Quantification of Fetal Brain Development from Ultrasound Images using Interpretable Deep Learning



Linde Sterre Hesse

St. Edmund Hall

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2023

"The journey never ends"
José Saramago

Acknowledgements

This thesis would not have been completed without the immense support I have received during my time in Oxford. There is no space to thank everyone by name, but to all those that supported me over the years, know that I am forever grateful.

Firstly, Ana, thank you for convincing me to do a DPhil in the first place and motivating me to apply to Oxford. I surely would not have come here without you. Also, my greatest gratitude for all the help, supervision and motivation you have given me along the way. Weidi and Mark, thank you so much for stepping in when Ana was on leave and guiding me through a challenging Covid year. Alison, thank you for your support in the final stages of this DPhil. Monique and Moska, my thanks for always providing me with advice from a clinical perspective.

To all the members of the OMNI group, you have made my DPhil so much more fun. Thank you for all your advice, insights and company. Special thanks to Hugo, Nicola and Maddy for proofreading large parts of this thesis. Felipe, thank you for always helping me out with my computer whenever I managed to break it. Maddy, you have kept me sane throughout the process of this thesis. I could not have done it without our morning coffee routine, 12 o'clock lunches and most of all, always having someone to chat to in the office.

Women of OxWoCS, you have given me a community in Oxford I could have only dreamed of. Spending time with you gave me the sometimes much-needed insight that there is more in life than DPhil work, and it has been an honour to work with every single one of you. Special shout-out to Catherine, Jimmy and Eleonora for making Covid times slightly more fun, and Nele, Ruby, Maddy, Yana and Kelsey for putting in all the hard work to continue the society.

To my other friends in Oxford, Karen, I am so thankful we met on one of our first days in Oxford. I have thoroughly enjoyed doing our DPhil together and will miss our regular library sessions (although I am mostly relieved this thesis is now almost completed). Lisa, meeting you has made my last two years in Oxford so much more enjoyable. Elisa, thank you for all the game nights and dinners, and the (maybe less successful) plant sitting.

To my friends at home, I might have missed many birthdays, parties and dinners, but I always enjoyed coming home and spending time with you. A special thank you to my best friend Rosa, for always being there for me.

My biggest thanks to Youp for always believing in me, and having my back no matter what. You were always able to cheer me up when this thesis brought me, at times, close to tears. I can not thank you enough for your endless support, the proofreading sessions and all the dinners you cooked for me, especially in the final stages of thesis writing.

And finally, to my family, Mum, Dad, and Tim, thank you for always supporting me and encouraging me to continue learning. Your love and support have made this journey possible, and I could not have done this without you.

I am grateful for the funding I received during my DPhil from the Engineering and Physical Sciences Research Council (EPSRC), and from the Microsoft Research Fellowship.

Abstract

Ultrasound images of the fetal brain are routinely acquired during pregnancy to assess the health and development of the fetus. It is standard clinical practice to obtain simple in-plane measurements in 2D images that can not capture the complex structural development of the fetal brain during gestation. Therefore, in this thesis, I propose deep learning-based methods that can improve the understanding of brain development from ultrasound.

Firstly, I studied the use of deep learning for subcortical structure segmentation. Deep learning models typically need a reasonably large number of samples to effectively learn the task. However, as subcortical structure segmentation is not a task typically performed in clinic, it is challenging to obtain large sample numbers with pixel-wise annotations. For this reason, I explored subcortical segmentation in a low-data regime, demonstrating that segmentation performance close to intra-observer variability can be obtained with only a handful of manual annotations. The developed segmentation models were then applied to a large number of volumes of a diverse, healthy population, generating ultrasound-specific growth curves of subcortical development.

Predicting the gestational age of a fetus based on brain morphology can also be used as a way to quantify developmental patterns. While conventional deep learning methods can be used for this task, they can typically not explain their reasoning process or provide insight into the image parts that contributed to the final prediction. However, for clinical applications, it is vital to understand model behaviour to identify failure modes and gain patients' trust. For this reason, I developed an age prediction model for fetal ultrasound that incorporates *guided attention* in the architecture to make *interpretable* and *local* brain age predictions. The attention is regularised with a segmentation loss, enforcing the network to focus on specific parts of the image. I demonstrate that guiding the network to focus on age-discriminative regions (the cortical plate and cerebellum) results in significantly improved prediction performance.

Finally, I propose an alternative approach to interpretable brain age prediction that uses an inherently interpretable network, as opposed to a *post-hoc* explanation. The network learns a set of representative examples from the training set (*prototypes*) and predicts the age of a new sample based on the distances

to these prototypes. The image-level distances are constructed from patch-level distances, which are structurally matched using optimal transport. The prototypes and distance computations can both be visualised, providing an understanding of the reasoning process of the model.

Contents

List of Figures **xi**

List of Tables **xiv**

List of Abbreviations **xv**

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Contributions . . . 3
- 1.3 Thesis Outline . . . 4
- 1.4 Statement of Originality . . . 5
- 1.5 List of Publications . . . 7

2 Ultrasonic Evaluation of the Developing Fetal Brain **10**
- 2.1 Fetal Brain Ultrasound . . . 10
  - 2.1.1 Characteristics of Fetal Brain Ultrasound . . . 10
  - 2.1.2 Appearance of the Fetal Brain . . . 12
  - 2.1.3 US Examinations in Clinic . . . 15
  - 2.1.4 Pregnancy Dating . . . 17
- 2.2 Automated Fetal Brain Analysis . . . 18
  - 2.2.1 Subcortical Structure Segmentation . . . 19
  - 2.2.2 Brain Age Prediction . . . 24
- 2.3 Conclusion . . . 30

3 Interpretable Deep Learning **31**
- 3.1 Visual Explanations . . . 34
  - 3.1.1 Gradient-based . . . 35
  - 3.1.2 Class Activation-based . . . 41
  - 3.1.3 Perturbation-based . . . 43
  - 3.1.4 Trainable Attention . . . 45
  - 3.1.5 Local Function Approximation . . . 46
  - 3.1.6 Limitations of Visual Attributions . . . 46
- 3.2 Example-based Explanations . . . 49

3.2.1 Prototype Learning 50
3.2.2 Counterfactuals 52
3.2.3 Distance-based 53
3.2.4 Limitations of Example-based Explanations 55
3.3 Interpretability for Brain Age Predictions 56

4 Datasets **58**
4.1 INTERGROWTH-21st Dataset 58
4.1.1 Previous Data Processing 59
4.1.2 Selection in this Thesis 61
4.1.3 Additional Preprocessing 62
4.2 LUMC Dataset 64
4.3 IXI MRI Dataset 66

5 Subcortical Segmentation of the Fetal Brain **67**
5.1 Introduction 68
5.2 Methods 71
5.2.1 Notation 71
5.2.2 Network Architecture 72
5.2.3 Few-Shot Training Labels 72
5.2.4 Manual Annotation Process 76
5.3 Experimental Set-Up 78
5.3.1 Dataset and Preprocessing 78
5.3.2 Experiments 79
5.3.3 Implementation 82
5.3.4 Evaluation 82
5.3.5 Volumetric Subcortical Growth Trajectories 84
5.4 Results 85
5.4.1 Weak Atlas Labels 85
5.4.2 Comparison between Atlas and Expert Labels 85
5.4.3 Varying the Number of Manual Annotations 86
5.4.4 Effect of Image Alignment 87
5.4.5 Comparison to Previous Work 88
5.4.6 Qualitative Results 91
5.4.7 Clustered Template Labels 91
5.4.8 Subcortical Growth Curves 92
5.4.9 Generalisability 94
5.5 Discussion 98

6 Local Brain Age Prediction using Guided Attention 100
6.1 Introduction . . . 101
6.2 Review of attention methods . . . 104
6.2.1 Attention Gates . . . 104
6.2.2 Guided Attention Gates . . . 106
6.3 Methods . . . 107
6.3.1 Problem Definition . . . 107
6.3.2 Network Architecture . . . 107
6.3.3 Attention . . . 109
6.3.4 Network Training . . . 111
6.4 Experimental Set-up . . . 112
6.4.1 Dataset . . . 112
6.4.2 2D Experiments . . . 114
6.4.3 3D Experiments . . . 116
6.4.4 Visualisation of Attention Maps . . . 116
6.4.5 Implementation . . . 117
6.5 Results . . . 118
6.5.1 2D Standard Planes . . . 119
6.5.2 3D Image Volumes . . . 128
6.5.3 Generalisability . . . 134
6.6 Discussion . . . 135

7 Interpretable Brain Age Prediction using Prototype Learning 138
7.1 Introduction . . . 139
7.2 Optimal Transport . . . 141
7.2.1 Assignment Problem . . . 142
7.2.2 Monge Formulation . . . 143
7.2.3 Kantorovich Relaxation . . . 145
7.2.4 Entropic Regularisation . . . 145
7.2.5 Optimal Transport in CNNs . . . 146
7.3 Methods . . . 147
7.3.1 Network Architecture . . . 148
7.3.2 Prototype Projection . . . 148
7.3.3 Distance Metric Loss . . . 149
7.3.4 Patch-based Distance Metric . . . 150
7.3.5 Consistency Loss . . . 152
7.3.6 Inference . . . 152
7.4 Experimental Set-Up . . . 153
7.4.1 Datasets . . . 153

7.4.2 Implementation 155
7.5 Results 156
7.5.1 Fetal Ultrasound 156
7.5.2 Adult MRI 162
7.6 Discussion 164

8 Conclusion **167**
8.1 Contributions 167
8.2 Limitations and Future Work 169
8.2.1 Datasets 170
8.2.2 Generalisability 171
8.2.3 Interpretability 173
8.2.4 Future Applications 174

Appendices

A INTERGROWTH-21st Data Selection Criteria **177**

B Additional Results Subcortical Segmentation **178**

C Optimal Transport Derivation **181**

Bibliography **183**

List of Figures

2.1 US acquisition example 11
2.2 Fissures of interest on the CoP 13
2.3 Example US image of the fetal brain at 22 GWs. 14
2.4 2D US standard planes for second-trimester scan. 15
2.5 Measurements to estimate US-based GA 18

3.1 Categorisation of interpretability methods reviewed in this chapter. 33
3.2 Schematic of different visual attribution based methods 36
3.3 Overview of ReLU on the backward pass for three gradient-based explanations 37
3.4 Overview of CAM explanations 42
3.5 Overview of perturbation-based explanations 44
3.6 Schematic overview of metric and prototype learning for categorical and continuous labels. 54

4.1 Data distributions of the INTERGROWTH-21st and LUMC datasets 60
4.2 Example US volumes over the GA range before size normalisation . 63
4.3 Example transventricular planes extracted from the extended dataset, and from the LUMC dataset 64
4.4 Example US volumes over the GA range from the LUMC dataset . 66
4.5 Example slices extracted from the 3D MR brain images in the IXI dataset 66

5.1 Schematic overview of segmentation pipeline 71
5.2 Overview of the 3D U-Net architecture used in this study. 73
5.3 Schematic overview of atlas label generation 75
5.4 Data distribution of the 537 US images used in this chapter. 78
5.5 Resulting performance values after post-processing for θ^{exp} and θ^{atl} 87
5.6 Varying the number of manual annotations (n_a) used for generating the training labels. 88
5.7 Example prediction of a test set volume at 22 GW. 89
5.8 A set of examples showing the most prominent segmentation errors. 90

5.9 Comparison of growth curves obtained for the CB in this study to previous work. . . . 94
5.10 Estimated structural volumes for subcortical structures as a function of GA . . . 95
5.11 Estimated relative volumes with respect to the whole brain volumes for subcortical structures as a function of GA . . . 96

6.1 Three different attribution-based explanation methods explaining the prediction of the same model for a single test set sample. . . . 102
6.2 Schematic overview of a AG. . . . 104
6.3 Schematic overview of the attention mechanism proposed by Jetley et al. [208] . . . 105
6.4 ResNet backbone encoder used for brain age predictions. . . . 108
6.5 Overview of different AG configurations. . . . 109
6.6 Data distribution of the 3359 INTERGROWTH-21st US volumes used in this chapter. . . . 113
6.7 Example of segmentation quality at 28 GWs. . . . 114
6.8 Comparison of the different attention architectures. . . . 118
6.9 Overview of the performance for different inputs to the AGs. . . . 120
6.10 Overview of average performance across 5 folds for different weights of the DSC in the training loss. . . . 121
6.11 Brain age predictions of the single best fold on the 2D images using guided attention. . . . 123
6.12 DSC segmentation performance of the learned attention mask with ground-truth CoP segmentations grouped by GW. The DSC performance is relatively stable accross gestation, showing a slow decline in performance at the end of the GA range. . . . 124
6.13 Qualitative results of guided attention for three test set images. . . . 126
6.14 Average masked feature representations per GW for 2D images. . . . 126
6.15 Average masked feature representations per GW for the network trained without segmentation loss. . . . 127
6.16 Averaged attribution maps per GW for three different saliency methods. 128
6.17 Hyperparameter tuning for the attention weight for the 3D volumes with the CP as ROI. . . . 129
6.18 True versus predicted GA for 3D models trained with different segmentation masks. . . . 132
6.19 Averaged masked feature representations per GW for the 3D model trained with the CoP as ROI. . . . 133
6.20 Average masked feature representations ($\hat{\mathbf{Z}}$) per GW for the 3D model trained with the CoP as ROI shown as 3D visualisations. . . . 133

6.21 Masked feature representation ($\hat{\mathbf{Z}}$) of three individual test set volumes trained with the CB as ROI 134

7.1 Schematic overview of the assignment problem. 143
7.2 Schematic overview of the Monge-Kantorovich problem. 144
7.3 Schematic of the decision space of Sinkhorn optimisation solving the OT problem. 147
7.4 Schematic overview of the *PRINT* architecture. 149
7.5 Overview of consistency loss between an image pair 151
7.6 Schematic overview of model inference of *PRINT*. 153
7.7 Overview of the INTERGROWTH-21st data across GA used in this chapter. 154
7.8 Overview of the MRI data across age used in this chapter. 155
7.9 Predicted versus ground-truth age 158
7.10 Example predictions of PRINT model for US and MRI 159
7.11 Illustration of OT matching between image sample and prototypes for US 160
7.12 Overview of hyperparameter tuning for the US dataset 161
7.13 Illustration of OT matching between image sample and prototypes for MRI 164

A.1 Inclusion criteria for the data selection of the INTERGROWTH-21st dataset 177

B.1 Segmentation performance when varying the weighing parameter between the $\mathcal{L}_{MD}$ and $\mathcal{L}_{C}E$. 179
B.2 Signed and unsigned relative volume differences after post-processing. 179
B.3 Signed and unsigned relative volume differences before the largest connected component post-processing. 180
B.4 Resulting performance values before the largest connected component post-processing. 180

List of Tables

2.1 Overview of existing automated subcortical segmentation methods for fetal brain US. 23
2.2 Overview of automated brain age prediction methods for fetal US . 27

4.1 Overview of images of the INTERGROWTH-21st dataset used in each contribution chapter. 62
4.2 Overview of images of the LUMC dataset used in each contribution chapter. 65

5.1 Overview of manual segmentation protocol. 77
5.2 Overview of the number of images and available labels for each of the data subsets in this chapter. 80
5.3 DSC performance of the segmentation networks using aligned images compared to results of previous studies 86
5.4 LPVH segmentation performance using two different types of atlas-based annotations 92
5.5 Segmentation results on LUMC test set 98

6.1 Performance on the test set with 2D images for the guided attention and several baselines. 121
6.2 Performance of 3D models on test set compared to several baselines. 130
6.3 Age prediction performance on an independent test set (LUMC). . . 135

7.1 Quantitative results for fetal brain age prediction from US on the test set. 157
7.2 Results of the *PRINT* model on the LUMC dataset without any fine-tuning. 162
7.3 Quantitative results for MRI brain age predictions on the test set. . 163

B.1 P-values obtained from statistical comparison of segmentation networks. 178

List of Abbreviations

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

AC . . . . . Abdominal circumference

BPD . . . . . Biparietal diameter

CaF . . . . . Calcarine fissure

CB . . . . . Cerebellum

CM . . . . . Cisterna magna

CNN . . . . . Convolutional neural network

CNS . . . . . Central nervous system

CoP . . . . . Cortical plate

CP . . . . . Choroid Plexus

CRL . . . . . Crown-rump length

CSPV . . . . . Cavum septum pellucidum et vergae

DeepLIFT . . Deep Learning Important FeaTures

FC . . . . . Fully connected

FGLS . . . . . Fetal Growth Longitudinal Study

FL . . . . . Femur length

GA . . . . . Gestational age

GAP . . . . . Global Average Pooling

GBP . . . . . Guided backpropagation

GW . . . . . Gestational week

HC . . . . . Head circumference

IntGrad . . . . Integrated gradients

IVF . . . . . In-vitro fertilisation

LIME . . . . . local interpretable model-agnostic explanations

LMP . . . . . Last menstrual period

LPVH . . . . . Lateral posterior ventricle horn

LRP . . . . . Layer-wise relevance propagation

LUMC . . . . Leiden University Medical Centre

MAE . . . . . Mean absolute error

MRI . . . . . Magnetic resonance imaging

MSE . . . . . Mean squared error

SF . . . . . Sylvian fissure

TCD . . . . . Transverse cerebellar diameter

OT . . . . . Optimal transport

POF . . . . . Parieto-occipital fissure

ReLU . . . . . Rectified linear unit

RMSE . . . . . Root mean squared error

ROI . . . . . Region of interest

SSD . . . . . Sum-of-squared-difference

US . . . . . Ultrasound

1 Introduction

1.1 Motivation

Pregnancies are routinely monitored using ultrasound (US) imaging due to its low cost and high availability. During these US examinations, the development of the fetal brain is assessed to measure growth and detect potential central nervous system (CNS) abnormalities [1]. However, this assessment is typically performed with a number of 2-dimensional (2D) in-plane measurements, which only provide a limited representation of the 3-dimensional (3D) structural development of the brain. In particular, intricate and dynamically evolving structures like the cortical plate (CoP) and choroid plexus (CP) are challenging to capture using conventional in-plane measurements. For this reason, the development of novel automated methods could facilitate a more in-depth analysis of fetal brain US, and may provide new insights into *in utero* brain development.

The subcortical brain structures, such as the CP and the cerebellum (CB), are typically included in routine and advanced examinations of the fetal brain [1, 2], indicating their importance in assessing healthy development. Their full structural development can be quantified by manually segmenting the structures of interest in 3D US, but this is time-consuming and labour-intensive and, therefore, not feasible in clinical practice. While deep learning techniques have been shown

to produce accurate segmentations in a range of medical imaging applications [3], such methods typically require a large number of manual annotations for training. As such annotations are not readily available for 3D US and are costly to acquire, this poses a key barrier to applying deep learning methods to subcortical structure segmentation in US. Furthermore, US images pose a number of unique characteristics that make the task especially challenging, such as the presence of shadowing artefacts and the free-hand nature of the acquisition, resulting in an unpredictable location of the fetal head in the acquired images. A number of previous works have studied automating subcortical segmentations in 3D US [4–8], but these all have their limitations, such as segmenting only a single structure [4, 6] or having a limited search region [5]. There is thus a need for automated methods that segment multiple subcortical structures from 3D whole brain US.

An alternative to structural segmentation is the use of brain age predictions to assess the development of the brain [9, 10]. In a healthy population, the development of the brain is found to be temporally and spatially consistent during gestation [11–20]. For this reason, gestational age (GA) predictions based on brain morphology can be used to quantity the stage of development. This prediction can then be compared to the ground-truth GA to determine whether the development is age appropriate. As ground-truth GA estimates are readily available in the clinic, this overcomes the difficulty in obtaining additional manual annotations. However, a major limitation of applying conventional deep learning-based models for the task of brain age prediction is that their decision-making process is not transparent [9, 10]. This makes it challenging to understand model behaviour and limits the additional insights that can be obtained alongside the model predictions [21, 22]. For example, a convolutional neural network (CNN) trained on the task of brain age prediction from US can use image features extracted from anywhere in the image, and it is not trivial to determine which image features contributed most to the prediction. This severely limits the practical use of such a model, as no meaningful conclusions about *local* structural development can be drawn.

It is thus vital to develop brain age prediction models that are *interpretable* and, hence, capable of providing insight into their reasoning process. This is essential for the successful translation of the developed model into clinical practice, as both healthcare practitioners and patients need to be able to understand *why* a certain prediction was made to build trust in the model. Additionally, interpretability can be essential to detect failure modes during model development [23], and can provide useful information on how to improve model design before deployment [22].

1.2 Contributions

This thesis proposes three deep learning-based methods to automate the analysis of fetal brain US. These methods make the analysis of large numbers of acquired images feasible and can, therefore, improve our understanding of brain development *in utero*. Furthermore, by developing deep learning models that are *interpretable* (Chapters 6 and 7), the proposed models can provide information on which image features were used, and how these features were used to generate the prediction.

To quantify subcortical development, I develop a method for the automated segmentation of several subcortical structures in Chapter 5 (*Contribution 1*). As obtaining manual annotations for training is costly, I specifically study training CNNs for this task in a low-data regime. I compare using a small number of individually annotated volumes versus using many weakly labelled volumes obtained from atlas-based segmentations. Due to the free-hand acquisition of fetal US images, the fetal brain position varies strongly across the acquired volumes: thus, I also study the impact of image alignment to a common coordinate system as a preprocessing step. Furthermore, I apply the final trained models to a cohort of healthy fetuses, obtaining US-specific growth curves for subcortical growth during the second trimester of gestation.

In Chapter 6, I then address the question of identifying *which* image features are used by fetal brain age prediction models (*Contribution 2*). I propose a network that incorporates *guided attention* in the architecture, which can not only identify relevant features but can also regularise the features that are used by the network.

By regularising the networks' attention with segmentation masks of the CB and CoP, I explore how the brain age prediction network can be enforced to focus on certain image areas and show that this regularisation even boosts prediction performance. In this chapter, a number of different attention architectures are investigated and experiments with both 2D and 3D US images are presented.

Finally, in Chapter 7, I propose a novel approach for interpretable brain age predictions that addresses *how* image features are used to generate a prediction (*Contribution 3*). The method is based on prototype learning and provides an example-based explanation, specifically for a continuous prediction (i.e. brain age). The proposed model makes a sample prediction from the distances to a set of learned *prototypes* in latent space, using a weighted mean of prototype labels. The prototypes are enforced to equate samples from the training set, and can, therefore, be easily visualised. The image-level distances are further constructed from patch-level distances, which are structurally matched using optimal transport (OT). During inference, the prototypes used for the prediction can be shown, and for each image-prototype distance, a detailed patch-level matching can be provided. In this chapter, the network is applied to 2D US images of the fetal brain, as well as to 2D magnetic resonance (MR) images of the adult brain, demonstrating that state-of-the-art performance can be obtained on both datasets with an interpretable CNN, thus providing insight in the reasoning process.

Automated methods hold the promise that they make the analysis of large amounts of data feasible. However, large US datasets are often acquired at multiple institutions or with different US scanners. For this reason, it is essential that the models can generalise well to new datasets. In this thesis, I demonstrate the generalisability of the developed models by testing the performance in each chapter on an independent test set acquired at a different site and with a different scanner, and show that good performance can be obtained on these US images.

1.3 Thesis Outline

This thesis is structured as follows:

Chapter 2 describes the characteristics of fetal brain US and how it is used to monitor brain development during gestation in clinical practice. It also contains a review of automated methods that have been developed in the past for the segmentation of subcortical structures and brain age predictions from fetal US images.

Chapter 3 provides an overview of the different types of interpretability methods used in medical imaging and summarises their respective benefits and limitations. Lastly, it also reviews previous studies using interpretable deep learning specifically developed for the task of brain age prediction.

Chapter 4 introduces the datasets used in this thesis and describes the main data selection and preprocessing steps.

Chapters 5–7 present the three contributions previously introduced. Each of these chapters will have a similar structure, starting with an introduction and a brief overview of any related literature that has not been discussed in Chapters 2 and 3. This will be followed by the methods and experimental set-up, describing the proposed methods in detail. Finally, the results will be presented and discussed, followed by a final conclusion.

Chapter 8 will summarise the main contributions and conclusions presented in this thesis. Furthermore, it will discuss the overall limitations of the work presented in this thesis and provide a number of recommendations for further work.

1.4 Statement of Originality

I declare that all the work presented in this thesis is my individual work unless stated otherwise. All figures and tables included in this thesis have been produced by me, or accurately referenced when other sources were used. Comments on the work were provided by Prof. Ana Namburete and by group members of

the OMNI lab. Additional comments on Chapter 5 were given by Prof. Mark Jenkinson and Prof. Weidi Xie.

Most of the data used in this thesis has been acquired previously, and some data selection and preprocessing steps have been adopted from earlier works. This is described in more detail in Chapter 4, but can be summarised as follows:

- The main US dataset used in this thesis was previously acquired as part of the INTERGROWTH-21st study [24]. Initial data selection and alignment to a common reference space were carried out by Prof. A.N., which will also be clearly stated in the data description Chapter 4. All manual ground-truth annotations of the subcortical structures used in Chapter 5, both on the atlas and individual images, were done by me.

- A second independent dataset was acquired during my DPhil by collaborators at the Leiden University Medical Centre (LUMC). I have performed all preprocessing for this dataset, and have created a data selection protocol together with Dr. Moska Aliasi, an ultrasonographer in the LUMC. Data selection was subsequently carried out by her.

- The final dataset used in this thesis consists of adult brain MR images and is only used in Chapter 7. This is a publicly available dataset, and the preprocessing of this dataset was carried out by Dr Nicola Dinsdale.

In addition to the dataset preprocessing, Chapter 5 uses an atlas constructed by A.N. in earlier work [25]. To propagate manual annotations on the atlas back to the individual volumes, and to reconstruct the atlas with a subset of volumes, I adapted code previously written by A.N. Lastly, the work in Chapter 6 is inspired by initial collaborative work between me and Dr Madeleine Wyburd, predicting local brain age from cropped brain images [26]. However, the work presented in this chapter is follow-up work representing my own contribution and does not contain collaborative work. Only the ground-truth annotations of the CoP used in this chapter were produced by M.W. (using an automated segmentation method

previously developed [27]), and she also assisted in creating the 3D visualisations of the CoP presented in Fig. 6.20.

1.5 List of Publications

A list of publications during my DPhil is listed below, followed by a description of how the publications relate to each of the contributions presented in this thesis.

Peer-Reviewed Publications

(i) **Linde S. Hesse**, Ana I. L. Namburete. "INSightR-Net: Interpretable Neural Network for Regression Using Similarity-Based Comparisons to Prototypical Examples" *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, 2022.

(ii) **Linde S. Hesse**, Moska Aliasi, Felipe Moser, Monique C. Haak, Weidi Xie, Mark Jenkinson, Ana IL Namburete, and the INTERGROWTH-21st Consortium. "Subcortical segmentation of the fetal brain in 3D ultrasound using deep learning." *NeuroImage*, 2022.

(iii) **Linde S. Hesse**[1], Madeleine K. Wyburd[1], Moska Aliasi, Mark Jenkinson, Aris T. Papageorghiou, Monique C. Haak and Ana I. L. Namburete "Assessment of Regional Cortical Development Through Fissure Based Gestational Age Estimation in 3D Fetal Ultrasound." *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis. Workshop at MICCAI 2021.* Springer, Cham, 2021.

(iv) **Linde S. Hesse**, and Ana I. L. Namburete. "Improving U-Net Segmentation with Active Contour Based Label Correction." In *Annual Conference on Medical Image Understanding and Analysis*, Springer, Cham, 2020,

[1]Indicates equal contribution

Preprints

(v) **Linde S. Hesse**, Nicola K. Dinsdale and Ana I. L. Namburete. "Prototype Learning for Explainable Regression," available on ArXiv: arXiv:2306.09858

Additional Conference Presentations

(vi) **Linde S. Hesse**[1], Madeleine K. Wyburd[1], Moska Aliasi, Mark Jenkinson, Aris T. Papageorghiou, Monique C. Haak and Ana I. L. Namburete "Automated grading of fetal cortical development using deep learning algorithms: a preliminary three-dimensional ultrasound study in healthy fetuses." *Ultrasound in Obstetrics & Gynecology,* 2021 (oral presentation)

(vii) Moska Aliasi, **Linde S. Hesse**, Madeleine K. Wyburd, Maartje Snoep, Ana I. L. Namburete, and Monique C. Haak. "Success rate of the subcortical 3D fetal brain segmentation using deep learning algorithms." *Ultrasound in Obstetrics & Gynecology,* 2022 (oral presentation)

(viii) **Linde S. Hesse**, Moska Aliasi, Monique C. Haak, Aris T. Papageorghiou, Weidi Xie, Mark Jenkinson and Ana IL Namburete "Few-shot subcortical brain structure segmentation in 3D fetal brain ultrasound." *Organisation of Human Brain Mapping (OHBM),* 2020

Relation to contributions The publications relate to each of the contributions in this thesis as follows:

- Contribution 1 (Chapter 5): The work in (ii) constitutes the first contribution of this thesis, and (viii) a technical conference presentation by me of the same work. The precursor to the work presented in this chapter was published in (iv). Lastly, (vii) contains a clinical conference presentation in collaboration with the LUMC, validating the segmentation performance on their data.

[1]Indicates equal contribution

- Contribution 2 (Chapter 6): Both (iii) and (vi) contain collaborative work between me and my colleague Dr M.W. which motivated the work presented in Chapter 6. However, the work presented in this chapter is distinct from the previously published work in (iii) and (vi).

- Contribution 3 (Chapter 7): The work presented in this chapter has been written into an article (preprint available on ArXiv (v)). Precursors to this work on a different dataset were published and presented at MICCAI 2022 (i).

2

Ultrasonic Evaluation of the Developing Fetal Brain

2.1 Fetal Brain Ultrasound

Pregnancies are routinely monitored with US examinations as it is cost-efficient, portable, and can acquire images in real-time. This section will provide an overview of the characteristics of fetal brain US, the developmental patterns that can be observed, and the types of US-based examinations that are performed. Additionally, it will provide an overview of how the due date is typically estimated in clinical practice, which is referred to as *pregnancy dating*.

2.1.1 Characteristics of Fetal Brain Ultrasound

US is acquired with a probe that emits a beam of sound waves that interacts with the different tissues in the human body. At tissue boundaries, part of the beam is reflected back to the probe, which can be measured and used to create an US image. The pattern at a certain boundary is determined by the difference in acoustic impedance between the tissues, meaning that boundaries between tissues with large differences in acoustic impedance, such as the skull and the surrounding fluid, are clearly visible on the image. Soft tissues on the other hand, which typically have a similar acoustic impedance as they mainly consist of water, show relatively low

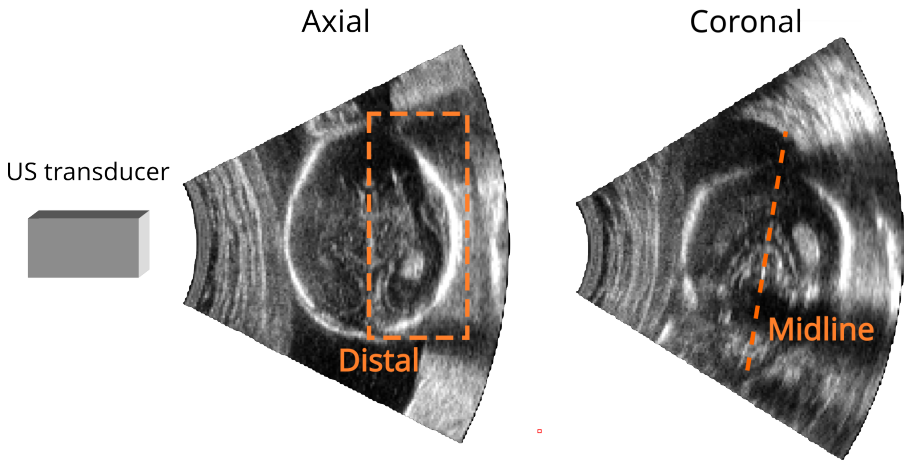


Figure 2.1: Example of US acquisition of the fetal brain. It can be seen that only the hemisphere distal to the probe is well visible (delineated with a dashed line). The other hemisphere is referred to as the proximal hemisphere. On the coronal plane, the approximate midline of the brain has been drawn, from which it can be observed that the orientation of the images can vary.

contrast on US. Furthermore, the appearance of a boundary depends on several factors, such as the angle of incidence, the distance to the probe and the tissue boundaries it has already encountered (reflecting part of the beam at each surface). For this reason, the boundaries seen in US images are typically discontinuous and dependent on the location and orientation of the image. In Fig. 2.1 an example US image is shown, illustrating the appearance of the fetal brain on US.

A typical characteristic of US is the presence of speckle. This refers to the granular textural patterns present throughout the whole image, and is a result of interference of the sound waves with each other [28]. While speckle is dependent on the tissue type and can therefore contain information, it is typically considered noise as it obscures anatomical details, especially of small structures. Another characteristic more specific to US images of the fetal brain is the shadowing of the fetal skull. The skull increasingly calcifies during gestation which makes it more difficult for the US beam to penetrate the skull. This is especially pronounced in the third trimester, where the deeper brain structures are often barely visible. Due to the strong acoustic shadowing of the skull, only the hemisphere distal to the US probe is well visible on scans acquired from the axial plane, as shown in Fig. 2.1.

Analysis of the fetal brain in US is therefore usually limited to a single hemisphere.

Finally, a key characteristic of fetal brain US acquired with a free-hand scanning protocol, which is typical at the bedside, is the varying position of the fetal brain due to the unpredictable fetal position in the womb, as well as the movement of the transducer relative to the fetal head. An example of this characteristic is shown in Fig. 2.1 in the coronal plane, where it can be seen that the midline is not well aligned (i.e. to an orthogonal reference system). As a result of the fetal position, the acquired visible hemisphere also varies across individuals.

In routine clinical practice, typically transabdominal 2D US is used to analyse fetal health. However, many modern US devices also have 3D probes, enabling the acquisition of 3D image volumes. These enable full volumetric analysis and facilitate the visualisation of fully orthogonal 2D planes. In clinical practice, 3D US of the fetal brain is recommended for the targeted neurosonography [2], which is performed when there is a suspicion of CNS abnormalities.

2.1.2 Appearance of the Fetal Brain

The description of fetal brain development is often divided into two parts: *cortical* and *subcortical* development. Cortical development is the development of the cortex, which, in its immature state, is referred to as the *cortical plate* (CoP), and subcortical development, which is the development of all structures encapsulated in the CoP, such as the CB, ventricular system and brainstem.

Cortical Development The cerebral cortex, which, in the adult brain, is heavily folded, begins as a smooth surface and gradually develops *sulci* (troughs) and *gyri* (peaks) during the fetal period. This process is known as *gyrification* and allows the cortical surface area to almost scale linearly with brain volume [29], whilst being contained in the skull, and is often linked to human intelligence [30]. Larger sulci can be referred to as *fissures*, and the emergence of such fissures is temporally and spatially consistent across healthy individuals. It has been well-timetabled with post-mortem studies [11–13] as well as with MRI [14–17] and US-based [18, 19] *in*

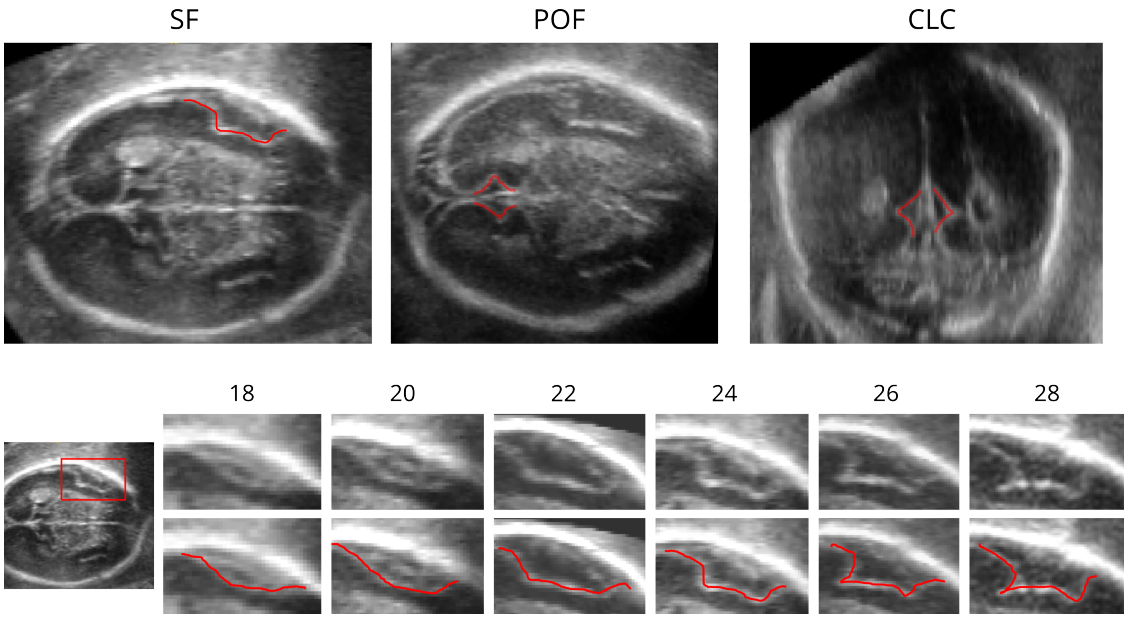


Figure 2.2: Fissures of interest on the CoP. Top: Examples of the SF and POF in the axial plane, and the CaF in the coronal plane. Bottom: SF development during the second trimester of gestation. The SF has been drawn onto the images in red in the bottom row. It can be seen that the SF becomes deeper and more pronounced during the second trimester, and starts operculising (i.e. folding over itself) around 26 GWs.

utero observations. For example, the Sylvian Fissure (SF) emerges around 14 weeks of gestation [31], and is first visible as a shallow indentation. With advancing GA the fissure starts folding over itself, referred to as *operculisation*, becoming narrower and deeper, until it eventually appears as a fully branched structure [32]. An example of the SF appearance on US between 18 and 28 gestational weeks (GWs) is shown in Fig. 2.2. Other fissures of interest that can be observed well on US during the second trimester are two midline fissures, the calcarine fissure (CaF), best observed in the coronal plane, and the parieto-occipital fissure (POF), usually visualised in the axial plane (see Fig. 2.2). The POF and CaF are located close together and can sometimes be challenging to differentiate, especially at earlier GWs.

Abnormal cortical development has been linked to a number of diseases such as congenital heart disease [33, 34], CNS abnormalities [16], agenesis of the corpus callosum [35–37] and Down syndrome [38]. For this reason, the degree of cortical folding can potentially be used as a biomarker for healthy or abnormal development.

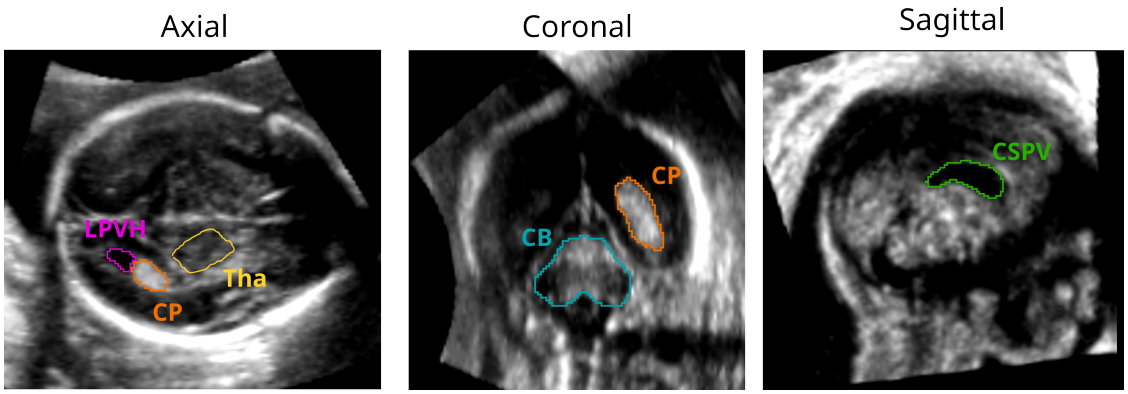


Figure 2.3: Example US image at 22 GWs with the visible subcortical structures drawn in: lateral posterior ventricle horn (LPVH), choroid plexus (CP), thalamus (Tha), cerebellum (CB) and the cavum septum pellucidum et vergae (CSPV). Due to the nature of US, with only one visible hemisphere, lateral structures are only drawn in one hemisphere.

Subcortical Development In contrast to the CoP, subcortical structure development is predominantly assessed with respect to their absolute and relative size in the brain. An example US image of the brain at 22 GWs with the key structures annotated is shown in Fig. 2.3. In early gestation, the most prominent brain structure is the CP, which can be seen as a large bright echogenic structure. The CP is positioned within the lateral ventricles, and almost completely fills these ventricles early in gestation. With increasing age, the relative size of the CP reduces and the fluid in the lateral ventricles becomes more visible. Especially the lateral posterior ventricle horn (LPVH) is usually well visible as a fluid-filled dark cavity next to the CP. The lateral ventricles extend as a comma-like structure encapsulating the CP to the midline, but due to their complex shape and indistinct boundaries, they are typically challenging to fully delineate on US scans. At the midline, the two lateral ventricles are separated by the cavum septum pellucidum (CSP), which is also a fluid-filled cavity between two thin membranes [39]. The CSP borders another fluid-filled cavity, the cavum vergae, and, as these two fluid-filled cavities can often not be clearly separated from each other on US images, I will refer to these together as the CSPV. At the start of the second trimester, the CSPV is not yet well visible, but it becomes clearly visible around 17 - 20 weeks [39], and can be best visualised in the sagittal plane as a clear dark cavity. On top of the CSPV, the corpus callosum is located. This is a pathway of white matter tracts connecting the two cerebral

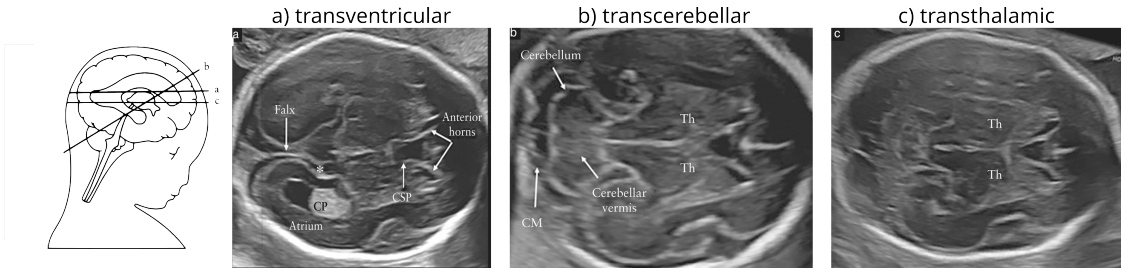


Figure 2.4: 2D US planes that are assessed during a routine second trimester of the fetal brain: (a) transventricular, (b) transcerebellar, and (c) transthalamic plane. Adapted from [41].

hemispheres with each other. However, this structure can be challenging to visualise on US [40], and is typically not visible on 3D scans from older US devices.

At the back of the brain, the cerebellum (CB) can be seen. On the axial plane, this can be seen as a butterfly-like structure (Fig. 2.4b), consisting of two hemispheric structures (the cerebellar hemispheres) connected by the cerebellar vermis [39]. The appearance of the CB is strongly affected by shadows from the occipital skull bone, and it is therefore often not well visible. However, the occlusion of a complete hemisphere is less strong at the back of the brain, and therefore, if visible, often both cerebellar hemispheres can be seen. Behind the CB, the cisterna magna (CM) can be seen, which is a small fluid-filled cavity at the midline, posteriorly to the CB. In the same axial plane as the CP and LPVH, the thalami are visible. These are two bean-like structures around the midline, but due to their poor soft tissue contrast, the boundaries are not very well-defined. Lastly, the brainstem is located below the thalami and appears as a darker structure around the midline, extending until the bottom of the brain where it connects to the developing spinal cord.

2.1.3 US Examinations in Clinic

Routine Examinations Pregnant women typically receive at least two US examinations during their pregnancy. An early examination, usually between 11 and 14 GWs, is performed to measure the growth of the fetus and confirm the delivery date [42]. This scan is therefore often referred to as the *dating scan*. During this scan also the brain is assessed, but due to the early stage of development, only a coarse visual inspection is performed [42].

A second routine scan in the middle of the second trimester, usually between 18 and 24 GWs, is performed to detect any fetal abnormalities [41]. During this scan, the fetal brain is studied in more detail, and several measurements are performed to quantify the development and rule out any CNS abnormalities. The examination is typically performed with transabdominal 2D US, and at least three axial planes are used to assess the brain: the transventricular, transcerebellar and transthalamic planes [41]. In Fig. 2.4examples of these three planes, often referred to as *standard* planes, are shown.

In these standard planes, the visual appearance of the lateral ventricles, CP, CB, CSPV and thalami is assessed. Furthermore, measurements of the atrial width, which is a measurement of the LPVH, transverse cerebellar diameter (TCD) and CM depth are taken. The TCD is predominantly used to quantify growth, whereas the atrial width, and CM depth are key in identifying CNS abnormalities [41]. For example, an atrial width larger than 10 mm indicates ventriculomegaly, which can be benign, or a sign of cerebral abnormality [43]. If during this scan there are any indications of a possible CNS abnormality, a referral for a targeted fetal neurosonography is given [1].

Targeted Neurosonography During a targeted neurosonography the brain is studied in more detail using a number of coronal and sagittal planes for visualisation. It is recommended to perform this examination with transvaginal US if technically feasible, but transabdominal US can be used when this is not possible, for example when the fetus is in breech position [2]. Recent guidelines also encourage the use of 3D US, as this facilitates the analysis of fully aligned views of orthogonal planes [2]. During the targeted neurosonography the emphasis is on a detailed *qualitative* analysis of brain development. This involves a more careful inspection of the subcortical structures examined during the routine scans, as well as the assessment of additional structures such as the brainstem and CC. Furthermore, during this assessment also the appearance of the CoP is qualitatively examined, which includes inspection of the SF, CaF and POF [2].

Cortical Assessment As described, the appearance of the CoP is not assessed during a routine examination, and, during a target neurosonography only a qualitative assessment is performed. While it is of clinical interest to quantify CoP development, due to the fact that it can potentially be used as a biomarker of health [16, 33–35], it is challenging to measure the complex shape changes it undergoes in a quantitative way. Some grading charts [32, 44] have been proposed that aim to divide fissure development into a number of distinct grades, ranging from a shallow indentation to fully operculised [32, 44]. Individual fissures can then be graded according to these charts, and compared to reference charts to determine whether the CoP is developing as expected. Other studies proposed measuring the depth [45–47] and angle [48] of the more prominent SF and POF. However, in practice such methods are very user-dependent, resulting in high inter-observer variability severely limiting their use in clinical practice [49].

2.1.4 Pregnancy Dating

The pregnancy duration is typically measured from the first day of the last menstrual period (LMP) and this is known as the *gestational age* (GA). The GA is about two weeks longer than the true fetal age measured from conception, the *conceptual age*, as ovulation occurs approximately two weeks after the first menstrual day. While the GA can also be estimated with US-based measurements of the fetus, it is, by definition, always expressed as the time since the LMP. This definition of GA, 'time since LMP', is thus distinct from the measurement technique used to estimate the GA.

In clinical practice, the GA is usually computed from the LMP, if known, or from an US-based crown rump length (CRL) measurement in the first trimester [50]. This measurement is depicted in Fig. 2.5a. In the first trimester, the CRL is the most accurate predictor of GA [51, 52], with an accuracy of $\pm 5 - 7$ days [53]. However, international guidelines recommend using the first available measure to estimate the initial GA, which is typically the LMP, and only adjust this estimate

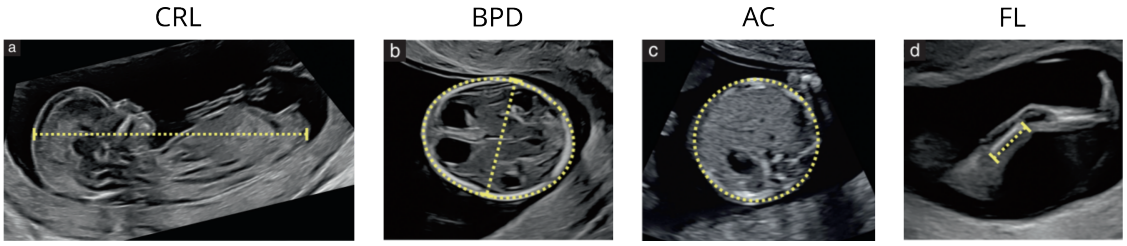


Figure 2.5: Measurements used to estimate GA from an US-based examination. Crown-rump length measurement is used during the first trimester (a), and a combination of the biparietal diameter (BPD) or head circumference (HC), abdominal circumference (AC), and femur length (FL) during the second and third trimester (b-d). Adapted from [42].

if the US-based CRL measurement in the first trimester deviates with more than 7 days from the LMP-based GA [53].

For women without a known LMP who receive their first US examination later in gestation (after approximately 14 GWs), a combination of multiple US-based measurements is recommended to estimate the GA (the head circumference (HC) or biparietal diameter (BPD), abdominal circumference (AC) and femur length (FL); see Fig. 2.5b-d). However, the accuracy of US-based measurements to estimate the GA decreases strongly in the second and third trimesters [53].

It should be noted that the GA computed from the known LMP assumes a two-week gap between the start of the menstrual period and ovulation, which does not always hold (i.e. due to irregular periods or variation in the menstrual cycle). On the other hand, the CRL is a physical measurement of the fetus and thus relates to the true post-conceptual age. Although this measurement is typically mapped to the GA using LMP-based reference charts (and therefore expressed as two weeks more than the conceptual age), these charts are population-based and will thus not be significantly affected by individual variation.

2.2 Automated Fetal Brain Analysis

In this thesis, I analyse the developing brain using subcortical structure segmentation and brain age predictions. For this reason, this section will provide an overview of past methods that have been proposed for these tasks and discuss their limitations.

2.2.1 Subcortical Structure Segmentation

To gain a deeper understanding of the growth of subcortical structures, it is necessary to use volumetric measurements instead of relying solely on 2D in-plane measurements, which are very plane-dependent and can only capture a limited representation of the full anatomy. Such 3D measurements can be obtained through manual pixel-wise annotations, but this is extremely time-consuming, especially for 3D volumes. For this reason, efforts have been put into developing automated segmentation methods. This section will first summarise all previous work aiming to automate subcortical segmentation in fetal US, and then provide an overview of the papers that have quantified volumetric subcortical growth during gestation in the past.

2.2.1.1 Automated Segmentation Methods

Automated subcortical structure segmentation has been widely studied in MRI volumes of the adult brain [54–56]. However, the adult brain is structurally different to the actively developing fetal brain during gestation, which contains structures in transient stages of morphogenesis. For example, certain structures, such as the CSPV, are only present during gestation and typically disappear after birth [57]. As MRI can also be safely used to visualise the fetal brain, but is clinically only recommended to confirm or complement a diagnosis following an US examination [2], some studies have performed subcortical structure segmentation in fetal MRI volumes [58–62]. Notably, Gholipour et al. [61] published an annotated atlas of the fetal brain from 21 to 39 GWs and used these labels to perform multi-atlas segmentation for new subjects. However, segmentation methods developed for MRI cannot be directly applied to US volumes due to the different nature of image acquisition resulting in large domain differences. While US is based on the interaction of the sound waves with tissue boundaries, MRI is based on protons in the tissue that react to magnetic fields. As a result, intensities on an MRI scan are rather uniform whereas the intensities on an US scan heavily depend on the interactions of the US beam with the surrounding tissues. For example, the two

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

A number of previous studies have developed methods specifically designed for subcortical segmentation in fetal US. An overview of these studies is shown in Table 2.1. From this table, it can be seen that the early methods all used 3D US for segmentation, and these will therefore be discussed first, followed by the more recent studies using 2D US image planes.

3D US Images To the best of my knowledge, the first automated subcortical segmentation method for fetal brain US was developed by Gutiérrez-Becker et al. [6]. In that study, a statistical shape model was used to segment the CB in 3D fetal US volumes. The performance was promising, but it was not clear how this method can be easily extended to other subcortical structures with more shape variation than the CB, making it challenging to develop an initial shape model. In another study, Yaqub et al. [5] performed segmentation of the CP, LPVH, CSP and CB in 3D US scans between 18 and 26 GWs. They proposed using a random decision forest using both appearance and distance features. Reported performance was good for the CP, LPVH and CSPV, but their method failed to generalise well to the CB. Furthermore, the search region for each structure was limited to the smallest cuboid enclosing all manual ground-truth annotations from the respective structure.

Since the introduction of deep learning for image segmentation, particularly with the U-Net architecture [63], there has been a notable improvement in the performance of various segmentation tasks in the field of medical imaging. However, only a few studies have applied deep learning methods to subcortical structure segmentation in 3D US, as usually a relatively large dataset with ground-truth labels is needed. Venturini et al. [7] used MRI atlas labels to overcome this need for manual annotations. Labels from the respective GW of the fetal brain MR atlas proposed

by Gholipour et al. [61] were registered to the individual US volumes, resulting in weak labels for the white matter, brainstem, thalami, and CB. These labels were subsequently used to train a multi-label 3D CNN to segment the respective subcortical structures in 3D US volumes. The reported performance was good, with extracted TCD measurements closely matching expected population-based averages, but a severe downside of this study was that testing was performed on the same weak MR atlas labels as used in training. As the registered MR atlas labels can be erroneous, and do not always align well with the structures visible in the US volumes, this therefore yields a biased assessment of performance and, as such, does not analyse the effect of using weak labels. In previous work [4], I used active contours to improve atlas-based labels as a preprocessing step before training a 3D U-Net, and showed that this resulted in increased performance for CSPV segmentations. However, I only validated that approach for an age range between 20 and 26 GWs, and do not expect this approach to translate well to all subcortical structures, as most have less distinct boundaries than the CSPV. Lastly, Venturini et al. [8] segmented the CB using unlabelled images in addition to fully annotated images to train a CNN. While the best performance was close to intra-observer variability, the number of labelled images required to obtain this performance (n=80) was relatively high. Furthermore, the approach was not extended to any other fetal subcortical structures.

2D US Images A number of other studies have developed segmentation models for 2D US planes [64–68]. Huang et al. [64] proposed to use a region-based based descriptor to segment the CSPV (referred to as corpus callosum [64]) and CP in 2D US planes. More recently, several papers all used modified versions of the U-Net for automated segmentation of the CB [65–67] and CSPV [69] in 2D standard planes. A similar approach was used to segment the CB in Maraci et al. [70], but a smaller CNN with only a single skip connection was used. Lastly, Chen et al. [68] aimed to obtain automated measurements of the atrial width by segmenting the LPVH, but no segmentation performance was reported. However, the sampled 2D slices

in all these studies were limited to standard US planes, making the segmentation task substantially less challenging than full 3D segmentation. Furthermore, the obtained segmentations can only provide a 2D measurement of subcortical growth, which only captures a limited part of the full volumetric growth.

Semi-automatic Methods In addition to the previously mentioned fully automated methods, some clinical studies report using *multiplanar* segmentation or *Virtual Organ Computer-Aided AnaLysis* (VOCAL) software (GE Healthcare) to perform their volumetric measurements in 3D US [71–74]. These methods estimate the structural volume from 2D contours drawn on parallel (multiplanar) or oblique (VOCAL) planes and are typically available in clinically used US devices. Although this does speed up analysis compared with fully annotating the volume, it still requires substantial manual annotation. Furthermore, the volume measurements have to be performed during US acquisition and take 2 to 10 minutes for a single structure [75]. These measurements would thus significantly lengthen an US session, and can not be used on a dataset of retroactively acquired 3D US volumes.

2.2.1.2 Volumetric Analysis of Subcortical Fetal Growth

Most studies analysing subcortical growth during gestation use 2D measurements routinely acquired in clinic [39]. Studies that do analyse volumetric development of subcortical structures during gestation most frequently use MRI volumes, either by manually annotating the volumes [76–78] or by using an atlas-based approach [79]. Theoretically, volumetric measures should be independent of the acquisition type (MRI or US), but in practice, these measures can vary due to the different appearance of tissues in each modality [80].

A number of studies have analysed 3D subcortical development using US [71–74, 81–84], and all have used the previously described multiplanar or VOCAL software. Both Benavides-Serralde et al. [73] and Zeng et al. [71] computed total intracranial, frontal, thalamic and cerebellar volumes to estimate the average volume differences for growth-restricted fetuses [73] and fetuses with congenital heart disease [71] compared against a normal control group. Several other studies measured thalamic

Table 2.1: Overview of existing automated subcortical segmentation methods for fetal brain US.

authors	**2D/3D**	**structures**	**age [GWs]**	**no. images**	**technique**
Gutiérrez-Becker et al. (2013) [6]	3D	CB	18 - 24	20	statistical shape model
Yaqub et al. (2013) [5]	3D	CP, CSPV CB, LPVH	18 - 26	51	machine learning
Venturini et al. (2019) [7]	3D	THA, BS, CB, WM	20 - 25	528	atlas-based and deep learning
Hesse et al. (2020) [4]	3D	CSPV	20 - 26	153	active contours and deep learning
Venturini et al. (2020) [8]	3D	CB	14 - 25	948*	deep learning
Huang et al. (2018) [64]	2D	CP CSPV	20 - 22 19 - 30	120 219	machine learning
Wu et al. (2020) [69]	2D	CSPV	18 - 37	224	deep learning
Chen et al. (2020) [68]	2D	LPVH	n.a.	2400	deep learning
Maraci et al. (2020) [70]	2D	CB	16 - 26	3736	deep learning
Singh et al. (2021) [65]	2D	CB	18 - 20	734	deep learning
Sreelakshmy et al. (2022) [66]	2D	CB	18 - 20	740	deep learning
Shu et al. (2022) [67]	2D	CB	18 - 26	959	deep learning

* Only 146 labelled volumes

[72, 74] and cerebellar volumes [72, 81–84] in order to construct normal ranges for subcortical development during gestation.

2.2.1.3 Summary Subcortical Segmentation

In summary, even though several studies have proposed automated methods for subcortical segmentation in fetal US, a single method obtaining good performance across multiple structures is lacking for 3D US. Furthermore, due to the absence of accurate automated segmentation methods, only thalamic and cerebellar volumetric development during gestation have been studied in previous work. In my first contribution (Chapter 5), I address these limitations by developing a segmentation method that obtains competitive segmentation performance across multiple subcortical structures (CB, LPVH, CSPV and CP). I then apply this method to obtain novel volumetric growth curves for the respective structures during the second trimester of gestation.

2.2.2 Brain Age Prediction

It is challenging to quantify the complete development of the brain only with structural segmentation methods, especially since segmentation labels are not usually available. Alternatively, prediction of the GA based on the brain morphology can be used to quantify the stage of development [10, 85]. The predicted age can be compared to the true GA to determine whether the development is age-appropriate or possibly advanced or delayed.

An important consideration when performing brain age predictions for the developing brain is whether to use size features for the prediction. Size features will typically increase the prediction accuracy as they are highly age-discriminative, but quantify growth rather than the stage of development. In this thesis, I am interested in quantifying the stage of development rather than growth, and for this reason, will not use any size-related features in the prediction model. In literature, studies that do not use size information (i.e. by only using appearance features [9] or scaling all

images to the same size [86–88]), and studies that do use size information to obtain the most accurate age prediction [89–92], which will both be discussed in this section.

Another aspect that should be considered is the determination of the ground-truth GA used for training and evaluating the model. As discussed earlier, it is not possible to determine the true post-conceptual age unless the exact day of conception is known, either from a known ovulation day or when conceived through h in-vitro fertilisation (IVF). However, estimates from the LMP, if known accurately, or a CRL measurement in the first trimester are typically a good approximation of the ground-truth (up to ± 5-7 days [53]). These approximations of GA are thus a *golden standard* rather than an actual ground-truth, but for clarity *ground-truth* will be used throughout this thesis.

A complete overview of prediction models developed for fetal brain age predictions from US is presented in Table 2.2. It can be seen that most methods have been developed for 2D or 2D+t (i.e. video) images, and that, except for a single older study [9], all age prediction models used deep learning methods. As for the segmentation methods, all early studies used 3D US and these will therefore be discussed first.

2.2.2.1 3D Ultrasound

There are only a few previous studies that perform automated brain age prediction from 3D US volumes [9, 26, 86]. Namburete et al. [9] used a machine learning approach to predict the brain age of healthy fetuses between 18 and 34 GWs. From the 3D US images, hand-crafted features were extracted that were then used to predict the GA of each fetus with a regression forest [93]. Two types of image-based features were proposed: *appearance-based* features that characterised shape patterns in cranial surface voxels, and *local size-based* features that computed the distance between sampled cranial surface points. In addition, the head circumference was used as an optional feature, which was extracted automatically from the cranial surface. Results were reported as root mean squared error (RMSE), resulting in an error of 7.52 days for including only appearance-based features, 6.72 days for both size and appearance-based features, and 6.20 when adding the head circumference

as a feature. In Table 2.2 only the performance for the appearance-based features is reported, as that model resembles the work in this thesis most closely.

In later work, the same dataset was used for brain age predictions using a shallow deep learning model [86]. Size-based image features were removed by spatially normalising all image volumes to a reference coordinate system, which involved scaling all image volumes to have a similar brain volume. A mean absolute error (MAE) of 7.72 days was reported, which was said to improve upon the earlier developed regression forest method, but this is difficult to verify due to the different measures used (RMSE versus MAE). Lastly, in Wyburd et al. [26], we proposed a deep learning model that received bounding boxes around fissures of interest as input to the prediction model. By predicting the age based on the appearance of a single fissure, this thus achieved *local* brain age predictions. That work [26] was carried out in collaboration between me and my colleague Dr. Wyburd, and has been the inspiration for the *guided attention* network for local brain age prediction proposed in this thesis (Chapter 6).

In all three of these studies [9, 26, 86] images from the same dataset were used (INTERGROWTH-21st). In that dataset, the ground-truth GA was computed from the LMP, which had to be known accurately, and confirmed with an US-based CRL measurement ($\leq$ 14 weeks). This dataset will also be used in this thesis and will be described in more detail in Chapter 4.

2.2.2.2 2D Ultrasound

It can be seen in Table 2.2 that more studies have explored age prediction for 2D US images. However, only two of these studies focussed on quantifying the stage of development by excluding scale features [87, 88].

No Scale Features In Lee et al. [87] a large number of 2D images were used to train a neural network that was able to predict the GA of the fetus and provide an uncertainty measure in the form of a standard deviation. The input images were acquired 2D transthalamic planes, which were directly used without any information on the pixel spacing. During acquisition, ultrasonographers will typically zoom

Table 2.2: Overview of automated brain age prediction methods for fetal US, *size feat* refers to whether any size-based features were used in the prediction model. If multiple results were reported in the paper, the one most similar to the work in this thesis is shown.

authors	**2D/3D**	**size feat**	**age [GWs]**	**no. images**	**MAE [days]**	**technique**
Namburete et al. (2015) [9]	3D	×	18 - 34	448	n.a.*	machine learning
Namburete et al. (2017) [86]	3D	×	18 - 34	447	7.72	deep learning
Wyburd et al. (2021) [26]	3D	×	19 - 30	811	4.1	
Lee et al. (2020) [87]	2D	×	14 - 42	8369	12.5	deep learning
Lee et al. (2023) [88]	2D	×	13 - 42	6666	4.5	deep learning
Maraci et al. (2020) [70]	2D	✓	16-26	3736	n.a.	deep learning†
Burgos-Artizzu et al. (2021) [89]	2D	✓	16 - 42	3065	4.76	deep learning
Pokaprakarn et al. (2022) [90]	2D+t	✓	9 - 39	6970	3.9	deep learning
Gomes et al. (2022) [91]	2D+t	✓	9 - 39	6656	3.8	deep learning
Dan et al. (2023) [92]	2D	✓	17 - 32	746	5.39	deep learning†

* RMSE of 7.52 days reported
† segmentation module followed by GA predictor, requires the availability of segmentation masks

in until the brain fills most of the image, before saving the view. For this reason, using these images as input is thus similar to upscaling all images to the same size as done by Namburete et al. [86]. An MAE of 12.5 days was obtained, which is substantially worse than the previously discussed 3D models [26, 86].

A considerable performance improvement was reported in Lee et al. [88], obtaining an MAE of 4.5 days for 2D images. That study used a dataset similar to [87], but in addition to the transthalamic planes of the brain (referred to as *head circumference* planes in the paper) also a femur and abdominal plane of each individual were used (as shown in Fig. 2.5). All three planes were given as input to a ResNet-like classifier that predicted the GW of the fetus. However, to consider it as a classification problem, the ages were binned per GW, removing some label information. On the test set of healthy fetuses, an MAE of 4.5 days was reported when using only the transthalamic plane as input to the model. The 2D images used in these studies [87, 88] also originated from the INTERGROWTH-21st dataset, and the ground-truths were thus determined from the LMP and confirmed with a CRL measurement.

Scale Features Several other papers did provide size as an input feature [70, 89–91, 94], either implicitly or explicitly, and thus focussed on determining the GA as accurately as possible (rather than quantifying the stage of development). Maraci et al. [70] developed a segmentation network to automatically segment the CB in 2D transcerebellar planes. From these segmentation masks the TCD was computed, which was then used to predict the GA using well-established parametric equations mapping cerebellar size to GA [94]. However, no performance measure for the final GA prediction was reported. In Burgos-Artizzu et al. [89] both 2D transthalamic planes and the image resolution (i.e. size feature) were given as input to a model predicting the GA. Additionally, manual biometric measurements were used as an optional additional input to the last layer of the network. The ground-truth GA was determined from a CRL measurement in the first trimester, and the reported MAE was 4.76 days (without the biometric measurements). Similarly to Maraci

et al. [70], Dan et al. [92] used a segmentation-based approach that first segmented the whole brain, abdomen and femur, and then used a regression network to predict the GA from these segmentation masks. Ground-truth GA was established at each individual examination using sonographic manual measurements of the HC, AC and FL [95]. While this is the recommended method to date a pregnancy in the second and third trimester [53], these estimates can have large errors (±7-21 days [53]) and this thus constitutes a very weak ground-truth GA. The additional value of training a segmentation model rather than directly predicting the GA from the images was also not specifiied, and the reported error was higher than in similar studies (5.39 days).

Lastly, both Pokaprakarn et al. [90] and Gomes et al. [91] used a similar dataset of 2D US sweeps (i.e. video data) to predict the GA between 9 and 39 GWs. These studies were both aimed at accurately dating the pregnancy, with an emphasis on resource-constrained settings where women typically only receive their first US examination later in pregnancy. Based on the site of acquisition, the ground-truth GAs were established either from the LMP and confirmed with a CRL measurement (USA), or with biometric measurements at the first US examination (Zambia). In both studies, a comparable MAE of 3.9 [91] and 3.8 [90] days was obtained on the main test set. Interestingly, Pokaprakarn et al. [90] also studied a subset of data in which the fetuses were conceived with IVF and for this subset the date of conception was thus exactly known. On this test set the reported MAE was only 2.8 days, demonstrating that inaccuracies in the ground-truth (i.e. *golden standard*) can contribute to a higher error.

2.2.2.3 Summary Age Prediction

In summary, a number of studies exist that predict fetal brain age from both 2D and 3D ultrasound, achieving high MAE performance. However, a large part of these studies aim to accurately date the pregnancy (using also size features) rather than quantifying the stage of development. Except for my previous work [26], most studies also only assess the developmental stage of the *whole* brain, whereas the *local* development could be more useful from a clinical perspective. For this reason,

in Chapter 6, I will propose a local brain age prediction model that predicts fetal age based on only part of the brain appearance.

Furthermore, almost all methods use deep learning models to perform fetal brain age prediction. While these methods resulted in improved prediction performance, a drawback is that it is challenging to understand why a certain prediction was made. In my last contribution (Chapter 7) I will address this limitation, and propose a deep learning model that is *interpretable*, and can provide insight into its reasoning process.

2.3 Conclusion

In this chapter, I introduced fetal brain US and how this is used in clinical practice to monitor the health and growth of the developing fetus. Furthermore, I reviewed past studies that addressed either automated subcortical segmentation or brain age predictions from fetal US images, and described their limitations. Finally, I summarised how these limitations motivated the contributions proposed in this thesis.

3
Interpretable Deep Learning

One of the limitations of conventional deep learning models is that they are not able to to provide an intuitive understanding of how the prediction was made, which is often also referred to as being a *black box*. To overcome this limitation, numerous methods have been developed that aim to provide human-understandable explanations for a prediction of a neural network while being faithful to the original model prediction [96]. In this context, faithfulness can be defined as being an accurate and complete representation of the decision-making process (i.e. highlighting the actual image features that are used by the model for the prediction) [96].

Interpretability of deep learning models is considered important for a variety of reasons [21, 23, 97]. For example, during model development, it is important to understand model behaviour for effective debugging and the identification of failure modes [23, 97, 98]. Such insights can then be used to iteratively improve the model before deployment. Similarly, insights into model behaviour can potentially lead to an improved understanding of the underlying causal relationships [23] (i.e. which parts of the image are most predictive for a certain task) or to the discovery of novel scientific knowledge [21]. Specifically for the healthcare domain, interpretability is also key in obtaining patients' and clinicians' trust, vital for the successful translation of developed methods into clinical practice [99].

A Note on Terminology The field of interpretable machine learning is not very well-defined [97], with definitions ranging from "providing transparency" (i.e. a sense of understanding the mechanism by which the model works) [97], or "the ability to present in understandable terms to a human" [21], to broader definitions like "extracting relevant knowledge about domain relationships contained in data" [96]. In this thesis, I will adhere to a broad definition of interpretability and consider interpretable deep learning methods as methods that can provide insight into the prediction process [97] or identify relevant patterns contained in the data [96]. In the literature, both 'interpretability' and 'explainability' are used, with 'interpretability' often used for methods that provide a sense of inherent understanding (i.e. such as decision trees), and 'explainability' preferred when a more active explanation is provided (i.e. *post-hoc* explanations) [100]. However, this difference is very subjective and there is a large grey area where both terms are used interchangeably. In this thesis, I will mainly use 'interpretability', as all contributions aim to incorporate a degree of inherent interpretability, but I will also use 'to explain' (as it has a different meaning than 'to interpret') and 'an explanation' when applicable.

Interpretability approaches can be categorised in many ways [101–103], however, there are a few key characteristics that are usually considered [101]:

- *Post-hoc* versus *inherently interpretable*: *post-hoc* methods analyse the prediction of a trained black-box model whereas inherently interpretable methods are prediction models designed to be simple enough to understand [96, 101]. Many traditional machine learning methods, such as linear regression or (small) decision trees, are considered to be self-explanatory, as their parameters (i.e. regression coefficients) can be provided as explanation. Such methods are by design faithful to the model prediction as they reflect the actual decision-making process. On the other hand, *post-hoc* methods typically approximate the complex function with a simpler human-understandable function, which may be less faithful to the actual model's behaviour [100].

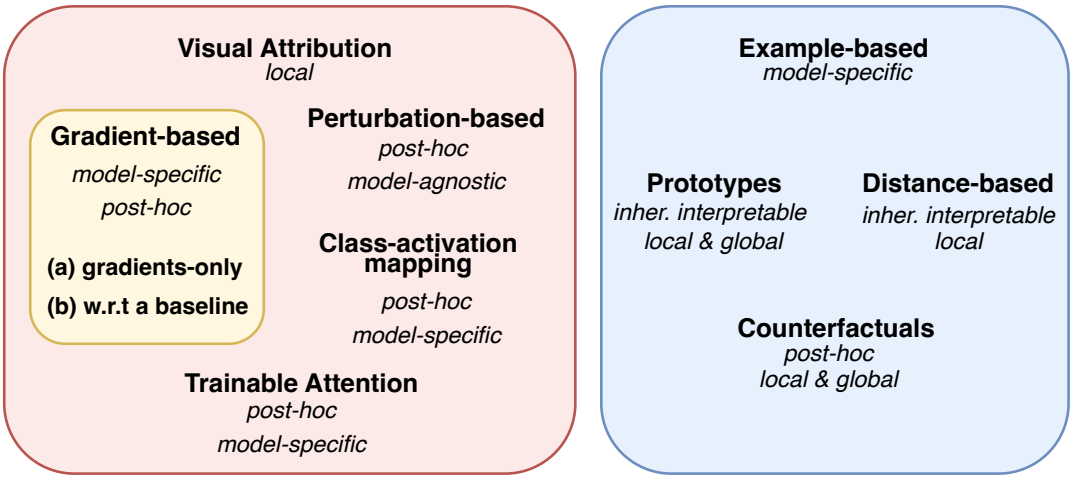


Figure 3.1: Categorisation of interpretability methods reviewed in this chapter. The characteristics given to each group (i.e. *post-hoc*) are the typical characteristics for a certain explanation class, but exceptions exist. The categories are also not mutually exclusive, for example, Grad-Cam [98] is a class-activation mapping method, but also uses the model gradients.

- *Model-agnostic* versus *model-specific*: model-agnostic explanation methods work for any prediction model, typically by probing the model with different inputs, and observing the change in output. The input can be varied, for example by occluding part of the image. These methods do not need access to internal model parameters or gradients. On the other hand, model-specific methods require certain network architectures and access to model gradients or weights.

- *Local* versus *global*: a local explanation explains a specific data sample, which is typically only valid in the small neighbourhood around that sample (i.e. for small perturbations of pixel values), whereas a global explanation provides an understanding of model behaviour across the dataset (e.g. what is the impact of a certain image feature across the whole dataset).

In the remainder of this review, the explainability approaches have been categorised into two main classes: *visual attribution* methods and *example-based* methods. An overview of these classes and their main characteristics is shown in Fig. 3.1.

Visual attribution methods aim to obtain a pixel-wise importance map indicating the importance of each pixel in the sample image, and are usually *post-hoc* and local.

Dependent on the implementation they can be either model-agnostic, e.g. gradient-based methods [104], or model-specific, e.g. perturbation-based methods [105].

Example-based methods provide similar images to the input image, either from the training set or model generated, as an explanation. These are typically model-specific, as they require a method for generating the examples, and can be either *post-hoc* or inherently interpretable. Dependent on the method, they can provide local explanations (i.e. the most similar image for a single sample) or global (i.e. the most representative examples for a certain class).

Finally, there are methods that do not fall into either of these two categories, such as concept-based explanations [106, 107] or text-based explanations [108]. However, as these typically require additional labels for training, these are considered out of scope for this review.

This review aims to provide a summary of the interpretability methods used in medical imaging, rather than being an exhaustive review of all applications (this has been extensively reviewed in previous work [101]). After discussing the individual approaches, the last section (Section 3.3) will provide a review of which methods have been applied specifically to the task of brain age prediction.

3.1 Visual Explanations

The most commonly used explanation methods in medical imaging are visual attribution-based methods [101]. These methods aim to obtain the contribution of each image pixel with respect to the prediction. Given an image $\mathbf{X} \in \mathbb{R}^{w \times h}$, the black-box prediction model f makes a prediction as $\hat{y} = f(\mathbf{X})$. In this section, I assume the prediction to be a single value $\hat{y} \in \mathbb{R}^1$, either the predicted scalar value (regression) or the logit of a single class. A feature attribution method then assigns an importance score to each pixel, or to groups of pixels, in $\mathbf{X}$, based on their contribution to the prediction, resulting in an attribution, or saliency, map: $\mathbf{A} \in \mathbb{R}^{w \times h}$. For classification, this thus results in class-specific attribution maps explaining the contribution of pixel values to the prediction of a certain

class, whereas for regression a single attribution map is generated that explains the contribution of pixels to the predicted value.

Attribution-based methods are often divided into two categories: gradient-based methods and perturbation-based methods. Gradient-based methods use the networks' gradients to generate attribution maps, and are thus model-specific. On the other hand, perturbation-based methods perturb the input image (i.e. by occluding part of the image) and observe the effect of this perturbation on the prediction. These methods are thus model-agnostic, and can be applied to any prediction model.

While commonly considered under the gradient-based explanation methods, in this review I will discuss Class Activation Maps (CAM) as a third category, as, in their simplest form, they do not require gradient computations. These methods are only suitable for certain model architectures and are therefore also considered model-specific.

3.1.1 Gradient-based

Gradient-based explanations typically compute local gradients with respect to the input image to generate an explanation. Formally, this can be written as:

$$\mathbf{G} = \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \tag{3.1}$$

with $\mathbf{G} \in \mathbb{R}^{w \times h}$ the pixel-wise gradient map. As this is a linear approximation of the function f at $\mathbf{X}$, which is highly non-linear, this only holds for very small changes in $\mathbf{X}$. Rather than estimating the global contribution of pixels in $\mathbf{X}$ (i.e. how much contributed a certain pixel value to the prediction), the resulting attribution maps thus only describe the sensitivity of the prediction to changes in the pixel values of $\mathbf{X}$.

Explanation methods using gradients can be classified into methods that provide the gradient map directly ($\mathbf{A_{grad}} = \mathbf{G}$, with $\mathbf{A_{grad}}$ the attribution map) and methods that compute feature contributions with respect to a baseline image, $\mathbf{X_0} \in \mathbb{R}^{w \times h}$. This is illustrated in Fig. 3.2 (a) and (b). It can be seen that for

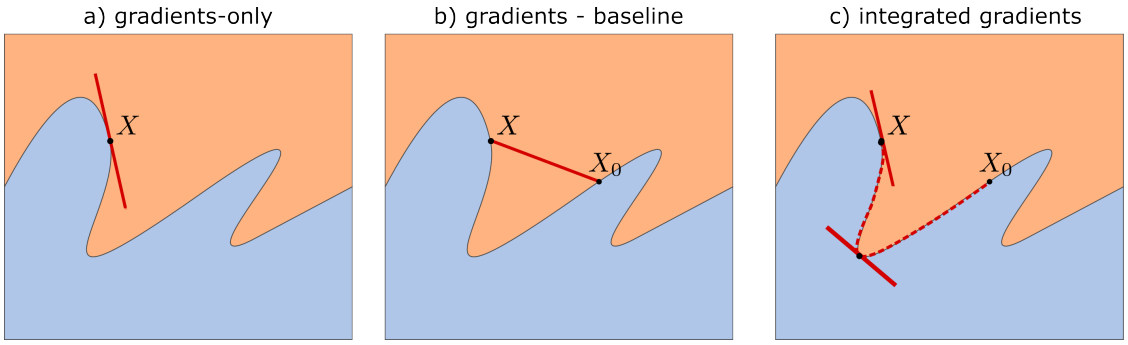


Figure 3.2: Schematic of different visual attribution based methods for (a) gradients only (e.g. backpropagation [104]), (b) gradients w.r.t. a baseline (e.g. LRP [109]), and (c) integrated gradients [110]. The red lines are different approximations of the (local) decision boundary and the two colours represent two classes. The dashed line for the integrated gradients is the path along which the gradients are computed. The decision space is drawn as a binary 2-class problem for visualisation purposes but is, in reality, the continuous value of an output node. In practice, the line orthogonal to the decision boundary is computed (i.e. the direction of most change).

the gradient-only approach, the linearisation around $\mathbf{X}$ is only valid for a small neighbourhood and does not translate well to a global decision boundary.

On the other hand, methods that compute feature contributions with respect to a baseline typically linearise the contributions of features between $\mathbf{X}$ and $\mathbf{X_0}$ (see Fig. 3.2b). As this considers not only the local neighbourhood of $\mathbf{X}$, but considers the change between $\mathbf{X}$ and $\mathbf{X_0}$, these methods are sometimes referred to as *global* gradient methods [111] (i.e. Fig. 3.2b can be considered as a better approximation of the overall decision boundary than (a)). However, it has been shown [110, 112] that under certain conditions such methods can be simplified to the element-wise multiplication of the local gradient with the input ($\mathbf{A_{grad-base}} = \mathbf{G} \odot \mathbf{X}$). This demonstrates that these methods are thus not always truly global, and I will therefore avoid this terminology and refer to these methods as *gradients w.r.t. a baseline.*

3.1.1.1 Gradients Only

The *gradients only* approach ($\mathbf{A_{grad}} = \mathbf{G}$) can be considered as computing the effect of changing a certain image pixel by a small amount (e.g. *locally*) on the prediction. Around the same time, three well-known gradient-based explanations methods were proposed: Deconvolution [113], Backpropagation [104] and Guided

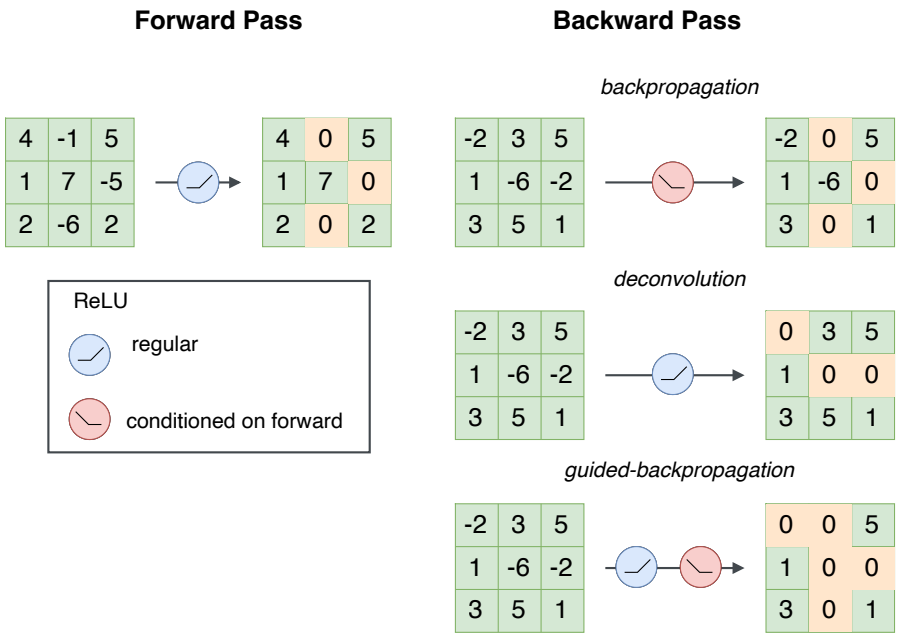


Figure 3.3: Schematic overview of the handling of the ReLU on the backward pass for backpropagation [104], deconvolution [113] and GBP [114]. The green grids represent example feature maps. Numbers that are changed by the ReLU (i.e. set to 0) on each pass are shown in orange.

Backpropagation (GBP) [114]. While each was introduced as a somewhat different method, they only differ in their handling of the Rectified Linear Unit (ReLU) non-linearity in the backward pass.

A schematic example of the backward ReLU for all three methods is shown in Fig. 3.3. Backpropagation uses the vanilla back-propagation gradient, which implies passing through all pixel values that were positive in the forward pass, and setting the other values to 0, regardless of their values in the backward pass [104]. It should also be noted that this differs from backpropagation during network training, where the derivatives with respect to the *network weights* are computed. Deconvolution treats the ReLU in the backward pass the same as in the forward pass and only passes through positive signals [113]. Finally, GBP is a combination of both approaches, essentially having a regular ReLU followed by a ReLU conditioned on the forward pass [114].

While each of these methods is widely used to explain neural networks [101], it has been shown that only back-propagation results in valid class-specific attribution

maps. Both Deconvolution and GBP result in sharper, less noisy, attention maps, but are not a faithful representation of the model prediction [115]. That work also showed that both Deconvolution and GBP do partial image recovery (i.e. reconstruct part of the input image) [115], which, while resulting in crisp attribution maps, is not the desired or expected outcome of an explanation method. This was also confirmed in [116], where it was shown that GBP results in visually appealing attribution maps for neural networks initialised with noise.

To reduce the noise in attribution maps obtained with backpropagation, SmoothGrad [117] proposed to use the average attribution over a set of noisy input images. The noisy input images were sampled around the original image $\mathbf{X}$ with additive Gaussian noise: $\mathbf{X} + \mathcal{N}(0, \sigma^2)$. This essentially attempted to approximate the average gradient in the neighbourhood of $\mathbf{X}$, and resulted in visually sharpened attribution maps.

3.1.1.2 Gradients with respect to a Baseline

Explanation methods that use a reference baseline aim to compute the contribution of each feature value (i.e. pixel values) with respect to this baseline. This is typically done using a linearisation between $\mathbf{X}$ and $\mathbf{X_0}$ (see Fig. 3.2), and aims to find the contribution of each pixel value to the prediction $\hat{y}$. This contribution can be considered as the $\delta\hat{y}$ that each pixel induces, which, for a linear approximation, can be obtained as follows:

$$\mathbf{A_{grad-base}} = \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} (\mathbf{X} - \mathbf{X_0}) \tag{3.2}$$

in which the elements of $\mathbf{A}$ represent the δy of each pixel ($a_{i,j} = \delta y$). It can be seen that Eq. 3.2 boils down to $\mathbf{G} \odot \mathbf{X}$ for a zero-valued baseline ($\mathbf{X_0} = 0$) [111].

Gradients w.r.t. a baseline methods are often said to fulfil the completeness axiom [110, 112], which states that the pixel-wise attributions should sum up to the difference in output ($f(x)$) between the sample $\mathbf{X}$ and the baseline $\mathbf{X_0}$. This can be formalised as follows:

$$\sum_{i \in n_p} a_i = f(\mathbf{X}) - f(\mathbf{X_0}) \tag{3.3}$$

with $a_{i,j}$ the attribution at pixel (i, j) and n_p the total number of pixels in the image.

Several gradient-based methods w.r.t. a baseline have been proposed in the literature that I will now briefly discuss, and show, under which conditions, they reduce to $\mathbf{G} \odot \mathbf{X}$. It should also be noted that choosing a baseline for images is non-trivial, as it is not possible to simply remove a certain feature. The baseline is therefore often chosen to be a black, white image or mean intensity image (i.e. with all pixel values set to the mean intensity of the image, or images in the dataset), but this might not always be the most optimal choice.

Relevance Propagation Layer-wise relevance propagation (LRP) [109] and Deep Learning Important FeaTures (DeepLIFT) [112] both use the concept of relevance propagation to find the contribution of each pixel value to the final prediction. Relevance propagation involves determining the relative contribution, referred to as relevance, of all inputs to a certain network node. For example, for the output node (i.e. the prediction), the contribution of each node in the previous layer has to be determined. This is performed consecutively until the input (image) space is reached. At each node the distribution of relevance across the previous layer nodes is performed according to the conservation of relevance rule: the relevance of a single node should be equal to the sum of the relative relevance of all input nodes. This conservation of relevance rule is equivalent to the completeness axiom in Eq. 3.3.

Implicitly, LRP computes the relevance of each pixel with respect to a zero-valued baseline, i.e. how much did each pixel contribute compared to a prediction of zero with an input image of only zeros [118] (Eq. 3.2 with $\mathbf{X_0} = 0$). Similarly, DeepLIFT explains the difference in prediction $\delta\hat{y}$, between $\mathbf{X}$ and a reference input, $\mathbf{X_0}$ [112] (Eq. 3.2 with a chosen baseline). For both methods, a set of rules determine how the relevance is distributed at the non-linearities, which have to be selected heuristically for each layer.

While the relevance propagation has a different intuition to the previously discussed gradient computations, LRP is equivalent to a scaled version of $\mathbf{G} \odot \mathbf{X}$ when only ReLU activations are used [111]. Further, DeepLIFT with a zero-baseline also equals $\mathbf{G} \odot \mathbf{X}$ for a network with only ReLU activations and no bias terms in the layers [111].

Deep SHAP An extension to DeepLIFT was proposed in [119], in which the rules were adapted so that the pixels' relevance would match the Shapley values. Shapley values are a classical concept from cooperative game theory quantifying the individual contribution of each player in a game [120]. This concept can be extended to determining the importance of individual image pixels and provides theoretical guarantees over the heuristically chosen rules in the original paper [112].

Integrated Gradients Integrated gradients (IntGrad) attempts to overcome the limitations of local gradients by using an average gradient on the path between the image $\mathbf{X}$ and the baseline $\mathbf{X_0}$ [110] (Fig. 3.2c). Considering the path between $\mathbf{X}$ and $\mathbf{X_0}$, integrated gradients computes the gradient across all points on the path.

Defining a point on the path as $\mathbf{X_0} + \alpha(\mathbf{X} - \mathbf{X_0})$, with α ranging from 0 to 1, the average gradient can be obtained by integration over that path. It is then multiplied with the difference between the baseline and image to obtain the final attribution map. This can be formulated as follows:

$$\mathbf{A_{IntGrad}}(\mathbf{X}) = (\mathbf{X} - \mathbf{X}_0) \odot \int_{\alpha=0}^{1} \frac{\partial F(\mathbf{X_0} + \alpha(\mathbf{X} - \mathbf{X_0}))}{\partial \mathbf{X}} d\alpha \tag{3.4}$$

which, for a zero-valued baseline simplifies to $\mathbf{G}_{mean} \odot \mathbf{X}$ with $\mathbf{G}_{mean}$ the average gradient on the path between $\mathbf{X}$ and $\mathbf{X_0}$.

In practice, the integral in IntGrad is computed with a *Riemann* sum, which requires the gradients to be evaluated at multiple values of $\mathbf{X}$. This makes IntGrad slower to compute than most of the aforementioned methods using only a single local derivative.

3.1.2 Class Activation-based

A different approach to obtaining attribution maps was proposed by Zhou et al. [121]. In their work, they introduced Class Activation Maps (CAM), that were specific to CNNs ending with a Global Average Pooling (GAP) layer followed by a single fully connected (FC) layer (e.g. a ResNet architecture [122]). The generation of CAMs is depicted in Fig. 3.4. The feature representation after the last convolutional layer, defined by $\mathbf{Z} \in \mathbb{R}^{h_z \times w_z \times c_z}$ with h_z and w_z the latent spatial dimensions of the respective layer and c_z the channel dimension, is used to generate the CAM attribution. These representations are combined with the FC layer weights connecting each feature map to the output node (or the output node of interest for classification), denoted by $\mathbf{w} \in \mathbb{R}^{c_z}$, to generate the attribution map as follows:

$$\mathbf{A_{CAM}} = \sum_{k \in c_z} \mathbf{Z}_k w_k \tag{3.5}$$

with $\mathbf{Z}_k$ being the k$^{\text{th}}$ feature map in the channel dimension, w_k the weight connecting the k$^{\text{th}}$ feature map with the output node of interest, and $\mathbf{A_{CAM}} \in \mathbb{R}^{h_z \times w_z}$ the resulting attribution map. The attribution map is therefore of the same resolution as the last convolutional layer, which can subsequently be upsampled to the input size of the image for visualisation. However, this means that, depending on the resolution of the last convolutional layer, the resolution of the explanation can be coarse compared to gradient-based pixel-level attribution maps.

As discussed, CAM can only be applied to CNNs ending with a GAP layer followed by a single FC layer. To enable the application of CAM to other CNN architectures Grad-CAM was proposed [98]. This can be considered as a generalisation of the CAM method and uses gradients to compute importance scores for each of the feature maps in $\mathbf{Z}$, as opposed to the FC layer weights. The gradients are computed from the output node with respect to the last convolutional feature representations, and are subsequently pooled with GAP to obtain scalar weights for each of the feature maps:

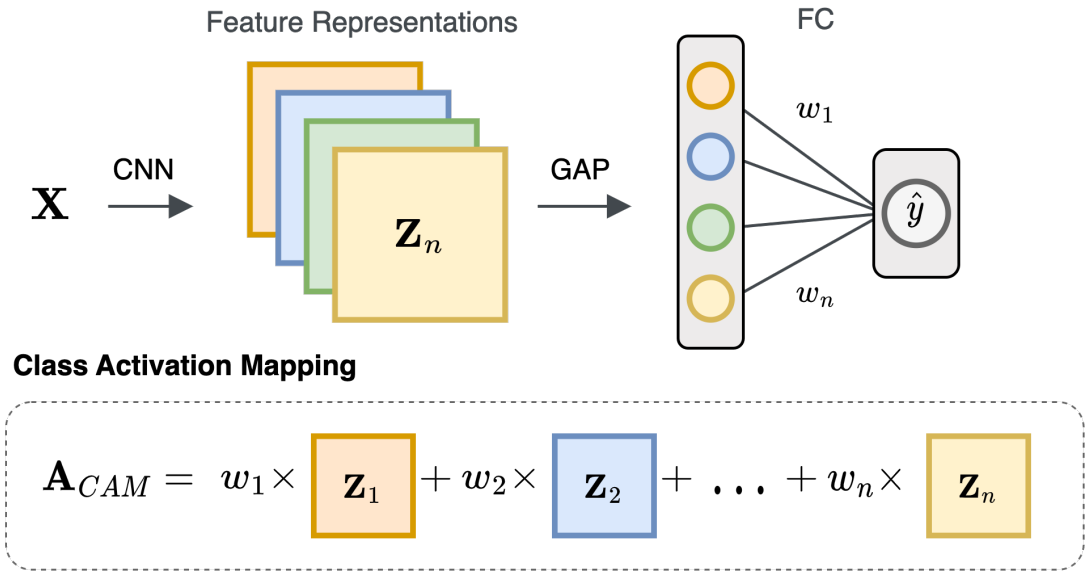


Figure 3.4: Schematic overview of class activation maps (CAM) [121]. CNN is restricted to be fully convolutional as the CAMs are generated from the extracted feature representations.

$$\mathbf{w}^{\mathbf{Grad-CAM}} = \text{GAP}(\frac{\partial \hat{y}}{\partial \mathbf{Z}}) \tag{3.6}$$

with $\mathbf{w}^{\mathbf{Grad-CAM}} \in \mathbb{R}^{c_z}$. The total representation is then filtered with a ReLU to only keep the positive attributions for each class.

$$\mathbf{A}_{\mathbf{Grad-CAM}} = \text{ReLU}(\sum_{k \in c_z} \mathbf{Z}_k w_k^{Grad-CAM}) \tag{3.7}$$

It can clearly be seen that this is a generalisation of CAM in which the weights of each feature map are computed with gradients. For a CNN with a GAP layer followed by a single FC, the gradients over the last layer are simply the FC layer weights, resulting in the same formulation as CAM in Eq. 3.5 (except for the ReLU).

In addition to Grad-CAM, Selvaraju et al. [98] also proposed guided Grad-CAM, which is a simple element-wise multiplication of $\mathbf{A}_{\mathbf{Grad-CAM}}$, upsampled to the image resolution, and the attribution map obtained with GBP [114]. The motivation for this was to combine the high resolution of GBP with the class-specificity of Grad-CAM [98].

Several extensions of Grad-CAM have also been proposed [123, 124], which mainly vary in how the weights ($\mathbf{w}^{\mathbf{Grad-CAM}}$) of each feature map are generated from the gradients (i.e. instead of using GAP). Another work [125] criticised the use of gradients over the non-convolutional layers to generate the weights, due to their limited local validity. They instead proposed to use importance scores obtained from masking the image with the respective activation map and observing the change in the output logit [125]. This approach essentially generalises CAM with a perturbation-based approach instead of a gradient-based approach. Perturbation-based approaches will be discussed in more detail in the following section.

3.1.3 Perturbation-based

Perturbation-based explanation methods apply a certain perturbation to the input image (e.g. occluding a certain area) and observe the effect of this on the predicted value. They can thus be applied to any prediction model, and are, therefore, considered model-agnostic. However, perturbation-based explanations are typically slow to generate as the model predictions must be evaluated for several perturbed images. A schematic overview of perturbation-based explanation methods is shown in Fig. 3.5.

The attribution maps resulting from perturbation-based methods depend heavily on the type of perturbation used. Typically, we are interested in explaining which image regions contributed most to the final prediction, and it thus makes sense to *remove* certain image regions. On the other hand, region removal might not be able to explain a predictor that relies on image intensities throughout the whole image (e.g. differences between images acquired with different MR scanners).

Occlusion One of the first works using perturbation-based explanations measured the occlusion sensitivity by occluding parts of the image with overlapping grey squares (i.e. a pixel value of 128 for 8-bit images) and observing the resulting classifier output [113]. An attribution map was then created by showing the average change in logit value (i.e. $\Delta\hat{y}$) for each occluded region.

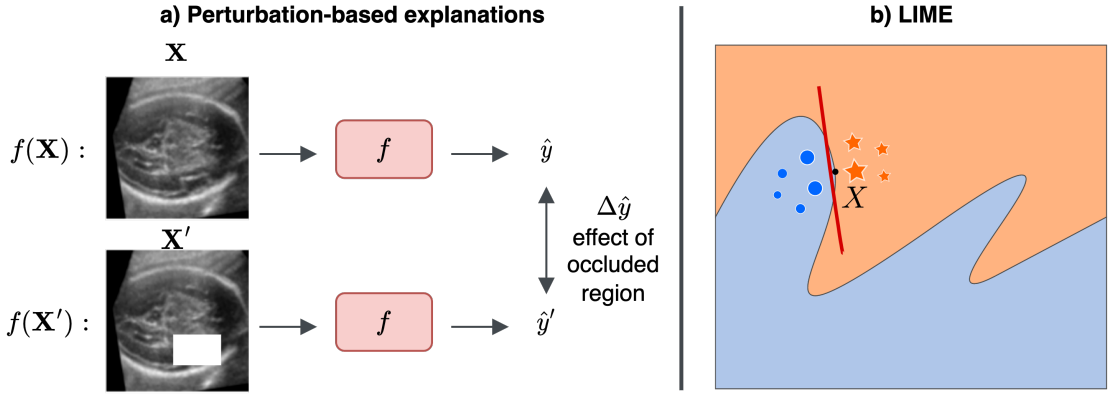


Figure 3.5: Overview of (a) perturbation-based explanations, and (b) local function approximation with LIME [105]. For LIME the coloured stars and circles show perturbed samples ($\mathbf{X}'$) in the neighbourhood of $\mathbf{X}$ to which the (local) approximation of the decision boundary is fitted.

LIME Another approach using occlusion is Local Interpretable Model-agnostic Explanations (LIME) [105]. In LIME, the main idea is to approximate the complex model, denoted by f, that we want to explain by a local linear (or other interpretable) function. The local function is estimated by drawing samples, denoted by $\mathbf{X}'$, in the local neighbourhood of the image $\mathbf{X}$. These samples are generated by greying out parts of the image with superpixels, which are groups of perceptually similar pixels obtained by the oversegmentation of an image.

For each perturbed sample, the prediction is obtained: $\hat{y} = f(\mathbf{X}')$, as well as a sample weight based on a kernel function of its distance to the original sample. In the original LIME formulation an exponential kernel was used resulting in the following formulation for each sample weight:

$$w_{\mathbf{X}'} = \exp\left(\frac{-d(\mathbf{X}, \mathbf{X}')^2}{\sigma^2}\right) \tag{3.8}$$

with $d(\mathbf{X}, \mathbf{X}')^2$ the L_2 distance between $\mathbf{X}$ and $\mathbf{X}'$, σ the standard deviation of the exponential kernel and $w_{\mathbf{X}'}$ the resulting weight of sample $\mathbf{X}'$.

Both the weights and predictions of all $\mathbf{X}'$ can then be used to fit a local linear classifier or regressor (see Fig. 3.5b). This will result in coefficients for each of the superpixels that can be used to visualise the contribution of each image region.

A key element of LIME, when used for image-based predictions, is the choice of superpixel method. The original LIME paper used an out-of-the-box segmentation algorithm (Quickshift [126]), but any segmentation method generating superpixels could be used. Notably, in [127] it was proposed to use nuclei segmentations as superpixels for the explanations of histopathology image predictions, resulting in sharpened attribution maps that correlated better with tissue boundaries.

KernelSHAP Combining LIME with Shapley values, Lundberg and Lee [119] modified the exponential kernel so that Shapley values were generated for the importance scores of each superpixel.

Meaningful Perturbations To make the explanations less dependent on a certain selected occlusion method, Fong and Vedaldi [128] proposed to use meta-learning to learn the mask that maximally reduced the output activation of the class of interest. The learned masks were constrained to be smooth to limit the artefacts that could be encoded in the masks.

3.1.4 Trainable Attention

Trainable attention was not specifically developed as an interpretability method and is, therefore, not always described in interpretebility reviews. For this reason, it is also not clearly defined as belonging to one of the classes in Fig. 3.1. As it assigns importance scores to individual pixels (or groups of pixels) [128], in this review, it has been classified as a visual attribution method. However, in contrast to all other visual attribution methods, trainable attention is inherently interpretable and not *post-hoc*. Trainable attention incorporates an attention module into the architecture that learns to generate a pixel-wise mask for each sample and applies this to the image to mask out irrelevant features [129]. While this was initially developed to boost performance, the attention weights can also be visualised as a type of explanation. A more detailed review of trainable attention will be provided in Chapter 6.

3.1.5 Local Function Approximation

Recent work [130] showed that many gradient and perturbation-based explanation methods can be considered as local (simple) function approximators. The main difference between these methods is how the local function is approximated, which can be summarised by three aspects:

- The loss function used to approximate the local function (gradient matching for gradient-based methods and mean squared error for perturbation-based methods, see Fig. 3.2a and 3.5b).
- The type of noise used to define the neighbourhood (for example infinitesimal small additive noise for gradients only, multiplicative noise for $\mathbf{G} \odot \mathbf{X}$ and element-wise multiplication with a mask for perturbation-based methods).
- How the neighbourhood is weighted: normally distributed (gradients), uniform ($\mathbf{G} \odot \mathbf{X}$) or a kernel function (LIME and KernelSHAP).

This perspective shows that while the different visual attribution methods might appear very different at first glance, there are in fact many similarities between them. Most importantly, almost all attribution methods compute *local* linear approximations of the complex decision model which are only valid in the neighbourhood of the sample that is being explained. A notable class of models that do not fit in this framework are all gradient-based explanation methods w.r.t. a baseline that do not simplify to $\mathbf{G} \odot \mathbf{X}$ (i.e. DeepLIFT [112]).

3.1.6 Limitations of Visual Attributions

Visual attribution explanations have been the most popular explanation method in the field of medical imaging [101]. While they can be a very powerful explanation technique, in recent years numerous works have appeared highlighting the inherent limitations of such methods [115, 116, 123, 131, 132].

Gradient-based Methods As already mentioned in the previous section, most visual attribution methods compute *local* gradients at the point of the sample image, effectively approximating the non-linear network with a local linear approximation. This explanation differs from the one typically desired and should, therefore, be interpreted with care. For example, a particular feature can be very important around the decision boundary, but as the function flattens (resulting in a gradient of close to 0), the feature might lose its local importance, and is, therefore, not highlighted in a feature attribution map.

Numerous papers published in the last couple of years have also questioned the reliability and effectiveness of gradient-based feature attribution methods in explaining highly non-linear neural networks [115, 116, 131, 132]. Adebayo et al. [116] proposed a set of sanity checks was to verify the sensitivity of explanations to model weights and training data. *Model weight sensitivity* was tested by randomising layer weights of the network (either layers individually or cascading randomisation) and *data sensitivity* by training with true and completely randomised labels. Experiments were performed for a number of gradient-based explanation methods, showing that all tested methods computing an explanation with respect to an (implicit) baseline, produced explanations that were dominated by the input when the gradients were noisy (as is the case when the model parameters or data labels are randomised). This made the visual appearance of the attribution map insensitive to both model parameters and training data, which is clearly undesirable for an explanation method. Furthermore, it was shown that GBP also had very low sensitivity to these factors with its explanations resembling an edge detector output. This confirmed earlier work [115] that showed that both GBP and Deconvolution reconstruct part of the input image rather than providing a faithful explanation.

Another sanity check was introduced by Kindermans et al. [132], stating that reliable explanation methods should be *shift invariant*. This means that the attribution map should be unchanged when applying a shift to the data, such as adding 1 to all pixel values, and changing the prediction model accordingly (i.e. by adjusting the bias terms). In their work they demonstrated that all tested

gradient-based methods with a baseline (LRP, IntGrad, and $\mathbf{G} \odot \mathbf{X}$) were not shift invariant, failing the sanity check. In the same work [132], it was also shown that the resulting explanations of such methods were highly dependent on the choice of baseline. Finally, a recent paper [131] proved theoretically that gradient-based attribution methods fulfilling the completeness axiom (Eq. 3.3) are not able to correctly explain model behaviour for complex neural networks (i.e. models that can not be well approximated by a linear function), further questioning the use of such attribution methods.

In summary, vanilla backpropagation-based [104] methods (without reference baseline) produce valid explanations, but these are only valid in a very small neighbourhood around the sample. They should therefore be interpreted with care. On the other hand, gradient methods with respect to a baseline produce unfaithful explanations under several conditions [116, 131, 132] and might, therefore, be better avoided until a better understanding of these methods becomes available.

Class Activation-based Explanation methods based on CAM [121] have been shown to produce faithful explanations, that are sensitive to model weights and data labels [116]. Their main limitations are that they can only be used for CNNs, and that the resulting attribution maps have a resolution equal to the output size of the last convolutional layer. As this is typically a number of times smaller than the original image due to max-pooling layers, the resulting attribution maps are coarse, limiting the level of detail in the explanation.

The generalisation of CAM, Grad-CAM [98], has received criticism in that the averaging of the pixel-wise gradients by a GAP into a single weight per feature map is heuristically chosen, which sparked some alternative implementations [123, 124]. Furthermore, given that GBP can produce unfaithful attribution maps, guided Grad-CAM (which is an element-wise multiplication of Grad-CAM and GBP) can thus also be considered unfaithful.

Perturbation-based Perturbation-based explanations are highly dependent on the types of perturbations applied. Ideally, the image would be perturbed in a way that removes a certain image feature while still representing a realistic image. This is however challenging, especially for medical imaging. Methods like LIME [105] use superpixel methods to produce segmentations used for the explanation. While showing good results on natural images, such segmentation methods do not always translate well to medical images, especially in a modality like US where object boundaries appear visually very different from in natural images. Dedicated segmentation methods can be used instead [127], but the drawback of such methods is that they are potentially prone to confirmation bias [116, 133], since an explanation matching all structural boundaries might look visually appealing by default, but does not reveal any useful model behaviour. Furthermore, the removed regions need to be replaced with a sensible value, which is not trivial for medical images but can have a large effect on the resulting explanation.

Meaningful perturbations do not require a choice of segmentation method as the mask is learned by optimisation. However, the mask is applied to the image with either of three options: a constant value replacement, noise injection or Gaussian blur. The results are therefore dependent on this choice, and they do not always translate well to medical imaging tasks [134]. Furthermore, optimizing an additional meta-learner to produce explanations for each sample is impractical in practice.

3.2 Example-based Explanations

Example-based explanation methods provide example images as explanation for a prediction. For example, this could be the n most similar samples in the training set, or an image generated from the latent space to visualise changes in the prediction. This class of explanation models is typically model-specific, as it requires specific architectures or loss functions to provide the examples with the prediction.

In this section, I will give a brief overview of example-based explanation methods, their applications to medical imaging tasks and their limitations.

3.2.1 Prototype Learning

In prototype learning, a set of feature representations in latent space is learned during model training, the so-called *prototypes*, which can be considered representative examples for a certain class label. Subsequently, inference for a new sample is performed using the distances to the learned prototypes in latent space, for example by using nearest-neighbour classification [135]. In most early deep prototype learning approaches, the learned prototypes were unconstrained points in latent space [136, 137], making it challenging to visualise or interpret the learned prototypes. These early prototype models therefore cannot really be considered true interpretability methods, but they inspired later methods.

To achieve visualisation of the prototypes, Li et al. [138] introduced an autoencoder [139] to reconstruct an image from the latent representation. In their approach, they simultaneously learned a set of representative prototypes in latent space and a decoder network that was able to reconstruct a learned prototype into image space. Furthermore, a sample prediction was made by passing the prototype-sample distances through an FC layer, effectively weighting the prototypes for the final prediction, followed by a softmax layer generating the class probabilities. In that approach, the prototypes did not have associated class labels, but their contribution to each of the classes was learned by the weights in the FC layer. The learned prototypes looked visually plausible, but their methods were only demonstrated on three small toy datasets: MNIST [140], Fashion MNIST [141] and a simple car dataset [142].

To extend these methods to more challenging datasets, Chen et al. [143] proposed to constrain the prototypes in the final model to be equal to latent representations of images in the training set, which could easily be visualised. This was achieved by training the prototype model, referred to as *ProtoPNet*, in three stages: (1) regular training of network weights with the last FC layer weights frozen, (2) projection of prototypes, and (3) optimization of the last FC layer weights. The projection of the prototypes entailed replacing all prototype representations (points in the latent space) with the closest latent representation of an actual training image. In contrast

to the work by [138], the prototypes were also assigned a class label at the beginning of training. Further, a latent image representation $\mathbf{Z} \in \mathbb{R}^{h_z \times w_z \times c_z}$ was considered as a set of $h_z w_z$ vectors with length c_z, each representing part of the image. The prototypes were then defined as vectors of the same size: representing an image *part*. This enabled more fine-grained comparison, and resulted in explanations of the form of 'which *part* of the image looks like which prototype'. The single image-prototype distances required to make the final prediction were obtained by taking the minimum overall distances between the respective image parts and the prototype. To regularise the latent space, distances between samples and prototypes of the same class were maximized whereas distances between different classes were maximized, as illustrated in Fig. 3.6b.

Variations on the *ProtoPNet* architecture have been applied to several classification tasks in medical imaging [144–148]. Kim et al. [144] applied such an architecture to the classification of chest X-rays. Their network was extended with an *occurrence module*, predicting where in the image a certain prototype pattern was likely to occur. These prototype-specific weights were then used to pool the image representation into a single vector before computing a single image-prototype distance (as opposed to taking the minimum distance over all patches). Classification of chest X-rays was also considered by [146] and Singh and Yow [147], where they modified the distance computations between the image and prototypes [146] or the loss function to train the last FC layer weights [147]. Barnett et al. [148] proposed a variation on *ProtoPNet* for skin melanoma classification, using fine-grained annotations to enforce prototypes to be similar to medically relevant image regions. This resulted in more representative prototypes, but required the availability of such annotations during training. Finally, Wolf et al. [149] proposed an interpretable model for the classification of adult MR brain images, combining a prototype-based model for images with the interpretable analysis of tabular data.

Importantly, prototype-based explanation models are inherently interpretable as they are able to provide the distances and prototype images the prediction is based on. While not being completely transparent, due to the mapping from the

image to the latent space, the explanations are by definition faithful to the model prediction as they represent the actual reasoning process, rather than approximating it *post-hoc*. Furthermore, they can provide both local (the decision-making process for a single sample) and global explanations (the set of learned prototypes).

The *ProtoPNet* architecture has not been applied for regression as the regularisation of the latent space requires the notion of *same* and *different* class. An extension of a prototype-based model for regression will be presented in Chapter 7 of this thesis.

3.2.2 Counterfactuals

Counterfactual visual explanations aim to generate an image representation that demonstrates what changes to an image would result in a different prediction [150]. This thus not only demonstrates *which* image areas are important, as in attribution-based explanations, but also *how* the areas affect the prediction (i.e. by showing how the image would look for a different prediction).

One of the first works applying counterfactuals in medical imaging was by Baumgartner et al. [151]. In that work, counterfactuals were produced by finding a (minimal) additive map that, when added to the sample image, changed the prediction to a given class (i.e. diseased versus healthy). The additive maps were created by training a generative adversarial network (GAN) [152] to produce realistic images that would be classified as the other class. The approach was demonstrated by transforming MR images of patients with Alzheimer's disease to resemble Mild Cognitive Impaired images. Similarly, Narayanaswamy et al. [153] and Schutte et al. [154] used GANs to generate images that changed the prediction with minimal image changes. Effectively, this also transformed images from one class to the other. Narayanaswamy et al. [153] applied this for the detection of diabetic macular oedema from retinal images, whereas Schutte et al. [154] to knee osteoarthritis grading from X-ray images. However, a disadvantage of such approaches is that only a single deterministic translated image is created per sample, whereas multiple different changes could result in a different prediction.

A different approach was introduced by Pawlowski et al. [155], which integrated a causal graph with a generative model. The causal graph was designed based on the prediction problem (e.g. including nodes for age and sex for brain MR images), and the generative model was able to generate counterfactual images with respect to changes in the causal graph (i.e. what would the brain image look like for an increased age). This approach was demonstrated on a toy MNIST dataset and a brain MRI dataset. For the brain MRI dataset, whole brain volume, ventricular volume, sex and age were all incorporated into the causal graph, enabling counterfactual image generation with respect to these parameters. Around the same time, Interpretable Classification via disentangled representations and feature attribution Mapping (ICAM) was proposed by Bass et al. [156]. In their approach, a continuous latent space was learned in which class-specific and class-irrelevant (i.e. subject-specific) features were disentangled. Image reconstructions were learned by a combination of a variational autoencoder (VAE) [157] and a GAN loss. The reconstructions were then used to analyse the prediction during test time by sampling from the latent space. Like Baumgartner et al. [151], they applied it to the detection of Alzheimer's disease from brain MRI scans. To make ICAM also suitable for the explanation of continuous variables, an extension of this model was proposed by Bass et al. [158], named ICAM-reg. This method was demonstrated on several regression tasks for adult brain MRI, including brain age prediction.

3.2.3 Distance-based

One of the most straightforward interpretable models is learning to map the data to a meaningful latent representation space in which inference can be done through distance-based predictions. Explanations can then be provided in the form of the closest samples in the training set. While not specifically designed for interpretability, such methods can thus provide an intuitive explanation of their prediction. These methods are also closely related to prototype learning as both approaches regularise distances in latent space (Fig. 3.6). While reviewing all such

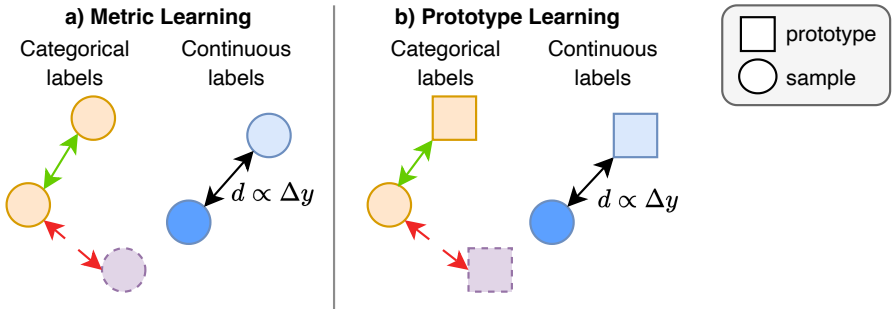


Figure 3.6: Schematic overview of (a) metric and (b) prototype learning for categorical and continuous labels, illustrating the similarities between the two approaches. Colours represent different classes or continuous values. The training objective for classification is minimizing the distances between samples of the same class (green arrows) while maximizing those of different classes. For regression, the distances are trained to be proportionate to their label difference.

approaches is considered out of scope for this review, I will give a summary of how metric learning approaches have been applied for regression and interpretability.

Learning a mapping to a meaningful representation space is often referred to as (deep) metric learning, and aims to generate a representation space in which distances between similar samples are small and, inversely, distances between dissimilar samples are large [159]. In metric learning, the mapping by a neural network is typically achieved by training with contrastive or triplet losses [160]. These losses were initially developed for tasks with a clear distinction between similar and dissimilar samples, such as classification, and were less suitable for regression tasks with continuous labels [159, 161]. Early work simply quantised the continuous differences [162, 163], but more recent extensions have also adapted the triplet loss for continuous labels [164–167]. Specifically designed for regression, Chao et al. [168] suggested regularising the latent space by enforcing the feature distances between samples in a batch, computed with the Euclidean distance, to be proportionate to their difference in label space. Furthermore, samples were weighted by a Gaussian function so as to weigh samples close to each other more than those far apart. This was motivated by the fact that distances in the feature space should be computed along the manifold (*geodesic distances*), and that simple distance measures, such as the Euclidean distance, are only a good approximation

for small distances. Inference was then done using a weighted mean of neighbouring training sample labels, and these could be visualised as an explanation.

Notable are also some recent methods developed for natural images that used metric learning in combination with patch-based matching to provide improved interpretability [169–171]. This approach was used for several tasks such as image retrieval [170], face identification [171] and few-shot learning [169]. To provide patch-based matching, OT optimisation was used, which will be discussed in more detail in Chapter 7.

3.2.4 Limitations of Example-based Explanations

While example-based explanations can be very effective, there are also a number of limitations of current methods that should be considered.

Challenges in Training Both prototype and counterfactual models can be challenging to train due to the large number of hyperparameters that need to be tuned. For example, ICAM-reg [158] has 10 tunable parameters just to weight the loss elements, whereas the *ProtoPNet* architecture requires three stages for training, each having separate learning rates and training durations that need to be optimised. Furthermore, counterfactual methods (as well as early prototype-based models using generative models) require training generative models to produce realistic explanations. This is known to be a challenging task requiring large amounts of training data, which might not always be available, especially when working with medical images.

Distance-based explanations are typically easier to train, but all latent representations of the training set images need to be available at inference to provide a nearest neighbour explanation. These can either be stored in memory, or accessed at inference, slowing down the predictions substantially. Additionally, during training, the distance-based losses normally rely on regularising distances within a batch, requiring large batch sizes. Both of these factors can be problematic for

large datasets (or datasets consisting of large images) due to memory constraints, especially for 3D volumes, such as 3D US.

Level of Detail in Explanation Both distance and prototype-based approaches typically provide comparisons to whole images or to image parts as an explanation. This is by definition of a lower resolution than visual attribution methods, which typically provide pixel-level importance scores. Counterfactual explanations can, on the other hand, provide pixel-level maps by computing difference maps between the sample image and the generated image.

Classification Only While showing promising results for medical imaging, the *ProtoPNet* has so far only been applied for classification tasks [143]. Due to loss elements that are based on the notion of *same class* and *different class*, these can not be directly applied to regression.

3.3 Interpretability for Brain Age Predictions

Several studies have attempted to introduce interpretability into brain age prediction models, predominantly for adult MRI. Most commonly, visual attribution methods were used to explain brain age predictions [10, 172, 173].

Dinsdale et al. [10] used backpropagation [104] to create saliency maps. The saliency maps were averaged across the whole test set and used to illustrate the differences between brain age models trained with linearly and non-linearly registered MR images. Alternatively, Hofmann et al. [173] used LRP to explain brain age predictions from MR images. The relevance with respect to the mean prediction value was computed (the mean was encoded in the bias term of the last layer), meaning that positive relevance indicated an increase, and negative a decrease from the mean value. However, due to the aforementioned problems with baseline-based attributions [131, 132] these results have to be interpreted with care. Finally, Lombardi et al. [172] used SHAP and LIME to explain a brain age prediction model from MR images. However, the brain age prediction model was not trained with the

MR image data but with a set of morphological features previously extracted from the images. The resulting explanations are therefore the attributions of higher-level features as opposed to image pixels.

In a different study, explaining brain MR classification rather than regression, several explanation methods were compared [174] ($\mathbf{G} \odot \mathbf{X}$, LRP, GBP and occlusion). This study showed that the visual attribution methods are not always consistent or comparable between different methods, confirming earlier introduced limitations of visual attribution methods (Section 3.1.6).

While not specifically a method for interpretability, predicting *local* brain age can also be considered under the umbrella of interpretability. These approaches attempt to determine which image areas (or features) are used most during the prediction. Both Bintsi et al. [175] and Ballester et al. [176] used patch- or slice-based approaches to predict local brain age. While these methods provide a more detailed prediction, separate networks had to be trained for each slice or patch, increasing the computational overhead. Using only a single model, Popescu et al. [177] used a U-Net to predict brain age voxel-wise. The predictions were more fine-grained than previous slice- or patch-based approaches, but the reported prediction error was considerably higher than the error obtained with baseline models (average MAE ~10 years vs. 3 years).

Alternatively, as described in Section 3.2.2, in [151, 155, 158], generative models were used to demonstrate the changes that would be expected in the image for different ages. However, as discussed, training generative models is challenging and typically requires large amounts of training data which is not always available.

4 Datasets

In this thesis, two datasets containing 3D US images of the fetal brain from healthy fetuses were used: the INTERGROWTH-21st dataset [24] and a dataset from the Leiden University Medical Centre (LUMC). The last contribution of this thesis (Chapter 7) also demonstrates the proposed method on an adult brain MRI dataset (the IXI dataset). Each of these datasets will now be introduced in more detail.

4.1 INTERGROWTH-21st Dataset

The INTERGROWTH-21st dataset is the main dataset used throughout this thesis. This dataset contains 3D US images acquired as part of the INTERGROWTH-21st Fetal Growth Longitudinal Study [24] (FGLS). That study aimed to describe human growth and neurodevelopment from early pregnancy up to 2 years of age in an optimally healthy population drawn from eight urban areas worldwide. Inclusion criteria specified healthy women who met a list of strict criteria based on, for example, age, BMI and medical history [24]. The initial inclusion was done before 14 GW, and subsequently, the women received an US examination every 5 weeks until birth. A total of 4607 women were initially included, of which 4321 delivered a baby without congenital malformations. Due to the rigorous selection criteria,

this population can be considered as optimally healthy, which enables the study of healthy development in this thesis.

Each woman included in INTERGROWTH-21st FGLS underwent multiple scanning sessions, resulting in a total of 19702 sessions that were included in the dataset. During each session, both 2D and 3D US images of the fetal brain were acquired on a Philips US machine (Philips HD-9, Philips Ultrasound, USA) with a curvilinear abdominal transducer. In this thesis, I will only use the 3D image volumes and perform any 2D experiments on planes extracted from the 3D volumes. All 3D brain images in this dataset were acquired from the axial plane.

Age Estimation In this thesis, fetal brain development is assessed with respect to the GA. For this reason, it is important to have an accurate measure of ground-truth GA. The ground-truth GA in the INTERGROWTH-21st dataset was determined by the LMP, and confirmed with an US-based CRL measurement before 14 GWs. If this measurement deviated by more than 7 days from the LMP, the fetus was excluded from the dataset. Furthermore, to be included in the study, women needed to have a regular menstrual cycle and know their LMP accurately [24]. While always being an estimate of the true post-conceptual age (as described in Section 2.1.4), these criteria thus ensured an accurate estimation of ground-truth GA.

4.1.1 Previous Data Processing

Due to the characteristics of US, such as the presence of shadows and speckle (described in Section 2.1.1), the quality of the acquired US volumes can vary considerably. In order to develop algorithms based on the presence of structural information, it is important to only use the US volumes in which sufficient anatomical information is present. Both initial data selection and initial processing of the images in the INTERGROWTH-21st dataset were carried out by Prof. Namburete in previous work [25]. From the full dataset, two subsets were selected: the *atlas* subset and the *validation* subset. An overview of the age distribution in the two data subsets is shown in Fig. 4.1a.

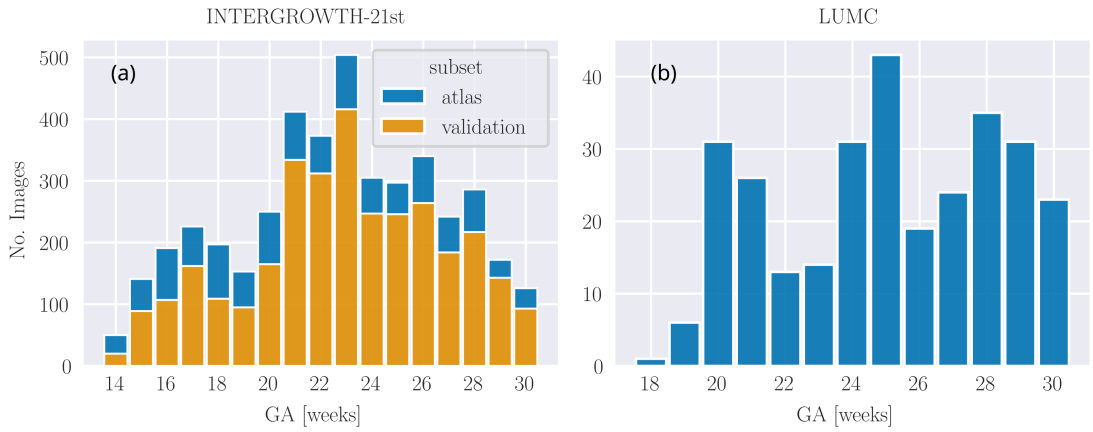


Figure 4.1: Data distribution across GA in the (a) INTERGROWTH-21st and (b) LUMC datasets, Chapter 5 uses images between 18 and 26 GWs, Chapter 6 between 18 and 28 GWs, and Chapter 7 the full age range. The combination of the atlas and validation subsets in the INTERGROWTH-21st dataset is referred to as the *extended* subset.

Atlas Subset A subset of 1062 3D image volumes with optimal US quality was selected from the INTERGROWTH-21st dataset to create a weekly US-specific atlas between 14 and 30 GWs [25]. For that selection, an extensive list of selection criteria was used, which is detailed in Appendix A Fig. A.1. In summary, the most important selection criteria were:

(i) Absence of motion artefacts

(ii) Head fully contained in the image

(iii) Minimal acoustic shadows in the distal hemisphere

(iv) High contrast between fluid-filled and dense (i.e. CP, skull) regions

To create the atlas, all volumes were manually aligned to the same reference coordinate system with a rigid transformation, resampled to an isotropic voxel size of 0.6 mm and resized to a size of 160x160x160 voxels. The atlas was then created using Demons diffeomorphic groupwise registration per GW [25], which mapped all images from the rigidly aligned orientation to a reference atlas space. The registered images were subsequently averaged to obtain a template image for each GW (i.e. average intensity images). Due to the diffeomorphic nature of the transformations,

the inverse transformation can be used to map annotations from the atlas space back to individual volumes, which will be used in Chapter 5.

For each of the images in the atlas subset, whole brain masks are available, which were used to compute whole brain volumes in Chapter 5. These masks were obtained by binarising template images in an MRI fetal brain atlas [61], and subsequently manually aligning these to the individual US images using an affine transform (rigid + scale). The MRI atlas contained template images for each GW between 21 and 31. For each US image, the MRI-derived mask of the corresponding GW was used, and for fetuses younger than 21 GWs, the template image of 21 GWs was used.

Extended Subset To extend the selected subset with US images of moderate quality, another subset was selected in previous work [9], which I will refer to as the *validation* subset. To perform this selection, some selection criteria were relaxed, such as allowing a lower contrast between fluid-filled and dense regions and accepting a higher number of acoustic shadows. In Appendix A Fig. A.1 the selection criteria that were relaxed for the validation dataset are indicated with a (*).

For the validation subset a total of 3203 image volumes were selected. These volumes were also resampled to an isotropic voxel size of 0.6 mm and resized to a size of 160x160x160. All volumes were then aligned to a common reference system using an automated algorithm, referred to as BEAN [178]. The combination of the *atlas* and *validation* subset will be referred to as the *extended* subset. To ensure equal processing between the two data parts, the *atlas* volumes in this subset were re-aligned with BEAN [178] (as opposed to the previous manual alignment).

4.1.2 Selection in this Thesis

An overview of the images selected for each of the chapters in this thesis can be found in Table 4.1. In Chapter 5 the atlas subset was used, and in Chapter 6 and 7 the extended subset. In Chapter 5, I used manual annotations in the atlas space to generate weak segmentation labels for the individual images. These were obtained by propagating the labels using the inverse transformation obtained during

Table 4.1: Overview of images of the INTERGROWTH-21st dataset used in each contribution chapter. 2D images consist of transventricular planes extracted from 3D image volumes.

	2D/3D	task	age range [weeks]	subset	no. images
Chapter 5	3D	subc. segm.	18-26	atlas	537
Chapter 6	2D&3D	brain age	18-28	extended	3359
Chapter 7	2D	brain age	14-30	extended	4265

atlas registration. For this reason, only image volumes of the atlas subset could be used for model development in this chapter. Furthermore, when completing the work in Chapter 5, the extended subset was not readily aligned to a common reference system, as BEAN had not yet been developed. For this reason, only the atlas subset was used in Chapter 5. In Chapter 6 and 7 no label propagation was needed, and therefore the larger extended subset was used.

Due to the different requirements, I used a dedicated age range for each chapter. In the subcortical segmentation chapter (Chapter 5), I used an age range between 18 and 26 GWs. Before 18 weeks the structures of interest were too small to consistently segment, whereas past 26 weeks the (manual) segmentation task became too challenging due to a decrease in the quality of the volumes (i.e. due to acoustic shadowing). In Chapter 6 I regularised the attention of a brain age prediction model with segmentation masks, however, the masks do not have to be pixel-level accurate. To make clinically meaning predictions, it is desired to have an age range as large as possible, and the age range was, therefore, extended to 28 weeks. Lastly, in Chapter 7 the full extended subset was used, ranging from 14 to 30 GWs. Each chapter will contain an overview of the number of images available in the respective age range. In Fig. 4.2, example volumes for both the atlas and validation set are shown across the GA range.

4.1.3 Additional Preprocessing

In Chapter 5 the atlas dataset was used without any additional processing. However, for the extended dataset used in Chapter 6 and 7, I performed some additional preprocessing steps.

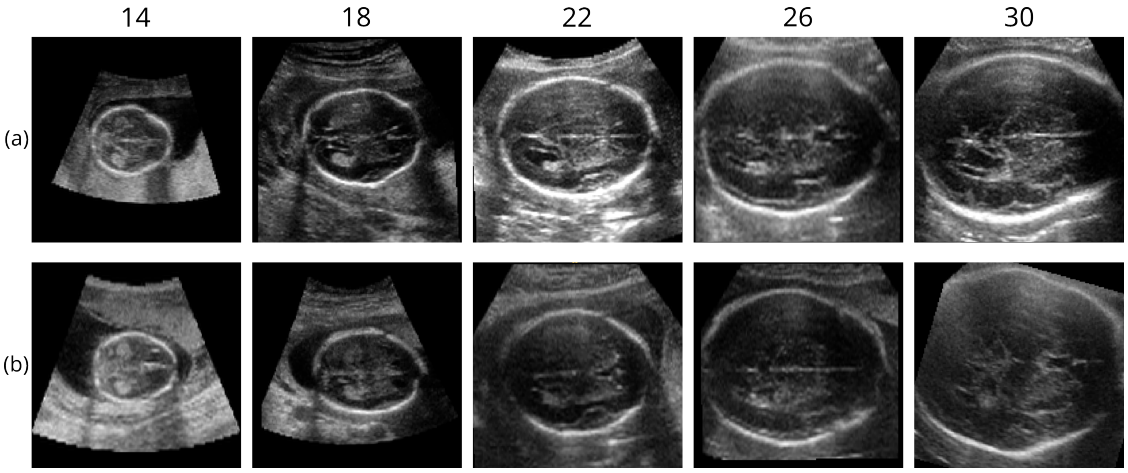


Figure 4.2: Example US volumes from the INTERGROWTH-21st dataset over the GA range. Images shown have been aligned to a common coordinates system but not size normalised. For visualisation the images were also mirrored across the midline so that always the bottom hemisphere is visible. (a) shows examples from the atlas subset, and (b) from the validation subset. It can be seen that the atlas images are of slightly better average quality than the validation images (i.e. better contrast and visibility of anatomical details).

Size Normalisation As described in Section 2.2.2, an important consideration for brain age prediction models is whether to use size features [86, 88]. In this thesis, I am interested in quantifying the stage of brain development and therefore do not want to use size information. To enforce the network to focus on structural developmental patterns rather than on size alone, all image volumes were spatially normalised to the average brain volume at 30 GWs. This was done using the same automated algorithm used for the alignment of the volumes (BEAN) [178].

2D Slice Extraction To perform experiments with 2D images, transventricular standard planes were extracted from the aligned and size-normalised 3D volumes. The planes were extracted using a set of fixed coordinates, and interpolated with bilinear interpolation. Examples of extracted transventricular planes across gestation for the INTERGROWTH-21st dataset are shown in Fig. 4.3a.

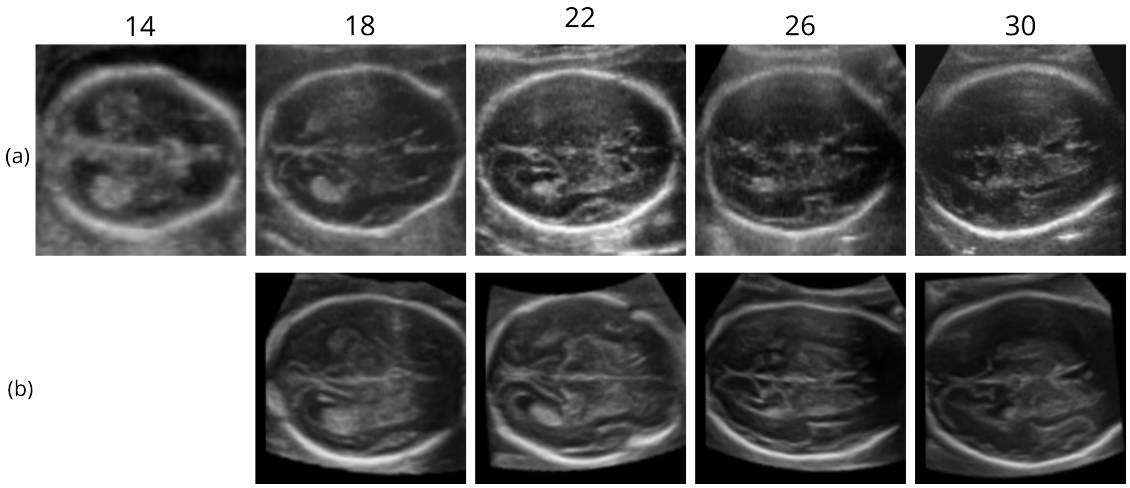


Figure 4.3: Example transventricular planes extracted from the (a) extended dataset, and (b) from the LUMC dataset. The LUMC dataset spans an age range from 18 GWs, and is therefore not shown at 14 GW. 2D images consist of transventricular planes extracted from 3D image volumes.

4.2 LUMC Dataset

To test the generalisability of the methods developed in this thesis, I also tested the performance on an independent dataset, namely the LUMC dataset. This dataset was acquired in the LUMC as part of a collaboration during my DPhil to study the brain development of children with CHD. For that study, 3D brain images of both healthy fetuses and fetuses with congenital heart disease were acquired with a Canon Aplio I800. The women were recruited from around 20 GWs and subsequently scanned at 4-week intervals. At each scanning session, multiple 3D volumes were typically acquired. In this thesis, the aim of an independent dataset is to validate the generalisability performance and therefore only images of the healthy fetuses were used. The images used for testing in each chapter were selected to match the age range used during model training. Example images of the LUMC dataset across the GA range are shown in Fig. 4.4.

Data Selection Data selection was carried out by Dr Moska Aliasi, an ultrasonographer in the LUMC, after a joint discussion between us to determine the selection protocol. During the data selection, four subcortical structures (CB, CP, LPVH and CSPV) were assessed, along with the SF. For each individual structure,

Table 4.2: Overview of images of the LUMC dataset used in each contribution chapter. Images were selected based on the visibility of the structures noted in the selection criteria column. For Chapter 5 the no. of images indicates the number of images on which any one of the subcortical structures were visible.

	2D/3D	age range [weeks]	selection criteria	no. images
Chapter 5	3D	18-26	CP, CB, LPVH, CSPV	149
Chapter 6	2D&3D	18-28	SF	156
Chapter 7	2D	14-30	SF	185

a single 3D image was selected (of the multiple acquired images during a single session) on which the respective structure was most clearly visible. If a structure was not well visible in any of the image volumes acquired in a session, the structure was excluded from the respective scan session. This resulted in a different subset of included image volumes for each of the structures.

In Chapter 5 the subset of volumes in which any of the subcortical structures were visible was selected. The segmentation algorithm was applied to all these volumes, and the performance is only reported for the images where the respective structure was visible. In Chapter 6 and 7 the subset of images on which the SF is visible was used as this was considered the most important structure for brain age predictions. An overview of the images from the LUMC dataset used in each chapter is shown in Fig. 4.2.

Preprocessing The LUMC image volumes were preprocessed in the same way as the INTERGROWTH-21st volumes. This involved resampling them to an isotropic voxel size, resizing them and then aligning them to the same coordinate system using BEAN [178]. Lastly, for Chapter 6 and 7 all volumes were spatially normalised and 2D transventricular planes were extracted. An overview of example extracted planes from this dataset is shown in Fig 4.3b.

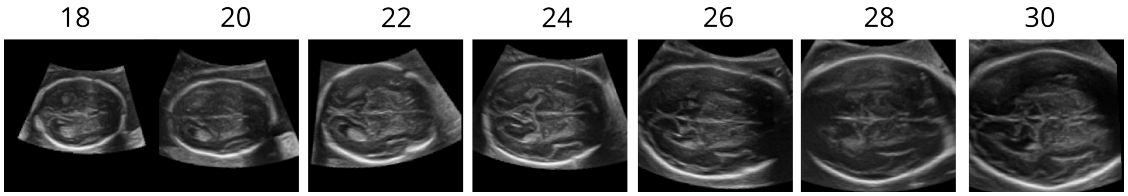


Figure 4.4: Example US volumes over the GA range from the LUMC dataset. It can be observed that the volumes have slightly better soft tissue contrast than the INTERGROWTH-21st volumes, especially at the higher GAs.

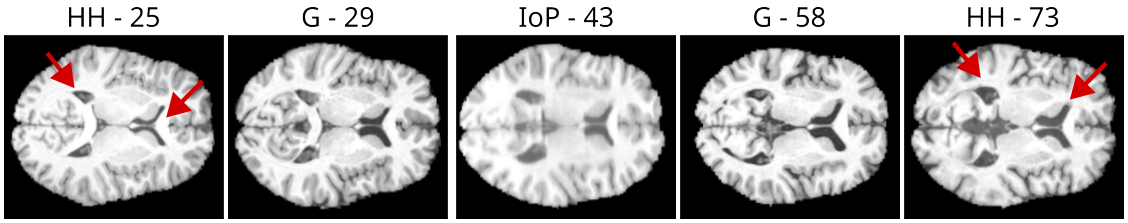


Figure 4.5: Example slices extracted from the 3D MR brain images in the IXI dataset. The title of each image indicates the institution where the image was acquired (HH: Hammersmith Hospital, G: Guy's Hospital, IoP: Institute of Psychiatry), followed by the chronological age in years. The ventricles have been indicated with red arrows in the first and last image, from which the difference in ventricular volume can be clearly observed.

4.3 IXI MRI Dataset

In Chapter 7, I also demonstrate the developed method on a dataset of T1 MR images of normal, healthy subjects. This dataset is publicly available[1] and has been acquired in three different institutions in London, UK. Preprocessing of this dataset has been performed by my colleague Dr Dinsdale using the FSL Anat pipeline[2]. This pipeline involved skull stripping, bias field correction and linear registration to the 1 mm MNI template [179]. Subjects that did not successfully complete this pipeline were excluded for further analysis, which resulted in a total of 561 images between the age of 20 and 86 years. From each preprocessed 3D image volume, I then extracted an axial plane containing the ventricles. These 2D images were then used in Chapter 7. Example slices from the IXI MR images are shown in Fig. 4.5.

[1] `https://brain-development.org/ixi-dataset/`
[2] `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/fsl_anat`

5

Subcortical Segmentation of the Fetal Brain

Contents

5.1 Introduction 68
5.2 Methods 71
5.2.1 Notation 71
5.2.2 Network Architecture 72
5.2.3 Few-Shot Training Labels 72
5.2.4 Manual Annotation Process 76
5.3 Experimental Set-Up 78
5.3.1 Dataset and Preprocessing 78
5.3.2 Experiments 79
5.3.3 Implementation 82
5.3.4 Evaluation 82
5.3.5 Volumetric Subcortical Growth Trajectories 84
5.4 Results 85
5.4.1 Weak Atlas Labels 85
5.4.2 Comparison between Atlas and Expert Labels 85
5.4.3 Varying the Number of Manual Annotations 86
5.4.4 Effect of Image Alignment 87
5.4.5 Comparison to Previous Work 88
5.4.6 Qualitative Results 91
5.4.7 Clustered Template Labels 91
5.4.8 Subcortical Growth Curves 92
5.4.9 Generalisability 94
5.5 Discussion 98

In this chapter, I study the the quantifications of subcortical development using automated deep learning-based segmentations in 3D US. As obtaining manual ground-truth annotations is costly, I explored how a segmentation network can be optimally trained using only a limited number of annotations. Subsequently, I applied this network to a large cohort of healthy fetuses, generating novel US-specific growth curves for four subcortical structures of interest. The work presented in this chapter has been published in:

> **Linde S. Hesse**, Moska Aliasi, Felipe Moser, Monique C. Haak, Weidi Xie, Mark Jenkinson, Ana IL Namburete, and the INTERGROWTH-21st Consortium. "Subcortical segmentation of the fetal brain in 3D ultrasound using deep learning." *NeuroImage* (2022): 119117.

Early versions of this work have also been presented at:

> Moska Aliasi, **Linde S. Hesse**, Madeleine K. Wyburd, Maartje Snoep, Ana I. L. Namburete, and Monique C. Haak. "Success rate of the subcortical 3D fetal brain segmentation using deep learning algorithms." *Ultrasound in Obstetrics & Gynecology,* 2022

> **Linde S. Hesse**, Moska Aliasi, Monique C. Haak, Aris T. Papageorghiou, Weidi Xie, Mark Jenkinson and Ana IL Namburete "Few-shot subcortical brain structure segmentation in 3D fetal brain ultrasound." *Organisation of Human Brain Mapping (OHBM),* 2020

5.1 Introduction

During routine US examinations, the size and appearance of several subcortical structures in the fetal brain are assessed. The abnormal development of these structures can be a potential sign of CNS abnormalities, and it is, therefore, important to monitor them during gestation. Their full structural development is best quantitatively assessed with 3D volumetric information. However, as described in Section 2.2.1.2, most studies analysing these structures during gestation only use

in-plane measurements obtained from 2D US [180, 181]. While these are routinely acquired during a basic examination of the fetal brain, they only provide a limited representation of anatomical development. For this reason, it is desirable to use 3D US to obtain volumetric measures of subcortical development. As manually annotating several structures is not feasible in routine clinical practice, automated subcortical segmentation methods for 3D US could facilitate this analysis and may provide new insights into *in utero* subcortical development.

Structural segmentation in fetal brain US is a challenging task due to the low soft tissue contrast, reverberation artefacts and the characteristic presence of speckle. Consequently, precise structural boundaries can be hard to distinguish, resulting in high inter- and intra-observer variability in manual annotations. Since manual segmentation is not a task usually performed in clinical practice, even trained ultrasonographers can have difficulty in accurately segmenting subcortical structures in 3D US volumes.

While only limited previous studies have explored automated subcortical segmentation in 3D fetal US (as described in Section 2.2.1), subcortical segmentation has also been studied in MR images of the fetal brain [61, 182–185]. These studies use predominantly atlas-based segmentations, in which annotations from an *atlas* space are used to segment individual image volumes [186]. The atlas space can consist of a number of individually annotated images [183–185], or of an average intensity image (referred to as *template* image) that is annotated [61]. These annotations are then used to segment a new image by using image registration methods. Due to the lack of an annotated atlas for fetal brain US, atlas-based methods have not been applied widely for brain structure segmentation in fetal US.

On the other hand, it has been shown that deep learning methods can be successfully applied to other segmentation tasks in 3D US volumes of the fetal brain [4, 8, 187, 188], resulting in higher performance than traditional segmentation methods. However, due to the difficulty of obtaining manual annotations for subcortical structures, a key barrier to applying deep learning methods to this task is obtaining sufficient ground-truth annotations for training.

To overcome the need for a large manually annotated dataset, few-shot learning can be used, in which only a small number of manual annotations are used to train a CNN. Several few-shot learning approaches have been proposed for segmentation tasks in the medical image domain [189–191], showing that good segmentation performance can be obtained using only a very limited amount of voxel-wise manual annotations. In this chapter, I will develop a deep learning-based method for the segmentation of several subcortical structures in 3D US in a low-data regime. Inspired by previous atlas-based segmentation methods in MRI [61], I will compare two types of ground-truth labels for training, where both have been obtained using only a few manual annotations: (1) a small number of individually annotated US volumes (referred to as *expert labels*), and (2) a large number of weakly labelled US volumes obtained by propagating annotated template images (referred to as *atlas labels*).

Additionally, I investigate the impact of image alignment to a common coordinate system as a preprocessing step. As the anatomical orientation of the fetal brain varies in US image volumes, initial global (affine) registration of the brain is expected to have a positive effect on the segmentation performance. Although the alignment of US volumes is a non-trivial task, methods have been developed for the automated alignment of fetal brain volumes [192, 193]. Furthermore, compared to the voxel-wise annotation of several subcortical structures, manual image alignment requires considerably less effort. For these reasons, alignment of the volumes to a common coordinate can be preferred over the additional voxel-wise manual annotation, and, as such, it is important to explore global alignment as a requisite preprocessing step.

In this work, I aim to segment the CP, LPVH, CB, and CSPV during the second trimester of gestation. This trimester is of particular interest since this is when women undergo an US examination as part of routine care to screen for anomalies [194]. Furthermore, in this trimester most subcortical structures have developed enough to be visible on US but are less affected by acoustic shading due to progressing calcification of the skull and amniotic fluid reduction later in gestation.

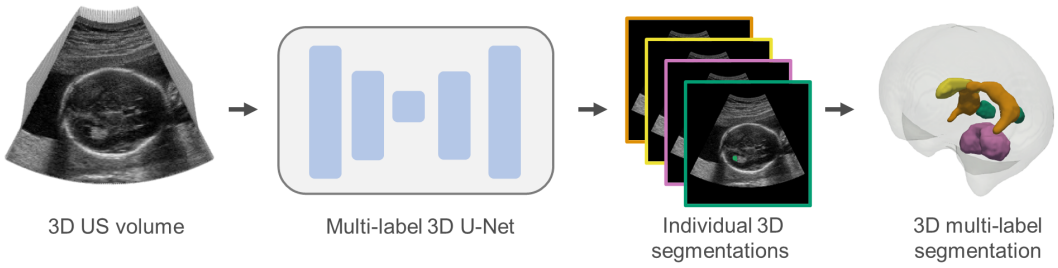


Figure 5.1: Schematic overview of segmentation pipeline. The segmented structures shown are the CB (pink), LPVH (green), CSPV (yellow) and CP (orange). The LPVH and CP are shown in both hemispheres for visualisation purposes but are only segmented in the visible hemisphere. A detailed overview of the U-Net architecture is shown in Fig. 5.2.

Finally, to obtain an improved understanding of volumetric subcortical development during the second trimester, the developed segmentation model will be applied to a large cohort of healthy fetuses. The resulting model predictions will be validated for the CB, as volumetric growth curves for the CB have been reported in previous studies [72, 76–78, 82, 84, 195], and subsequently used to generate novel US-specific growth curves of subcortical volume development during the second trimester of gestation.

5.2 Methods

In this chapter, a single segmentation network is developed that segments the CB, LPVH, CP and CSPV in 3D US volumes using a minimal number of manual annotations for training. A schematic overview of this pipeline is shown in Fig. 5.1.

5.2.1 Notation

Given a dataset of m image volumes, the set of image volumes in their original acquired orientation is denoted as $\breve{\mathcal{X}} = \{\breve{\mathbf{X}}_1, ..., \breve{\mathbf{X}}_m\}$ with $\breve{\mathbf{X}} \in \mathbb{R}^3$, and the set of image volumes rigidly aligned to the same reference coordinate system as $\dot{\mathcal{X}} = \{\dot{\mathbf{X}}_1, ..., \dot{\mathbf{X}}_m\}$. Referring to either $\breve{\mathcal{X}}$ or $\dot{\mathcal{X}}$ will simply be done by using $\mathcal{X}$, and the original and manually aligned images will also be referred to as *unaligned* and *aligned*, respectively. The ground-truth multi-label segmentation masks are denoted by $\mathbf{Y}$. The binary mask of a single class will be denoted by a subscript c, with c

being the class (structure), and the type of the ground-truth label (abbreviated as *atl* for atlas labels and *exp* for expert labels) is marked with a superscript ($\mathbf{Y}_c^{type}$). Similarly to the images, the set of labels will be denoted by $\mathcal{Y}$.

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

5.2.2 Network Architecture

A multi-label 3D U-Net [63] with batch normalisation [196] was used as network architecture for the subcortical segmentation. The network, denoted by θ, predicts a multichannel segmentation, denoted by $\hat{\mathbf{Y}}$, for a 3D US input image: $\hat{\mathbf{Y}} = \theta(\mathbf{X})$, with each channel corresponding to a class. A schematic overview of this pipeline is given in Fig. 5.1. Following the same convention as for the images, $\breve{\theta}$ and $\dot{\theta}$ correspond to networks trained with $\breve{\mathcal{X}}$ and $\dot{\mathcal{X}}$ respectively. To obtain probabilities for each class, a softmax function is applied across the channel dimension of $\hat{\mathbf{Y}}$, which will be denoted as $\hat{\mathbf{Y}}_c$. Final predictions for a single class are subsequently obtained by assigning each voxel to the class with the highest probability. Based on initial experiments, the network was designed with a depth of five, and with 16 feature maps in the initial block of the encoder. A schematic overview of the full network architecture is shown in Fig. 5.2. While this network obtained the best performance in initial experiments, the main objective of this chapter is not to study the effect of the model architecture, but to explore the optimal way of training a CNN with a small number of manual annotations.

5.2.3 Few-Shot Training Labels

Defining n_a as the number of images that are manually annotated, two types of training labels will be considered, both obtained from an equivalent number of manual annotations:

- **expert-based labels**: a set of n_a individually annotated images.
- **atlas-based labels**: a large number of images with weak propagated atlas labels obtained from annotating n_a 3D template images.

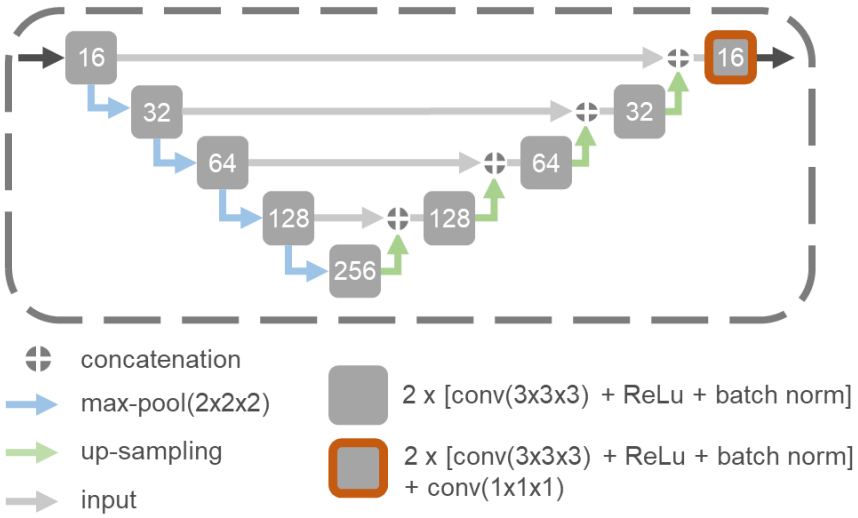


Figure 5.2: Overview of the 3D U-Net architecture used in this study. The numbers on top of the grey blocks indicate the number of feature maps in the respective layer. The input is a 3D ultrasound volume and the network outputs a multi-class segmentation of the same dimensions.

These labels will, in mathematical notation, be abbreviated by *exp* and *atl* respectively. The set of expert-based labels can be obtained by directly annotating the set of individual images. However, to generate the atlas-based labels some additional steps were performed that will now be detailed.

5.2.3.1 Atlas Labels

Atlas-based labels were obtained for the images by manually annotating 3D template images as opposed to annotating individual images (Fig. 5.3). Template images were created by registering a subset of s rigidly aligned images $\dot{\mathcal{S}} \in \dot{\mathcal{X}}$ to each other using a Demons diffeomorphic groupwise registration approach [25], resulting in a set of non-rigid transformations that can transform each image from the rigidly aligned orientation to a reference space R defined by: $\mathcal{T} = \{T_i : \dot{\mathbf{X}}_i \rightarrow \mathbf{X}_{i,R} | i = 1, ..., s\}$. The registered images were subsequently averaged to obtain a template image $\mathbf{K} = \frac{1}{s} \sum_{i=1}^{i=s} \mathbf{X}_{i,R}$.

In this template $\mathbf{K}$, the four subcortical structures of interest were manually annotated (see Section 5.2.4), resulting in a set of labelled template images. For simplicity, the annotated template images will also be referred to with $\mathbf{K}$. Subsequently, these labels were non-linearly registered back to the individual images

using the inverse transform of the registration to the reference space, given by: $\mathcal{T}_{inv} = \{T_i^{-1} : \mathbf{K} \rightarrow \dot{\mathbf{Y}}_i | i = 1, ..., s\}$. Due to anatomical variations between individuals and potential errors in the registration, the resulting labels may contain imperfections and are thus considered to be weak labels. To clearly differentiate between the annotated template images (which could be referred to as an atlas) and the propagated atlas-based labels, I will refer to the former only as *annotated template images* and to the latter as *atlas labels*.

Standard Template Construction In order to generate template images, rigidly aligned images from a GA range of 1 week were used, resulting in one template image per GW. To capture possible differences in anatomy between the left and right cerebral hemispheres, template images for each hemisphere were constructed separately using the respective images and fused together at the midsagittal plane by joining the interhemispheric fissure in the two templates (Fig. 5.3 top row). The set of annotated template images (one per GW) is denoted by $\mathcal{K}$, containing annotations of all segmented structures (LPVH, CB, CP and CSPV).

Cluster-Based Template Construction During initial experiments with the atlas labels, it was seen that the LPVH shape was not well captured by a single template image per GW. For this reason, I extended the atlas construction to generate multiple templates for this structure using a clustering approach (Fig. 5.3 bottom row). I selected images from an age range of two GWs, grouped them using k-medoids clustering and subsequently constructed n_c template images of the LPVH using this grouping. In our experiments, n_c was empirically set to four as for this number the resulting template images captured the most important shape variations (based on a visual assessment of the templates). As any anatomical variation between left and right can be captured by the clustering, images of the left and right hemispheres were combined by flipping the right side images across the midline. In the resulting clustered template images (denoted by $\mathbf{K}^{\mathbf{clust}}$) only the LPVH was manually annotated, resulting in an additional set of atlas labels for the LPVH. Unless explicitly stated otherwise, the labels propagated from the

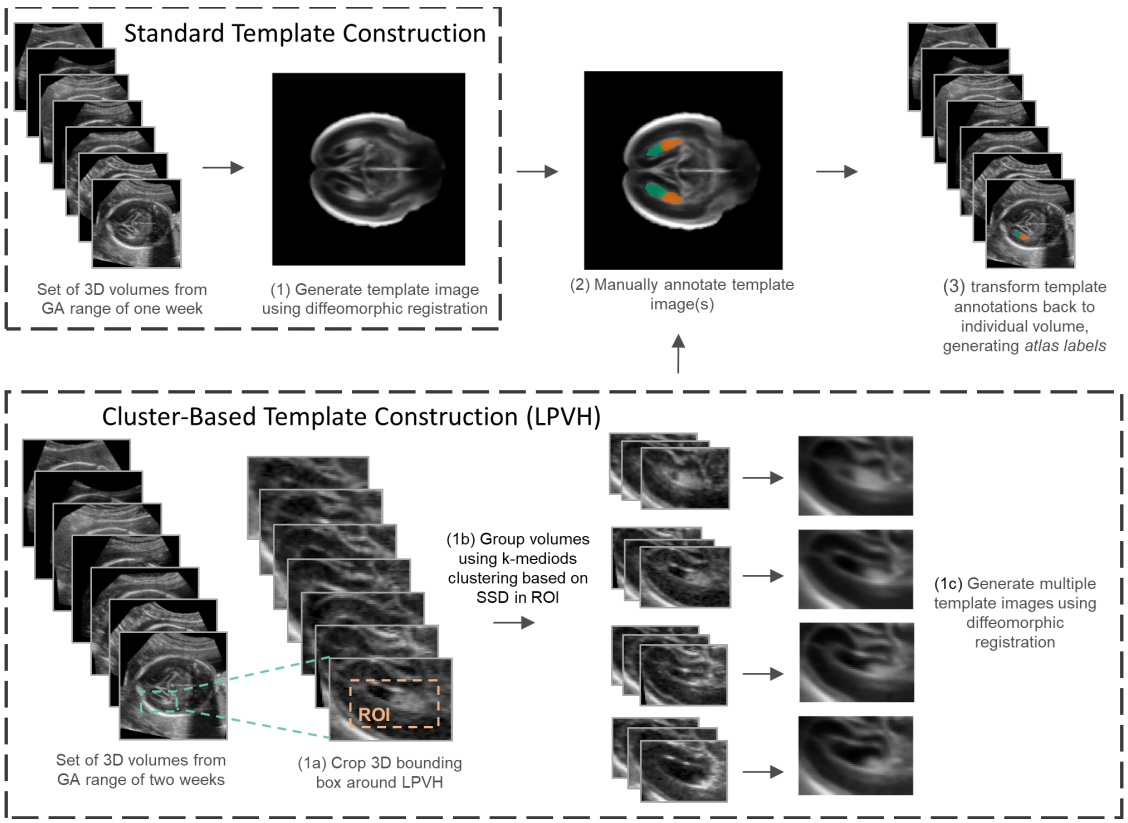


Figure 5.3: Schematic overview of atlas label generation. Standard whole brain templates (top row) were constructed for each GW and all structures (CB, CP, CSPV and LPVH) were annotated in these templates. Cluster-based template construction (bottom row) was only performed for the LPVH.

cluster-based template images are used for the LPVH in the remainder of this study (i.e. $\mathcal{Y}^{atl}_{LPVH}$ is obtained from the cluster-based template annotation).

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

to outliers, I excluded a few outliers based on their visual appearance prior to clustering. After the clustering step, each outlier was assigned to the nearest cluster based on the distance to the cluster centre. Once all images were assigned to a cluster, templates were created from the images in each cluster, including the outliers, as described earlier. The k-medoids clustering was set to repeat 500 times, and results containing clusters with less than three images were discarded.

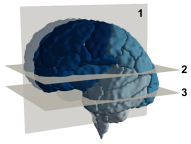
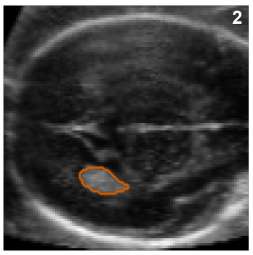
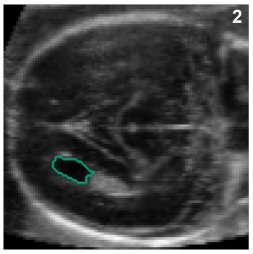
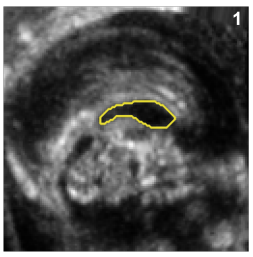
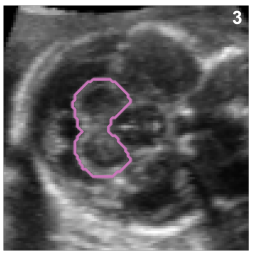
5.2.4 Manual Annotation Process

Both the constructed template images, $\mathcal{K}$ for all structures and $\mathcal{K}^{clust}$ for the LPVH, as well as a subset of individual images, $\dot{\mathcal{X}}$, had to be manually annotated. Since subcortical segmentation is not a task usually performed clinically, and structural boundaries can be very challenging to distinguish in 3D US images, I designed a bespoke segmentation protocol in consultation with experienced ultrasonographers (M. Aliasi and M.C. Haak). An overview of the manual segmentation protocol is shown in Table 5.1. Based on this protocol, I performed the manual annotations (for both template and individual images), and these were subsequently verified by M. Aliasi.

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

The annotation of the LPVH in both the left and right hemisphere in the standard template images also means that the annotation effort required to annotate the

Table 5.1: Overview of manual segmentation protocol.

Structure	Segmentation Protocol	Example
CP (*Choroid Plexus*) 	• Identify in transventricular plane as circular echogenic (bright) structure inside the ventricular cavity. • Segment complete CP, which moves from the midline (posteriorly) to a more lateral location (inferior). • Use all three planes, predominantly the axial and coronal, to ensure consistent segmentation.	
LPVH (*Lateral Posterior Ventricle Horn*)	• Ventricular cavity posterior to the CP. • Identify in transventricular plane as a hypoechoic (dark) cavity. • Segment inside the echogenic boundary. • Use all three planes to ensure consistent segmentation.	
CSPV (*Cavum Septum Pellucidum et Vergae*)	• Identify in the midsagittal plane as a fluid-filled cavity. • In older fetuses the cavum septum pellucidum (anterior) and cavum vergae (posterior) can be seen as two adjoining cavities but are segmented together. • Use the axial plane to determine the lateral boundaries of CSPV.	
CB (*Cerebellum*)	• Include bright echogenic boundary*. • Identify in the axial plane as the two cerebellar hemispheres connected by the vermis. • Determine superior and inferior boundaries in sagittal planes. • Use all three planes to ensure consistent segmentation.	

* Cerebellar boundary included for consistency as the TCD is clinically measured including this boundary

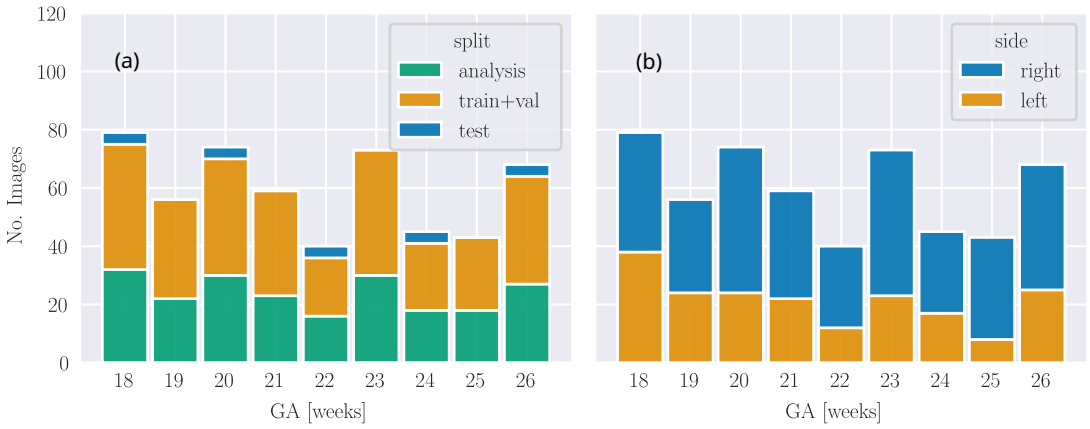


Figure 5.4: Data distribution of the 537 US images used in this chapter (a) shows the distribution over the subsets and (b) the proportion of images with a visible right or left hemisphere.

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

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

5.3 Experimental Set-Up

5.3.1 Dataset and Preprocessing

In this chapter, 3D US images from the *atlas subset* of the INTERGROWTH-21st dataset were used (as described in Chapter 4). All images in this subset between 18 and 26 weeks of gestation were selected, resulting in a total of 537 images. This age range in the second trimester was selected as the subcortical structures of interest are too small to accurately segment before 18 GWs, whereas after 26 GWs the images in the dataset were heavily affected by reduced amniotic fluid and acoustic shadowing of the fetal skull that becomes increasingly calcified. In Fig. 5.4 the data distribution over the GA range is shown.

Of these 537 images, 321 were used for model development and the remaining 216 to generate volumetric growth curves of the four subcortical structures (referred to as the *analysis* subset or $\mathcal{X}_{analysis}$). Of the 321 images used for model development, 20 were used for testing ($\mathcal{X}_{test}$) and the remaining for training ($\mathcal{X}_{train}$) and validation ($\mathcal{X}_{val}$). These 20 test set images were evenly distributed across the GA range, with four images for every second GW (at 18, 20, 22, 26, and 26 GWs). Furthermore, it was ensured that the training, validation, and analysis subsets were also evenly distributed across the GA range. I manually annotated all test images and all of these thus had expert ground-truth labels (as described in Section 5.2.4). Ten test images were twice manually annotated, in order to obtain intra-observer variability.

To obtain atlas labels for the dataset $\mathcal{Y}^{atl}$, template images had to be generated for the dataset ($\mathbf{K}$ and $\mathbf{K}^{clust}$). These template images were created using all images in the dataset ($\mathcal{X}_{train}$,$\mathcal{X}_{val}$,$\mathcal{X}_{test}$, $\mathcal{X}_{analysis}$), as these were available from earlier work [25]. However, the effect on the appearance of the template images of using all images versus only the one used for training is very small, and as such not expected to have a significant effect on the resulting atlas labels.

As the GA of the dataset used in this study ranges between 18 and 26, n_a for the training data was set to 9. This yielded two overlapping training sets, one with 301 images containing weak atlas labels, and one with 9 images containing expert labels. During training of the networks, 15% of the 301 was used as validation. For the 9 expertly labelled individual images, no validation set was used, but all settings were kept the same as when training with atlas labels. An overview of how the data were separated into different subsets, and the number of available labels in each set is summarised in Table 5.2.

5.3.2 Experiments

In order to study the effect of using only a small number of manual annotations, multiple experiments were performed. Models trained with either expert or atlas labels are denoted by θ^{atl} and θ^{exp}.

Table 5.2: Overview of the number of images and available labels for each of the data subsets in this chapter. All images originate from the atlas subset of the INTERGROWTH-21st dataset (as described in Chapter 4).

	volumes	atlas labels	expert labels
Analysis subset	216	0	0
Model Development subset			
Testing	20	20	20
Training	256	256	9
Validation	45	45	0

Comparison between Atlas and Expert Labels Firstly, I aimed to study the effect of using n_a expertly annotated images versus many weakly labelled images, obtained from annotating n_a template images. This first experiment was performed with images from the full GA range in the data, resulting in $n_a = 9$. Defining n_x as the number of training images used during training, two separate networks were trained: one that used all atlas-labelled images (θ^{atl}, $n_x = 215$) and a second that only used the expertly labelled images (θ^{exp}, $n_x = 9$). As the alignment of the images was expected to influence the performance, this experiment was repeated for the aligned images, $\dot{\mathcal{X}}$, as well as for images in their original orientation, $\breve{\mathcal{X}}$. Models trained with aligned images were evaluated on aligned test data, and vice-versa.

Varying the Number of Manual Annotations Next, n_a was reduced to study the performance decrease when using even fewer manual annotations for training. For $\mathcal{Y}^{exp}$, a simple subset of n_a images were used whereas for $\mathcal{Y}^{atl}$ all images obtained from n_a annotated templates were used. The models trained with these labels are noted as $\theta^{exp}_{n_a}$ and $\theta^{atl}_{n_a}$, and in the experiments n_a ranged from 2 to 9. The images used for training were selected to be uniformly distributed over the whole GA range, e.g. for $n_a = 2$ the images from 20 and 24 GWs were selected. Even though these models were only trained with images from selected GWs, evaluation was performed on all test images. As for the previous experiment, this was repeated for both $\breve{\mathcal{X}}$ and $\dot{\mathcal{X}}$.

Cluster-based atlas labels During initial testing, I observed that using atlas labels from multiple LPVH templates, $\mathbf{K}^{clust}$, increased performance compared

to using the labels from a single annotated template per GW, **K** (see Table 5.4). For this reason, all aforementioned experiments using atlas labels were performed with the LPVH annotations from the clustered template images. To quantify the performance increase from these improved template images, I also annotated the LPVH in **K**, and trained a network with these labels. As this only applies to the atlas labels, no expert training labels were used in this experiment.

5.3.2.1 Network Training

For training, I used a combination of multi-class squared DSC ($\mathcal{L}_{MD}$) [197] and Cross-Entropy (CE) loss ($\mathcal{L}_{CE}$), which are, for a single sample, defined by:

$$\mathcal{L}_{total} = \mathcal{L}_{MD} + \lambda\mathcal{L}_{CE} = (1 - \sum_{c=1}^{n_{class}} \frac{2|\mathbf{Y_c} \cap \hat{\mathbf{Y}}_\mathbf{c}|}{|(\mathbf{Y_c})^2| + |(\hat{\mathbf{Y}}_\mathbf{c})^2|}) - \lambda \sum_{c=1}^{n_{class}} \mathbf{Y_c} \log\Big(\hat{\mathbf{Y}}_\mathbf{c}\Big) \quad (5.1)$$

with λ as a relative weighting parameter between both loss terms, n_{class} the total number of classes (which equals 4 in this study), and $\hat{\mathbf{Y}}_c$ the predictions for each class before binarisation. It was observed during initial experiments that the DSC loss term resulted in higher performance for individual structures due to the high unbalance between the structures and background, whereas the addition of the CE loss term ensured convergence for all structures. Based on these experiments the λ was set to 1, but it did not have a strong effect on performance for values within the same order of magnitude (see Appendix B, Fig. B.1 for the quantitative result of varying λ).

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

$$\begin{aligned} s_{min}(w) &= max\left(\sqrt[3]{\frac{0.9 \cdot (v_{avg}(18))}{v_{avg}(w)}}, 1/1.5\right) \\ s_{max}(w) &= min\left(\sqrt[3]{\frac{1.1 \cdot v_{avg}(26)}{v_{avg}(w)}}, 1.5\right) \end{aligned} \quad (5.2)$$

with w the GA of the fetus in weeks and $v_{avg}(w)$ the average whole brain volume at a certain GW. This ensured that scaling was only performed to brain sizes within the dataset (i.e. images at 18 GW were predominantly up-scaled whereas images at 26 GW were mostly down-scaled). I applied augmentation for training with both $\dot{\mathbf{X}}$ and $\breve{\mathbf{X}}$. For $\dot{\mathbf{X}}$, this meant that the network was trained with images slightly deviating from the initial alignment to the same coordinate system, making it more robust to imperfect alignment.

As the size of the training set varies strongly across experiments, all models were trained for the same number of iterations (defined as passing a single batch of data through the network). This number was empirically set to be equivalent to training with all atlas-labelled images for 100 epochs. I used the ADAM optimiser [198] with an initial learning rate of 0.001 for training, and a batch size of four. This relatively small batch size was constrained by the GPU memory due to the large input images ($160 \times 160 \times 160$). The results reported in this chapter are the performance after the last training epoch averaged over three training runs.

5.3.3 Implementation

All models were implemented in Python 3.7 using PyTorch (version 1.7.1) and trained on an NVIDIA Tesla V100 with 32 GB of memory. To generate the template images used to obtain the atlas labels Matlab (version 9.8) was used, and all manual annotation was performed using the freely available MITK-Workbench.

5.3.4 Evaluation

5.3.4.1 Post-processing

During visual inspection of the resulting segmentations, it was noticed that for some images (about 25% of the test images for $\dot{\mathbf{X}}$ and 50% for $\breve{\mathbf{X}}$), θ^{exp} made predictions that were also in the hemisphere proximal to the transducer (which is partly occluded due to shadowing), whereas the ground-truth only contained manual annotations of the distal hemisphere for both the CP and LPVH. However, these predictions were always smaller in size than the predictions in the distal hemisphere, as only part

of the structure was clearly visible. Furthermore, for $\breve{\mathbf{X}}$ some very small spurious areas were predicted for a few test images. Quantitatively, this can also be seen from the resulting large Hausdorff distances before post-processing (see Appendix B, Fig. B.4). For these reasons, I post-processed the binarised predictions $\hat{\mathbf{Y}}_c$ by taking the largest connected component as the final prediction, denoted by $\hat{\mathbf{Y}}_{c,post}$. As the evaluation metrics are computed both with and without post-processing, for simplicity the notation $\hat{\mathbf{Y}}_c$ is used in the remainder of this section to represent either case. For evaluation, $\hat{\mathbf{Y}}_c$ will always refer to the binarised predictions per class.

5.3.4.2 Metrics

The binary predictions for each class, $\hat{\mathbf{Y}}_c$, were separately evaluated against the binary manual expert labels for the respective class, $\mathbf{Y}_c^{exp}$. For evaluation, the Dice Similarity Coefficient (DSC) and the 95^{th} percentile Hausdorff distance (H_{95}) were used.

DSC The DSC measures the overlap of the prediction with the ground-truth, and is defined by:

$$\mathrm{DSC}(\mathbf{Y}_c^{exp}, \hat{\mathbf{Y}}_c) = \frac{2|\mathbf{Y}_c^{exp} \cap \hat{\mathbf{Y}}_c|}{|\mathbf{Y}_c^{exp}| + |\hat{\mathbf{Y}}_c|} \tag{5.3}$$

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

Hausdorff The 95^{th} percentile Hausdorff distance (H_{95}), which is the maximum (or 95^{th} percentile for this case) over the set of boundary points of the minimum distances (between a boundary point in one set and any boundary point in the other set). If $\mathbf{B_c}$ and $\hat{\mathbf{B}}_\mathbf{c}$ represent the surface points of $\mathbf{Y}_c^{exp}$ and $\hat{\mathbf{Y}}_c$ respectively, the H_{95} is given by [199]:

$$H_{95}(\mathbf{B_c}, \hat{\mathbf{B}}_c) = \max(h_{95}(\mathbf{B}_c, \hat{\mathbf{B}}_\mathbf{c}), h_{95}(\hat{\mathbf{B}}_c, \mathbf{B}_c)) \tag{5.4}$$

where h is the directional Hausdorff distance given by:

$$h_{95}(\mathbf{A}, \mathbf{B}) = \underset{\mathbf{a}\in\mathbf{A}}{95^{\text{th}}} \min_{\mathbf{b}\in\mathbf{B}} d(\mathbf{a}, \mathbf{b}) \tag{5.5}$$

with $d(\mathbf{a}, \mathbf{b})$ the Euclidean distance between points $\mathbf{a}$ and $\mathbf{b}$, and 95^{th} being the 95^{th} percentile. I used the 95^{th} percentile distance as it is more robust to very small outliers than the standard maximum Hausdorff distance. As the Hausdorff distance represents the distance between prediction and ground-truth, lower distances indicate better segmentation performance.

In addition to the DSC and H_{95} also the signed and unsigned relative volume differences were computed. These results are presented in Appendix B (Fig. B.2 and B.3).

5.3.5 Volumetric Subcortical Growth Trajectories

In order to generate growth curves for the subcortical structures, I finally applied the trained models to the US images in the analysis subset. Subcortical volumes ($V_{\hat{Y}_{c,post}}$) were predicted for $\dot{\mathcal{X}}_{analysis}$ using both $\dot{\theta}^{exp}$ and $\dot{\theta}^{atl}$ in order to analyse the differences. Subsequently, a linear or quadratic polynomial was fit to the predicted volumes as a function of the GA. The quadratic term was only added if it proved to be statistically significant for both $\dot{\theta}^{exp}$ and $\dot{\theta}^{atl}$ (determined by a two-sided t-test), as the underlying structural growth should correspond between the two networks.

In addition to structural volumes, also the relative structural volume with respect to the whole brain volume was computed, defined by:

$$V_{rel_brain} = \frac{V_{\hat{Y}_{c,post}}}{V_{brain}} \tag{5.6}$$

with V_{brain} the whole brain volume of the respective image. These whole brain volumes were computed from whole brain masks derived from an MRI fetal brain atlas [61] (as described in Section 4.1).

5.4 Results

All performance values presented in this section were obtained from $\hat{\mathbf{Y}}_{c,post}$ (predictions post-processed to only keep the largest connected component). Results without post-processing are given in Appendix B (Figs. B.4 and B.3).

5.4.1 Weak Atlas Labels

In order to quantify the number of inaccuracies in the weak atlas training labels, the DSC overlap between the naive atlas propagated labels ($\mathcal{Y}^{atl}$) and the manual ground-truth labels ($\mathcal{Y}^{exp}$) was determined in the test set. These results are shown in Table 5.3 as *prop. atlas*. For the LPVH, the DSC overlap of the atlas labels obtained from the clustered template images, $\mathbf{K}^{clust}$, is shown. It can be observed that the naive propagated atlas labels obtained relatively low DSC, ranging from 0.68 for the LPVH to 0.80 for the CB, indicating that the atlas labels used as weak ground-truth labels contain a substantial amount of label noise.

5.4.2 Comparison between Atlas and Expert Labels

The resulting segmentation performance of the trained networks θ^{atl} ($n_x = 239$) and θ^{exp} ($n_x = 9$) are shown in Fig. 5.5 and Table 5.3. For both networks, the number of images that were manually annotated to obtain the training labels was the same ($n_a = 9$). The results shown were based on experiments that were trained and tested on images with the same alignment, either $\breve{\mathcal{X}}$ or $\dot{\mathcal{X}}$. Table 5.3 shows the performance for the aligned images (statistical results for these comparisons are provided in Appendix B, Table B.1), whereas in Fig. 5.5 results for both the aligned and unaligned settings are presented.

It can be observed that, somewhat surprisingly, in the aligned setting, the best performance is obtained using only nine expertly annotated images, as opposed to a much larger number of weakly annotated images. This indicates that even though segmentation in 3D US by a human annotator is a challenging task, a network is able to learn the correct features from only a small set of manual annotations, provided that the images are all aligned to the same coordinate space.

Table 5.3: DSC performance of the segmentation networks using aligned images compared to results of previous studies (obtained on different datasets). Numbers in between brackets indicate the standard deviation and the best performance for each structure is shown in bold (only regarding results on 3D images). For both $\dot{\theta}^{exp}$ and $\dot{\theta}^{atl}$ results of statistical testing with respect to the propagated atlas labels are shown with (**) $p < 0.005$ and (*) $p < 0.05$. Full overview of statistical results can be found in Appendix B (Table B.1).

	CP	LPVH	CSPV	CB
2D image planes				
Huang et al. [64]	0.76 (0.08)		0.81 (0.06)	
Wu et al. [69]			0.78 (0.12)	
Singh et al. [65]				0.87 (0.13)
Sreelakshmy et al. [66]				0.91 (0.08)
Shu et al. [67]				0.91 (0.03)
3D image volumes				
Gutiérrez-Becker et al. [6]				0.80 (0.05)
Yaqub et al. [5][a]	0.79 (0.09)	0.82 (0.10)	0.74 (0.11)	0.63 (0.15)
Venturini et al. [8]				0.73 (0.014)
This work				
prop. atlas	0.79 (0.07)	0.68 (0.10)	0.72 (0.10)	0.80 (0.09)
$\dot{\theta}^{atl}$	0.82 (0.05)**	0.77 (0.11)**	0.76 (0.07)	0.86 (0.03)*
$\dot{\theta}^{exp}$	**0.85** (0.04)**	**0.85**(0.05)**	**0.78** (0.13)	**0.90** (0.02)**
Intra-observer variability	0.85 (0.05)	0.86 (0.03)	0.86 (0.04)	0.91 (0.03)

[a] Search region limited to cuboid around ground-truth annotation
[b] Segmentation in 2D standard planes

5.4.3 Varying the Number of Manual Annotations

The results of reducing the number of manual annotations, n_a, are shown in Fig. 5.6. It can be seen that for $\theta^{atl}_{n_a}$, only a slight decrease in performance is observed when decreasing n_a. Even by training with labels generated from just 2 GWs (coming from 2 manual annotations), the performance is close to the maximum performance (average DSC of 0.78 versus maximum DSC of 0.80 for the aligned case).

On the other hand, for $\theta^{exp}_{n_a}$ a performance decrease was observed when reducing the number of training images, showing a stronger decrease for the *unaligned* images than for the *aligned* images. As in this experiment as little as two images were used for training, while evaluating on images from the whole GA range, this drop in performance was expected.

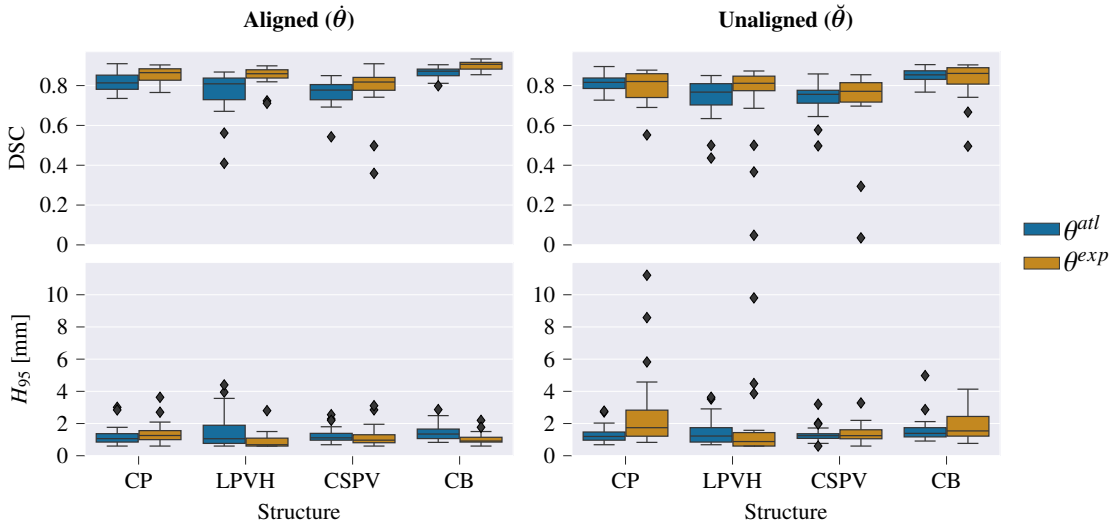


Figure 5.5: Resulting performance values after post-processing for θ^{exp} and θ^{atl}. For the unaligned images, larger H_{95} are observed than for the aligned set-up, most pronounced for the θ^{exp}. For θ^{atl}, the segmentation performance between the aligned and unaligned settings shows only a small difference.

5.4.4 Effect of Image Alignment

To explore the sensitivity of the models to brain pose, I also studied the effect of image alignment on the segmentation performance. From Fig. 5.5 it is evident that higher performance was obtained for the *aligned* images compared to the *unaligned* images, which is most pronounced in the lower H_{95} for these images. As segmenting the structures in *aligned* images can be considered an easier task, this was in line with the expectations.

When comparing the performance between the *aligned* and *unaligned* settings for both models in more detail, a large performance difference can be observed for θ^{exp} between the *unaligned* and *aligned setting*, being most pronounced in Fig 5.6. On the other hand, this difference was not as pronounced for θ^{atl}. This can be explained by the fact that in the *unaligned* setting the segmentation task consists of a combination of structure localisation and subsequent segmentation. The weak, atlas labels contain imperfections at the boundaries of the segmentations, but are always approximately in the correct location. For this reason, these labels provide more information for the localisation task than the few expertly annotated images and, as such, obtain better segmentation performance in the *unaligned* case.

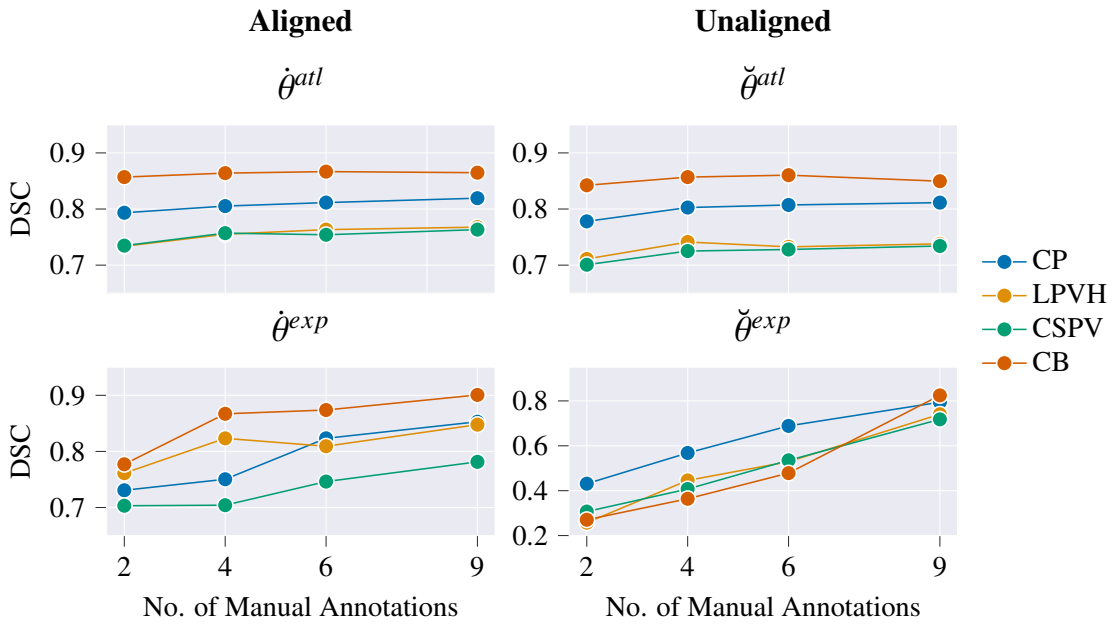


Figure 5.6: Varying the number of manual annotations (n_a) used for generating the training labels. Reducing n_a has only a small effect on the performance of θ^{atl} whereas a drop in performance is observed for θ^{exp}. Note the different ranges of the y-axis of the bottom right plot.

The few expertly annotated images in this setting are not sufficient to provide both localisation and segmentation, especially when reducing n_a below nine (Fig 5.6). These results thus suggest that a large amount of weak labels is better for localisation, whereas few expert labels perform better if the task consists of solely delineating a structure in approximately the same location.

5.4.5 Comparison to Previous Work

In Table 5.3, the resulting DSC performance using *aligned* images is compared to the intra-observer variability and to previous work performing subcortical structure segmentation. The intra-observer variability was obtained by annotating ten images from the test set twice. Furthermore, for $\dot{\theta}^{atl}$ and θ^{exp} the significance with respect to the naive propagated atlas labels is shown. The p-values shown were obtained by a repeated measures Analysis of Variance (ANOVA) for each structure, followed by *post-hoc* testing with a paired t-test. Reported p-values are from the *post-hoc* testing and underwent Bonferroni correction for the four structures, as well as for the three model comparisons (see Table B.1 in Appendix B for full statistical results).

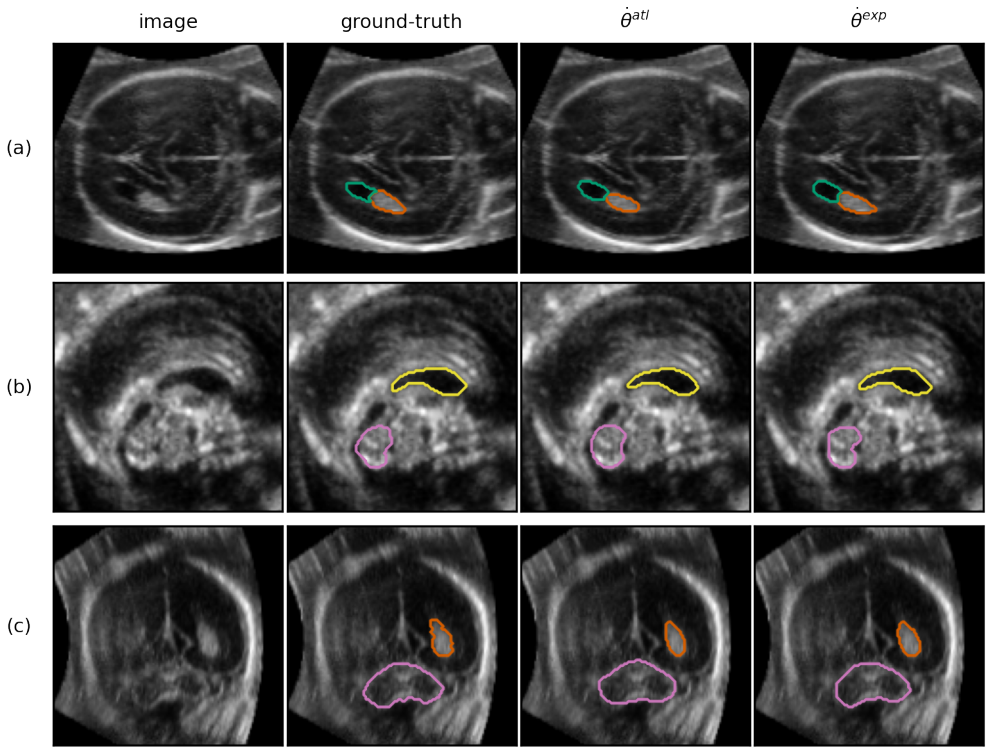


Figure 5.7: Example prediction of a test set volume at 22 GW (randomly chosen from 5 best-performing images, average DSC across structures of 0.85), visualised in the (a)axial, (b) sagittal and (c) coronal plane (segmentation was performed in 3D). The structures shown are the LPVH (green), CP (orange), CSPV (yellow), and CB (pink). Only the outer boundary of the segmentation is shown in colour for visualisation purposes.

However, it should be noted that performance metrics reported in different studies have to be carefully interpreted. Factors such as US image quality, structural definition (i.e. what is included in the segmentation), and GA range can all affect the difficulty of the task. The GA range affects the difficulty of the segmentation tasks as acoustic shadowing increases with advancing GA, due to calcification of the fetal skull. Furthermore, overlap measures, such as the DSC, generally report higher values for larger structures because a single (erroneous) voxel has a smaller effect on the resulting DSC value for these structures. I also want to stress that small performance increases are very difficult to quantify in 3D US segmentation due to the high intra- and inter-observer variability predominantly resulting from the ambiguity of the exact location of structural boundaries.

For all structures, superior performance is obtained compared to previous work

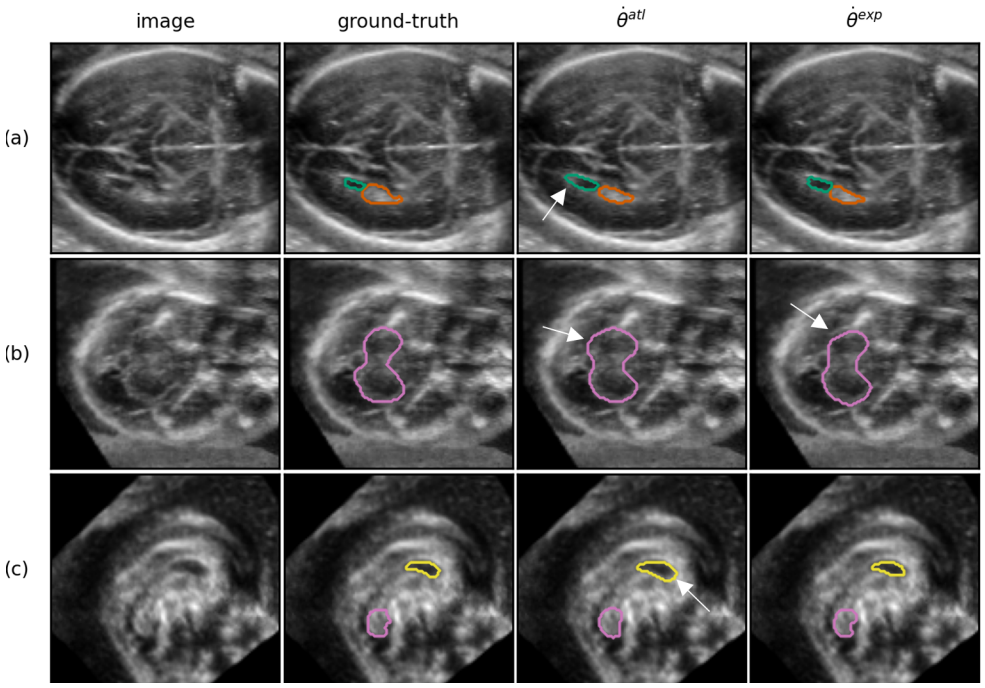


Figure 5.8: A set of examples showing the most prominent segmentation errors from $\dot{\theta}^{exp}$ and $\dot{\theta}^{atl}$ after post-processing (a) axial plane of an image from a subject at 26 GW demonstrating over-segmentation of the LPVH (green) (b) axial plane of an image from a subject at 22 GW demonstrating over-segmentation of the CB (pink) for $\dot{\theta}^{atl}$ and slight under-segmentation for $\dot{\theta}^{exp}$ (c) sagittal plane of an image from a subject at 18 GW, demonstrating over-segmentation for $\dot{\theta}^{atl}$ as well as the difficulty of manually segmenting the CSPV (yellow) at young ages.

on 3D US images. The best performance obtained for the CB, LPVH and CB is also very close to the intra-observer variability in the manual annotations (DSC of 0.85, 0.85 and 0.90 versus an intra-observer variability of 0.86, 0.85 and 0.91 for the CP, LPVH and CB, respectively). An especially large performance increase was observed for the CB (DSC of 0.90 versus DSC of 0.80, 0.63 and 0.73 in previous studies). However, part of this increase can be attributed to the fact that in this study the CB was manually annotated including the bright echogenic boundaries, thus resulting in a larger segmented region, whereas this boundary was not included in earlier work [5, 8]. This choice was made based on the fact that the TCD, which is clinically used to assess growth during a standard fetal examination, also includes this boundary [200]. Furthermore, manual segmentation is more consistent when including this boundary as the outside edge is generally clearly visible in the US images.

For the CSPV a larger difference with the intra-observer variability is observed

(DSC of 0.78 versus 0.86). This can be largely attributed to poor segmentation performance at 18 GWs, when the CSPV is small and the boundaries are ill-defined (Fig. 5.8), as well as to the difficulty of defining the lateral boundaries of the CSPV in the 3D volumes.

Previous work on 2D US images showed superior performance for both the CSPV [64] (0.81 versus 0.78) and CB [66, 67] (0.91 versus 0.90). However, these are segmentations in 2D standard planes (midsagittal and transcerebellar), making the task substantially less challenging. Furthermore, due to differences in structural definition [8], a limited search region [5], as well as unavailability of code, it was not possible to do a direct evaluation of past methods on the current dataset. Although the reported comparisons should thus be interpreted with care, they do show that competitive segmentation performance was obtained.

5.4.6 Qualitative Results

In Fig. 5.7, an example prediction is shown for an aligned image at 22 GWs. The manual ground-truth as well as the predictions from both $\dot{\theta}^{exp}$ and $\dot{\theta}^{atl}$ are shown. A qualitative inspection suggests that the predictions of the two networks closely resemble each other as well as the ground-truth. In Fig. 5.8, slices are shown containing the most prominent error modes in the aligned test set. From Fig. 5.8a-c, it can be observed that $\dot{\theta}^{atl}$ tends to over-segment in some areas whereas the predictions from $\dot{\theta}^{exp}$ align better with the ground-truth. In Fig. 5.8c a CSPV segmentation at 18 GW with a very low DSC (0.51) for $\dot{\theta}^{atl}$ is shown. It can be observed from this figure that the CSPV is challenging to (manually) segment at this young GA, due to the small structural volumes as well as to the boundaries that are not yet very well-defined.

5.4.7 Clustered Template Labels

I extended the atlas-based labels with a clustering approach for the LPVH. This approach was only applied for the LPVH, as the other structures did not display as much shape variation within GWs. In Table 5.4 the LPVH segmentation performance

Table 5.4: LPVH segmentation performance using two different types of atlas-based annotations: LPVH labels obtained from annotating a single template per week ($\mathcal{K}$) or annotating a set of clustered template images ($\mathcal{K}^{clust}$).

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

of training $\dot{\theta}^{atl}$ with atlas labels either propagated from $\mathcal{K}$ or $\mathcal{K}^{clust}$ is shown. Training a network with the propagated clustered template annotations significantly improved the DSC performance from 0.69 to 0.77 ($p<0.001$) and H_{95} from 2.2 to 1.6 ($p<0.05$). Significance values were calculated using a paired t-test. This thus suggests that the additional variation captured by the clustered templates contributed to an improved segmentation of the LPVH.

5.4.8 Subcortical Growth Curves

The growth curves obtained from applying the trained networks to 278 images of healthy fetuses are presented in Figs. 5.10 and 5.11. In Fig. 5.10 the predicted structural volumes are shown, whereas in Fig. 5.11 the predicted relative volumes with respect to the whole brain volume (V_{rel_brain}), are shown. The growth curves were fitted with a linear or quadratic fit based on the significance of the quadratic component computed with a two-sided t-test.

To illustrate the level of consistency between $\dot{\theta}^{exp}$ and $\dot{\theta}^{atl}$, in Figs. 5.10 and 5.11 the samples were coloured based on their residual value in the growth curves obtained from $\dot{\theta}^{exp}$. For a certain structure, samples with matching colours between the growth curves of the $\dot{\theta}^{exp}$ and $\dot{\theta}^{atl}$ thus correspond to the same image. The choice of colouring the samples based on $\dot{\theta}^{exp}$ was made because of its better DSC performance (Table 5.3), but as it is only used to visualise the matching images, this choice is relatively arbitrary.

It can be observed that both of the models, $\dot{\theta}^{exp}$ and $\dot{\theta}^{atl}$, result in a similar growth trend and that the relative sample distribution is largely preserved (i.e. values above or below the growth curve are respectively above or below the curve for the other

method, by a similar amount). It can also be seen that $\dot{\theta}^{atl}$ predicts larger volumes than $\dot{\theta}^{exp}$, which was also confirmed by the fact that $\dot{\theta}^{atl}$ resulted in positive unsigned relative volume differences, corresponding to over-segmentation (see Appendix B Fig. B.2). However, as the visualisation of structural boundaries can be subjective in US, a consistent (albeit overestimating) segmentation demonstrates the utility of this method in comparison to manual delineation.

Previous work on volumetric subcortical development during gestation from US has largely been limited to cerebellar and thalamic growth [71–74]. Therefore, this study provides novel US-specific volumetric growth curves of the CSPV, CP and LPVH for a geographically diverse, healthy population. To validate the volumetric measures obtained in this study, I compared the growth curves of the structural CB volume to previously reported CB growth curves from US and MRI [72, 76–78, 82, 84, 195], which are shown in Fig. 5.9. This showed that the generated curves showed excellent agreement with past US studies, especially in the first half of the GA range. After 24 GWs the growth rates of the CB (slope of the growth curves) found in this study are slightly lower than in previous work. However, as described previously, the CB becomes more challenging to segment with advancing GA and, as such, this could explain the slightly deviating volume measurements at the end of the second trimester. Compared to the MRI studies a larger volume is predicted for most of the GA range, which matches the observation that the CB, when including the bright echogenic boundary, appears slightly larger on the US volumes than on MR images.

From the resulting growth curves (Figs. 5.10 and 5.11), it can be observed that the CB undergoes rapid growth during the second trimester of gestation, whereas the relative cerebellar volume, with respect to the total brain volume, remains mostly consistent (between 1.5% and 2.5 %). As the TCD is expected to increase linearly during the second trimester [39], this matches the quadratic growth of CB volume. The CP only shows a very small increase in structural volumes during the second trimester. However, the relative volumes show a rapid decline with respect to the total brain volume, ranging from 12% of the total brain volume at 18 GWs to 3% at 27 GWs. This aligns with the observation that early in the second trimester

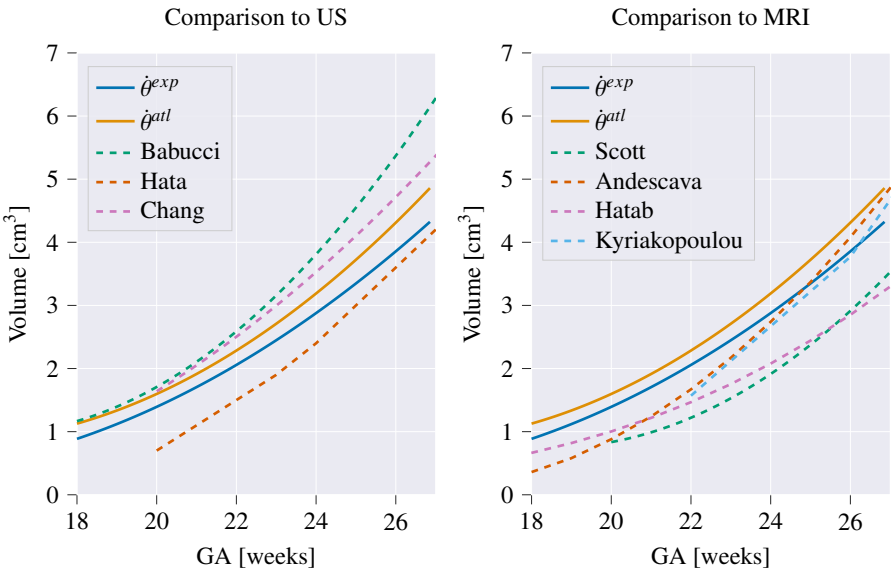


Figure 5.9: Comparison of growth curves obtained for the CB in this study to previous work. In the left panel the CB growth curves from our study ($\dot{\theta}^{exp}$ (blue) and $\dot{\theta}^{atl}$ (orange)) are compared to other studies reporting cerebellar growth curves from US: Babucci et al. [72] (green), Hata et al. [82] (dark orange), Chang et al. [84] (pink), and the right panel compares our curves to previously reported MRI growth curves: Scott et al. [76] (green), Andescavage et al. [78] (dark orange), Hatab et al. [77] (pink), and Kyriakopoulou et al. [195] (light blue).

the CP almost completely fills the lateral ventricles and comprises a large part of the fetal brain, whereas later in gestation it appears as a small structure only partly filling the lateral ventricles. The LPVH structural volume increases linearly from about 0.1 cm^3 to 0.3 cm^3 in the studied GA range, which is in agreement with the fact that the atrial width remains stable in healthy fetuses during the second trimester [39]. The CSPV shows a linear increase in structural volume during the studied GA range (from about 0.01 cm^3 at 18 GWs to 0.6 cm^3 at 26 GWs) but its relative volume increases early in the second trimester and remains constant after approximately 24 GWs. As a deviating measurement of the CSPV has been related to agenesis of the corpus callosum [201] but is not widely studied, the presented growth curves might facilitate this in future work.

5.4.9 Generalisability

In order to validate the segmentation performance on a completely unseen dataset, the developed models were also applied to the LUMC test set. For this dataset, no

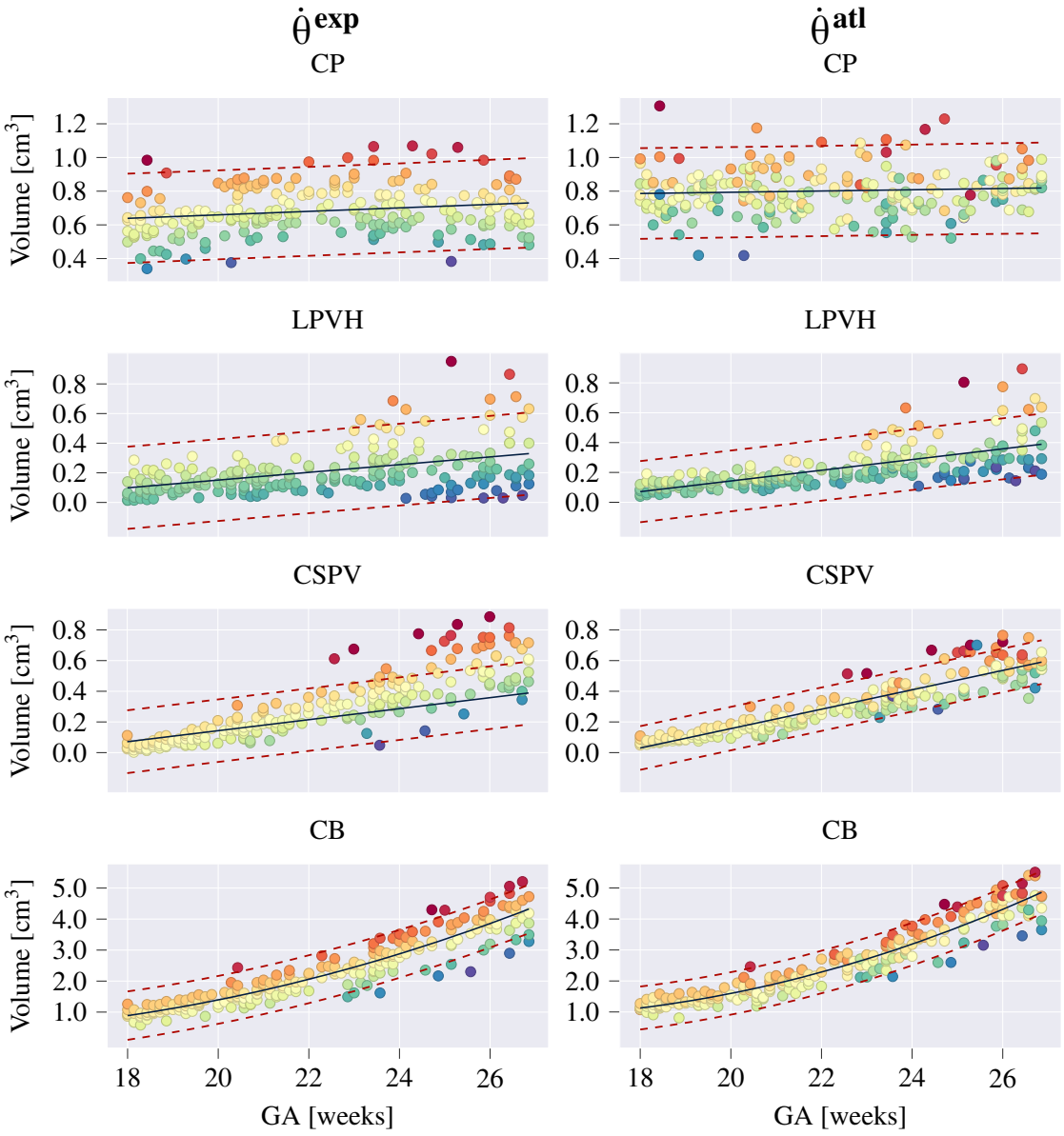


Figure 5.10: Estimated structural volumes for subcortical structures as a function of GA. Volumes were fitted with a linear or quadratic fit (black), in which the quadratic term was only added if it was significant for both networks (per structure). The 95% prediction confidence intervals were also computed and are shown with red dashed lines. For each structure, samples were coloured based on their residual for $\dot{\theta}^{exp}$, and the same colours (per sample) were used for $\dot{\theta}^{atl}$.

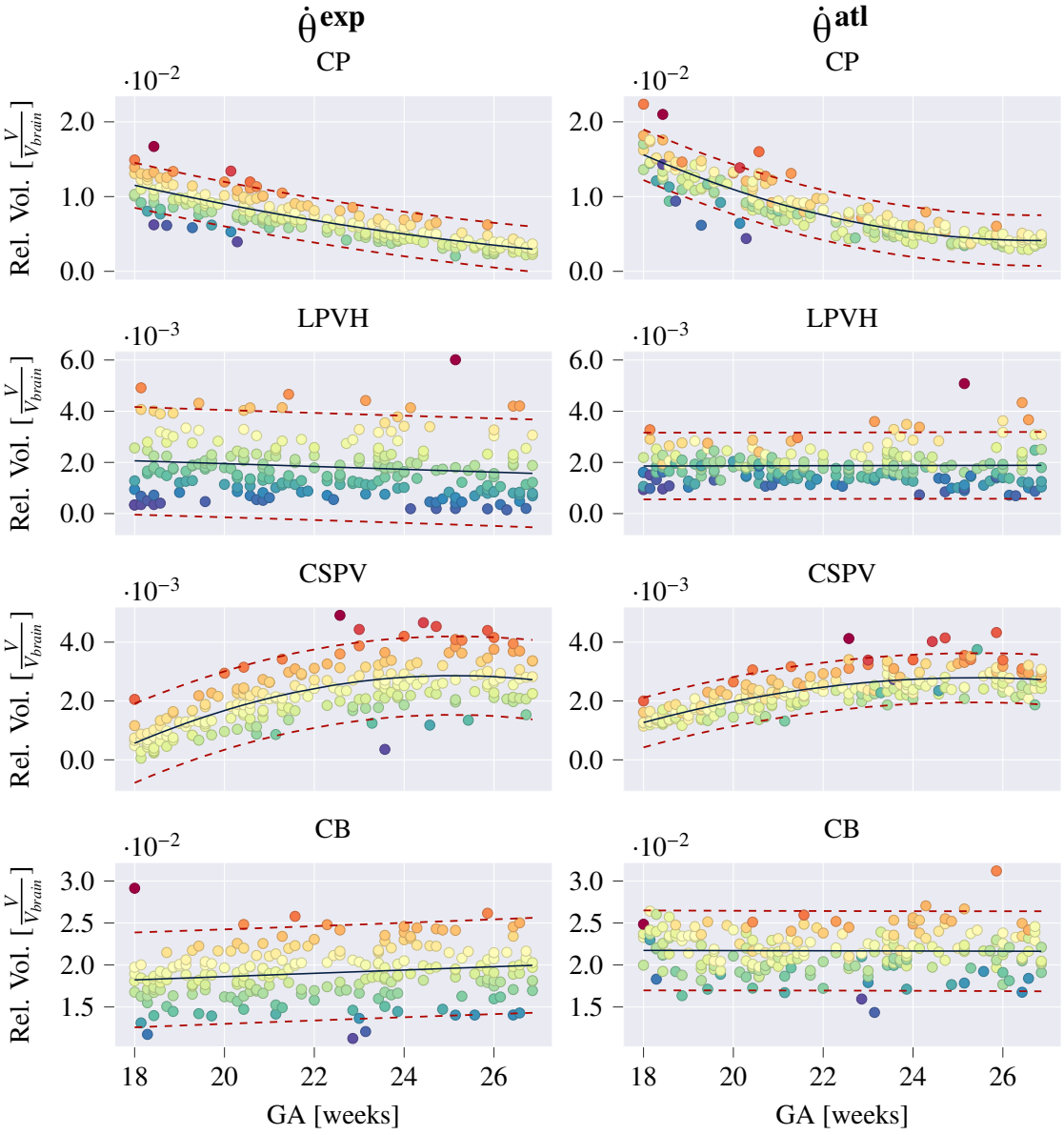


Figure 5.11: Estimated relative volumes (Rel. Vol.) with respect to the whole brain volumes (V_{rel_brain}) for subcortical structures as a function of GA. Volumes were fitted with a linear or quadratic fit (black), in which the quadratic term was only added if it was significant for both models (per structure). The 95% prediction confidence intervals were also computed and are shown with red dashed lines. For each structure, samples were coloured based on their residual for $\dot{\theta}^{exp}$, and the same colours (per sample) were used for $\dot{\theta}^{atl}$.

voxel-level manual ground-truths were available, and the segmentation performance was therefore qualitatively assessed and graded by a trained ultrasonographer (Dr M. Aliasi) using the following four grades:

1. High segmentation quality (correct segmentation of the structure)
2. Intermediate segmentation quality (small segmentation boundary errors, less than 20% of the structure is over- or under-segmented)
3. Low segmentation quality (major segmentation boundary errors, 20-50% of the structure is over- or under-segmented)
4. Very low segmentation quality (incorrect identification of the structure and/or >50% of the structure is over- or under-segmented)

The results on the INTERGROWTH-21st dataset showed that optimal performance on aligned images was obtained using a few individually annotated images for training, and that increasing the number of training images (from 1 to 9) increased segmentation performance (Fig. 5.6). It could thus be inferred that including more than 9 training volumes will likely improve the segmentation performance further. As the analysis on the LUMC dataset was intended for clinical research in future work (i.e. comparing cohorts of healthy fetusses with fetusses with congenital heart disease), optimal performance on this dataset was desired. Each model-generated segmentation had to be manually graded by an ultrasonographer, and the generalisation performance on this dataset was therefore only evaluated using a model retrained with 20 individually annotated image volumes. These 20 training images contained the 9 training images used in the rest of this chapter as well as 11 images from the original test set.

The results of this retrained model on the LUMC dataset are shown in Table 5.5. It can be seen that a majority of the images have a high segmentation quality, and that, except for the LPVH, close to 90% of the images have been graded as either having high or intermediate segmentation quality. Considering that this dataset was acquired with a different US scanner, with different settings and at a different institution, this shows that good segmentation performance can be

obtained without fine-tuning the model with images from the new domain. The LPVH obtained a lower performance, with only about 70% of the images being graded as having high or intermediate segmentation quality. However, in the majority of the failure cases, the ventricular cavity appeared substantially less dark than in the INTERGROWTH-21st images, which most likely explains the difficulty of the model to segment these cases accurately.

Table 5.5: Segmentation results on the LUMC test set, the grades represent (1) high segmentation quality, (2) intermediate segmentation quality, (3) low segmentation quality, and (4) very low segmentation quality. The numbers represent the percentages of each grade, with the number of images between brackets. The total number of images per structure varies as only images where a certain structure was visible were graded.

	CP	LPVH	CSPV	CB
no. images	114	116	116	109
Grade 1	85.1 (97)	63.8 (74)	83.6 (97)	79.8 (87)
Grade 2	4.4 (5)	11.2 (13)	6.0 (7)	10.1 (11)
Grade 3	7.0 (8)	11.2 (13)	6.0 (7)	7.3 (8)
Grade 4	3.5 (4)	13.8 (16)	4.3 (5)	2.8 (3)

5.5 Discussion

This chapter showed that only a small number of annotated images are needed to successfully train a network for subcortical segmentation in 3D US images. To obtain optimal performance, alignment of the images is required beforehand. However, on unaligned testing images, a high performance can still be achieved using several weakly annotated images. By applying the trained networks to a large cohort of fetuses, novel US-specific growth trajectories of the CP, LPVH, CSPV and CB were generated for a geographically diverse, healthy population. This chapter thus demonstrates the feasibility of subcortical segmentation in 3D US using deep learning, and shows that volumetric measures obtained from these models can be used to obtain an improved understanding of subcortical growth during gestation.

A limitation of the work presented in this chapter is that no methods were used that were specifically tailored for training with noisy labels (the *atlas* labels), or

for training with only a handful of annotations (the *expert* labels). However, due to the high intra-observer variability as well as the relatively small test set, it is challenging to measure small increases in segmentation performance potentially obtained by these methods. Additionally, the experiments demonstrate the feasibility of subcortical segmentation in fetal US by showing that good performance (i.e. close to human-level performance) can be obtained using simple methods that can easily be reproduced.

Even though the segmentation performance was shown to be close to intra-observer variability, the test set used in this chapter contained a limited number of 20 image volumes. Furthermore, the annotations of a single annotator were considered to be the ground-truth in this study. In further work, the presented models and growth charts could be validated in more detail by increasing the size of the test set and by obtaining more accurate ground-truths from consensus segmentations of multiple annotators.

The resulting performance on the LUMC test set showed that without fine-tuning, good segmentation performance for most structures can be obtained on a dataset acquired at a different site. While the evaluated model on the LUMC dataset slightly deviates from the models evaluated in the rest of the chapter (i.e. using 20 individually annotated images for training instead of 9), it nevertheless demonstrates the ability of the network to make accurate predictions for a completely new domain. This indicates that the domain gap in US might be smaller than that seen in, for example, MRI [202, 203]. In further work, the models developed in this chapter offer the promise of comparing subcortical development between different fetal cohorts without doing time-consuming manual annotations. However, more work is necessary to extensively validate the segmentation performance for datasets acquired at a larger number of sites and containing a range of pathological conditions.

6

Local Brain Age Prediction using Guided Attention

Contents

6.1 Introduction 101
6.2 Review of attention methods 104
6.2.1 Attention Gates 104
6.2.2 Guided Attention Gates 106
6.3 Methods 107
6.3.1 Problem Definition 107
6.3.2 Network Architecture 107
6.3.3 Attention 109
6.3.4 Network Training 111
6.4 Experimental Set-up 112
6.4.1 Dataset 112
6.4.2 2D Experiments 114
6.4.3 3D Experiments 116
6.4.4 Visualisation of Attention Maps 116
6.4.5 Implementation 117
6.5 Results 118
6.5.1 2D Standard Planes 119
6.5.2 3D Image Volumes 128
6.5.3 Generalisability 134
6.6 Discussion 135

In this chapter, I study the development of the brain by quantifying the stage of development with brain age predictions. Specifically, I introduce a method that

can quantify *local* brain age by enforcing the network to focus on only part of the brain using guided attention. The focus of the network can be regularised with a segmentation mask of a specific structure, or region of interest, obtaining information about the structural development of specific brain regions. This chapter demonstrates that guiding the network to known age-discriminative regions such as the CoP improves the prediction performance significantly while providing *local* age predictions (i.e. based primarily on the CoP appearance).

6.1 Introduction

Deep learning models that predict the GA of a fetus from US volumes have the potential to be used to quantify the stage of development of the fetal brain *in utero* [9]. To draw any meaningful clinical conclusions from the predicted ages, it is important to gain insight into the factors that contributed to the final model output, or, more specifically, how deviations in structural appearance contributed to the final age prediction. For example, a particular trained model might only predict the GA based on the skull appearance, which is known to change in appearance during gestation (i.e. due to the increased calcification). While this model might result in accurate age predictions and can inform us about the stage of development of the skull, it would be incorrect to use this model to conclude anything about, for example, subcortical development.

The most commonly used explanation methods for CNNs are *post-hoc* attribution-based, which are not always able to consistently and accurately explain the prediction of a model [116] (as described in Chapter 3). To illustrate this point, Fig. 6.1 shows three different attribution methods 'explaining' the age prediction from the same model for the same test set image. It can be seen that there is significant disagreement between the methods, making them hard to interpret and prone to confirmation bias (i.e. selecting the explanation that confirms a certain hypothesis)[116, 204].

Attribution-based methods can, at best, only provide the contributing regions for an already trained model. For an US brain age prediction model, these regions can

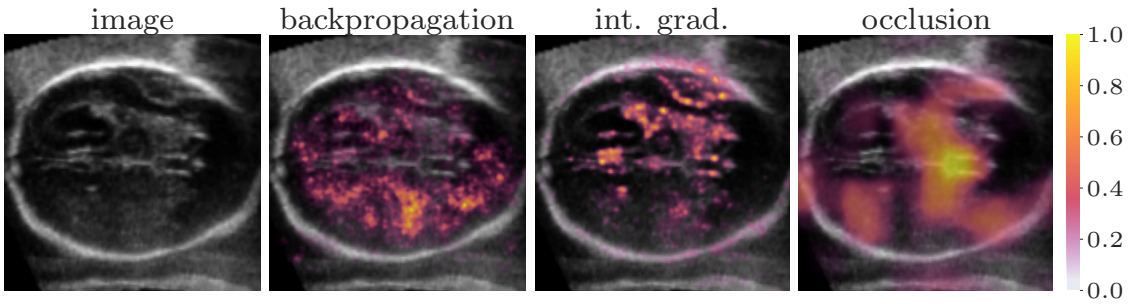


Figure 6.1: Three different attribution-based explanation methods explaining the prediction of the same model for a single test set sample: backpropagation [104], integrated gradients [110], and occlusion [113]. While explaining the same sample of the same model, the three methods clearly assign importance to different image regions.

be expected to correspond to age-discriminative regions in the visible hemisphere, such as the SF or the CP. However, the attribution is likely to cover large parts of the image (as seen in Fig 6.1), as multiple brain structures are, to a degree, age-discriminative. Additionally, the attribution might not match the structures of interest for a certain research question (i.e. assessing the stage of development of the CoP), limiting their practical use.

For this reason, in this chapter I propose to regularise which features the network uses for the age prediction task by incorporating *guided Attention Gates (AG)* in the architecture [205, 206]. An AG aims to learn an attention mask that can exclude irrelevant feature activations in intermediate feature representations to focus on relevant areas. Furthermore, a *guided* AG applies a segmentation loss to the learned masks, which regularises them with a Region of Interest (ROI), thereby effectively enforcing the network to focus on that area. Importantly, segmentation masks are only required during training and not at inference time, as the attention mask is learned by the network.

In this chapter, I will primarily use the CoP as the ROI. As discussed in Chapter 2.1.4, CoP development during gestation is well timetabled, and it has been shown that in a healthy population individual fissures develop in a temporally consistent way [32, 44]. Furthermore, abnormal cortical folding has been correlated with fetal abnormalities, such as congenital heart disease [33], illustrating the importance of assessing cortical folding. However, assessing the CoP folding in

clinic is challenging and current assessment methods, such as grading charts [32, 44], are very subjective and therefore user-dependent [49]. For this reason, it is of clinical interest to develop prediction models that can assess brain development by predicting the GA while focussing on the CoP.

In addition, experiments with the CB as the ROI are performed. As described in Chapter 2.1.4, the CB is always assessed during routine US examinations to measure the TCD, used to quantify growth. While size information will mostly be removed from the US images by size normalising them (to enforce the network to focus on structural development, as described in Chapter 2), it is of interest whether the CB appearance alone can be age discriminative.

I will start this chapter with a short overview of the different types of *attention* architectures that have been proposed in previous studies, followed by a detailed overview of the methods and experimental set-up. The results will then first be presented for 2D US images, where the attention is regularised with CoP segmentation masks. This will be followed by the results for the 3D images, which contain experiments for regularising both the CoP and CB individually, as well as a combination of the two.

The contributions in this chapter are as follows:

- It demonstrates how guided attention can be applied to train age prediction models that focus on specific brain structures of interest.
- A number of different attention architectures are explored to determine the optimal architecture for the task.
- The methods are applied to both 2D and 3D fetal brain US images, achieving significantly better performance than baseline CNNs for both domains, while limiting the number of features that are used.
- Average attention maps per GW are created that demonstrate that the learned attention closely matches the known areas of development in the CoP.

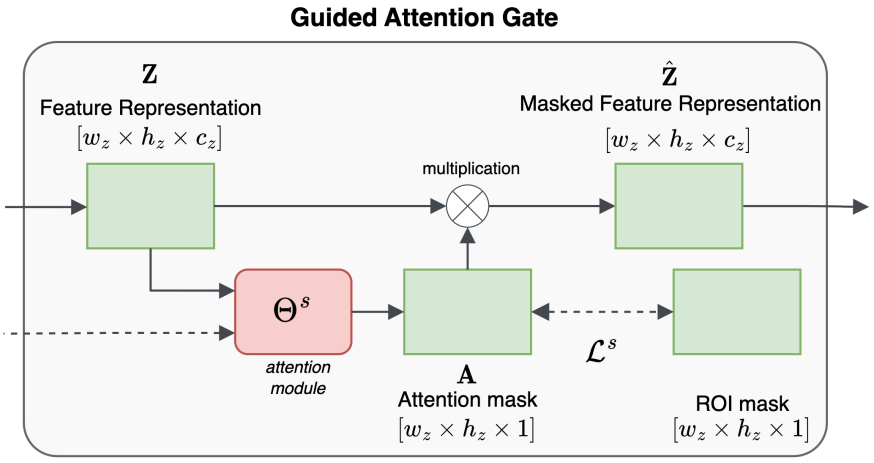


Figure 6.2: Schematic overview of a trainable AG. The parameters of Θ^s are learned during training. The dashed line going into the attention module indicates an optional second input to this module. The dashed line between the attention mask and the ROI ground-truth mask indicates the (optional) segmentation loss, which makes the AG into a *guided* AG. The multiplication between the attention mask and the feature representation is element- and channel-wise.

6.2 Review of attention methods

In this section, I will give a short overview of how AGs have been used in previous studies.

6.2.1 Attention Gates

In this work, I mainly focus on *soft trainable attention gates* that adhere to the following definition. Given a feature representation $\mathbf{Z} \in \mathcal{R}^{h_z \times w_z \times c_z}$, a trainable AG aims to generate an attention mask, $\mathbf{A} \in \mathcal{R}^{h_z \times w_z}$ with $a_{ij} \in [0, 1]$, that modulates the importance of each spatial location in $\mathbf{Z}$. A masked, or attended, feature representation can then be generated by multiplying $\mathbf{Z}$ and $\mathbf{A}$ element- and channel-wise: $\hat{\mathbf{Z}} = \mathbf{A} \odot \mathbf{Z}$. Intuitively, the AG thus enables the network to adjust the relative importance of spatial features in the representation $\mathbf{Z}$, and weight key features more than undescriptive features (e.g. background). The attention mask $\mathbf{A}$ is generated by an *attention module*, denoted by Θ^s, which is a function of the local feature representation $\mathbf{Z}$. A schematic overview of a trainable AG is shown in Fig. 6.2.

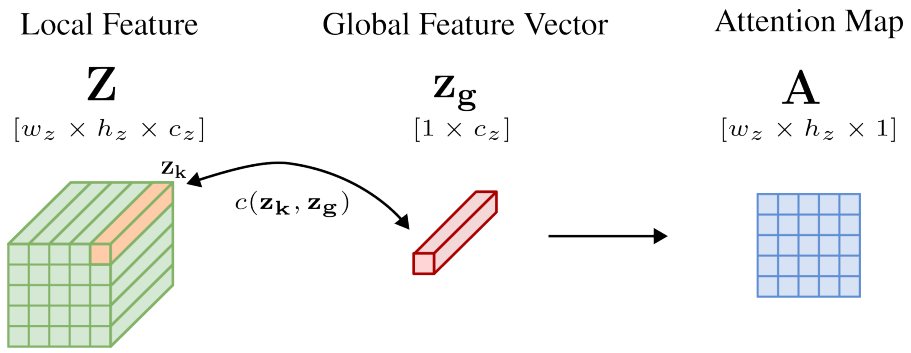


Figure 6.3: Schematic overview of the attention mechanism proposed by Jetley et al. [208]. The local feature represents an intermediate feature representation of the network, and the global feature vector represents the vector representation before the last FC layer. The attention map is generated by computing the compatibility ($c(\boldsymbol{z_k}, \boldsymbol{z_g})$) between each vector in the local feature representation $\boldsymbol{z_k}$ and the global feature vector $\boldsymbol{z_g}$.

In *soft* AGs the learned attention map **A** is directly multiplied with the feature representation, without any form of thresholding. This makes them differentiable and therefore suitable for end-to-end training. On the other hand, *hard* AGs typically apply thresholding and use the resulting masks for intermediate steps such as region proposals [207]. As thresholding operations are non-differentiable, they can usually not be trained end-to-end. Furthermore, a *trainable* AG refers to the learning of the attention mask during training, as opposed to extracting feature attribution maps *post-hoc* (e.g., as is the case for attribution-based methods).

Using trainable soft AGs for the classification of natural images was first proposed by Jetley et al. [208]. In that work, the local feature representation was considered as a set of $h_z w_z$ feature vectors, each of size c_z. The attention mask was then generated by computing the *compatibility* of each local spatial feature vector with the global feature representation, which was the vector representation before the last FC layer. This is illustrated in Fig. 6.3. The motivation for this approach was that spatial features that align well with the global feature vector must contain important features for the prediction and should thus be upweighted. The compatibility between the vectors was computed either by their dot product or by a simple addition [208]. Increased performance was reported for classification tasks on several natural image datasets (e.g. CIFAR [209]), demonstrating the use of a trainable attention mechanism in a CNN.

Trainable attention was introduced in medical imaging tasks by Schlemper et al. [129]. The key difference with the work by Jetley et al. [208] was the introduction of an image-grid feature representation (with the same number of dimensions as the local feature representation) as a global feature, as opposed to a flattened vector. Grid-based gating enabled the global representation to be location specific, providing a richer gating signal. In the AG both local and global representations were linearly mapped to the same number of channels, summed together and sampled to the same spatial dimensions. This was followed by a ReLU activation and another linear mapping to a single channel. Finally, a sigmoid activation normalised the attention coefficients between 0 and 1, resulting in the final attention mask. The AGs were incorporated in a variant of VGG-Net for fetal US scan plane detection [129] and in a U-Net for abdominal organ segmentation from computed tomography [129]. Improved segmentation and classification performance was reported, and the resulting attention mask demonstrated good overlap with ground-truth structures of interest. However, to obtain optimal performance the network had to be trained separately per AG, followed by fine-tuning of the whole network.

6.2.2 Guided Attention Gates

All aforementioned works have so far incorporated AGs but did not directly regularise the learned attention masks [129, 208]. On the other hand, Yan et al. [205] used a similar AG as by Schlemper et al. [129], but the learned attention masks were also regularised with a segmentation loss. This was done to guide the attention towards regions known to be important for the prediction tasks, in this case, lesion segmentation masks. I will refer to AGs regularised with a segmentation loss as *guided* AGs.

Similar to Yan et al. [205], Yang et al. [206] regularised the AGs towards diagnostically relevant regions in breast histopathology images. However, their AG differed from previous work [129, 205] in that the feature representations of all intermediate layers were first concatenated, which was then used as input to the AG. The AG computed an attention mask using a single convolutional layer

and a sigmoid activation, and subsequently multiplied that with the concatenated feature representation. The final prediction was then made using a GAP operator and a single FC layer.

Slightly different from the previous guided AGs, but worth mentioning for completeness, are works that directly regularise feature maps (as opposed to learning an attention mask). In Yin et al. [210] the feature maps before the last FC layer were extracted and summed together with a weighted sum, the weights being the weight of the respective channel in the FC layer. The resulting summed map was then regularised with a segmentation loss. However, the disadvantage of such an approach is that the regularisation has a strong effect on all features, forcing the activations of all regions of interest to be high. On the other hand, the masking approach only enforces activations outside the regions of interest to be low, and does not constrain the attention inside the ROI. Furthermore, because the feature maps are extracted at the end of the network, these are of a lower resolution than the original image resolution, when using a standard encoder (i.e. ResNet).

6.3 Methods

6.3.1 Problem Definition

The guided attention network aims to make a prediction for a sample image $\mathbf{X}$ focusing on a certain ROI. This is achieved by incorporating an AG that masks out irrelevant regions with an attention mask learned with an attention module (Θ^s). Furthermore, the attention mask is regularised with a segmentation loss to match the desired ROI. In this section notation is used for 2D images, however, by replacing the respective 2D by 3D operations they also apply for the 3D image volumes.

6.3.2 Network Architecture

Given an image $\mathbf{X} \in \mathbb{R}^{w \times h}$, with w and h the spatial dimensions, the networks predicts a continuous value, in this case the GA, denoted by $\hat{y}$. The network backbone is based on the ResNet-18 encoder [122] that extracts intermediate feature representations, denoted by $\mathbf{Z}_i$, with a number of consecutive layer blocks.

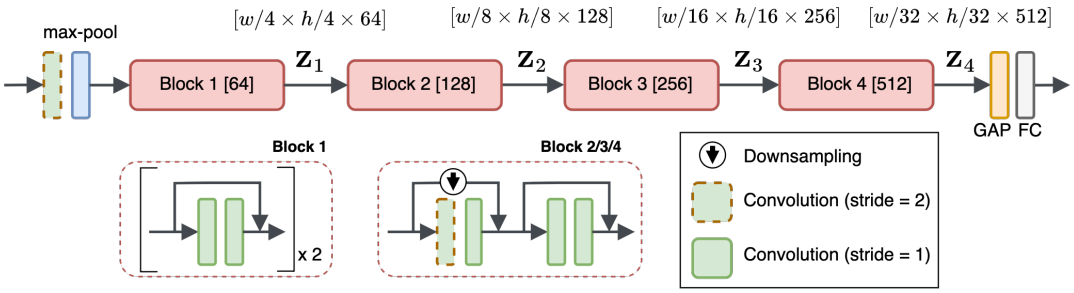


Figure 6.4: ResNet backbone encoder extracting intermediate feature representations, Z_i, of an input image $\mathbf{X} \in \mathbb{R}^{w \times h}$. All convolutional layers have a kernel size of 3×3 except for the first one that has a larger 7×7 kernel. The numbers in square brackets in each block indicate the number of feature maps in the convolutional layers. Each convolutional layer is followed by a ReLU activation and batch normalisation. Downsampling is performed with 1×1 convolutions with a stride of 2.

All illustrated in Fig. 6.4, the ResNet encoder consists of four blocks, each containing four consecutive convolutional layers with a 3×3 kernel. The number of filters per convolutional layer is 64 for the first block, and doubles for each consecutive block. Within the blocks, each set of two convolutional layers has a residual connection summing the output of the two layers with the input, allowing gradients to pass through directly. The first convolutional layer of the last three blocks has a stride of 2, gradually downsampling the extracted feature representations. In contrast to the standard ResNet-18 architecture, no stride of 2 is applied in the first layer block. This choice was made to maintain a higher resolution in the extracted feature representations, later used in the AGs. The first residual connection in each block also contains a single 1×1 convolution to match the feature size for summing.

Before the first block, the encoder also contains one convolutional layer with a 7×7 kernel and a stride of 2, as well as a max pooling layer that downsamples the input by two. All convolutional layers are followed by batch normalisation [196] and a ReLU activation. In the case of the residual connection, a single ReLU is applied after summing the feature representations. Finally, the output of the last block is pooled into a vector using a GAP operation, followed by a single FC layer to make the final prediction.

The extracted intermediate feature representation after block i is denoted by $\mathbf{Z}_i \in \mathbb{R}^{w_z \times h_z \times c_z}$, with w_z and h_z the latent spatial dimensions and c_z the channel

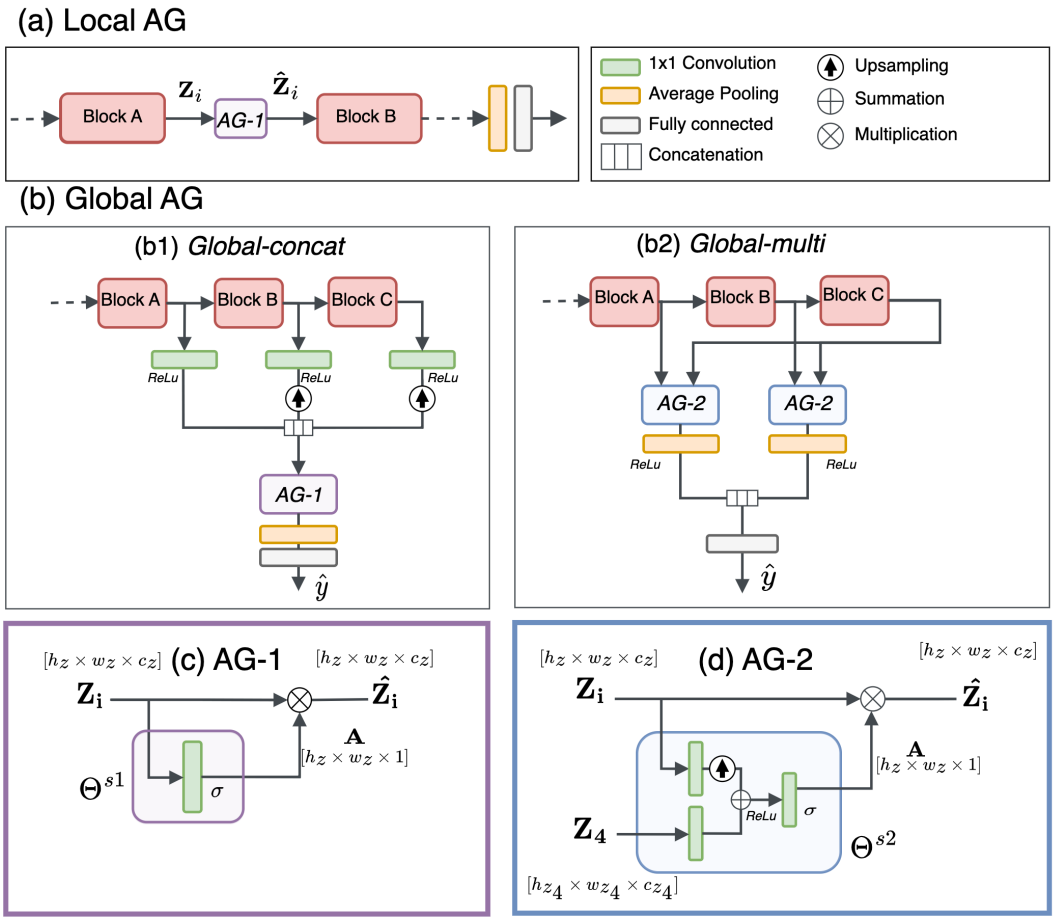


Figure 6.5: Overview of different AG configurations. (a) and (b) show the local and global AG architectures, and (c) and (d) the structure of the AGs with (c) one and (d) two inputs respectively. The AG schematics include the attention module (Θ^s), which learns the attention mask, in a shaded box.

dimension. In Fig. 6.4 these are annotated with their respective sizes.

6.3.3 Attention

In this chapter, I explored different types of AGs, which can be classified into *local* and *global* AGs. Here I define *local* AGs as AGs that only receive a single intermediate feature representation as input and output a masked feature representation of the same size. These can be positioned in between blocks of the encoder without any additional changes to the architecture (Fig. 6.5a). On the other hand, *global* AGs are defined as AGs that receive more than one intermediate feature representation as input. As such, these need to be positioned on top of the encoder architecture,

and the prediction is made from the output of the AG (or from the combination of outputs of several AGs), as illustrated in Fig. 6.5b.

6.3.3.1 Local Attention Gates

A local AG multiplies a local feature representation $\mathbf{Z}_i$ with a learned attention mask, $\mathbf{A}_i$, to obtain a masked feature representation: $\hat{\mathbf{Z}}_i = \mathbf{Z}_i \cdot \mathbf{A}_i$. The masked feature representation is subsequently passed through the remainder of the network. The attention mask is learned by an attention module, Θ^{s1}, that consists of a convolutional layer with a 1×1 kernel mapping the feature representation to a single channel, followed by a sigmoid activation to produce a mask with values between 0 and 1.

6.3.3.2 Global Attention Gates

A global AG receives multiple intermediate feature representations as input. Based on the literature [205, 206], two variations of global attention were explored: *global-concat*, where the intermediate feature representations are concatenated before providing them to a single AG [206], and *global-multi*, where multiple AGs are combined and each of them receives exactly two feature representations as input [205], of which one is typically the last ($\mathbf{Z}_4$).

Global-concat The *global-concat* AG (Fig. 6.5b1) closely resembles the *local* AG architecture. However, instead of using a single intermediate representation as input, the concatenation of a set of intermediate representations is used as input for the attention module (Θ^{s1}). Before concatenation, all intermediate representations are upsampled to the same size, and convolved with a set of additional convolutional layers (one for each $\mathbf{Z}_i$). After the attention module, the masked feature representation, $\hat{\mathbf{Z}}_i$ is pooled with GAP and passed to a single FC layer, resulting in the final prediction. As the concatenated feature representations are given to a single attention module, this type of AG results in only a single learned attention mask.

Global-multiple The *global-multiple* AG (Fig. 6.5b2) follows the architecture proposed in [129], in which each attention module receives the feature representation after the last block, $\mathbf{Z}_4$, and a local intermediate feature representation, $\mathbf{Z}_i$ as input. As this type of attention module receives two inputs, its architecture differs and will be denoted by Θ^{s2}. Each attention module learns an individual attention mask and multiplies this with the respective intermediate feature representations to obtain the masked feature representations. To make a single prediction, the masked feature representations are pooled with a GAP operation, resulting in a single vector per attention module. These vectors are then concatenated and given as input to a single FC layer to make the final prediction.

The architecture of this attention module, Θ^{s2}, is illustrated in Fig. 6.5d. First, both representations are linearly mapped to the same number of feature channels using 1×1 convolutions. The feature map with the smallest spatial dimensions is then upsampled to match the size of the other representation. Subsequently, both feature representations are summed together and activated with a ReLU activation. The summed representation is then convolved with another 1×1 convolution, reducing the number of channels to 1. Finally, a sigmoid activation is applied resulting in the final attention mask ($\mathbf{A}$).

6.3.4 Network Training

The network is trained with a combination of a regression loss, regularising the predicted ages, and a segmentation loss, regularising the learned attention masks. For the regression loss, the mean squared error (MSE) ($\mathcal{L}^{\mathcal{MSE}}$) was used, and for the segmentation loss the DSC overlap ($\mathcal{L}^{DSC}$). The DSC loss was computed between the predicted attention mask ($\mathbf{A}$) and ground-truth attention mask down-sampled to the same size using nearest neighbour interpolation. The total training loss was then given by:

$$\mathcal{L} = \mathcal{L}^{\mathcal{MSE}} + \alpha \mathcal{L}^{\mathcal{DSC}} \tag{6.1}$$

with α as a weighing factor between the two loss components. It was hypothesised that the network should first learn to attend to the desired ROI, which is enforced with the segmentation loss, and then focus on learning accurate age predictions rather than optimising the segmentation loss further. For this reason, a weight scheduler was used for α that decayed during training, so that the weight was high early in training and low later.

6.4 Experimental Set-up

Experiments were performed with 2D US images and 3D US volumes, with the task being GA prediction.

6.4.1 Dataset

INTERGROWTH-21st For the main experiments 2D images and 3D volumes from the INTERGROWTH-21st study were used (as described in Chapter 4). The volumes used in this chapter were all aligned to the same coordinate reference system and size normalised by scaling them to the average brain volume at 30 GWs [178]. Size normalisation was performed to enforce the network to focus on structural-developmental patterns rather than predominantly on size differences. The 2D images used were transventricular standard planes extracted from the 3D volumes.

A total of 3359 volumes from the INTERGROWTH-21st dataset were used, which were split into train/validation (80%) and test volumes (20%). For the 2D experiments, the train/validation volumes were used for 5-fold cross-validation where for each fold 10% was used for validation. Due to computational constraints, for the 3D experiments only a single fold was trained. The images were evenly split across the GWs, with each split containing a comparable number of images showing a visible left (or right) hemisphere. An overview of the distribution of data in this chapter is shown in Fig. 6.6.

Segmentations of the ROI were required to train the network with the segmentation loss. In this work, CoP and CB segmentations were used. The CB segmentations were obtained using the subcortical segmentation network presented in Chapter

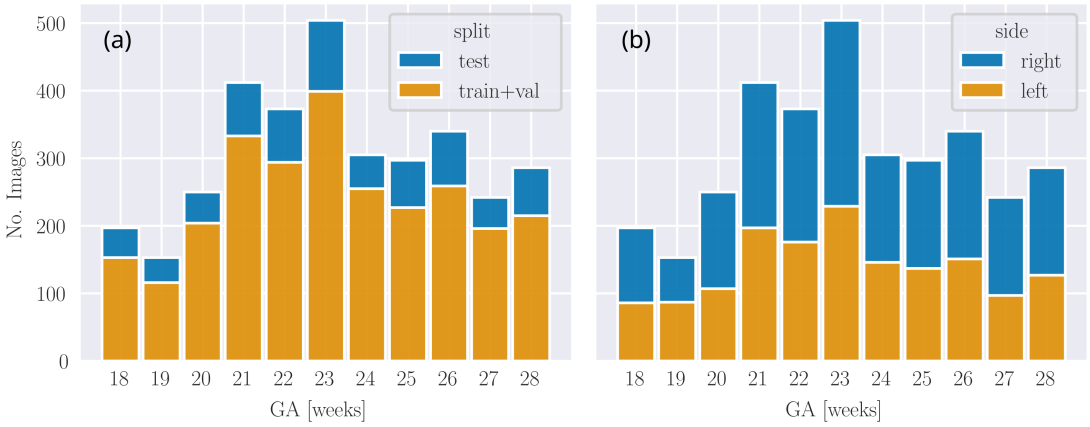


Figure 6.6: Data distribution of the 3359 INTERGROWTH-21st US volumes used in this chapter, (a) shows the train/test data split across gestation, and (b) the visible hemispheres of each of the image volumes.

5. As it is well known that an increased number of training volumes results in increased performance, the segmentation network was retrained using the nine training volumes as well as part of the test set volumes used in Chapter 5 (resulting in a total of 20 training volumes). This corresponds to the network used to evaluate the segmentation generalisation performance on the LUMC dataset in Section 5.4.9.

The CoP segmentations were obtained by applying TEDS-Net to the 3D volumes [211], an algorithm developed for topologically correct segmentations that can be used to segment the CoP in 3D volumes. Both segmentation algorithms were applied to the aligned (non-size normalised) 3D volumes. The segmentations were then upscaled to the size normalised volumes using a scaling transform with the known scaling parameters. Segmentations for the 2D images were extracted from the 3D images using the corresponding plane coordinates.

Both the TEDS-Net and subcortical segmentation networks were originally developed for US volumes from 18 to 26 GWs. However, it was observed that up until 28 GWs the resulting segmentations were of reasonably good quality, which is illustrated in Fig. 6.7 As the segmentations are only used to *guide* the networks' attention to the desired ROI, and segmentation performance is not the main objective, the quality of the masks does not need to be perfect. Furthermore, the DSC segmentation loss is computed on the resolution of the feature representation

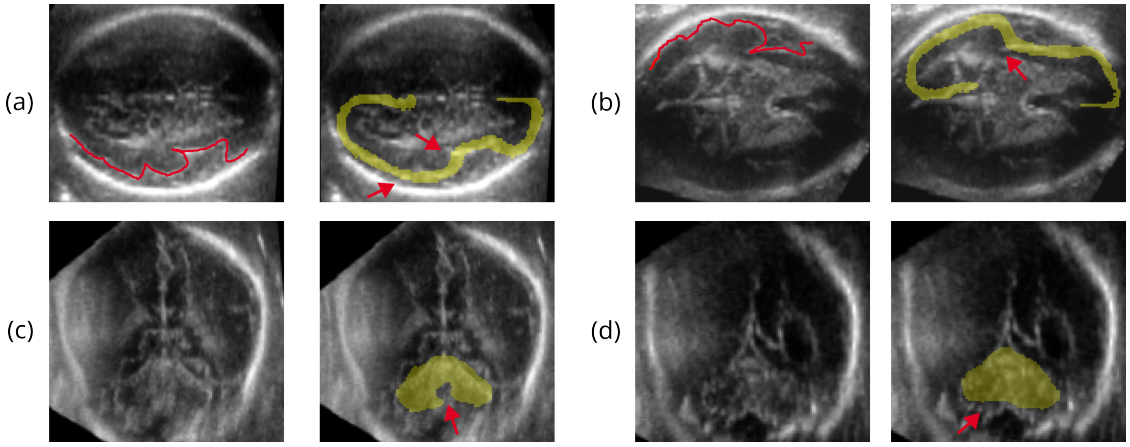


Figure 6.7: Example of segmentation quality at 28 GWs, (a) and (b) show axial planes with the CoP segmentation produced by TEDS-Net [211], and (c) and (d) coronal planes with the CB segmentation (produced with the subcortical segmentation model from Chapter 5). For clarity, an approximate outline of the CoP has been drawn in red, and arrows indicate where the produced segmentations show errors. It can be observed that most errors are for high-resolution details, and will have little effect on the loss computed in the downsampled resolution.

($\mathbf{Z}_i$), with the ground-truth mask downsampled to match this resolution. As the earliest feature representation, $\mathbf{Z}_1$, is already downsampled by a factor of 4, high-resolution details in the GT mask will not have any effect on the computed loss. As a larger GA range is desired to increase the clinical application of the model, it was therefore decided to use an age range between 18 and 28 GWs for all experiments.

LUMC Dataset To assess the generalisability of the proposed network, I also tested the final 2D and 3D models on a second independent test set, namely the LUMC dataset (as described in Section 4.2). For this analysis, all available image volumes between 18 and 28 GWs on which the SF was visible were used, resulting in a total of 156 volumes. As for the INTERGROWTH-21st dataset, all volumes were aligned and size normalised as preprocessing steps [178].

6.4.2 2D Experiments

For the 2D images, standard transventricular planes were used, as this is one of the most important planes in routine US examinations. On this plane, the CB is not visible, and therefore all 2D experiments were only performed with the CoP as ROI.

Model Configuration Five-fold cross-validation was performed to determine the optimal model configuration. As introduced in Section 6.3.3, several AG architectures were explored. For each of the architectures, the best configuration was determined using cross-validation and was as follows:

- *local*: Two AGs were used, one after the second layer block, and one after the third ($\mathbf{Z}_2$ and $\mathbf{Z}_3$).

- *global-concat*: The AG received all intermediate feature representations as input ($\mathbf{Z}_{1-4}$) and each of the individual convolutions before concatenation had a number of filters equal to half of the number of input channels (i.e. $c_z/2$ for $\mathbf{Z}_i \in \mathbb{R}^{h_z \times w_z \times c_z}$)

- *global-multi*: three AGs were used, after blocks 1, 2 and 3. Each AG received the respective intermediate feature representation ($\mathbf{Z}_{1-3}$), and the feature representation after the last layer ($\mathbf{Z}_4$) as input. The number of filters in the first two layers of the attention module was set to half the number of channels in the respective intermediate representation.

The *global-concat* AG configuration was found to result in optimal performance and was therefore used for the remainder of the experiments.

The weight of the attention loss, α, is another key hyperparameter as it balances the contribution of the MSE prediction loss and the DSC segmentation loss. For this reason, experiments were performed with a range of weight values for α, both with and without decay of the weight.

Baseline Comparisons After determining the optimal model configuration, the best cross-validation fold was selected based on the validation MAE and applied to the test set. The resulting performance was compared to several baseline models:

- ResNet baseline: network architecture without AGs

- ResNet baseline - masks only: ResNet baseline trained with binary segmentation masks as input instead of US images. This experiment was performed to determine the amount of information contained within the mask itself.
- ResNet baseline - masked images: ResNet baseline trained with US images masked with the ground-truth segmentations. This experiment resembles the baseline of masks only, but also provides the network with intensity information in the ROI.
- *unguided* attention: optimal AG architecture but trained without the segmentation loss (i.e. $\alpha = 0$).

6.4.3 3D Experiments

In the 3D volumes both the CB and CoP segmentations were used as ROIs individually, as well as their combined region.

Model Configuration The optimal architecture was assumed to be the same for 2D and 3D images, and was thus not individually adjusted but rather set to the optimal 2D architecture (*global-concat* with all intermediate representations as input). Due to the importance of the loss weighing value α, this parameter was tuned specifically for the 3D experiments using a single training run per parameter value.

Baseline Comparisons For the 3D experiments, the same baseline comparisons were carried out as for the 2D volumes. For the experiments trained with masks or masked images, the CB and CoP masks were used individually as well as the union of both regions. The reported performance for this comparison is the test set performance of the single training run.

6.4.4 Visualisation of Attention Maps

To visualise the learned attention, both the learned masks ($\mathbf{A}$) and the masked feature representation $\hat{\mathbf{Z}}$ will be visualised. To visualise the multichannel masked feature representations, I used the same approach as proposed for the visual

attribution method CAM [121], and summed the representations across the channel dimension using a weighted sum. The weights are determined from the FC weight of each respective channel after the (GAP) layer. All individual average maps were then normalised between 0 and 1.

Furthermore, weekly average attention maps were created. For the 2D images and the 3D volumes with the CoP as ROI, these were created by grouping the attention maps by GW and visible hemisphere, and then averaging per group by taking an average of the attention intensities. The averages for the left and right hemispheres were then concatenated together at the midline, and again normalised between 0 and 1 to obtain a single mean representation per GW. For the 3D volumes trained with the CB as ROI, the attention maps were only grouped by GW and not by hemisphere as typically both CB hemispheres are well visible.

While a simple average is likely a good approximation of the mean representation, it does not account for shape differences between the individual volumes. Therefore, it would be better to non-linearly register the activations to a common atlas space before taking the average. The network used for segmenting the CoP generates a deformation from an initial prior shape, resulting in a segmentation. The inverse of this deformation can be used to register all activation maps back to the prior shape, in which they can be averaged more accurately. While resulting in a more accurate average, and thus demonstrating the locations of high attention more precisely, this mapping does lose information about the width (i.e. spread) of the attention, as all activations maps are registered to the same prior shape. For this reason, it will only be used to generate 3D surfaces of the attention across gestation (Fig. 6.20).

6.4.5 Implementation

The networks were implemented with Pytorch Lightning using Python 3.8 and Pytorch 1.13. All experiments were performed on an Nvidia A10 with 24 GB RAM.

A 2D ResNet-18 pre-trained on ImageNet was used as a backbone encoder for the 2D experiments. For the 3D experiments, a 3D ResNet-18 encoder was used without any pre-training. The learning rate was set to 5e-4, decaying with a factor

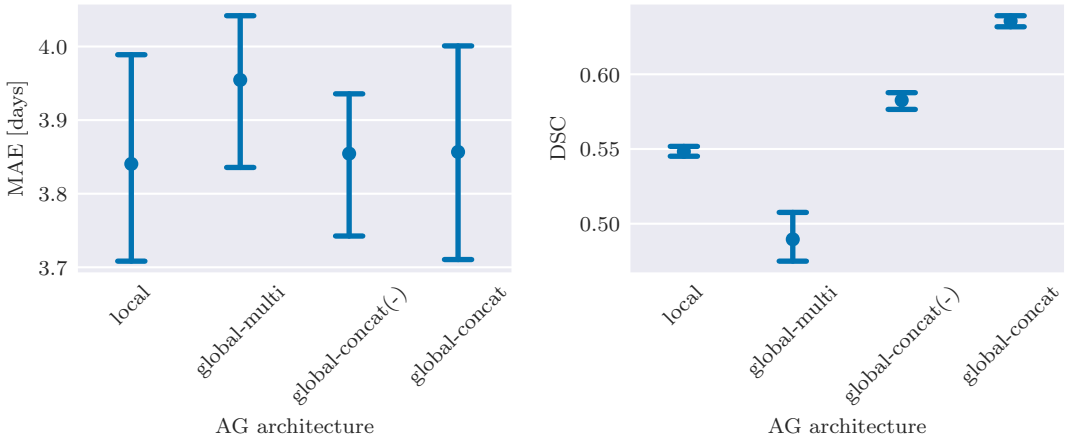


Figure 6.8: Comparison of the different attention architectures. For each type, the average validation performance across 5 folds is shown with 95% confidence intervals. On the left, the MAE is shown for each AG type, and on the right the DSC between the predicted attention mask (**A**) and the ground-truth ROI mask. If multiple AGs were present in the architecture, a mean of the individual DSC is reported. The *global-single* model outperforms the other models in terms of DSC significantly ($p < 0.001$, paired t-test with Bonferroni correction for multiple comparisons), whereas the MAE performance is not significantly different across the model architectures.

of 2 every 50 epochs. The 2D networks were trained for 200 epochs with a batch size of 30. Due to GPU memory constraints, the 3D volumes were only trained with a batch size of 8, and for a total of 250 epochs as it was observed that these converged more slowly than the 2D networks. For the 2D experiments, the initial weight of the segmentation loss, α, was set to 10^3, and decayed by a factor of 10 after every 20 epochs. This α was set to 10^4 for the 3D experiments, decaying only every 40 epochs, also due to the slower convergence.

6.5 Results

In this section, an overview of the experimental results will be presented. The results for the 2D standard planes are presented first, in which the CoP is used as a ROI mask. This is followed by results on the 3D volumes, using both the CoP and CB as segmentation masks (combined and individually).

6.5.1 2D Standard Planes

6.5.1.1 Model Configuration

An extensive hyperparameter search was carried out to determine the optimal AG model configuration using 5-fold cross-validation. The results presented in this section are the validation performance averaged across the 5 folds.

Architecture In Fig. 6.8 the performance of the different AG architectures (as presented in 6.3.3) is shown. Additionally, also the *global-concat* architecture without representation-specific convolutional layers is shown as *global-concat*$^{(-c)}$.

It can be observed that the MAE is largely comparable across most AG architectures (MAE of 3.84, 3.85, 3.86 for the local, *global-concat*$^{(-c)}$ and *global-concat* AG, respectively), with the *global-multi* AG performing somewhat worse than the others (MAE of 3.95). On the other hand, the DSC performance shows a larger variation between the architectures (DSC ranging from 0.48 for the *global-multi* to 0.63 for the *global-concat*), indicating that some AGs are able to learn more accurately (i.e. matching the ground-truth ROI) attention masks. As both the *local* and *global-multi* architecture have more than one AG, and therefore more than one learned attention mask, the reported DSC is an average of the individual AG masks.

The *global-concat* architecture significantly ($p < 0.001$, a paired t-test with Bonferroni correction for multiple comparisons) outperformed the other architectures in terms of DSC. This architecture can, in contrast to the other configurations, use all intermediate feature representations to learn a single attention mask, which, as expected, results in more accurate attention masks. For this reason, this architecture was used in the remainder of the experiments, and will simply be referred to as *guided attention*.

Attention Inputs The effect of varying the number of input features to the *global-concat* AG is shown in Fig. 6.9. It can be observed that including more intermediate feature representations results in an improvement of the MAE. Further, using the last three feature representations ($\mathbf{Z}_{2-4}$, MAE of 3.88) as inputs clearly

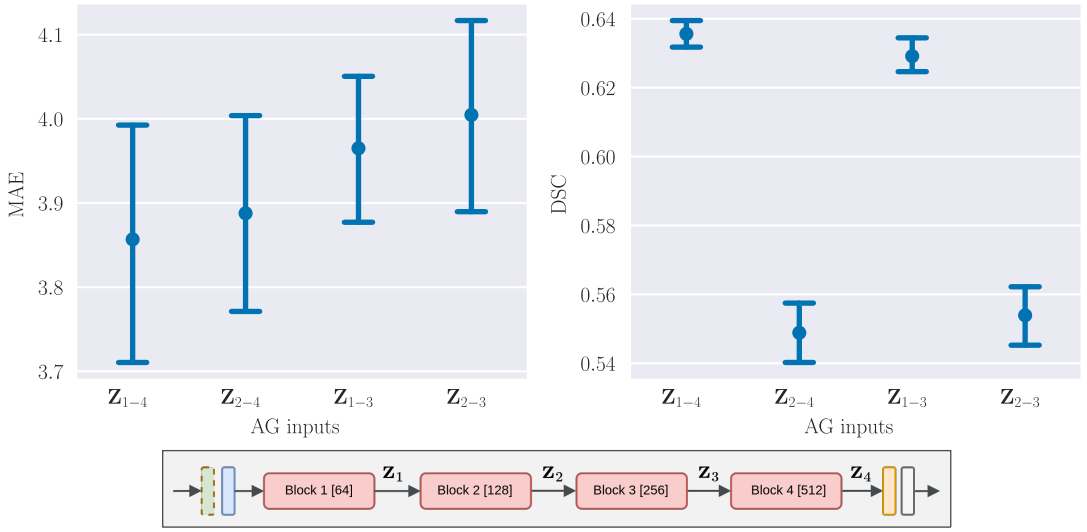


Figure 6.9: Overview of the performance for different inputs to the AGs. The ResNet architecture with the respective intermediate feature representations is illustrated underneath. Providing all intermediate feature representations ($\mathbf{Z}_{1-4}$) as input results in the best MAE performance and DSC overlap of the attention map.

result in better prediction performance than using only the first three ($\mathbf{Z}_{1-3}$, MAE of 3.96). This indicates that the deeper feature representations are more important for the MAE, aligning with the fact that representations of deeper layers are typically more informative for the task objective. On the other hand, the DSC performance shows an opposite effect, with a higher DSC for including the $\mathbf{Z}_{1-3}$ (DSC of 0.63) than for the $\mathbf{Z}_{2-4}$ (DSC of 0.55). This could be explained by the fact that earlier representations contain high-resolution information that is lost in deeper layers, therefore contributing to a better segmentation performance. For all experiments in the remainder of this chapter, all feature representations were used as inputs ($\mathbf{Z}_{1-4}$).

Attention Loss Weight The performance for different weighting values α between the MSE and DSC loss is shown in Fig. 6.10, demonstrating the performance for experiments with and without attention weight decay. It can be seen that when using weight decay, the MAE performance improves up to an α of about 10^3 (MAE of 3.86). This confirms the hypothesis that by *guiding* the network to known areas of interest (i.e. the CoP), it is able to better learn the discriminative age patterns, resulting in improved performance. The DSC performance also behaves

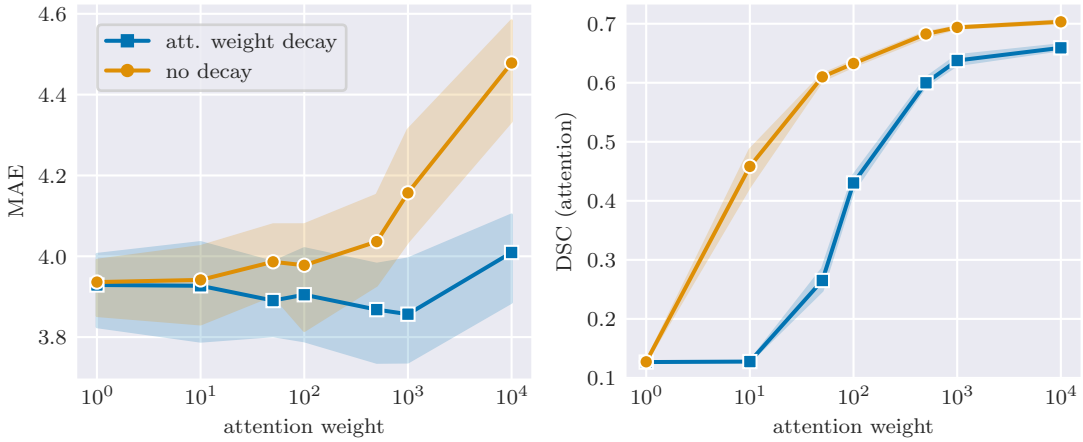


Figure 6.10: Overview of average performance across 5 folds for different weights of the DSC in the training loss. The shaded areas indicate the 95% confidence intervals. The left panel shows the MAE performance and the right the DSC of the resulting binarised attention masks with the ground-truth segmentation masks. The performance is shown for a model trained with decay of the weight during training (*att. weight decay*), and for a model with a continuous weight (*no decay*). For the experiments with weight decay, the initial α is shown. The attention weight decay can be seen to result in better MAE for a similar DSC.

Table 6.1: Performance on the test set with 2D images for the guided attention and several baselines. The attention DSC is computed between the learned attention mask and the ground-truth CoP segmentations, (*) indicates significance with respect to the ResNet baseline trained with images (1a, $p<0.01$).

	MAE [days] ↓	< 1 GW [%] ↑	DSC ↑
Lee et al [88]	3.9	86.9	-
1. ResNet baselines			
1a. images	3.98 (3.21)	83.73	-
1b. masks only (CoP)	4.67 (3.82)	77.37	-
1c. masked images (CoP)	4.73 (3.98)	74.96	-
2. Attention experiments			
2a. Unguided AG	3.88 (3.19)	85.71	0.13 (0.01)
2b. Guided AG (CoP)	**3.73** (2.99)*	**87.69**	**0.63** (0.05)

as expected, with higher weights resulting in better segmentation performance. Furthermore, the segmentation performance obtained with the attention decay is close to that without, especially at higher initial weight values (DSC of 0.64 versus 0.69, respectively, at an α of 10^3). For all remaining experiments on the 2D images, a weight of 10^3 was used with attention weight decay.

6.5.1.2 Performance Analysis

After determining the optimal model configuration, the performance on the test set was obtained for the guided attention and several other baselines. These results are shown in Table 6.1.

Baselines It can be observed that the guided attention significantly outperforms the ResNet baseline trained with the same 2D US images (MAE of 3.73 vs. 3.98 for the guided attention (2b) and ResNet baseline (1a) respectively, $p < 0.01$ from a paired t-test), showing that the addition of the guided AG results in improved performance. The baseline only trained with CoP masks (1b), achieves a lower performance than those trained with the 2D images itself. However, the MAE of 4.67 days demonstrates that the shape of the CoP contains a considerable amount of age-predictive information. As the CoP shape changes notably during the second trimester, this is as expected and confirms the importance of this structure for the age prediction task. The effect of providing the network with not only the CoP shape, but also with intensity features, can be observed from the model trained with masked CoP images (1c). While these should contain more information than only the binarised masks, the comparable performance (MAE of 3.67 vs. 4.73 for the masks only and masked images respectively) demonstrates that the intensity features are either less age-discriminative or that the network is not able to successfully use this information.

Unguided Attention To assess the effect of the segmentation loss, the AG network was also trained without this loss component (referred to as *unguided AG*). It can be seen in Table 6.1 that the *guided* attention obtains better performance than the *unguided* attention (MAE of 3.73 versus 3.88, respectively), clearly demonstrating that guiding the network's focus to age discriminative patterns improves the performance. For the *unguided* attention also the DSC with the CoP was computed, resulting in a low DSC of 0.13. This illustrates that the AG does

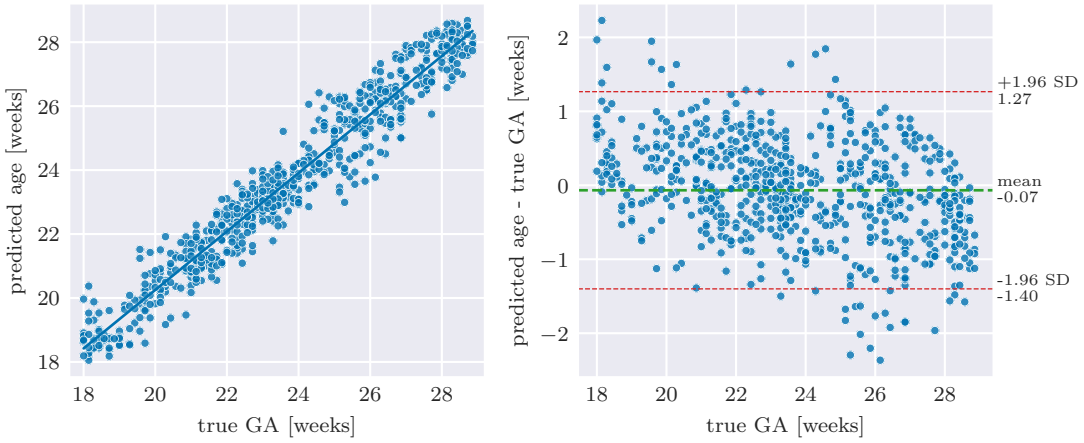


Figure 6.11: Brain age predictions of the single best fold on the 2D images using guided attention. The left panel shows the true GA versus the predicted age with a linear fit. The right panel shows the difference between predicted and true age across gestation in a modified Bland-Altman plot, where the true GA has been plotted on the x-axis as opposed to the mean of the two measures. The dashed lines in the Bland-Altman plot indicate the mean difference and the 1.96 SD range.

not learn to attend to the CoP without the AG segmentation loss (qualitative results are shown in Section 6.5.1.3).

Comparison to Other Work The guided AG obtains a competitive MAE of 3.73 days with 87.69 % of the errors being within 1 GW. This is better than the previously reported error (MAE of 3.9 days) for fetal brain age prediction from 2D US images in the second trimester [88]. In [88] also images without size or scale information were used, making it most similar to this work. However, the images used in that study [88] were 2D images acquired with a 2D probe, as opposed to extracted 2D planes from 3D volumes used in this study. It is difficult to assess whether using directly acquired 2D images makes the age prediction task more or less challenging, and a completely fair comparison is therefore difficult to make.

Individual Predictions In Fig. 6.11 the true versus predicted GA for all test set samples is shown with a linear fit. It is evident that the model is able to accurately predict the GA, and that the linear fit is close to the line of equality. On the right in Fig. 6.11 a modified Bland-Altman plot is shown with the prediction error versus

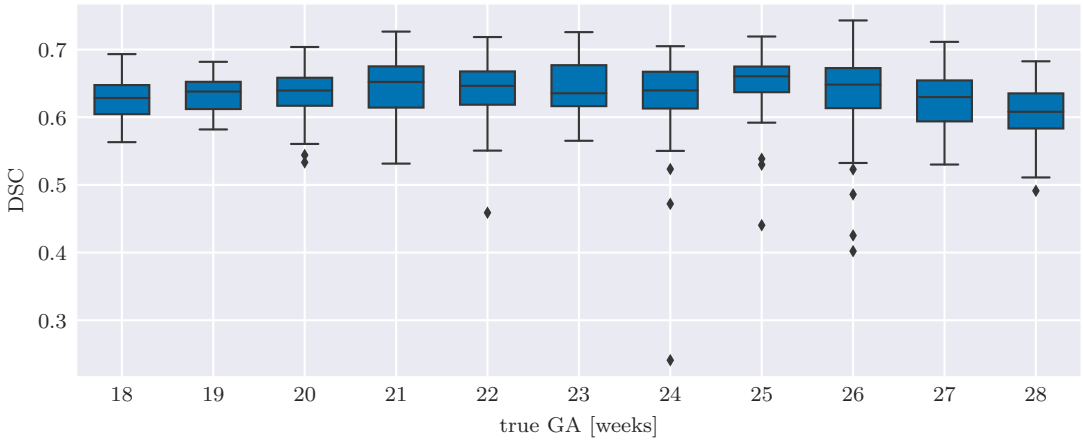


Figure 6.12: DSC segmentation performance of the learned attention mask with ground-truth CoP segmentations grouped by GW. The DSC performance is relatively stable accross gestation, showing a slow decline in performance at the end of the GA range.

the true GA. It can be seen that there is a small deviation towards the mean, with more positive errors for the lower ages (i.e. prediction higher than the true age) and negative for the higher ones. This indicates a slight deviation towards the mean, but is to be expected for a model trained with regression loss.

Segmentation Performance In Fig. 6.12 the DSC performance on the test set is shown across the GA range. DSC were computed between the learned attention masks and ground-truth CoP segmentations. It can be observed that the DSC is relatively stable across the GWs with few outliers. However, a slight performance decrease can be observed at both ends of the age range, most pronounced at the later GWs. This can partly be explained by the fact that the model has seen more comparable training samples in the middle of the age range. Furthermore, with increasing GA, the shape of the CoP becomes gradually more complex (due to operculisation, as described in Section 2.1.2) and there is an increase in shadowing artefacts due to the calcification of the skull, which both make it more challenging to compute an accurate attention mask.

6.5.1.3 Qualitative Results

Having shown that the proposed guided AG architecture obtains improved prediction performance, the learned attention can be visualised to obtain an improved understanding of the features used for the prediction.

Individual Images In Fig. 6.13 the results for three test set volumes with a different GA are shown. The sample image is shown on the left, followed by the ground-truth CoP mask, the learned attention mask ($\mathbf{A}$) and the masked feature representation after the AG ($\hat{\mathbf{Z}}$). It can be seen that the learned attention masks closely match the ground-truth masks in overall shape. The learned masks do appear less fine-grained, especially at the SF, but this is a result of the lower-resolution features used in the AG. From the masked feature representations it can be observed that the network mainly focuses on the SF for the first and third image, while it attends more to the midline (around the POF) in the second image. Both of these fissures undergo significant shape changes in the second trimester [32], showing that the model is able to learn salient image features.

Weekly Averages To achieve a better understanding of the attention throughout the GA range, Fig. 6.14 shows the learned masked attention maps ($\hat{\mathbf{Z}}$) of the test set averaged per GW (see Section 6.4.4). The attention maps were averaged and then scaled back to their original size. In Fig. 6.14 the average attention maps are overlaid on top of age-specific atlas images. It can clearly be seen that there is high activation around the SF, most concentrated at the earlier GWs (i.e. 18-20 GW).

The SF is one of the first fissures to develop and can be seen from about 14 GWs on fetal US [31]. It is therefore in line with the expectation that the network focuses on this area at the beginning of the age range. From about 23 GWs the attention shifts towards the POF, located around the midline. Then, at later weeks, the attention seems to shift back its focus towards the SF. This matches the observation that the POF is barely visible at the beginning of the second

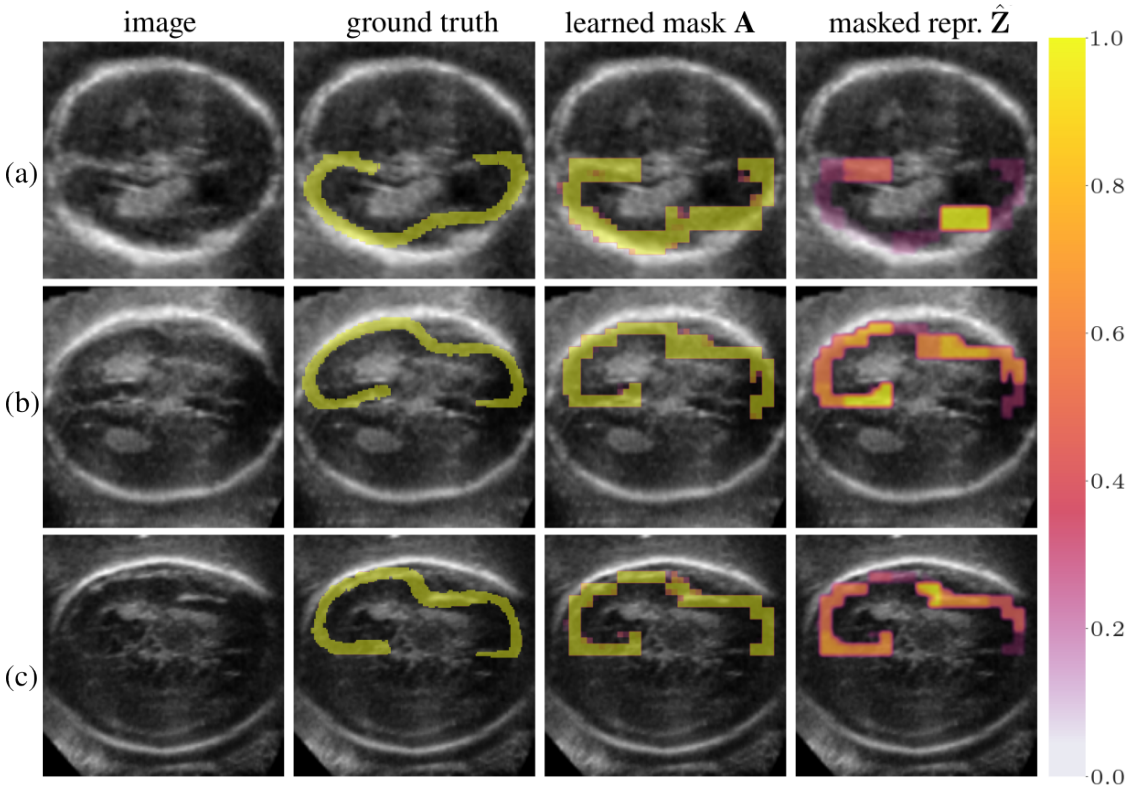


Figure 6.13: Qualitative results of guided attention for three test set images. Images have a ground-truth GA of (a) 18.3, (b) 22.7, and (c) 25.7 GWs, and are predicted (a) 19.3, (b) 22.1, and (c) 25.5. Shown is the ground-truth CoP mask, the predicted attention mask ($\mathbf{A}$) and the masked feature representation ($\hat{\mathbf{Z}}$). The masked feature representation ($\hat{\mathbf{Z}}$) is used for most other visualisations in this chapter.

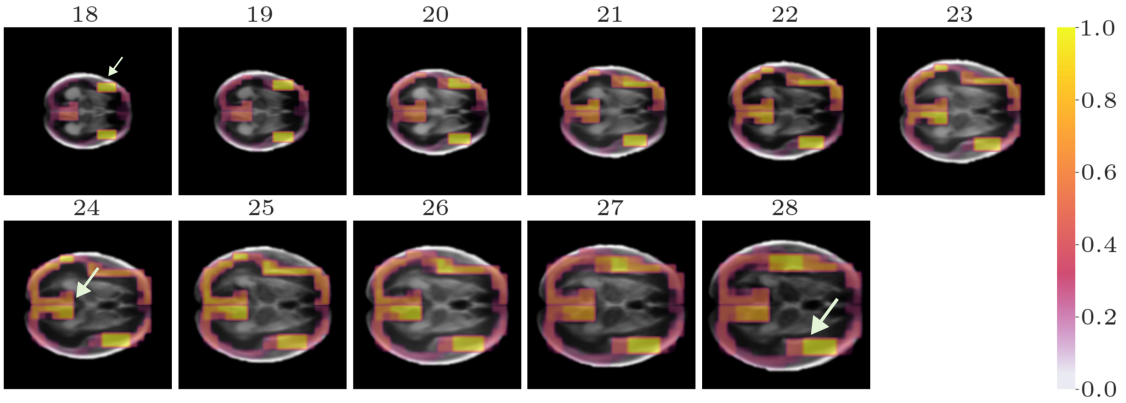


Figure 6.14: Average masked feature representations per GW ($\hat{\mathbf{Z}}$) for 2D images, higher values (i.e. yellow) indicate high average attention. Maps were obtained by averaging the attention per GW grouped by visible hemisphere, and concatenating them at the midline for visualisation. After computing the averages, these were scaled back to their true size and visualised as an overlay on weekly atlas images [9]. The attention is high at the SF (indicated with an arrow at 18 and 28 GWs), and at the POF (indicated with an arrow at 24 GWs).

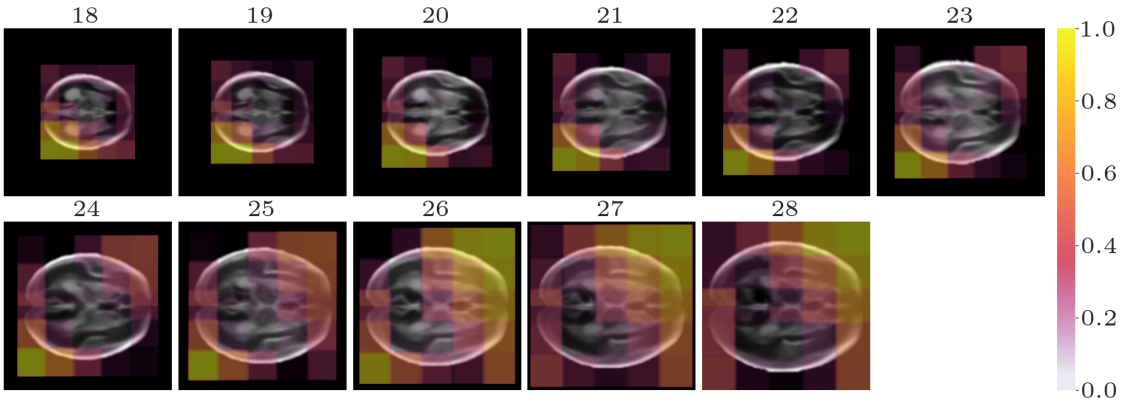


Figure 6.15: Average masked feature representations ($\hat{\mathbf{Z}}$) per GW for the network trained without segmentation loss. Averages were downscaled to their original size and shown overlaid on the atlas images.

trimester, and undergoes rapid chances from about 21 GWs, developing from a shallow indentation to a well-defined fissure in a few GWs [32].

To illustrate the differences between the *guided* and *unguided* attention, Fig. 6.15 shows the same weekly averages for the model trained without segmentation loss. The smaller size of the attention at younger ages is due to the fact that they were downsampled to their original size for visualisation. It can be seen that the attention shows a very block-like pattern, which is likely caused by the fact that the feature representation at the lowest resolution ($\mathbf{Z}_4$) has the largest contribution to the learned attention masks (due to their higher number of channels). The attention up to about 26 GWs is mainly focused on a background or skull pattern in the left bottom corner, whereas this shifts to the right top corner later in gestation. Furthermore, the attention between different training runs varied significantly, indicating that the areas of focus are largely determined by random initialisation of the model rather than developmental patterns. This thus clearly shows the advantage of the addition of a segmentation loss to guide the attention.

Finally, Fig. 6.16 shows the averaged saliency maps across gestation for the baseline ResNet model (1a in Table 6.1) with three explanation methods: (a) backpropagation [104], (b) integrated gradients [110], and (c) occlusion [113]. It can be seen that the attribution is quite spread out through the images, and differs substantially between the three methods. For all methods, there appears to be some

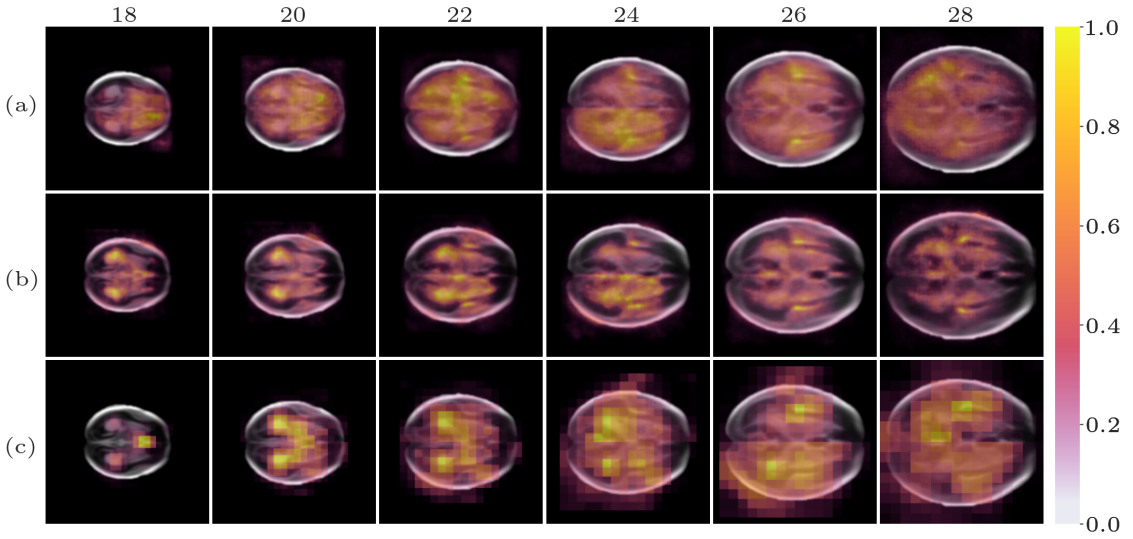


Figure 6.16: Average attribution maps per GW for three different saliency methods: (a) backpropagation [104], (b) integrated gradients [110], and (c) occlusion [113]. It is evident that, even when averaged across all test set samples, the attribution methods assign importance to different regions in the image.

focus on the SF and CP, but, due to the large variability, no meaningful conclusions about developmental patterns can be drawn. Furthermore, the differences between the explanation methods make them susceptible to confirmation bias, as the explanation that best matches the hypothesis could be preferred.

6.5.2 3D Image Volumes

While experiments with the 2D standard plane images provided interesting results, the ROI that can be selected is limited as it only contains a single slice of the brain. Therefore, experiments were also performed with 3D volumes, using both the CoP and the CB as regions of interest.

6.5.2.1 Model Configuration

Due to the computational complexity of training 3D image networks, the same AG architecture was adopted as for the 2D image models, namely the *global-concat* architecture.

Attention Loss Weight As learning a 3D segmentation is more challenging than a 2D, the attention weight loss was tuned specifically for the 3D volumes

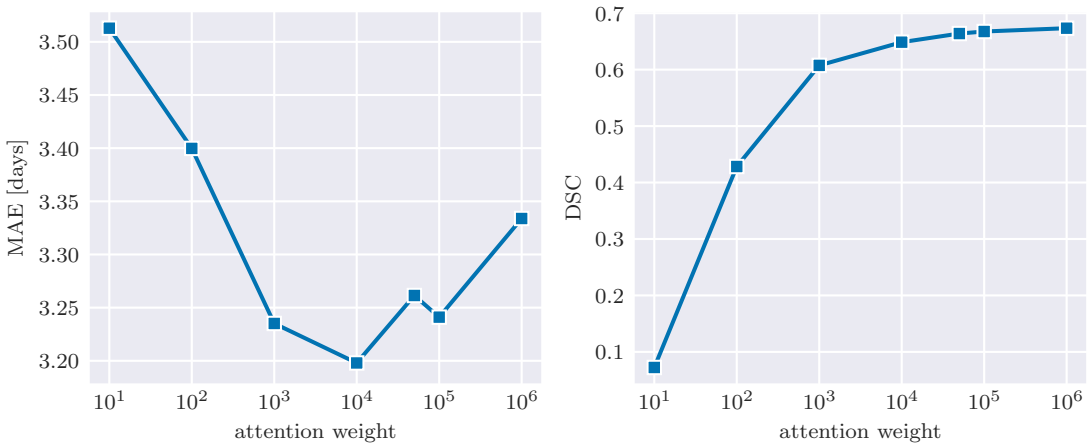


Figure 6.17: Hyperparameter tuning for the attention weight (α) for the 3D volumes with the CP as ROI. Only a single fold was trained for each α, and therefore no error bars are shown. The experiments shown were with a decaying α with a factor of 10 every 40 epochs.

using the CoP as ROI. These results are presented in Fig. 6.17. It can be seen that there is a large drop in MAE when increasing the weight up to about 10^4, already demonstrating the positive effect of the segmentation loss on the prediction performance. After 10^4 the MAE increases only very slowly, especially considering the large range of values used (10^1 - 10^6, note the difference in range with the 2D images). It was observed that training converged slower with the 3D volumes than with the 2D images, likely due to the more challenging task. This suggests that the network requires a higher initial weight to learn an accurate attention mask, which then improves prediction performance later in training.

6.5.2.2 Prediction Performance

An overview of all quantitative results for the 3D volumes on the test set images is shown in Table 6.2. It can clearly be seen that the guided AG obtains the highest performance, with all models (5a-c, trained with CoP, CB or both) obtaining better MAE performance than the baselines (significantly better for the CoP and both). The highest performing model is obtained using both the CB and CoP masks as ROI, resulting in an MAE of 3.45 days and 88.56% of errors within 1 GW. This thus demonstrates that the guided attention can significantly

Table 6.2: Performance of 3D models on test set compared to several baselines. The values between brackets indicate the standard deviations and '< 1 GW' refers to the percentage of test set volumes having a prediction error less than 1 GW. Stars indicate significance with respect to the ResNet baseline trained with images (1, (*) p<0.05, (**) p<0.01)

	MAE [days] ↓	< 1 GW [%] ↑	DSC[1]↑
Namburete et al. (2017) [86]	7.72		
Wyburd et al. (2021) [26]	4.1		
ResNet baselines			
1. Volumes	3.77 (3.03)	84.75	-
2. Masks only			
2a. CoP	4.48 (3.66)	79.10	-
2b. CB	4.91 (4.00)	75.28	-
2c. Both	4.37 (3.57)	79.66	-
3. Masked images			
3a. CoP	4.08 (3.29)	81.92	
3b. CB	4.89 (4.08)	76.00	
3c. Both	4.00 (3.29)	81.92	
Attention experiments			
4. Unguided Attention	3.78 (2.96)	85.73	-
5. Guided Attention			
5a. CoP	3.50* (2.84)	87.85	0.64 (0.05)
5b. CB	3.61 (2.76)	87.57	0.79 (0.09)
5c. Both	**3.45**** (2.82)	**88.56**	0.67 (0.05)

[1] DSC is computed with each of the respective structures

improve prediction performance when regularising the attention with known age-descriminative areas of the brain.

Baselines The same types of baselines were trained as for the 2D images, but experiments were performed with not only the CoP as ROI, but also the CB, as well as a combination of both the CB and the CoP (*both*). It can be seen that for all experiments performed with the three mask options (*masks only*, *masked images* and *guided attention*), the combination of both structures obtains the highest performance (MAE of 4.37, 4.00, and 3.45, respectively), closely followed by the CoP (4.48, 4.08, 3.50), and, with a larger margin, the CB (4.91, 4.89, 4.61). This aligns with the observation that the CoP undergoes more significant

shape changes during the second trimester than the CB. The CB mainly grows in volumetric size during the second trimester, but these changes were removed by upscaling all volumes to the same size.

In Table 6.2 it can also be seen that from all baselines, training a ResNet with the image volumes itself results in the highest performance (MAE of 3.75). Training with *masks only* obtains the lowest performance (MAE of 4.48 for CoP), followed by training with *masked images* (MAE of 4.03 for CoP). This suggests that the 3D ResNet is able to use the additional intensity features present in the masked images, resulting in better prediction performance. This is in contrast to the 2D results, where both *masks only* and *masked images* obtained similar results.

The unguided attention experiment (Table 6.2-4) results in a similar performance as the ResNet baseline, which is significantly lower than the guided attention experiment. This confirms that the AG without segmentation loss is not able to recognise salient image regions, thereby not improving the performance.

Comparison of Models with Different ROIs In Fig. 6.18 the true versus predicted GA is shown for three models trained with different ROIs in the segmentation loss: the CoP, the CB and the union of both. For each model, a linear fit is shown, and the individual samples have been coloured according to their residual in the first plot (trained with *both*). In the two other plots, the same colours have been used per sample, demonstrating the agreement between the three models. From the colours, it is evident that there is good agreement between the models, with outliers being largely preserved between them. This suggests that some of these outliers might show patterns of delayed or advanced development, resulting in larger MAE errors. In a healthy population, brain development is temporally consistent [11, 13], and therefore good agreement between models trained with different areas of interest is expected. This thus validates the performance of the guided attention model.

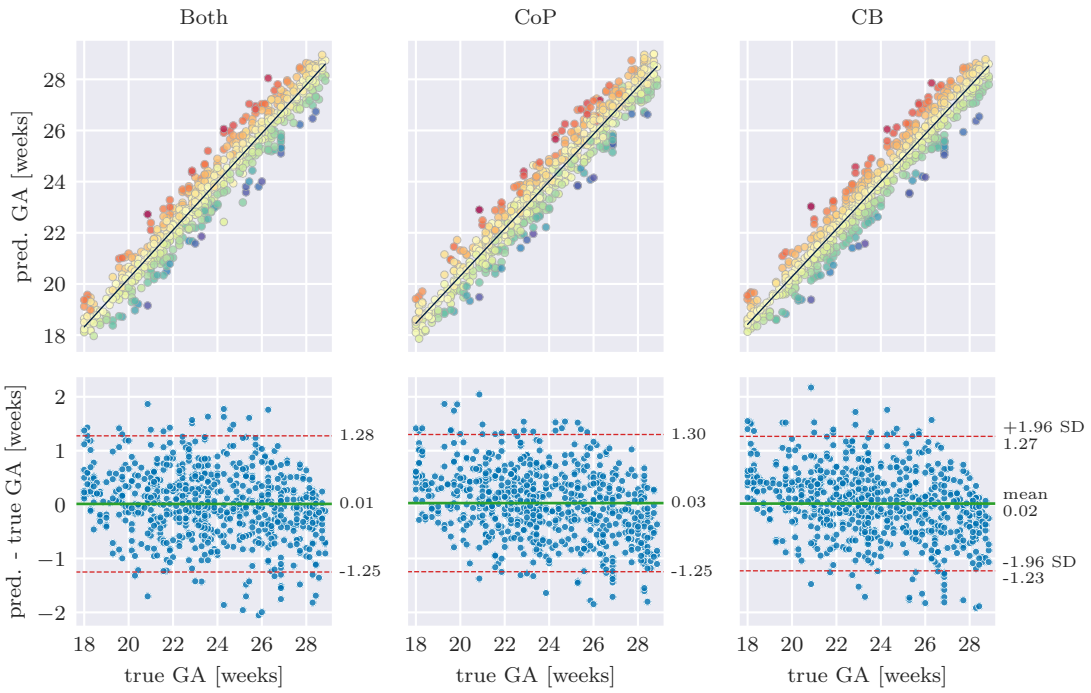


Figure 6.18: True vs predicted GA for models trained with the following segmentation masks: CoP, CB and the union of both. Linear fits are shown for each plot. Points in the top row were coloured based on their residual in the first column (*both*), and shown in the same colour in the other two plots. It can be seen that there is good agreement between the three models.

6.5.2.3 Qualitative Results

In Fig. 6.19 and Fig. 6.20 the average attention per GW is shown for the 3D volumes trained with a CoP ROI. In Fig. 6.19 the averages were obtained with an arithmetic mean of the individual activation intensities, and for each week a single axial plane is shown. Fig. 6.20 shows average attention maps that were obtained by projecting the individual attention maps back to a prior shape, corresponding to the CoP shape at 22 GWs, and averaging them there (as described in Section 6.4.4). For these attention maps, 3D visualisations are shown, for which both the left and right hemispheres are displayed side-by-side. It can be seen that for all GWs there is some focus on the SF, most pronounced at the start of the GA range and at 25 and 26 GWs. In the first half of the GWs, there is also a clear focus on the front of the brain. While this area does not correspond to a specific fissure, substantial changes in this area of the CoP can be observed in the second trimester.

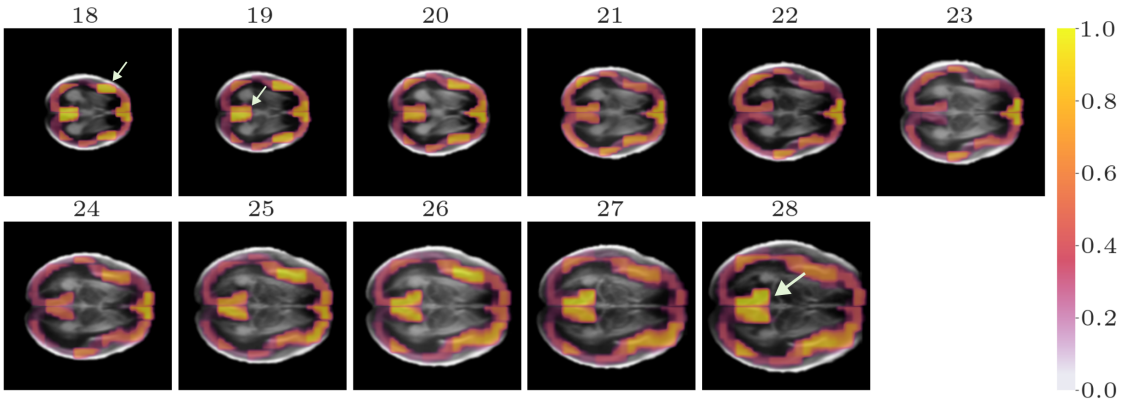


Figure 6.19: Averaged masked feature representations ($\hat{\mathbf{Z}}$) per GW for the 3D model trained with the CoP as ROI. Higher values (i.e. yellow) indicate high average attention. Maps were obtained by averaging the attention per GW grouped by visible hemisphere, and concatenating them at the midline for visualisation. After computing the averages, these were scaled back to their true size and visualised as an overlay on weekly atlas images [9]. Axial planes are shown for each GW. Arrows point towards the SF (at 18 GWs) and the POF (19 and 28 GWs) where the attention is high.

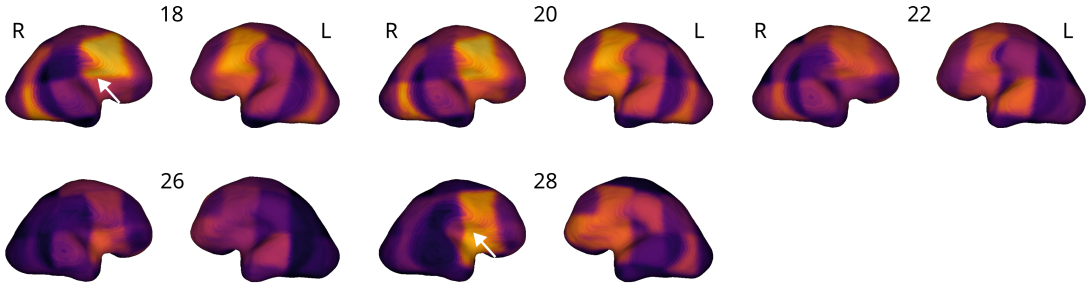


Figure 6.20: Average masked feature representations (($\hat{\mathbf{Z}}$)) per GW for the 3D model trained with the CoP as ROI. 3D visualisations of the attention on the CoP are shown for each second GW.

Lastly, clear attention can also be observed around the midline at the back of the brain, which corresponds to the location of the POF.

In Fig. 6.21 an axial plane for three individual test set volumes is shown for a model trained with CB attention masks. It can clearly be seen that the learned attention focuses solely on the CB, showing a relatively uniform spread across the CB. This demonstrates that guided attention is not only able to learn the GA from the CoP but also from the CB.

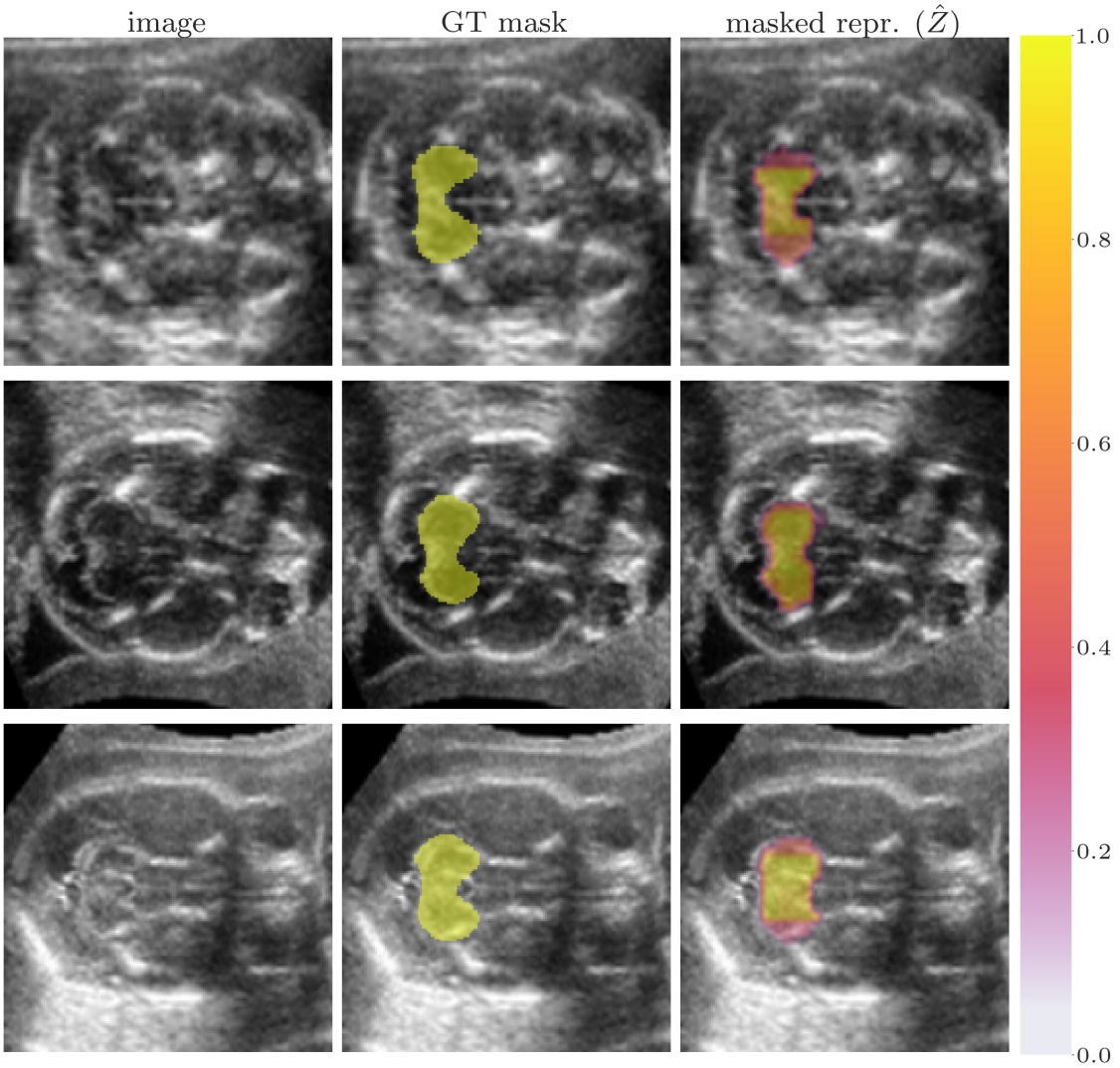


Figure 6.21: Masked feature representation ($\hat{\mathbf{Z}}$) of three individual test set volumes trained with the CB as ROI. Axial planes on which the CB is visible are shown for each sample.

6.5.3 Generalisability

So far, I have demonstrated the performance on the INTERGROWTH-21st test set volumes. In order to validate the performance on an external test set, the trained models were also applied to the LUMC dataset without any fine-tuning. The results on this dataset (N=156) are shown in Table 6.3. It can be seen that for both the 2D images and 3D volumes the best performance is obtained using the *guided* attention with the union of CoP and CB masks. This confirms the hypothesis that *guided* attention can improve the performance on unseen domains

Table 6.3: Age prediction performance on an independent test set (LUMC). The values between brackets indicate the standard deviations and '< 1 GW' refers to the percentage of test set volumes having a prediction error less than 1 GW. All guided attention results (6) were found to be not significant with respect to the ResNet Baseline.

	MAE [days] ↓	error < 1 GW [%] ↑
2D Images		
1. ResNet Baseline	5.37 (4.89)	69.87
2. Unguided Attention	5.52 (4.77)	71.15
3. Guided Attention (CoP)	**5.24** (4.74)	**75.00**
3D Volumes		
4. ResNet Baseline	4.21 (4.21)	80.77
5. Unguided Attention	4.67 (4.70)	80.13
6. Guided Attention		
6a. CoP	4.01 (4.11)	88.46
6b. CB	3.83 (3.94)	89.74
6a. Both	**3.73** (4.11)	**90.38**

by learning more transferable features.

While the reported MAE on the LUMC dataset is slightly lower than that on the INTERGROWTH-21st dataset (MAE of 3.73 for the LUMC dataset versus 3.45 on INTERGROWTH-21st, using both CoP and CB as masks), this is a minimal drop in performance considering that the volumes were acquired with different scanners and at different institutions. Furthermore, the percentage of scans with an error lower than 1 GW is more than 90%, indicating that the larger MAE can be attributed to a few samples with large errors.

Interestingly, the performance with only the CB as ROI is better than that with the CoP only for the LUMC dataset. This suggests that the CB appearance might be more similar across the two datasets than that of the CoP. However, the combination of both segmentation masks results in the best performance, which is in line with the experiments on the INTERGROWTH-21st dataset.

6.6 Discussion

In this chapter, I have shown that guided attention enables a model to focus on a specific ROI, and that this can also lead to improved prediction performance

compared to a vanilla ResNet (5 and 10% MAE improvement for 2D and 3D images respectively). These results indicate that a vanilla ResNet does not optimally use the available information present in an image. This might be explained by the fact that multiple regions in the fetal brain can, to some degree, be age-discriminative, and that the vanilla network is not always able to identify the most discriminative ones. By *guiding* the network in early training to recognise the CoP (or CB), the network is forced to focus on that area, resulting in improved age prediction performance later in training. Furthermore, by constraining the model to focus on specific parts of the brain, more detailed conclusions can be drawn about the development of individual structures.

The models presented in this chapter were developed with US images of healthy fetuses, with the motivation of learning the developmental patterns of an optimally healthy population. In future work, the developed models could be applied to assess structural brain development in diseased cohorts by comparing the predicted to the true GA, also referred to as the *brain age gap* [10, 212]. As the guided attention is able to attend to specific regions of interest, *local* brain ages can be obtained, providing potential novel insight into fetal brain development in new cohorts.

In the proposed *guided* attention approach segmentation masks of the ROI are required during training, but, importantly, not during inference. This makes the *guided* attention more practical than a two-step approach where first the ROI is detected and segmented, followed by age prediction based on this region. Furthermore, in this chapter, I have shown that using masked volumes (with the ROI) results in considerably lower performance than using the *guided* attention. This suggests that the network needs additional image features to achieve optimal performance. For example, when using the CoP as ROI, an important feature could be the distance to the skull.

While the attention masks can be regularised to focus on specific brain regions, the large receptive field of the convolutions, especially at deeper layers, makes it impossible to ensure the exclusion of all other image features. Care should therefore be taken when interpreting the final results. However, it has been shown that

the contribution of image features on a value in a CNN (the *effective* receptive field) is typically Gaussian distributed, and covers only a fraction of the theoretical receptive field [213]. Furthermore, convolutional layers usually preserve spatial relationships, as illustrated by their ability to perform weakly supervised object detection [214] (i.e. with only whole image labels). This property is, for example, also used in CAM (as introduced in Section 3.1.2).

While obtaining segmentation masks for training can be a bottleneck, the masks are not required to be pixel-level accurate. Rather, they should contain the approximate region that the network should focus on. For this reason, the required segmentation masks could be obtained with automated segmentation models such as presented in Chapter 5. Furthermore, future work could explore training the AGs semi-supervised, where segmentation masks are only available for a subset of the data. As shown in Chapter 5, a small set of annotated image volumes can be sufficient to train a segmentation model for subcortical segmentation. For this reason, I expect a small subset of masks to be sufficient to *guide* the AG, thereby relaxing the need for a fully annotated dataset.

In this chapter, I showed that *local* brain age predictions can be obtained using guided attention, significantly improving the prediction performance. Further, I demonstrated that guided attention also improved the prediction performance on an independent test set (MAE of 3.73 days, compared to a MAE of 4.21 for the ResNet baseline). This demonstrates that the developed models can be used to analyse independent datasets, enabling the comparison of local brain age in cohorts with specific conditions in further research.

7

Interpretable Brain Age Prediction using Prototype Learning

Contents

7.1 Introduction 139
7.2 Optimal Transport 141
7.2.1 Assignment Problem 142
7.2.2 Monge Formulation 143
7.2.3 Kantorovich Relaxation 145
7.2.4 Entropic Regularisation 145
7.2.5 Optimal Transport in CNNs 146
7.3 Methods 147
7.3.1 Network Architecture 148
7.3.2 Prototype Projection 148
7.3.3 Distance Metric Loss 149
7.3.4 Patch-based Distance Metric 150
7.3.5 Consistency Loss 152
7.3.6 Inference 152
7.4 Experimental Set-Up 153
7.4.1 Datasets 153
7.4.2 Implementation 155
7.5 Results 156
7.5.1 Fetal Ultrasound 156
7.5.2 Adult MRI 162
7.6 Discussion 164

Similarly to Chapter 6, this chapter uses brain age predictions to quantify the

stage of development of the fetal brain. However, while the last chapter addressed *which* image features were used for the prediction, this chapter addresses *how* these image features are used to make the prediction. It introduces a novel inherently interpretable model architecture of which the reasoning process can be visualised, providing an understanding of the prediction. The work presented in this chapter has been written into a paper, of which a preprint is available online:

> **Linde S. Hesse**, Nicola K. Dinsdale and Ana I. L. Namburete. "Prototype Learning for Explainable Regression," available on ArXiv: arXiv:2306.09858

7.1 Introduction

Deep learning models are typically considered to be black boxes, meaning that they produce output predictions without providing insights into the decision-making process. This severely limits the adoption of such methods in clinical practice, as the decision-making process needs to be transparent to understand model behaviour and gain patients' trust [99]. It is therefore vital to develop models that are interpretable and hence capable of providing insight into their reasoning process [100].

The most frequently used methods to explain a model's prediction are attribution-based, which explain the prediction of a trained model *post-hoc* [101]. However, as described in Chapter 3, these explanations are not always a faithful representation of the original prediction and can, for example, resemble edge maps rather than being dependent on the trained model [116]. Furthermore, they can only show *which* features were used for the prediction, and not *how* these features were used to generate the prediction.

A promising example-based interpretable model for the classification of natural images was proposed by Chen et al. [143]. That method was able to show how the image features were used to generate the prediction. Their model architecture (*ProtoPNet*) learned a set of embeddings in latent space, referred to as *prototypes*, and used the distances to each of these prototypes to classify a new sample. Each prototype was assigned a class label, and could thus be considered as a representative

example (in latent space) for that class. The prototypes in the final model were enforced to equate representations of actual training images, which made it possible to visualise the prototypes. In contrast to *post-hoc* methods, this type of architecture is thus **inherently interpretable**, as the final prediction is directly generated from the prototype distances.

However, many important medical image tasks are continuous regression problems, such as brain age prediction [10, 26]. As *ProtoPNet* uses the categorical labels to pull (or push) samples together (or apart), this architecture cannot be directly applied for regression. Furthermore, the prototypes in *ProtoPNet* represent only *image parts*, with the intuition of discovering whether the pattern in a prototype is present in any of the image patches. This provides patch-level detail in the prediction but is less suitable for continuous regression problems where more gradual structural changes have to be detected.

In this work, I propose *PRINT*: a Prototype-based model for Regression using an Interpretable neural Network and optimal Transport. Motivated by the similarities to metric learning, the images are mapped to an inherently continuous representation space in which the distances between images and prototypes in *latent* space are proportionate to their differences in *label* space. The prediction for a new sample is subsequently made from a weighted average of prototype labels within a given distance. As all prototypes can be visualised, this provides an intuitive understanding of the prediction for a regression task.

In contrast to *ProtoPNet*, the prototypes in *PRINT* are latent representations of whole training images. To also incorporate spatial detail into the model's decision-making process, the image-level distances are decomposed into patch-wise similarities between image patches. The patch-wise similarities are computed in latent space and then structurally matched using optimal transport (OT). The OT finds an optimal matching matrix that contains the *soft assignment* scores between the image patches of the sample and prototype. This matrix is then used to compute a single image-level distance. The resulting matches can be inspected to verify

whether anatomically corresponding patches are matched together between an image and prototype, thus providing a detailed decomposition of the image-level distance.

The network is trained using a combination of two losses: A *metric loss*, which enforces the prototype-image distances to be proportionate to their label differences, and a *consistency loss* that encourages the distances between identical image patches under an anatomically justified transformation to be small. The network and all loss elements are fully differentiable and, thus, the network can be trained end-to-end.

In this chapter, I demonstrate the proposed model on the task of brain age prediction for both fetal US images and adult MR images.

In summary, the key contributions of this chapter are:

- I propose a novel interpretable model for continuous regression and demonstrate that it obtains competitive performance for brain age prediction on two datasets: fetal US and adult MRI.
- I demonstrate how OT can be incorporated into a prototype-based model, providing patch-level matching scores while computing a single image-level distance.
- I introduce a geometric consistency loss which improves the model performance and leads to more anatomically correct matching between patches of the prototype and sample image.

7.2 Optimal Transport

In the proposed architecture, OT is used to match the image patches of the sample image and prototype with each other. This section will introduce the concept of OT and describe how it can be efficiently solved.

OT aims to find the matching matrix, also referred to as *optimal flow* or coupling matrix, that results in the minimal distance (or *lowest cost*) between two distributions. In this section, I will focus on discrete distributions, but it can also be applied to continuous distributions. In literature, OT is often explained as

'moving one pile of sand (*distribution 1*), into another pile of sand (*distribution 2*), in the most efficient way'. For this reason, it can also be referred to as the Earth Movers Distance [215]. However, moving piles of sand is still a rather abstract concept, and does not give a very intuitive understanding of the problem. In this section, I will explain OT from its simplest formulation: the assignment problem, and generalise it from there to the broader concept.

7.2.1 Assignment Problem

The assignment problem assigns a number of n tasks to m workers in the most optimal way, given a cost matrix. In this example, the cost matrix could denote how much time each person takes for a certain task, and the objective to minimize the total time needed to complete all tasks by the workers. The formulation as *tasks* and *workers* aims to provide an intuitive understanding of the problem, but these can in practice represent any concepts. Given a set of $m = 4$ workers, $[w_1, w_2, w_3, w_4]$, a set of $n = 4$ tasks, $[t_1, t_2, t_3, t_4]$, and a cost matrix, $\mathbf{C} \in \mathbb{R}^{m \times n}$, the objective is to find the assignment matrix, $\mathbf{Q} \in \mathbb{R}^{m \times n}$, that assigns each worker to a task, resulting in the lowest total cost (c_{tot}). The total cost can be defined by:

$$c_{\text{tot}} = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} q_{ij} \tag{7.1}$$

where c_{ij} and q_{ij} are the elements of $\mathbf{C}$ and $\mathbf{Q}$, respectively. This cost is minimised with respect to $\mathbf{Q}$, resulting in the following objective:

$$\min_{\mathbf{Q}} \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} q_{ij} \tag{7.2}$$

An additional constraint in this optimisation problem is that all tasks need to be assigned to a worker, and all workers need to be assigned a task. Specifically, each row and column of the assignment matrix, $\mathbf{Q}$, needs to sum up to 1. This can formally be written as:

$$\sum_{i=1}^{n} q_{ij} = 1 \forall j \in m, \quad \sum_{j=1}^{m} q_{ij} = 1 \forall i \in n \tag{7.3}$$

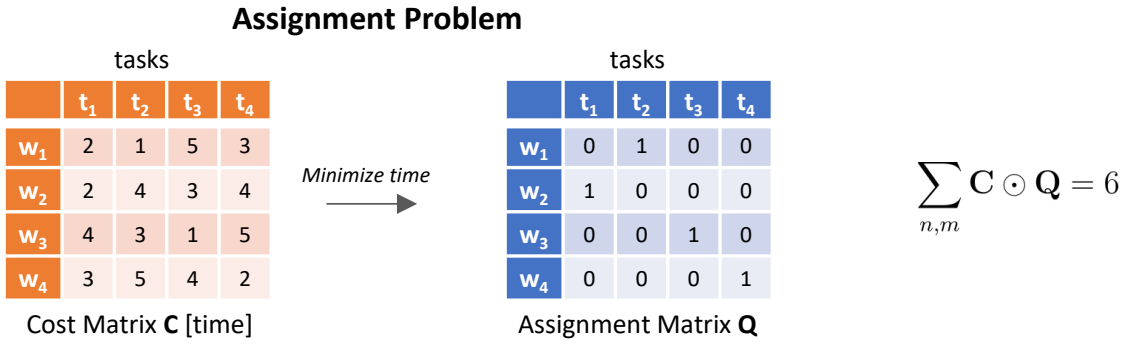


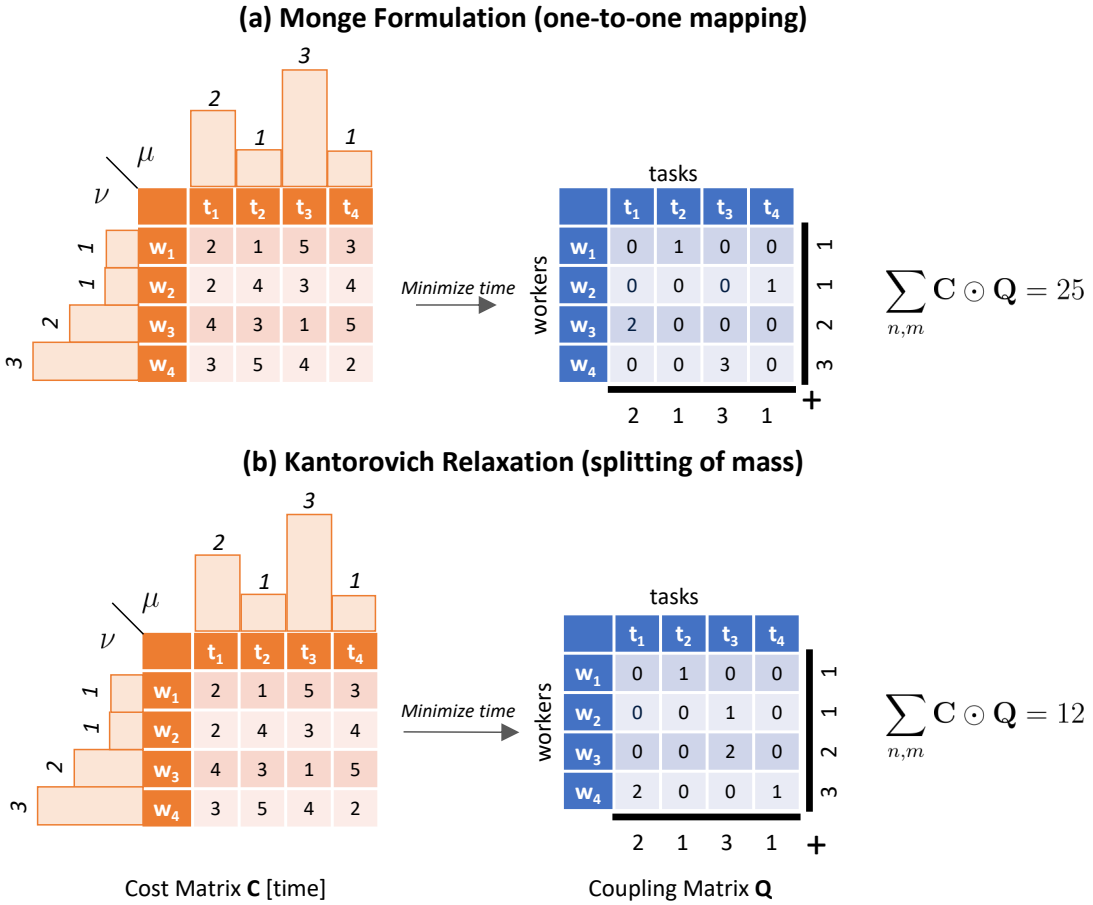
Figure 7.1: Schematic overview of the assignment problem. Given a set of $m = 4$ workers, $[w_1, w_2, w_3, w_4]$, and a set of $n = 4$ tasks, $[t_1, t_2, t_3, t_4]$, the aim is to assign tasks to workers in the most optimal way. This is defined as the assignment with the lowest cost (i.e. time). The problem is constrained in that each worker needs to be assigned one task, and each task one worker (i.e. the rows and columns of $\mathbf{Q}$ need to sum up to 1).

An example assignment problem is shown in Fig. 7.1, with the cost matrix, $\mathbf{C}$, shown in orange and the optimal assignment matrix, $\mathbf{Q}$, in blue. In the assignment problem setting the entries of the assignment matrix can only be 1 or 0, with a 1 representing that that worker is assigned to the respective task. The assignment problem can be solved with brute force by computing the cost for all possibilities, but, more efficient algorithms exist to solve this problem, such as the Hungarian method [216].

In order to satisfy both constraints given in Eq. 7.3, the number of workers needs to be equal to the number of tasks ($m = n$), which is referred to as a *balanced* assignment problem. If this is not the case, dummy columns or rows can be added to transform the problem into a balanced one.

7.2.2 Monge Formulation

In the assignment problem, individual tasks and workers need to be assigned to each other. However, consider now the scenario where the set of tasks represents task *types*, and there can be more than one item of a single type. For example, given the list of task types as $[t_1, t_2, t_3, t_4]$, we now also have a vector representing the number of tasks of each type: $\mu = [2, 1, 3, 1]$. This can be pictured as a frequency histogram of tasks or a *distribution* of tasks. For the workers, we can make the same



$$\sum_{n,m} \mathbf{C} \odot \mathbf{Q} = 25$$

$$\sum_{n,m} \mathbf{C} \odot \mathbf{Q} = 12$$

Figure 7.2: Schematic overview of the Monge-Kantorovich problem. Given a set of workers and a set of tasks, the aim is to assign tasks to workers in the most optimal way. This is defined as the assignment with the lowest cost (i.e. time). The problem is constrained in that each worker needs to be assigned one task, and each task one worker. (a) shows the Monge formulation, in which a one-to-one mapping is required, and (b) the Kantorovich relaxation where a worker type can be assigned to multiple tasks.

assumption and denote these counts by: $\nu = [3, 2, 1, 1]$. This is schematically shown in Fig. 7.2, where the counts are drawn as histograms around the cost matrix.

For this formulation, the constraints in Eq. 7.3 have to be adapted so that the rows and columns now sum up to the respective counts (instead of to 1) This can be written as:

$$\sum_{i=1}^{m} q_{ij} = \mu_j \forall j \in m, \quad \sum_{j=1}^{n} q_{ij} = \nu_i \forall i \in n \tag{7.4}$$

This formulation is referred to as the Monge formulation [217]. In this formulation, like in the assignment problem, the objective is to find a one-to-one mapping

between the two distributions. In our example, this means that all workers of a certain type (i.e. all three w_3) must be assigned to the same task. It can be seen that there are only two possible solutions that satisfy this constraint for the cost matrix shown in Fig. 7.2, as w_4 must be assigned to t_3, and, likewise, w_3 to t_1.

In order for a solution to exist, the initial distributions, ν and μ, need to match up with each other (i.e. be a permutation of each other in the discrete case). The resulting matrix $\mathbf{Q}$ in this scenario is usually no longer referred to as the *assignment* matrix but as the *coupling* or *optimal flow* matrix. In this chapter, I will also refer to this matrix as the *matching* matrix, as it matches two distributions with each other.

7.2.3 Kantorovich Relaxation

The Monge formulation (Section 7.2.2) only allows a one-to-one mapping, which makes it ill-defined for non-matching distributions. For this reason, the Kantorovich relaxation was proposed [218], which states that the mass at a certain point (i.e. a number of workers) can be split across targets (i.e. the tasks). This means that there is no longer a one-to-one mapping required between the distributions. However, the two distributions are still required to have the same area under the curve (i.e. the sum of all counts in the histogram). In many cases, probability distributions are used as initial distributions, which always sum up to 1.

The optimal solution for the example under the Kantorovich relaxation is shown in Fig. 7.2b. It can be seen that the workers w_3 are now split across tasks t_1 and t_2, and that the optimal solution can achieve a lower total cost than under the Monge formulation.

This complete formulation is often referred to as the Monge-Kantorovich problem, and in this thesis, I will refer to this as simply the OT problem.

7.2.4 Entropic Regularisation

The OT problem is computationally expensive to solve, and therefore an *entropic regularisation* term was added to smooth the optimisation problem [219]. En-

tropic regularisation transforms the optimisation objective in Eq. 7.1 into the following objective:

$$\min_{\mathbf{Q}} \sum_{i=1}^{m} \sum_{j=1}^{m} c_{ij} q_{ij} + \epsilon \mathrm{H}(\mathbf{Q}) \tag{7.5}$$

where $\mathrm{H}(\mathbf{Q})$ denotes the entropy function, given by:

$$\mathrm{H}(\mathbf{Q}) = -\sum_{i=1}^{m} \sum_{j=1}^{n} q_{ij} \log(q_{ij}) \tag{7.6}$$

and ϵ the strength of the entropy regularisation. The advantage of this formulation is that it can be rewritten as follows[1]:

$$\min_{\mathbf{Q}} \sum_{i=1}^{m} \sum_{j=1}^{m} c_{ij} q_{ij} + \epsilon \mathrm{H}(\mathbf{Q}) = \min_{Q} (D_{\mathrm{KL}}(\mathbf{Q} \parallel \mathbf{K})), \text{with } \mathbf{K} = e^{-\mathbf{C}/\epsilon} \tag{7.7}$$

with $\mathbf{K}$ a kernel function of the cost matrix ($\mathbf{C}$), k_{ij} an element of $\mathbf{K}$, and D_{KL} the Kullback-Leibner divergence between $\mathbf{Q}$ and $\mathbf{K}$ [220].

The objective can now be considered as minimising the KL divergence between $\mathbf{K}$ and $\mathbf{Q}$, under the constraints given in Eq. 7.4. This can efficiently be solved using the Sinkhorn algorithm that initialises $\mathbf{Q}$ with $\mathbf{K}$, and then iteratively projects the matrix on the individual constraints, converging into the solution that satisfies both constraints [221]. This is shown in Fig. 7.3, where both constraints are shown as coloured semicircles in a simplified schematic decision space. The objective is to find the closest $\mathbf{Q}$ to $\mathbf{K}$ that satisfies both constraints, which can be found with Sinkhorn iterations [221].

7.2.5 Optimal Transport in CNNs

By formulating the OT optimisation problem with entropic regularisation, the optimisation is provably convex and fully differentiable, and can thus be incorporated into deep learning architectures [219, 222]. Most comparable to the work presented in this chapter are previous studies that used OT to compute patch-based distances between pairs of natural images [169–171]. In those studies, each image was

[1] the full derivation can be found in Appendix C

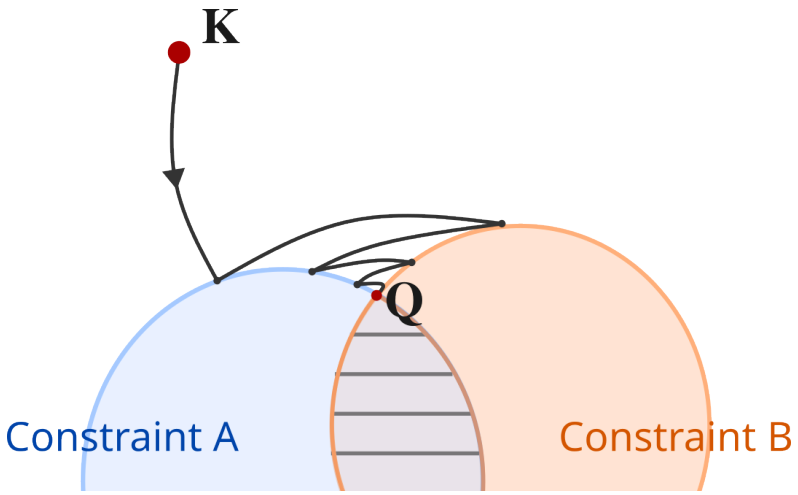


Figure 7.3: Schematic of the decision space of Sinkhorn optimisation solving the OT problem. The solution is initialised with **K** and then iteratively projected on the constraints A and B to arrive at the optimal solution. The coloured semicircles represent the constraints introduced in Eq. 7.4.

considered as a distribution of patches, and the latent distances between the patches of both images were computed as the cost matrix. OT was then applied to match the patches of both images with each other, resulting in the matching matrix. Both the cost matrix and matching matrix were then used to compute a single image-level distance (i.e. as described by Eq. 7.1). Both Zhao et al. [170] and Phan and Nguyen [171] used these patch-based distances to improve the distance computations in latent space between sets of images without retraining any network weights. This resulted in improved performance for metric learning [170] and face identification [171], and also added interpretability as the patch-based distance computations could be analysed in detail. On the other hand, Zhang et al. [169] used OT in an end-to-end trainable architecture and applied it for few-shot learning, also resulting in improved performance and interpretability.

7.3 Methods

The proposed *PRINT* model aims to learn a set of prototypes in latent space, each representing a whole image from the training set. Each prototype has a continuous label (e.g. age), and sample predictions are made using a weighted mean of distances to these prototypes. To incorporate patch-level detail, the image-level distances

between the image and prototypes are decomposed into patch-level distances and structurally matched using OT matching [169–171]. A schematic overview of the *PRINT* model architecture is shown in Fig. 7.4.

7.3.1 Network Architecture

Given a 2D image $\mathbf{X} \in \mathbb{R}^{w \times h \times 1}$, a feature extractor f is trained that can extract a latent representation of $\mathbf{X}$, denoted by $\mathbf{Z} \in \mathbb{R}^{w_z \times h_z \times c_z}$, with w_z and h_z the spatial dimensions and c_z the channel dimension. The feature extractor f is composed of a base encoder (g), and an additional block (h) with two convolutional layers with 1×1 kernels and a sigmoid as the last activation. Simultaneously, a set of *prototypes* is learned, which are essentially learned embeddings in latent space, each of the same size as $\mathbf{Z}$: $\mathcal{P} = \{\mathbf{P_k} \in \mathbb{R}^{w_z \times h_z \times c_z}, \forall k \in [1, n_p]\}$, with n_p as the number of prototypes. Each of the prototypes also has an assigned label, resulting in a vector of prototype labels, denoted by $\boldsymbol{y^{proto}}$. These prototype labels are fixed at the beginning of training and are not optimised. On the other hand, the prototypes themselves are considered model parameters, and can therefore be trained end-to-end with the feature extractor. For each sample, the network computes the distances to each of the prototypes in latent space, resulting in a vector of prototype distances of length n_p, denoted by $\boldsymbol{d}$.

7.3.2 Prototype Projection

During training, the prototypes are updated with each step and can therefore be located throughout the latent space. However, as proposed in [143], every N epochs these are replaced by the closest latent representation of an image in the training dataset, which is referred to as the *prototype projection*. The model checkpoints are saved only straight after the projection, ensuring that in the saved model each prototype can be visualised using the corresponding image from the training set.

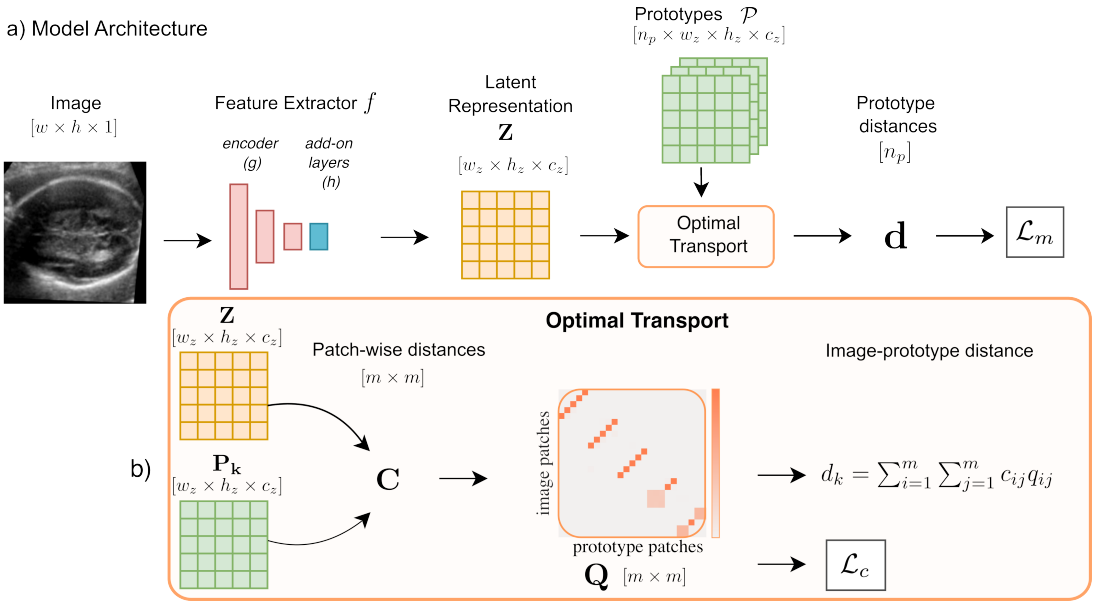


Figure 7.4: Detailed overview of the (a)*PRINT* architecture and (b) patch-based matching with OT. The OT matching is shown for one sample and prototype, and is repeated for each prototype to obtain $\boldsymbol{d}$. The m is the total number of patches per sample or prototype, given by $w_z h_z$.

7.3.3 Distance Metric Loss

The distances between samples and prototypes need to be regularised, so that for a certain sample with ground-truth label y, the distance in latent space to a prototype k is related to its difference in label space: $d_k \propto |y_k^{proto} - y|$, with d_k and y_k^{proto} as the distance to and label of the kth prototype (i.e. the kth element of $\boldsymbol{d}$ and $\boldsymbol{y^{proto}}$), respectively.

However, as the latent space exists on a high-dimensional manifold, computing the feature distances on the manifold between prototype and sample, the geodesic distance, is non-trivial. In this work, the geodesic distance in the local neighbourhood of a sample representation is approximated using the Euclidean distance, which is a good approximation for small distances on the manifold.

The proposed network is trained using an adapted version of a loss proposed by Chao et al. [168]. Instead of regularising distances over the samples in the batch [168], the distances between samples in the batch and each of the prototypes in the *local neighbourhood* are regularised. The neighbourhood is determined by differences

in labels between prototypes and samples, as opposed to the feature distance. The total loss for a single sample with ground-truth label y is thus given by:

$$\mathcal{L}_m(\boldsymbol{d}, \boldsymbol{y^{proto}}, y) = \sum_{k=1}^{n_p}(|s \cdot d_k - (|y_k^{proto} - y|)|)w_k^{train} \tag{7.8}$$

where s is a learnable parameter scaling the feature distances to the label differences and w_k^{train} the weight of the kth prototype. w_k^{train} weights the prototype with a Gaussian function based on the label differences, and is defined by:

$$w_k^{train} = e^{-\frac{|y_k^{proto} - y|}{2\sigma^2}} + \alpha \tag{7.9}$$

where σ controls the size of the neighbourhood (i.e. the standard deviation of the Gaussian kernel), and α is a small number to prevent the latent embeddings from tangling. Without it, samples and prototypes with a large label difference have a negligible effect on each other and could therefore embed close together in feature space despite being dissimilar.

7.3.4 Patch-based Distance Metric

In order to train the network with the distance metric loss, distances need to be computed between the sample and each of the prototypes, both of size $h_z \times w_z \times c_z$. A common approach is to use an average pooling operator that compresses the spatial dimensions, resulting in a vector of size c_z. The Euclidean distance can then be computed between these vectors [164, 168] but this discards all spatial information. To provide information about the spatial makeup of the distance between a prototype and a sample, I use a patch-based distance metric based on OT matching instead.

The latent representation $\mathbf{Z}$ and a single prototype $\mathbf{P_k}$ can both be considered as sets of m feature vectors: $[\boldsymbol{z_1}, \boldsymbol{z_2}, ..., \boldsymbol{z_m}]$ and $[\boldsymbol{p_1}, \boldsymbol{p_2}, ..., \boldsymbol{p_m}]$, with $m = h_z w_z$, and each vector of size c_z. The proposed distance metric computes the Euclidean distances between both sets of feature vectors, $d(\boldsymbol{z_i}, \boldsymbol{p_j}) \forall i, j \in \{1, ..., m\}$, resulting in a cost matrix $\mathbf{C} \in \mathbb{R}^{m \times m}$. As not all pairwise distances should contribute equally to the image-level distance (i.e. the distance between a skull patch and patch containing the ventricles is not meaningful), I use OT to obtain a matching matrix

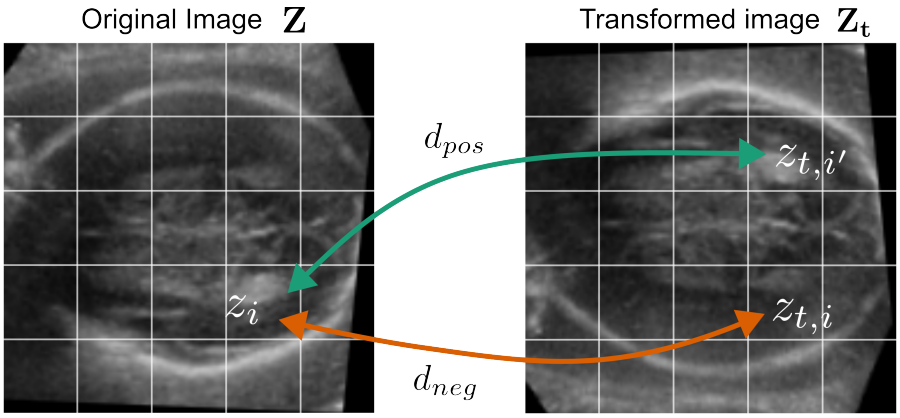


Figure 7.5: Overview of the used consistency loss between a latent image representation ($\mathbf{Z}$) and the representation of a transformed image ($\mathbf{Z_t}$) for a single triplet. The grid indicates the size of the latent pixels. The objective is to encourage the d_{pos} to be larger than d_{neg} (by a certain margin γ).

between the patches of $\mathbf{Z}$ and $\mathbf{P_k}$, $\mathbf{Q} \in \mathbb{R}^{m \times m}$ (as described in detail in Section 7.2.5). This matrix $\mathbf{Q}$ can be considered as a *soft assignment* matrix, in which each feature vector $\boldsymbol{z_i}$ can be partly matched with more than one feature vector $\boldsymbol{p_j}$. These matching and cost matrices can subsequently be used to compute the image-level distance between the sample and the kth prototype as follows (in line with Eq. 7.1):

$$d_k = \sum_{i=1}^{m} \sum_{j=1}^{m} c_{ij} q_{ij} \tag{7.10}$$

where c_{ij} and q_{ij} are the elements of $\mathbf{C}$ and $\mathbf{Q}$, respectively. Intuitively, OT thus assigns image patches to each other, and then only uses distances between assigned patches to compute the final image-level distance.

The OT optimisation is constrained by the initial distributions (i.e. the number of task types or worker types in the example given in Section 7.2.5). These distributions can be considered as importance scores for each of the image patches, and are in this work set to uniform distributions for both the image and prototype patches. The OT problem is formulated with entropic regularisation so that it can be solved using Sinkhorn iterations [219].

7.3.5 Consistency Loss

The OT optimisation aims to find structural matches between the image and prototype patches. To achieve anatomically correct matching, I also introduced a consistency loss that acts on the distances between feature vectors in the latent representation of an image $\boldsymbol{z_i} \in \mathbf{Z} = f(\mathbf{X})$ and the feature vectors of the same image under an anatomically justified transformation $T : \boldsymbol{z_{t,i}} \in \mathbf{Z_t} = f(T(\mathbf{X}))$. The consistency loss encourages distances of the same image content, $d(\boldsymbol{z_i}, \boldsymbol{z_{t,i'}})$ with $i' = T(i)$, to be small, while encouraging distances between patches of the same spatial location, $d(\boldsymbol{z_i}, \boldsymbol{z_{t,i}})$, to be large. This resembles a contrastive learning problem where certain feature vectors are pulled together (positive pairs) whereas others are pushed apart (negative pairs). Therefore, the contrastive loss is formulated as a triplet loss, which encourages the distance between an anchor and a positive vector to be larger than the distance between an anchor and a negative vector by a certain margin γ [160]. To create triplets, all vectors of $\boldsymbol{z_i}$ were used as anchors, and the respective vectors in $\boldsymbol{z_t}$ as positive ($\boldsymbol{z_{t,i'}}$) and negative ($\boldsymbol{z_{t,i}}$) samples, excluding any triplets for which the positive and negative sample were the same. The Euclidean distance was then computed from each anchor to the positive sample, d_{pos}, and the negative sample, d_{neg}. The total consistency loss per image computes the sum over all triplets, and can be described by:

$$\mathcal{L}_c = \sum_{i=1}^{m} \max(d(\boldsymbol{z_i}, \boldsymbol{z_{t,i'}}) - d(\boldsymbol{z_i}, \boldsymbol{z_{t,i}}) + \gamma, 0)[\boldsymbol{z_{t,i'}} \neq \boldsymbol{z_{t,i}}] \tag{7.11}$$

For training, I used a combination of the distance metric loss (Eq. 7.8) and consistency loss as $\mathcal{L}_{total} = \mathcal{L}_m + \beta\mathcal{L}_c$, where β is the weight of the consistency loss.

7.3.6 Inference

At inference, a prediction for a new sample is made using a weighted average of the prototype labels within a certain radius r, given by:

$$\hat{y} = \frac{\sum_{k=1}^{n_p} w_k^{test} y_k^{proto}}{\sum_{k=1}^{n_p} w_k^{test}}, \quad w_k^{test} = \begin{cases} e^{-\frac{s \cdot d_k}{2(r/3)^2}}, & \text{if } s \cdot d_k \leq r \\ 0, & \text{otherwise} \end{cases} \tag{7.12}$$

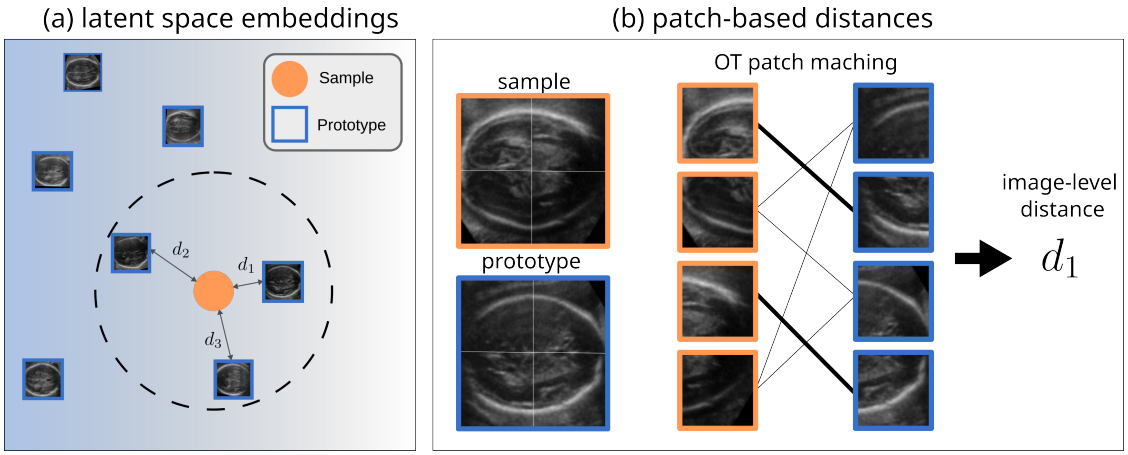


Figure 7.6: Schematic overview of model inference of *PRINT*. (a) The prediction for a new sample is made using a weighted sum of prototype labels within a certain radius (shown by the dashed circle around the sample). All prototypes represent real images from the training set and can thus be visualised. (b) Each image-level distance in latent space (i.e. d_1) is based on patch-level matchings between the patches of the image and prototype, which results in a single image-level distance. Patches can be partly matched with each other, which is indicated by thinner lines connecting them. For visualisation, only four patches per image are shown (i.e. $h_z = 2$, and $w_z = 2$), but this will be a higher number in practice.

in which the weight of each prototype is thus determined by a Gaussian with standard deviation $r/3$. This is schematically shown in Fig. 7.6a, where it can be seen that the prototypes (latent representations of training images) within a certain radius in latent space contribute to the model prediction for a sample. Fig. 7.6b shows how the image-level distances can be decomposed into patch-level matches between the image and prototype.

7.4 Experimental Set-Up

7.4.1 Datasets

7.4.1.1 Fetal Ultrasound

In this chapter, all 4265 2D images from the extended subset of the INTERGROWTH-21st dataset were used. As described in Chapter 4, the 3D volumes used were aligned to the same coordinate system and scaled to the average brain size at 30 GW using an automated alignment method [178]. After alignment, the 2D transventricular plane was extracted from each of the 3D volumes using fixed

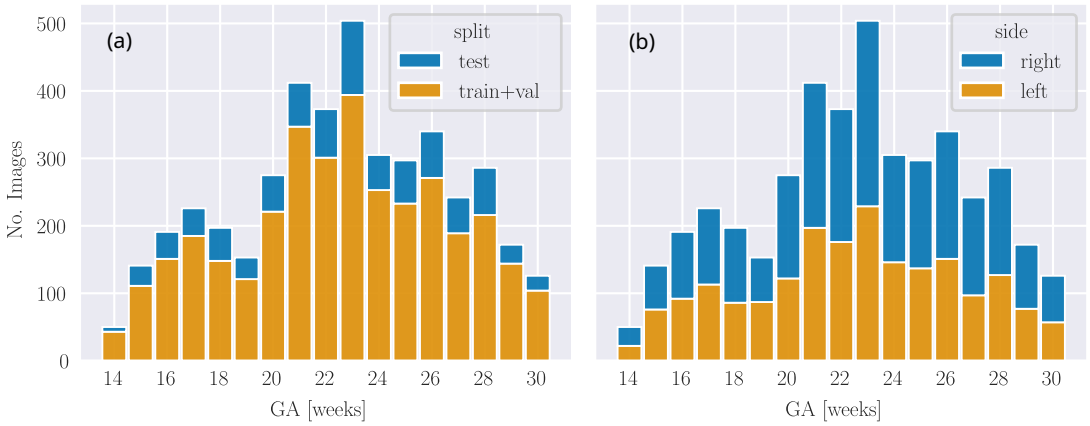


Figure 7.7: Overview of the INTERGROWTH-21st data across GA used in this chapter, (a) shows the distribution for the train and test split, and (b) for images with a visible left and right hemisphere.

coordinates, resulting in a total set of 4265 2D images of size 160×160 with an isotropic pixel size of 0.6 mm. For the experiments, this set was split into test (20%) and train/validation (80%). An overview of the data distribution used in this chapter across gestation is shown in Fig. 7.7.

7.4.1.2 Adult MRI

In this chapter, I also demonstrate the proposed approach on T1 MR images of the adult brain from the IXI dataset[2], which contains MR images from healthy subjects between the age of 19 and 86 years. All 3D images were linearly registered to the same template orientation, as well as skull stripped (as described in Section 4). From each 3D image volume an axial plane was extracted, which was selected to contain the ventricles, as the increase in ventricle size with ageing and the corresponding brain atrophy is well established [223]. This resulted in a total of 561 MR images, each of size 160×192, which was also split into test (20%) and train/validation set (80%). An overview of the LUMC data used in this study is shown in Fig. 7.8.

[2] `https://brain-development.org/ixi-dataset/`

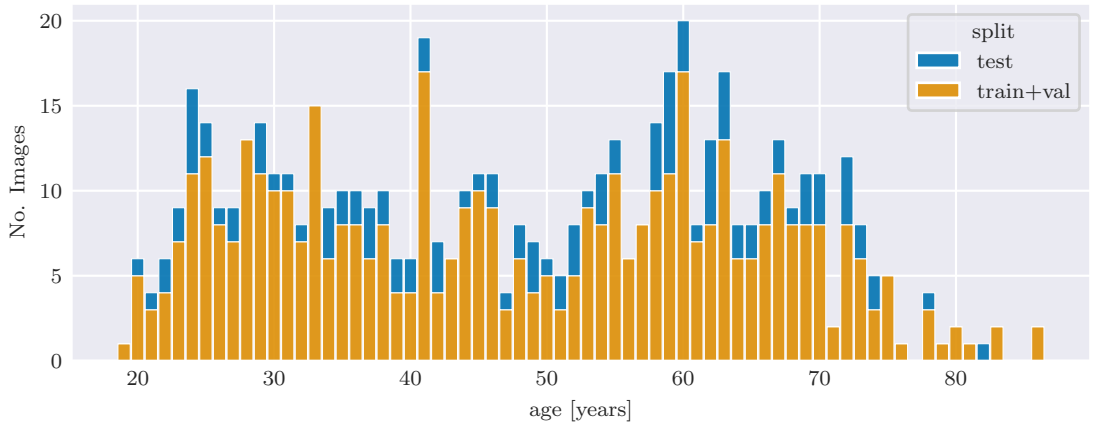


Figure 7.8: Overview of the MRI images from the IXI dataset used in this chapter. The distribution across the chronological age is shown, with the train and test split indicated.

7.4.2 Implementation

I implemented the proposed network architecture with Pytorch Lightning, using PyTorch 1.13 and Python 3.10. All experiments were performed on a single Nvidia A10 with 24 GB RAM. The OT optimisation was performed with an entropic regularisation strength of 0.1 (ϵ in Eq. 7.5), and it was optimised using Sinkhorn iterations implemented in logarithmic space to prevent numerical instabilities. A ResNet-18 was used as a base encoder, g, pre-trained on ImageNet.

Five-fold cross-validation was used to tune all hyperparameters, of which an overview is given in Section 7.5.1. The reported test set performance was then obtained by applying the best fold (on the validation set) to the test set. All experiments were run for 300 epochs for the US data and 150 for the MRI, with a learning rate of 5×10^{-4} and 1×10^{-4}, respectively. The prototypes were projected to the closest sample in the training set every 25 epochs. Unless otherwise reported, for both datasets α and σ in Eq. 7.9 were set to 0.05 and 1, respectively, and r in Eq. 7.12 to 3; the number of channels in the additional layer block, h, and prototypes, c_z, to 512, and, the number of prototypes, n_p, to 100. The latent space dimensions w_z and h_z were both 5 for the US dataset, and 5 and 6, respectively, for the MRI dataset. Prototype labels were assigned uniformly at the beginning of training within the label range present in the training dataset.

Due to shadowing artefacts of the fetal skull, usually only one of the two brain hemispheres is visible in fetal brain US images [1]. As it is desired to match the two visible hemispheres with each other, the consistency loss for this dataset was implemented with horizontal flips along the midline of the brain as a geometric transformation (see Fig. 7.5). The weight of the consistency loss, β, was set to 10, and the margin, γ, to 0.1. For the adult MR images, however, both hemispheres are visible and so a consistency loss is not required. This was confirmed by cross-validation, and therefore this loss element was set to 0. During inference, only prototypes within a certain radius r of a sample are considered when generating the final prediction. Therefore, it can occur that for a certain sample, no prototypes are within this area, resulting in an invalid prediction. To report also a performance value for these samples, I used nearest-neighbour predictions and assigned the label of the closest prototype to these samples.

7.5 Results

7.5.1 Fetal Ultrasound

Baseline Comparisons The MAE prediction performance of the fetal US dataset is shown in Table 7.1. The proposed *PRINT* architecture has been compared to two baseline models: *ResNet Baseline* – a vanilla ResNet 18 model (pre-trained on ImageNet) trained with mean squared error loss on the brain age task, and *AvgPool Baseline* – a network with the same architecture as the proposed model but computing distances not with patch-based OT but by first pooling the feature representations into a single 1D vector and computing Euclidean distances between these. It can be seen that the proposed method outperforms the ResNet baseline by 0.11 days, showing that the common assumption that increased interpretability results in a decrease in prediction performance does not hold. However, the performance improvement did not pass a paired t-test ($p = 0.15$). The reported prediction performance of several previous studies is also shown [26, 86, 88], demonstrating that the prediction errors are in line with previously reported values

Table 7.1: Quantitative results for fetal brain age prediction from US on the test set. The values between brackets indicate the standard deviations. The ablations show the results without consistency loss ($\mathcal{L}_c$) and without add-on layers in the feature extractor (h). The GA range of each study is given in weeks. (1a-c) show age prediction performance reported in other studies, (2a-b) baselines implemented in this work, and (3) the proposed network architecture with ablation study (3a-c)

		Ablations		
	GA range	$\mathcal{L}_c$	h	MAE [days]
(1a) Namburete et al. (2017) [86]†	18-36	-	-	7.72 (6.01)
(1b) Wyburd et al. (2021) [26]†	19-31	-	-	4.1 (3.7)
(1c) Lee et al. (2023) [88]	13-41	-	-	4.5
	18-28	-	-	3.9
(2a) ResNet Baseline	14-30	-	-	4.04 (3.16)
(2b) AvgPool Baseline	14-30	-	-	4.09 (3.24)
(3a) PRINT (no $\mathcal{L}_c$, h)	14-30	×	×	4.16 (3.50)
(3b) PRINT (no $\mathcal{L}_c$)	14-30	×	✓	3.96 (3.19)
(3c) PRINT (no h)	14-30	✓	×	4.08 (3.53)
(3) PRINT	14-30	✓	✓	**3.93** (3.12)

† Study used 3D image volumes

for fetal brain age predictions. For a fair comparison, only studies that did not use size features were included in this table (as described in Section 2.2.2).

Predicted versus True GA In Fig. 7.9a the predicted GA versus the ground-truth GA is shown for the fetal US dataset. The samples have been fitted with a linear fit, which is close to the identity line (i.e. $x = y$), demonstrating good prediction performance throughout the whole GA range used.

Example Prediction Fig.7.10a shows an example prediction of the *PRINT* model for the US dataset. The prediction is composed of a weighted average of the labels of prototypes within a certain radius, and these neighbouring prototypes are shown next to the sample image. The weight of each prototype is inferred from its distance using a Gaussian function and shown on the right. The colour of each point indicates the prototype label, which is also shown with a coloured border around each image.

In addition to visualising the prototypes used to make a prediction, the computed

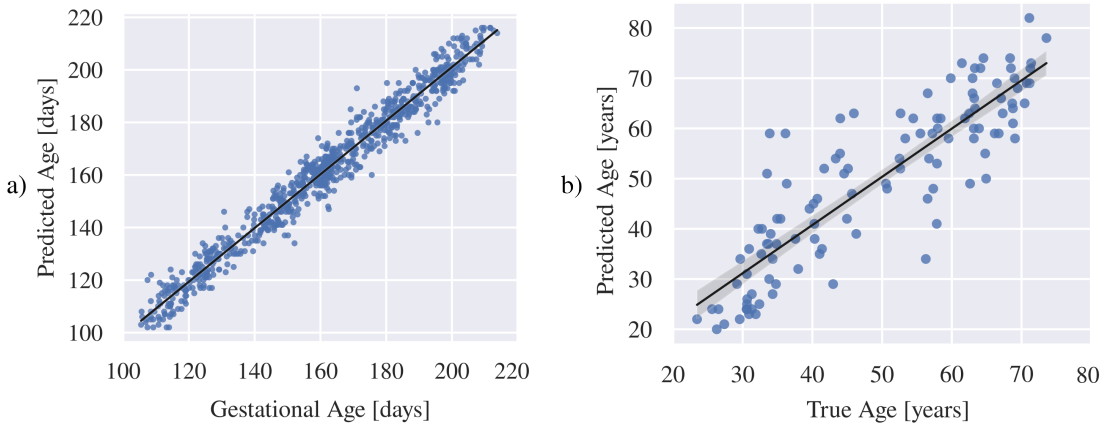


Figure 7.9: Predicted versus ground-truth age for (a) fetal US and (b) adult MRI. The fetal age estimation shows a high correlation between the true and predicted age, whereas the errors are larger for the adult age prediction.

distances can also be decomposed into patch-level matching matrices, which are shown in Fig. 7.11. For each sample-prototype pair, the matching (*soft assignment*) matrix is given in the last column of each panel, indicating which patches are most similar between the prototype and sample. The *reshuffled* prototype shows the OT matching more intuitively, in which at each image location the prototype patch with the highest matching score is shown. It is evident that the two visible hemispheres are correctly matched (Fig. 7.11.1a, b), illustrating the advantage of using OT matching in the network.

Effect of Consistency Loss As evidenced by the ablation study presented in Table 7.1, both the consistency loss and the add-on layers improve the prediction performance for the US dataset, demonstrating the need for both components in the model. The effect of the consistency loss on the US images is further illustrated in Fig. 7.11, showing sample-prototype pairs for (a and b) the full model trained with consistency loss and, (c) for the model without consistency loss. It can be seen that for the pair in Fig. 7.11a the image and prototype both have a visible right hemisphere (with the SF annotated), whereas in Fig. 7.11b and c, the visible hemisphere of the prototype is opposite to that of the image. The full model correctly identifies this and matches the patches in the visible hemisphere with

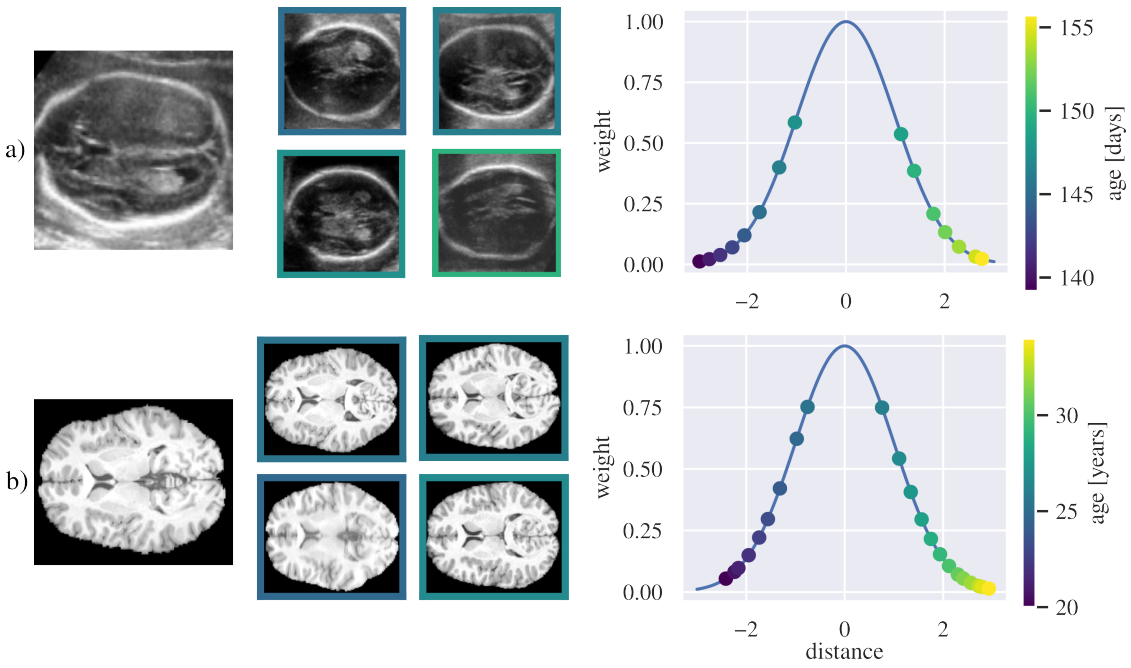


Figure 7.10: Example predictions for (a) US and (b) MRI. The image sample is shown on the left, and the four closest prototypes are shown in the middle. The right-most graphs indicate the weight of each of the prototypes in the final prediction. Prototypes with a label below the predicted value were plotted with negative distances for visualisation purposes, but unsigned distances were used in the model itself. For the US image, both the predicted and ground-truth age were 148 days and for the MRI the ground-truth age was 24 years, and the predicted age 25.6 years.

each other. This can also be seen in the *reshuffled* prototype, in which the sides are flipped. The same pattern of matching between hemispheres was found for all test set images indicating that the model correctly matches the patches of the visible hemispheres. Important to note here is that no annotations for the visible hemispheres were used during training, as the consistency loss is unsupervised.

On the other hand, in Fig. 7.11c it is evident that for the model trained without consistency loss, no matching occurs between the two visible hemispheres. Instead, the same spatial location in both the image and prototype are matched, as shown by the diagonal matching matrix. This thus illustrates that the consistency loss is responsible for matching the correct hemispheres. In this work only flipping across the midline of the brain has been used as a geometric transformation in the consistency loss. In future work, other transformations could be considered, based on the geometric variation present in the dataset.

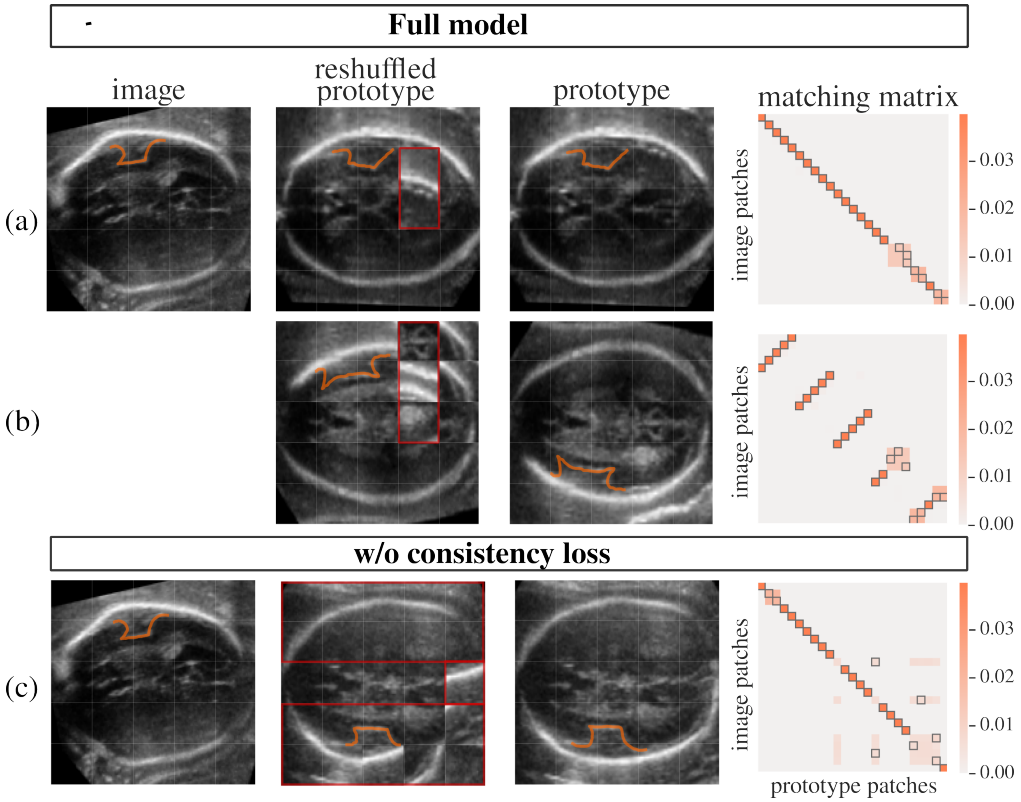


Figure 7.11: Illustration of OT matching between image and prototypes patches for the US dataset. The obtained OT matching matrix (**Q**) is given in the last column of each panel, with the most similar prototype patch to each image patch given a black border. For each image patch, the most similar prototype patch is shown at the respective image location in the *reshuffled* prototype. Patches that are incorrectly (according to the structural anatomy) matched are shown with red borders, and in all images, the SF is annotated in red to clearly indicate the visible hemisphere. For (a) and (b), the results are obtained from the full *PRINT* model, in which (a) illustrates the case where for both image and prototype the same hemisphere is visible, and (b) where a different hemisphere is visible. Row (c) is generated from the model trained without consistency loss.

Sensitivity to Hyperparameter Choice The average validation performance across the 5 folds for each of the important hyperparameters is shown in Fig. 7.12 for the US dataset. For the consistency weight (Fig. 7.12a), the top plot shows the patch-matching accuracy, which is computed from the overlap with approximated ground-truth matching matrices determined from the known visible hemisphere in each of the samples. It is evident that increasing the consistency weight results in an improvement of both the MAE and patch-matching accuracy up to a weight of 10 after which it slowly starts deteriorating again. It should be noted that the consistency loss is about four orders of magnitudes smaller than the metric loss

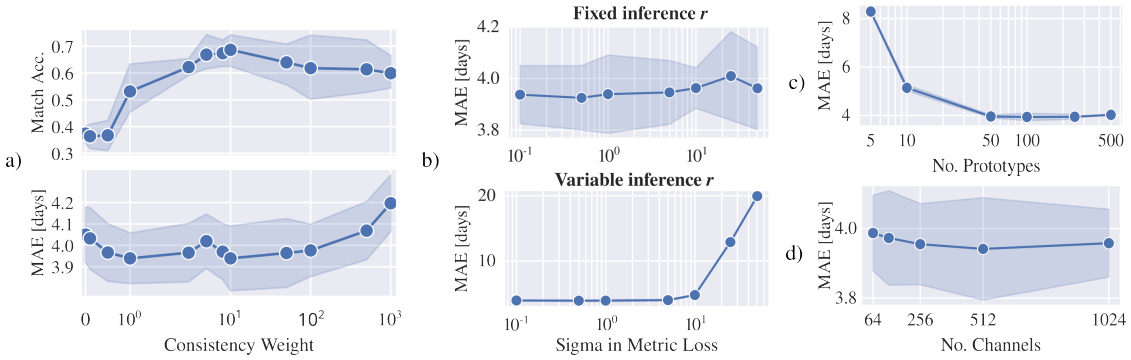


Figure 7.12: Overview of hyperparameter tuning for the US dataset for the (a) consistency weight (β), (b) standard deviation in metric loss (σ), (c) number of prototypes and (d) number of channels in the add-on layers and prototype representations. For σ two types of inference are shown, keeping the inference radius, r, constant (top) or varying it in line with the training sigma, $r = 3\sigma$, (bottom).

during training, which explains the small effect for very low weights.

In Fig. 7.12b the standard deviation (σ) in the metric loss (Eq. 7.9) is varied. The top and bottom panels show the results of the same training runs, but during inference, the radius is either fixed ($r = 3$, top panel) or adapted based on the σ during training ($r = 3\sigma$, bottom panel). The training sigma has only a small effect on performance throughout the range of values tried, whereas the inference radius does considerably affect the performance, most notably at higher values. This is beneficial as the inference r can be easily adjusted after training, and can thus be optimised offline.

The number of prototypes and the number of channels in the prototype representations (and in the add-on layers) are shown in Fig. 7.12c and d, respectively. Both show improved performance when increasing the number of prototypes or channels, levelling off for higher values.

Overall, these results show that the hyperparameters introduced in the proposed model behave expectedly. This confirms the stability of the method and shows that the results are not too sensitive to the hyperparameter selection.

Generalisability to LUMC Dataset To determine the generalisability of the proposed model to a new US dataset, the model trained on the INTERGROWTH-21st US dataset was also applied to the LUMC dataset. The model was applied

Table 7.2: Results of the *PRINT* model on the LUMC dataset without any fine-tuning, the values indicate standard deviations and the stars indicate significance with respect to the ResNet baseline ($p < 0.01$)

	MAE (days)
ResNet	8.9 (7.54)
ProtoNet	7.15* (5.7)

with frozen weights to the new dataset without any fine-tuning. The results of this experiment are shown in Table 7.2, demonstrating that *PRINT* outperforms the ResNet baseline significantly ($p < 0.01$ using a paired t-test). However, the reported error is substantially larger than the MAE reported on the INTERGROWTH-21st US dataset (MAE of 7.15 vs. 3.93), which is mainly caused by large errors for images late in the GA range. As the images of the LUMC dataset show a large domain gap with the INTERGROWTH-21st images later in gestation, due to the higher image quality of the LUMC images, this is in line with the expectation.

7.5.2 Adult MRI

Baseline Comparisons In Table 7.3 the performance for the MRI dataset is shown, with the same two baselines as implemented for the US dataset (a vanilla ResNet-18 and the *AvgPool* baseline). In brain MR images, both hemispheres are visible, and, therefore, there is no need to enforce inter-hemispheric patch similarity with a consistency loss. As such, this loss was not applied to the MRI dataset, and the only ablation performed for this dataset were the add-on layers (h). It is evident that these, in line with the US findings, improve the prediction performance (an MAE of 6.19 vs. 6.37 without). Furthermore, the *PRINT* architecture performs on par with the two implemented baseline architectures. It has been argued that due to differences in the age range in the data, the weighted MAE, which simply divides the MAE by the age range in years, is a better measure to compare MR brain age prediction studies with each other [224]. For this reason, the weighted MAE is also shown in Table 7.3. When considering the large age range of the IXI dataset (in the weighted MAE), and the relatively small size of the training dataset

Table 7.3: Quantitative results for MRI on the test set. The values between brackets indicate the standard deviations. The wMAE represents the weighted MAE which is obtained by normalising for the age range of the dataset by dividing by it (MAE/age range) [224]. The ablations show the results without add-on layers in the feature extractor(h). All MRI experiments were performed without consistency loss ($\mathcal{L}_c$)

	Ablation		
	h	MAE [years]	wMAE
ResNet Baseline	-	6.29 (4.79)	0.094
AvgPool Baseline	-	6.41 (5.47)	0.096
PRINT (no h)	×	6.37 (5.11)	0.095
PRINT	✓	6.19 (4.95)	0.092

(i.e. compared to using UK Biobank images), the reported errors are also in range with previous studies for adult brain MRI age prediction [224].

Predicted versus True Age In Fig. 7.9b the true versus predicted chronological age is shown for the MRI dataset. It can be observed that, compared to the US predictions, larger errors are observed. This can be attributed to the smaller training dataset size as well as to the larger variability present in the IXI dataset, as it is acquired at three different sites.

Example Predictions An example prediction showing the nearest prototypes for an adult brain MR image is shown in Fig. 7.10. It can be seen that the closest prototypes have a similar ventricular size to the sample image, aligning with the observation that the ventricular size is highly age-discriminative for adult brain MR images.

The patch-based matching results for three sample-prototype pairs are shown in Fig. 7.13. It is evident that the matching matrices are mostly according to the diagonal, suggesting that each patch in the image is matched with the same spatial location in the prototype. As the images were aligned to the same coordinate system, this matching is expected and indicates that the model can learn the similarity between corresponding patches.

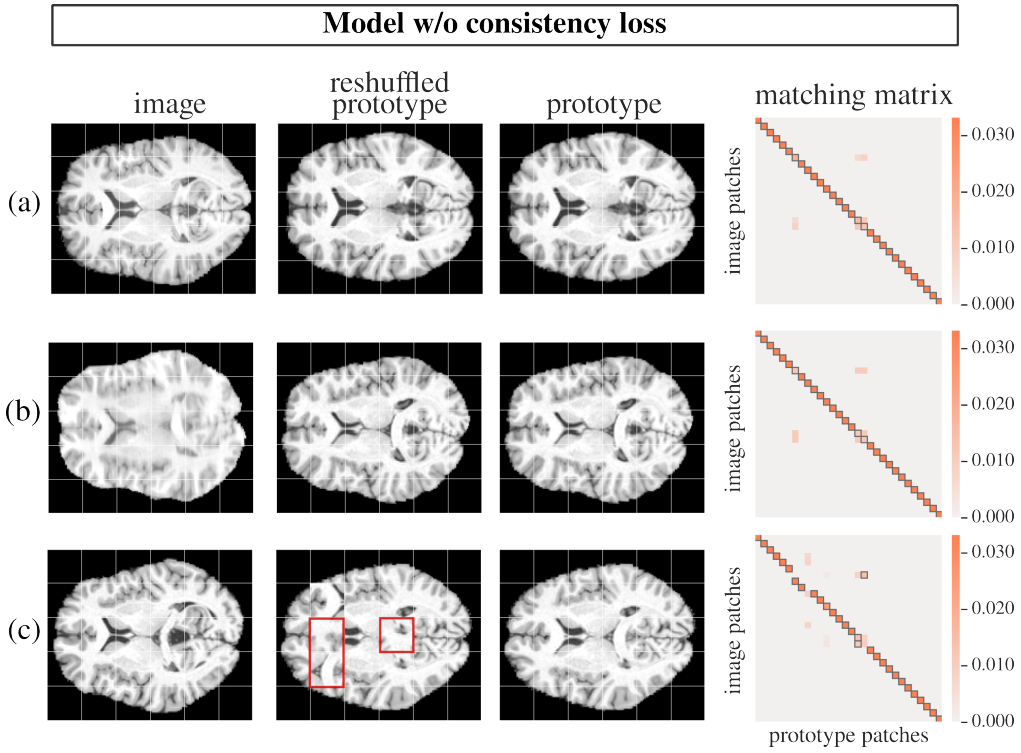


Figure 7.13: Illustration of OT matching between image sample and prototypes for MRI. The MR images are shown in the same orientation as the US images, with the midline on the horizontal axis. The obtained OT matching matrix ($\mathbf{Q}$) is given in the last column of each panel, with the most similar prototype patch to each image patch given a black border. These patches are shown in order of the sample image in the reconstructed prototype, with patches that are incorrectly matched shown with red borders. For MRI three examples are shown for the model without consistency loss.

For almost all test set volumes, the sample-prototype pairs were matched correctly (i.e. according to the structural anatomy), but Fig. 7.13c shows one of the few samples where a few patches were mismatched (shown with red borders). However, even in this case, most patches were correctly identified. This thus demonstrates that anatomically correct image patches are matched, even when the volumes are acquired on different MRI scanners (Fig. 7.13b), despite known harmonisation effects [99].

7.6 Discussion

In this chapter, I proposed a novel interpretable model for continuous regression tasks based on prototype learning and patch-based OT matching. I showed that

the model obtained competitive prediction performance on two brain age prediction datasets; fetal US and adult MRI. On both datasets, the patch-based distance metric was able to correctly learn the structural matches between the sample image and prototypes. The proposed method is versatile and can in future work be applied to other continuous regression problems.

The primary focus of this chapter is not merely achieving competitive age prediction performance. Instead, I aimed to demonstrate that it is possible to attain a more interpretable model without sacrificing prediction accuracy. The type of explanations obtained in this chapter differs from the traditional attribution-based methods, where pixel-level importance scores are generated. Instead of answering the question of *which* image features are used, this chapter sought to address *how* these features are used to make a prediction. It thus provides an alternative approach to making neural networks more interpretable, and could, in future work, be combined with attention- or attribution-based methods. Furthermore, while in this study uniform initial distributions were used in Eq. 7.4 (i.e. all patches get the same importance weight), this could in further work be replaced by other options such as cross-correlation between the patches [170] to generate importance scores for each of the patches, further improving the interpretability of the model.

The images used in this chapter were all aligned to a common coordinate system as preprocessing step and, as such, the only geometric variation present in these images is which of the two hemispheres is visible in the US images. For this reason, the consistency loss was only applied to the US dataset, using horizontal flips across the midline as geometric augmentation. In future work, the consistency loss can easily be adapted to other geometric variations present in the dataset. For example, the $PRINT$ architecture could be trained with US images in their original orientation (i.e. not aligned to a common coordinate system), and consistency could then be enforced for translation, rotation and flipping.

A limitation of the consistency loss is that it regularises the distances between patches in latent space, where the spatial resolution is several times lower than in the image space. The consistency can thus only be enforced for larger geometric

augmentations. However, future work could investigate the effect of a higher-resolution latent space on performance, as this would enable the regularisation of smaller geometric augmentations.

This chapter provided experiments with the interpretable *PRINT* model on 2D images only. In theory, the proposed architecture can also directly be applied to 3D image volumes by replacing all 2D operations with the corresponding 3D operations. However, as the complexity of the Sinkhorn algorithm scales with $O(n^2)$, with n the number of patches, this will lead to a substantial increase in training time. While not investigated in detail in this chapter, the Sinkhorn algorithm typically converged within a few iterations to a relatively good estimate of the optimal matching matrix. It could therefore be considered to reduce the maximum number of iterations to reduce the computational time.

In summary, this chapter demonstrated the feasibility of an inherently interpretable network for brain age prediction. In further work, it can be extended to 3D images as well as to US images acquired in their original orientation.

8 Conclusion

In this thesis, I developed several methods for the quantitative analysis of fetal brain US during the second trimester of gestation. In Chapter 5 I made use of automated segmentation masks to measure the volumetric growth during gestation, whereas both Chapter 6 and 7 quantify the stage of development of the brain by using brain age predictions. All of the developed methods make the analysis of large datasets feasible, and can, as such, provide novel insight in brain development during gestation. This section will summarise the main contributions presented in this work, and identify the limitations and directions for further research.

8.1 Contributions

The main contributions presented in this thesis can be summarised as follows:

1. **Subcortical Segmentation in 3D US**

   In Chapter 5 I have shown how a limited number of manual annotations can be used to train a subcortical segmentation network that achieves performance close to the intra-observer variability for four structures of interest in 3D US: the CB, LPVH, CP and CSPV. This chapter demonstrated that few individually annotated images resulted in better performance when the images were aligned to the same reference system, whereas a larger number of weakly

labels achieved better performance for images in their original orientation. The developed models were then applied to a large subset of healthy fetuses, generating novel US-specific growth curves. The growth curves were validated by comparing the CB growth to previously reported volumetric measures of the CB, showing excellent agreement between them. This contribution thus demonstrated the feasibility of automated subcortical structure segmentation in 3D US, even when manual annotations are costly to obtain.

2. **Local Brain Age Prediction from US with Guided Attention**
   In Chapter 6 I introduced a network architecture that is able to predict *local* brain age. The network architecture incorporates a guided attention module that forces the network to focus on specific structures in the brain. Similar guided attention architectures have been used in two previous works [205, 206], but, to the best of my knowledge, this is the first work that uses guided attention to achieve *local* predictions (i.e. based on only part of the image). I performed experiments with both 2D images and 3D image volumes and showed that there was a significant increase in brain age prediction performance (5 - 10 % MAE improvement) when regularising the attention with known age-discriminative regions such as the CoP and CB. This clearly demonstrated the benefit of guiding the networks' focus towards structures of interest. The obtained attention maps for the CoP-guided training also showed high attention at the location of two fissures (the SF and POF) that are known to actively develop during the second trimester of gestation, confirming that the developed age prediction network focuses on age-discriminative areas.

3. **Protoype Learning for Interpretable Brain Age Prediction**
   In Chapter 7 I proposed a novel architecture for the interpretable prediction of brain age. The network can provide inherently interpretable predictions based on patch-level similarities to a set of learned prototypes. The architecture was specifically tailored towards continuous regression, and predicts the brain age of a sample image based on a weighted sum of prototype labels in the

vicinity in latent space. In the architecture, OT is used to match the patches of the image and prototype with each other, and it was demonstrated that this enables the opposite sides of the brain hemispheres to match with each other (i.e. the visible hemispheres). Experiments were performed on 2D US images of the fetal brain, as well as on a dataset of 2D MR images of the adult brain, resulting in competitive performance for both datasets, demonstrating that increased interpretability can be achieved without sacrificing prediction accuracy.

4. **Generalisability of Developed Models**

   In each of the Chapters 5–7 I also demonstrated the performance on an independent test set (the LUMC dataset, as described in Chapter 4) without any fine-tuning. This dataset was acquired at a different institute and with a different US scanner than the images in the training dataset, and therefore provided a good measure of generalisation performance across datasets. The predictions of the segmentation network were examined and graded according to the quality of the segmentation by a trained ultrasonographer. This showed that for most structures, almost 90% of the images had a high or intermediate segmentation quality, indicating that the learned features transferred well to a new US dataset. Furthermore, the brain age prediction performance (Chapter 6 and 7) on the independent test set was high, especially for the 3D images in the second trimester (MAE of 3.73 days). This clearly demonstrates the feasibility of using the developed models for datasets acquired at different sites or with different scanners.

8.2 Limitations and Future Work

The individual algorithms and methods developed as part of this thesis have a number of limitations, which have already been addressed in the individual chapters. However, the work presented in this thesis also has some overall limitations. These

limitations often motivate future work and these will therefore be discussed together in this section.

8.2.1 Datasets

Image Quality and Age Range The models developed in this thesis were mainly trained with images from the INTERGROWTH-21st dataset during the second trimester of gestation. In this dataset the image quality strongly reduced in the 3rd trimester, primarily due to an increase in shadowing artefacts as a result of the increased calcification of the skull. For this reason, it was not possible to train the segmentation and brain age algorithms far into the third trimester.

Furthermore, as the quality of the US images decreases during gestation, the image quality is thus, to some degree, dependent on the GA. As a result, image quality can possibly be learned by the network as an age-discriminative feature, especially for the larger GA range used in Chapter 7 (14 - 30 GWs). However, the learned attention masks shown in Chapter 6 (Fig. 6.14) demonstrated a clear focus on age-discriminative regions of the CoP (i.e. the SF and CoP), suggesting a limited contribution of image quality to the prediction.

In future work, the developed models can be extended into the third trimester using a dataset with higher image quality in that trimester. For example, the LUMC dataset shows a substantially better image quality in the third trimester than the images in the INTERGROWTH-21st dataset, due to the acquisition with a newer US device.

Size Normalisation An important consideration in this thesis is that size information was removed from the US images for brain age prediction (Chapter 6 and Chapter 7), by upscaling them to the same whole brain size. This was motivated by the fact that the aim is to develop a model that can assess structural brain development rather than volumetric size, as size can be readily inferred with 2D or 3D measurements. While it is not possible to completely separate structural development from volumetric size, as both are inter-dependent, scaling the images

to the same brain size should remove most of the size information and is in line with previous works [86, 88]. A possible drawback of this approach is that interpolation artefacts can be introduced into the images that could then be used by the network to predict the age. While this can not be ruled out completely, early experiments for brain age prediction with strong up- and downsampling scaling did not indicate any such effects. In addition, the high attention at age-discriminative regions also suggests that the network does not predominantly use interpolation patterns. In future work, other methods could be considered to remove the volumetric size information such as dedicated confound removal techniques [10, 225], however, due to the strong dependency between age and size, this is a non-trivial task.

Data Dimensionality All experiments with 2D images in this thesis were performed with 2D planes extracted from aligned 3D US volumes with a set of fixed coordinates. While this provided a direct way of obtaining a large number of 2D standard plane images, this is not completely representative of clinical 2D US acquisition. In a routine US examination typically 2D+t videos are acquired, and certain planes are saved as 2D images. However, due to factors like the fetal position and the ultrasonographers' experience, such images might not always capture exactly the same anatomical plane. As a result, the variation between such images is usually larger than that seen in the 2D images used in this thesis. Further work should look into adapting the developed algorithms to create an end-to-end pipeline that can make predictions from routinely acquired US images in clinical practice. As the 2D+t videos contain more structural information than the individually acquired planes, it would be of interest to integrate the developed models with a 3D volumetric reconstruction algorithm [226], so that the 3D models developed in this thesis can be applied to the 2D+t videos.

8.2.2 Generalisability

Domain Performance In this thesis, I have demonstrated the generalisability performance of the developed models on an independent test set, the LUMC dataset.

While achieving good segmentation performance for most subcortical structures, the segmentation of the LPVH was graded considerably lower than the other structures. Likewise, brain age prediction accuracy on this dataset was high for the second-trimester images in Chapter 6 (MAE of 5.24 and 3.73 for the 2D and 3D images, respectively), but substantially lower for the larger GA range used in Chapter 7 (MAE of 7.15). This decrease in brain age prediction performance can likely be attributed to the large differences in US quality between the INTERGROWTH-21st and LUMC dataset later in the GA range (i.e. from about 28 GWs), as described above. In this thesis, the developed models were applied to the LUMC dataset without fine-tuning. Further work could investigate the use of dedicated domain adaptation techniques to further improve the segmentation and brain age prediction performance on external datasets [203, 227].

Age Prediction Ground-Truth Another consideration, specifically for the brain age prediction models, is the difference in the estimation of ground-truth GA between the datasets. In the INTERGROWTH-21st dataset the GA was computed from the LMP, but only women with an accurately known LMP were included in the study. On the other hand, in the LUMC dataset the ground-truth GA was determined with an US-based CRL measurement in the first trimester. These differences in ground-truth estimation might thus also contribute to the difference in age prediction performance. This is also an important aspect to consider when applying the age prediction models to study the structural development in new cohorts, as the difference between predicted and ground-truth GA (the brain age gap) is only meaningful if the ground-truth GA is an accurate estimate. In future work, it would be of interest to validate the brain age prediction models with a dataset of fetuses conceived with IVF [90], as for this group the day of conception is exactly known. However, such datasets are challenging to obtain and this might therefore not be very feasible.

Population Research The reported average performance for the task of brain age prediction in this thesis is high, with the best MAE being only 3.45 days. However, the relatively large standard deviations (2.82 days) indicate that the predictions on an individual level can vary considerably. This can partly be attributed to the natural variation present in the structural brain development of individuals in a healthy population, and individual predictions can thus not directly be linked to normal or abnormal brain development. For this reason, the proposed brain age prediction methods are, at this time, intended for population research rather than for predictions on an individual level. For example, the methods could be applied to study brain development in a cohort of fetuses with congenital heart disease. To further improve performance and reduce the variability introduced by the stochastic nature of training deep learning models, also an ensemble of models could be considered, as these typically result in better prediction performance.

8.2.3 Interpretability

Visualisations In Chapter 6 and 7 dedicated methods have been designed to improve the interpretability of brain age prediction models. Both of these chapters addressed a different question, with Chapter 6 focussing on *which* images features are used in the prediction, and Chapter 7 on *how* features within individual images are used for the prediction. The masked feature representations obtained in Chapter 6 could be directly visualised, providing an intuitive way to present the importance of individual image features. On the other hand, it was more challenging to visualise the complete decision-making process of *PRINT*, as it consists of similar prototypes as well as of patch-based distances between all image-prototype pairs. In further work, interactive graphical user interfaces could be developed that allow the user to explore the visualisations in more detail. This could provide insight into model behaviour that can then be used to improve the prediction model. For example, the consistency loss was motivated by the observation that the visible hemispheres in the US images were not correctly matched with each other.

Receptive Field As the methods in Chapter 6 are based on CNNs, it is important to consider that information from outside the region of interest (which is enforced with a segmentation loss) could contribute to the prediction to some degree, due to the large receptive fields of the convolutional filters. A similar consideration is present in Chapter 7, where distances between patches in latent space might also be influenced by image features outside the image patch obtained by upsampling. Both of these chapters use the assumption that convolutions largely preserve spatial relationships, but this might not always hold in practice. In future work, experiments can be carried out that study this relationship in more detail, for example by occluding parts of the image outside the region of interest (Chapter 6) or patch (Chapter 7), and observing the effect on the model predictions.

8.2.4 Future Applications

Abnormal Development In this thesis, the developed models have only been trained using a dataset of healthy fetuses. Therefore, only changes in the *speed* of development (i.e. whether a certain structural pattern is age-appropriate) can be detected, as such images will resemble images from the training set. On the other hand, the models will not be able to quantify structural patterns that bear little resemblance to healthy development. For example, the segmentation algorithm will most likely fail to segment the LPVH in fetuses with severe ventriculomegaly, as the model has not seen such cases during training. In future work, the segmentation algorithm can be extended with non-normal training data so that it is more robust to changes in brain appearance. This is more difficult for brain age predictions, as the concept of the brain age gap can only be reliably used to quantity whether the stage of development is age-appropriate. However, strong abnormalities are often straightforward to detect by an ultrasonographer, and as such, these cases can be excluded from analysis. More subtle changes are more challenging to detect visually, and brain age prediction models can thus aid the analysis of these cases.

Analysis of Non-Normal Cohorts The models developed in this thesis enable the analysis of large cohorts of data, without the need for costly manual annotations. While the developed models were only applied to images of healthy fetuses, in future work these models could be applied to non-normal cohorts to study brain development in those groups. For example, children born with congenital heart disease often show delayed cognitive development in childhood, but the developmental origin of this delay is unknown. While earlier work has studied brain age predictions in this group [228], the *local* brain age model proposed in this work (Chapter 6) can identify local changes and, as such, might provide novel insight in the structural brain development in this group. An important consideration when applying the developed models to study non-normal cohorts, is that a control (healthy) group is required with the same characteristics (i.e. same site and scanner). This ensures that any observed changes in the predictions can be correctly attributed to the abnormality, rather than being a result of domain shift.

Appendices

A

INTERGROWTH-21st Data Selection Criteria

Overall 3D image	**Midsagittal Plane**
Absence of maternal or fetal motion artifacts	Cavum septi pellucidi visible
Head fully contained within the image	Eye sockets not visible
Minimal acoustic shadows in the distal hemisphere*	Facial profile clearly visible (<24 weeks' gestation)*
High contrast between fluid-filled (cavum septi pellucidi, ventricles) and dense (e.g. skull, choroid plexus) regions*	Choroid plexus not visible (<18 weeks)

Transthalamic plane	**Transcerebellar plane**
Symmetrical plane	Cerebellum visible*
Head occupying at least 30% of image	Cerebellar vermis visible*
Cavum septi pellucidi visible	Cisterna magna visible
Choroid plexus visible within ventricular cavity	
Thalami visible	
Posterior ventricle cavity visible	
Sylvian fissure visible	

* relaxed for validation subset selection (i.e. minimal shadows → minimal large shadows, high contrast → medium-high contrast , visible → partly visible)

Figure A.1: Inclusion criteria for the data selection of the INTERGROWTH-21st dataset. All criteria were used for selection of the *atlas* subset, and a relaxed set of criteria (those marked with a * were relaxed) for the *validation* subset. These criteria were created by Prof. Namburete for earlier work.

B

Additional Results Subcortical Segmentation

Table B.1: P-values obtained from comparing the DSC segmentation performance of both networks, $\dot{\theta}^{atl}$ and $\dot{\theta}^{exp}$, with each other as well as with the naive propagated atlas masks (*prop. atlas*). For each structure a repeated measures ANOVA was performed followed by post-hoc testing with a paired t-test. Reported p-values are *post-hoc* tests corrected with Bonferonni correction for the four structures, as well as for the three model comparisons. For the CSPV the ANOVA returned non-significant differences (p=0.08), and as such the *post-hoc* tests are not reported.

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

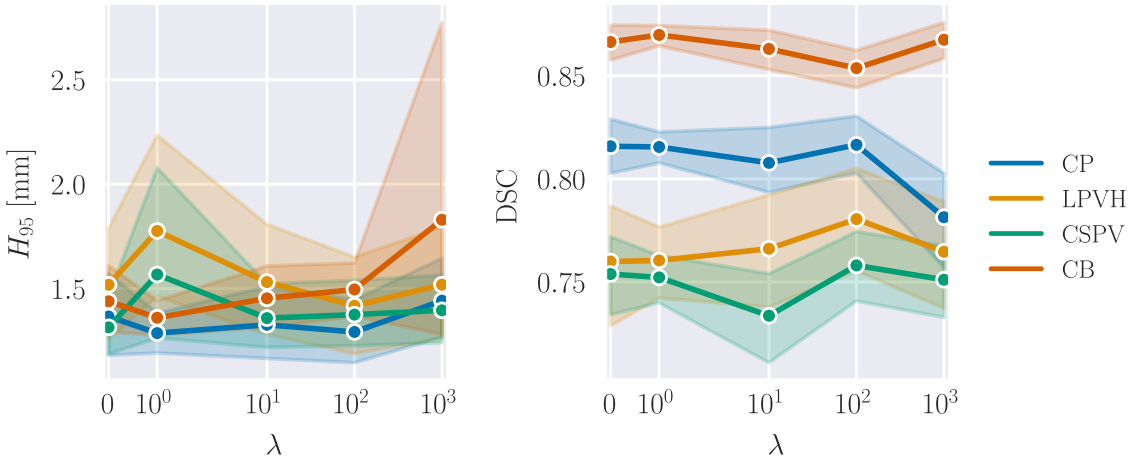


Figure B.1: Segmentation performance when varying the weighing parameter between the $\mathcal{L}_{MD}$ and $\mathcal{L}_{CE}$, λ in Eq. 5.1.

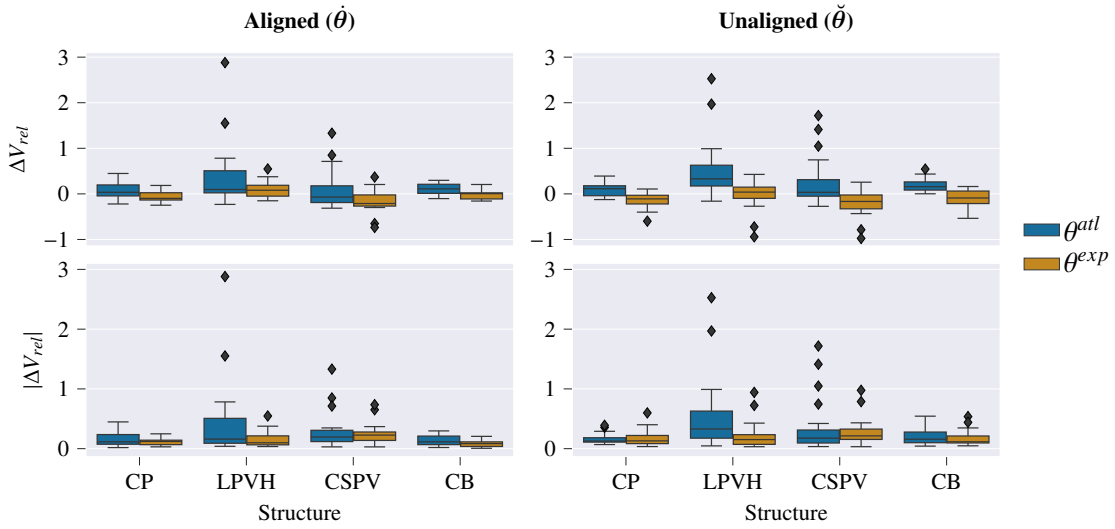


Figure B.2: Signed and unsigned relative volume differences after post-processing for θ^{exp} and θ^{atl} after largest component post-processing. θ^{atl} shows an average ΔV_{rel} higher than 0 (except for the CSPV in the aligned setting), corresponding to over-segmentation. θ^{exp} shows, on average, a slight negative ΔV_{rel}, thus under-segmentation, for the unaligned images while the differences are centred around zero for the aligned data. The clear outliers for the CSPV and LPVH correspond to images at 18 or 20 GW where the total ground-truth volume of the respective structures is very small, resulting in large relative volume differences.

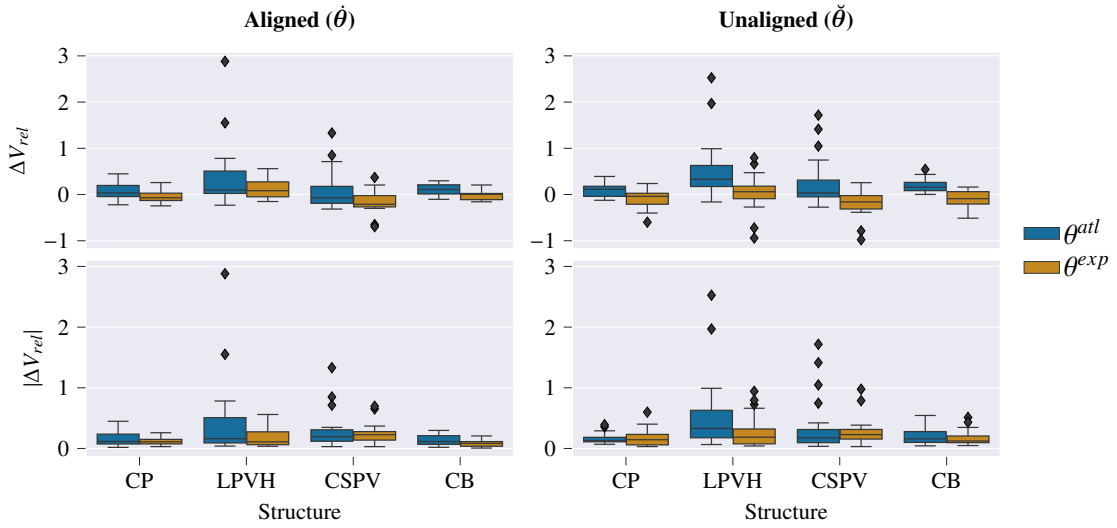


Figure B.3: Signed and unsigned relative volume differences before the largest connected component post-processing step was applied for both θ^{atl} and θ^{atl}. It can be seen that post-processing has only a very small effect on the resulting volume measurements (i.e. compared to Fig B.2).

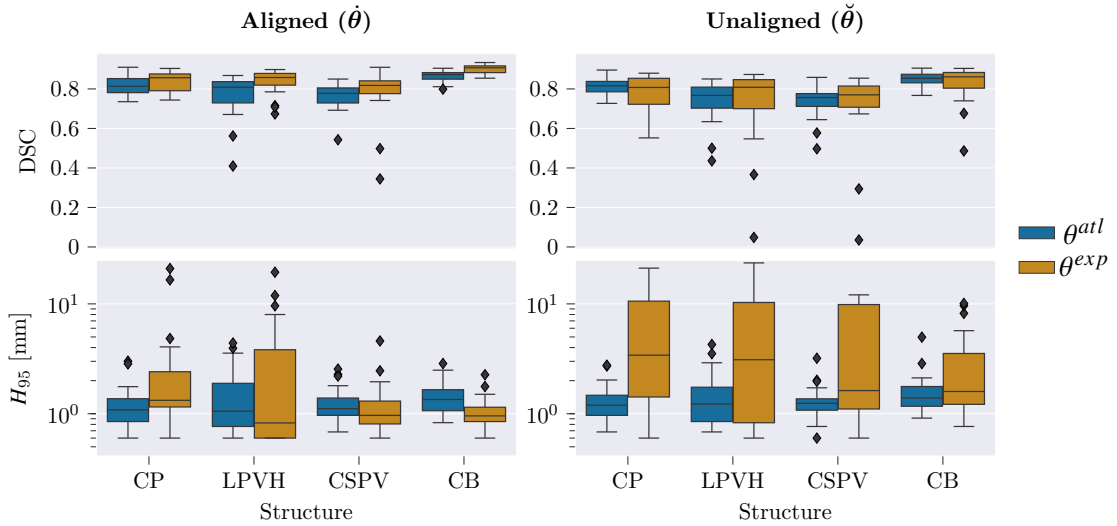


Figure B.4: Resulting performance values before the largest connected component post-processing step was applied for θ^{atl} and θ^{atl}. Note the log scale for the y-axis of the bottom row. It can be seen that the reported Hausdorff distances are without post-processing are large, which is due to spurious segmented regions, predominantly in the less-visible hemisphere.

C

Optimal Transport Derivation

The optimal transport (OT) objective between a cost matrix, $\mathbf{C}$, and a coupling matrix, $\mathbf{Q}$, with entropic regularisation in Eq. 7.5 was given as:

$$\min_{\mathbf{Q}}(\sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij}q_{ij} - \epsilon \mathrm{H}(\mathbf{Q})) = \min_{\mathbf{Q}}(\sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij}q_{ij} - \epsilon \sum_{i=1}^{m}\sum_{j=1}^{n} q_{ij}\log(q_{ij})) \tag{C.1}$$

This objective can be rewritten in matrix notation with inner products as:

$$\min_{\mathbf{Q}}(\langle \mathbf{T}, \mathbf{Q}\rangle + \epsilon\langle \mathbf{Q}, \log(\mathbf{Q})\rangle) \tag{C.2}$$

with all $\log(\cdot)$ and $e^{(\cdot)}$ functions in this section being element-wise. Leaving the $\min_{\mathbf{Q}}$ out for clarity, and using the linearity of inner products, this formulation can be written as:

$$= \langle \mathbf{Q}, \mathbf{C} + \epsilon\log(\mathbf{Q})\rangle = \epsilon\langle \mathbf{Q}, \frac{\mathbf{C}}{\epsilon} + \log(\mathbf{Q})\rangle \tag{C.3}$$

Considering that $x = \log(e^{-x})$, this can be simplified to:

$$= \epsilon\langle \mathbf{Q}, -\log\Big(e^{-\mathbf{C}/\epsilon}\Big) + \log(\mathbf{Q})\rangle = \epsilon\langle \mathbf{Q}, \log\Big(\frac{\mathbf{Q}}{e^{-\mathbf{C}/\epsilon}}\Big)\rangle \tag{C.4}$$

Defining a new matrix $\mathbf{K}$ as a kernel function of $\mathbf{C}$:

$$\mathbf{K} = e^{-\mathbf{C}/\epsilon} \tag{C.5}$$

then Eq. C.4 simplifies to:

$$= \epsilon \langle \mathbf{Q}, \log\left(\frac{\mathbf{Q}}{\mathbf{K}}\right) \rangle \tag{C.6}$$

which is the definition of the KL divergence between $\mathbf{Q}$ and $\mathbf{K}$, scaled by ϵ. As the objective is minimized with respect to $\mathbf{Q}$, the scaling factor will not have an effect, and can be left out of the expression. The full objective, written as a KL divergence is now given by:

$$\min_{Q}(D_{KL}(\mathbf{Q} \parallel \mathbf{K})), \text{ with } \mathbf{K} = e^{-\mathbf{C}/\epsilon} \tag{C.7}$$

Bibliography

[1] G. Malinger, D. Paladini, K. K. Haratz, A. Monteagudo, G. L. Pilu, and I. E. Timor-Tritsch. "ISUOG Practice Guidelines (updated): sonographic examination of the fetal central nervous system. Part 1: performance of screening examination and indications for targeted neurosonography". In: *Ultrasound Obstetrics & Gynecology* 56 (3 2020), pp. 476–484.

[2] D Paladini, G Malinger, R Birnbaum, A Monteagudo, G Pilu, LJ Salomon, and IE Timor-Tritsch. "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.

[3] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42 (2017), pp. 60–88.

[4] Linde S Hesse and Ana IL Namburete. "Improving U-Net Segmentation with Active Contour Based Label Correction". In: *Annual Conference on Medical Image Understanding and Analysis (MIUA)*. Springer. 2020, pp. 69–81.

[5] Mohammad Yaqub, Remi Cuingnet, Raffaele Napolitano, David Roundhill, Aris Papageorghiou, Roberto Ardon, and J Alison Noble. "Volumetric segmentation of key fetal brain structures in 3D ultrasound". In: *International Workshop on Machine Learning in Medical Imaging (MLMI)*. Springer. 2013, pp. 25–32.

[6] Benjamín Gutiérrez-Becker, Fernando Arámbula Cosío, Mario E Guzmán Huerta, Jesús Andrés Benavides-Serralde, Lisbeth Camargo-Marín, and Verónica Medina Bañuelos. "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.

[7] Lorenzo Venturini, Aris T Papageorghiou, J Alison Noble, and Ana IL Namburete. "Multi-task CNN for Structural Semantic Segmentation in 3D Fetal Brain Ultrasound". In: *Annual Conference on Medical Image Understanding and Analysis (MIUA)*. Springer. 2019, pp. 164–173.

[8] Lorenzo Venturini, Aris T Papageorghiou, J Alison Noble, and Ana IL Namburete. "Uncertainty Estimates as Data Selection Criteria to Boost Omni-Supervised Learning". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2020, pp. 689–698.

[9] Ana I.L. Namburete, Richard V. Stebbing, Bryn Kemp, Mohammad Yaqub, Aris T. Papageorghiou, and J. Alison Noble. "Learning-based prediction of gestational age from ultrasound images of the fetal brain". In: *Medical Image Analysis* 21 (1 2015), pp. 72–86.

[10] Nicola K Dinsdale, Emma Bluemke, Stephen M Smith, Zobair Arya, Diego Vidaurre, Mark Jenkinson, and Ana IL Namburete. "Learning patterns of the ageing brain in MRI using deep convolutional networks". In: *NeuroImage* 224 (2021), p. 117401.

[11] P. E. Hansen, M. C. Ballesteros, K. Soila, L. Garcia, and J. M. Howard. "MR imaging of the developing human brain. Part 1. Prenatal development". In: *RadioGraphics* 13 (1 1993), pp. 21–36.

[12] Je G. Chi, Elizabeth C. Dooling, and Floyd H. Gilles. "Gyral development of the human brain". In: *Annals of Neurology* 1 (1 1977), pp. 86–93.

[13] K. Dorovini Zis and C. L. Dolman. "Gestational development of brain". In: *Archives of Pathology & Laboratory Medicine* 101 (4 1977), pp. 192–195.

[14] C. Garel, E. Chantrel, H. Brisse, M. Elmaleh, D. Luton, J. F. Oury, G. Sebag, and M. Hassan. "Fetal Cerebral Cortex: Normal Gestational Landmarks Identified Using Prenatal MR Imaging". In: *American Journal of Neuroradiology* 22 (1 2001), p. 184.

[15] Vidya Rajagopalan, Julia Scott, Piotr A. Habas, Kio Kim, James Corbett-Detig, Francois Rousseau, A. James Barkovich, Orit A. Glenn, and Colin Studholme. "Local Tissue Growth Patterns Underlying Normal Fetal Human Brain Gyrification Quantified In Utero". In: *The Journal of Neuroscience* 31 (8 2011), p. 2878.

[16] 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.

[17] Hyuk Jin Yun, Hyun Ju Lee, Joo Young Lee, Tomo Tarui, Caitlin K. Rollins, Cynthia M. Ortinau, Henry A. Feldman, P. Ellen Grant, and Kiho Im. "Quantification of sulcal emergence timing and its variability in early fetal life: Hemispheric asymmetry and sex difference". In: *NeuroImage* 263 (2022), p. 119629.

[18] 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 in Obstetrics & Gynecology* 24 (7 2004), pp. 706–715.

[19] A. Monteagudo and I. E. Timor-Tritsch. "Development of fetal gyri, sulci and fissures: a transvaginal sonographic study". In: *Ultrasound in Obstetrics & Gynecology* 9 (4 1997), pp. 222–228.

[20] Bero Olof Verburg, EAP Steegers, Maria De Ridder, RJM Snijders, E Smith, Albert Hofman, HA Moll, VWV Jaddoe, and JCM Witteman. "New charts for ultrasound dating of pregnancy and assessment of fetal growth: longitudinal data from a population-based cohort study". In: *Ultrasound in Obstetrics & Gynecology* 31.4 (2008), pp. 388–396.

[21] Finale Doshi-Velez and Been Kim. "Towards A Rigorous Science of Interpretable Machine Learning". In: *arXiv preprint arXiv:1702.08608* (2017).

[22] Amina Adadi and Mohammed Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160.

[23] Adrian Weller. "Transparency: Motivations and Challenges". In: *Explanainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Vol. 11700 LNCS. Springer Verlag, 2019, pp. 23–40.

[24] 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: *Lancet* 384.9946 (2014), pp. 869–879.

[25] Ana IL Namburete, Raquel van Kampen, Aris T Papageorghiou, and Bartłomiej W Papież. "Multi-channel groupwise registration to construct an ultrasound-specific fetal brain atlas". In: *International Workshop on Preterm, Perinatal and Paediatric Image Analysis*. Springer, 2018, pp. 76–86.

[26] Madeleine K Wyburd, Linde S Hesse, Moska Aliasi, Mark Jenkinson, Aris T Papageorghiou, Monique C Haak, and Ana IL Namburete. "Assessment of Regional Cortical Development Through Fissure Based Gestational Age Estimation in 3D Fetal Ultrasound". In: *International Workshop on Preterm, Perinatal and Paediatric Image Analysis*. Springer, 2021, pp. 242–252.

[27] Madeleine K Wyburd, Nicola K Dinsdale, Ana IL Namburete, and Mark Jenkinson. "TEDS-Net: enforcing diffeomorphisms in spatial transformers to guarantee topology preservation in segmentations". In: *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer. 2021, pp. 250–260.

[28] Christoph B. Burckhardt. "Speckle in Ultrasound B-Mode Scans". In: *IEEE Transactions on Sonics and Ultrasonics* 25 (1 1978), pp. 1–6.

[29] John W Prothero and John W Sundsten. "Folding of the cerebral cortex in mammals: a scaling model". In: *Brain, Behavior and Evolution* 24.2-3 (1984), pp. 152–167.

[30] Michael D Gregory, J Shane Kippenhan, Dwight Dickinson, Jessica Carrasco, Venkata S Mattay, Daniel R Weinberger, and Karen F Berman. "Regional variations in brain gyrification are associated with general cognitive ability in humans". In: *Current Biology* 26.10 (2016), pp. 1301–1305.

[31] TP Naidich, JL Grant, N Altman, RA Zimmerman, SB Birchansky, B Braffman, and JL Daniel. "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.

[32] L. R. Pistorius, P. Stoutenbeek, F. Groenendaal, L. de Vries, G. Manten, E. Mulder, and G. Visser. "Grade and symmetry of normal fetal cortical development: a longitudinal two- and three-dimensional ultrasound study". In: *Ultrasound in Obstetrics & Gynecology* 36.6 (2010), pp. 700–708.

[33] C Clouchoux, AJ Du Plessis, M Bouyssi-Kobar, W Tworetzky, DB McElhinney, DW Brown, A Gholipour, D Kudelski, SK Warfield, RJ McCarter, et al. "Delayed cortical development in fetuses with complex congenital heart disease". In: *Cerebral cortex* 23.12 (2013), pp. 2932–2943.

[34] Cynthia M. Ortinau et al. "Early-Emerging Sulcal Patterns Are Atypical in Fetuses with Congenital Heart Disease". In: *Cerebral Cortex* 29 (8 2019), p. 3605.

[35] Ernst Schwartz et al. "The Prenatal Morphomechanic Impact of Agenesis of the Corpus Callosum on Human Brain Structure and Asymmetry". In: *Cerebral Cortex* 31 (9 2021), pp. 4024–4037.

[36] D. J. Warren, D. J.A. Connolly, and P. D. Griffiths. "Assessment of sulcation of the fetal brain in cases of isolated agenesis of the corpus callosum using in utero MR imaging". In: *American Journal of Neuroradiology* 31 (6 2010), pp. 1085–1090.

[37] Tomo Tarui et al. "Disorganized Patterns of Sulcal Position in Fetal Brains with Agenesis of Corpus Callosum". In: *Cerebral Cortex* 28 (9 2018), pp. 3192–3203.

[38] Hyuk Jin Yun et al. "Regional Alterations in Cortical Sulcal Depth in Living Fetuses with Down Syndrome". In: *Cerebral Cortex* 31 (2 2021), pp. 757–767.

[39] G Malinger, D Paladini, KK Haratz, A Monteagudo, GL Pilu, and IE Timor-Tritsch. "ISUOG Practice Guidelines (updated): sonographic examination of the fetal central nervous system. Part 1: performance of screening examination and indications for targeted neurosonography". In: *Ultrasound in Obstetrics & Gynecology* 56.3 (2020), pp. 476–484.

[40] A. Youssef, T. Ghi, and G. Pilu. "How to image the fetal corpus callosum". In: *Ultrasound in Obstetrics & Gynecology* 42 (6 2013), pp. 718–720.

[41] LJ Salomon, Z Alfirevic, V Berghella, CM Bilardo, GE Chalouhi, F Da Silva Costa, E Hernandez-Andrade, G Malinger, H Munoz, D Paladini, et al. "ISUOG Practice Guidelines (updated): performance of the routine mid-trimester fetal ultrasound scan". In: *Ultrasound in Obstetrics & Gynecology* 59.6 (2022), pp. 840–856.

[42] International Society of Ultrasound in Obstetrics et al. "ISUOG Practice Guidelines (updated): performance of 11–14-week ultrasound scan". In: *Ultrasound in Obstetrics & Gynecology* 61.1 (2023), pp. 127–143.

[43] C Garel and C Alberti. "Coronal measurement of the fetal lateral ventricles: comparison between ultrasonography and magnetic resonance imaging". In: *Ultrasound in Obstetrics & Gynecology* 27.1 (2006), pp. 23–27.

[44] E Quarello, J Stirnemann, Y Ville, and L Guibaud. "Assessment of fetal Sylvian fissure operculization between 22 and 32 weeks: a subjective approach". In: *Ultrasound in Obstetrics & Gynecology* 32.1 (2008), pp. 44–49.

[45] R. Napolitano et al. "International standards for fetal brain structures based on serial ultrasound measurements from Fetal Growth Longitudinal Study of INTERGROWTH-21st Project". In: *Ultrasound in Obstetrics & Gynecology* 56 (3 2020), pp. 359–370.

[46] I. Alonso, M. Borenstein, G. Grant, I. Narbona, and G. Azumendi. "Depth of brain fissures in normal fetuses by prenatal ultrasound between 19 and 30 weeks of gestation". In: *Ultrasound in Obstetrics & Gynecology* 36 (6 2010), pp. 693–699.

[47] Sofie C. Husen, Irene V. Koning, Attie T.J.I. Go, Anne W. Van Graafeiland, Sten P. Willemsen, Irene A.L. Groenenberg, and Régine P.M. Steegers-Theunissen. "Three-dimensional ultrasound imaging of fetal brain fissures in the growth restricted fetus". In: *PLOS ONE* 14 (5 2019), e0217538.

[48] L.C. Poon, D.S. Sahota, P. Chaemsaithong, T. Nakamura, M. Machida, K. Naruse, Y.M. Wah, T.Y. Leung, and R.K. Pooh. "Transvaginal three-dimensional ultrasound assessment of Sylvian fissures at 18-30 weeks' gestation". In: *Ultrasound in Obstetrics & Gynecology* 54 (2 2019), pp. 190–198.

[49] MA Coelho Neto, P Roncato, Carolina Oliveira Nastri, and Wellington de Paula Martins. "True Reproducibility of UltraSound Techniques (TRUST): systematic review of reliability studies in obstetrics and gynecology". In: *Ultrasound in Obstetrics & Gynecology* 46.1 (2015), pp. 14–20.

[50] H. P. Robinson and J. E.E. Fleming. "A critical evaluation of sonar crown-rump length measurements". In: *BJOG: An International Journal of Obstetrics & Gynaecology* 82 (9 1975), pp. 702–710.

[51] Pekka Taipale and Vilho Hiilesmaa. "Predicting delivery date by ultrasound and last menstrual period in early gestation". In: *Obstetrics & Gynecology* 97 (2 2001), pp. 189–194.

[52] David A. Savitz, James W. Terry, Nancy Dole, John M. Thorp, Anna Maria Siega-Riz, and Amy H. Herring. "Comparison of pregnancy dating by last menstrual period, ultrasound scanning, and their combination". In: *American Journal of Obstetrics and Gynecology* 187 (6 2002), pp. 1660–1666.

[53] "Committee Opinion No 700: Methods for Estimating the Due Date". In: *Obstetrics & Gynecology* 129 (5 2017), E150–E154.

[54] Bruce Fischl. "FreeSurfer". In: *NeuroImage* 62.2 (2012), pp. 774–781.

[55] Jose Dolz, Christian Desrosiers, and Ismail Ben Ayed. "3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study". In: *NeuroImage* 170 (2018), pp. 456–470.

[56] Zeynettin Akkus, Alfiia Galimzianova, Assaf Hoogi, Daniel L Rubin, and Bradley J Erickson. "Deep learning for brain MRI segmentation: state of the art and future directions". In: *Journal of Digital Imaging* 30.4 (2017), pp. 449–459.

[57] MHG Dremmen, RH Bouhuis, LME Blanken, RL Muetzel, MW Vernooij, HE Marroun, VWV Jaddoe, FC Verhulst, Henning Tiemeier, and Tonya White. "Cavum septum pellucidum in the general pediatric population and its relation to surrounding brain structure volumes, cognitive function, and emotional or behavioral problems". In: *American Journal of Neuroradiology* 40.2 (2019), pp. 340–346.

[58] Antonios Makropoulos, Serena J Counsell, and Daniel Rueckert. "A review on automatic fetal and neonatal brain MRI segmentation". In: *NeuroImage* 170 (2018), pp. 231–248.

[59] Nadieh Khalili, Nikolas Lessmann, Elise Turk, N Claessens, Roel de Heus, Tessel Kolk, Max A Viergever, Manon JNL Benders, and Ivana Išgum. "Automatic brain tissue segmentation in fetal MRI using convolutional neural networks". In: *Magnetic Resonance Imaging* 64 (2019), pp. 77–89.

[60] Gerard Sanroma, Oualid M Benkarim, Gemma Piella, Karim Lekadir, Nadine Hahner, Elisenda Eixarch, and Miguel A González Ballester. "Learning to combine complementary segmentation methods for fetal and 6-month infant brain MRI segmentation". In: *Computerized Medical Imaging and Graphics* 69 (2018), pp. 52–59.

[61] Ali Gholipour, Caitlin K Rollins, Clemente Velasco-Annis, Abdelhakim Ouaalam, Alireza Akhondi-Asl, Onur Afacan, Cynthia M Ortinau, Sean Clancy, Catherine Limperopoulos, Edward Yang, et al. "A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth". In: *Scientific reports* 7.1 (2017).

[62] Kelly Payette, Priscille de Dumast, Hamza Kebiri, Ivan Ezhov, Johannes C Paetzold, Suprosanna Shit, Asim Iqbal, Romesa Khan, Raimund Kottke, Patrice Grehten, et al. "An automatic multi-tissue human fetal brain segmentation benchmark using the Fetal Tissue Annotation Dataset". In: *Scientific Data* 8.1 (2021), pp. 1–14.

[63] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2015, pp. 234–241.

[64] Ruobing Huang, Ana Namburete, and Alison Noble. "Learning to segment key clinical anatomical structures in fetal neurosonography informed by a region-based descriptor". In: *Journal of Medical Imaging* 5.1 (2018), p. 014007.

[65] Vishal Singh, Pradeeba Sridar, Jinman Kim, Ralph Nanan, N. Poornima, Shanmuga Priya, G. Sameera Reddy, Sathyabama Chandrasekaran, and Ramarathnam Krishnakumar. "Semantic Segmentation of Cerebellum in 2D Fetal Ultrasound Brain Images Using Convolutional Neural Networks". In: *IEEE Access* 9 (2021), pp. 85864–85873.

[66] R. Sreelakshmy, Anita Titus, N. Sasirekha, E. Logashanmugam, R. Benazir Begam, G. Ramkumar, and Raja Raju. "An Automated Deep Learning Model for the Cerebellum Segmentation from Fetal Brain Images". In: *BioMed Research International* 2022 (2022).

[67] Xin Shu, Feng Chang, Xin Zhang, Changbin Shao, and Xibei Yang. "ECAU-Net: Efficient channel attention U-Net for fetal ultrasound cerebellum segmentation". In: *Biomedical Signal Processing and Control* 75 (2022), p. 103528.

[68] Xijie Chen, Miao He, Tingting Dan, Nan Wang, Meifang Lin, Lihe Zhang, Jianbo Xian, Hongmin Cai, and Hongning Xie. "Automatic Measurements of Fetal Lateral Ventricles in 2D Ultrasound Images Using Deep Learning". In: *Frontiers in Neurology* 11 (2020), p. 526.

[69] Yuzhou Wu, Kuifang Shen, Zhigang Chen, and Jia Wu. "Automatic Measurement of Fetal Cavum Septum Pellucidum From Ultrasound Images Using Deep Attention Network". In: *IEEE International Conference on Image Processing (ICIP)*. IEEE. 2020, pp. 2511–2515.

[70] Mohammed A. Maraci, Mohammad Yaqub, Rachel Craik, Sridevi Beriwal, Alice Self, Peter von Dadelszen, Aris Papageorghiou, and J. Alison Noble. "Toward point-of-care ultrasound estimation of fetal gestational age from the trans-cerebellar diameter using CNN-based ultrasound image analysis". In: *Journal of Medical Imaging* 7.1 (2020), p. 014501.

[71] S Zeng, QC Zhou, JW Zhou, M Li, C Long, and QH Peng. "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.

[72] Giulia Babucci, Karl Rosen, Benito Cappuccini, and Graziano Clerici. "3D evaluation of fetal brain structures: reference values and growth curves". In: *The Journal of Maternal-Fetal & Neonatal Medicine* 34.21 (2021), pp. 1–6.

[73] Andrés Benavides-Serralde, Edgar Hernández-Andrade, J Fernández-Delgado, Walter Plasencia, Matthias Scheier, Fátima Crispi, Francesc Figueras, Kypros H Nicolaides, and Eduard Gratacós. "Three-dimensional sonographic calculation of the volume of intracranial structures in growth-restricted and appropriate-for-gestational age fetuses". In: *Ultrasound in Obstetrics & Gynecology* 33.5 (2009), pp. 530–537.

[74] Alexandros Sotiriadis, Ioannis Dimitrakopoulos, Makarios Eleftheriades, Theodoros Agorastos, and George Makrydimas. "Thalamic volume measurement in normal fetuses using three-dimensional sonography". In: *Journal of Clinical Ultrasound* 40.4 (2012), pp. 207–213.

[75] Juan Pedro Kusanovic, Jyh Kae Nien, Luís F Gonçalves, Jimmy Espinoza, Wesley Lee, M Balasubramaniam, Eleazar Soto, Offer Erez, and Roberto Romero. "The use of inversion mode and 3D manual segmentation in volume measurement of fetal fluid-filled structures: comparison with Virtual Organ Computer-aided AnaLysis (VOCAL™)". In: *Ultrasound in Obstetrics & Gynecology* 31.2 (2008), pp. 177–186.

[76] Julia A Scott, Kia S Hamzelou, Vidya Rajagopalan, Piotr A Habas, Kio Kim, A James Barkovich, Orit A Glenn, and Colin Studholme. "3D morphometric analysis of human fetal cerebellar development". In: *The Cerebellum* 11.3 (2012), pp. 761–770.

[77] Mustapha R Hatab, Salwa W Kamourieh, and Diane M Twickler. "MR volume of the fetal cerebellum in relation to growth". In: *Journal of Magnetic Resonance Imaging* 27.4 (2008), pp. 840–845.

[78] Nickie N Andescavage, Adre du Plessis, Robert McCarter, Ahmed Serag, Iordanis Evangelou, Gilbert Vezina, Richard Robertson, and Catherine Limperopoulos. "Complex trajectories of brain development in the healthy human fetus". In: *Cerebral Cortex* 27.11 (2017), pp. 5274–5283.

[79] Julia A Scott, Piotr A Habas, Kio Kim, Vidya Rajagopalan, Kia S Hamzelou, James M Corbett-Detig, A James Barkovich, Orit A Glenn, and Colin Studholme. "Growth trajectories of the human fetal brain tissues estimated from 3D reconstructed in utero MRI". In: *International Journal of Developmental Neuroscience* 29.5 (2011), pp. 529–536.

[80] Anastasija Arechvo, Göran Lingman, Lars Thurn, Tomas Jansson, and Ligita Jokubkiene. "Fusion imaging in brain structure measurements on a fetus phantom, combining real-time ultrasound with magnetic resonance imaging". In: *Australasian Journal of Ultrasound in Medicine* 24.3 (2021), pp. 161–172.

[81] MJ Rutten, LR Pistorius, EJH Mulder, Ph. Stoutenbeek, LS De Vries, and GHA Visser. "Fetal cerebellar volume and symmetry on 3-D ultrasound: volume measurement with multiplanar and vocal techniques". In: *Ultrasound in Medicine & Biology* 35.8 (2009), pp. 1284–1289.

[82] Toshiyuki Hata, Atsushi Kuno, Shu-Yan Dai, Eisuke Inubashiri, Uiko Hanaoka, Kenji Kanenishi, Chizu Yamashiro, Hirokazu Tanaka, and Toshihiro Yanagihara. "Three-dimensional sonographic volume measurement of the fetal cerebellum". In: *Journal of Medical Ultrasonics* 34.1 (2007), pp. 17–21.

[83] Edward Araujo Júnior, Hélio A Guimarães Filho, Cláudio R Pires, Luciano M Nardozza, Antonio F Moron, and Rosiane Mattar. "Validation of fetal cerebellar volume by three-dimensional ultrasonography in Brazilian population". In: *Archives of Gynecology and Obstetrics* 275.1 (2007), pp. 5–11.

[84] Chiung-Hsin Chang, Fong-Ming Chang, Chen-Hsiang Yu, Huei-Chen Ko, and Hsi-Yao Chen. "Assessment of fetal cerebellar volume using three-dimensional ultrasound". In: *Ultrasound in Medicine & Biology* 26.6 (2000), pp. 981–988.

[85] James H. Cole and Katja Franke. "Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers". In: *Trends in Neurosciences* 40 (12 2017), pp. 681–690.

[86] Ana I.L. Namburete, Weidi Xie, and J. Alison Noble. "Robust regression of brain maturation from 3D fetal neurosonography using CRNs". In: *International Workshop on Fetal and Infant Image Analysis*. Vol. 10554. Springer Verlag, 2017, pp. 73–80.

[87] Lok Hin Lee, Elizabeth Bradburn, Aris T. Papageorghiou, and J. Alison Noble. "Calibrated Bayesian Neural Networks to Estimate Gestational Age and Its Uncertainty on Fetal Brain Ultrasound Images". In: *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis (ASMUS and PIPPI)*. Vol. 12437. Springer Science and Business Media Deutschland GmbH, 2020, pp. 13–22.

[88] Lok Hin Lee, Elizabeth Bradburn, Rachel Craik, Mohammad Yaqub, Shane A Norris, Leila Cheikh Ismail, Eric O Ohuma, Fernando C Barros, Ann Lambert, Maria Carvalho, et al. "Machine learning for accurate estimation of fetal gestational age based on ultrasound images". In: *npj Digital Medicine* 6.1 (2023), p. 36.

[89] Xavier P. Burgos-Artizzu, David Coronado-Gutiérrez, Brenda Valenzuela-Alcaraz, Kilian Vellvé, Elisenda Eixarch, Fatima Crispi, Elisenda Bonet-Carne, Mar Bennasar, and Eduard Gratacos. "Analysis of maturation features in fetal brain ultrasound via artificial intelligence for the estimation of gestational age". In: *American Journal of Obstetrics and Gynecology MFM* 3 (6 2021), p. 100462.

[90] Teeranan Pokaprakarn et al. "AI Estimation of Gestational Age from Blind Ultrasound Sweeps in Low-Resource Settings". In: *NEJM Evidence* 1 (5 2022).

[91] Ryan G. Gomes et al. “A mobile-optimized artificial intelligence system for gestational age and fetal malpresentation assessment”. In: *Communications Medicine* 2 (1 2022), p. 128.

[92] Tingting Dan et al. “DeepGA for automatically estimating fetal gestational age through ultrasound imaging”. In: *Artificial Intelligence in Medicine* 135 (2023), p. 102453.

[93] Antonio Criminisi, Jamie Shotton, Ender Konukoglu, et al. “Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning”. In: *Foundations and Trends in Computer Graphics and Vision* 7.2-3 (2012), pp. 81–227.

[94] V. Donadono et al. “OC06.01: Development of normal fetal brain structures: international standards based on ultrasound measurements from the INTERGROWTH-21st Project”. In: *Ultrasound in Obstetrics & Gynecology* 50.S1 (2017), pp. 10–11.

[95] F. P. Hadlock, R. L. Deter, R. B. Harrist, and S. K. Park. “Estimating fetal age: computer-assisted analysis of multiple fetal growth parameters.” In: *Radiology* 152 (2 1984), pp. 497–501.

[96] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. “Definitions, methods, and applications in interpretable machine learning”. In: *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 116.44 (44 2019), pp. 22071–22080.

[97] Zachary C Lipton. “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.” In: *Queue* 16.3 (2018), pp. 31–57.

[98] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. “Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization”. In: *IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626.

[99] Nicola K. Dinsdale, Emma Bluemke, Vaanathi Sundaresan, Mark Jenkinson, Stephen M. Smith, and Ana I.L. Namburete. “Challenges for machine learning in clinical translation of big data imaging studies”. In: *Neuron* 110.23 (2022), pp. 3866–3881.

[100] Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.

[101] Bas H.M. van der Velden, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, and Max A. Viergever. “Explainable artificial intelligence (XAI) in deep learning-based medical image analysis”. In: *Med. Image Anal.* 79 (2022), p. 102470.

[102] Zohaib Salahuddin, Henry C. Woodruff, Avishek Chatterjee, and Philippe Lambin. “Transparency of deep neural networks for medical image analysis: A review of interpretability methods”. In: *Computers in Biology and Medicine* 140 (2022), p. 105111.

[103] Arjan M. Groen, Rik Kraan, Shahira F. Amirkhan, Joost G. Daams, and Mario Maas. "A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: Limited use of explainable AI?" In: *European Journal of Radiology* 157 (2022), p. 110592.

[104] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: *arXiv preprint arXiv:1312.6034* (2014).

[105] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. Association for Computational Linguistics (ACL), 2016, pp. 97–101.

[106] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)". In: *International Conference on Machine Learning (ICML)*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2668–2677.

[107] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. "Concept Bottleneck Models". In: *International Conference on Machine Learning (ICML)*. PMLR, 2020, pp. 5338–5348.

[108] "From Chest X-Rays to Radiology Reports: A Multimodal Machine Learning Approach". In: *Digital Image Computing: Techniques and Applications (DICTA)*. Institute of Electrical and Electronics Engineers Inc., 2019.

[109] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PlOS ONE* 10.7 (2015), e0130140.

[110] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks". In: *International Conference on Machine Learning (ICML)*. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 3319–3328.

[111] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. "Towards better understanding of gradient-based attribution methods for Deep Neural Networks". In: *International Conference of Representation Learning (ICLR)*. 2018.

[112] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences". In: *International Conference on Machine Learning (ICML)*. Vol. 70. PMLR. 2017, pp. 3145–3153.

[113] Rob Zeiler Matthew D. and Fergus. "Visualizing and Understanding Convolutional Networks". In: *IEEE European Conference on Computer Vision (ECCV)*. Springer International Publishing, 2014, pp. 818–833.

[114] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. "Striving for simplicity: The all convolutional net". In: *arXiv preprint arXiv:1412.6806* (2014).

[115] Weili Nie, Yang Zhang, and Ankit Patel. "A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations". In: *International Conference on Machine Learning (ICML)*. Vol. 80. PMLR, 2018, pp. 3809–3818.

[116] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. "Sanity Checks for Saliency Maps". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 31. 2018.

[117] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. "Smoothgrad: removing noise by adding noise". In: *arXiv preprint arXiv:1706.03825* (2017).

[118] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. "Not just a black box: Learning important features through propagating activation differences". In: *arXiv preprint arXiv:1605.01713* (2016).

[119] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30. 2017, pp. 4768–4777.

[120] Lloyd S. Shapley. "A value for n-person games". In: *Contributions to the Theory of Games* (1953), pp. 307–318.

[121] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning Deep Features for Discriminative Localization". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2921–2929.

[122] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[123] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks". In: *IEEE Winter Conference on Applications of Computer Vision, (WACV)*. 2018, pp. 839–847.

[124] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. "Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs". In: *The British Machine Vision Conference (BMVC)*. 2020.

[125] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. "Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2020, pp. 111–119.

[126] Andrea Vedaldi and Stefano Soatto. "Quick shift and kernel methods for mode seeking". In: *European Conference on Computer Vision (ECCV)*. Springer Verlag, 2008, pp. 705–718.

[127] Mara Graziani, Iam Palatnik de Sousa, Marley M.B.R. Vellasco, Eduardo Costa da Silva, Henning Müller, and Vincent Andrearczyk. "Sharpening Local Interpretable Model-Agnostic Explanations for Histopathology: Improved Understandability and Reliability". In: *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer International Publishing, 2021, pp. 540–549.

[128] Ruth C Fong and Andrea Vedaldi. "Interpretable explanations of black boxes by meaningful perturbation". In: *IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 3429–3437.

[129] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. "Attention gated networks: Learning to leverage salient regions in medical images". In: *Medical Image Analysis* 53 (2019), pp. 197–207.

[130] Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. "Which Explanation Should I Choose? A Function Approximation Perspective to Characterizing Post Hoc Explanations". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 35. 2022, pp. 5256–5268.

[131] Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. "Impossibility theorems for feature attribution". In: *arXiv preprint arXiv:2212.11870* (2022).

[132] Pieter Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. "The (Un)reliability of Saliency Methods". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Vol. 11700 LNCS. Springer Verlag, 2019, pp. 267–280.

[133] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. "The false hope of current approaches to explainable artificial intelligence in health care". In: *The Lancet Digital Health* 3.11 (2021), E745–E750.

[134] Hristina Uzunova, Jan Ehrhardt, Timo Kepp, and Heinz Handels. "Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders". In: *SPIE Medical Imaging*. Vol. 10949. International Society for Optics and Photonics. SPIE, 2019, p. 1094911.

[135] Nicolas Papernot and Patrick McDaniel. "Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning". In: *arXiv:1803.04765* (2018).

[136] Been Kim, Cynthia Rudin, and Julie A. Shah. "The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 27. 2014.

[137] Jacob Bien and Robert Tibshirani. "Prototype selection for interpretable classification". In: *Annals of Applied Statistics* 5.4 (2011), pp. 2403–2424.

[138] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. "Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions". In: *AAAI Conference on Artificial Intelligence*. Vol. 32. 1. AAAI press, 2018, pp. 3530–3537.

[139] Geoffrey E Hinton and Ruslan R Salakhutdinov. "Reducing the dimensionality of data with neural networks". In: *Science* 313.5786 (2006), pp. 504–507.

[140] Yann LeCun, Corinna Cortes, and CJ Burges. *MNIST handwritten digit database*. `http://yann.lecun.com/exdb/mnist`. 2010.

[141] Han Xiao, Kashif Rasul, and Roland Vollgraf. *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. https://github.com/zalandoresearch/fashion-mnist. 2017.

[142] Sanja Fidler, Sven Dickinson, and Raquel Urtasun. "3D object detection and viewpoint estimation with a deformable 3D cuboid model". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 25. 2012.

[143] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. "This Looks Like That: Deep Learning for Interpretable Image Recognition". In: *Sanity Checks for Saliency Maps*. 2019.

[144] Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. "XProtoNet: diagnosis in chest radiography with global and local explanations". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 15719–15728.

[145] Sanaz Mohammadjafari, Mucahit Cevik, Mathusan Thanabalasingam, and Ayse Basar. "Using ProtoPNet for Interpretable Alzheimer's Disease Classification." In: *Canadian Conference on Artificial Intelligence*. 2021.

[146] Gurmail Singh and Kin-Choong Yow. "An interpretable deep learning model for COVID-19 detection with chest X-ray images". In: *IEEE Access* 9 (2021), pp. 85198–85208.

[147] Gurmail Singh and Kin-Choong Yow. "These do not look like those: An interpretable deep learning model for image recognition". In: *IEEE Access* 9 (2021), pp. 41482–41493.

[148] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y. Lo, and Cynthia Rudin. "A case-based interpretable deep learning model for classification of mass lesions in digital mammography". In: *Nature Machine Intelligence* 3.12 (2021), pp. 1061–1070.

[149] Tom Nuno Wolf, Sebastian Pölsterl, and Christian Wachinger. "Don't PANIC: Prototypical Additive Neural Network for Interpretable Classification of Alzheimer's Disease". In: *International Conference on Information Processing in Medical Imaging (IPMI)*. Springer. 2023, pp. 82–94.

[150] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. "Counterfactual Visual Explanations". In: *International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 2376–2384.

[151] Christian F Baumgartner, Lisa M Koch, Kerem Can Tezcan, Jia Xi Ang, and Ender Konukoglu. "Visual Feature Attribution Using Wasserstein GANs". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 8309–8319.

[152] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 27. 2014.

[153] Arunachalam Narayanaswamy, Subhashini Venugopalan, Dale R. Webster, Lily Peng, Greg S. Corrado, Paisan Ruamviboonsuk, Pinal Bavishi, Michael Brenner, Philip C. Nelson, and Avinash V. Varadarajan. "Scientific discovery by generating counterfactuals using image translation". In: *International Conference on Medical Image Computing and computer-Assisted Intervention (MICCAI)*. Springer Science and Business Media Deutschland GmbH, 2020, pp. 273–283.

[154] Kathryn Schutte, Olivier Moindrot, Paul Hérent, Jean-Baptiste Schiratti, and Simon Jégou. "Using stylegan for visual interpretability of deep learning models on medical images". In: *arXiv preprint arXiv:2101.07563* (2021).

[155] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. "Deep Structural Causal Models for Tractable Counterfactual Inference". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 857–869.

[156] Cher Bass, Mariana Da Silva, Carole Sudre, Petru-Daniel Tudosiu, Stephen M Smith, and Emma C Robinson. "ICAM: Interpretable Classification via Disentangled Representations and Feature Attribution Mapping". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 7697–7709.

[157] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations, ICLR, 2013.

[158] Cher Bass, Mariana da Silva, Carole Sudre, Logan Z J Williams, Petru-Daniel Tudosiu, Fidel Alfaro-Almagro, Sean P Fitzgibbon, Matthew F Glasser, Stephen M Smith, and Emma C Robinson. "ICAM-reg: Interpretable Classification and Regression with Feature Attribution for Mapping Neurological Phenotypes in Individual Scans". In: *IEEE Transactions in Medical Imaging* 42.4 (2023).

[159] Jiwen Lu, Junlin Hu, and Jie Zhou. "Deep metric learning for visual understanding: An overview of recent advances". In: *IEEE Signal Processing Magazine* 34 (6 2017), pp. 76–84.

[160] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet: A unified embedding for face recognition and clustering". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823.

[161] Mahmut Kaya and Hasan Şakir Bilge. "Deep metric learning: A survey". In: *Symmetry* 11 (9 2019), p. 1066.

[162] Chien Liang Liu and Qing Hong Chen. "Metric-Based Semi-Supervised Regression". In: *IEEE Access* 8 (2020), pp. 30001–30011.

[163] Greg Mori, Caroline Pantofaru, Nisarg Kothari, Thomas Leung, George Toderici, Alexander Toshev, and Weilong Yang. "Pose embeddings: A deep architecture for learning to match human poses". In: *arXiv preprint arXiv:1507.00302* (2015).

[164] Kang Zheng et al. "Semi-supervised Learning for Bone Mineral Density Estimation in Hip X-Ray Images". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2021, pp. 33–42.

[165] Xiaonan Zhao, Huan Qi, Rui Luo, and Larry Davis. "A weakly supervised adaptive triplet loss for deep metric learning". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 3177–3180.

[166] Weihang Dai, Xiaomeng Li, Wan Hang Keith Chiu, Michael D. Kuo, and Kwang Ting Cheng. "Adaptive Contrast for Image Regression in Computer-Aided Disease Assessment". In: *IEEE Transactions on Medical Imaging* 41 (5 2022), pp. 1255–1268.

[167] Sungyeon Kim, Minkyo Seo, Ivan Laptev, Minsu Cho, and Suha Kwak. "Deep metric learning beyond binary supervision". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 2288–2297.

[168] Hanqing Chao, Jiajin Zhang, and Pingkun Yan. "Regression Metric Loss: Learning a Semantic Representation Space for Medical Images". In: *International Conference on Medical Image Computing and computer-Assisted Intervention (MICCAI)*. Springer, 2022, pp. 427–436.

[169] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. "DeepEMD: Differentiable Earth Mover's Distance for Few-Shot Learning". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.5 (2022), pp. 5632–5648.

[170] Wenliang Zhao, Yongming Rao, Ziyi Wang, Jiwen Lu, and Jie Zhou. "Towards Interpretable Deep Metric Learning With Structural Matching". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 9887–9896.

[171] Hai Phan and Anh Nguyen. "DeepFace-EMD: Re-Ranking Using Patch-Wise Earth Mover's Distance Improves Out-of-Distribution Face Identification". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 20259–20269.

[172] Angela Lombardi, Domenico Diacono, Nicola Amoroso, Alfonso Monaco, João Manuel RS Tavares, Roberto Bellotti, and Sabina Tangaro. "Explainable deep learning for personalized age prediction with brain morphology". In: *Frontiers in Neuroscience* 15 (2021).

[173] Simon M Hofmann, Frauke Beyer, Sebastian Lapuschkin, Ole Goltermann, Markus Loeffler, Klaus-Robert Müller, Arno Villringer, Wojciech Samek, and A Veronica Witte. "Towards the interpretability of deep learning models for multi-modal neuroimaging: Finding structural changes of the ageing brain". In: *NeuroImage* 261 (2022), p. 119504.

[174] Fabian Eitel and Kerstin Ritter. "Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification". In: *International Workshop on Interpretability of Machine Intelligence in Medical Image Computing*. Ed. by Springer. Vol. 11797. 2019, pp. 3–11.

[175] Kyriaki Margarita Bintsi, Vasileios Baltatzis, Arinbjörn Kolbeinsson, Alexander Hammers, and Daniel Rueckert. "Patch-Based Brain Age Estimation from MR Images". In: *International Workshop on Machine Learning in Clinical Neuroimaging*. Vol. 12449. 2020, pp. 98–107.

[176] Pedro L. Ballester, Laura Tomaz da Silva, Matheus Marcon, Nathalia Bianchini Esper, Benicio N. Frey, Augusto Buchweitz, and Felipe Meneguzzi. "Predicting Brain Age at Slice Level: Convolutional Neural Networks and Consequences for Interpretability". In: *Frontiers in Psychiatry* 12 (2021).

[177] Sebastian G Popescu, Ben Glocker, David J Sharp, and James H Cole. "Local Brain-Age: A U-Net Model". In: *Frontiers in Aging Neuroscience* 13 (2021).

[178] Felipe Moser, Ruobing Huang, Bartłomiej W Papież, Ana IL Namburete, INTERGROWTH-21st Consortium, et al. "BEAN: Brain Extraction and Alignment Network for 3D fetal neurosonography". In: *NeuroImage* 258 (2022), p. 119341.

[179] Günther Grabner, Andrew L Janke, Marc M Budge, David Smith, Jens Pruessner, and D Louis Collins. "Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2006, pp. 58–66.

[180] Satoshi Shinohara, Yasuhiko Okuda, Shuji Hirata, and Kohta Suzuki. "Predictive possibility of the transverse cerebellar diameter to abdominal circumference ratio for small-for-gestational-age fetus suspected as a cause of maternal placental syndromes: a retrospective cohort study". In: *Hypertension in Pregnancy* 39.2 (2020), pp. 145–151.

[181] T Tongsong, C Wanapirak, and T Thongpadungroj. "Sonographic diagnosis of intrauterine growth restriction (IUGR) by fetal transverse cerebellar diameter (TCD)/abdominal circumference (AC) ratio". In: *International Journal of Gynecology & Obstetrics* 66.1 (1999), pp. 1–5.

[182] Ali Gholipour, Alireza Akhondi-Asl, Judy A Estroff, and Simon K Warfield. "Multi-atlas multi-shape segmentation of fetal brain MRI for volumetric and morphometric analysis of ventriculomegaly". In: *NeuroImage* 60.3 (2012), pp. 1819–1831.

[183] Piotr A Habas, Kio Kim, James M Corbett-Detig, Francois Rousseau, Orit A Glenn, A James Barkovich, and Colin Studholme. "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.

[184] Eva Dittrich, Tammy Riklin Raviv, Gregor Kasprian, René Donner, Peter C Brugger, Daniela Prayer, and Georg Langs. "A spatio-temporal latent atlas for semi-supervised learning of fetal brain segmentations and morphological age estimation". In: *Medical Image Analysis* 18.1 (2014), pp. 9–21.

[185] A Serag, V Kyriakopoulou, MA Rutherford, AD Edwards, JV Hajnal, P Aljabar, SJ Counsell, J Boardman, and D Rueckert. "A multi-channel 4D probabilistic atlas of the developing brain: application to fetuses and neonates". In: *Annals of the BMVA* 2012.3 (2012), pp. 1–14.

[186] Mariano Cabezas, Arnau Oliver, Xavier Lladó, Jordi Freixenet, and Meritxell Bach Cuadra. "A review of atlas-based segmentation for magnetic resonance brain images". In: *Computer Methods and Programs in Biomedicine* 104.3 (2011), e158–e177.

[187] Felipe Moser, Ruobing Huang, Aris T Papageorghiou, Bartłomiej W Papież, and Ana IL Namburete. "Automated fetal brain extraction from clinical Ultrasound volumes using 3D Convolutional Neural Networks". In: *Annual Conference on Medical Image Understanding and Analysis (MIUA)*. Springer. 2019, pp. 151–163.

[188] Madeleine K Wyburd, Mark Jenkinson, and Ana IL Namburete. "Cortical plate segmentation using CNNs in 3D fetal ultrasound". In: *Annual Conference on Medical Image Understanding and Analysis (MIUA)*. Springer. 2020, pp. 56–68.

[189] Abhijit Guha Roy, Shayan Siddiqui, Sebastian Pölsterl, Nassir Navab, and Christian Wachinger. "'Squeeze & excite'guided few-shot segmentation of volumetric images". In: *Medical Image Analysis* 59 (2020), p. 101587.

[190] Arnab Kumar Mondal, Jose Dolz, and Christian Desrosiers. "Few-shot 3D multi-modal medical image segmentation using generative adversarial learning". In: *arXiv preprint arXiv:1810.12241* (2018).

[191] Dawood Al Chanti, Vanessa Gonzalez Duque, Marion Crouzier, Antoine Nordez, Lilian Lacourpaille, and Diana Mateus. "IFSS-Net: Interactive Few-Shot Siamese Network for Faster Muscle Segmentation and Propagation in Volumetric Ultrasound". In: *IEEE Transactions on Medical Imaging* 40.10 (2021), pp. 2615–2628.

[192] Ana IL Namburete, Weidi Xie, Mohammad Yaqub, Andrew Zisserman, and J Alison Noble. "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.

[193] Rémi Cuingnet, Oudom Somphone, Benoit Mory, Raphael Prevost, Mohammad Yaqub, Raffaele Napolitano, Aris Papageorghiou, David Roundhill, J Alison Noble, and Roberto Ardon. "Where is my baby? A fast fetal head auto-alignment in 3D-ultrasound". In: *International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2013, pp. 768–771.

[194] National Institute for Health and Care Excellence. *Antenatal care for uncomplicated pregnancies (NICE Guideline NG201)*. `https://www.nice.org.uk/guidance/ng201/chapter/Recommendations`. Accessed: 01-07-2023. 2021.

[195] Vanessa Kyriakopoulou, Deniz Vatansever, Alice Davidson, Prachi Patkee, Samia Elkommos, Andrew Chew, Miriam Martinez-Biarge, Bibbi Hagberg, Mellisa Damodaram, Joanna Allsop, et al. "Normative biometry of the fetal brain using magnetic resonance imaging". In: *Brain Structure and Function* 222.5 (2017), pp. 2295–2307.

[196] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International Conference on machine learning (ICML)*. PMLR. 2015, pp. 448–456.

[197] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation". In: *Fourth International Conference on 3D Vision (3DV)*. 2016, pp. 565–571.

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

[199] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. "Comparing images using the Hausdorff distance". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15.9 (1993), pp. 850–863.

[200] D Prayer, G Malinger, PC Brugger, C Cassady, Luc De Catte, Bart De Keersmaecker, GL Fernandes, P Glanc, LF Gonçalves, GM Gruber, et al. "ISUOG Practice Guidelines: performance of fetal magnetic resonance imaging". In: *Ultrasound in Obstetrics & Gynecology* 49.5 (2017), pp. 671–680.

[201] K Karl, T Esser, KS Heling, and R Chaoui. "Cavum septi pellucidi (CSP) ratio: a marker for partial agenesis of the fetal corpus callosum". In: *Ultrasound in Obstetrics & Gynecology* 50.3 (2017), pp. 336–341.

[202] Linde S Hesse, Grey Kuling, Mitko Veta, and Anne L Martel. "Intensity augmentation to improve generalizability of breast segmentation across different MRI scan protocols". In: *IEEE Transactions on Biomedical Engineering* 68.3 (2020), pp. 759–770.

[203] 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.

[204] David Alvarez-Melis and Tommi S Jaakkola. "On the robustness of interpretability methods". In: *arXiv:1806.08049* (2018).

[205] Yiqi Yan, Jeremy Kawahara, and Ghassan Hamarneh. "Melanoma recognition via visual attention". In: *Information Processing in Medical Imaging (IPMI)*. Springer. 2019, pp. 793–804.

[206] Heechan Yang, Ji-Ye Kim, Hyongsuk Kim, and Shyam P Adhikari. "Guided soft attention network for classification of breast cancer histopathology images". In: *IEEE Transactions on Medical Imaging* 39.5 (2020), pp. 1306–1315.

[207] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. "Thorax disease classification with attention guided convolutional neural network". In: *Pattern Recognition Letters* 131 (2020), pp. 38–45.

[208] Saumya Jetley, Nicholas A. Lord, Namhoon Lee, and Philip H.S. Torr. "Learn to pay attention". In: *International Conference on Learning Representations (ICLR)*. 2018, pp. 1–14.

[209] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images". In: (2009).

[210] Chong Yin, Siqi Liu, Rui Shao, and Pong C. Yuen. "Focusing on Clinically Interpretable Features: Selective Attention Regularization for Liver Biopsy Image Classification". In: *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Cham: Springer International Publishing, 2021, pp. 153–162.

[211] Madeleine K Wyburd. "Preserving known anatomical topology in medical image segmentation using deep learning". PhD thesis. University of Oxford, 2022.

[212] James H Cole, Stuart J Ritchie, Mark E Bastin, Valdés Hernández, S Muñoz Maniega, Natalie Royle, Janie Corley, Alison Pattie, Sarah E Harris, Qian Zhang, et al. "Brain age predicts mortality". In: *Molecular psychiatry* 23.5 (2018), pp. 1385–1392.

[213] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. "Understanding the effective receptive field in deep convolutional neural networks". In: *Advances in Neural Information Processing Systems (NeurIPS)* 29 (2016).

[214] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Object detectors emerge in deep scene CNNs". In: *International Conference on Learning Representations (ICLR)*. 2015.

[215] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. "Earth mover's distance as a metric for image retrieval". In: *International Journal of Computer Vision* 40.2 (2000), pp. 99–121.

[216] Harold W Kuhn. "The Hungarian method for the assignment problem". In: *Naval Research Logistics Quarterly* 2.1-2 (1955), pp. 83–97.

[217] Cédric Villani et al. *Optimal transport: old and new*. Vol. 338. Springer, 2009.

[218] Leonid V Kantorovich. "On the translocation of masses". In: *Dokl. Akad. Nauk SSSR*. Vol. 37. 1942, pp. 227–229.

[219] Marco Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 26. 2013.

[220] Solomon Kullback and Richard A Leibler. "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86.

[221] Philip A. Knight. "The Sinkhorn-Knopp algorithm: Convergence and applications". In: *SIAM Journal on Matrix Analysis and Applications* 30.1 (2008), pp. 261–275.

[222] Luis Caicedo Torres, Luiz Manella Pereira, and M. Hadi Amini. "A Survey on Optimal Transport for Machine Learning: Theory and Applications". In: *arXiv preprint arXiv:2106.01963* (2021).

[223] Helen Creasey and Stanley I Rapoport. "The aging human brain". In: *Annals of Neurology* 17.1 (1985), pp. 2–10.

[224] James H Cole, Katja Franke, and Nicolas Cherbuin. "Quantification of the biological age of the brain using neuroimaging". In: *Biomarkers of human aging* (2019), pp. 293–328.

[225] Qingyu Zhao, Ehsan Adeli, and Kilian M Pohl. "Training confounder-free deep learning models for medical applications". In: *Nature Communications* 11.1 (2020), p. 6010.

[226] Pak-Hei Yeung, Linde Hesse, Moska Aliasi, Monique Haak, Weidi Xie, Ana IL Namburete, et al. "ImplicitVol: sensorless 3D ultrasound reconstruction with deep implicit representation". In: *arXiv preprint arXiv:2109.12108* (2021).

[227] Hao Guan and Mingxia Liu. "Domain adaptation for medical image analysis: a survey". In: *IEEE Transactions on Biomedical Engineering* 69.3 (2021), pp. 1173–1185.

[228] Sheila MP Everwijn, Ana IL Namburete, Nan van Geloven, Fenna AR Jansen, Aris T Papageorghiou, Aalbertine K Teunissen, Lieke Rozendaal, Nico Blom, Jan M van Lith, and Monique C Haak. "The association between flow and oxygenation and cortical development in fetuses with congenital heart defects using a brain-age prediction algorithm". In: *Prenatal Diagnosis* 41.1 (2021), pp. 43–51.