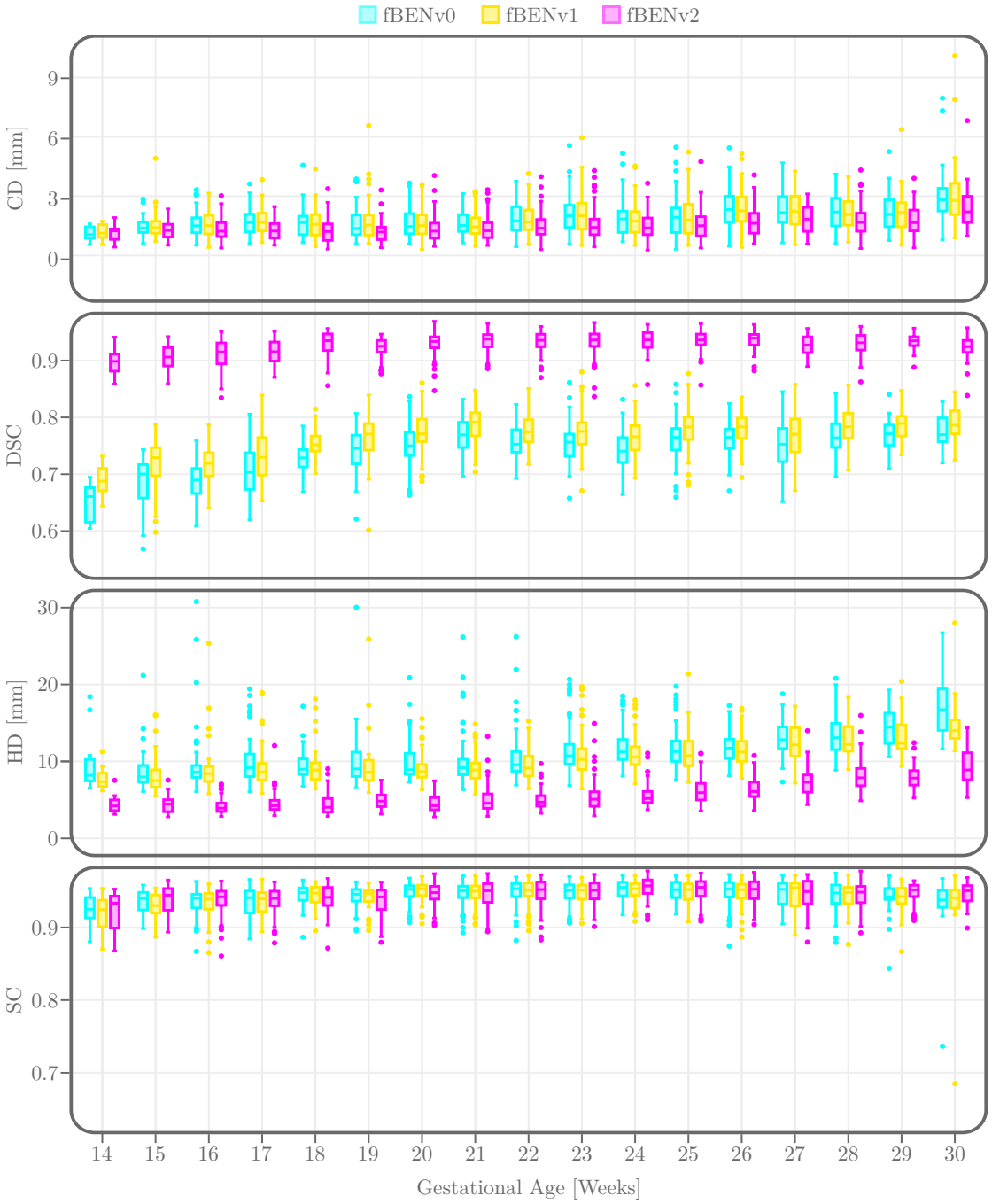

Figure 4.5: Performance of fBENv0, fBENv1, and fBENv2 for Centroid Distance CD($\mathbf{M}, \widehat{\mathbf{M}}$), Dice Similarity Coefficient DSC($\mathbf{M}, \widehat{\mathbf{M}}$), Hausdorff Distance HD($\mathbf{M}, \widehat{\mathbf{M}}$), Symmetry Coefficient SC($\widehat{\mathbf{M}}$), separated by GW.

performance for CD($\mathbf{M}, \widehat{\mathbf{M}}$), DSC($\mathbf{M}, \widehat{\mathbf{M}}$), and HD($\mathbf{M}, \widehat{\mathbf{M}}$) is significantly more consistent, with a smaller standard deviation, and fewer outliers.

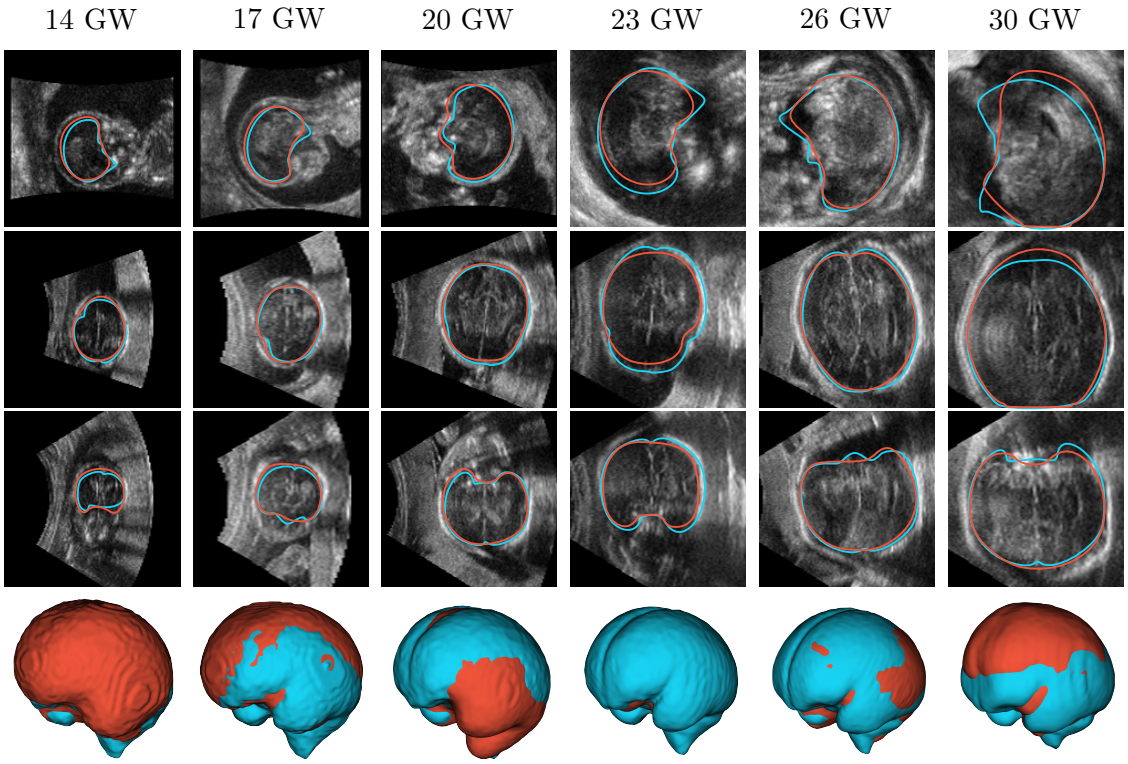

Figure 4.6: Examples of the reference mask $\mathbf{M}$ (blue) and the predicted mask $\widehat{\mathbf{M}}$ (red) generated by fBENv2 superimposed onto the mid-planes of the corresponding reference scan $\mathbf{S}$, for multiple GW. These examples were specifically selected since their DSC($\mathbf{M}, \widehat{\mathbf{M}}$) performance is closest to the mean performance of fBENv2 for that GW. Top: XY-plane. Middle: XZ-plane. Bottom: YZ-plane. Bellow each example is an aligned and scaled 3D rendering of the predicted mask overlayed on the ground-truth mask.

Example predictions $\widehat{\mathbf{M}}$ generated by fBENv2 for several gestational ages are shown in Fig.4.6, alongside the reference masks $\mathbf{M}$, overlaid onto the corresponding scan $\mathbf{S}$. In order for these examples to be a representative as possible, I have selected the predictions that had a DSC($\mathbf{M}, \widehat{\mathbf{M}}$) performance closest to the mean of their respective GW. The results show that fBENv2 performs accurately and consistently for the entire gestational range, closely matching the reference masks in most cases, as the 3D renders at the bottom of the figure show. These examples also show that the network manages to accurately use the structural information of the scan to adapt its prediction, rather than predicting the same mask for each GW. As such, the predictions generally offer a better result than the reference masks. This is

particularly evident in cases where the manual alignment was poor, such as the example at 30 GW. The network seems to achieve a similar accuracy, regardless of the location, orientation, or size of the brain of the subject. Similarly, aberrations, shadows, and artefacts appear to have no impact on the quality of the predictions.

4.3.2 Regional performance

Figure 4.7 shows a sliced axial view of the difference between the aligned and scaled reference mask $\mathbf{M}^p$ and the mean aligned and scaled predicted mask $\widehat{\mathbf{M}}^p$, for 14, 22, and 30 GW. The red areas indicate regions where the predictions segment less (under-segment) than the reference mask, while blue areas represent regions where the predictions segment more (over-segment), with darker shades indicating stronger discrepancies. The results show that the predictions generated by fBENv2 are generally consistent with the reference masks, with a low level of discrepancies around the borders. This is an expected consequence of using a single reference brain mask for every subject in a specific GW, as this does not take into account the morphological variability between subjects, something that is particularly clear for the 23 GW example of Fig. 4.6, where the occipito-frontal distance of the subject is smaller than that of reference mask, in spite of a similar biparietal distance.

Additionally, inconsistencies in the manual alignment, such as seen in the 30

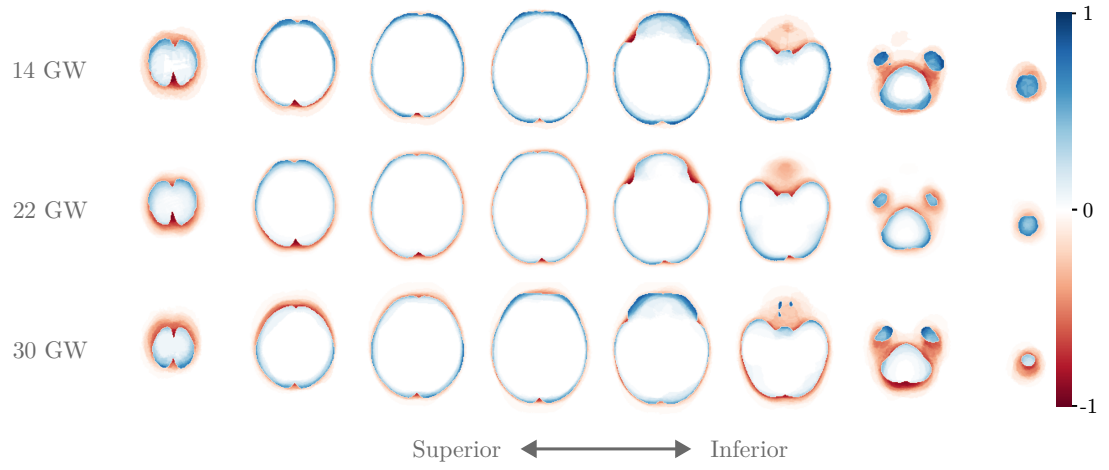

Figure 4.7: Mean regional performance of fBENv2 against the reference masks **M**. Positive and negative values indicate regions where fBENv2 is over-predicting and under-predicting, respectively. Gestational age is shown on the left.

GW example of Figure 4.6, also causes some discrepancies around the borders. However, there are three regions where stronger discrepancies are observed: the brain stem, as well as around the lateral and longitudinal sulci. These represent regions where the structural information acquired by MRI differs from US. While these structures are clearly visible when acquired with MRI, they are difficult to observe with 3D US. The higher amount of ossification around the brain stem due to the skull and spine limits what US can capture, while the low contrast between the brain and the cerebrospinal fluid makes them almost indistinguishable in US. Since the spatiotemporal atlas [159] used to generate the labels represents the structural information obtained through MRI, these discrepancies are not only expected, but welcomed, as they confirm that fBENv2 is creating a mask based on the observed structural information, and is not overfitting to the reference masks.

4.3.3 Performance vs. misalignment

So far I have analysed the mean performance of fBENv2 against alternative methods, its performance across the entire gestational range, and its regional performance. In this section I examine the impact that the misalignment of the fetal brain in the 3D US scan has on performance. The alignment parameters generated during the manual alignment of the data (see Sec. 3.2) can be used for this purpose by separating the assessment into three components. The first of these components is the *translation* misalignment, which will be defined as the Euclidean distance between the centre of the fetal brain and the centre of the 3D scan. The second is the *rotation* misalignment, which will be defined as the angle of the difference rotation of two quaternions. Finally, the *scaling* misalignment is the scaling required to match the brain volume to the mean brain volume at GW 30.

The Pearson correlation coefficient is used to determine whether there is a significant correlation between the $\mathrm{CD}(\mathbf{M}, \widehat{\mathbf{M}})$, $\mathrm{DSC}(\mathbf{M}, \widehat{\mathbf{M}})$, $\mathrm{HD}(\mathbf{M}, \widehat{\mathbf{M}})$, and $\mathrm{SC}(\widehat{\mathbf{M}})$ results, and the three misalignments. As the results in Fig. 4.8 show, the value of the correlation coefficient for the *translation* misalignment is below 0.12 for every measure, meaning that the performance of fBENv2 is nearly invariant to

the location of the fetal brain within the 3D scan. A similar situation is observed for the *rotation* misalignment, where the correlation coefficient is nearly zero for most metrics, confirming that the performance of fBENv2 is also nearly invariant to the orientation of the brain inside the 3D scan, in spite of the asymmetric structural information of the brain (see Sec. 1.1). In contrast, there is a much stronger correlation between the *scaling* misalignment and performance. As the *scaling* misalignment is directly related to the gestational age, this result is not surprising, since the results shown in Fig. 4.5 show that the performance is not perfectly even. However, while there is a correlation between the *scaling* misalignment and the performance of fBENv2, the results of that figure also show that this correlation results in a minimal variation in performance, making the performance of fBENv2 nearly invariant all types of misalignment.

4.3.4 Performance consistency

As a final assessment of the performance of fBENv2, I analyse the effect that the orientation of the input scan has on the prediction. In other words, how much do the predictions of the same scan differ, depending on which orientation the scan was passed onto the network. To assess this, each scan was augmented to its 24 possible orientations, and fBENV2 was used to predict the brain mask for each one, which were subsequently reoriented to match the original orientation of the input scan. For each one of the 24 predictions, the CD, DSC, and HD were measured against the other 23, as well as their SC, and the corresponding mean, median, and standard deviation values were calculated. The results of Tab. 4.7 show the average results for the entire hold-out testing split of the $\mathfrak{D}_B$ dataset, which show significantly better results than those obtained against the reference masks, shown in Tab.4.6. The same can be seen when separated by GW, as shown in Fig. 4.9. For every GW, the mean consistency of fBENv2 is significantly higher than its average performance.

Since the consistency of fBENv2 is significantly higher than its state-of-the-art performance, its performance can be considered to be nearly invariant to the orientation of the input scan $\mathbf{S}$.

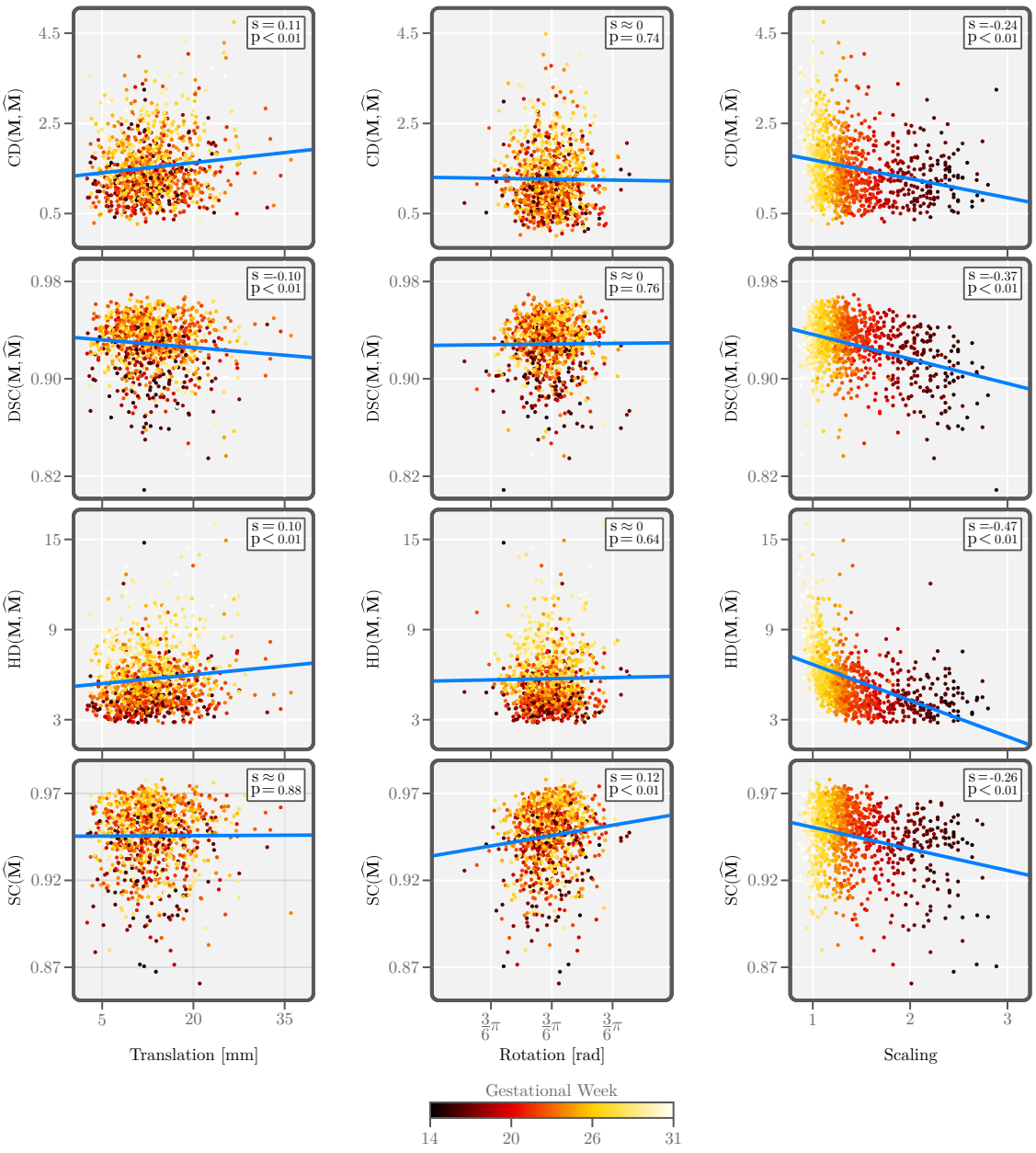

Figure 4.8: Performance of fBENv2 for Centroid Distance CD, Dice Similarity Coefficient DSC, Hausdorff Distance HD, Symmetry Coefficient SC, against brain misalignment. The Pearson correlation coefficient s, and the corresponding p-value are provided for each pair, as well as a linear fit for visualisation purposes. Translation is the Euclidean distance between the centre of the brain and the centre of the volume, Rotation is the cosine distance between the orientation of the head and the canonical space, and Scaling is the factor needed to scale the brain volume to match the average brain volume at 25 GWs.

Table 4.7: Performance consistency of fBENv2 against input orientation. The brain masks $\widehat{\mathbf{M}}$ for the 24 possible orientations of each scan $\mathbf{S}$ were generated by fBENv2, and subsequently reoriented to match the orientation of the original scan. The Symmetry Coefficient SC was measured for each mask, while the Centroid Distance CD, Dice Similarity Coefficient DSC, and Hausdorff Distance HD were measured for each mask against the other 23. Finally, their mean, median, and standard deviation was calculated. The table shows the average results for the entire hold-out testing split of dataset $\mathfrak{D}_B$.

	CD [mm] ↓	DSC ↑	HD [mm] ↓	SC ↑
Mean	0.70 ± 0.11	0.968 ± 0.006	3.0 ± 1.1	0.948 ± 0.016
Median	0.68 ± 0.10	0.968 ± 0.006	2.8 ± 0.9	0.948 ± 0.017
Std. dev.	0.30 ± 0.06	0.007 ± 0.002	0.9 ± 0.9	0.009 ± 0.003

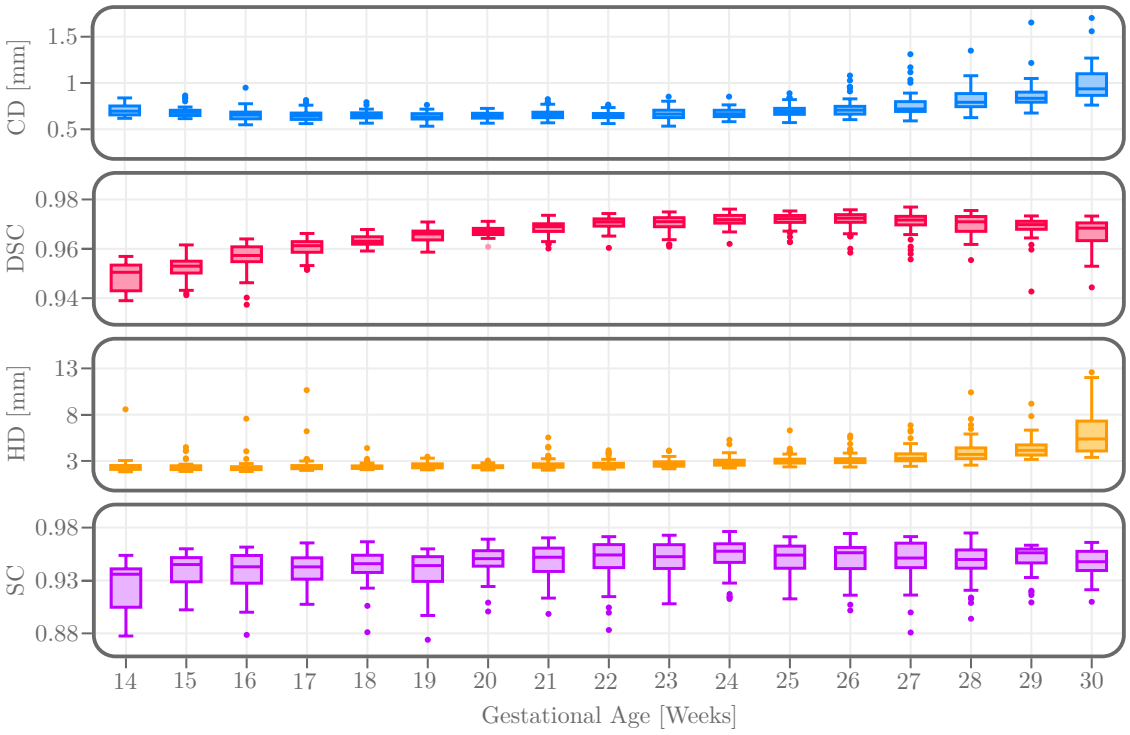

Figure 4.9: Performance consistency of fBENv2 against input orientation, separated by GW. The brain masks $\widehat{\mathbf{M}}$ for the 24 possible orientations of each scan $\mathbf{S}$ were generated by fBENv2, and subsequently reoriented to match the orientation of the original scan. The Symmetry Coefficient SC was measured for each mask, while the Centroid Distance CD, Dice Similarity Coefficient DSC, and Hausdorff Distance HD were measured for each mask against the other 23. The values on each plot represents their mean performance.

4.4 Discussion and Conclusions

In this chapter I have proposed the first module of the Fully-Automated DL Pipeline for 3D Fetal Brain Ultrasound: the fetal Brain Extraction Network (fBEN). I have shown the process of developing the fBEN network, as well as its iterative evolution, concluding with the development of fBENv2.

I have thoroughly analysed the performance of fBEN to fully grasp its strengths and weaknesses, as well as to determine whether the initial goals were achieved.

The performance of fBENv2 is significantly higher than any current alternative method, setting the new state-of-the-art performance for the task of automated fetal brain extraction from 3D US scans, and achieving a similar performance to the equivalent state-of-the-art solutions for fetal MRI. The network also manages to perform consistently for the entire gestational range of 14.0 to 30.9 weeks, showing only minor variations throughout that range. Similarly, the regional performance of fBENv2 is consistent throughout the brain, adapting its prediction based on the structural information of the input scan. As such, its predictions are more accurate than the generated labels.

I also showed that the performance of fBENv2 is not only robust regardless of the quality of the input scan, unlike its previous iteration, but it is also virtually invariant to the location, orientation, and size of the brain inside the scan.

Finally, I demonstrated that the predictions of fBENv2 are consistent, regardless of the orientation in which the scan is passed. This reinforces its robustness against the orientation of the brain inside the scan, but also ensures the same performance regardless of the orientation in which a particular user stores their scans.

All of this was achieved with rough, noisy labelling, as well as manual alignments performed by non-experts. While it would be interesting to compare this approach against more accurate labels generated by experts, the generation of such data is currently unfeasible.

5 Automated Fetal Brain Alignment

Contents

5.1 Introduction

After successfully developing a state-of-the-art solution for the task of fetal brain extraction from 3D US scans in Ch. 4, I continue the development of the proposed pipeline by focusing on the second module presented in this chapter, which addresses

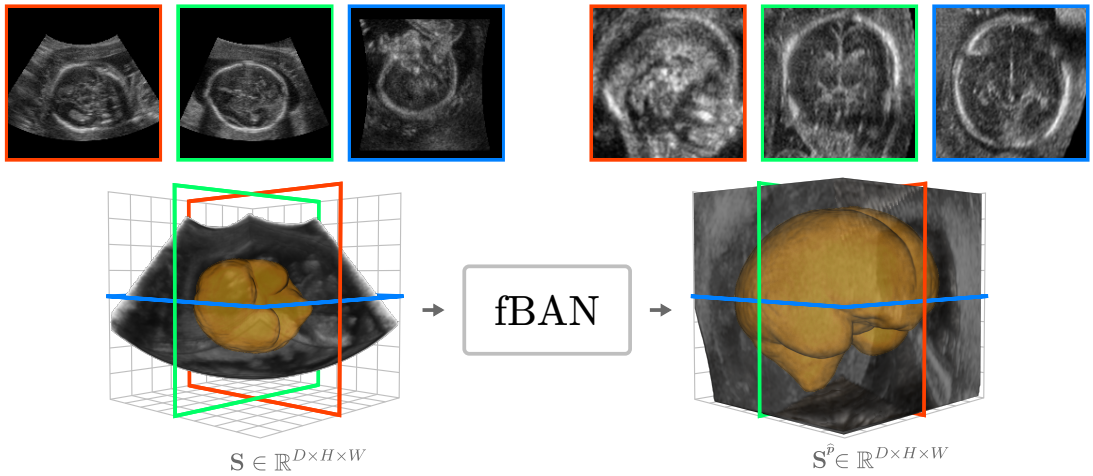

Figure 5.1: Graphical abstract of the fetal Brain Alignment Network (fBAN).

the challenging task of aligning the 3D brain to a canonical reference space.

While clinical guidelines for acquisition of 3D neurosonograms have been proposed with the aim of minimising the positional variability of the brain between scans, unpredictable location of the fetal head inside the womb, combined with the inherent limitations of US, makes it difficult for them to be effective. Due to acoustic shadows and artefacts, the optimal positioning of the probe in order to capture as much structural information as possible is often at odds with the consistent alignment of the brain relative to the scan. Nevertheless, an accurate alignment of the 3D fetal brain is crucial for inter- and intra-subject comparisons, as improperly aligned scans can result in the misrepresentation of structural similarities and differences. However, the manual alignment of 3D US scans of the fetal brain is a tedious, time-consuming tasks that require a high level of expertise. Therefore, a robust and accurate automated solution for this task is paramount for the widespread adoption of this imaging modality.

Just like for the extraction task of the previous chapter, the importance of brain alignment has resulted in the development of a multitude of robust, automated solutions for more traditional 3D neuroimaging modalities such as MRI. In contrast, only a handful of works have been proposed for fetal 3D US, most of which rely on specific anatomical landmarks to be visible, which is not always feasible. Additionally, no solutions have been proposed for gestational ages earlier than 17 GWs.

In this chapter I propose the second module of the Fully-Automated DL Pipeline for 3D Fetal Brain Ultrasound: the fetal Brain Alignment Network (fBAN). This is an end-to-end, cascade regression-based CNN that predicted the alignment parameters required to rigidly align minimally pre-processed, standard clinical 3D US scans, to a canonical reference space, as shown in Fig. 5.1.

Through exhaustive analysis, I show that fBAN achieves state-of-the-art performance, comparable with the current state-of-the-art solutions for fetal MRI. I demonstrate that fBAN performs consistently for the entire gestational range of 14.0 to 30.9 GW. Finally, I demonstrate that the performance of the network is invariant to the location and orientation of the fetal brain in the 3D scan.

The contributions in this chapter are:

- Development of fBAN, an end-to-end, cascade regression-based CNN that predicted the alignment parameters required to rigidly align minimally pre-processed, standard clinical 3D US scans, to a canonical space.
- Demonstration that the extraction network (fBEN) implicitly learns most of the alignment information through its encoder section.
- Propose Spatial Landmarks (SL) as a of a novel representation of the alignment task.
- Development of a novel SL loss function for training fBAN.
- Development of a novel Transfer Learning approach to ensure that the network is making its prediction based on the understanding of the structural information of the brain, hindering it from "cheating".

5.2 Methods

In this section I describe the iterative process of developing fBAN, from its initial developmental stages up to its final refinement. I also cover the data used, the measures chosen for a multifaceted assessment of performance, and the implementation details of the method.

5.2.1 Data

Just like with fBEN, discussed in Ch. 4, most of the development of fBAN (Sec. 5.2.4, 5.2.6, and 5.2.7) was performed with the $\mathfrak{D}_A$ dataset, which consists of a set of $n_A = 1185$ high-quality 3D US scans of the fetal brain $\mathbb{S}_A$, and their corresponding masks $\mathbb{M}_A$ and alignment parameters $\mathbb{P}_A$. These parameters are comprised of three translation parameters $\boldsymbol{p}_T = (p_x, p_y, p_z)$, normalised to $(-0.5, 0.5)$, three rotation parameters (Euler angles) $\boldsymbol{p}_R = (p_\alpha, p_\beta, p_\gamma)$, normalised to $(-\pi, \pi)$, and one scaling parameter $\boldsymbol{p}_S = (p_s)$ in the range $(0.92, 2.72)$. Dataset $\mathfrak{D}_A$ was randomly split (proportionally to the gestational ages) into a 356 hold-out set for testing, and an 829 set for training and validation.

For the final refinement of fBAN, discussed in Sec. 5.2.8, dataset $\mathfrak{D}_B$ is used instead, which consists of a set of $n_B = 4290$ scans $\mathbb{S}_B$, a set of corresponding masks $\mathbb{M}_B$, and a set of alignment parameters $\mathbb{P}_B$. The parameters consist of three translation parameters $\boldsymbol{p}_T = (p_x, p_y, p_z)$, normalised to $(-0.5, 0.5)$, four rotation parameters (normalised quaternions) $\boldsymbol{p}_R = (p_{q_w}, p_{q_x}, p_{q_y}, p_{q_z})$, and one scaling parameter $\boldsymbol{p}_S = (p_s)$ in the range $(0.86, 3.12)$. In contrast to $\mathfrak{D}_A$, the scans of $\mathfrak{D}_B$ have been selected to represent the variability of the INTERGROWTH-21$^{\text{st}}$dataset, containing both high and low quality scans. $\mathfrak{D}_B$ was split into a 1073 hold-out set for testing, and a 3217 set for training and validation. However, unlike $\mathfrak{D}_A$, this split was performed using an iterative stratification approach that evenly distributed the data based on the gestational age of the scans, as well as the structural similarity of the brain.

Both datasets span the gestational age range of 14.1 to 30.9 GW (99 to 216 gestational days). The pre-processing of the scans was limited to resampling to an isotropic voxel size of 0.6 mm/vxl, subsequently centre-cropping to a size of (160,160,160), and finally normalising their features to within 0 and 1.

Note that the manual alignment of $\mathfrak{D}_A$, and the manual alignment of $\mathfrak{D}_B$ were performed by different users, and are in different canonical spaces. Therefore, the parameters of $\mathbb{P}_A$ and $\mathbb{P}_B$ and are not directly comparable. Consequentially,

while the scans of $\mathbb{S}_A$ are contained in $\mathbb{S}_B$, their corresponding masks in $\mathbb{M}_B$ are not identical to those in $\mathbb{M}_A$.

A more detailed description of the data used in this thesis can be found in Chapter 3.

5.2.2 Implementation details

Similarly to the development of fBEN described in Sec.4, for most of the development of fBAN the networks were implemented in Python using the Tensorflow [165] and Keras [166] libraries. The networks were developed on an Intel Xeon E-2146G CPU (3.50GHz, 6 cores) and an Nvidia GTX 1080 Ti. For the final refinement, the networks were re-implemented in Python using the Pytorch [167] library, and an Nvidia A10. Unless specified, all fBAN networks were trained with a batch size $bs = 4$, the Adam optimiser, and an empirically selected learning rate $lr = 0.001$. All fBAN networks were trained using a 3-fold cross-validation, and tested against a hold-out dataset.

5.2.3 Evaluation measures

In order to perform a thorough evaluation of the predictions of fBAN, I have chosen to rely on several different measures that reflect different aspects of performance, allowing for the quick understanding of the strengths and weaknesses of the network.

First, it is important to assess how close the alignment parameters $\widehat{\boldsymbol{p}}$ predicted by fBEN are to the reference parameters $\boldsymbol{p}$ generated through the manual alignment described in Sec. 3.2. For this, I will rely on the Mean Squared Error (MSE), which is calculated as shown in Eq. (5.1), where $n_{\boldsymbol{p}}$ is the number of parameters.

$$\mathrm{MSE}(\boldsymbol{p}, \widehat{\boldsymbol{p}}) = \frac{1}{n_{\boldsymbol{p}}} \sum_{i=1}^{n_{\boldsymbol{p}}} (p_i - \widehat{p}_i)^2 \tag{5.1}$$

As mentioned in Sec. 5.2.1, these parameters represent different transform types, and their values are in different ranges. Therefore, for a more detailed assessment of the predicted parameters $\widehat{\boldsymbol{p}}$, I also calculate the MSE of the translation parameters

$\text{MSE}_T = \text{MSE}(\boldsymbol{p}_T, \widehat{\boldsymbol{p}}_T)$, the rotation parameters $\text{MSE}_R = \text{MSE}(\boldsymbol{p}_R, \widehat{\boldsymbol{p}}_R)$, and the scaling parameter $\text{MSE}_S = \text{MSE}(\boldsymbol{p}_S, \widehat{\boldsymbol{p}}_S)$, separately.

However, while the MSE is very useful to assess how far the predicted parameters are from their expected values, it does not adequately represent the alignment performance of the network. The effect each parameter has on the transform of the scan is not only dependent on their value, but also the type of transform, the order in which the transforms are applied, and even the location of each voxel in the scan, among other factors. These complex interactions make it difficult to analyse the performance of the network based on the predicted parameters alone.

Instead, the binary brain masks $\mathbf{M}$ can be used to directly assess the effect of the transform on the voxels of the network. These masks can be aligned with the reference parameters $\boldsymbol{p}$ and predicted parameters $\widehat{\boldsymbol{p}}$, resulting in the reference aligned mask $\mathbf{M}^{\boldsymbol{p}}$ and the predicted aligned mask $\mathbf{M}^{\widehat{\boldsymbol{p}}}$, respectively. Some of the same measures used to develop fBEN (see Sec. 4.2.3) can then be used to assess the similarity between these two aligned masks to assess the alignment performance of fBAN, since the relationships between parameters are implicitly contained in the location of the aligned voxels. The Dice Similarly Coefficient $\text{DSC}(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}})$ will indicate the overall performance for each prediction, while the Hausdorff Distance $\text{HD}(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}})$ will help assess the regional performance.

The combined use of $\text{MSE}(\boldsymbol{p}, \widehat{\boldsymbol{p}})$, $\text{DSC}(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}})$ and $\text{HD}(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}})$ results in a thorough, detailed assessment of the performance of fBAN, that allows for the efficient analysis and comparison the results throughout development, as well as against alternative methods.

To assess the statistical significance when comparing these measures, I first determined the normality with a D'Agostino and Pearson's test [169], followed by a paired Student's t-test [170] or a Wilcoxon signed-rank test [171], for normal and non-normal samples, respectively. A significance threshold of $p < 0.05$ was used for both tests.

5.2.4 Initial development

After successfully developing a state-of-the-art solution for the task of automated fetal brain extraction from 3D US scans in Ch. 4, I continued the pipeline development focusing on the automated alignment of the fetal brain. However, the results shown in Sec. 4.3 indicated that fBEN was able to accurately predict the brain mask, regardless of the misalignment of the brain relative to the scan, as shown in Fig. 4.8. Therefore, not only was fBEN able to handle the high intrinsic variability of the 3D US data, but it must also have some degree of understanding of the location, orientation, and scaling of the brain. Given the encoder-decoder architecture of fBEN, I hypothesised that the majority of this information must be captured in the encoder. Therefore, I decided to start the development of fBAN by adapting the encoder portion of fBEN into a regression network that predicts the alignment parameters needed to construct a similarity transform that aligns the input scan to a canonical space.

However, there are multiple ways to generate the same alignment, depending on the order in which the translation, rotation, and scaling transforms are applied, or where the centre of origin is defined. The values for the translation, rotation, and scaling parameters will be different for each approach, as might be the similarity transform itself, but the transformed volume will be the same. However, the task of predicting the necessary parameters, and therefore the difficulty of developing the corresponding network, is different depending on the approach used. To make the task as simple as possible for the network, I decided to define the geometric centre of the 3D volume as the origin of the coordinate system around which the translation transform $\boldsymbol{T}_T$, the rotation transform $\boldsymbol{T}_R$, and the scaling transform $\boldsymbol{T}_S$ are applied, in that order. For a voxel at position $\boldsymbol{x}$, its transformed position $\boldsymbol{y}$ is defined as shown in Eq. (5.2).

$$\boldsymbol{T} \cdot \boldsymbol{x} = \boldsymbol{y} \tag{5.2}$$

Therefore, the similarity transformed is defined as shown in Eq. (5.3).

$$\boldsymbol{T} = \boldsymbol{T}_S \cdot \boldsymbol{T}_R \cdot \boldsymbol{T}_T \tag{5.3}$$

Since the centre of mass of the aligned brain is at the origin of the canonical space, applying $\boldsymbol{T}_T$ first allows for the transforms $\boldsymbol{T}_R$ and $\boldsymbol{T}_S$ to be performed around the centre of mass of the unaligned brain. In addition to being easier to visualise, this approach significantly reduces the complexity of the task for the network, since the resulting translation, rotation, and scaling parameters are independent of each other. In other words, the translation parameters (p_x, p_y, p_z) are only dependent on the location of the brain in the scan and not its orientation or size of the brain, while the rotation parameters $(p_\alpha, p_\beta, p_\gamma)$ are only dependent on the orientation of the brain, and the scaling parameter p_s is only dependent on its size.

After defining the task, I focused on adapting the encoder architecture of fBEN into a regression network. As shown in Fig. 5.2, the initial architecture of fBAN consisted of 5 Convolutional Blocks, 5 MaxPooling layers, and 2 Dense layers, with each Convolutional Block comprised of two subsequent sets of a Convolutional layer, a Batch Normalisation layer, and a ReLU activation layer. I started with an initial kernel size $ks = (3, 3, 3)$ for the Convolutional Blocks, and hidden dimensions $hd = (16, 32, 64, 128, 256)$, with the final Dense layer returning the 7 predicted parameters $\boldsymbol{p} = (p_x, p_y, p_z, p_\alpha, p_\beta, p_\gamma, p_s)$.

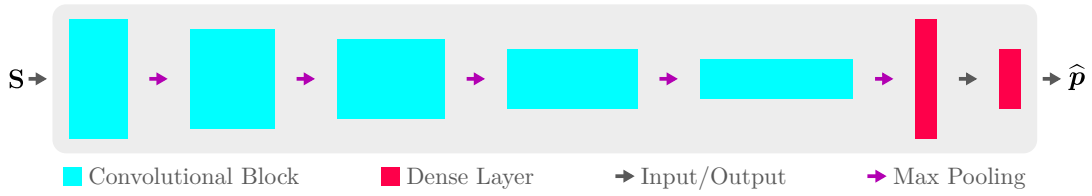

Figure 5.2: Schematics of the initial architecture of fBAN, adapted from fBEN.

At this stage, I had defined the task and architecture of the network, but still needed to find an appropriate loss function to train it. However, the non-linear relationships between parameters, transforms, and the aligned scan, make it very challenging to find a suitable loss function that is robust and yields highly accurate predictions, particularly when the network has not been optimised for the task. Therefore, I continued the development focusing only on finding a robust loss function that would allow me to test the hypothesis, leaving the search for a more refined function for later. For the initial loss function, I decided to simply

focus on the accuracy of the predicted parameters $\widehat{\boldsymbol{p}}$, for which I tested three well-known functions: the Mean Squared Error (Eq. 5.4), the Mean Absolute Error (Eq. 5.5), and the Logcosh (Eq. 5.6).

$$\mathfrak{L}_{\text{MSE}} = \frac{1}{n_{\boldsymbol{p}}} \sum_{i=1}^{n_{\boldsymbol{p}}} (p_i - \widehat{p}_i)^2 \tag{5.4}$$

$$\mathfrak{L}_{\text{MAE}} = \frac{1}{n_{\boldsymbol{p}}} \sum_{i=1}^{n_{\boldsymbol{p}}} |p_i - \widehat{p}_i| \tag{5.5}$$

$$\mathfrak{L}_{\text{MSE}} = \sum_{i=1}^{n_{\boldsymbol{p}}} log(cosh(p_i - \widehat{p}_i)) \tag{5.6}$$

As an additional constrain, I reduced the spatial resolution of the $\mathfrak{D}_A$ dataset from 0.6 mm/vxl to 1.2 mm/vxl, allowing for faster development thanks to a greatly reduced computational cost. This makes the initial optimisation of the network significantly faster, after which the full resolution scans can be reintroduced for the final refinement. Nevertheless, for consistency, all low-resolution predictions have been assessed using the full-resolution reference masks. Since the translation parameters $\boldsymbol{p}_T$ are normalised, the same alignment parameters work at any spatial resolution, so no further changes are needed.

As Tab. 5.1 shows, $\mathfrak{L}_{\text{MAE}}$ resulted in the best performance for every measure used, outperforming the alternatives by 8.3% for MSE($\boldsymbol{p}, \widehat{\boldsymbol{p}}$), 3% for DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$), and 7.8% for HD($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$), all of which are statistically significant. The same trend is seen when separately analysing the translation, rotation, and scaling parameters. Therefore, $\mathfrak{L}_{\text{MAE}}$ was chosen for the next step of development. It is also notable that this unoptimised network achieved a DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) of 0.82, suggesting that the hypothesis was indeed correct, and that the encoder architecture of fBEN is adequate for predicting the alignment parameters.

After choosing the initial loss function, I focused on optimising the network architecture. I performed the same iterative approach as the one described in Sec. 4.2.4, optimising kernel size, hidden dimensions, and number of Pooling layers. The optimised settings consisted of a kernel size of 3, a total of 5 Pooling layers, and $(32, 64, 128, 256, 512, 1024)$ hidden dimensions.

Table 5.1: Performance comparison of using different loss functions for training fBAN: Mean Squared Error $\mathfrak{L}_{\text{MSE}}$, Mean Absolute Error $\mathfrak{L}_{\text{MAE}}$, and the Logarithm of the Hyperbolic Cosine $\mathfrak{L}_{\text{LCH}}$. The best performance for each measure (Mean Squared Error MSE, Dice Similarity Coefficient DSC, Hausdorff Distance HD) is highlight in bold. The arrows indicate whether a higher (up) or lower (down) value is preferred.

	Loss fn.		
Measure	$\mathfrak{L}_{\text{MSE}}$	$\mathfrak{L}_{\text{MAE}}$	$\mathfrak{L}_{\text{LCH}}$
$\text{MSE}(\boldsymbol{p}, \widehat{\boldsymbol{p}}) \downarrow$	0.12 ± 0.29	$\mathbf{0.09 \pm 0.32}$	0.10 ± 0.29
$\text{MSE}(\boldsymbol{p}_T, \widehat{\boldsymbol{p}}_T) \downarrow$	0.0041 ± 0.0033	$\mathbf{0.0022 \pm 0.0029}$	0.0028 ± 0.0037
$\text{MSE}(\boldsymbol{p}_R, \widehat{\boldsymbol{p}}_R) \downarrow$	0.24 ± 0.66	$\mathbf{0.19 \pm 0.74}$	0.21 ± 0.67
$\text{MSE}(\boldsymbol{p}_S, \widehat{\boldsymbol{p}}_S) \downarrow$	0.047 ± 0.073	$\mathbf{0.038 \pm 0.071}$	0.048 ± 0.085
$\text{DSC}(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}) \uparrow$	0.79 ± 0.08	$\mathbf{0.82 \pm 0.08}$	0.80 ± 0.09
$\text{HD}(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}})$ [voxels] $\downarrow$	25.6 ± 7.6	$\mathbf{22.4 \pm 7.6}$	24.3 ± 7.1

As Tab. 5.2 shows, the optimised network obtained a significant improvement in performance for every measure tested. There was a 39.7% improvement in the $\text{MSE}(\boldsymbol{p}, \widehat{\boldsymbol{p}})$ performance, with a more modest improvement of 4.9% and 16.5% for $\text{DSC}(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}})$ and $\text{HD}(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}})$. However, improvement in the accuracy of the predictions was entirely focused on the rotation and scaling parameters, with no improvement in the accuracy of the translation parameters which were already accurately predicted. This is likely due to higher difficulty of predicting the rotation and scaling parameters. Nevertheless, with a $\text{DSC}(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}})$ of 0.86 and a $\text{HD}(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}})$ of 18.4 voxels, the network was clearly managing to align the brain to a reasonable degree of accuracy. While this performance is still short far from the goal state-of-the-art performance, it is good enough to confirm the initial hypothesis regarding the encoder architecture of the fBEN network, justifying the use of its architecture for fBAN.

At this stage, it became apparent that relying entirely on the predicted parameters for the loss function was limiting performance. While $\mathfrak{L}_{\text{MAE}}$ is very useful to help the network get close to its goal, it does not take into account the complex relationships between the alignment parameters and the actual alignment of the scans. Thankfully, there already was a function available that reflected these relationships: $\text{DSC}(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}})$. Since the reference masks $\mathbf{M}$ are aligned before

Table 5.2: Performance comparison of fBAN trained with $\mathfrak{L}_{\text{MAE}}$, before and after iterative optimisation. The best performance for each measure (Mean Squared Error MSE, Dice Similarity Coefficient DSC, Hausdorff Distance HD) is highlight in bold. The arrows indicate whether a higher (up) or lower (down) value is preferred.

	$\mathfrak{L}_{\text{MAE}}$	
Measure	Unoptimised	Optimised
MSE($\boldsymbol{p}, \widehat{\boldsymbol{p}}$) ↓	0.09 ± 0.32	**0.06 ± 0.34**
MSE($\boldsymbol{p}_T, \widehat{\boldsymbol{p}}_T$) ↓	0.0022 ± 0.0029	**0.0020 ± 0.0018**
MSE($\boldsymbol{p}_R, \widehat{\boldsymbol{p}}_R$) ↓	0.19 ± 0.74	**0.14 ± 0.79**
MSE($\boldsymbol{p}_S, \widehat{\boldsymbol{p}}_S$) ↓	0.038 ± 0.071	**0.017 ± 0.031**
DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) ↑	0.82 ± 0.07	**0.86 ± 0.06**
HD($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) [voxels] ↓	22.4 ± 7.6	**18.7 ± 7.4**

calculating the DSC, the impact of each parameter is incorporated implicitly into the aligned voxels, which should translate to better predictions. The corresponding loss function $\mathfrak{L}_{\text{DSC}}$ is calculated as shown in Eq. 5.7, with $m = 160^3$ being the number of voxels in the scan, and i-th voxels $\mathbf{M}_i^{\boldsymbol{p}}$ and $\mathbf{M}_i^{\widehat{\boldsymbol{p}}}$ having the binary value $\mathbf{M}^{\boldsymbol{p}}(x_i, y_i, z_i)$ and $\mathbf{M}^{\widehat{\boldsymbol{p}}}(x_i, y_i, z_i)$, respectively. Just like in Sec. 5.2.4, the smoothing factor $s = 1$ is added to account for the possibility of both $\mathbf{M}^{\boldsymbol{p}}$ and $\mathbf{M}^{\widehat{\boldsymbol{p}}}$ having no positive voxels, which would result in dividing by zero.

$$\mathfrak{L}_{\text{DSC}} = 1 - \text{DSC}(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}) = 1 - \frac{2 \cdot \sum_{i=1}^{m} (\mathbf{M}_i^{\boldsymbol{p}} \cdot \mathbf{M}_i^{\widehat{\boldsymbol{p}}}) + s}{\sum_{i=1}^{m} \mathbf{M}_i^{\boldsymbol{p}} + \sum_{i=1}^{m} \mathbf{M}_i^{\widehat{\boldsymbol{p}}} + s} \tag{5.7}$$

However, relying on $\mathfrak{L}_{\text{DSC}}$ alone for training is not possible, since it requires some degree of overlap to help guide the alignment. If there is no overlap between $\mathbf{M}^{\boldsymbol{p}}$ and $\mathbf{M}^{\widehat{\boldsymbol{p}}}$, the DSC will always be zero, and the network will fail to converge. Therefore, I combined the $\mathfrak{L}_{\text{DSC}}$ and $\mathfrak{L}_{\text{MAE}}$ into a new loss function $\mathfrak{L}_{\text{MAE+DSC}}$, as defined in Eq. 5.8, where w_{DSC} is a weighting factor. By combining the properties of both, the drawbacks of each are compensated by the other. During the early stages of training, it is guided entirely by $\mathfrak{L}_{\text{MAE}}$, since there is no overlap between $\mathbf{M}^{\boldsymbol{p}}$ and $\mathbf{M}^{\widehat{\boldsymbol{p}}}$. Once the network predictions are close enough for the overlap to appear, the $\mathfrak{L}_{\text{DSC}}$ term starts contributing to the optimisation. However, since $\mathfrak{L}_{\text{DSC}} \in [0, 1]$, the $\mathfrak{L}_{\text{MAE}}$ term remains the main guide for the training until $\mathfrak{L}_{\text{MAE}} < 1$. Eventually, the

$\mathfrak{L}_{\text{MAE}}$ term becomes small enough and the $\mathfrak{L}_{\text{DSC}}$ term becomes the main contributor to the loss. This balance can be manipulated by changing the weighting factor w_{DSC}, as it will depend on the magnitude of the parameter values. In this case, $w_{DSC} = 1$ resulted in a good balance.

$$\mathfrak{L}_{\text{MAE+DSC}} = \mathfrak{L}_{\text{MAE}} + w_{DSC} \cdot \mathfrak{L}_{\text{DSC}} \tag{5.8}$$

The results shown in Tab. 5.3 confirm the limitations of relying entirely on the parameters to train the network, and the benefits of relying on the implicit information of the transformed voxels. While the addition of $\mathfrak{L}_{\text{DSC}}$ resulted in no statistically significant changes for MSE($\boldsymbol{p}, \widehat{\boldsymbol{p}}$) (or its subdivisions), it did result in statistically significant performance improvements of 3.4% for the DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$), and a 15.5% for the HD($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$). Additionally, with a DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) of 0.89 and a HD($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) of 15.8 voxels, fBAN was approaching state-of-the-art performance.

Table 5.3: Performance comparison of fBAN trained with loss functions $\mathfrak{L}_{\text{MAE}}$ and $\mathfrak{L}_{\text{MAE+DSC}}$. The best performance for each measure (Mean Squared Error MSE, Dice Similarity Coefficient DSC, Hausdorff Distance HD) is highlight in bold. The arrows indicate whether a higher (up) or lower (down) value is preferred.

	Loss fn.	
Measure	$\mathfrak{L}_{\text{MAE}}$	$\mathfrak{L}_{\text{MAE+DSC}}$
MSE($\boldsymbol{p}, \widehat{\boldsymbol{p}}$) ↓	**0.06 ± 0.34**	0.07 ± 0.35
MSE($\boldsymbol{p}_T, \widehat{\boldsymbol{p}}_T$) ↓	0.0020 ± 0.0018	**0.0012 ± 0.0014**
MSE($\boldsymbol{p}_R, \widehat{\boldsymbol{p}}_R$) ↓	**0.14 ± 0.79**	0.15 ± 0.81
MSE($\boldsymbol{p}_S, \widehat{\boldsymbol{p}}_S$) ↓	**0.017 ± 0.031**	0.005 ± 0.018
DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) ↑	0.86 ± 0.06	**0.89 ± 0.05**
HD($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) [voxels] ↓	18.7 ± 7.4	**15.8 ± 6.6**

However, in spite of the benefits, the reliance on DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) to implicitly incorporate the parametric relationships of the alignment also results in several disadvantages.

As mentioned before, the use of $\mathfrak{L}_{\text{DSC}}$ requires the predicted parameters $\widehat{\boldsymbol{p}}$ to be accurate enough to result in an overlap between $\mathbf{M}^{\boldsymbol{p}}$ and $\mathbf{M}^{\widehat{\boldsymbol{p}}}$, since its value will otherwise be zero. However, if the predicted parameters do not manage to adequately correct the orientation of the brain by the time the $\mathfrak{L}_{\text{DSC}}$ becomes

a significant component of $\mathfrak{L}_{\text{MAE+DSC}}$, its addition to the loss function can be detrimental to the training of the network. For example, if the predicted parameters $\widehat{\boldsymbol{p}}$ result in a brain mask $\mathbf{M}^{\widehat{\boldsymbol{p}}}$ that is overlapping $\mathbf{M}^{\boldsymbol{p}}$, but rotated by more than $\frac{\pi}{2}$ around one of the axes, the $\mathfrak{L}_{\text{DSC}}$ component of $\mathfrak{L}_{\text{MAE+DSC}}$ will hinder the network from correcting this rotation, since the transitional stages would result in a higher cost. Instead, the network will try to maximise the overlap of the rotated mask $\mathbf{M}^{\boldsymbol{p}}$ and the reference mask $\mathbf{M}^{\boldsymbol{p}}$, entrenching itself in a local minimum. Therefore, the use of $\text{DSC}(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}})$ as part of the loss function $\mathfrak{L}_{\text{MAE+DSC}}$ is only beneficial once the network has achieved a certain level of alignment accuracy.

The use of $\mathfrak{L}_{\text{MAE+DSC}}$ is also quite computationally expensive, since it requires aligning every reference mask $\mathbf{M}$ to $\mathbf{M}^{\boldsymbol{p}}$ and $\mathbf{M}^{\widehat{\boldsymbol{p}}}$, as well as comparing every voxel between them, for every epoch. As a result, training with $\mathfrak{L}_{\text{MAE+DSC}}$ took 4 times longer than training with $\mathfrak{L}_{\text{MAE}}$.

In addition to being computationally expensive, relying on $\text{DSC}(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}})$ to implicitly incorporate the parametric relationships of the alignment is also quite inefficient. Due to the nature of the binary brain masks, most of the voxels in $\mathbf{M}$ are actually empty, and do not offer any information, yet they still need to be transformed. Furthermore, while all of the positive voxels are needed to calculate the $\text{DSC}(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}})$, most of them provide redundant information about the parameters and the transform. Finally, while the use of $\mathfrak{L}_{\text{MAE+DSC}}$ resulted in a significant improvement over $\mathfrak{L}_{\text{MAE}}$, the performance was still not sufficient to fulfil the state-of-the-art performance goal of the pipeline. Therefore, I continued the development by focusing on improving the loss function used to train the network.

5.2.5 Spatial Landmarks Loss

After the initial development of fBAN, I had confirmed initial hypothesis that the architecture of the encoder section of fBEN was adequate for the task of fetal brain alignment, as well as robust against the high intrinsic data variability of the data. However, the performance achieved by fBAN was still underwhelming. Given the robustness against misalignment that fBEN had exhibited (see Sec. 4.3.3), I

was inclined to believe that this was a consequence of the training methodology, rather than an architectural limitations. Therefore, I continued the development by focusing on developing a more efficient, robust, and refined approach to train fBAN.

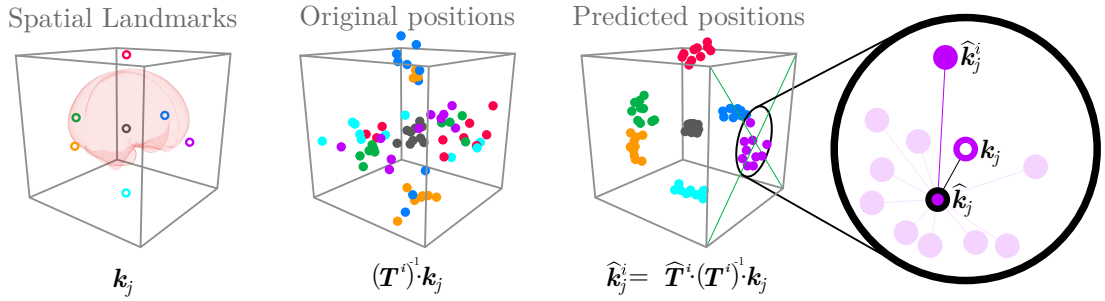

Figure 5.3: Schematics of the Spatial Landmark (SL) representation of the alignment task. The aligned volume is represented by seven SLs $\boldsymbol{k}_j$. The SLs are transformed to their original, unaligned positions using the inverse of the ground-truth similarity transform $\boldsymbol{T}^i$, and subsequently transformed to their predicted positions $\widehat{\boldsymbol{k}}_j^i$ using the predicted similarity transform $\widehat{\boldsymbol{T}}^i$.

First, I focused on developing a better representation of the alignment task. For this, I defined the Spatial Landmarks (SL): a series of points $\boldsymbol{k}_j$ in 3D space that represent the spatial characteristics of the canonical space. As shown in Fig. 5.3, 1 SL represents the geometric centre of the scan aligned to the canonical space, and 6 SLs represent the centre of each of its faces, for a total of $n_{SL} = 7$. For a scan of size (D_x, D_y, D_z), these are therefore defined in homogeneous coordinates as shown in Eq. (5.9)-(5.15)

$$\boldsymbol{k}_1 = (0, 0, 0, 1)^T \tag{5.9}$$

$$\boldsymbol{k}_2 = (-\frac{D_x}{2}, 0, 0, 1)^T \tag{5.10}$$

$$\boldsymbol{k}_3 = (\frac{D_x}{2}, 0, 0, 1)^T \tag{5.11}$$

$$\boldsymbol{k}_4 = (0, -\frac{D_y}{2}, 0, 1)^T \tag{5.12}$$

$$\boldsymbol{k}_5 = (0, \frac{D_y}{2}, 0, 1)^T \tag{5.13}$$

$$\boldsymbol{k}_6 = (0, 0, -\frac{D_z}{2}, 1)^T \tag{5.14}$$

$$\boldsymbol{k}_7 = (0, 0, -\frac{D_z}{2}, 1)^T \tag{5.15}$$

The original positions of the SLs relative to i-th unaligned scan $\mathbf{S}_i$ can be calculated by simply applying the inverse of the similarity transform $\boldsymbol{T}^i$ generated with the reference parameters $\boldsymbol{p}^i$. Furthermore, by applying the similarity transform $\widehat{\boldsymbol{T}}^i$ generated with the predicted parameters $\widehat{\boldsymbol{p}}^i$, the predicted position $\widehat{\boldsymbol{k}}_j^i$ of the SLs is obtained, as shown in Eq. (5.16).

$$\widehat{\boldsymbol{k}}_j^i = \widehat{\boldsymbol{T}}^i \cdot (\boldsymbol{T}^i)^{-1} \cdot \boldsymbol{k}_j \tag{5.16}$$

Since $\boldsymbol{k}_j$ are defined by the dimensions of the scan, the predicted positions $\widehat{\boldsymbol{k}}_j^i$ offer an accurate representation of the effects that the predicted parameters $\widehat{\boldsymbol{p}}$ would have on the transformed volume. This approach manages to implicitly represent the complex relationships between the parameters and the aligned scan without the computational cost associated with transforming the entire volume, as was the case with $\mathfrak{L}_{\text{DSC}}$. Additionally, the quality of the alignment achieved using the predictions of fBAN can be observed by analysing the predicted positions $\widehat{\boldsymbol{k}}_j^i$. A more accurate alignment will results in smaller distances between $\widehat{\boldsymbol{k}}_j^i$ and their corresponding goal position $\boldsymbol{k}_j$.

Now that I had a better representation of the alignment task, I needed a suitable loss function. Since better alignments will translate into smaller distances, my initial idea was to simply use the Euclidean distance between the predicted positions $\widehat{\boldsymbol{k}}_j^i$ and their goal positions $\boldsymbol{k}_j$ as a loss function. However, this proved to be quite unstable, causing the network to quickly converge to a local minimum of $\widehat{p}_s = 0$. The reason for this is likely that it is initially much easier for the network to minimise the distances by placing all the SLs at the origin, than it is to understand the task at hand. However, even if fBAN was pre-trained using $\mathfrak{L}_{\text{MAE}}$, the network still showed a tendency to under-predict the scaling parameter $\widehat{p}_s$, since predicting a scaling parameter that is slightly too small results in shorter distances than if it is slightly too big.

Instead, I decided to reconceptualise the goal. Rather than considering them as individual points, the predicted positions $\widehat{\boldsymbol{k}}_j^i$ of each SL can be treated as a separate cluster. This is shown in Fig. 5.3 as clusters of different colours. Therefore, in order to guide the training of fBAN, a suitable loss function must fulfil two tasks:

- Accuracy: it must minimise the distance between each cluster and their goal position $\boldsymbol{k}_j$
- Precision: It must minimise the spread of each cluster

Therefore, I developed the Spatial Landmarks Loss $\mathfrak{L}_{\text{SLL}}$ to focus specifically on fulfilling these tasks.

For each SL $\boldsymbol{k}_j$, let $\widehat{\boldsymbol{k}}_j$ be the mean of the predicted positions $\widehat{\boldsymbol{k}}_j^i$, as shown in Eq. (5.17).

$$\widehat{\boldsymbol{k}}_j = \frac{1}{n_b} \sum_{i=1}^{n_b} \widehat{\boldsymbol{k}}_j^i \tag{5.17}$$

I define the Accuracy of Spatial Landmarks loss as the mean Euclidean distance between $\boldsymbol{k}_j$ and $\widehat{\boldsymbol{k}}_j$, as shown in Eq. (5.18).

$$\mathfrak{L}_{\text{ASL}} = \frac{1}{n_{SL}} \sum_{j=1}^{n_{SL}} \|\boldsymbol{k}_j - \widehat{\boldsymbol{k}}_j\| \tag{5.18}$$

Similarly, I define the Precision of Spatial Landmarks loss as the mean Euclidean distance between each predicted position $\widehat{\boldsymbol{k}}_j^i$ and their corresponding mean $\widehat{\boldsymbol{k}}_j$, as shown in

$$\mathfrak{L}_{\text{PSL}} = \frac{1}{n_{SL} \cdot n_b} \sum_{j=1}^{n_{SL}} \left(\sum_{i=1}^{n_b} \|\widehat{\boldsymbol{k}}_j - \widehat{\boldsymbol{k}}_j^i\| \right) \tag{5.19}$$

Finally, the Accuracy and Precision of Spatial Landmarks are combined to create the Spatial Landmark Loss function $\mathfrak{L}_{\text{SLL}}$, as shown in Eq. (5.20).

$$\mathfrak{L}_{\text{SLL}} = \frac{1}{n_{SL}} \sum_{j=1}^{n_{SL}} \|\boldsymbol{k}_j - \widehat{\boldsymbol{k}}_j\| + \frac{1}{n_{SL} \cdot n_b} \sum_{j=1}^{n_{SL}} \left(\sum_{i=1}^{n_b} \|\widehat{\boldsymbol{k}}_j - \widehat{\boldsymbol{k}}_j^i\| \right) \tag{5.20}$$

By relying on the Accuracy and Precision of Spatial Landmarks, the previous situation where the network converges to a local minimum of $\widehat{p}_s = 0$ can be avoided. As the results in Tab. 5.4 show, the use of $\mathfrak{L}_{\text{SLL}}$ resulted in significant improvements of 32.3% for the MSE($\boldsymbol{p}, \widehat{\boldsymbol{p}}$), and 8.9% for HD($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$). However, no significant improvement was observed for DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$). While the additional computational cost of calculating $\mathfrak{L}_{\text{SLL}}$ resulted in a 3.5% increase of training time over $\mathfrak{L}_{\text{MAE}}$,

Table 5.4: Performance comparison of fBAN trained with loss functions $\mathfrak{L}_{\text{MAE+DSC}}$ and $\mathfrak{L}_{\text{SLL}}$. The best performance for each measure (Mean Squared Error MSE, Dice Similarity Coefficient DSC, Hausdorff Distance HD) is highlight in bold. The arrows indicate whether a higher (up) or lower (down) value is preferred.

	Loss fn.	
Measure	$\mathfrak{L}_{\text{MAE+DSC}}$	$\mathfrak{L}_{\text{SLL}}$
MSE$(\boldsymbol{p}, \widehat{\boldsymbol{p}}) \downarrow$	0.07 ± 0.35	$\mathbf{0.04 \pm 0.11}$
MSE$(\boldsymbol{p}_T, \widehat{\boldsymbol{p}}_T) \downarrow$	0.0012 ± 0.0014	$\mathbf{0.0004 \pm 0.0002}$
MSE$(\boldsymbol{p}_R, \widehat{\boldsymbol{p}}_R) \downarrow$	0.15 ± 0.81	$\mathbf{0.10 \pm 0.27}$
MSE$(\boldsymbol{p}_S, \widehat{\boldsymbol{p}}_S) \downarrow$	0.005 ± 0.018	$\mathbf{0.004 \pm 0.008}$
DSC$(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}) \uparrow$	0.89 ± 0.05	$\mathbf{0.89 \pm 0.03}$
HD$(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}})$ [voxels] $\downarrow$	15.8 ± 6.6	$\mathbf{14.4 \pm 6.1}$

it resulted in a 79.3% decrease over $\mathfrak{L}_{\text{MAE+DSC}}$. Therefore, the $\mathfrak{L}_{\text{SLL}}$ managed to surpass $\mathfrak{L}_{\text{MAE+DSC}}$ in every aspect.

Nevertheless, in spite of the significant performance improvements, it became apparent that there was something limiting fBAN. In particular, while the network was accurately predicting the translation and scaling parameters, it was struggling to achieve a similar performance for the rotation parameters. While it was possible that this was the result of an architectural limitation, I was inclined to think that this performance bottleneck might be the result of "cheating".

5.2.6 Transfer learning against cheating

Due to the difficulty associated with analysing the inner-workings of DL networks, they are often considered "black boxes". As a result, understanding how a DL network reached their predictions is not trivial. Therefore, it is difficult to know if a network is cheating.

By cheating, what I mean is a network that managed to find a way of minimising the loss function not by understanding the task at hand, but through alternate ways. For example, take a network that has been train to classify pictures of cats and dogs. Ideally, the network would achieve this by learning to understand the features of each animal. However, if all the cat pictures are indoors, while all the dog pictures are outdoors, the network might find it easier to rely on this information instead.

While the resulting predictions might be just as accurate, the network did not learn to understand the features of each animal, i.e., it cheated.

In the case of fBAN, I believed that the network had managed to find a way to cheat when predicting the rotation parameters. As discussed in Sec. 2.2, the clinical guidelines try to minimise the positional variability of the fetal head inside the 3D US scan through the acquisition protocol used by sonographers to place the probe. However, this protocol does not consider the Inferior-Superior, Posterior-Anterior, or Left-Right orientations of the head. This can be seen in the original positions shown in Fig. 5.3 as clusters of the same colours on opposite sides. As a result, the bulk of the rotation part of the alignment task consists of correcting these orientations.

Therefore, the concern was that rather than learning to fully understand the spatial position of the brain in the scan, fBAN was cheating by learning to predict corrections for these orientations rather than an accurately rotating the brain into the canonical space. This is likely to be a much simpler task for the network to learn, as it doesn't require an understanding of the structural information of the brain and can likely be achieved by relying on a handful of salient features, such as the neck location or the shape of the US beam. Additionally, since correcting these orientations makes up most of the rotation needed to align the scan, it would result in a significant reduction of the loss function, especially during the early stages of training. However, this approach would be too coarse to accurately correct the rotation of each scan, which would explain the lower accuracy of the predicted rotation parameters, as shown in Tab. 5.4. Thus, I needed a method that could ensure that the predictions generated by fBAN were performed based on the structural information of the brain, and not through such a cheating approach.

I achieved this through a Transfer Learning approach. Since cheating in the same manner would not be beneficial for the task of brain extraction, the fBEN has no incentive in learning to extract the required features from the scan. Instead, it learns to extract the relevant structural information of the head in order to generate its prediction. Therefore, by transferring the knowledge learned by fBEN to fBAN, the latter can be forced to predict the alignment parameters based only on this

structural information. This can be easily achieved by simply freezing the rest of the network and training only the two Dense layers of the regression architecture of fBAN to predict the alignment parameters $\hat{\boldsymbol{p}}$ based on the output of the encoder section of the trained fBEN. However, this approach would limit the alignment task to be performed based only on the structural information that was useful for the extraction task, which is likely to limit the performance.

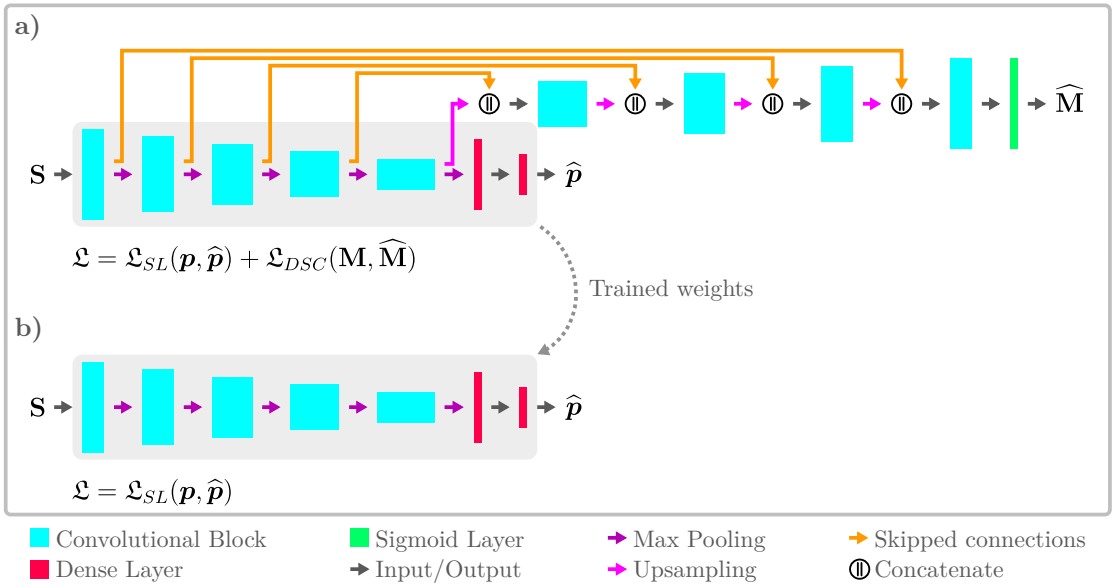

Figure 5.4: Schematics of the Transfer Learning method used to train fBAN. **a)** Two Dense layers are appended to the encoder of the trained fBEN. The network is then retrained to predict the brain mask $\widehat{\mathbf{M}}$ and the alignment parameters $\hat{\boldsymbol{p}}$. **b)** The decoder and skipped connection are removed, and the remaining regression network is refined with a final training.

Instead, I developed a method to use the learned knowledge of fBEN to simply encourage fBAN to find the relevant structural information to solve its own task. As shown in Fig. 5.4a, I started by attaching the Dense layers to the output of the encoder section of the trained fBEN, thus creating fBAN. I then retrained the combined network to predict the alignment parameters $\hat{\boldsymbol{p}}$ in addition to the extraction mask $\widehat{\mathbf{M}}$, using a combined loss function $\mathfrak{L} = \mathfrak{L}_{\text{SLL}} + \mathfrak{L}_{\text{DSC}}$. This allows for the learned knowledge of fBEN to be transferred to fBAN, while still allowing the network to learn to extract additional structural information that might be beneficial. Afterwards, as shown in Fig. 5.4b, the decoder section is removed and fBAN is refined through a final training.

The results shown in Tab. 5.5 show that the Transfer Learning approach resulted in a significant improvement for every measure, when compared to training the network from scratch. There was a 58.3% improvement for MSE($\boldsymbol{p}, \widehat{\boldsymbol{p}}$), of which the most substantial change was observed for the rotational parameters, with the MSE($\boldsymbol{p}_R, \widehat{\boldsymbol{p}}_R$) reducing by 58.7%. This translated into statistically significant improvements of 30.0% for HD($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) and 1% for DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$). Since forcing fBAN to rely on the structural information of the head resulted in a disproportionate improvement of predicted rotation parameters $\widehat{\boldsymbol{p}}_R$, it is likely that my concerns with regards to cheating were correct. However, without the means to analyse the inner workings of fBAN, it is not possible to confirm this theory, as the network remains a black box model.

Table 5.5: Performance comparison of fBAN trained with and without the Transfer Learning approach. The best performance for each measure (Mean Squared Error MSE, Dice Similarity Coefficient DSC, Hausdorff Distance HD) is highlight in bold. The arrows indicate whether a higher (up) or lower (down) value is preferred.

	Training method	
Measure	Without Transfer Learning	With Transfer Learning
MSE($\boldsymbol{p}, \widehat{\boldsymbol{p}}$) ↓	0.04 ± 0.11	**0.02 ± 0.02**
MSE($\boldsymbol{p}_T, \widehat{\boldsymbol{p}}_T$) ↓	0.0004 ± 0.0002	**0.0002 ± 0.0002**
MSE($\boldsymbol{p}_R, \widehat{\boldsymbol{p}}_R$) ↓	0.10 ± 0.27	**0.04 ± 0.05**
MSE($\boldsymbol{p}_S, \widehat{\boldsymbol{p}}_S$) ↓	0.004 ± 0.008	**0.002 ± 0.006**
DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) ↑	0.89 ± 0.03	**0.90 ± 0.04**
HD($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) [voxels] ↓	14.4 ± 6.1	**10.1 ± 3.9**

In spite of the significant improvements achieved through the Transfer Learning approach to training fBAN, the DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) performance was underwhelming when compared to the performance achieved by fBEN. Since I had already worked on improving the loss function, and the training methodology, I concluded that the performance might be limited by the network architecture.

5.2.7 Cascade architecture

At this point of development, I had a well optimised network, a robust loss function, and a reliable training methodology, achieving state-of-the-art performance for

the task of aligning the fetal brain from 3D US scans. However, the performance achieved by the extraction network fBEN lead me to believe that there was still room for improvement.

So far, I had worked with an architecture that was derived from the encoder section of the fBEN network, under the hypothesis that most of the alignment information would be captured there, which has been categorically confirmed by the results. However, in doing so, I had ignored the proportion of information that is being passed through the skipped connections and retrieved by the decoder, potentially constraining the performance of fBAN.

fBAN was managing to correct the bulk of the translation, rotation, and scaling, but the final alignment was somewhat coarse. I reasoned that the intrinsic variability of the data and the difficulty of the task were overwhelming the learning ability of the network. Therefore, the simplest solution would be to append a secondary alignment network, resulting in a cascade architecture [180][181][182]. The assumption being that this cascade architecture would allow most of the complexity of the task to be handled by the first network, leaving enough learning bandwidth for the second network to focus solely on the task of refining the prediction done by the first.

As shown in Fig. 5.5, the original network is trained using the same Transfer Learning method discussed in Sec. 5.2.6. A final fBAN network with a cascade architecture is created, and the learned knowledge of the original network is transferred onto the first subnetwork (X1), after which its weights are frozen. The same knowledge is also transferred to the second subnetwork (X2), to encourage it to rely on similar structural information of the head for its predictions.

During training, the unaligned scan $\mathbf{S}$ is passed as input to X1, which predicts a first set of alignment parameters $\widehat{\boldsymbol{p}}_1$. The similarity transform $\mathbf{M}_1$ generated with these parameters is then used to align $\mathbf{S}$. The resulting scan $\mathbf{S}^{\widehat{p}_1}$ is then passed as input to X2, which in turn predicts a second set of alignment parameters $\widehat{\boldsymbol{p}}_2$ used to generate the similarity transform $\mathbf{M}_2$. The combination of both transforms results in the final transform $\mathbf{M}$, as shown in Eq. (5.21).

$$\mathbf{M} = \mathbf{M}_2 \cdot \mathbf{M}_1 \tag{5.21}$$

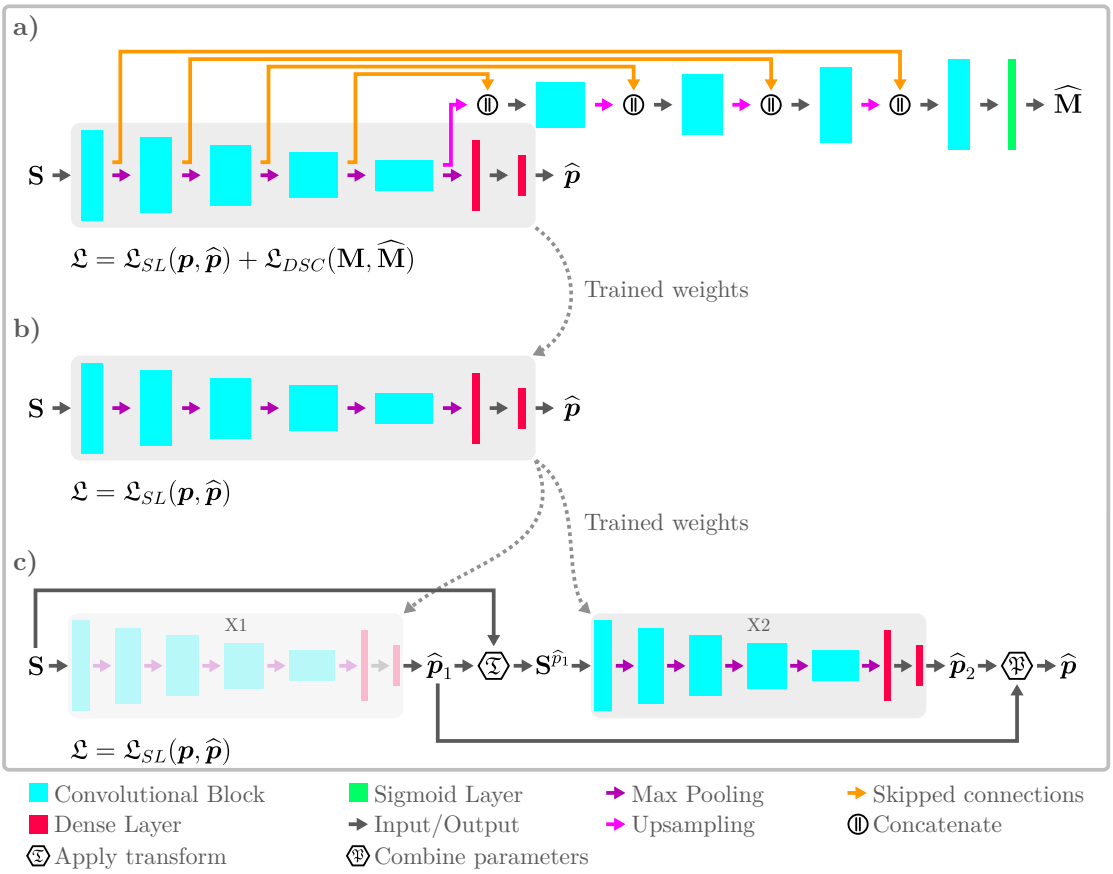

Figure 5.5: Schematics of the training method used for the cascade architecture of fBAN. **a)** Two Dense layers are appended to the encoder of the trained fBEN. The network is then retrained to predict the brain mask $\widehat{\mathbf{M}}$ and the alignment parameters $\widehat{\boldsymbol{p}}$. **b)** The decoder and skipped connection are removed, and the remaining regression network is refined with a final training. **c)** The trained regression network is duplicated, and the cascade architecture is created. The first subnetwork of fBAN (X1) receives the unaligned scan $\mathbf{S}$ and predicts a first set of parameters $\widehat{\boldsymbol{p}}_1$. The corresponding aligned scan $\mathbf{S}^{\widehat{\boldsymbol{p}}_1}$ is passed as input to the second subnetwork (X2), which predicts a second set of parameters $\widehat{\boldsymbol{p}}_2$ that are combined with the first set to produce the final prediction $\widehat{\boldsymbol{p}}$. The trained weights of X1 are frozen, before training the network a final time.

Finally, $\mathbf{M}$ is deconstructed into the corresponding predicted parameters $\widehat{\boldsymbol{p}}$.

The results comparing the performance of using a single network (X1), a two-network cascade architecture (X2), and a three-network cascade architecture (X3) are shown in Tab. 5.6. As expected, X2 resulted in statistically significant improvements across all measures. MSE$(\boldsymbol{p}, \widehat{\boldsymbol{p}})$ was reduced by 71.0%, most of which due to a 72.1% improvement in the MSE$(\boldsymbol{p}_R, \widehat{\boldsymbol{p}}_R)$ of the rotation parameters. These improvements in the accuracy of the predicted parameters resulted in a 3.3% improvement of the DSC$(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}})$, and a 13.9% of the HD$(\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}})$, finally reaching a performance

similar to that of fBEN. The results also show that a third network was not beneficial, with no statistically significant differences in performance between X2 and X3.

Table 5.6: Performance comparison of fBAN with three different number of subnetworks. The best performance for each measure (Mean Squared Error MSE, Dice Similarity Coefficient DSC, Hausdorff Distance HD) is highlight in bold. The arrows indicate whether a higher (up) or lower (down) value is preferred.

	Subnetwork		
Measure	X1	X2	X3
MSE($\boldsymbol{p}, \widehat{\boldsymbol{p}}$) [$10^{-3}$] ↓	18.6 ± 21.2	5.3 ± 5.6	5.3 ± 5.5
MSE($\boldsymbol{p}_T, \widehat{\boldsymbol{p}}_T$) [$10^{-3}$] ↓	0.24 ± 0.19	**0.11 ± 0.11**	**0.11 ± 0.11**
MSE($\boldsymbol{p}_R, \widehat{\boldsymbol{p}}_R$) [$10^{-3}$] ↓	42.4 ± 46.6	**11.8 ± 12.5**	11.9 ± 12.6
MSE($\boldsymbol{p}_S, \widehat{\boldsymbol{p}}_S$) [$10^{-3}$] ↓	2.2 ± 5.9	1.4 ± 4.0	**1.3 ± 4.0**
DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) ↑	0.90 ± 0.04	**0.94 ± 0.02**	0.93 ± 0.02
HD($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) [voxels] ↓	10.1 ± 3.9	**8.7 ± 3.3**	**8.7 ± 3.2**

5.2.8 Final refinement

At this stage of development I had successfully developed a state-of-the-art solution for the task of automated fetal brain alignment from minimally-preprocessed 3D US scans. However, in order for fBAN to be a reliable module of the pipeline, it must be able to maintain this level of performance, regardless of the scan quality.

So far, the development of fBAN has been performed using dataset $\mathfrak{D}_A$, which consists of scans selected specifically avoiding particularly low-quality examples. Therefore, just as discussed in Sec. 4.2.5, I ensured that fBAN was robust to scan quality by performing a final refinement of the network with dataset $\mathfrak{D}_B$. In addition to containing more than three times the amount of scans, $\mathfrak{D}_B$ purposefully contains scans of varied quality, which should result in a more robust network performance. I expanded upon this by also randomising the orientation of the input scans during training and validation, effectively increasing from 4290 scans contained in dataset $\mathfrak{D}_B$ to 102960 scans in dataset $\mathfrak{D}_{B^*}$. This augmentation has the additional benefit of ensuring that the network performs consistently, regardless of the initial orientation of the input scan.

Accordingly, fBENv2 was used as the base network for the Transfer Learning approach. However, that network was trained with full-resolution (0.6 mm/vxl) scans, while the fBAN development was constrained to their low-resolution (1.2 mm/vxl) counterparts. Therefore, I removed this constraint as part of this final refinement.

I also opted for using the unity quaternion $(p_{q_w}, p_{q_x}, p_{q_y}, p_{q_z})$ instead of the Euler angles $(p_\alpha, p_\beta, p_\gamma)$ for the rotation parameters. While Euler angles are generally easier to interpret, and can be beneficial for some applications, they have several limitations. Firstly, they allow for multiple possible solutions to the same rotation transform, making it difficult to analyse the accuracy of the predicted parameters. Additionally, since the final rotation transform is comprised of three subsequent rotations for each Euler angle, as shown in Eq. (5.3), the network needs to understand that the effect a rotation parameter has on the overall alignment can depend on the previous parameters, making the task more challenging. Finally, the use of Euler angles can also result in gimbal lock [183], i.e., the loss of one degree of freedom in 3D rotations. While this was not an issue during the initial development, it became a problem when using dataset $\mathfrak{D}_B$. Since its canonical space involves two π rotations more than the canonical space of $\mathfrak{D}_A$ (See Sec.3.4), it significantly increased the likelihood of a gimbal lock occurring. In contrast, by using unit quaternion, these issues can be eliminated without any significant drawback. While it is still possible to have two solutions to the same rotation transform, this can be avoided entirely by simply ensuring that the first parameter p_{q_w} is always positive.

Finally, I performed a final optimisation of fBAN in the same iterative manner as discussed in Sec. 5.2.4. This resulted in the removal of one of the Dense layers, as it provided no additional benefits. The final architecture of fBAN is shown in Fig. 5.6. Each subnetwork has a convolutional kernel size of $ks = (3, 3, 3)$, a depth of $l = 4$, and $hd = (16, 32, 64, 128, 256)$ hidden dimensions, and was trained for 100 epochs using the AdamW optimiser with a learning rate of $lr = 0.001$.

An exhaustive analysis of the performance of fBAN can be found in Sec. 5.3.

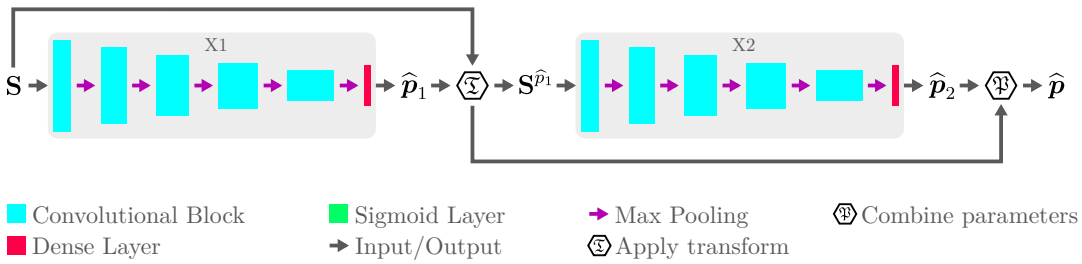

Figure 5.6: Schematics of the final cascade architecture for fBAN.

5.3 Results

In this section I perform an exhaustive analysis of the performance of fBAN. I compare its performance against alternative methods, and evaluate its performance by gestational age, and brain misalignment. I also assess the robustness of the network against input orientation. Finally, I show the usefulness of SL as an evaluation measure.

5.3.1 Mean performance

Table 5.7: Average performance of fBAN compared with multiple alternative methods. All methods were evaluated using the hold-out testing split of the $\mathfrak{D}_B$ dataset. The best performance for each measure (Mean Squared Error MSE, Dice Similarity Coefficient DSC, Hausdorff Distance HD) is highlight in bold. The arrows indicate whether a higher (up) or lower (down) value is preferred.

Method	MSE($\boldsymbol{p}, \widehat{\boldsymbol{p}}$) [$10^{-3}$] ↓	DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) ↑	HD($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) [voxels] ↓
Unaligned	164 ± 7	0.53 ± 0.20	41.1 ± 7.30
SimpleElastix	175 ± 354	0.44 ± 0.21	103.1 ± 34.5
ANTS	214 ± 320	0.53 ± 0.20	67.7 ± 29.1
Namburete	157 ± 259	0.67 ± 0.18	34.3 ± 8.2
fBAN	**0.86 ± 1.11**	**0.93 ± 0.02**	**8.9 ± 3.8**

The mean performance of fBAN on the hold-out testing split of dataset $\mathfrak{D}_B$ is shown in Tab. 5.7. As the results show, in spite of the more challenging nature of $\mathfrak{D}_B$, fBAN managed to achieve a similar performance to what it had previously achieved with $\mathfrak{D}_A$, seen in Tab. 5.6. However, it is worth mentioning that a direct comparison of these results is not possible, since each dataset has a different

canonical space and rotational parameters (Euler angles vs quaternions), and was tested on a different number of scans.

Table 5.7 also shows the performance of several alternative methods. First, the *Unaligned* performance, i.e. the results achieved relying only on the positioning of the probe by the sonographer, is shown. This is not intended to assess the accuracy of this positioning, but rather as a baseline to contextualise the results achieved by the other methods. The *SimpleElastix* [178] and *ANTS* [179] are the same methods discussed in Sec. 4.3.1. Each method was used to register the average masked and aligned scan for the corresponding GW to each scan, from which the alignment parameters can be extracted. For each of the two methods, an exhaustive evaluation of the possible configurations was performed, including several different metrics and transform types, and the highest performance achieved are shown. However, neither method was able to register the template to the scans in an accurate and reliable manner, which translated to a lower performance than what was achieved with the *Unaligned* scans. While these methods have been extensively used in the literature, with state-of-the-art performance for multiple imaging modalities, the high intrinsic variability of the 3D US data, and the large positional variability seem to have hinder their performance. Finally, the performance achieved by *Namburete* [104] is shown. In contrast to *SimpleElastix* and *ANTS*, this method did result in a significant improvement over the *Unaligned* scans. Nevertheless, this performance is significantly lower to that achieved in [104], where they obtained a DSC of 0.9. This is likely due to the significantly more challenging dataset $\mathfrak{D}_B$, as that network as trained on a high-quality dataset of similar to $\mathfrak{D}_A$.

The results shown in Tab. 5.7 confirm that fBAN significantly outperforms every other method, improving over its closest competitor by 99.5% for MSE($\boldsymbol{p}, \widehat{\boldsymbol{p}}$), 39.2% for DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$), and 74.1% for HD($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$). A significant part of this performance is thanks to the focus on robustness during the final refinement discussed in Sec. 5.2.8. To highlight this, the same network was trained with dataset $\mathfrak{D}_B$, but using the training and validation split of $\mathfrak{D}_A$, i.e., only the high-quality scans. As Tab. 5.8 shows, this results in a significant decrease in performance across

all measures, confirming the importance of conveying the true quality variability of the data during training.

Table 5.8: Average performance of fBAN trained with the training and validation split of $\mathfrak{D}_A$, compared to the full splits of $\mathfrak{D}_B$. Note that both networks were trained with the same data, with the only difference being that the $\mathfrak{D}_A$ splits contain only high-quality scans, while the $\mathfrak{D}_B$ splits contain mixed-quality scans. Both methods were evaluated using the hold-out testing split of the $\mathfrak{D}_B$ dataset. The best performance for each measure (Mean Squared Error MSE, Dice Similarity Coefficient DSC, Hausdorff Distance HD) is highlight in bold. The arrows indicate whether a higher (up) or lower (down) value is preferred.

Method	MSE($\boldsymbol{p}, \widehat{\boldsymbol{p}}$) [$10^{-3}$] ↓	DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) ↑	HD($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) [voxels] ↓
fBAN ($\mathfrak{D}_A$)	20.73 ± 22.61	0.79 ± 0.19	15.2 ± 6.5
fBAN ($\mathfrak{D}_B$)	**0.86 ± 1.11**	**0.93 ± 0.02**	**8.9 ± 3.8**

5.3.2 Performance vs. gestational week

Figure 5.7 shows the performance of fBAN by GW. The results show that fBAN is robust against the gestational age of the scan, managing to achieve consistently high performance for every measure throughout the entire gestational range, with only minor drops in performance around the edges of the range.

The MSE($\boldsymbol{p}, \widehat{\boldsymbol{p}}$) performance of fBAN shows a slight decrease in performance at 14 GW and 15 GW, with a higher mean and standard deviation. When separated by parameter type, MSE($\boldsymbol{p}_R, \widehat{\boldsymbol{p}}_R$) and MSE($\boldsymbol{p}_S, \widehat{\boldsymbol{p}}_S$) reveal a similar behaviour for the predicted rotation and scaling parameters. In contrast, the opposite is observed for the translation parameters MSE($\boldsymbol{p}_T, \widehat{\boldsymbol{p}}_T$), with a slight drop in performance observed after 26 GW. This behaviour can be explained by the size of a voxel relative to the fetal brain. At earlier GWs the smaller size of the fetal brain results in a proportionally larger voxel. Therefore, a difference of 1 voxel when defining the brain size and shape results in a proportionally larger impact in the rotation and scaling required to align it to a canonical space, making it harder to accurately predict the corresponding parameters. However, the relatively coarser voxels also make it easier to determine which voxel is closest to the centre of the brain. Conversely, the larger fetal brain at later GWs makes this less clear, resulting

in a slightly larger variation in predicted translation parameters required to align the centre of the brain with the centre of the 3D scan.

The effect of these trends can be seen reflected in the DSC($\mathbf{M}^{\boldsymbol{p}}, \widehat{\mathbf{M}^{\boldsymbol{p}}}$) and HD($\mathbf{M}^{\boldsymbol{p}}, \widehat{\mathbf{M}^{\boldsymbol{p}}}$) performance, which also show a slight drop at earlier GWs, and a smaller drop around 30 GWs. However, the lowest performing week still yields accurate results, with a mean DSC($\mathbf{M}^{\boldsymbol{p}}, \widehat{\mathbf{M}^{\boldsymbol{p}}}$) of 0.90 and a mean HD($\mathbf{M}^{\boldsymbol{p}}, \widehat{\mathbf{M}^{\boldsymbol{p}}}$) of 13.4 voxels (which translates to 3.2 mm without scaling).

Figure 5.8 shows representative examples of the mid-axial, mid-coronal, and mid-sagittal planes for several GWs. These examples have been chosen because their DSC($\mathbf{M}^{\boldsymbol{p}}, \widehat{\mathbf{M}^{\boldsymbol{p}}}$) performance is the closest to the mean performance of fBAN for the corresponding GW. These examples show that the alignment predicted by fBAN is remarkably close to the manual alignment. The accuracy and consistency that fBAN achieves for these examples helps reiterate the robustness of the network against scan quality, age, and positional variability. It also helps illustrate the limited effect that the drops in performance around the edges of the gestational range shown in Fig. 5.7 have on the performance of the network.

Additionally, these examples show that fBAN manages to surpass the manual alignment in some cases. This is particularly clear for the axial view of the 14 GW example. The sagittal views also show that the alignment of fBAN can be more consistent, something that is reflected in the location and orientation of the CSP. This is not surprising, as the lack of symmetric landmarks makes it especially difficult to manually align the sagittal plane in a consistent manner.

To better visualise the higher consistency of fBAN, Fig. 5.8 shows the orthogonal midplanes of the mean aligned scan of the same GWs as before. For every GW, the results generated by fBAN are sharper, and have higher contrast than those generated through manual alignment. While these improvements are relatively minor and hard to see in some cases, such as the coronal and axial views at 20-23 GWs, it is evident at the edges of the gestational range, where the improved consistency allows for some structures to be resolved that were not well defined before, such as the HG in the axial view at 14 GW and the left SF in the coronal

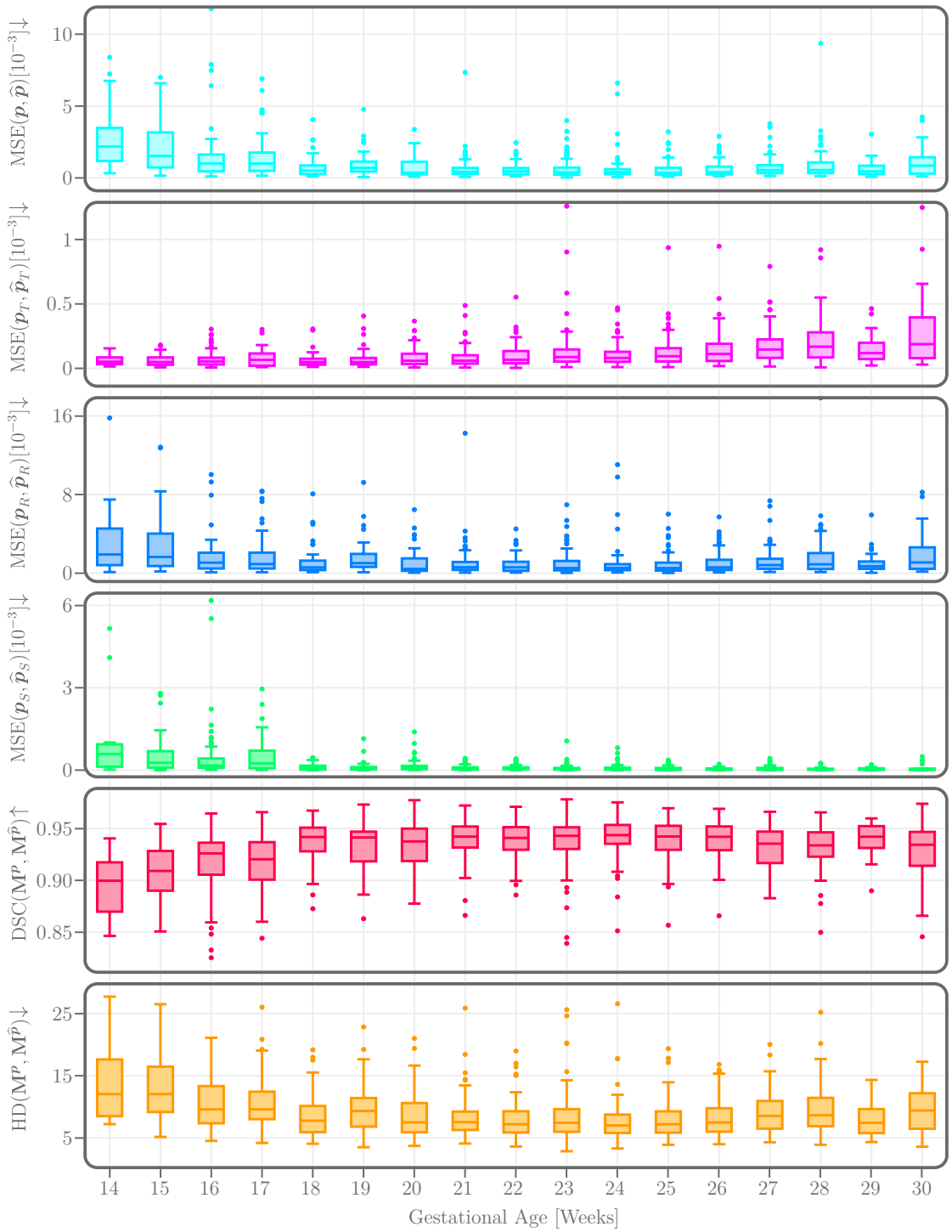

Figure 5.7: Performance measures of fBAN, separated by GW. From top to bottom: Mean Squared Error of predicted parameters MSE($\boldsymbol{p}, \widehat{\boldsymbol{p}}$), Mean Squared Error of predicted translation parameters MSE($\boldsymbol{p}_T, \widehat{\boldsymbol{p}}_T$), Mean Squared Error of predicted rotation parameters MSE($\boldsymbol{p}_R, \widehat{\boldsymbol{p}}_R$), Mean Squared Error of predicted scaling parameter MSE($\boldsymbol{p}_S, \widehat{\boldsymbol{p}}_S$), Dice Similarity Coefficient of aligned mask DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$), and Hausdorff Distance of aligned mask HD($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$). The arrows indicate whether a higher (up) or lower (down) value is preferred.

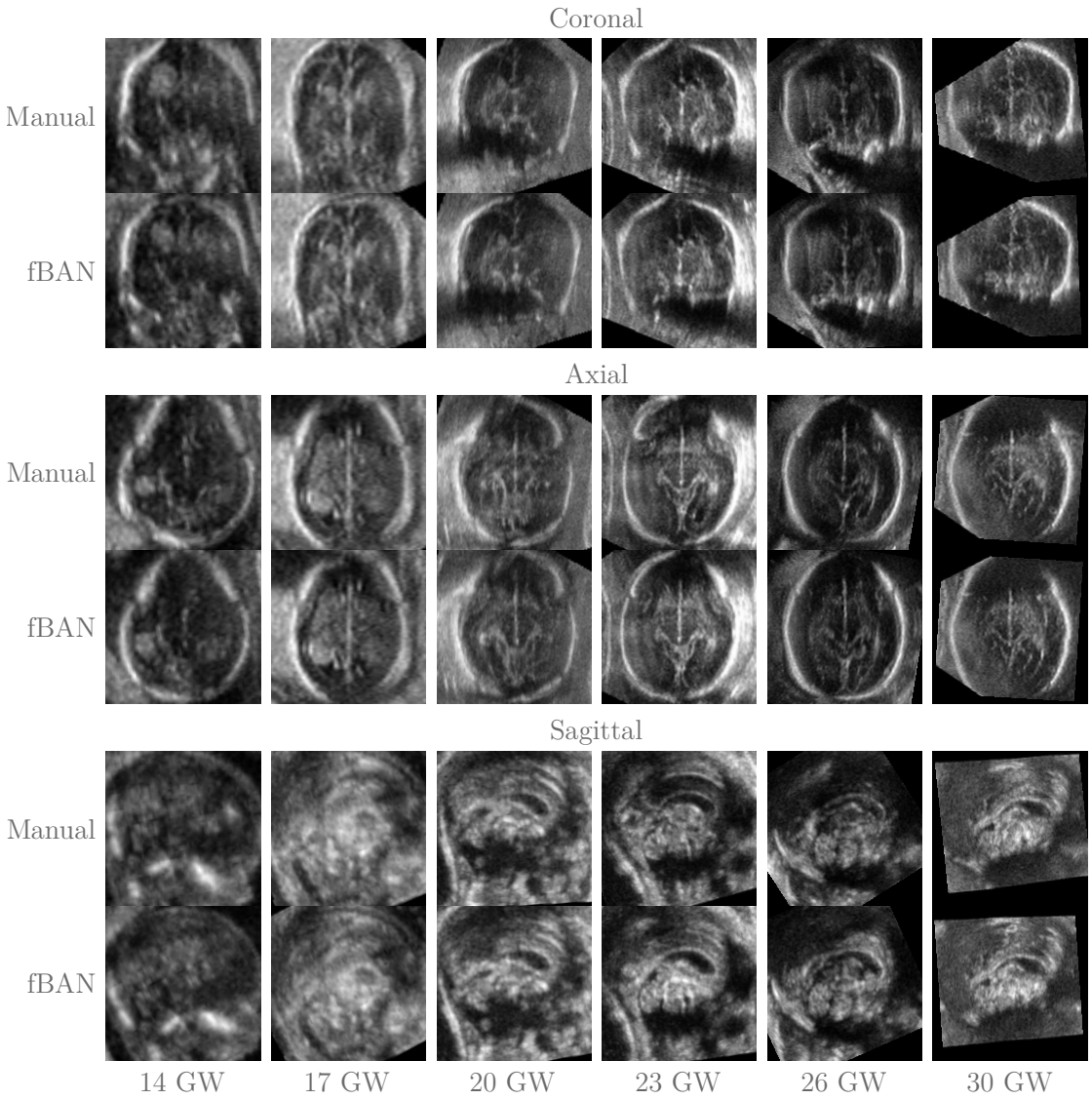

Figure 5.8: Examples of the alignment predicted by fBAN compared against the manual alignment, for multiple GW. These examples were specifically selected since their DSC($\mathbf{M}^{\boldsymbol{p}}, \widehat{\mathbf{M}^{\boldsymbol{p}}}$) performance is closest to the mean performance of fBAN for that GW. Top: Coronal midplane. Middle: Axial midplane. Bottom: Sagittal midplane.

view at 30 GWs. This could indicate that the slight decrease in performance shown in Fig. 5.7 are a reflection of the decreased quality of the manual alignment, instead.

5.3.3 Performance vs. misalignment

The results have shown that fBAN performs accurately and consistently for the entire gestational range, with a performance that is nearly invariant to the GW of the fetus. However, in order for the module to perform consistently in the pipeline,

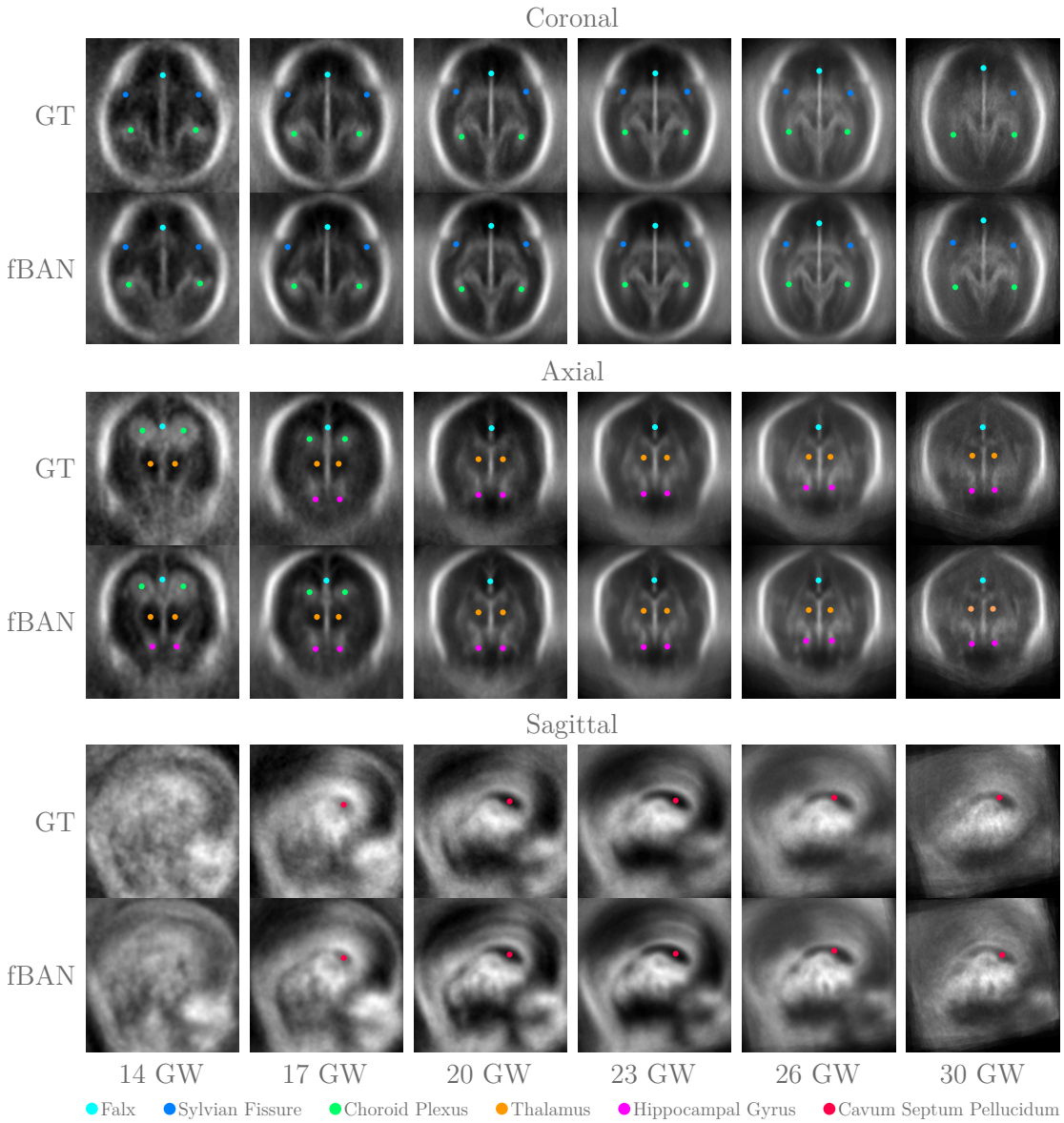

Figure 5.9: Mean aligned scans predicted by fBAN compared against the manual alignment, for multiple GW. Top: Coronal midplane. Middle: Axial midplane. Bottom: Sagittal midplane.

it is also crucial that the performance is not dependent on the initial misalignment of the brain relative to the canonical space.

To do this, I first analyse the performance of fBAN for each alignment parameter separately. This is shown in Fig. 5.10, where they have been plotted against their target value, along with their linear regression and the corresponding R^2 score. The predicted translation parameters $\widehat{p}_x$, $\widehat{p}_y$, and $\widehat{p}_z$ are highly accurate, lying almost entirely along the line of equality, with a minimal amount of outliers. The

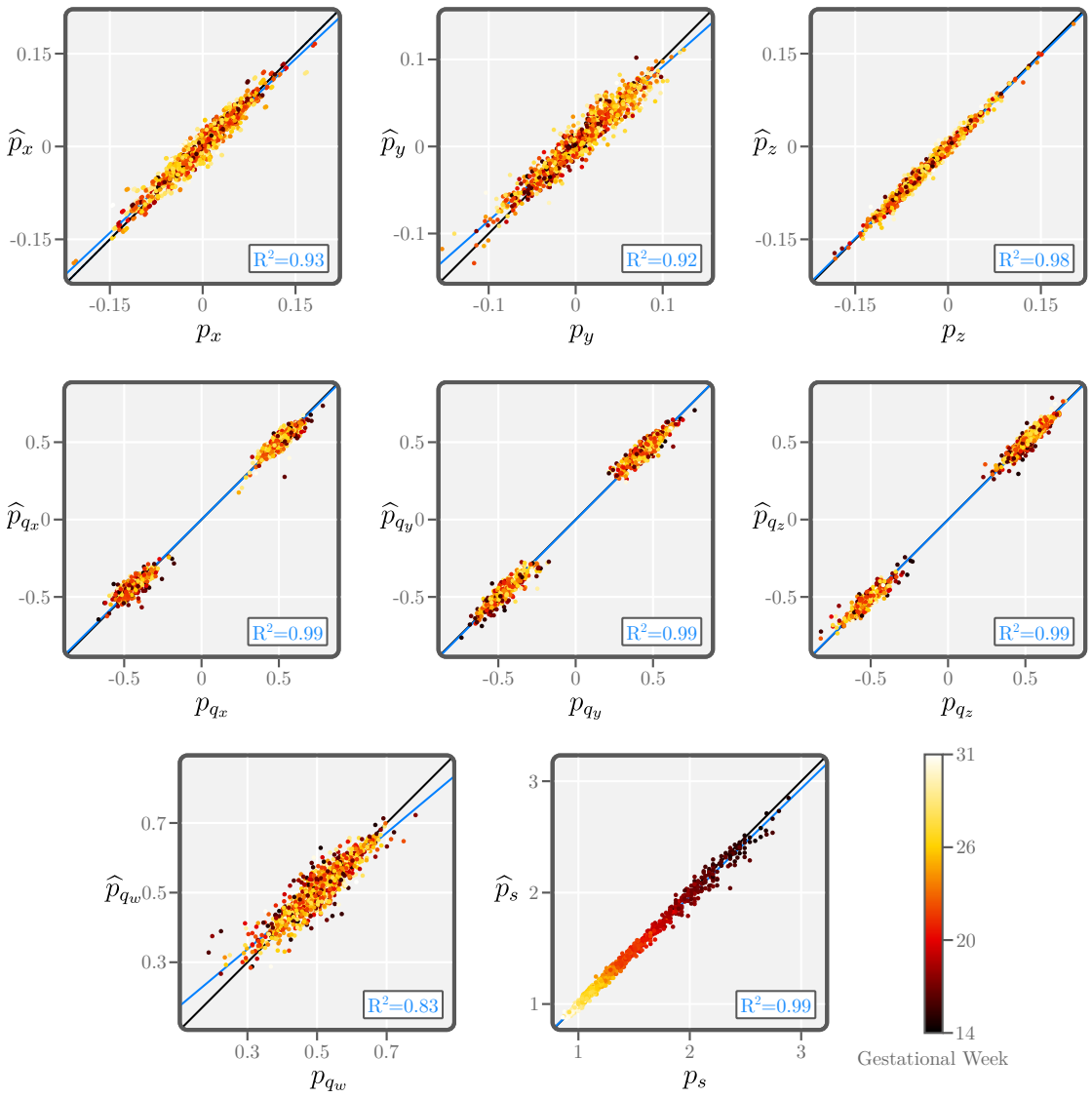

Figure 5.10: Predicted parameters plotted against their target value. The line of equality is shown in black, and the linear regression for each parameter is shown in blue, with its R^2 score displayed in the bottom-right corner.

absolute error of the predicted parameters also shows only a weak correlation with the magnitude of the references, with Pearson's correlation coefficients lying between -0.04 and -0.06.

The predicted rotation parameters $\widehat{p}_{q_w}$, $\widehat{p}_{q_x}$, $\widehat{p}_{q_y}$, and $\widehat{p}_{q_z}$ also lie consistently along the line of equality, with the former showing slightly more variability than the rest. The latter three are also arranged in two distinct clusters, as a consequence of the clinical acquisition protocol, which does not take into account the Inferior-

Superior, Posterior-Anterior, or Left-Right orientations of the head. With correlation coefficients lying between -0.004 and -0.20, the correlation between the absolute error of the predicted parameter the magnitude of the references is also weak.

Finally, the predicted scaling parameter $\widehat{p}_s$ are also consistent with p_s, lying along the line of equality, with a minimal amount of outliers. However, unlike the other parameters, absolute error of $\widehat{p}_s$ shows a modest correlation to the magnitude of p_s, with a correlation coefficient value of 0.48. However, this is not surprising, as the magnitude of the p_s is directly related to gestational age of the fetus. Therefore, this is consistent with the results discussed in Sec. 5.3.2, and only represents a slight decrease in accuracy.

After analysing the accuracy of each predicted parameter separately, I proceeded to analyse the performance of fBAN relative to the misalignment that these parameters represent. Since a straightforward definition of misalignment is not trivial, I separated the assessment into three components, in the same manner discussed in Sec. 4.3.3. The first of these components is the translation misalignment, which was defined as the Euclidean distance between the centre of the fetal brain and the centre of the 3D scan. The second is the rotation misalignment, which was defined as the angle of the difference rotation of two quaternions. Finally, since the scaling predicted by fBAN is isometric, the scaling misalignment is simply the scaling parameter p_s. For each of these three components, the performance of fBAN through the MSE($\boldsymbol{p}, \widehat{\boldsymbol{p}}$), DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$), and HD($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$) measures is analysed, as shown in Fig. 5.11.

The results show that the performance of fBAN is weakly correlated to the translation misalignment, with Pearson's correlation coefficient of 0.12, 0.15, and -0.21 for MSE($\boldsymbol{p}_T, \widehat{\boldsymbol{p}}_T$), DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$), and HD($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$), respectively. In other words, the performance of fBAN shows a slight decrease in performance the further the centre of the brain is from the centre of the scan. Nevertheless, this difference is very minor, as the previous results have already shown.

The correlation is even weaker for the rotation misalignment, with correlation coefficients of 0.07, 0.06, and -0.02. Along with the results shown in Fig. 5.10, this

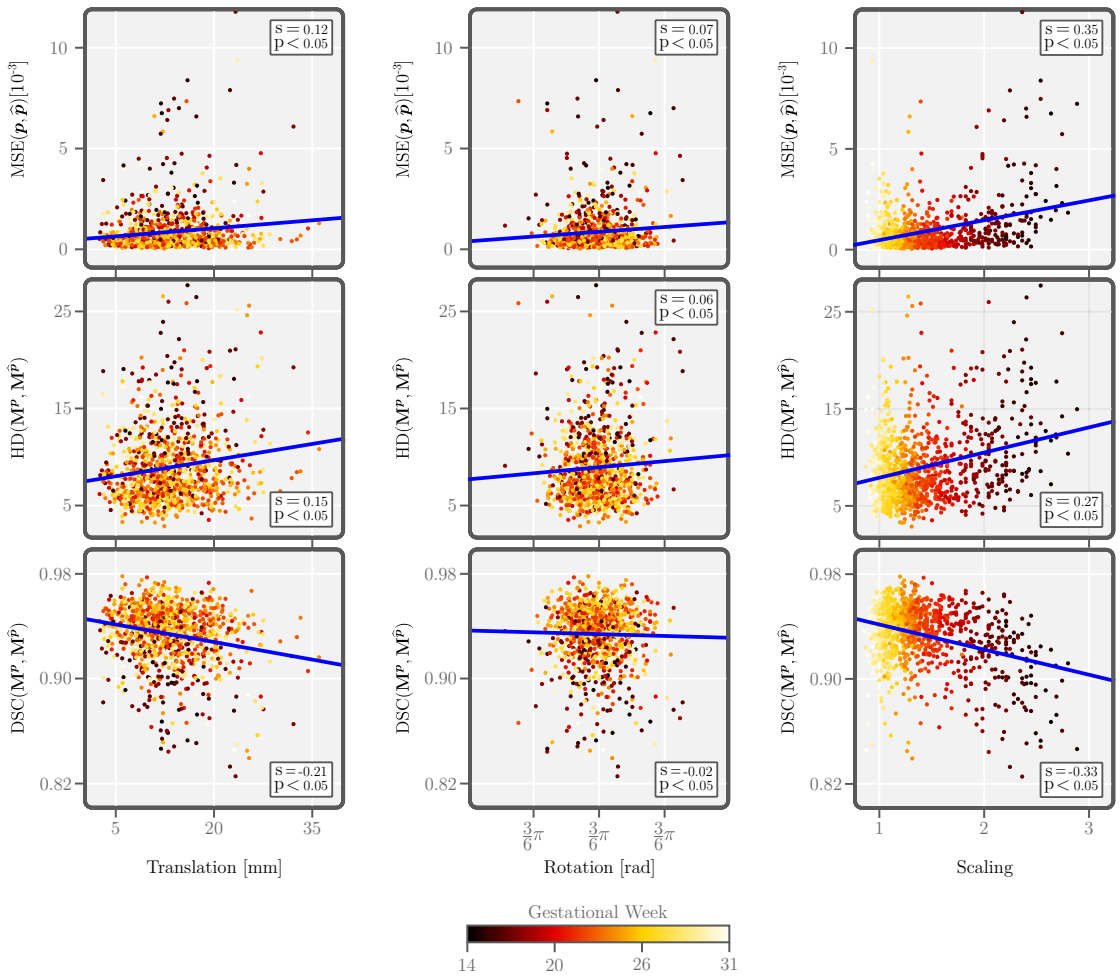

Figure 5.11: Performance of fBAN for Mean Squared Error of predicted parameters MSE($\boldsymbol{p}, \widehat{\boldsymbol{p}}$), Dice Similarity Coefficient of aligned mask DSC($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$), and Hausdorff Distance of aligned mask HD($\mathbf{M}^{\boldsymbol{p}}, \mathbf{M}^{\widehat{\boldsymbol{p}}}$), against brain misalignment. The Pearson correlation coefficient s, and the corresponding p-value are provided for each pair, as well as a linear fit for visualisation purposes. Translation misalignment is the Euclidean distance between the centre of the brain and the centre of the volume, the Rotation misalignment is the cosine distance between the orientation of the head and the canonical space, and the Scaling misalignment is the factor needed to scale the brain volume to match the average brain volume at 30 GWs.

confirms that the performance of fBAN is essentially invariant to the orientation of the head inside the scan.

Finally, the results show that the performance of fBAN is modestly correlated to the scaling misalignment, with correlation coefficients 0.35, 0.27, and -0.33. This is consistent with the results discussed in Sec. 5.7, as the scaling is directly correlated with the gestational age of the fetus. However, as Fig. 5.7 shows, this drop in performance for earlier GW is minor.

5.3.4 Performance consistency

Just as in Sec. 4.3.4, the consistency of the alignment predicted by fBAN is assessed in relation to the orientation of the input scan $\mathbf{S}$. Each scan of the hold-out test split of dataset $\mathfrak{D}_B$ was augmented to its 24 possible orientations and fBAN was used to predict the alignment parameters $\widehat{\boldsymbol{p}}$ for each one. The inverse augmentation was then applied, resulting in 24 sets of parameters that aim to align the scan in its original orientation to the canonical space. For each set, MSE($\widehat{\boldsymbol{p}}$), DSC($\mathbf{M}^{\widehat{\boldsymbol{p}}}$), and HD($\mathbf{M}^{\widehat{\boldsymbol{p}}}$) were measured against the other 23, after which their mean, median, and standard deviation values were calculated. As Tab. 5.9 shows, the average results are significantly better than those obtained against the reference parameters $\boldsymbol{p}$, indicating that the consistency of the network is higher than its performance.

Table 5.9: Performance consistency of fBAN against input orientation. Each scan $\mathbf{S}$ of the hold-out test split of $\mathfrak{D}_B$ was augmented to its 24 possible orientations, for which fBAN predicted the alignment parameters $\widehat{\boldsymbol{p}}$. These parameters were then transformed to match the original orientation of $\mathbf{S}$. For each set of parameters, the Mean Squared Error MSE($\widehat{\boldsymbol{p}}$), Dice Similarity Coefficient DSC($\mathbf{M}^{\widehat{\boldsymbol{p}}}$), and Hausdorff Distance HD($\mathbf{M}^{\widehat{\boldsymbol{p}}}$) were measured against each of the other 23, and the resulting mean, median, and standard deviations were calculated. The table shows the average results for the entire hold-out testing split.

Measure	Mean	Median	Standard deviation
MSE($\widehat{\boldsymbol{p}}$) [10^{-3}] ↓	0.42 ± 2.16	0.14 ± 0.31	0.84 ± 5.51
MSE($\widehat{\boldsymbol{p}}_T$) [$10^{-3}$] ↓	0.02 ± 0.02	0.02 ± 0.02	0.02 ± 0.02
MSE($\widehat{\boldsymbol{p}}_R$) [$10^{-3}$] ↓	0.69 ± 4.08	0.16 ± 0.42	1.57 ± 10.90
MSE($\widehat{\boldsymbol{p}}_S$) [$10^{-3}$] ↓	0.51 ± 1.67	0.22 ± 0.65	0.73 ± 2.36
DSC($\mathbf{M}^{\widehat{\boldsymbol{p}}}$) ↑	0.97 ± 0.01	0.97 ± 0.01	0.01 ± 0.01
HD($\mathbf{M}^{\widehat{\boldsymbol{p}}}$) [voxels] ↓	4.0 ± 1.8	3.7 ± 1.6	1.4 ± 1.0

When separated by GW, the same trend is observed, as shown in Fig. 5.12. For every GW, the mean consistency of fBAN is significantly higher than its average performance. While significantly higher variability in the consistency of MSE($\widehat{\boldsymbol{p}}$) is observed at 14 GW, this does not translate to the other measures, reinforcing the limitations of relying entirely on the accuracy of the predicted parameters to assess the quality of the alignment.

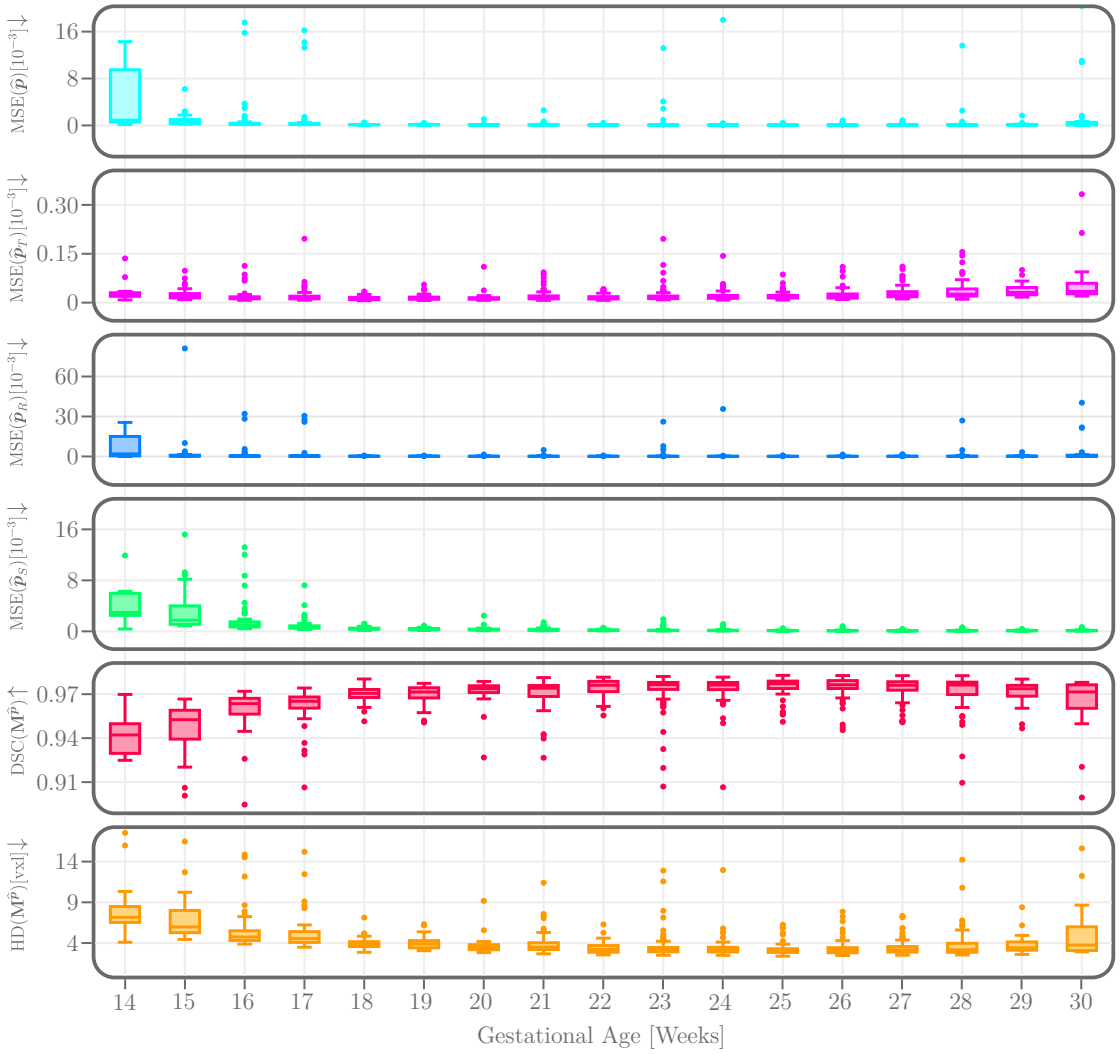

Figure 5.12: Performance consistency of fBAN against input orientation, separated by GW. Each scan $\mathbf{S}$ of the hold-out test split of $\mathfrak{D}_B$ was augmented to its 24 possible orientations, for which fBAN predicted the alignment parameters $\widehat{\boldsymbol{p}}$, which were subsequently transformed to match the original orientation of $\mathbf{S}$. For each set of parameters, the measures were calculated against each of the other 23, and averaged. From top to bottom: Mean Squared Error of predicted parameters MSE($\widehat{\boldsymbol{p}}$), Mean Squared Error of predicted translation parameters MSE($\widehat{\boldsymbol{p}}_T$), Mean Squared Error of predicted rotation parameters MSE($\widehat{\boldsymbol{p}}_R$), Mean Squared Error of predicted scaling parameter MSE($\widehat{\boldsymbol{p}}_S$), Dice Similarity Coefficient of aligned mask DSC($\mathbf{M}^{\widehat{\boldsymbol{p}}}$), and Hausdorff Distance of aligned mask HD($\mathbf{M}^{\widehat{\boldsymbol{p}}}$). The arrows indicate whether a higher (up) or lower (down) value is preferred.

Therefore, since the consistency of fBAN is significantly higher than its state-of-the-art performance, its performance can be considered to be nearly invariant to the orientation of the input scan $\mathbf{S}$.

5.3.5 Spatial landmarks

As a final assessment of the performance of fBAN, the accuracy of the aligned SL introduced in Sec.5.2.5 is analysed, as shown in Fig. 5.13, along with the augmented version. A translation misalignment of the scans is observed in the *Original* positions of the SLs represented by the spread of the central SL (dark grey). The rotation misalignment is represented by the angle between each SL and their goal position, with the centre of the volume as the vertex. Thanks to the colouring of each SL, it can be observed that the canonical space is different from the target space of the clinical acquisition protocol. As mentioned before, this protocol does not take into account the Inferior-Superior, Posterior-Anterior, or Left-Right orientations of the head, which is visualised as two distinct clusters of each SL, mirrored around the centre of the volume. Finally, the scaling misalignment is represented by the axial spread of the SLs.

The predictions generated by the first subnetwork of fBAN (X1) greatly reduce the translation and scaling misalignment. It also fixes the orientations, which results in every SL neatly clustered around their goal position. However, there is still a significant angular spread of each cluster, indicating a still significant rotational misalignment. Additionally, there is a significant number of outliers.

The refinement performed by the second subnetwork of fBAN (X2) results in a significantly smaller spread of each cluster, improving translation, rotation, and scaling misalignment. Additionally, the number of outliers is greatly reduced. However, the Superior (red), Anterior (purple), Inferior (cyan), and Posterior (green) SL still show a significant amount of spread. As mentioned in Sec. 5.3.1, the lack of symmetric landmarks makes it especially difficult to manually correct the rotation misalignment of the sagittal plane in a consistent manner. Since the alignment performed by fBAN is more consistent, as shown in Fig.5.8, the discrepancies result in a rotational spread of these four SLs, around the Left-Right axis.

The network shows a nearly identical performance when performing an orientation augmentation of the input scans. The mirrored clusters are no longer visible in the *Augmented* positions, as every possible orientation of each scan is

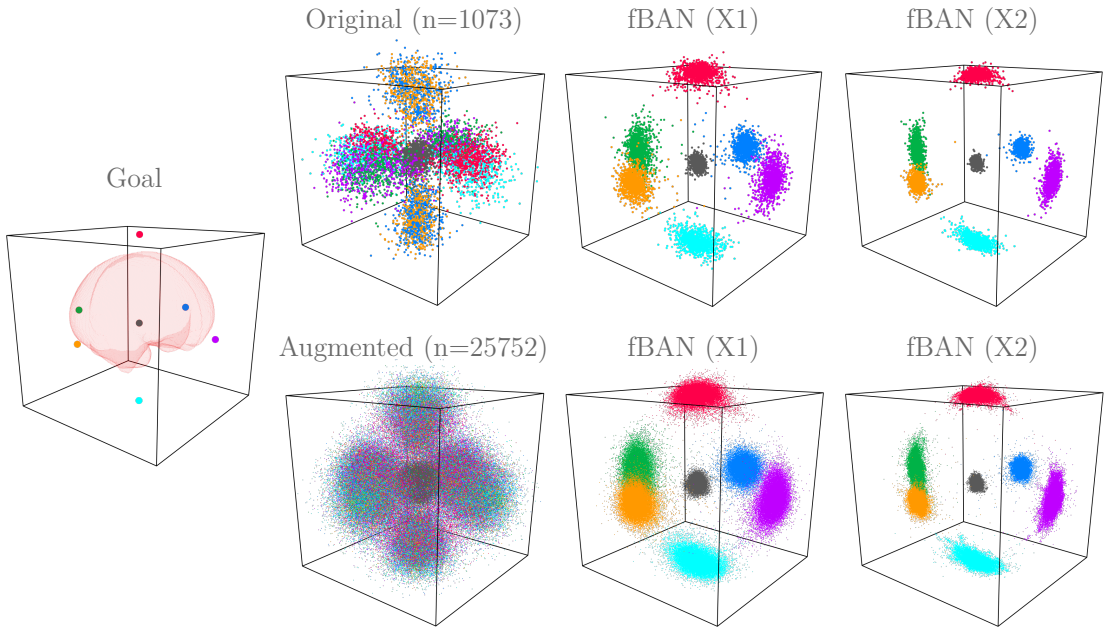

Figure 5.13: Visualization of the performance of fBAN through the use of Spatial Landmarks (SL). Left: SL of the aligned scans, i.e., their goal positions. Top: Original positions in the unaligned scans of the hold-out testing split of $\mathfrak{D}_B$, as well as their aligned positions as predicted by the first (X1) and second (X2) subnetworks of fBAN. Bottom: Equivalent results for scans augmented to represent the 24 possible orientations.

represented. However, after the first subnetwork of fBAN (X1), every orientation is corrected, resulting in clusters that are nearly identical to those obtained without augmentation. Finally, the refinement performed by the second subnetwork fBAN (X2) removes most of the outliers, and results in a similar distribution of SLs as well.

To reinforce these results, the corresponding Spatial Landmarks Loss $\mathfrak{L}_{\text{SLL}}$ of each is shown in Tab. 5.10, along with the Accuracy of Spatial Landmarks $\mathfrak{L}_{\text{ASL}}$ and the Precision of Spatial Landmarks $\mathfrak{L}_{\text{PSL}}$. Due to the mirroring of SLs around the centre due to the orientations of the fetal head in the scan, the mean position of each SL in the *Original* positions is roughly located around the centre of the volume. Therefore, the mean position of the central SL is close to its goal, while the other six are around 80 voxels away from their goal, which results in a $\mathfrak{L}_{\text{ASL}}$ of 69.6 voxels. Thus, the $\mathfrak{L}_{\text{PSL}}$ of 53.56 voxels represents roughly the mean Euclidean distance between each SL and the centre of the volume. In the *Augmented* positions, the situation remains the same, which explains the nearly identical results.

Table 5.10: Accuracy ($\mathfrak{L}_{\text{ASL}}$) and Precision ($\mathfrak{L}_{\text{PSL}}$) of Spatial Landmarks, as well as the Spatial Landmarks Loss $\mathfrak{L}_{\text{SLL}}$, for the Spatial Landmarks in their original positions (unaligned), the positions predicted by the first (X1), and the positions predicted by the second (X2) subnetworks of fBAN. For each, the performance of the augmented dataset is also provided.

Method	$\mathfrak{L}_{\text{ASL}}$ [voxels] ↓	$\mathfrak{L}_{\text{PSL}}$ [voxels] ↓	$\mathfrak{L}_{\text{SLL}}$ [voxels] ↓
Original	69.59	53.56	123.14
Original (augmented)	68.57	55.41	123.99
X1	1.56	11.91	13.47
X1 (augmented)	1.40	12.05	13.45
X2	1.02	7.64	8.66
X2 (augmented)	1.03	7.65	8.67

However, after the first subnetwork of fBAN (X1), the $\mathfrak{L}_{\text{ASL}}$ is reduced to 1.56 voxels, indicating that the mean position of each SL is almost exactly at their goal position. In other words, the alignment space predicted by X1 is nearly identical to the canonical space. The $\mathfrak{L}_{\text{PSL}}$ of 11.91 voxels indicates that the clusters are significantly more compact than before. Finally, while the second subnetwork of fBAN (X2) only results in a minor reduction of $\mathfrak{L}_{\text{ASL}}$ to 1.02, it results in a $\mathfrak{L}_{\text{PSL}}$ of 7.64, greatly reducing the spread of the clusters. The performance of X1 and X2 for the *Augmented* positions is nearly identical, with no statistically significant differences.

As expected, the behaviour of the $\mathfrak{L}_{\text{ASL}}$ and $\mathfrak{L}_{\text{PSL}}$ results are consistent with the performance observed in Fig. 5.13, corroborating the robustness of $\mathfrak{L}_{\text{SLL}}$ as a loss function during training. However, the results in this section also show that the SLs can be very useful to visually understand the alignment performance of a network, facilitating the observation of patterns that were not evident from the other measures.

5.4 Discussion and Conclusions

In this chapter, I have proposed the second module of the Fully-Automated DL Pipeline for 3D Fetal Brain Ultrasound: the fetal Brain Alignment Network (fBAN).

As in the previous chapter, I have shown the process of developing the fBAN, describing its iterative evolution.

I have also developed a new representation of the alignment task through the use of Spatial Landmarks, along with a new loss function $\mathfrak{L}_{\text{SLL}}$ that can rely on these SLs to robustly guide the training of fBAN.

Through the use of Transfer Learning, I have addressed the potential problem of fBAN cheating to solve the alignment task, forcing the network to rely on the structural information of the brain for its predictions.

I addressed the architectural limitations by expanding fBAN into a cascade architecture, and optimise the number of cascade networks, showing that a second subnetwork is enough to achieve the maximum performance.

I have thoroughly analysed the performance of fBAN to fully grasp its strengths and weaknesses, as well as to determine whether the performance goals were achieved. The performance of fBAN is significantly higher than any current alternative method, setting the new state-of-the-art performance for the task of automated fetal brain alignment from 3D US scans.

The network also manages to perform consistently for the entire gestational range of 14.0 to 30.9 weeks, showing only minor variations throughout that range. The performance of fBAN is also robust to the quality of the input scan, as well as virtually invariant to the location, orientation, and size of the brain inside the scan. Additionally, the results show that the alignment predicted by fBAN is more consistent and accurate than my own non-expert manual alignment.

Finally, I demonstrated that the predictions of fBAN are consistent, regardless of the orientation in which the scan is passed, reinforcing its robustness against the orientation of the brain inside the scan, in addition to ensuring the same performance regardless of the orientation in which a particular user stores their scans.

All of this was achieved with manual alignments performed by non-experts. While it would be interesting to compare this approach against more accurate labels generated by experts, the generation of such data is currently unfeasible.

6

Automated Fetal Brain Fingerprinting

Contents

6.1 Introduction

In the previous chapter I discussed the successful development of the first two modules of the Fully-Automated DL Pipeline for 3D Fetal Brain Ultrasound. The

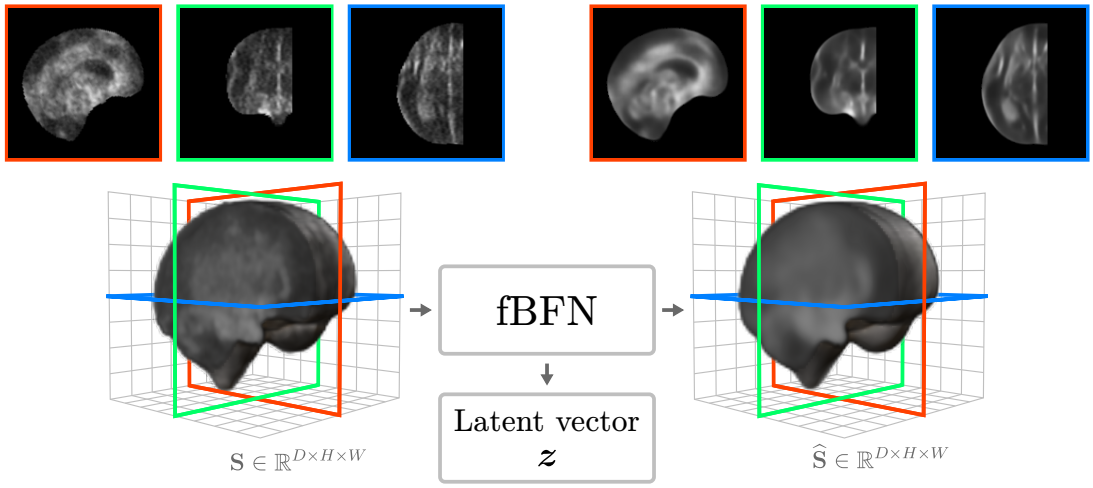

Figure 6.1: Graphical abstract of the fetal Brain Fingerprinting Network (fBFN). The network condenses the structural information of the input scan $\mathbf{S}$ into the latent vector $\boldsymbol{z}$, from which the network generates the reconstructed scan $\widehat{\mathbf{S}}$.

first module (fBEN) focused on the task of extracting the 3D brain from the scans, while the second one (fBAN) focused on aligning it to a canonical space. For the third and final module presented in this chapter, I focus on the challenging task of analysing the 3D fetal neurosonograms.

As mentioned in Sec. 1.1, one of the main challenges limiting the widespread adoption of 3D US for the assessment of in-utero cerebral development lies on the difficulty of analysing the 3D scans. In addition with the typical challenges associated with US as an imaging modality, such as acoustic shadows and artefacts, low contrast, and the high intrinsic variability of the data, extracting the 3D structural context of the brain is very difficult since a clinician can only analyse the data one cross-sectional slice at a time, obfuscating the benefits of this modality over its 2D counterpart.

Other neuroimaging modalities such as MRI have tackled this challenged by developing automated methods to represent the entire 3D brain in a simpler, easier to analyse form. A common approach is to represent the brain by the quantification the volumes of cerebral substructures [149]. However, while this approach makes the analysis easier, the volume of cerebral structures is a crude oversimplification of the brain information contained in the scan. Alternatively, characterising the 3D brain by the morphology of its cerebral structures offers a significantly more detailed representation [152], but still dismisses important information such as the

tissue texture, or their relative intensities. Additionally, both approaches rely on tools such as FreeSurfer[150] for a complete and accurate segmentation of the brain structures in order to capture the correct information. While several works have been published on the topic of segmentation of 3D US scans [35][36][37][38][39][40], many of which show good performance and reliability, such a complete solution for the segmentation of the 3D brain is not currently available for 3D US.

In this chapter I propose the third module of the Fully-Automated DL Pipeline for 3D Fetal Brain Ultrasound: the fetal Brain Fingerprinting Network (fBFN). This network is an end-to-end, general-purpose brain fingerprinting solution that recharacterises the 3D brain by encoding the structural information from the 3D US scan (∼4 million voxels) into a relatively small (500) set of descriptive parameters. As shown in Fig. 6.1, fBFN is based on a VAE architecture [184], and aims to provide a condensed representation of the 3D brain in a continuously distributed latent space, facilitating structural comparisons and subsequent analysis steps.

Through exhaustive analysis, I show that fBFN manages to accurately encode the structural information of the brain into a latent vector, as well as decode this vector into an accurate reconstruction of the original scan. Additionally, I show that fBFN manages to predict the gestational age of the input scan with state-of-the-art accuracy. I demonstrate that fBFN performs consistently for the entire gestational range of 14.0 to 30.9 GW, as well as throughout all regions of the fetal brain. I show that manipulating the gestational age of the latent vector can be used to generate an artificially aged reconstruction of the input scan, which is consistent with the structural information of subsequent scans of the same subject. I show that fBFN can be used to quickly retrieve structurally similar scans with similar performance as relying on the Structural Similarity Index Measure but at a fraction of the computational cost. Finally, I show the potential that fBFN has as a tool to analyse the standard development of structures in the fetal brain throughout gestation.

The contributions in this chapter are:

- Development of fBFN, an end-to-end DL network based on the VAE architecture, that accurately encodes the structural information of the 3D fetal brain into a latent vector in a continuously distributed latent space.
- Demonstration that the gestational age information can be mostly disentangled from the rest of the latent space through the addition of a simple age-prediction task to an information-constrained representation.
- Development of a state-of-the-art solution for the prediction of gestational age from 3D US scans.
- Demonstration thar fBFN can be used to artificially manipulate the gestational age of a scan
- Demonstration that artificially aged scans are structurally consistent with subsequent scans of the same subject
- Demonstration that fBFN can be used as a similarity comparison, achieving a similar performance to SSIM [163] at a fraction of the computational cost.
- Demonstration that fBFN can be used to analyse the structural development of the brain without the need for additional labels.

6.2 Methods

6.2.1 Data

For the initial development of fBFN, dataset $\mathfrak{D}_C$ was used, which is the manually aligned and masked version of dataset $\mathfrak{D}_B$. It consists of a set of $n_C = 4290$ manually aligned and masked scans $\mathbb{S}_C$, and set of corresponding masks $\mathbb{M}_C$. The scans in this dataset have been selected to represent the variability of the INTERGROWTH-21$^{\text{st}}$dataset, containing both high and low quality scans. $\mathfrak{D}_C$ was split into a 1073 hold-out set for testing, and a 3217 set for training and validation, using an iterative stratification approach that evenly distributed the data based on the gestational age of the scans, as well as the structural similarity of the brain.

Subsequently, the task was constrained by using $\mathfrak{D}_D$. For this dataset, the scans (and masks) of $\mathfrak{D}_C$ were mirrored across the sagittal plane, so that the information-rich hemisphere was always on the left. These mirrored scans were subsequently cropped to remove most of the low-information hemispheres.

For the final refinement of fBFN, dataset $\mathfrak{D}_E$ was used. This dataset consists of masked and aligned scans $\mathbb{S}_E$, and set of corresponding masks $\mathbb{M}_E$, generated by the fBEN and fBAN networks, which were subsequently mirrored and cropped in the same manner as $\mathfrak{D}_D$. $\mathfrak{D}_E$ consists of the 13779 scans available from the INTERGROWTH-21$^{\text{st}}$dataset, split into a 3445 hold-out set for testing, and a 10334 set for training and validation, using a stratification approach that evenly distributed the data based on the gestational age of the scans.

Both datasets span the gestational age range of 14.1 to 30.9 GW (99 to 216 gestational days) and the features of their respective scans were normalised to within 0 and 1.

A more detailed description of the data used in this thesis can be found in Chapter 3.

6.2.2 Implementation details

The development of fBFN was implemented in Python using the Pytorch [167] library. The networks were developed on an Intel Xeon E-2146G CPU (3.50GHz, 6 cores) and an Nvidia GTX 1080 Ti, as well as an Nvidia A10. All fBFN networks were trained using a 3-fold cross-validation, and tested against a hold-out dataset.

6.2.3 Evaluation measures

In this chapter I will cover the development of fBFN, which is based on a VAE architecture [184]. As the schematics in Fig. 6.2 show, this network architecture consists of two subsequent sub-networks: an Encoder and a Decoder. The Encoder receives the input scan $\mathbf{S}$ and condenses its features into a mean vector $\boldsymbol{\mu}$ and a log-variance vector $\boldsymbol{\varphi} = ln(\boldsymbol{\sigma^2})$, which describe the (ideally standard normal) distribution of the latent space generated by the network. These vectors are then

used to resample a latent vector $\boldsymbol{z}$, which is subsequently passed as input to the Decoder, which then generates a predicted reconstruction $\widehat{\mathbf{S}}$ of the original scan.

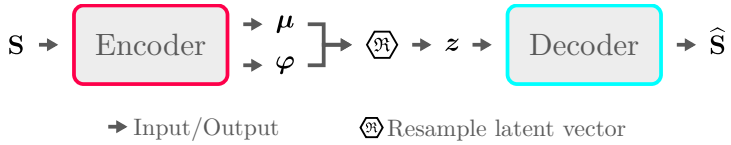

Figure 6.2: Schematics of the basic architecture of a VAE network.

Later in development (Sec. 6.2.6) I will modify the network to specifically require it to predict the gestational age $d_{\mathbf{S}}$ of the scan $\mathbf{S}$ as the first parameter z_0 of the latent vector $\boldsymbol{z}$.

Therefore, in order to evaluate the performance of fBFN, three main evaluation measures are needed: one to evaluate the accuracy of the reconstructed scan, one to evaluate the latent space distribution against the standard normal distribution, and one to evaluate the accuracy of the predicted gestational age.

To evaluate the performance of the reconstruction $\widehat{\mathbf{S}}$, I will rely on the Structural Similarity Index Measure (SSIM) [163]. While there is a plethora of alternative image quality assessment methods, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Feature Similarity Index (FSIM) [185], I have chosen to rely on SSIM for two main reasons. Firstly, rather than measuring absolute errors, its perception-based approach combines multiple perceptual characteristics of the images (luminance, contrast, and structure), which results in a robust, accurate assessments that generally outperform most alternative methods [186][187][188][189]. Secondly, its output values are normalised to between 0 and 1, which makes the results easier to interpret, as well as facilitating the balancing of the combined loss functions during training. While FSIM also shares these two characteristics, I opted for SSIM due to its ubiquity in medical imaging. To calculate the SSIM between two 3D volumes $\boldsymbol{x}$ and $\boldsymbol{y}$, the means μ_x and μ_y, the standard deviations σ_x and σ_y, and the covariance σ_{xy} are calculated. With these values, the luminance similarity $\mathrm{l}(\boldsymbol{x}, \boldsymbol{y})$, the contrast similarity $\mathrm{c}(\boldsymbol{x}, \boldsymbol{y})$, and the structure similarity $\mathrm{s}(\boldsymbol{x}, \boldsymbol{y})$ between both volumes can be calculated, as shown in

Equations (6.1), (6.2), and (6.3), where $c_1 = 0.0001$, $c_2 = 0.0003$, and $c_3 = 0.00015$ are empirically chosen constants used to stabilise the division by a weak denominator.

$$\mathrm{l}(\boldsymbol{x}, \boldsymbol{y}) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \tag{6.1}$$

$$\mathrm{c}(\boldsymbol{x}, \boldsymbol{y}) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \tag{6.2}$$

$$\mathrm{s}(\boldsymbol{x}, \boldsymbol{y}) = \frac{2\sigma_{xy} + c_3}{\sigma_x^2 + \sigma_y^2 + c_3} \tag{6.3}$$

The $\mathrm{SSIM}(\boldsymbol{x}, \boldsymbol{y})$ is comprised of the product of these three similarities, as shown in Eq. (6.4).

$$\mathrm{SSIM}(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{l}(\boldsymbol{x}, \boldsymbol{y}) \cdot \mathrm{c}(\boldsymbol{x}, \boldsymbol{y}) \cdot \mathrm{s}(\boldsymbol{x}, \boldsymbol{y}) \tag{6.4}$$

For this particular implementation, a sliding-box approach with a box of size (9,9,9) was used, weighted by a box-centred normal distribution with a standard deviation of 0.5 in all axes. This yields a 3D volume, where the value of each voxel represents the SSIM of the sliding-box around the equivalent voxel in the original scan $\mathbf{S}$ and the reconstructed scan $\widehat{\mathbf{S}}$. To ignore any extra-cerebral voxels, only the values that lie within the brain mask $\mathbf{M}$ are averaged to obtain the value of the reconstruction measure $\mathrm{SSIM}(\mathbf{S}, \widehat{\mathbf{S}}, \mathbf{M})$.

To assess the distribution of the latent space I rely on the Kullback-Leibler Divergence (KLD) [190][191]. The KLD, also known as the relative entropy, is the distance measure between two probabilistic distributions. For two distributions P and Q over the same sample space $\mathcal{X}$, the KLD is defined as shown in Eq. (6.5).

$$\mathrm{KLD}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \ln\left(\frac{P(x)}{Q(x)}\right) \tag{6.5}$$

For two multivariative normal distributions $P = \mathcal{N}(\mu_1, \Sigma_1)$ and $Q = \mathcal{N}(\mu_2, \Sigma_2)$ of dimension $n_{\boldsymbol{z}}$, with μ_i and Σ_i being the respective mean and covariance matrices, the equation can be derived as shown in Eq. (6.6) .

$$\mathrm{KLD}(P||Q) = \frac{1}{2}\left[\ln\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) - n_{\boldsymbol{z}} + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T\Sigma_2^{-1}(\mu_2 - \mu_1)\right] \tag{6.6}$$

In the particular case of VAEs, $P = \mathcal{N}(\mu, \Sigma)$, where $\Sigma = diag(\sigma_1^2, ..., \sigma_{n_z}^2)$, and $Q = \mathcal{N}(0, 1)$, which yields the equation shown in Eq. (6.7).

$$\mathrm{KLD}(P||Q) = \frac{1}{2}\left[-\sum_{i}^{n_z}(\ln(\sigma_i^2) + 1) + \sum_{i}^{n_z}\sigma_i^2 + \sum_{i}^{n_z}\mu_i^2\right] \tag{6.7}$$

However, as shown in Fig. 6.2, the Encoder predicts the mean μ and the log-variance $\boldsymbol{\varphi} = ln(\boldsymbol{\sigma^2})$, resulting in Eq. (6.8)

$$\mathrm{KLD}(\boldsymbol{\mu}, \boldsymbol{\varphi}) = \frac{1}{2}\left[-\sum_{i}^{n_z}(\boldsymbol{\varphi}_i + 1) + \sum_{i}^{n_z}\exp(\boldsymbol{\varphi}_i) + \sum_{i}^{n_z}\mu_i^2\right] \tag{6.8}$$

As a final modification, I add a factor $1/n_z$ to this equation, which facilitates the performance comparison between distributions of different dimensionalities. This results in the final equation of the $\mathrm{KLD}(\boldsymbol{\mu}, \boldsymbol{\varphi})$ as shown in Eq.(6.9).

$$\mathrm{KLD}(\boldsymbol{\mu}, \boldsymbol{\varphi}) = \frac{1}{2n_z}\left[-\sum_{i}^{n_z}(\boldsymbol{\varphi}_i + 1) + \sum_{i}^{n_z}\exp(\boldsymbol{\varphi}_i) + \sum_{i}^{n_z}\mu_i^2\right] \tag{6.9}$$

Finally, to measure the accuracy of the predicted gestational age I rely on $\mathrm{MAE}(d_{\mathbf{S}}, z_0)$ as defined in Eq. (6.10), with $d_{\mathbf{S}}$ being the standardised gestational age in days and z_0 the first parameter of the resampled latent vector $\boldsymbol{z}$.

$$\mathrm{MAE}(d_{\mathbf{S}}, z_0) = |d_{\mathbf{S}} - z_0| \tag{6.10}$$

6.2.4 Initial development

The first step of developing fBFN was to adapt the basic VAE architecture into a network that would perform well with the high intrinsic variability of the 3D US data. Initially, rather than developing the network from the ground up, I opted to adapt the architectures of fBEN and fBAN into a VAE. However, the large number of variables in this un-optimised network proved too difficult to refine into a working network. Therefore, in order to facilitate this task, I first focused on developing and optimising an AutoEncoder (AE), which can later be adapt into a VAE. This allowed me to initially focus entirely on the reconstruction performance of the network, without having to address the properties of the latent space. As an additional constrain, I reduced the spatial resolution of the $\mathfrak{D}_C$ dataset from 0.6

mm/vxl to 1.2 mm/vxl, which reduces the effective noise-like speckle variability of the scans, allowing the network to focus on the stronger structural landmarks.

In a similar manner to the initial development of the previous networks, I performed an iterative optimisation of the AE, in which I compared the performance of different architectures, activation layers (Tanh, ReLU, LeakyReLU, Sigmoid), normalisation layers (Batch vs Instance normalisation), number of hidden dimensions, and network depth. For the dimensions of the latent vector $\boldsymbol{z}$, I used an empirically selected value of 1000. The reconstruction loss function was $\mathfrak{L}_{\text{SSIM}}$, as shown in Eq. (6.11)

$$\mathfrak{L}_{\text{SSIM}} = 1 - \text{SSIM}(\mathbf{S}, \widehat{\mathbf{S}}, \mathbf{M}) \tag{6.11}$$

All networks were trained for 200 epochs keeping the trained weights of the best validation performance, and using the Adam optimiser with a learning rate of 0.0005.

The schematics of the final optimised AE network are shown in Fig. 6.3. The Encoder section of the network consists of five subsequent convolutional blocks, each containing a convolutional layer, a batch normalisation layer, and a LeakyReLU layer, followed by a fully connected layer. Rather than relying on pooling for down-sampling, each convolutional layer has a stride size of 2. The Decoder section is essentially a mirrored version of the Encoder with transposed convolutional layers and a final convolutional block consisting of a convolutional layer, a batch normalisation layer, a LeakyReLU activation layer, a second convolutional layer, and a Sigmoid activation layer. The transposed convolutional layers also have a stride of 2 to achieve an effective upsampling without the use of interpolation. The hidden dimensions of the Encoder were (32,64,128,256,512,1000), while for the Decoder they were (1000,512,256,128,64,32,32).

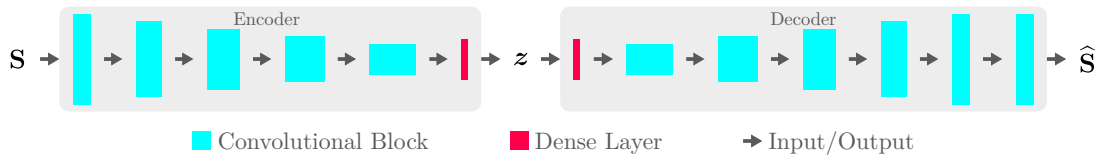

Figure 6.3: Schematics of the optimised AE network.

Figure 6.4 shows an example scan $\mathbf{S}$ at 25 GW, along with the predicted reconstruction $\widehat{\mathbf{S}}$ generated by the optimised AE trained with the $\mathfrak{L}_{\text{SSIM}}$. Additionally, the reconstruction of the same scan generated by an identical AE trained with the loss $\mathfrak{L}_{\text{MAE}} = \text{MAE}(\mathbf{S}, \widehat{\mathbf{S}})$ has been included. The prediction $\widehat{\mathbf{S}}$ generated by the AE resembles the structural qualities of the original scan $\mathbf{S}$. While most of the fine details are lost in the process, the bulk of the structural information is clearly maintained. When comparing the two loss functions, it can be observed that the $\mathfrak{L}_{\text{SSIM}}$ results in a significantly sharper, and more accurate reconstruction. This is particularly evident when observing the CSP in the sagittal view. Additionally, the reconstructions of the network trained with $\mathfrak{L}_{\text{MAE}}$ exhibit a grain-like texture. Therefore, the development continued with $\mathfrak{L}_{\text{SSIM}}$ as the reconstruction loss.

Now that I had an optimised AE, I continue the development of fBFN by reintroducing the resampling step into the network, resulting in the VAE architecture shown in Fig. 6.5. For the loss function, the combination shown in Eq. (6.12) was used, with α being a weighting constant.

$$\mathfrak{L} = \alpha\mathfrak{L}_{\text{KLD}} + \mathfrak{L}_{\text{SSIM}} \tag{6.12}$$

I optimised fBFN using the same iterative approach as before, with the addition of comparing multiple values for the weighting constant α. Aside for the resampling step, the architecture of fBFN remained largely the same as that of the optimised AE, with the only differences being a higher number of hidden dimensions (64,128,256,512,1024,1000,1024,512,256,128,64), and a final convolutional block that contained a single convolutional layer and a Sigmoid activation layer, as this resulted in the best overall performance.

The value of the weighting constant α proved to be critical for achieving the optimal reconstructions. The optimal value for fBFN was $\alpha = 0.05$, but this will change depending on the reconstruction function used. If α is too low, the network effectively performs similarly to the AE, since the distribution of the latent space is weakly affected by the training. On the other hand, if α is too high, the network prioritises shaping the distribution to a standard normal distribution

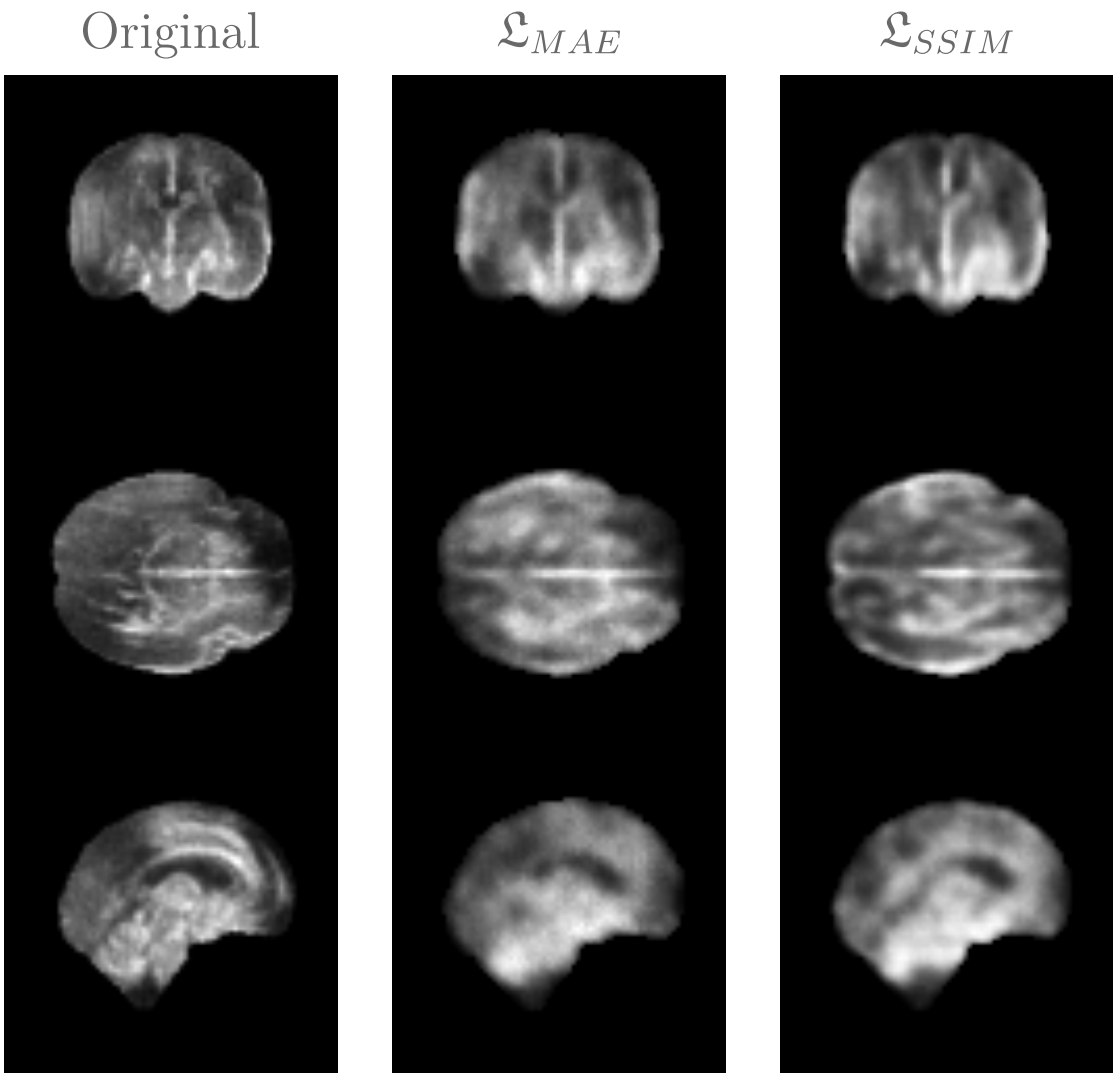

Figure 6.4: Example reconstruction $\widehat{\mathbf{S}}$ of a 25 GW scan $\mathbf{S}$ (left) of an optimised AE trained using $\mathfrak{L}_{\text{SSIM}}$ (middle) and $\mathfrak{L}_{\text{MAE}}$ (right) as the reconstruction loss.

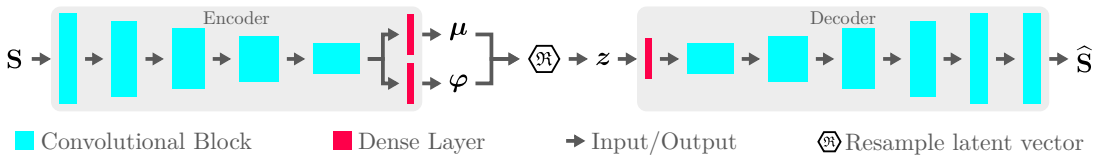

Figure 6.5: Schematics of the fBFN architecture.

$\mathcal{N}$ over accurate reconstructions. An example of this effect is shown in Fig. 6.6, where the same example scan as in Fig. 6.4 has been reconstructed by a fBFN with two different α values. fBFN trained with the optimal value $\alpha = 0.05$ results in a significantly more accurate reconstruction than what was achieved with the optimised AE. In particular, the asymmetry of the structural information discussed in Sec. 1.1 is better maintained, and the reconstruction is overall sharper. In contrast, $\alpha = 0.25$ results in reconstruction that is more symmetrical, losing the individual structural characteristics of the input scan.

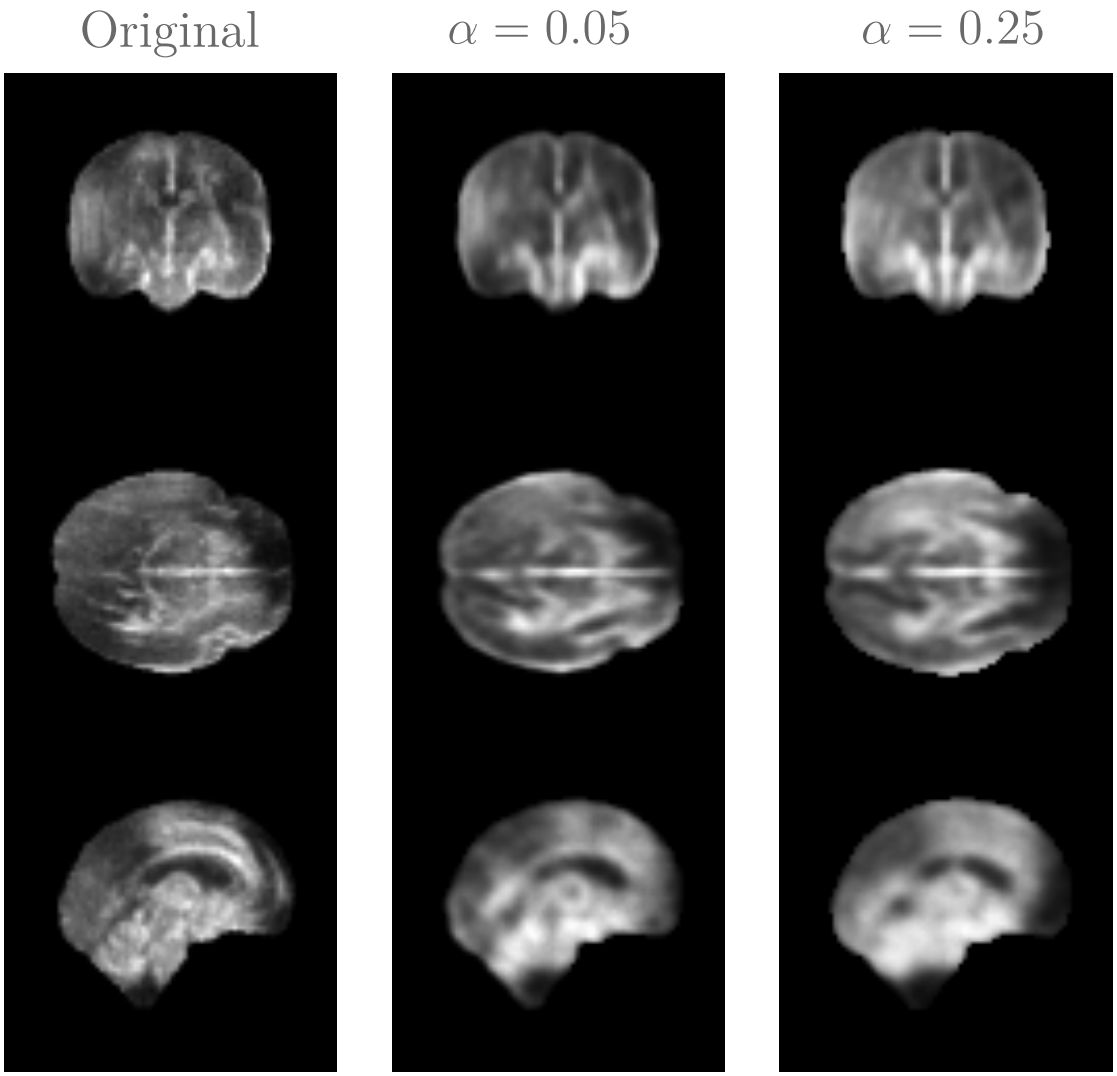

Figure 6.6: Example reconstructions $\widehat{\mathbf{S}}$ of a 25 GW scan $\mathbf{S}$ (left) achieved with the same network, but with a different α weight for the loss function shown in Eq. 6.12.

As a final step in the initial development, I focused on optimising the number

of latent dimensions. Ideally, fBFN would manage to condense the structural information of the 3D brain into the smallest number of parameters possible. However, reducing the number of parameters too much could result in an significant loss of structural information. Therefore, I iteratively reduced the number of latent dimensions until a drop in performance was observed. As Fig. 6.7 shows, the performance of the network was barely affected by the reduction of latent dimensions until 500 parameters were reached. There was a significant drop in performance across all measures for 400, a trend that continued for lower numbers. The conclusion is that 500 parameters represents the minimum amount of parameters that fBFN needs to store the structural information of the 3D brain without a significant loss in information.

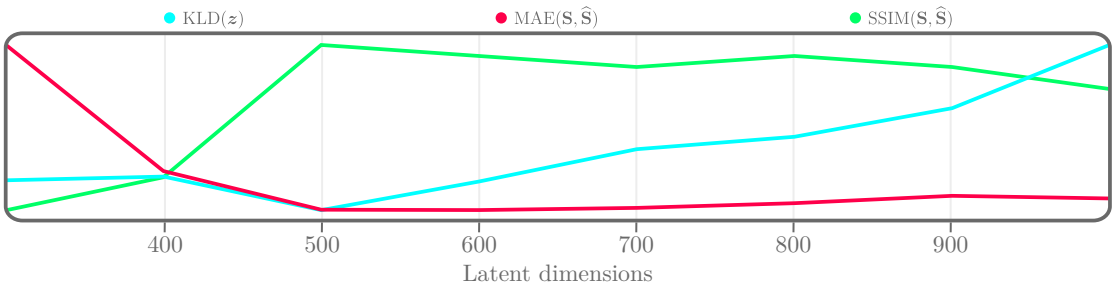

Figure 6.7: Performance comparison of the same fBFN trained with different numbers of latent dimensions. The three plots have been normalised for easier visualisation.

6.2.5 Task constraint

At this stage of development I had developed a functional fBFN that managed to condense most of the structural information of the 3D brain into 500 parameter. However, as the example shown in Fig. 6.6 highlights, while the reconstructions $\widehat{\mathbf{S}}$ manage to capture the majority of the structural characteristics of the input scan $\mathbf{S}$, there is still a significant amount of finer detail that is lost. This is likely a result of the high intrinsic variability of the scans, which makes it challenging for the network to focus on the important structures. For example, the asymmetry of the scans discussed in Sec. 1.1 results in a hemisphere that contains significantly less structural information. In spite of this, that same hemisphere has the equal impact on the reconstruction loss as the one containing most of the structural information. As a

result, the network focuses resources to accurately reconstruct regions with little useful information. A similar effect will be caused by the noise-like characteristics of speckle, which fBFN will attempt to accurately reconstruct at the cost of resources.

To minimise these issues, and help fBFN focus on the most relevant structural information found on the 3D scan, I constrained the task by using dataset $\mathfrak{D}_D$ (see Sec. 6.2.1), effectively removing the low-information hemisphere entirely. Additionally, in order to address the challenges associated with speckle, the reconstruction loss $\mathfrak{L}_{\text{SSIM}}$ of the predicted scan $\widehat{\mathbf{S}}$ is calculated against an anisotropically filtered version $\mathbf{S}^*$ of the input scan $\mathbf{S}$.

Constraining the task results in a significant improvement in the accuracy of the reconstructions. This can be clearly see in Fig. 6.8, which shows the reconstructions of the same 25 GW example as before, reconstructed with and without the constrained task. In addition to a more accurate reconstruction, the constrained task results in significantly sharper structure. Additionally, the intensity distribution of the original scan is better preserved.

It is important to mention that while this task constrain results in better, more accurate reconstructions, it comes at the cost of the loss of any relevant structural information that might have been available in the low-information hemisphere of the imaged 3D brain, as well as information that might be lost during the anisotropic filtering of the input scan. Since the goal of this particular chapter is mostly exploratory in nature, this is not a significant concern. However, a more robust solution that avoids this information loss would be beneficial in the future.

6.2.6 Soft age disentanglement

So far I had focused mainly on the reconstruction performance of fBFN. After all, a higher accuracy of the reconstructed scan $\widehat{\mathbf{S}}$ confirms a higher amount of structural information condensed into the latent vector $\boldsymbol{z}$. However, analysing the latent space is also of great importance, since a crucial advantage of the VAE is the manner in which the latent space is distributed.

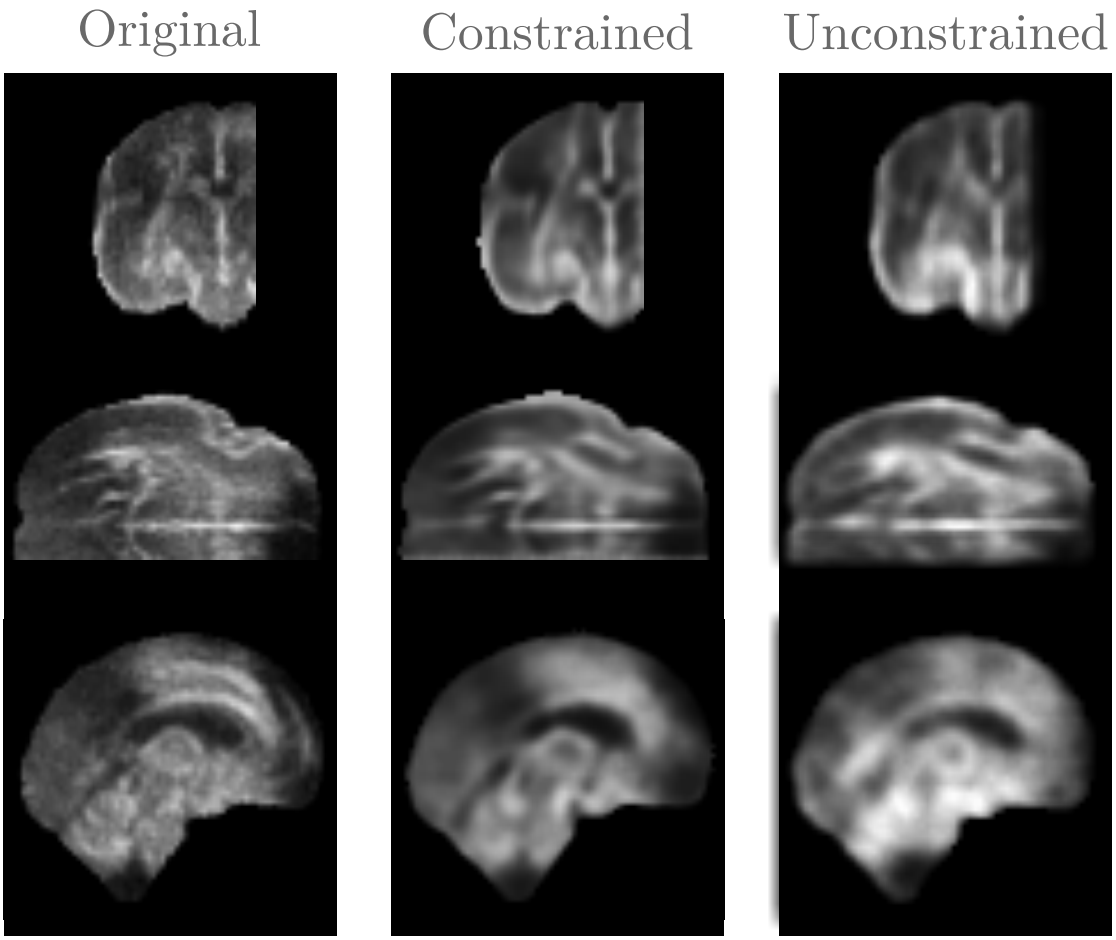

Figure 6.8: Example reconstructions $\widehat{\mathbf{S}}$ of a 25 GW scan $\mathbf{S}$ performed by a network trained with a constrained and unconstrained dataset. The constrained dataset allows the network to focus on the structurally relevant regions, resulting in better reconstruction accurate. Note that the unconstrained example has been cropped and mirrored after reconstruction to facilitate comparisons.

Figure 6.9 shows a 2D projection of a Principal Component Analysis (PCA) of the latent vectors of the testing split of dataset $\mathfrak{D}_D$. By colouring each point based on the GW of the input scan $\mathbf{S}$, it is immediately apparent that fBFN has managed to implicitly learn a linear age distribution in its latent space. This is ideal, as it means that in addition to learning to encode the structural information of the 3D brain, fBFN has learnt that the changes of structural information follow an age-dependent development and has shaped its latent space accordingly.

However, without additional guidance during training, the representation of gestational age in latent space is entangled in latent space. This can be observed in Fig 6.10, where the absolute value of the Pearson correlation coefficient $|s|$ of

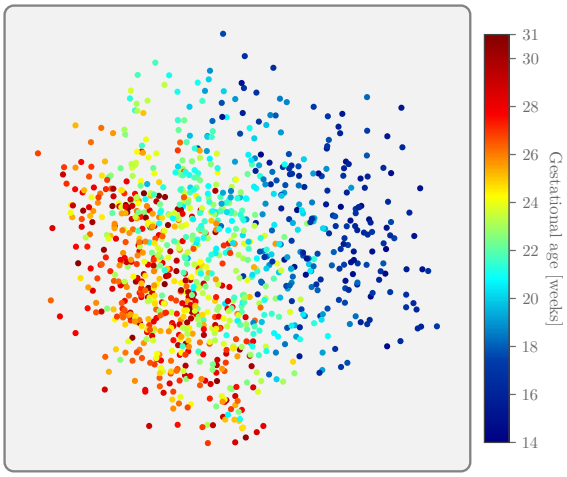

Figure 6.9: 2D projection of a PCA of the latent vectors $\boldsymbol{z}$ encoded by fBFN.

each parameter and the GW is displayed in cyan. The results show that there are 5 parameters showing a moderate degree of correlation with gestational age ($0.3 < |s| < 0.5$), and one parameter with a strong correlation ($0.5 < |s|$).

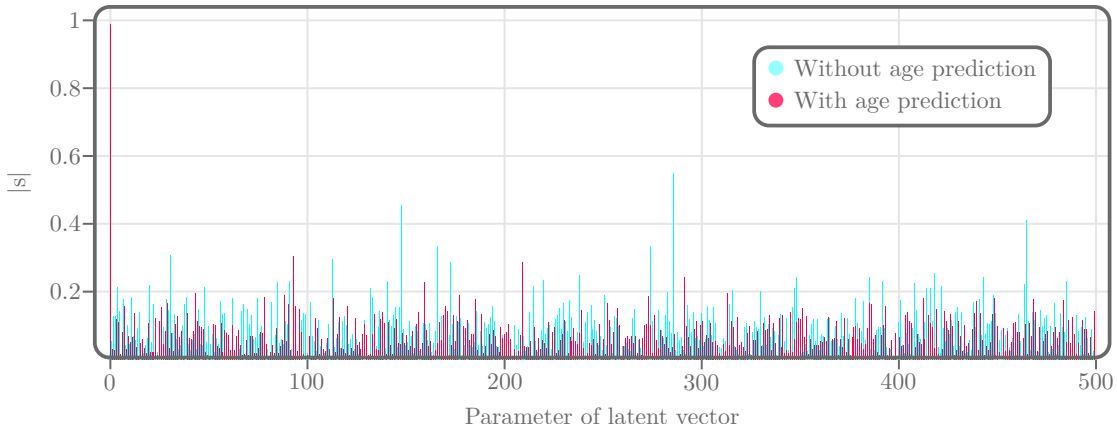

Figure 6.10: Absolute value of the Pearson correlation coefficient s between each of the 500 parameters of the encoded vectors $\boldsymbol{z}$ and the gestational age (in days) of their corresponding scan $\mathbf{S}$.

In order to encourage fBFN to disentangle the gestational age in latent space, I have added the task of predicting the gestational age in the first parameter of the predicted latent vector $\boldsymbol{z}$. Since the results showed that 500 parameters is the minimum number of parameters needed to condense the structural information of the scan, I expected this approach to force fBFN to encode most of the gestational age representation in this first parameter, minimising the age correlation of the

remaining parameters. The corresponding loss function to train fBFN is defined in Eq. (6.13), where $\beta = 1$ is a weighting constant, $d_{\mathbf{S}}$ is the standardised gestational age of scan $\mathbf{S}$, and $\mathbf{S}^*$ is its anisotropically filtered version.

$$\mathfrak{L} = \alpha\mathfrak{L}_{\text{KLD}}(\boldsymbol{\mu}, \boldsymbol{\varphi}) + \beta\mathfrak{L}_{\text{MAE}}(d_{\mathbf{S}}, z_0) + \mathfrak{L}_{\text{SSIM}}(\mathbf{S}^*, \widehat{\mathbf{S}}, \mathbf{M}) \tag{6.13}$$

As Fig. 6.10 shows (coloured in crimson), adding the task of predicting the gestational age in the first latent parameter results in a nearly perfect correlation with a absolute Pearson correlation coefficient of $|s| = 0.99$. Additionally, the correlation scores of the remaining 499 parameters is significantly reduced, with no parameter achieving an absolute coefficient higher than 0.3.

Figure 6.11 shows that fBFN still creates a linear distribution of gestational age in latent space. However, such a distribution is not found in the 2D projection of the PCA of the remaining 499 parameters, further reinforcing the notion that most of the gestational age information has been disentangled to the first parameter.

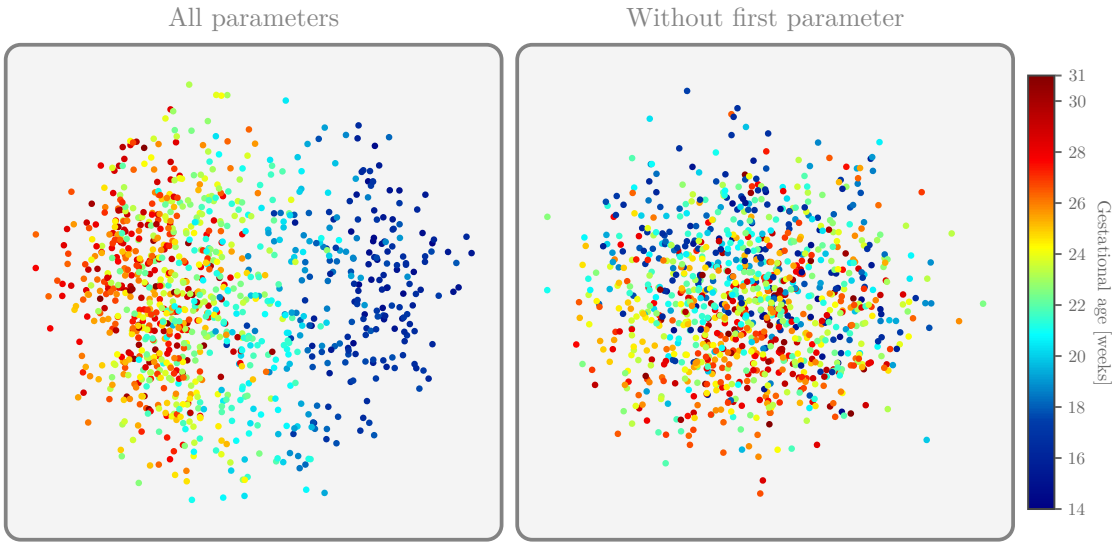

Figure 6.11: 2D projection of a PCA of the latent vectors $\boldsymbol{z}$ encoded by fBFN. Left: PCA of all 500 parameters. Right: PCA of 499 parameters, excluding the first parameter z_0 where the gestational age is encoded.

In addition to the soft disentanglement of the latent space, this method resulted in an mean absolute error of 3.36 ± 3.12 days of the predicted gestational age. However, no statistically significant impact was observed in the accuracy of the reconstructed scans.

6.2.7 Final refinement

In a similar manner to the development of fBEN and fBAN, the final step in the development of fBFN was to perform a final refinement. So far the entirety of the development has been performed under the constraint of using low-resolution versions of dataset $\mathfrak{D}_D$, which reduces computational cost in addition to smoothing out high-frequency intensity variabilities, such as those resulting from speckle. However, to maximise the sharpness of the reconstructions, the first part of the final refinement consisted of removing this constraint in favour of the full-resolution 0.6 mm/vxl voxel spacing.

The second part focused on expanding the available data. The development of fBFN has thus far been performed with scans and masks from the manually aligned and masked dataset $\mathfrak{D}_C$, with $\mathfrak{D}_D$ being a further processing of the same data. However, unlike during the training of the previous networks, data augmentations through similarity transforms are not possible, since only using one half of the 3D brain is being used. This results in a total of only 4290 scans, which is a significantly limiting factor considering the large amount of structural variability present in the data. The manual alignment of more scans would be time overly consuming. Thankfully, the high performance of the previous networks allowed for their use to automatically align and mask the entirety of the available INTERGROWTH-21$^{\text{st}}$dataset, which consisted of 13779 scans spanning the gestational range of 14.0 to 30.9 GW. This new dataset $\mathfrak{D}_E$ was split using a stratification approach that evenly distributed the data based on the gestational age of the scans, resulting in 10334 training scans, and a 3445 hold-out set for testing.

After a final optimisation, I performed a final training of fBFN for 200 epochs, with the loss function shown in Eq. (6.14), using the AdamW optimiser with a learning rate of $lr = 0.0001$ and a weight decay of $wd = 0.0001$.

$$\mathfrak{L} = 0.00025 \cdot \mathfrak{L}_{\text{KLD}}(\boldsymbol{\mu}, \boldsymbol{\varphi}) + 1 \cdot \mathfrak{L}_{\text{MAE}}(d_{\mathbf{S}}, z_0) + 1 \cdot \mathfrak{L}_{\text{SSIM}}(\mathbf{S}^*, \widehat{\mathbf{S}}, \mathbf{M}) \tag{6.14}$$

An exhaustive analysis of the performance of fBFN can be found in Sec. 6.3.

6.3 Results

In this section I perform an exhaustive analysis of the performance of fBFN. I evaluate its mean performance, its performance by gestational age, and its regional performance. I compare the mean of scans $\mathbf{S}$ against the mean of reconstructions $\widehat{\mathbf{S}}$, as well as the reconstruction of the mean latent vector $\boldsymbol{z}$. Finally, I analyse the latent space distribution, as well as the correlation between parameters and gestational age.

6.3.1 Mean performance

Table 6.1: Average performance of fBFN evaluated with the hold-out testing split of two datasets: $\mathfrak{D}_D$ and $\mathfrak{D}_E$. Note that while the scans of the testing split of $\mathfrak{D}_D$ are a subset of the testing split of $\mathfrak{D}_E$, the former has been manually aligned while the later has been automatically generated with fBEN and fBAN. The arrows indicate whether a higher (up) or lower (down) value is preferred.

Dataset	SSIM$(\mathbf{S}, \widehat{\mathbf{S}}, \mathbf{M})$ ↑	SSIM$(\mathbf{S}^*, \widehat{\mathbf{S}}, \mathbf{M})$ ↑	MAE$(d_{\mathbf{S}}, z_0)$ ↓	KLD$(\boldsymbol{\mu}, \boldsymbol{\varphi})$ ↓
$\mathfrak{D}_D$	0.896 ± 0.046	0.927 ± 0.030	2.29 ± 2.62	0.041 ± 0.015
$\mathfrak{D}_E$	0.858 ± 0.055	0.893 ± 0.053	4.22 ± 4.42	0.045 ± 0.023

In order to assess the mean performance of fBFN, the evaluation performed in this subsection will be performed using the hold-out testing splits from datasets $\mathfrak{D}_D$ and $\mathfrak{D}_E$. While the network was trained using $\mathfrak{D}_E$, this dataset was generated entirely using the predictions of fBEN and fBAN. Therefore, in addition to the larger, more challenging dataset, the performance of fBFN is likely to also be affected by the performance of the other two networks. In contrast, Dataset $\mathfrak{D}_D$ was generated entirely through manual alignment, which ensures a level of quality and consistency. Please note that raw scans used to generate the hold-out split of $\mathfrak{D}_D$ is a subset of the raw scans used to generate the hold-out split of $\mathfrak{D}_E$, ensuring that fBFN has not encounter them before. Given the high performance and consistency shown in Sections 4.3 and 5.3, I expect only small differences in the performance of fBFN. After confirming this, I will continue the analysis of fBFN in the following subsections relying only on the hold-out split of $\mathfrak{D}_E$.

The results shown in Tab 6.1 shown that the predictions $\widehat{\mathbf{S}}$ generated by fBFN are structurally consistent with the original scans $\mathbf{S}$ of $\mathfrak{D}_D$, with a mean SSIM$(\mathbf{S}, \widehat{\mathbf{S}}, \mathbf{M})$ of 0.896 out of a maximum of 1. This performance is even higher if calculated against the anisotropically filtered scan $\mathbf{S}^*$, where it achieves a value of 0.927. For this dataset, fBFN also manages to achieve a state-of-the-art performance for the prediction of the gestational age of the input scan $\mathbf{S}$, with a mean absolute error of 2.29 days.

As predicted, a slight performance drop for fBFN is observed when evaluated against $\mathfrak{D}_D$, showing a 4.2% drop of performance for SSIM$(\mathbf{S}, \widehat{\mathbf{S}}, \mathbf{M})$, and 3.7% when calculated against the anisotropically filtered scan $\mathbf{S}^*$. While these drops are statistically significant, the performance is still high. A more significant drop is observed when assessing the predicted gestational age, when 84.3% lower performance. However, at an mean absolute error of 4.22 days, the performance still achieves state-of-the-art performance.

6.3.2 Performance vs. gestational week

Figure 6.12 shows the performance of fBFN separated by GW, for datasets $\mathfrak{D}_D$ and $\mathfrak{D}_E$. The performance of fBFN is consistent between datasets, with only a minor decrease in performance for $\mathfrak{D}_E$, which is consistent with the mean results shown in Tab. 6.1. The network shows virtually no performance correlation with gestational age for SSIM$(\mathbf{S}, \widehat{\mathbf{S}}, \mathbf{M})$, SSIM$(\mathbf{S}^*, \widehat{\mathbf{S}}, \mathbf{M})$, and MAE$(d_{\mathbf{S}}, z_0)$, all of which have a Pearson correlation coefficient with an absolute value of $s = 0.09$ or smaller. The KLD$\boldsymbol{z}$ does exhibit a stronger correlation with gestational age, with a slight drop in performance for the GW 14 and 15. However, with a coefficient of $s_{\mathfrak{D}_D} = -0.26$ and $s_{\mathfrak{D}_E} = -0.13$, the results only show a weak negative correlation.

Representative examples of the reconstructed scans $\widehat{\mathbf{S}}$ of several GWs are shown in Fig. 6.13. These examples have been chosen because their SSIM$(\mathbf{S}, \widehat{\mathbf{S}}, \mathbf{M})$ performance is the closest to the mean performance of fBFN for the corresponding GW. They confirm that the reconstructions $\widehat{\mathbf{S}}$ predicted by the network are structurally consistent with the input scan, across the entire gestational range.

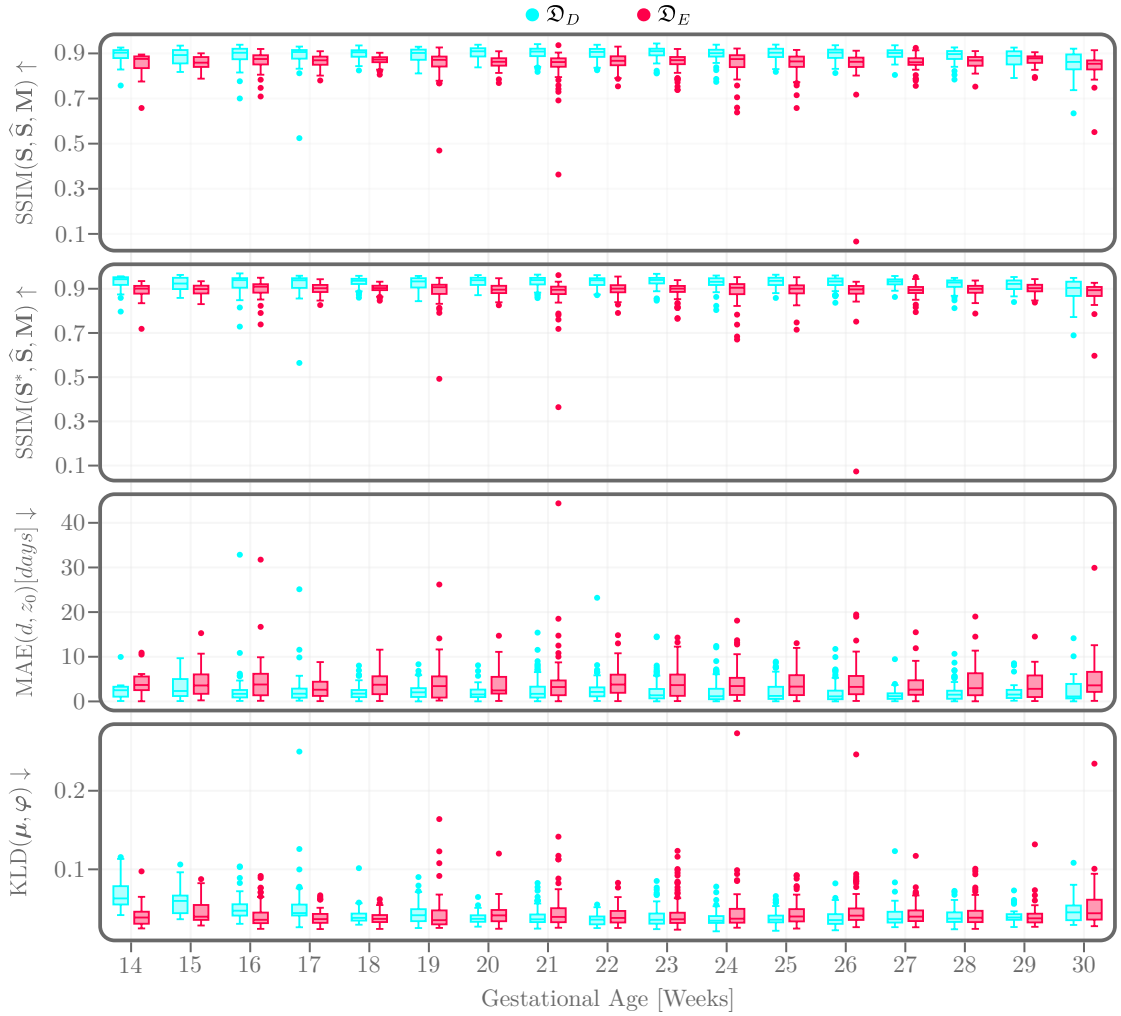

Figure 6.12: Performance measures of fBFN evaluated using the hold-out split of datasets $\mathfrak{D}_D$ and $\mathfrak{D}_E$, separated by GW. From top to bottom: Structural Similarity Index Measure SSIM$(\mathbf{S}, \widehat{\mathbf{S}}, \mathbf{M})$ between reconstruction $\widehat{\mathbf{S}}$ and the original scan $\mathbf{S}$, SSIM$(\mathbf{S}^*, \widehat{\mathbf{S}}, \mathbf{M})$ between reconstruction $\widehat{\mathbf{S}}$ and the anisotropically filtered original scan $\mathbf{S}^*$, Mean Absolute Error MAE$(d_{\mathbf{S}}, z_0)$ between the reported gestational age d and the predicted gestational age z_0, Kullback-Leibler Divergence KLD$(\boldsymbol{\mu}, \boldsymbol{\varphi})$ between the latent space of fBFN and the standard multivariative normal distribution $\mathcal{N}(0, 1)$. The arrows indicate whether a higher (up) or lower (down) value is preferred.

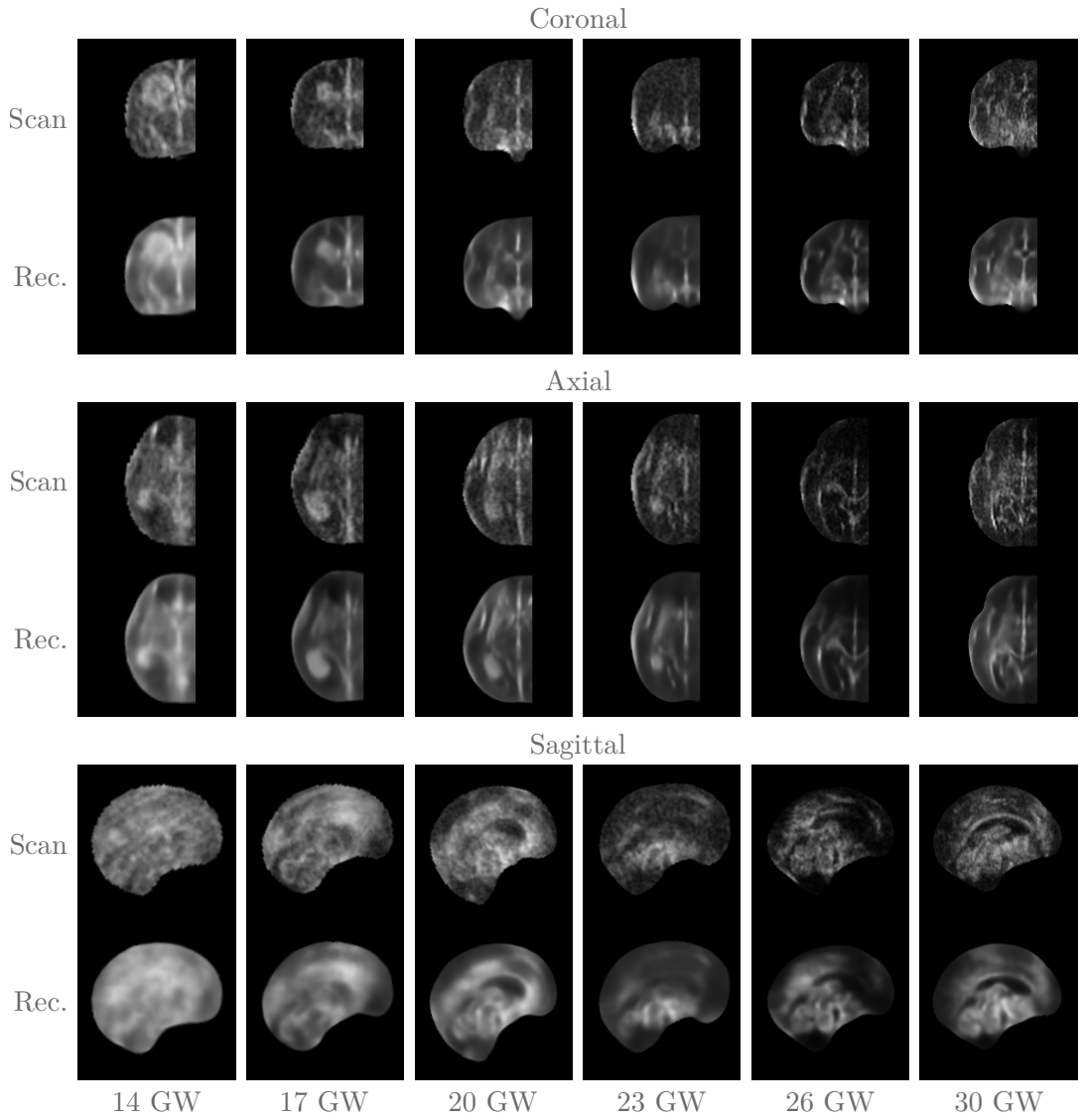

Figure 6.13: Examples of the reconstruction $\widehat{\mathbf{S}}$ predicted by fBFN compared against the input scan $\mathbf{S}$, for multiple GW. These examples were specifically selected since their SSIM$(\mathbf{S}, \widehat{\mathbf{S}}, \mathbf{M})$ performance is closest to the mean performance of fBFN for that GW. Top: Coronal midplane. Middle: Axial midplane. Bottom: Sagittal midplane.

The most salient features on the input scan $\mathbf{S}$ are preserved, as are the general luminance and contrast. However, the noise-like patterns of speckle is not conserved, instead being represented as a regional blur. This is a direct result of calculating the reconstruction loss $\mathfrak{L}_{\text{SSIM}}$ against the anisotropically filtered scan $\mathbf{S}^*$. Nevertheless, in spite of the high degree of accuracy with which fBFN reconstructs the original scan, these are not perfect. Some regions of the reconstructed scan are not entirely accurate, as can be seen in the axial view of the 14 GW example.

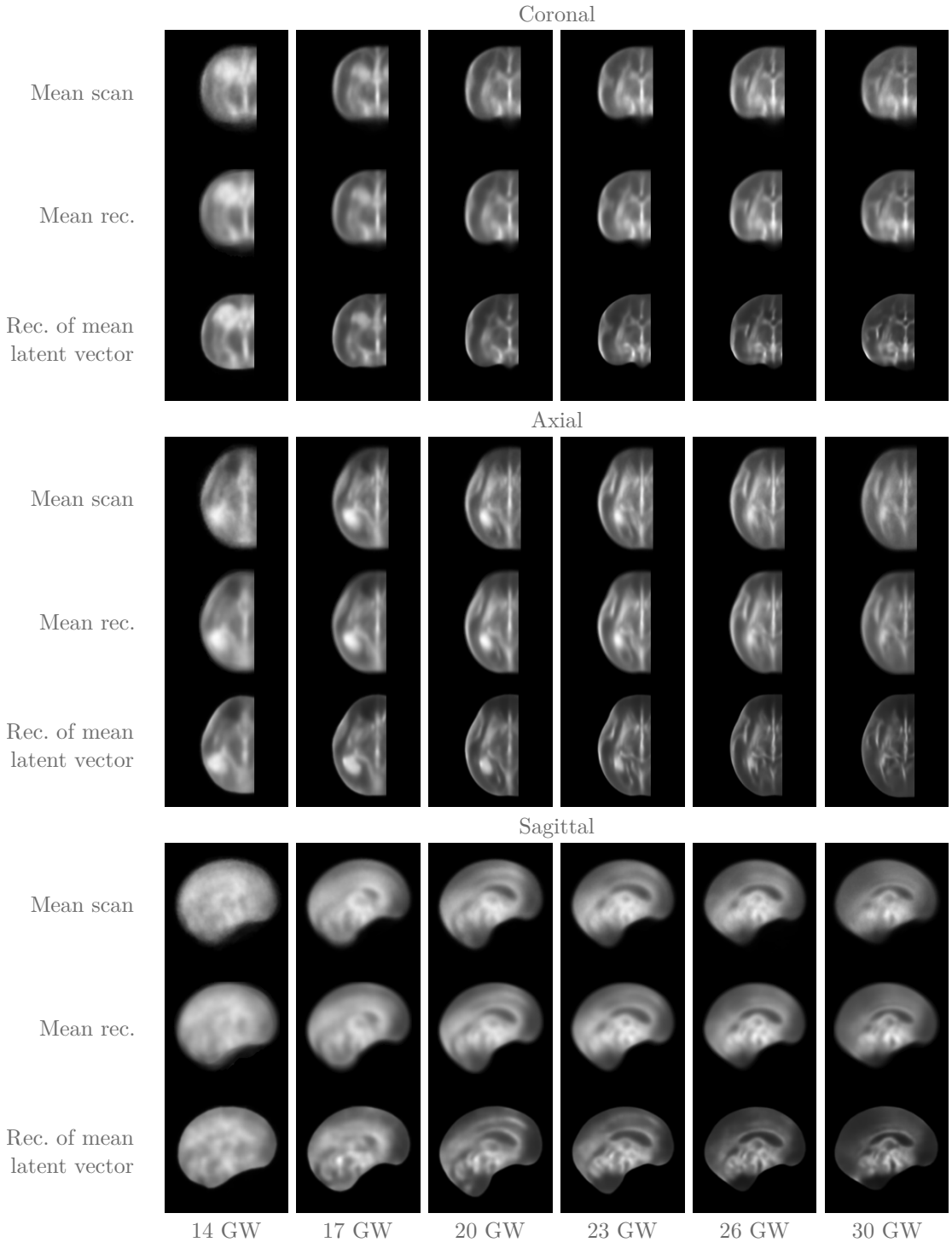

Figure 6.14: Mean scans $\mathbf{S}$ compared against the mean of reconstructed scans $\widehat{\mathbf{S}}$, and the reconstruction of the mean latent vector $\boldsymbol{z}$ predicted by fBFN compared, for multiple GW. Top: Coronal midplane. Middle: Axial midplane. Bottom: Sagittal midplane.

To better visualise the consistency of fBFN, Fig. 6.14 shows the orthogonal midplanes of the mean scan of the same GWs as before, as well as the mean reconstruction, and the reconstruction of the mean latent vector. The mean of the reconstructions is nearly identical to the mean scan, for the entire gestational range, once again reflecting the high reconstruction accuracy of fBFN. Only at 14 GW are the differences more significant, with the mean scan exhibiting a more grainy texture, while the mean reconstruction is smoother in nature. This is most likely due to interpolation artefacts generated when scaling the scan, which are not condensed into the latent vector $\boldsymbol{z}$. The reconstruction of the mean latent vector is also structurally consistent with the mean scan. However, this reconstruction is significantly sharper, and shows a higher contrast throughout. This is a direct benefit of the VAE architecture of fBFN, which creates a continuously distributed latent space. When averaging the scans of a particular GW, the variations in contrast, luminance, and structural information caused by factors such as the positioning of the probe relative to the fetal head are also averaged, which results in blurry regions. However, since these differences between scans are merely different locations in the continuously distributed latent space, their geometric centre represents an absence of such variations, in addition to representing the average structural information of the brain.

6.3.3 Regional performance

To assess the regional reconstruction performance of fBFN, the mean difference between the input scan $\mathbf{S}$ and the reconstruction $\widehat{\mathbf{S}}$ is calculated for several GWs representative of the entire gestational range, as shown in Fig. 6.15. Overall, these results are consistent with those shown in Fig. 6.12 and 6.14, showing that the reconstructions achieve remarkable structural fidelity across the entire gestational range, with only as small decrease in performance at 14 GWs. As mentioned in the previous section, this is likely due to interpolation artefacts generated when scaling the scan, which are not condensed into the latent vector.

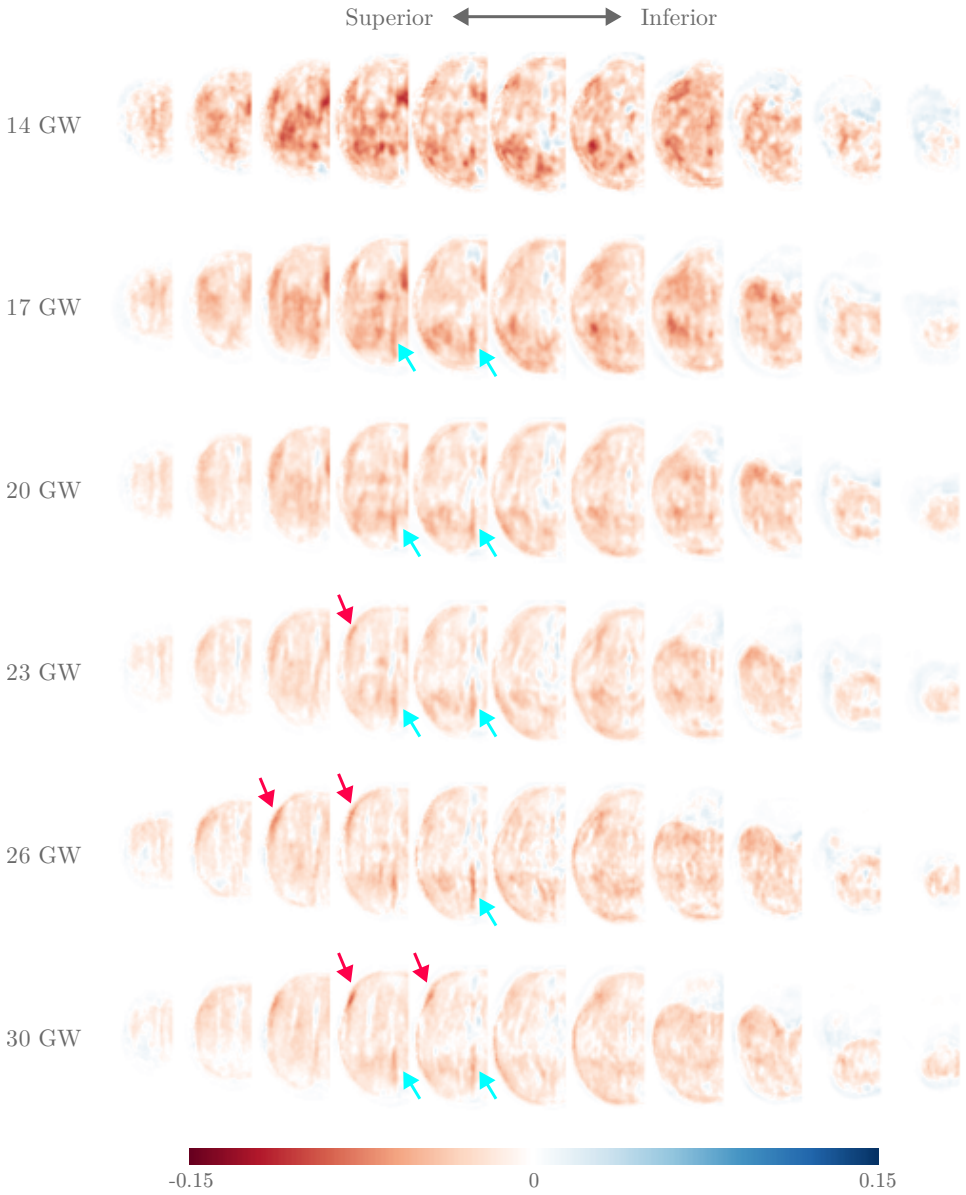

Figure 6.15: Mean regional performance of the reconstructed scans $\widehat{\mathbf{S}}$ predicted by fBFN against the input scan $\mathbf{S}$. Positive and negative values indicate regions where fBFN is over-predicting and under-predicting intensities, respectively. Gestational age is shown on the left. The results show a tendency to under-predict intensities evenly throughout the brain, with arrows indicating minor hotspots around the posterior portion of the Falx (cyan), as well as the border between the middle-frontal gyrus and the skull (crimson).

This reconstruction accuracy is reflected in a mean absolute intensity difference of 0.012 ± 0.013, which raises to 0.025 ± 0.021 at 14 GWs. Since the scans $\mathbf{S}$ are normalised to a range of 0 to 1, this represents a relative difference of only 1.2% and 2.5%, respectively.

In general, there is a tendency to under-predict intensities evenly throughout the brain, with the highest accuracy located around the regions that contain high amounts of contrast, such as the edges of bright structures. This is consistent with the mildly blurry quality of the reconstruction scans $\mathbf{S}$, which is characteristic of VAE networks [192][193][194]. However, Fig. 6.15 shows that there are some mild under-prediction hotspots around the posterior portion of the Falx, as well as the border between the middle-frontal gyrus and the skull, while the frontal portion of the Falx tends to be mildly over-predicted. This is likely due to the shadow artefacts created by the interaction between the US beam and the skull, which tend to affect these regions.

6.3.4 Latent space

After thoroughly analysing the reconstruction performance of fBFN, I continue the analysis by focusing on the characteristics of the latent space. Fig. 6.16 shows the mean value of each parameter of the latent vector $\boldsymbol{z}$ for each day of gestation within the gestational range. As expected, since fBFN encodes the gestational age in the 0-th parameter, a strong correlation is observed. However, the distribution of the intensities of most parameters is also correlated with age to some degree. Most parameter show peaks (or valleys) only for a certain age range, with weaker values in between. This is likely due to certain structural characteristics being only visible at certain GWs, or other characteristics of the scans that are related to the gestational age but not to the structural development of the brain, such as the acoustic shadows, angle of incidence of the beam, and distance between the US probe and the brain.

A 2D projection of the principal components of the latent vectors of the hold-out split of dataset $\mathfrak{D}_E$ is shown in Fig 6.17. As expected, the distribution of the latent space follows roughly a normal distribution thanks to the $\mathfrak{L}_{\mathrm{KLD}}(\boldsymbol{\mu}, \boldsymbol{\varphi})$ of

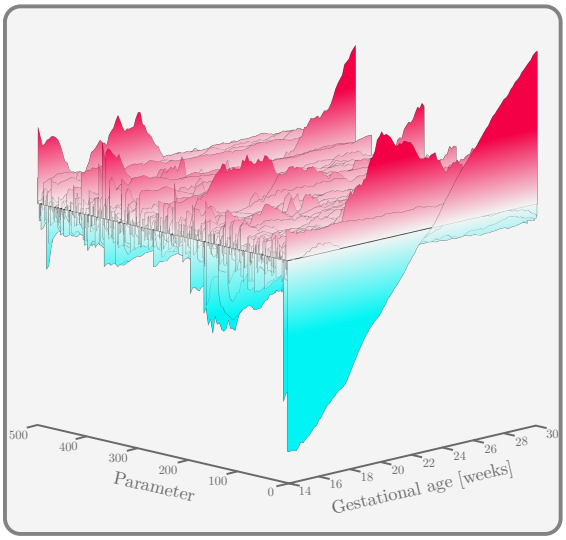

Figure 6.16: Relative values of the parameters of the latent vector $\boldsymbol{z}$, averaged by gestational age in days. The predicted gestational age z_0 is encoded in parameter 0.

the loss function. Additionally, the projection shows a clear linear distribution of gestational age. Thanks to the soft disentanglement approach introduced in

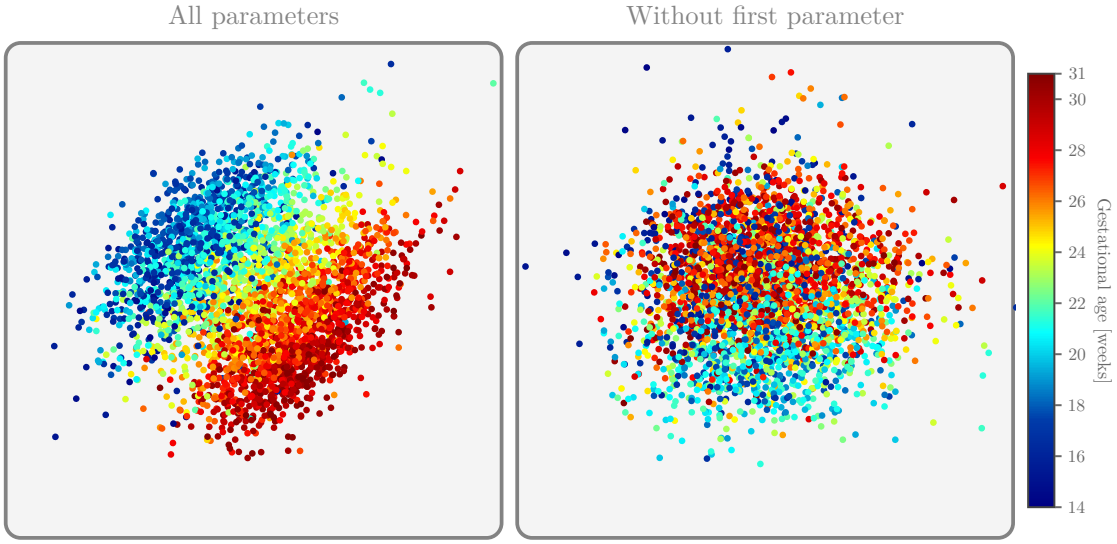

Figure 6.17: 2D projection of a PCA of the latent vectors $\boldsymbol{z}$ encoded by fBFN. Left: PCA of all 500 parameters. Right: PCA of 499 parameters, excluding the first parameter z_0 where the gestational age is encoded.

Sec. 6.2.6, the majority of the age information is contained in the first parameter of the latent vector, which is evident in the loss of a clear age distribution if this parameter is excluded from the PCA. This is corroborated when calculating the Pearson correlation factor s between each parameter and the gestational age of the

scan. The mean absolute values $|s|$ are shown in Fig. 6.18, which shows that the first parameter of the latent vector achieves a near perfect correlation with a score of 0.982, with the remaining 499 parameters obtaining a mean score of 0.054 ± 0.049. However, while the majority of these parameters show only a weak correlation, there are three parameters that achieve a modest correlation: parameter 201 with a score of 0.36, parameter 226 with a score of 0.32, and parameter 363 with a score of 0.31. However, this is consistent with the results already seen in Fig. 6.16.

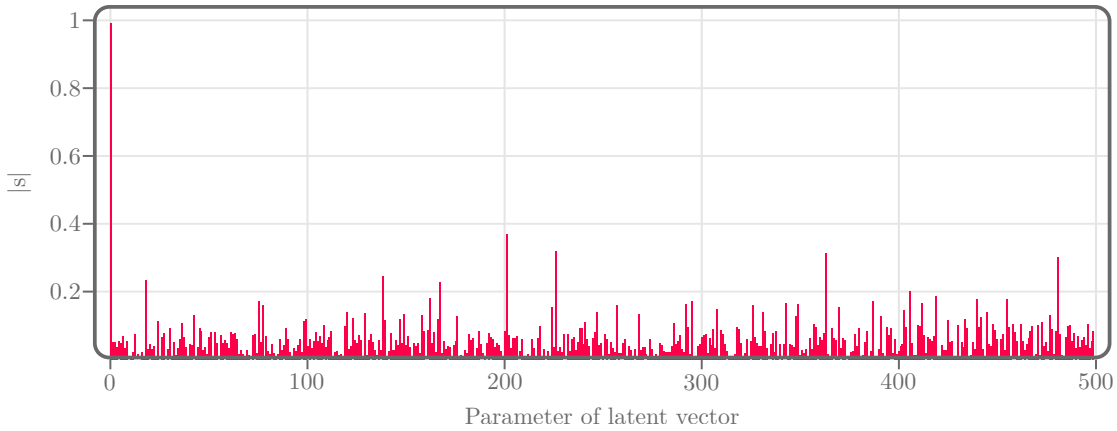

Figure 6.18: Absolute value of the Pearson correlation coefficient s between each of the 500 parameters of the encoded vectors $\boldsymbol{z}$ and the gestational age (in days) of their corresponding scan $\mathbf{S}$.

Nevertheless, by calculating the 2D projection of the principal components of these 4 parameters, as shown in Fig. 6.19, the same behaviour as before can be observed, with the age distribution no longer observed if the first parameter is removed.

6.3.5 Age manipulation

So far, I have focused on analysing the performance of fBFN, either analysing the accuracy of the predicted reconstructions or analysing the characteristics of the latent space and the parameters of the latent vectors. In this subsection I will instead analyse the effects that the manipulation of the latent vector has on the predicted reconstruction. In particular, I will focus on the manipulation of the gestational age of the scan. If the latent space were completely disentangled, this could be achieved by simply changing the value of the first parameter of the latent vector. However, the results in Sec. 6.3.4 have shown that this is not the case.

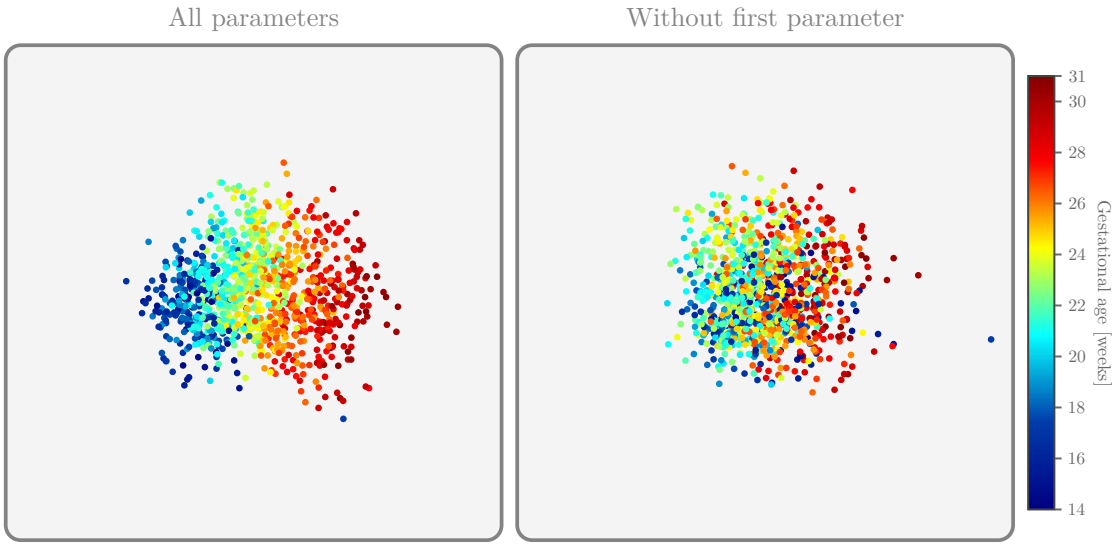

Figure 6.19: 2D projection of a PCA of the latent vectors $\boldsymbol{z}$ encoded by fBFN. Left: PCA of parameters 0, 201, 226, and 363. Right: PCA of parameters 201, 226, and 363, excluding the first parameter z_0 where the gestational age is encoded.

Therefore, in order to manipulate the gestational age of a scan, the required vector must first be determined. This can easily be done by calculating the difference vector between the mean latent vector at the goal gestational age and the mean latent vector at the gestational age of the original scan. This difference vector can then simply be added to the latent vector before being passed to the decoder, in order to reconstruct the manipulated scan.

Examples of age manipulated reconstructions are shown in Fig. 6.20, with the original reconstruction highlighted in red. The manipulated reconstruction keeps the general characteristics of the original scan, such as brightness and shadows, while being consistent with the structural development of the manipulated age. Additionally, the structural characteristics of the manipulated scan are dependent on the information provided in the original scan. Therefore, the reconstructions of different two scans manipulated to the same GW are also structurally different.

To determine how accurate the manipulated reconstructions are to the real brain development the manipulated reconstruction of a scan is compared against a real scan of the same subject, as shown in Fig. 6.21. As expected, the manipulated scan conserves the non-structural characteristics of the original scan. However, the

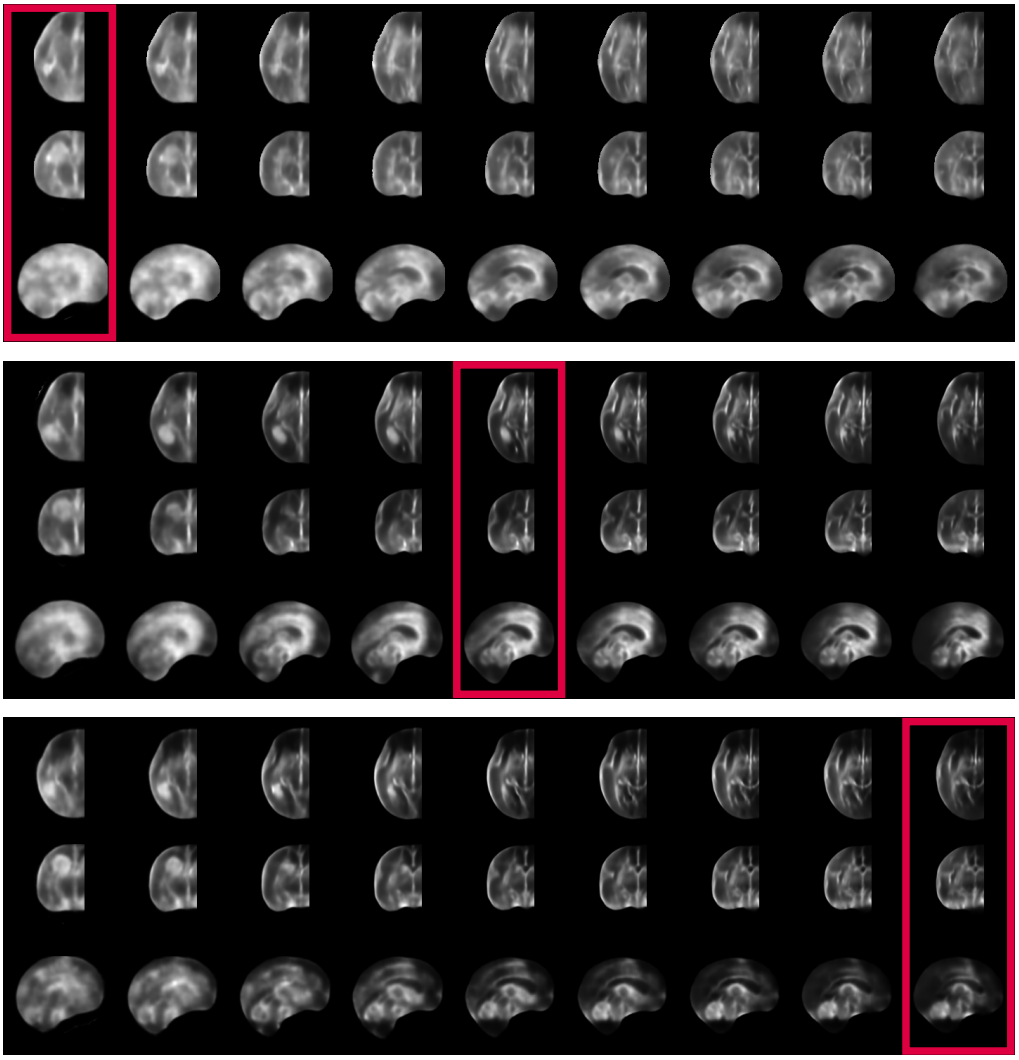

Figure 6.20: Examples of age-manipulated reconstructions. Each row shows the axial, coronal, and sagittal midplanes of a different scan, reconstructed at multiple GW (labelled at the bottom). The reconstruction at the original gestational age of each example is highlighted in red.

structural characteristic of the manipulated scan are remarkably close to the real scan at the same age, while being structurally distinct from other real scans.

These results highlight the potential that fBFN has a predictive tool. By comparing the manipulated scan with a real subsequent scan of the same subject, it would be possible to assess whether the development of the brain followed as expected. This would allow for clinical assessments to be longitudinal in nature, as well as personalised rather than chart based. Nevertheless, while these results are promising, its worth noting that this is only a small sample size and that

further research is required.

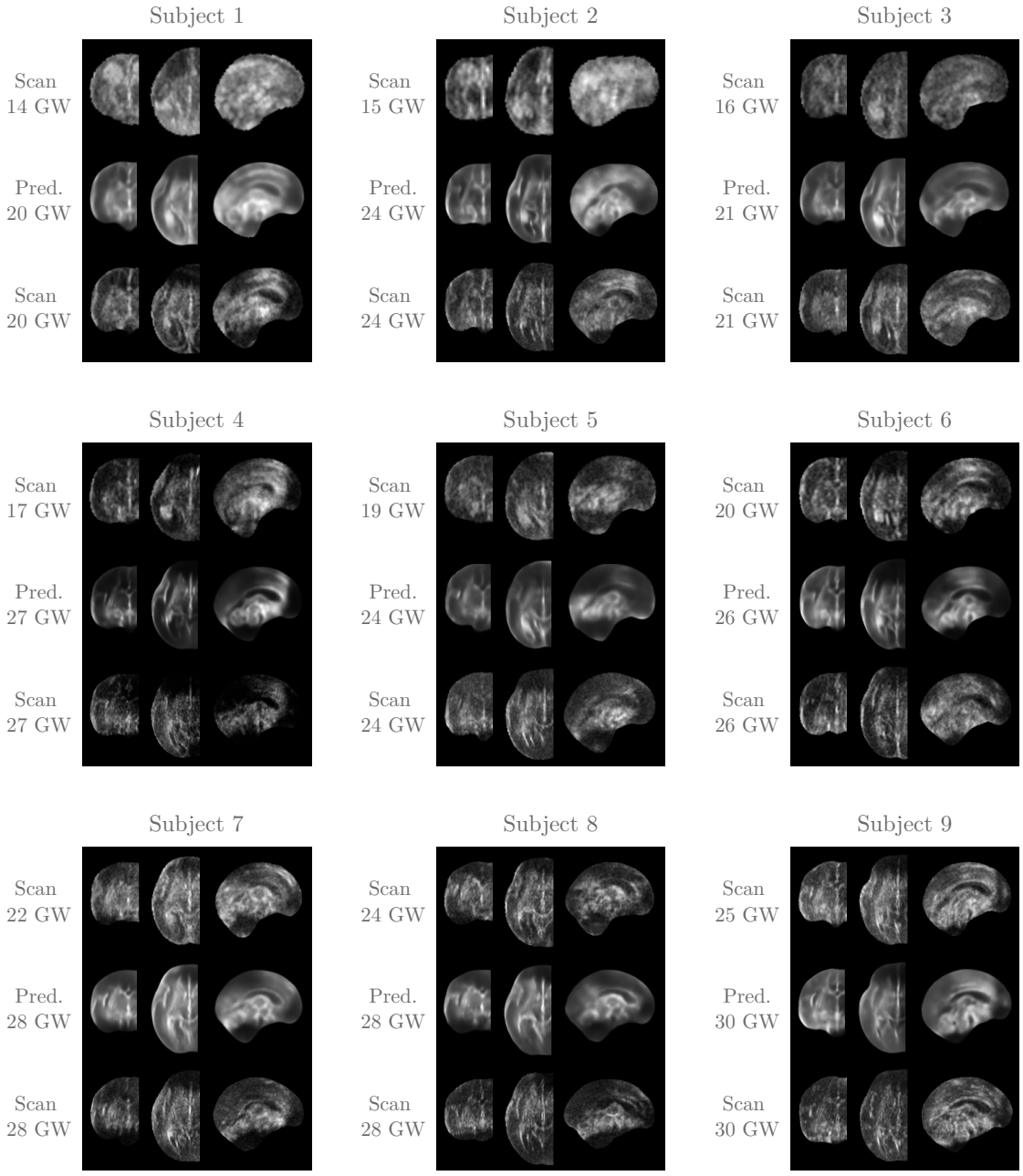

Figure 6.21: Examples of artificially aged reconstructions compared against subsequent scans of the same subject. For each subject, the top and bottom rows show the coronal, axial, and sagittal midplanes of 3D US scans acquired at different GWs, while the middle row shows the reconstruction of the earlier scan artificially aged to match the GW of the later scan.

6.3.6 Scan similarity comparison

After evaluating the effects of manipulating the gestational age of a scan, this subsection focuses on the potential of using fBFN as an effective method to quickly compare the structural characteristics of different scans. Due to the nature of the latent space of a VAE, it stands to reason that the closer two latent vectors are in latent space, the more similar the structural information between the original scans must be.

However, thoroughly evaluating this is rather difficult, as it would require an expert to assess and score the similarities between every pair of the 3445 scans of the hold-out split of dataset $\mathfrak{D}_E$ in order to generate a baseline. This would require 5932290 pairs to be compared, which would be unfeasible.

In order to circumvent this problem, the performance of fBFN is compared against that of SSIM. As the accuracy of relying on SSIM to find the structurally closest scans is also unknown, the gestational age difference will be used as a proxy. First, the SSIM($\mathbf{S}_A, \mathbf{S}_B$) between every pair of scans $\mathbf{S}_A$ and $\mathbf{S}_B$ is calculated, as well as the Euclidean distance $\Delta z = ||\boldsymbol{z}_B - \boldsymbol{z}_A||$ between their latent vectors. Then, for every scan $\mathbf{S}_A$, the closest matching $\mathbf{S}_B$ according to either method is found, i.e. the highest SSIM($\mathbf{S}_A, \mathbf{S}_B$) score or the smallest Δz distance. Finally, the MAE(d_A, d_B) between the reported gestational age of the first scan d_A (in days), and the age d_B of the closest match is calculated. Since scans of similar gestational age should be relatively close in terms of structural development, a smaller MAE(d_A, d_B) would indicate better performance as a scan similarity comparison.

Figure 6.22 shows the closest match according to either method for the same representative examples used in Fig. 6.13. With the exception of the example at 17 GW, each method resulted in a different closest match. The $\Delta \boldsymbol{z}$ method yielded a poor match at 14 GW, which is consistent with previous results. Overall, these examples suggest that both method performed similarly well, with the matches yielded by either method being structurally consistent with the original scan, as well as with each other.

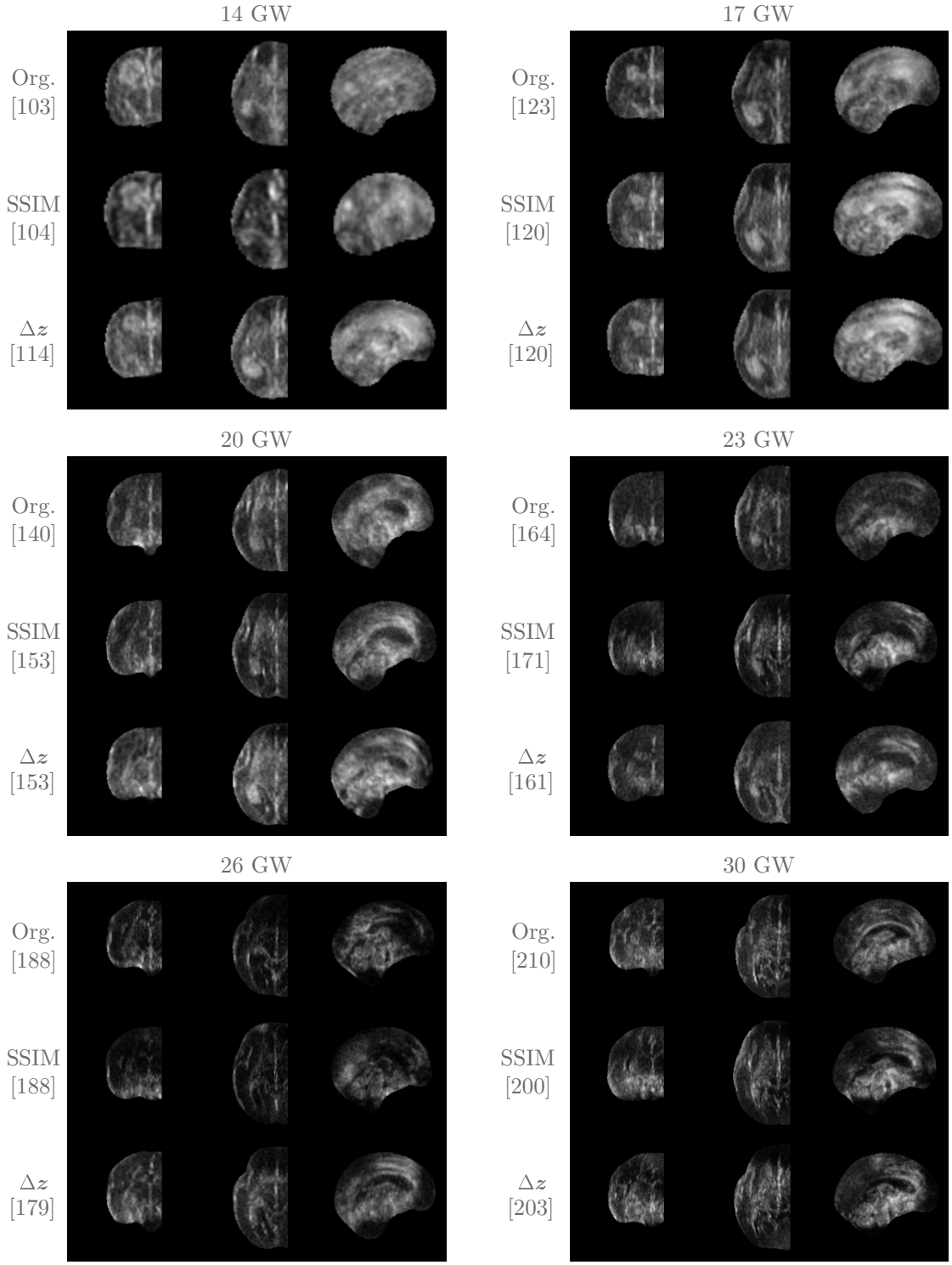

Figure 6.22: Examples of the closest match to the original scan (top), according to the SSIM (middle) and Δz (bottom) methods, for multiple GWs. The gestational age in days of each scan is provided in brackets on the left.

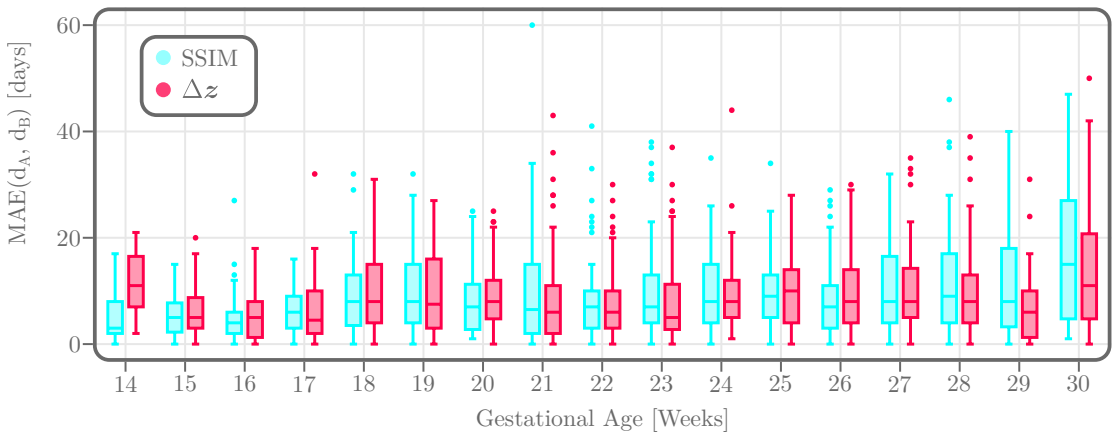

Figure 6.23: The Mean Absolute Error MAE(d_A, d_B) between the gestational age (in days) of each scan, and the gestational age of the closest match found, separated by GW.

The MAE(d_A, d_B) results for each GW are shown in Fig.6.23. While SSIM($\mathbf{S}_A, \mathbf{S}_B$) performed better between 14 GW and 18 GW, Δz surpassed it for most of the other GWs, outperforming it significantly for 29 GW and 30 GW. However, overall the performance of both methods is very similar.

This similar performance is confirmed by Tab. 6.2, which shows that the average performance of both methods is nearly identical, with the difference being non-statistically significant. However, in spite of achieving as similar performance, the condensed nature of the latent vectors meant that measuring all 5932290 Δz values took only 5.8 seconds. This is a stark contrast with the SSIM($\mathbf{S}_A, \mathbf{S}_B$), which took over 16 hours to be calculated with a state-of-the-art NVidia A10 GPU.

Table 6.2: Mean performance of the scan similarity comparison using the SSIM and $\Delta \boldsymbol{z}$ methods. The MAE(d_A, d_B) is the Mean Absolute Error between the gestational age (in days) of each scan, and the gestational age of the closest match found. The Time is the total time it took to compare every pair of scans in the hold-out test split of dataset $\mathfrak{D}_E$.

Method	MAE(d_A, d_B) [days] ↓	Time [s] ↓
SSIM	9.1 ± 8.0	58419.6
$\Delta \boldsymbol{z}$	8.8 ± 7.5	5.8

Therefore, while the results indicate that relying on the latent vectors generated by fBFN to perform a scan similarity comparison yields similar results as relying on the SSIM, the significant difference in computational cost and time make it a preferable option.

6.3.7 Structural development analysis

As a final step in the analysis, this subsection will focus on the potential use of fBFN as a general developmental analysis tool. To do this, the mean latent vector for each GW of the gestational range is reconstructed, subsequently comparing the observed structural development with what has been observed in literature. In particular, I will focus on the SF, the LV, and the CP, as their morphological development has been thoroughly described in the literature [195][76][5][64][196].

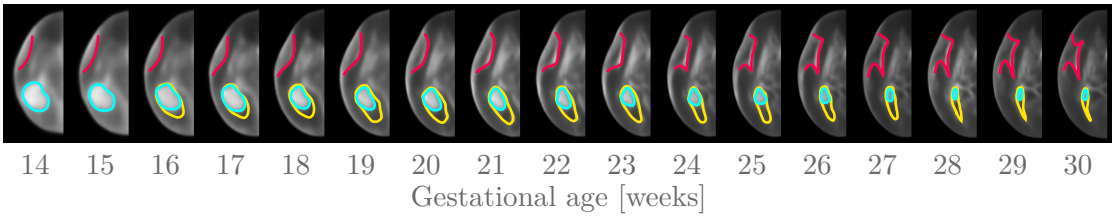

Figure 6.24: Axial midplane of the reconstructed mean latent vector $\boldsymbol{z}$ for each GW of the entire gestational range. The Sylvian Fissure (SF), Lateral Ventricle (LV), and Choroid Plexus (CP) have been manually outlined in crimson, cyan, and yellow, respectively.

Figure 6.24 shows the transventricular plane of the reconstruction of the mean latent vector, separated by GW. On each frame, the SF (crimson), the LV (yellow), and the CP (cyan) have been manually outlined, to illustrate their morphological development. Note that the LV was not clearly visible between 14 GW and 15 GW, and have therefore not been annotated.

At 14 GW, the CP covers a relatively large region of the brain, with a roughly elliptical or "bean" shape. As the fetal brain develops, the relative size of the CP becomes smaller, a process which continues until the end of the gestational range. At 16 GW, the walls of the LV are better defined, and are seen closely enveloping the CP with a visible gap towards the posterior of the brain. As the fetal brain develops, the length of the LV remains consistent relative to the size of the brain but its relative width shrinks, closely matching that of the CP. As a result, the gap between the CP and the LV increases in size, until around 27 GW, were walls of the LV seem to collapse at the furthest point. This morphological development of the LV and the CP is consistent with the descriptions found in [195], [76], and [5].

The SF is seen as a shallow indentation at 14 GW which increases in size as the brain develops. Around 20 GW, the shape begins to flatten, resulting in an obtuse angular shape. At 23 GW, the SF shows characteristics of operculisation, with the sides of the shape indenting towards its centre. The progress of this process is visible until 30 GW, where the operculisation is almost complete, with the short sides of the shape converging towards each other. This behaviour matches the development of the SF described in [64] and [196] almost perfectly.

While the analysis in this subsection only focuses on three cerebral structures, the performance of fBFN shown in Sections 6.3.2, 6.3.3, and 6.3.5 shows that the network has the potential to be used to research the development of other structures, without requiring the tedious and time consuming manual labelling of structures.

6.4 Discussion and Conclusions

In this chapter, I have proposed the third module of the Fully-Automated DL Pipeline for 3D Fetal Brain Ultrasound: the fetal Brain Fingerprinting Network (fBFN).

As in the previous chapters, I have shown the iterative process of developing the fBFN. Through exhaustive analysis, I have confirmed that fBFN manages to accurately and reliably encode the vast majority of the structural information of the 3D brain into a small latent vector $\boldsymbol{z}$ of 500 parameters. Additionally, the network is able to accurately decode the vector, predicting a reconstruction $\widehat{\mathbf{S}}$ that is structurally consistent with the input scan $\mathbf{S}$. I have also shown that the latent space created by fBFN is continuous, follows a roughly normal distribution, and exhibits a linear distribution of consistent with the gestational age of the scans.

By minimising the number of latent dimensions that are available for the network to encode the structural information of the scan, I have also shown that a soft-disentanglement of the latent space is possible by simply forcing the first parameter to be a prediction of the standardised gestational age $d_{\mathbf{S}}$ of the input scan. Additionally, the network managed to achieve a state-of-the-art accuracy of ± 2.29 days for the prediction of the gestational age of the input scan, surpassing the

current state-of-the-art solutions [32][33][34] when tested against dataset $\mathfrak{D}_D$. While this performance was lower for dataset $\mathfrak{D}_E$ at ± 4.22 days, this is still matching the current state-of-the-art performance achieved by Wyburd et al. [34] for whole-brain prediction. Additionally, this drop in performance is likely to be caused by inaccuracies in the predictions of fBEN and fBAN (see Sec. 3.4).

The results show that the performance of network is consistent throughout the entire gestational range of 14.0 to 30.9 GW, as well as throughout all regions of the fetal brain.

I have also demonstrated the potential for practical applications of fBFN. By manipulating the encoded vector, it could be used as a predictive tool to generate and expected development scan that can be compared with subsequent scans of the same subject, allowing for a more personalised assessment of the developmental progress of the brain. Alternatively, the Euclidean distance between latent vectors can be used to locate structurally similar scans with a similar performance as using SSIM, at a fraction of the computational cost. Furthermore, the characteristics of the latent space can be exploited to study the normal development of cerebral structures without the need for further data labelling.

Overall, fBFN has managed to achieve the goal of recharacterising the structural information of the 3D brain in order to facilitate its analysis. However, unlike the previous two modules, this requires further testing and refinement to be reliably deployed for general use. In particular, further disentanglement of the latent space would be of great benefit, as it would allow for an easier understanding of the structural information. Additionally, while the current reconstructions are remarkably accurate, most of the high-frequency information, such as speckle, is lost. Therefore, future work needs to focus on minimising the information that is lost in the encoding-decoding process. However, the most important focus for future work will be aimed at analysing the current characteristics of the latent space, and the developmental information that can be extracted from it.

7 Conclusion

Contents

7.1 Contributions

In this thesis, I have presented the three first fundamental modules that comprise the proposed fully-automated DL pipeline for 3D fetal brain ultrasound. The development of these modules constitute several contributions to the field, which I describe in this section. These contributions have been achieved strictly following the five criteria (Sec. 1.1) of performance, usability, robustness to age, robustness to misalignment, and robustness to quality.

7.1.1 fBEN

The first contribution of this thesis is the development of the fetal Brain Extraction Network fBEN (Ch. 4), a fully-automated, end-to-end 3D CNN with an encoder-decoder architecture for the automated extraction of the fetal brain from minimally pre-processed 3D US scans. Through the exhaustive assessment discussed in Sec. 4.3, I have demonstrated that fBEN achieves state-of-the-art performance for this task, significantly outperforming all current alternatives for 3D US, and achieving a performance that is similar to the equivalent state-of-the-art solution for fetal MRI. fBEN achieves this level of performance while being the first, and currently only 3D US solution that extracts the fetal brain directly from the 3D scan, without relying on templates, shape approximations, or a slice-by-slice approach. Additionally, fBEN manages to achieve consistent performance throughout the entire gestational range of 14.0 to 30.9 GW, currently being the only proposed solution for extracting the fetal brain earlier than 18 GW. Furthermore, the performance achieved by fBEN is virtually invariant to the location and orientation of the fetal brain within the 3D US scan, as well as to the orientation of the scan itself, facilitating its usability for a large range of users.

7.1.2 fBAN

The second contribution of this thesis is the development of the fetal Brain Alignment Network fBAN (Ch. 5), a fully-automated, end-to-end regression network with a cascade architecture that accurately predicts the alignment parameters required to rigidly align minimally pre-processed, standard clinical 3D US scans, to a canonical reference space. Through the exhaustive assessment discussed in Sec. 5.3, I have demonstrated that fBAN achieves state-of-the-art performance for this task, significantly outperforming all current alternatives for 3D US, and achieving a performance that is similar to the equivalent state-of-the-art solution for fetal MRI. Just like fBEN, fBAN achieves this level of performances by directly predicting the alignment parameters from the 3D US scan, without relying templates, shape approximations, slice-by-slice approaches, or the visibility of additional structures

such as the eye or the neck. The results discussed in Sec. 5.3 confirm that fBAN achieves this performance consistently throughout the entire gestational range of 14.0 to 30.9 GW, currently being the only proposed solution for alignment of the fetal brain earlier than 18 GW. Additionally, I have also demonstrated the robustness of the performance achieved by fBAN, which showed to be virtually invariant to the initial misalignment of the brain within the scan, as well as to the orientation of the scan itself.

7.1.3 fBFN

The third contribution of this thesis is the development of the fetal Brain Fingerprinting Network fBFN (Ch. 6), an end-to-end, general-purpose brain fingerprinting solution based on a VAE architecture, that recharacterises the 3D brain by encoding the structural information from the 3D US scan into a latent vector containing a relatively small set of descriptive parameters. fBFN is the first and currently only brain fingerprinting method proposed for fetal 3D US. Through the exhaustive assessment discussed in Sec. 6.3, I have demonstrated that fBFN manages to predict a structurally accurate reconstruction of the input 3D US fetal brain scan, confirming that the entire structural information is being encoded into the latent vector. Additionally, the results show that the high-performance of fBFN is achieved consistently across the entire gestational range of 14.0 to 30.9 GW, and is robust against a large image quality variability. Furthermore, in addition to achieving a state-of-the-art accuracy for the task of gestational age prediction, I have shown that the manner in which fBFN has encoded its latent space allows for the manipulation of the gestational age of the input scan. This has the potential to be used for predictive, personalised medicine, since artificially aged reconstructions have been shown to be structurally consistent with subsequent scans of the same subject. I have also shown that this learnt latent space contains a large amount of information regarding the normal structural development of the fetal brain. This was confirmed by analysing the learnt development of the SF, the LV, and the CP, which closely matches the descriptions find in the literature. Finally, I have demonstrated that

Euclidean distance between the latent vectors encoded from the 3D scans can be used as a fast, efficient solution for the task of image retrieval of structurally similar scans. This method achieved a similar performance to established measures such as the SSIM [163], at a fraction of the computational cost.

7.2 Limitations

In this section I cover the main current limitations of the three fundamental modules of the proposed pipeline.

7.2.1 Data

A shared limitation of fBEN, fBAN, and fBFN can be found in the manner in which the INTERGROWTH-21stwas collected. While this was a multi-site study spanning multiple countries, every 3D scan used for training and testing was collected using the same model of US machine and transducer, there is a risk that the performance of the network will not translate to scans collected with different equipment. However, although multi-site datasets of other modalities are being harmonised to remove scanner biases [197], in 3D US the operator settings (e.g. time-gain compensation) have a much larger impact on the characteristics of the image. Therefore, the variability introduced by the settings used in different sites should counterbalance this limitation.

Additionally, the INTERGROWTH-21ststudy was specifically designed to image healthy women that experienced a normal pregnancy. Therefore, the 3D scans used to develop these three modules excludes scans exhibiting particularly abnormal brain development. This could potentially hamper their applicability in such cases and needs to be explored.

Finally, the method in which the datasets were split for training and testing may cause a certain degree of data leakage. For some subjects, multiple 3D US scans were acquired during the same session, which was not accounted for when splitting the data. As a result, some of the scans of the testing set represent the same subject at the same gestational age as scans of the training set. In particular, this accounts

for 0 scans of Dataset $\mathfrak{D}_A$, 85 scans of Datasets $\mathfrak{D}_B$, $\mathfrak{D}_C$, and $\mathfrak{D}_D$, and 135 scans of $\mathfrak{D}_E$. Nevertheless, the effect this oversight may had on the assessed performance of each module of the pipeline is unlikely to be significant. Firstly, they account for only a small fraction of their respective testing set. Secondly, while these scans represent the same subject at the same gestational age, the information contained is qualitatively and quantitatively distinct, due to the data variability caused by probe positioning, head location and orientation, and acquisition settings.

7.2.2 fBEN

The main current limitations of fBEN the lack of expert-generated labels. As discussed in Sec. 3.3, the ground-truth labels used to train and evaluate fBEN have been generated by rigidly aligning the same binary mask to every scan of the corresponding gestational age. This results in noisy labelling, as this approach does not consider the normal variability of the fetal head. Additionally, the morphology of these masks is slightly different than the morphology of the brain that can be observed in 3D US since they were derived from the MRI-based, normative spatiotemporal atlas created by Gholipour et al. [159], which allows for the cerebrospinal fluid and the brain stem to be clearly distinguished from the rest of the scan. Furthermore, the reliance of the 21 GW atlas for the earlier GW results in additional labelling noise.

While this noisy labelling approach proved to be effective for training fBEN, it also limits the accuracy of the performance analysis. Therefore, in spite of the thorough assessment performed in Sec. 4.3, without expert-generated labels, the accuracy of the predictions generated by fBEN can only be assessed up to a certain point. However, this limitation is difficult to address, as the manual labelling of thousands of brain extraction masks by expert clinicians would be prohibitively time-consuming whilst requiring a high degree of expertise, making it unfeasible.

7.2.3 fBAN

Similarly to the main limitation of fBEN, one of the main limitations of fBAN lies on the lack of expert-generated ground-truth labels. In spite of the GUI I developed to facilitate the quality and consistency of my manual alignment (see Sec. 3.2), my lack of experience is certain to limit the accuracy of the ground-truth alignment parameters against which the predicted parameters were tested. While this noisy labelling approach proved to be effective for training fBAN, it also limits the accuracy of the performance analysis. Therefore, without expert-generated labels, the accuracy of the alignment performed by fBAN can only be assessed up to a certain point. However, while this would be significantly more feasible than expert-generated brain extraction labels, the manual alignment of thousands of scans by expert clinicians would still be prohibitively time-consuming.

Another limitation of fBAN is its reliance on a similarity transform for the predicted alignment. While this rigid registration approach to a canonical reference space was the goal of the module, a subsequent non-rigid registration step is likely to be necessary when comparing multiple scans.

7.2.4 fBFN

One current limitation of fBFN lies on the quality of the predicted reconstructed scan. While the results discussed in Sec. 6.3 confirm that these reconstructions achieve a high level of accuracy when compared to the original scan, there is still room for improvement. This would also ensure that complete structural information is encoded into the latent vector, something that cannot be directly assessed.

Another limitation is the constraint of only relying on the information-rich hemisphere of the brain. In addition to requiring an extra amount of pre-processing, this results in a loss of any structural information that might still be captured in the other hemisphere.

Finally, while I have managed to achieve a soft-disentanglement of the latent space by mostly separating the gestational age from the rest of the encoded

information, the lack of a true disentangled space makes the assessment and interpretation of the values encoded into the latent vector difficult.

7.3 Future Work

As mentioned in Sec 7.2.1, one of the current limitations of the fBEN, fBAN, and fBFN modules that comprise the proposed pipeline is the potential generalisation limitations due to relying entire on the INTERGROWTH-21stdata for their development. Therefore, one of the main focal areas of future works will be to obtain additional 3D US scans of the fetal brain acquired with different machines and probes, as well as scans that represent abnormal brain development.

Another focus for future work is with regards to the availability of the pipeline. The pipeline need to be easily accessible for researchers. The current plan is to make it freely available online, once the necessary approval of all involved parties has been received. However, current data-privacy concerns need to be addressed first. Additionally, an easy-to-use GUI will be essential for this pipeline to reach the widest possible audience. The development of this is already under way.

Finally, there is further improvement and analysis to be done for fBFN. As mentioned in Sec. 7.2.4, there is currently still room for improvement regarding the reconstruction accuracy of the network. Additionally, while the current results show that the artificially aged reconstructions are structurally consistent with future scans of the same subject, this needs to be assessed in-depth in order to determine the true potential that fBFN has for predictive, personalised medicine. Similarly, there current information of the learnt latent space of fBFN remains mostly unexplored. Given the quality of the mean reconstructions shown in Fig.6.14, it is clear that there is a significant amount of information regarding the normal structural development of the fetal brain waiting to be investigated. Furthermore, given the current accuracy of the reconstructions predicted by fBFN, it would be interesting to attempt a similar fingerprinting approach for adult brain MRI. While this is outside the scope of the proposed pipeline, it remains an interesting potential focus for future research.

References

[1] Joan Stiles and Terry L. Jernigan. 'The Basics of Brain Development'. In: *Neuropsychol Rev* 20.4 (2010), pp. 327–348. URL: https://doi.org/10.1007%2Fs11065-010-9148-4.

[2] 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?' In: *Ultrasound Obstet Gynecol* 24.7 (2004), pp. 706–715. URL: https://doi.org/10.1002%2Fuog.1802.

[3] International Society of Ultrasound in Obstetrics & Gynecology Education Committee et al. 'Sonographic examination of the fetal central nervous system: guidelines for performing the'basic examination'and the'fetal neurosonogram''. In: *Ultrasound in obstetrics & gynecology: the official journal of the International Society of Ultrasound in Obstetrics and Gynecology* 29.1 (2007), pp. 109–116.

[4] Stuart Campbell. 'A short history of sonography in obstetrics and gynaecology'. In: *Facts, views & vision in ObGyn* 5.3 (2013), p. 213.

[5] D Paladini et al. 'ISUOG Practice Guidelines (updated): sonographic examination of the fetal central nervous system. Part 2: performance of targeted neurosonography'. In: *Ultrasound in Obstetrics & Gynecology* 57.4 (2021), pp. 661–671.

[6] Donna Kirwan. 'Nhs fetal anomaly screening programme'. In: *National Standards and Guidance for England* 18.0 (2010).

[7] Mshe Bronshtein and Zeev Weiner. 'Prenatal diagnosis of dilated cava septi pellucidi et vergae: associated anomalies, differential diagnosis, and pregnancy outcome'. In: *Obstetrics & Gynecology* 80.5 (1992), pp. 838–842.

[8] Selami Serhatlioglu et al. 'Sonographic measurement of the fetal cerebellum, cisterna magna, and cavum septum pellucidum in normal fetuses in the second and third trimesters of pregnancy'. In: *J. Clin. Ultrasound* 31.4 (2003), pp. 194–200. URL: https://doi.org/10.1002%2Fjcu.10163.

[9] Lyndon M. Hill et al. 'The role of the transcerebellar view in the detection of fetal central nervous system anomaly'. In: *American Journal of Obstetrics and Gynecology* 164.5 (1991), pp. 1220–1224. URL: https://doi.org/10.1016%2F0002-9378%2891%2990686-l.

[10] Joanna J Phillips et al. 'Dandy-Walker malformation complex: correlation between ultrasonographic diagnosis and postmortem neuropathology'. In: *Obstetrics & Gynecology* 107.3 (2006), pp. 685–693.

[11] KH Nicolaides et al. 'Ultrasound screening for spina bifida: cranial and cerebellar signs'. In: *The Lancet* 328.8498 (1986), pp. 72–74.

[12] David A Nyberg et al. 'Enlarged cisterna magna and the Dandy-Walker malformation: factors associated with chromosome abnormalities.' In: *Obstetrics and gynecology* 77.3 (1991), pp. 436–442.

[13] G. Pilu et al. 'The clinical significance of fetal isolated cerebral borderline ventriculomegaly: report of 31 cases and review of the literature'. In: *Ultrasound Obstet Gynecol* 14.5 (1999), pp. 320–326. URL: https://doi.org/10.1046%2Fj.1469-0705.1999.14050320.x.

[14] P. Gaglioti et al. 'Fetal cerebral ventriculomegaly: outcome in 176 cases'. In: *Ultrasound Obstet Gynecol* 25.4 (2005), pp. 372–377. URL: https://doi.org/10.1002%2Fuog.1857.

[15] Israel Goldstein et al. 'Cerebellar measurements with ultrasonography in the evaluation of fetal growth and development'. In: *American journal of obstetrics and gynecology* 156.5 (1987), pp. 1065–1069.

[16] Catherine Garel et al. 'Fetal MRI: normal gestational landmarks for cerebral biometry, gyration and myelination'. In: *Child's Nervous System* 19.7 (2003), pp. 422–425.

[17] B S Mahony et al. 'The fetal cisterna magna.' In: *Radiology* 153.3 (1984), pp. 773–776. URL: https://doi.org/10.1148%2Fradiology.153.3.6387792.

[18] Roy A Filly et al. 'Detection of fetal central nervous system anomalies: a practical level of effort for a routine sonogram.' In: *Radiology* 172.2 (1989), pp. 403–408.

[19] Faye C Laing et al. 'Sonography of the fetal posterior fossa: false appearance of mega-cisterna magna and Dandy-Walker variant.' In: *Radiology* 192.1 (1994), pp. 247–251.

[20] I Sarris et al. 'Intra-and interobserver variability in fetal ultrasound measurements'. In: *Ultrasound in obstetrics & gynecology* 39.3 (2012), pp. 266–273.

[21] L. J. Salomon et al. 'The impact of choice of reference charts and equations on the assessment of fetal biometry'. In: *Ultrasound Obstet Gynecol* 25.6 (2005), pp. 559–565. URL: https://doi.org/10.1002%2Fuog.1901.

[22] Cynthia Ortinau et al. 'Regional alterations in cerebral growth exist preoperatively in infants with congenital heart disease'. In: *The Journal of thoracic and cardiovascular surgery* 143.6 (2012), pp. 1264–1270.

[23] S Zeng et al. 'Volume of intracranial structures on three-dimensional ultrasound in fetuses with congenital heart disease'. In: *Ultrasound in Obstetrics & Gynecology* 46.2 (2015), pp. 174–181.

[24] Richard W Prager et al. 'Three-dimensional ultrasound imaging'. In: *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine* 224.2 (2010), pp. 193–223.

[25] Sean P Fitzgibbon et al. 'The developing Human Connectome Project (dHCP) automated resting-state functional processing framework for newborn infants'. In: *Neuroimage* 223 (2020), p. 117303.

[26] Cathie Sudlow et al. 'UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age'. In: *PLoS medicine* 12.3 (2015), e1001779.

[27] Mark Jenkinson et al. 'Fsl'. In: *Neuroimage* 62.2 (2012), pp. 782–790.

[28] Yann Cointepas et al. 'BrainVISA: software platform for visualization and analysis of multi-modality brain data'. In: *Neuroimage* 13.6 (2001), p. 98.

[29] Oscar Esteban et al. 'fMRIPrep: a robust preprocessing pipeline for functional MRI'. In: *Nature methods* 16.1 (2019), pp. 111–116.

[30] Matthew F Glasser et al. 'The minimal preprocessing pipelines for the Human Connectome Project'. In: *Neuroimage* 80 (2013), pp. 105–124.

[31] Bo-yong Park, Kyoungseob Byeon and Hyunjin Park. 'FuNP (fusion of neuroimaging preprocessing) pipelines: a fully automated preprocessing software for functional magnetic resonance imaging'. In: *Frontiers in neuroinformatics* 13 (2019), p. 5.

[32] Ana I.L. Namburete et al. 'Learning-based prediction of gestational age from ultrasound images of the fetal brain'. In: *Medical Image Analysis* 21.1 (2015), pp. 72–86. URL: https://doi.org/10.1016%2Fj.media.2014.12.006.

[33] Ana I. L. Namburete, Weidi Xie and J. Alison Noble. 'Robust Regression of Brain Maturation from 3D Fetal Neurosonography Using CRNs'. In: *Fetal, Infant and Ophthalmic Medical Image Analysis*. Springer International Publishing, 2017, pp. 73–80. URL: https://doi.org/10.1007%2F978-3-319-67561-9_8.

[34] Madeleine K Wyburd et al. 'Assessment of Regional Cortical Development Through Fissure Based Gestational Age Estimation in 3D Fetal Ultrasound'. In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis*. Springer, 2021, pp. 242–252.

[35] Benjamin Gutierrez Becker et al. 'Automatic segmentation of the cerebellum of fetuses on 3D ultrasound images, using a 3D Point Distribution Model'. In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010. URL: https://doi.org/10.1109%2Fiembs.2010.5626624.

[36] Juan J Cerrolaza et al. 'Fetal skull segmentation in 3D ultrasound via structured geodesic random forest'. In: *Fetal, Infant and Ophthalmic Medical Image Analysis: International Workshop, FIFI 2017, and 4th International Workshop, OMIA 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 4*. Springer. 2017, pp. 25–32.

[37] Benjamín Gutiérrez-Becker et al. 'Automatic segmentation of the fetal cerebellum on ultrasound volumes, using a 3D statistical shape model'. In: *Medical & Biological Engineering & Computing* 51.9 (2013), pp. 1021–1030. URL: https://doi.org/10.1007%2Fs11517-013-1082-1.

[38] Ruobing Huang, J. Alison Noble and Ana I. L. Namburete. 'Omni-Supervised Learning: Scaling Up to Large Unlabelled Medical Datasets'. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Springer International Publishing, 2018, pp. 572–580. URL: https://doi.org/10.1007%2F978-3-030-00928-1_65.

[39] Linde S Hesse et al. 'Subcortical segmentation of the fetal brain in 3D ultrasound using deep learning'. In: *NeuroImage* 254 (2022), p. 119117.

[40] Gustavo Velásquez-Rodríguez, Fernando Arámbula Cosío and Boris Escalate Ramírez. 'Automatic segmentation of the fetal cerebellum using spherical harmonics and gray level profiles'. In: *11th International Symposium on Medical Information Processing and Analysis*. Ed. by Eduardo Romero et al. SPIE, 2015. URL: https://doi.org/10.1117%2F12.2207833.

[41] Patricia E Hansen et al. 'MR imaging of the developing human brain. Part 1. Prenatal development.' In: *Radiographics* 13.1 (1993), pp. 21–36.

[42] Maitray D Patel, Ann E Swinford and Roy A Filly. 'Anatomic and sonographic features of the fetal skull.' In: *Journal of ultrasound in medicine* 13.4 (1994), pp. 251–257.

[43] G. Pilu et al. 'Three-dimensional ultrasound examination of the fetal central nervous system'. In: *Ultrasound Obstet Gynecol* 30.2 (2007), pp. 233–245. URL: https://doi.org/10.1002%2Fuog.4072.

[44] Julian Lombardi and Julian Lombardi. 'Embryogenesis'. In: *Comparative Vertebrate Reproduction* (1998), pp. 225–251.

[45] Jeremy Muhr and Kristin M Ackerman. 'Embryology, gastrulation'. In: (2020).

[46] Joan Stiles. *The fundamentals of brain development: Integrating nature and nurture*. Harvard University Press, 2008.

[47] Clemens Kiecker and Andrew Lumsden. 'Hedgehog signaling from the ZLI regulates diencephalic regional identity'. In: *Nat Neurosci* 7.11 (2004), pp. 1242–1249. URL: https://doi.org/10.1038%2Fnn1338.

[48] Harukazu Nakamura and Yuji Watanabe. 'Isthmus organizer and regionalization of the mesencephalon and metencephalon'. In: *Int. J. Dev. Biol.* 49.2-3 (2005), pp. 231–235. URL: https://doi.org/10.1387%2Fijdb.041964hn.

[49] Andrew Lumsden and Roger Keynes. 'Segmental patterns of neuronal development in the chick hindbrain'. In: *Nature* 337.6206 (1989), pp. 424–428.

[50] Anthony Gavalas et al. 'Neuronal defects in the hindbrain of Hoxa1, Hoxb1 and Hoxb2 mutants reflect regulatory interactions among these Hox genes'. In: (2003).

[51] Kathie M Bishop, Guy Goudreau and Dennis DM O'Leary. 'Regulation of area identity in the mammalian neocortex by Emx2 and Pax6'. In: *Science* 288.5464 (2000), pp. 344–349.

[52] Kathie M Bishop, John LR Rubenstein and Dennis DM O'Leary. 'Distinct actions of Emx1, Emx2, andPax6 in regulating the specification of areas in the developing neocortex'. In: *Journal of Neuroscience* 22.17 (2002), pp. 7627–7638.

[53] Tadashi Hamasaki et al. 'EMX2 Regulates Sizes and Positioning of the Primary Sensory and Motor Areas in Neocortex by Direct Specification of Cortical Progenitors'. In: *Neuron* 43.3 (2004), pp. 359–372. URL: https://doi.org/10.1016%2Fj.neuron.2004.07.016.

[54] Dennis D.M. O'Leary, Shen-Ju Chou and Setsuko Sahara. 'Area Patterning of the Mammalian Cortex'. In: *Neuron* 56.2 (2007), pp. 252–269. URL: https://doi.org/10.1016%2Fj.neuron.2007.10.010.

[55] Andreas Zembrzycki et al. 'Genetic interplay between the transcription factors Sp8 and Emx2 in the patterning of the forebrain'. In: *Neural Dev* 2.1 (2007). URL: https://doi.org/10.1186%2F1749-8104-2-8.

[56] Dennis DM O'Leary and Setsuko Sahara. 'Genetic regulation of arealization of the neocortex'. In: *Current Opinion in Neurobiology* 18.1 (2008), pp. 90–100. URL: https://doi.org/10.1016%2Fj.conb.2008.05.011.

[57] Abeer Al Mohtar et al. 'Direct measurement of the effective infrared dielectric response of a highly doped semiconductor metamaterial'. In: *Nanotechnology* 28.12 (2017), p. 125701. URL: https://doi.org/10.1088%2F1361-6528%2Faa5ddf.

[58] Je G Chi, Elizabeth C Dooling and Floyd H. Gilles. 'Gyral development of the human brain'. In: *Ann Neurol.* 1.1 (1977), pp. 86–93. URL: https://doi.org/10.1002%2Fana.410010109.

[59] TP Naidich et al. 'The developing cerebral surface. Preliminary report on the patterns of sulcal and gyral maturation–anatomy, ultrasound, and magnetic resonance imaging.' In: *Neuroimaging Clinics of North America* 4.2 (1994), pp. 201–240.

[60] Este Armstrong et al. 'The Ontogeny of Human Gyrification'. In: *Cereb Cortex* 5.1 (1995), pp. 56–63. URL: https://doi.org/10.1093%2Fcercor%2F5.1.56.

[61] Wally Welker. 'Why Does Cerebral Cortex Fissure and Fold?' In: *Cerebral Cortex*. Springer US, 1990, pp. 3–136. URL: https://doi.org/10.1007%2F978-1-4615-3824-0_1.

[62] K Dorovini-Zis and CL Dolman. 'Gestational development of brain.' In: *Archives of pathology & laboratory medicine* 101.4 (1977), pp. 192–195.

[63] A. Afif et al. 'Development of the human fetal insular cortex: study of the gyration from 13 to 28 gestational weeks'. In: *Brain Struct Funct* 212.3-4 (2007), pp. 335–346. URL: https://doi.org/10.1007%2Fs00429-007-0161-1.

[64] LR Pistorius et al. 'Grade and symmetry of normal fetal cortical development: a longitudinal two-and three-dimensional ultrasound study'. In: *Ultrasound in obstetrics & gynecology* 36.6 (2010), pp. 700–708.

[65] A. Monteagudo and I. E. Timor-Tritsch. 'Development of fetal gyri, sulci and fissures: a transvaginal sonographic study'. In: *Ultrasound Obstet Gynecol* 9.4 (1997), pp. 222–228. URL: https://doi.org/10.1046%2Fj.1469-0705.1997.09040222.x.

[66] B. Cohen-Sacher et al. 'Sonographic developmental milestones of the fetal cerebral cortex: a longitudinal study'. In: *Ultrasound Obstet Gynecol* 27.5 (2006), pp. 494–502. URL: https://doi.org/10.1002%2Fuog.2757.

[67] Deborah Levine and Patrick D Barnes. 'Cortical maturation in normal and abnormal fetuses as assessed with prenatal MR imaging'. In: *Radiology* 210.3 (1999), pp. 751–758.

[68] Li Mei Lan et al. 'Normal fetal brain development: MR imaging with a half-Fourier rapid acquisition with relaxation enhancement sequence'. In: *Radiology* 215.1 (2000), pp. 205–210.

[69] K Ruoss et al. 'Brain development (sulci and gyri) as assessed by early postnatal MR imaging in preterm and term newborn infants'. In: *Neuropediatrics* 32.02 (2001), pp. 69–74.

[70] Catherine Garel et al. 'Fetal cerebral cortex: normal gestational landmarks identified using prenatal MR imaging'. In: *American Journal of Neuroradiology* 22.1 (2001), pp. 184–189.

[71] Seiji Abe et al. 'Assessment of cortical gyrus and sulcus formation using MR images in normal fetuses'. In: *Prenat. Diagn.* 23.3 (2003), pp. 225–231. URL: https://doi.org/10.1002%2Fpd.561.

[72] Céline Fogliarini et al. 'Assessment of cortical maturation with prenatal MRI. Part I: normal cortical maturation'. In: *Eur Radiol* 15.8 (2005), pp. 1671–1685. URL: https://doi.org/10.1007%2Fs00330-005-2782-1.

[73] LJ Salomon et al. 'ISUOG practice guidelines: performance of first-trimester fetal ultrasound scan'. In: *Ultrasound in obstetrics & gynecology: the official journal of the International Society of Ultrasound in Obstetrics and Gynecology* 41.1 (2013), pp. 102–113.

[74] Laurent Julien Salomon et al. 'Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan'. In: *Ultrasound in Obstetrics & Gynecology* 37.1 (2011), pp. 116–126.

[75] G. Malinger, D. Lev and T. Lerman-Sagie. 'Normal and abnormal fetal brain development during the third trimester as demonstrated by neurosonography'. In: *European Journal of Radiology* 57.2 (2006), pp. 226–232. URL: https://doi.org/10.1016%2Fj.ejrad.2005.11.022.

[76] J D Cardoza, R B Goldstein and R A Filly. 'Exclusion of fetal ventriculomegaly with a single measurement: the width of the lateral ventricular atrium.' In: *Radiology* 169.3 (1988), pp. 711–714. URL: https://doi.org/10.1148%2Fradiology.169.3.3055034.

[77] BS Hertzberg et al. 'Choroid plexus-ventricular wall separation in fetuses with normal-sized cerebral ventricles at sonography: postnatal outcome.' In: *AJR. American journal of roentgenology* 163.2 (1994), pp. 405–410.

[78] R A Filly, D H Chinn and P W Callen. 'Alobar holoprosencephaly: ultrasonographic prenatal diagnosis.' In: *Radiology* 151.2 (1984), pp. 455–459. URL: https://doi.org/10.1148%2Fradiology.151.2.6709918.

[79] Catherine Garel, Catherine Fallet-Bianco and Laurent Guibaud. 'The fetal cerebellum: development and common malformations'. In: *Journal of child neurology* 26.12 (2011), pp. 1483–1492.

[80] Raffaele Napolitano et al. 'Pregnancy dating by fetal crown–rump length: a systematic review of charts'. In: *BJOG: An International Journal of Obstetrics & Gynaecology* 121.5 (2014), pp. 556–565.

[81] Aris T Papageorghiou et al. 'International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project'. In: *The Lancet* 384.9946 (2014), pp. 869–879. URL: https://doi.org/10.1016%2Fs0140-6736%2814%2961490-2.

[82] RJM Snijders and KH Nicolaides. 'Fetal biometry at 14–40 weeks' gestation'. In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 4.1 (1994), pp. 34–48.

[83] Lyn S Chitty et al. 'Charts of fetal size: 2. Head measurements'. In: *BJOG: An International Journal of Obstetrics & Gynaecology* 101.1 (1994), pp. 35–43.

[84] Juozas Kurmanavicius et al. 'Fetal ultrasound biometry: 1. Head reference values'. In: *BJOG: An International Journal of Obstetrics & Gynaecology* 106.2 (1999), pp. 126–135.

[85] GIANLUIGI Pilu et al. 'Sonographic evaluation of the normal developmental anatomy of the fetal cerebral ventricles: II. The atria.' In: *Obstetrics and gynecology* 73.2 (1989), pp. 250–256.

[86] ASM Vinkesteijn, PGH Mulder and JW Wladimiroff. 'Fetal transverse cerebellar diameter measurements in normal and reduced fetal growth'. In: *Ultrasound in obstetrics and gynecology* 15.1 (2000), pp. 47–51.

[87] Jeanne A Haimovici et al. 'Clinical significance of isolated enlargement of the cisterna magna (> 10 mm) on prenatal sonography.' In: *Journal of ultrasound in medicine* 16.11 (1997), pp. 731–734.

[88] Benjamín Gutiérrez Becker et al. 'Automatic segmentation of the fetal cerebellum on 3D ultrasound.' In: *Simposio Mexicano en Cirugía Asistida por Computadora y Procesamiento de Imágenes Médicas*. 2011.

[89] M. Yaqub et al. 'Automatic detection of local fetal brain structures in ultrasound images'. In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2012. URL: https://doi.org/10.1109%2Fisbi.2012.6235870.

[90] R Napolitano et al. 'OP26. 06: Automatic detection of fetal brain structures from ultrasound volumes'. In: *Ultrasound in Obstetrics & Gynecology* 40.S1 (2012), pp. 134–134.

[91] Mohammad Yaqub et al. 'Volumetric segmentation of key fetal brain structures in 3D ultrasound'. In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2013, pp. 25–32.

[92] Xinyu Liu et al. 'Automatic localization of the fetal cerebellum on 3D ultrasound volumes'. In: *Medical physics* 40.11 (2013), p. 112902.

[93] Gustavo Velásquez-Rodríguez et al. 'Automatic segmentation of the cerebellum in ultrasound volumes of the fetal brain'. In: *Revista mexicana de ingeniería biomédica* 36.2 (2015), pp. 121–129.

[94] Ruobing Huang, Weidi Xie and J. Alison Noble. 'VP-Nets : Efficient automatic localization of key brain structures in 3D fetal neurosonography'. In: *Medical Image Analysis* 47 (2018), pp. 127–139. URL: https://doi.org/10.1016%2Fj.media.2018.04.004.

[95] Hsin-Chen Chen et al. 'Registration-Based Segmentation of Three-Dimensional Ultrasound Images for Quantitative Measurement of Fetal Craniofacial Structure'. In: *Ultrasound in Medicine & Biology* 38.5 (2012), pp. 811–823. URL: https://doi.org/10.1016%2Fj.ultrasmedbio.2012.01.025.

[96] Joni-Kristian Kamarainen. 'Gabor features in image analysis'. In: *2012 3rd international conference on image processing theory, tools and applications (IPTA)*. IEEE. 2012, pp. 13–14.

[97] Rémi Cuingnet et al. 'Where is my baby? A fast fetal head auto-alignment in 3D-ultrasound'. In: *2013 IEEE 10th International Symposium on Biomedical Imaging*. IEEE. 2013, pp. 768–771.

[98] Benoît Mory et al. 'Real-time 3D image segmentation by user-constrained template deformation'. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012: 15th International Conference, Nice, France, October 1-5, 2012, Proceedings, Part I 15*. Springer. 2012, pp. 561–568.

[99] Ana IL Namburete, Richard V Stebbing and J Alison Noble. 'Cranial parametrization of the fetal head for 3D ultrasound image analysis'. In: *Medical Image Understanding and Analysis (MIUA)* (2013), pp. 196–201.

[100] Olaf Ronneberger, Philipp Fischer and Thomas Brox. 'U-net: Convolutional networks for biomedical image segmentation'. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[101] Juan J Cerrolaza et al. 'Deep learning with ultrasound physics for fetal skull segmentation'. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 564–567.

[102] JL Pérez González et al. 'Ultrasound fetal brain registration using weighted coherent point drift'. In: *12th International Symposium on Medical Information Processing and Analysis*. Vol. 10160. SPIE. 2017, pp. 48–54.

[103] Jorge Perez–Gonzalez et al. 'Probabilistic learning coherent point drift for 3D ultrasound fetal head registration'. In: *Computational and Mathematical Methods in Medicine* 2020 (2020).

[104] Ana I.L. Namburete et al. 'Fully-automated alignment of 3D fetal brain ultrasound to a canonical reference space using multi-task learning'. In: *Medical Image Analysis* 46 (2018), pp. 1–14. URL: https://doi.org/10.1016%2Fj.media.2018.02.006.

[105] Felipe Moser et al. 'Automated Fetal Brain Extraction from Clinical Ultrasound Volumes Using 3D Convolutional Neural Networks'. In: *Communications in Computer and Information Science*. Springer International Publishing, 2020, pp. 151–163. URL: https://doi.org/10.1007%2F978-3-030-39343-4_13.

[106] Felipe Moser et al. 'BEAN: Brain Extraction and Alignment Network for 3D Fetal Neurosonography'. In: *NeuroImage* (2022), p. 119341.

[107] Jérémie Anquez, Elsa D Angelini and Isabelle Bloch. 'Automatic segmentation of head structures on fetal MRI'. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE. 2009, pp. 109–112.

[108] Faliu Yi and Inkyu Moon. 'Image segmentation: A survey of graph-cut methods'. In: *2012 international conference on systems and informatics (ICSAI2012)*. IEEE. 2012, pp. 1936–1941.

[109] Piotr A Habas et al. 'A spatiotemporal atlas of MR intensity, tissue probability and shape of the fetal brain with application to segmentation'. In: *Neuroimage* 53.2 (2010), pp. 460–470.

[110] A. Gholipour et al. 'Multi-atlas multi-shape segmentation of fetal brain MRI for volumetric and morphometric analysis of ventriculomegaly'. In: *NeuroImage* 60 (2012), pp. 1819–1831.

[111] Youssef Taleb et al. 'Automatic Template-based Brain Extraction in Fetal MR Images'. In: 2013.

[112] Robert Wright et al. 'Automatic quantification of normal cortical folding patterns from fetal brain MRI'. In: *Neuroimage* 91 (2014), pp. 21–32.

[113] Vahid Taimouri et al. 'A template-to-slice block matching approach for automatic localization of brain in fetal MRI'. In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. 2015, pp. 144–147.

[114] Sébastien Tourbier et al. 'Automated template-based brain localization and extraction for fetal brain MRI reconstruction'. In: *NeuroImage* 155 (2017), pp. 460–472.

[115] Bernhard Kainz et al. 'Fast fully automatic brain detection in fetal MRI using dense rotation invariant image descriptors'. In: *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2014. URL: https://doi.org/10.1109%2Fisbi.2014.6868098.

[116] K. Keraudren et al. 'Automated fetal brain segmentation from 2D MRI slices for motion correction'. In: *NeuroImage* 101 (2014), pp. 633–643. URL: https://doi.org/10.1016%2Fj.neuroimage.2014.07.023.

[117] Martin Rajchl et al. 'Learning under distributed weak supervision'. In: *arXiv preprint arXiv:1606.01100* (2016).

[118] Martin Rajchl et al. 'Deepcut: Object segmentation from bounding box annotations using convolutional neural networks'. In: *IEEE transactions on medical imaging* 36.2 (2016), pp. 674–683.

[119] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus and Ali Gholipour. 'Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging'. In: *IEEE transactions on medical imaging* 36.11 (2017), pp. 2319–2330.

[120] Nadieh Khalili et al. 'Automatic segmentation of the intracranial volume in fetal MR images'. In: *Fetal, Infant and Ophthalmic Medical Image Analysis*. Springer, 2017, pp. 42–51.

[121] Nadieh Khalili et al. 'Automatic extraction of the intracranial volume in fetal and neonatal MR scans using convolutional neural networks'. In: *NeuroImage: Clinical* 24 (2019), p. 102061.

[122] Michael Ebner et al. 'An automated framework for localization, segmentation and super-resolution reconstruction of fetal brain MRI'. In: *NeuroImage* 206 (2020), p. 116324.

[123] Marta Ranzini et al. 'MONAIfbs: MONAI-based fetal brain MRI deep learning segmentation'. In: *arXiv preprint arXiv:2103.13314* (2021).

[124] Maria Kuklisova-Murgasova et al. 'Towards 3D registration of fetal brain MRI and ultrasound'. In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2012. URL: https://doi.org/10.1109%2Fisbi.2012.6235555.

[125] Maria Kuklisova-Murgasova et al. 'Registration of 3D Fetal Brain US and MRI'. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*. Springer Berlin Heidelberg, 2012, pp. 667–674. URL: https://doi.org/10.1007%2F978-3-642-33418-4_82.

[126] Ana I. L. Namburete et al. 'Multi-channel Groupwise Registration to Construct an Ultrasound-Specific Fetal Brain Atlas'. In: *Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis*. Springer International Publishing, 2018, pp. 76–86. URL: https://doi.org/10.1007%2F978-3-030-00807-9_8.

[127] Robert Wright et al. 'Complete fetal head compounding from multi-view 3D ultrasound'. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*. Springer. 2019, pp. 384–392.

[128] Robert Wright et al. 'LSTM spatial co-transformer networks for registration of 3D fetal US and MR brain images'. In: *Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis: First International Workshop, DATRA 2018 and Third International Workshop, PIPPI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*. Springer. 2018, pp. 149–159.

[129] Benjamin Hou et al. '3-D Reconstruction in Canonical Co-Ordinate Space From Arbitrarily Oriented 2-D Images'. In: *IEEE Transactions on Medical Imaging* 37.8 (2018), pp. 1737–1750.

[130] Shun Miao, Z. Jane Wang and Rui Liao. 'A CNN Regression Approach for Real-Time 2D/3D Registration'. In: *IEEE Transactions on Medical Imaging* 35.5 (2016), pp. 1352–1363.

[131] Evan M Gordon et al. 'Individual variability of the system-level organization of the human brain'. In: *Cerebral cortex* 27.1 (2017), pp. 386–399.

[132] J-F Mangin et al. 'A framework to study the cortical folding patterns'. In: *Neuroimage* 23 (2004), S129–S138.

[133] Deanna M Barch et al. 'Function in the human connectome: task-fMRI and individual differences in behavior'. In: *Neuroimage* 80 (2013), pp. 169–189.

[134] Evan M Gordon et al. 'Individual-specific features of brain systems identified with resting state functional correlations'. In: *Neuroimage* 146 (2017), pp. 918–939.

[135] Sophia Mueller et al. 'Individual variability in functional connectivity architecture of the human brain'. In: *Neuron* 77.3 (2013), pp. 586–595.

[136] Ruben Geevarghese et al. 'Subcortical structure volumes and correlation to clinical variables in Parkinson's disease'. In: *Journal of Neuroimaging* 25.2 (2015), pp. 275–280.

[137] Anita Thapar, Miriam Cooper and Michael Rutter. 'Neurodevelopmental disorders'. In: *The Lancet Psychiatry* 4.4 (2017), pp. 339–346.

[138] Sylvie Goldman et al. 'Motor stereotypies and volumetric brain alterations in children with Autistic Disorder'. In: *Research in autism spectrum disorders* 7.1 (2013), pp. 82–92.

[139] Harald Hampel et al. 'A precision medicine initiative for Alzheimer's disease: the road ahead to biomarker-guided integrative disease modeling'. In: *Climacteric* 20.2 (2017), pp. 107–118.

[140] Judit Ciarrusta et al. 'The developing brain structural and functional connectome fingerprint'. In: *Developmental Cognitive Neuroscience* 55 (2022), p. 101117.

[141] Jung-Hoon Kim et al. 'Towards A More Informative Representation of the Fetal-Neonatal Brain Connectome using Variational Autoencoder'. In: *bioRxiv* (2022), pp. 2022–06.

[142] Emily S Finn et al. 'Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity'. In: *Nature neuroscience* 18.11 (2015), pp. 1664–1671.

[143] Jin Liu et al. 'Chronnectome fingerprinting: Identifying individuals and predicting higher cognitive functions using dynamic brain connectivity patterns'. In: *Human brain mapping* 39.2 (2018), pp. 902–915.

[144] Dhiraj Ahuja and Bharat Singh. 'Brain fingerprinting'. In: *Journal of Engineering and Technology Research* 4.6 (2012), pp. 98–103.

[145] Henrique M Fernandes et al. 'Novel fingerprinting method characterises the necessary and sufficient structural connectivity from deep brain stimulation electrodes for a successful outcome'. In: *New Journal of Physics* 17.1 (2015), p. 015001.

[146] Kuldeep Kumar et al. 'Fiberprint: A subject fingerprint based on sparse code pooling for white matter fiber analysis'. In: *NeuroImage* 158 (2017), pp. 242–259.

[147] Oscar Miranda-Dominguez et al. 'Connectotyping: model based fingerprinting of the functional connectome'. In: *PloS one* 9.11 (2014), e111048.

[148] Fang-Cheng Yeh, David Badre and Timothy Verstynen. 'Connectometry: a statistical approach harnessing the analytical potential of the local connectome'. In: *Neuroimage* 125 (2016), pp. 162–171.

[149] Charles DeCarli et al. 'Measures of brain morphology and infarction in the framingham heart study: establishing what is normal'. In: *Neurobiology of aging* 26.4 (2005), pp. 491–510.

[150] Bruce Fischl. 'FreeSurfer'. In: *Neuroimage* 62.2 (2012), pp. 774–781.

[151] Matthew Toews et al. 'Feature-based morphometry: Discovering group-related anatomical patterns'. In: *NeuroImage* 49.3 (2010), pp. 2318–2327.

[152] Christian Wachinger et al. 'BrainPrint: A discriminative characterization of brain morphology'. In: *NeuroImage* 109 (2015), pp. 232–248.

[153] Martin Reuter, Franz-Erich Wolter and Niklas Peinecke. 'Laplace–Beltrami spectra as 'Shape-DNA'of surfaces and solids'. In: *Computer-Aided Design* 38.4 (2006), pp. 342–366.

[154] Benjamín Gutiérrez Becker et al. 'Automatic segmentation of the cerebellum of fetuses on 3D ultrasound images, using a 3D Point Distribution Model'. In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE. 2010, pp. 4731–4734.

[155] Jean-Marie Moutquin. 'Classification and heterogeneity of preterm birth'. In: *BJOG: An International Journal of Obstetrics & Gynaecology* 110 (2003), pp. 30–33.

[156] Robert E Black. 'Global prevalence of small for gestational age births'. In: *Low-Birthweight baby: born too soon or too small*. Vol. 81. Karger Publishers, 2015, pp. 1–7.

[157] Starting Matlab. 'Matlab'. In: *The MathWorks, Natick, MA* (2012).

[158] Leonhard Euler. 'Novi commentarii academiae scientiarum petropolitanae'. In: *Nr* 20 (1776), pp. 189–207.

[159] Ali Gholipour et al. 'A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth'. In: *Sci Rep* 7.1 (2017). URL: https://doi.org/10.1038%2Fs41598-017-00525-w.

[160] John C Mazziotta et al. 'A probabilistic atlas of the human brain: theory and rationale for its development'. In: *Neuroimage* 2.2 (1995), pp. 89–101.

[161] John Mazziotta et al. 'A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM)'. In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 356.1412 (2001), pp. 1293–1322.

[162] John Mazziotta et al. 'A four-dimensional probabilistic atlas of the human brain'. In: *Journal of the American Medical Informatics Association* 8.5 (2001), pp. 401–430.

[163] Zhou Wang et al. 'Image quality assessment: from error visibility to structural similarity'. In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[164] Kenji Suzuki. 'Overview of deep learning in medical imaging'. In: *Radiological physics and technology* 10.3 (2017), pp. 257–273.

[165] Martín Abadi et al. 'Tensorflow: Large-scale machine learning on heterogeneous distributed systems'. In: *arXiv preprint arXiv:1603.04467* (2016).

[166] François Chollet et al. *keras*. 2015.

[167] Adam Paszke et al. 'PyTorch: An Imperative Style, High-Performance Deep Learning Library'. In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[168] Diederik P Kingma and Jimmy Ba. 'Adam: A method for stochastic optimization'. In: *arXiv preprint arXiv:1412.6980* (2014).

[169] Ralph B d'Agostino. 'An omnibus test of normality for moderate and large size samples'. In: *Biometrika* 58.2 (1971), pp. 341–348.

[170] Student. 'The probable error of a mean'. In: *Biometrika* (1908), pp. 1–25.

[171] Frank Wilcoxon. 'Breakthroughs in statistics'. In: *Individual comparisons by ranking methods* (1992), pp. 196–202.

[172] Fausto Milletari, Nassir Navab and Seyed-Ahmad Ahmadi. 'V-net: Fully convolutional neural networks for volumetric medical image segmentation'. In: *2016 fourth international conference on 3D vision (3DV)*. Ieee. 2016, pp. 565–571.

[173] Fabian Isensee et al. 'nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation'. In: *Nature methods* 18.2 (2021), pp. 203–211.

[174] Shruti Jadon. 'A survey of loss functions for semantic segmentation'. In: *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*. IEEE. 2020, pp. 1–7.

[175] Seyed Sadegh Mohseni Salehi et al. 'Real-time automatic fetal brain extraction in fetal MRI by deep learning'. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018. URL: https://doi.org/10.1109%2Fisbi.2018.8363675.

[176] Danyang Xiao, Chengang Yang and Weigang Wu. 'Efficient DNN training based on backpropagation parallelization'. In: *Computing* 104.11 (2022), pp. 2431–2451.

[177] Ilya Loshchilov and Frank Hutter. 'Decoupled weight decay regularization'. In: *arXiv preprint arXiv:1711.05101* (2017).

[178] Kasper Marstal et al. 'SimpleElastix: A user-friendly, multi-lingual library for medical image registration'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2016, pp. 134–142.

[179] Brian B Avants, Nick Tustison, Gang Song et al. 'Advanced normalization tools (ANTS)'. In: *Insight j* 2.365 (2009), pp. 1–35.

[180] Jiahao Pang et al. 'Cascade residual learning: A two-stage convolutional neural network for stereo matching'. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017, pp. 887–895.

[181] Eddy Ilg et al. 'Flownet 2.0: Evolution of optical flow estimation with deep networks'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2462–2470.

[182] Shengyu Zhao et al. 'Recursive cascaded networks for unsupervised medical image registration'. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 10600–10610.

[183] Evan G Hemingway and Oliver M O'Reilly. 'Perspectives on Euler angle singularities, gimbal lock, and the orthogonality of applied forces and applied moments'. In: *Multibody System Dynamics* 44 (2018), pp. 31–56.

[184] Diederik P Kingma and Max Welling. 'Auto-encoding variational bayes'. In: *arXiv preprint arXiv:1312.6114* (2013).

[185] Lin Zhang et al. 'FSIM: A feature similarity index for image quality assessment'. In: *IEEE transactions on Image Processing* 20.8 (2011), pp. 2378–2386.

[186] Umme Sara, Morium Akter and Mohammad Shorif Uddin. 'Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study'. In: *Journal of Computer and Communications* 7.3 (2019), pp. 8–18.

[187] Quang-Khai Tran and Sa-kwang Song. 'Computer vision in precipitation nowcasting: Applying image quality assessment metrics for training deep neural networks'. In: *Atmosphere* 10.5 (2019), p. 244.

[188] Guangtao Zhai and Xiongkuo Min. 'Perceptual image quality assessment: a survey'. In: *Science China Information Sciences* 63 (2020), pp. 1–52.

[189] Alain Hore and Djemel Ziou. 'Image quality metrics: PSNR vs. SSIM'. In: *2010 20th international conference on pattern recognition*. IEEE. 2010, pp. 2366–2369.

[190] Solomon Kullback and Richard A Leibler. 'On information and sufficiency'. In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.

[191] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[192] Lu Mi, Macheng Shen and Jingzhao Zhang. 'A probe towards understanding gan and vae models'. In: *arXiv preprint arXiv:1812.05676* (2018).

[193] Chenrui Zhang and Yuxin Peng. 'Stacking VAE and GAN for context-aware text-to-image generation'. In: *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*. IEEE. 2018, pp. 1–5.

[194] Jialun Peng et al. 'Generating diverse structure for image inpainting with hierarchical VQ-VAE'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 10775–10784.

[195] Roy A Filly, Ruth B Goldstein and Peter W Callen. 'Fetal ventricle: importance in routine obstetric sonography.' In: *Radiology* 181.1 (1991), pp. 1–7.

[196] E Quarello et al. 'Assessment of fetal Sylvian fissure operculization between 22 and 32 weeks: a subjective approach'. In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 32.1 (2008), pp. 44–49.

[197] Nicola K Dinsdale, Mark Jenkinson and Ana IL Namburete. 'Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal'. In: *NeuroImage* 228 (2021), p. 117689.