TU WIEN Informatics

Conditional Spatiotemporal MRI Atlas of the Fetal Brain using Deep Learning

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Biomedical Engineering

eingereicht von

Johannes Tischer, B.Sc.
Matrikelnummer 12128292

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Prof. DI Dr.techn. Robert Sablatnig
Mitwirkung: Dipl.-Ing. Dr. techn. Roxane Licandro, BSc

Wien, 1. März 2024

Johannes Tischer	Robert Sablatnig

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Conditional Spatiotemporal MRI Atlas of the Fetal Brain using Deep Learning

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Biomedical Engineering

by

Johannes Tischer, B.Sc.
Registration Number 12128292

to the Faculty of Informatics

at the TU Wien

Advisor: Prof. DI Dr.techn. Robert Sablatnig
Assistance: Dipl.-Ing. Dr. techn. Roxane Licandro, BSc

Vienna, March 1, 2024

Johannes Tischer	Robert Sablatnig

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Johannes Tischer, B.Sc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 1. März 2024

Johannes Tischer

Kurzfassung

Die fetale Magnetresonanztomographie (MRT) bietet durch die hohe strukturelle Auflösung tiefe Einblicke in die Gehirnentwicklung. Die Beurteilung einer altersgerechten Entwicklung wird jedoch durch mehrere Faktoren erschwert. Dazu zählen komplexe neurologische Prozesse, individuelle Schwankungen in den Entwicklungsstadien sowie technische Unterschiede zwischen MRT-Geräten und Scan-Protokollen, die die Bildqualität beeinflussen. Zusätzlich erschwert die ungenaue Bestimmung der Schwangerschaftswoche den Vergleich mit gleichaltrigen Föten. Mit der Einführung von Atlanten wurde ein Referenzsystem geschaffen. Durch Bildregistrierung wird eine Korrespondez zwischen den neurotypischen Atlanten mit einem individuellen Föten hergestellt, das einen objektiven Vergleich sowohl innerhalb eines Individuums als auch zwischen verschiedenen Subjekten ermöglicht. In dieser Arbeit haben wir einen Datensatz mit neurotypischen fetalen Gehirnen von 308 Probanden im Gestationsalter von 21 bis 37 Wochen erstellt. Auf Basis dieses Datensatzes wurde zwei deep-learning Modelle entwickelt und implementiert, die Veränderungen abhängig von Attributen modellieren. *CAL-REG* umfasst ein CNN-basiertes Registrierungsmodell, das maximale Korrespondenz unter Einhaltung anatomischer Einschränkungen erzielt. *CAL-GAN* erweitert *CAL-REG* um einen Discriminator, der durch den direkten Vergleich zwischen generierten und echten Bildern, altersspezifische, scharfe Atlanten erzeugt, mit anatomisch realistischen Grenzen und Formvariationen. Beide Ansätze ermöglichen eine, zu den Atlasbildern zugehörige, robuste Segmentierung von sechs Gehirnregionen in Echtzeit, mit einem Dice-Koeffizienten von 85.5%. Zusätzlich erlaubt die volumetrische Analyse die Darstellung neurotypischer Wachstumstrajektorien. Die vorgeschlagene Methode ermöglicht die Generierung altersspezifischer, neurotypischer, fetaler Gehirne. Es erzielt eine Segmentierungsgenauigkeit, die mit konventionellen Verfahren vergleichbar ist, bietet dabei jedoch eine schnellere Verarbeitungsgeschwindigkeit. Darüber hinaus ist das Framework erweiterbar, beispielsweise für die individualisierte entwicklungsdiagnostische Beurteilung durch den direkten Vergleich zum Normkollektiv oder der Integration von anderen Modalitäten wie der Ultraschallbildgebung.

Abstract

Magnetic resonance imaging (MRI) of the fetal brain has emerged as an increasingly valuable tool for *in-vivo* assessment of brain development and maturation. However, assessing neurodevelopmental changes remains challenging, mainly due to inter-individual variability in brain development, different scanner and imaging protocols, and intensity inhomogeneities. Additionally, the uncertainty in determining gestational age further complicates the evaluation process, requiring profound expertise. To address these limitations, brain atlases provide a standardized reference system, enabling mapping between subjects, thus enable assessment and improved intra- and intersubject comparability. Here, we propose a continuous conditional atlas learning framework that enables both structural representation of the fetal brain and fast segmentation of new cases. Two generative deep learning models are introduced: CAL-REG, which uses a U-Net-based CNN in a direct registration approach to generate sharp, age-specific atlases, and CAL-GAN incorporating a discriminator to refine anatomical realism. The models are trained on our curated fetal brain dataset from the General Hospital of Vienna, encompassing 308 neurotypical subjects between 21 and 37 weeks of gestation. Our results demonstrate that the proposed method can generate age-specific atlases with sharp structural boundaries and realistic shape variance. In addition, the proposed approach allows robust, real-time segmentation of previously unknown subjects. Hereby, achieving a high overall Dice similarity coefficient of 85.5% across six selected tissue labels. Finally, we demonstrate how volumetric analysis of these atlases elucidates neurotypical growth trajectories, offering insights into fetal brain development. The deep learning framework proposed enables real-time, age-specific fetal brain template generation with minimal preprocessing, allowing for individualized developmental assessment. It achieves segmentation accuracy comparable to conventional approaches while operating significantly faster. Additionally, the framework allows further extensions, such as applications to specific diseases and integration with other imaging modalities such as ultrasound, expanding its potential for research and clinical use.

Contents

CHAPTER 1

Introduction

Advancements in Magnetic Resonance Imaging (MRI) acquisition, including techniques such as Single-Shot Fast Spin Echo (SSFSE), along with developments in ultra-high field strength and coil design and placement, have significantly increased the Signal-to-noise Ratio (SNR), leading to enhanced image quality and enabling advanced applications [Kab21]. Driven by these advancements, fetal MRI has emerged over the past two decades as an additional imaging modality alongside Ultrasound (US), thereby improving diagnostic accuracy and providing additional insights [RRAVTV+20].
However, fetal brain MRI encounters various non-trivial challenges, such as increased motion artifacts, caused by fetal movement and maternal breathing, necessitating the application of post-processing methods like Super-Resolution Reconstruction (SRR) [EWL+20]. In addition, the developing fetal brains vary in shape and appearance over the different Gestational Week (GW). Dynamic processes, such as cortical folding, maturing lamination, differences in cell density, and the predominate myelination of the White Matter (WM) [PKK+06], cause intensity differences and partial volume effects. Moreover, uncertainties in the determination and definition of the GW and consequently discrepancies concerning the time of appearance of anatomical structures (e.g., cortical sulcation [GCB+01]), require profound medical expertise to accurately assess neurodevelopmental changes and norms.
To address these limitations, fetal brain atlases (1) provide a common reference coordinate system for spatial normalization of individuals, enabling comparability across subjects or longitudinally over time within a subject, (2) enable intensity normalization to reduce variability due to scanner models or scanning protocols, and (3) serve as a guide for automatic brain region segmentation.
In fetal brain atlas construction, a single template cannot represent the high inter-group variability and would result in blurry templates, especially in regions with high alternation frequencies, such as the cerebral cortex [HKCD+10]. For this reason, spatiotemporal (3D + time) atlases encode both spatial and temporal variability, allowing, in comparison to

solely 3D, to better preserve anatomical variability across ages [HKCD+10, GRVA+17]. Additionally, automatic segmentation of individual subjects is provided by registration of the generated template to a subject. This process propagates the knowledge incorporated in the atlas (information about the anatomy or function) onto each individual's image. The anatomical label introduced onto each image enables measurements of the volume, shape, and intensity of the anatomical structure. Therefore, patterns of normal brain tissue growth are extracted *in utero* and relations to abnormal neurodevelopmental outcomes can be correctly identified. This, in turn, facilitates more precise consultations regarding potential prenatal interventions and treatments [RRAVTV+20].

1.1 Related Work

Over the past decade, several fetal brain atlases have been published [HKCD+10, SAB+12, GRVA+17], offering high-resolution representations of the developing brain. More recently, [UKM+23] introduced an atlas from seven different imaging modalities, including both structural and diffusion MRI, accompanied by detailed tissue parcellation. Initial applications of deep learning to fetal brain atlas construction have emerged in recent years [PCZ+21, LSM+21, ZHZ+25, DKCG+24]. However, existing approaches encounter limitations. For instance, [ZHZ+25] lacks segmentation label prediction or atlas-based segmentation of previously unseen subjects, while others [PCZ+21, LSM+21] rely on discrete GA intervals (e.g., <25, 26–28, 29–32, >33 GWs [LSM+21]), limiting temporal resolution, generalization, and clinical application.

In addition, state-of-the-art approaches require prior knowledge in the construction of the fetal brain atlas. For instance, [GRVA+17] initialized the training of age-specific segmentation maps, by fusing anatomical labels of the neonatal atlas (ALBERT) to the fetal brain. Others [PCZ+21, LSM+21], initialized the registration framework and generative model by incorporating a linear average (e.g., average of 100 samples or anatomical labels) as additional input of the model. Finally, [PCZ+21, LSM+21, ZHZ+25], required an already released atlas to normalize different brain sizes across Gestational Age (GA).

In this thesis, we address these gaps by proposing a continuous deep learning framework spanning 21–37 GW. Our approach integrates joint learning of anatomical representations and segmentation labels conditioned by age, leverages atlas-based segmentation to generalize to unseen individuals, and operates with minimal data preprocessing while requiring no anatomical priors.

1.2 Contribution of the Thesis

The contribution of this master's thesis is threefold: (1) the curation of a fetal brain imaging database for training and testing machine learning approaches (2) the construction of a novel approach for conditional fetal brain atlas learning and (3) extensive evaluation scheme to assess the performance of the method proposed.

Fetal Brain MRI Dataset Curation and Design As a first step, the database of the General Hospital Vienna (AKH Wien), containing over one thousand cases acquired between 2012 and 2022, is further extended by manual determination of sex, scanner field strength, and GWs of the fetuses. Subsequently, more than 300 neurotypical fetal MRI scans covering GWs 18-39, along with their associated attributes (sex, GW, field strength), are chosen based on radiological reports for the construction of the fetal brain atlas.

Conditional Fetal Brain Atlas Learning Inspired by the statistical diffeomorphic model VoxelMorph [BZS+18] and its extended work on conditional learning [DRGS19], [DRDG21], deformable registration and conditional template estimation are extended and implemented for fetal brains as an adversarial game, wherein the registration network acts as the generator, producing a deformation field and subsequently a template. Meanwhile, the discriminator functions as an image similarity measure, distinguishing between the generated template and the individual MRI scan of the fetus. This eliminates the requirement for a specific similarity metric, enhancing the precision related to attributes like age and gender, better fitting underlying group-wise spatiotemporal trends, and achieving improved sharpness and centrality [DRDG21].
Due to the limited amount of neurotypical fetal MRI scans, the training data is divided as follows: The majority of the available MRI scans (n $\approx$ 180) are used for training the registration framework, while the total of 34 neurotypical MRI scans (two scans of each GW) are utilized for interference of the conditional atlas.

Evaluation Strategy for Fetal Brain Atlas Learning and Segmentation Finally, the proposed conditional fetal brain atlas is evaluated and compared against state-of-the-art fetal brain atlas approaches, assessing the performance of automatic segmentation and the ability to accurately delineate areas of high developmental dynamics. Therefore, standard evaluation protocols are followed [TH15, DRDG21, PLdD+23], including average Dice Similarity Coefficient (DSC), 95th Percentile Hausdorff Distance (HD95), Volumetric Similarity (VS) for segmentation labels, and average deformation of the generated template to test dataset, Entropy Focus Criteria (EFC), average Jacobian determinant $J_\phi(p)$ to evaluate the registration framework.

1.3 Aim of the Thesis

The primary aim of this thesis is to investigate and advance the application of deep learning techniques for modeling the structural development of the fetal brain. In particular, this thesis seeks to explore the potential of end-to-end frameworks in capturing key developmental features such as morphological changes in shape and size across gestational stages. The research is guided by the following research questions:

- To what extent can the proposed conditional deep learning framework model and

predict fetal brain development, particularly in terms of structural and morphological evolution?

- What are the most effective data preprocessing strategies, and how do factors such as input image resolution (1.5 T vs. 3 T) and SRR quality influence model performance?

- How does the performance of the proposed deep learning approach compare to traditional (cpu-based) methods?

1.4 Summary of the Results

The models proposed within this thesis, CAL-REG and CAL-GAN, are trained on a curated dataset of neurotypical fetal brain MRI scans (n = 185) and are capable of visualizing continuous, age-specific representations of the developing fetal brain. The models provide smooth, regular deformations (average $|Jac_\phi| = 1.000$), achieving an EFC of 0.3, indicating a high image quality of the generated templates, as well as the warped templates into the subject space.
The integration of six anatomical labels into the framework provides further guidance for the registration framework and enables atlas-based segmentation of previously unseen individuals. Nevertheless, the traditional approach ANTs [AYP+10] outperformed all proposed models in terms of segmentation accuracy, achieving a mean DSC of 87.2 %, HD95 of 1.25 mm and VS of -0.003 across all tissue labels. In comparison, the best-performing model proposed in this thesis, CAL-GAN, achieves a DSC of 86.3 %, a HD95 of 1.60 mm, and a VS of 0.013. Inference on the test dataset (n = 34) required 16.3 hours using ANTs, whereas the proposed method completed the task in 34 seconds.

Volumetric analyses of neurodevelopmental growth patterns across GA revealed complex, non-linear trajectories that closely align with findings reported in recent literature [SAB+12, AdPM+17, UKM+23].

1.5 Outline

This thesis is structured as follows:

- **Chapter 2** begins with an overview of fetal brain anatomy and development. It then covers the fundamental principles of MRI physics and its most commonly used acquisition sequences. The chapter concludes with a discussion on the construction of fetal brain atlases and the challenges involved.

- **Chapter 3** introduces the fundamentals of image registration and medical image segmentation, with a particular emphasis on deep learning techniques.

- **Chapter 4** presents preliminary experiments, including an evaluation and discussion of commonly used loss functions, as well as an assessment of baseline model implementations.
- **Chapter 5** outlines the proposed methodology in detail, including the construction of the conditional fetal brain atlas and the evaluation strategies used. In addition, the chapter describes the curation of the fetal neurotypical dataset.
- **Chapter 6** elaborates on the conducted experiments and presents their results. It includes a quantitative evaluation of the predicted conditional templates and corresponding segmentations, visualizations of these predictions, and a qualitative assessment of the registration network.
- **Chapter 7** discusses the findings in relation to recent literature. Additionally, it addresses the limitations of the thesis and explores potential directions for future research. The chapter concludes with a summary of the thesis.

CHAPTER 2

Atlas of the Fetal Brain

In the case of in-vivo MRI, the Ground Truth (GT) of segmentation does not exist, because histological analysis cannot be performed. The most accurate approximation of a GT comes from manual segmentation carried out by one or more medical experts. However, the process of manually annotating a whole three-dimensional volume of the brain is highly time-consuming, prone to errors, and subject to both inter- and intra-variability [DGP15]. To address these challenges, automatic segmentation methods have been developed to assist manual segmentation, improving reproducibility and efficiency [DGP15].

One basic yet challenging application in fetal brain segmentation is the separation of the fetal brain from maternal tissue, a process known as fetal brain extraction [CSG+23]. This can be achieved using supervised methods that incorporate prior knowledge, a landmark or bounding box around the brain, or through completely unsupervised approaches. Further details can be found in [CSG+23].
Various methodologies have been developed to tackle the challenges associated with multi-tissue brain segmentation. These methods can broadly be categorized into intensity-based, atlas-based techniques, surface-based methods, and hybrid models [DGP15]. Intensity-based segmentation methods leverage the intensity values of image voxels to identify and separate different anatomical regions of interest. These methods include thresholding, region growing, classification, or clustering techniques [DGP15]. In fetal brain imaging, intensity-based methods are often constrained by the fact that the relative contrasts between different tissues vary spatially within a given tissue class and change over time as the fetal brain develops [HKCD+10]. As a result, intensity profiles overlap, complicating the segmentation process.
In surface-based segmentation of the brain, the focus lies on extracting the Cortical Gray Matter (cGM). Characteristics of the cortex, such as thickness, gyral, and sulcal depth, are associated with both healthy brain function and disease [DFS99, TSN+22]. By definition, surface-based segmentation is also an intensity-based segmentation technique, since the

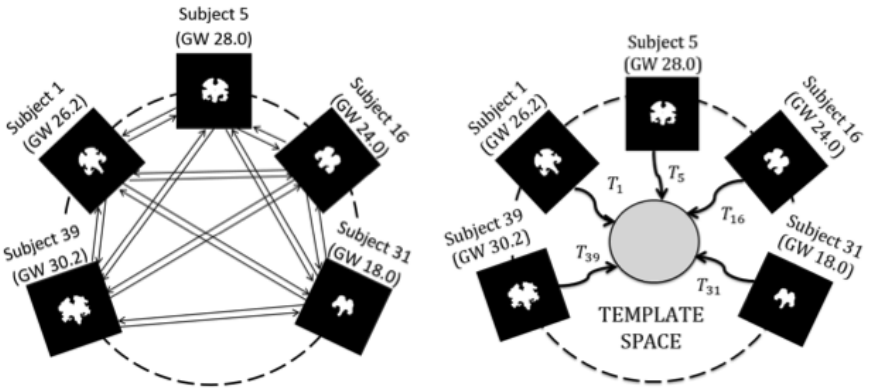


Figure 2.1: Systematic representation of **left:** pair-wise registration and **right:** group-wise registration. Figure adapted from [Lic15]

labels of the surfaces are determined by a strong gradient between the inner surface (cortex / total WM interface), pial surface (External Cerebrospinal Fluid (eCSF) and cortex interface), and central surface (geometric middle of pial and inner surface) [TRH+04]. While existing approaches target specific brain regions like the cortex [DFS99, TRH+04] or hippocampus [DKK21], a comprehensive framework for multi-tissue brain segmentation is still missing.

In contrast to the aforementioned approaches, atlas-based segmentation is not limited to specific brain regions and can, to some extent, compensate for poor image quality by leveraging prior anatomical knowledge. In the following paragraph, the principles of atlas-based segmentation are elaborated in more detail.

A "brain atlas" refers to an organized brain data collection mapped to a standardized reference space, designed to represent the anatomical structure and physiological features of the brain in a given population [CCM+24]. Much like a geographical map outlines areas and topology features, a brain atlas provides a comprehensive framework for understanding spatial relationships within brain images. Consequently, the brain atlas provides prior knowledge about the arrangements of brain regions. That prior knowledge, such as anatomical structure, can be transferred to a new, previously unknown subject [LSM+21]. This information is not limited to a modality and is leveraged to label the individual subject based on the atlas.

The variability of the brain over time is primarily addressed by introducing time-varying, or "spatiotemporal" brain atlases [GRVA+17]. This allows the longitudinal representation of the brain, such as those associated with aging or, as in our case, the development of the fetal brain.

In general, brain atlases can be differentiated into three subcategories: 1.) single-subject, 2.) multi-subject, and 3.) probabilistic atlases [CCM+24]:

1. A **single-subject** atlas corresponds to a reference system or volume image that

has been selected among an image dataset to be representative of the objects to be segmented in these. One pioneer atlas reference system was proposed by [TT88] to identify deep brain structures in stereotaxic coordinates. The information of the representative image can be transferred to an unknown subject via pair-wise registration (see Figure 2.1 left).

2. In **multi-subject atlases**, the reference system is derived from the combination of multiple subjects. Instead of registering each image to a fixed reference, groupwise registration aims to find a common reference frame or an average template that represents the group (see Figure 2.1 right). This approach avoids the bias introduced by selecting a single reference. Common applications include modeling of neonatal brain development [MSG+10], alterations caused by neurodegeneration [ZWC+18], or pathologies such as Alzheimer's disease [IILU+18].

3. Unlike multi-subject atlases, which register structural images across a population to create a final template, **probabilistic atlases** offer probability density functions (a degree of confidence) for multiple anatomical labels and each voxel [SMA+08], enabling the mapping of anatomical variability within a cohort [CCM+24].

To fully grasp the potential of a fetal brain atlas, it is essential to explore the development of the fetal brain in greater detail. Understanding the underlying data is equally important. Specifically, identifying the processes visible in the images and determining the level of detail they can reveal. To achieve this, we will first examine the fetal brain development and anatomy, followed by its respective MRI image sequences. The chapter will conclude with a discussion of methods for fetal brain atlas construction, as proposed in the literature.

2.1 Anatomy and Development of the Brain

The development of the fetal brain is a complex process involving multiple simultaneous streams. The evolution of the central nervous system can be structured into two periods. The embryonic phase (1 - 8 Post-conception Weeks (PCW)[1]), and the fetal period (9 PCW - birth, 35 to 40 PCW is considered an infant at term [Jud11]). Key processes are outlined by (1.) neural and glial proliferation, (2.) neuronal migration, (3.) neuronal differentiation (including subprocesses, such as molecular specification, neurochemical maturation, synaptogenesis, pruning and cell death and few more [KSJ19]), and (4.) cortical organization, and elaborated in the next subsections [LLSH22] [Ack92].

[1]Since the term GW can encompass three different starting points (last menstrual period, ovulation and/or fertilization, and implantation) the more precise and clinically correct term Post-conception Weeks (PCW) is used [Jud11]. PCW describes the exact timepoint when the embryo began to grow.

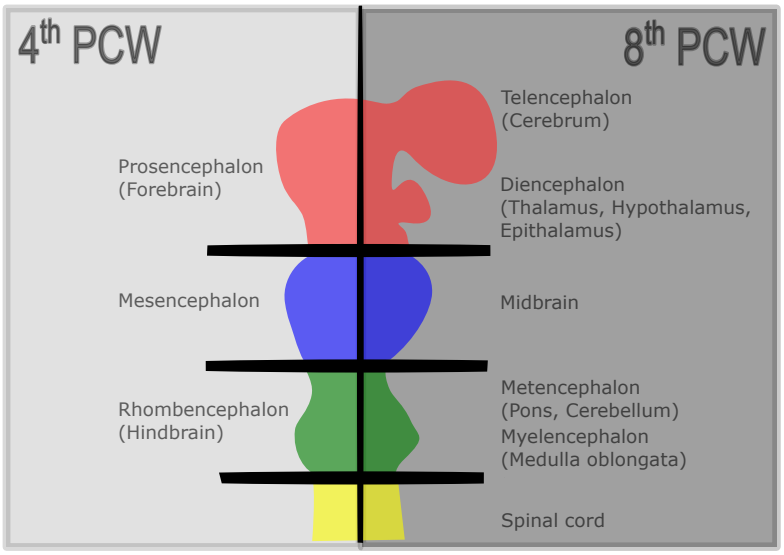


Figure 2.2: Symbolic representation of the neural tube. **Left:** three-vesicle stage reached around the fourth PCW. **Right:** the five-vesicle stage reached around the eighth PCW.

2.1.1 Embryonic phase (1 - 8 PCW)

With the end of the second PCW, the embryo is an oval-shaped two-layered structure. The upper layer, containing epiblast cells, is the origin of all structures of the developing embryo, while the lower layer, containing hypoblast cells, forms extraembryonic tissue (e.g., placenta) [Sti08].
The formation of the neural tube, a primitive three-ventricle state (see Figure 2.2) involves complex cascades of molecular and genetic signaling. It starts in the third PCW and consists of the prosencephalon (forebrain), mesencephalon (midbrain), and rhombencephalon (hindbrain) (see Figure 2.2 on the left) [APS12]. During the fifth PCW, the neural tube closes, trapping amniotic fluid within the central canal. This increases intraventricular pressure, promoting rapid brain enlargement [BSK15]. Until the seventh PCW the three primary vesicles give rise to five secondary brain vesicles [LLSH22] (see Figure 2.2 on the right):
The prosencephalon vesicle extends laterally to form the cerebral vesicles of the future cerebrum (telencephalon), while the medial portion of the prosencephalon remains small, forming the diencephalon vesicle. The mesencephalon vesicle shows minimal growth. The rhombencephalon vesicle divides into the myelencephalon vesicle, which later becomes the medulla oblongata, and the metencephalon vesicle, which forms the cerebellum and pons. Subsequently, the telencephalon enlarges significantly, connecting with the diencephalon and overhanging the mesencephalon (see Figure 2.2 on the right). The rhombencephalon and mesencephalon eventually form the brainstem, truncus encephali (cerebri) [APS12]. The eighth PCW describes the end of the embryonic phase, where the embryo is approximately 30mm in size and 2-3g in weight [Jud11].

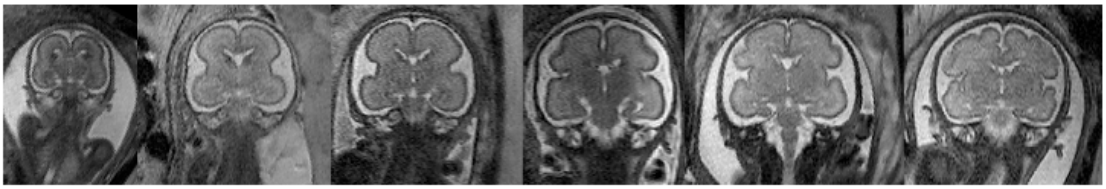


Figure 2.3: Coronal acquired T2-weighted MRI scans of the fetal brain at 20, 22, 24, 26, 28, 30 GW (from left to right) show the development of the cerebral cortex: While the cortex appears smooth at 20 GW, the complexity and size of the gyri and sulci increase over time. Image courtesy of the Medical University of Vienna.

2.1.2 Fetal period (9 PCW - birth)

The transition from the late embryonic phase to the fetal period is determined by neuronal and glial proliferation and describes the process of generating new neurons and glial cells in the developing brain [LLSH22]. Between the 12^{th} and 18^{th} PCW, most neural cells will move from the hollow, fluid-filled ventricular zone towards the outer zones and will create the cortex of the brain [Jud11]. Each migrating neuron is destined to reach its specific cortical layer and stay there for life. While tangentially migrating neurons, guided by axons, emerge from the ganglionic eminence and move toward the cerebral cortex, radially migrating neurons travel outward along glial processes that extend from the subependymal region to the cortex [Sti08]. Neuronal migration occurs in subsequent waves. Each wave of migrating neurons travels past their predecessors, resulting in the later neurons being the closest to the outer surface, finally forming the six-layered cortex at the 18^{th} PCW: the marginal zone (MZ), the cortical plate (CP), the presubplate (PSP), the intermediate zone (IZ), the subventricular zone (SVZ), and the ventricular zone (VZ) [BSK15][Jud11].
Once the neurons arrive in the cortex, they begin to differentiate and build the brain neural network through synaptogenesis, the formation of fiber pathways, and the production of neurotransmitters [Sti08]. The development of WM occurs later than the development of Gray Matter (GM) and is closely linked to the functional maturation of fiber systems [TSN+22, GS24]. Myelination of the axons begins in the third trimester of pregnancy continues postnatally into the first two years of life, and peaks around age 30 PCW [KV18, GS24]. The myelinated axons enhance the efficiency of the information transmission, mainly due to higher speed [Sti08].

Around the 24^{th} PCW, the first cortical folding of the cortical surface appears between the frontal and temporal lobes (see Figure 2.3) [BSK15]. The process of gyrification typically begins slightly earlier in the right hemisphere than in the left [BSK15]. Although the underlying neurobiological mechanisms of gyri and sulci remain unclear, evidence already suggests that larger surface areas—and therefore an increased amount of GM- are associated with greater computational power [Jud11]. Other theories indicate that gyrification is the result of unequal growth of particular brain regions or structural adaptations for species with larger brains [Jud11].

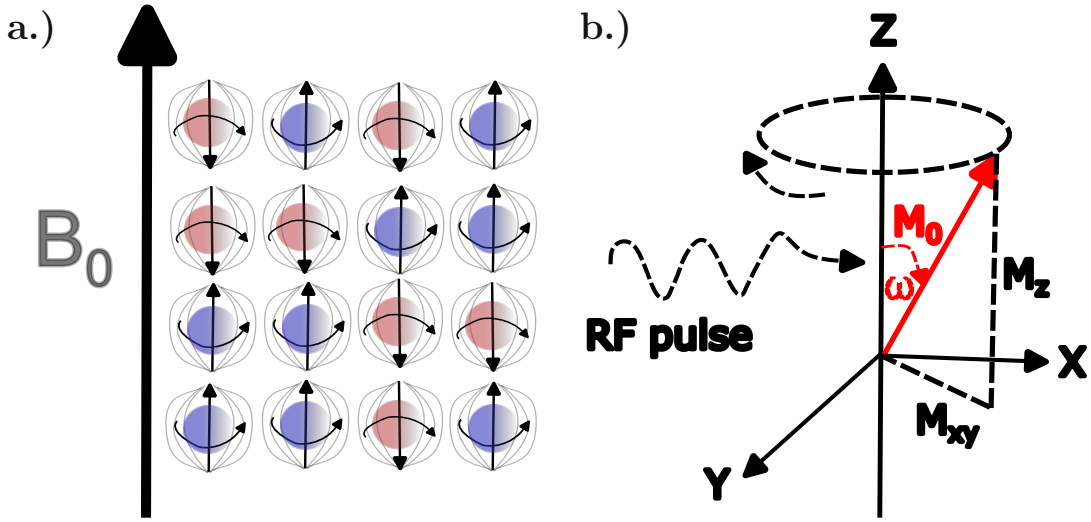


Figure 2.4: **a.)** The protons are aligned by the external magnetic field B_0 creating a net magnetization in the direction of B_0. **b.)** RF pulse flips the magnetization vector by the angle ω into the transverse plane.

2.2 Image acquisition of the fetal brain

The screening modality of choice to quantify and visualize fetal brain development and structures during pregnancy is sonography or US. US performs in real-time, its application is low-cost, and is able to detect the vast majority of fetal and placental abnormalities [WEA+16]. However, in cases when diagnostic information about potential anomalies cannot be confirmed solely by a US examination, additional fetal MRI provides further insights and improved diagnostic accuracy [RRAVTV+20].
Fetal MRI brain scans are recommended to be acquired by 18 GW (with the formation of the six-layered cortex, see Section 2.1), although their common application is between the 22 and 32 GW, when more evolved morphologic brain abnormalities can be obtained [PMDC+23]. The high resolution of MRI, compared to US, enhances tissue contrast and enables a more accurate assessment of volumetric measurements and soft tissue characteristics [PMDC+23]. This makes MRI particularly valuable in assessing pathologies such as dysgenesis of the corpus callosum, malformation of cortical development, and posterior fossa anomalies [Sal14]. While in the US - in a simplified manner - the echo of deflected sound waves indicates the underlying tissue and structures, MRI is based on the interaction of hydrogen protons with a strong, external magnetic field B_0 and Radiofrequency (RF) pulses. MRI is an active and broad research field encompassing research in physics, electrical engineering, and image processing. The following section dives into the basics of MRI physics and the fundamental fetal image sequences.

2.2.1 Physical Basics of MRI

MRI leverages the quantum mechanical property of spin, restricted to nuclei with an odd number of protons and neutrons. The high natural abundance of protons ($\approx 100\%$) and their concentration in human tissue (88 Molar) in the form of water ($^1H_2\,^8O$), allow images with good spatial resolution and quality [PMDC+23]. By nature, each hydrogen proton possesses a magnetic moment and spins around its axis with the Larmor frequency of 42.58 MHz (for magnetic field $B_0 = 1\,T$) [PK12]. Three recurring steps are needed to acquire an image based on these physical attributes. First, the hydrogen protons are aligned by a homogeneous magnetic field B_0. They can align with the field in two ways - parallel and anti-parallel - with the majority ending up in the parallel state, creating a net magnetization vector in the direction of B_0 (see Figure 2.5 a) [PK12]. The protons are now aligned and rotating at the same speed, yet not "in phase". Secondly, an RF pulse tuned to the Larmor frequency tips the nuclei, resulting in a rotation of the net magnetization vector into the transverse plane (see Figure 2.5 b). The total transverse redirection is described by the flip angle ω (cf. Eq. 2.1: $s(t)$ reaches its maximum at $\omega = \pi/2$) [PK12]. In addition to the redirection, the RF pulse also synchronizes the spins of the protons [BCH+14].
Finally, the RF signal is turned off, resulting in the relaxation of the protons into equilibrium, which can be described by two separate but simultaneously happening processes, denoted as T_1 and T_2 relaxation [PK12]. The protons re-emit the absorbed energy and realign with the magnetic field B_0. T_1 relaxation is a time constant and defines the time it takes for the longitudinal magnetization M_Z (parallel to B_0) to reach 63% of the original magnetization [PK12]. T_2 relaxation is defined as the time it takes for the spins to dephase to 37% of the original value [PK12]. There are many causes for the loss of coherence in protons, i.e., spin-spin interaction and magnetic field inhomogeneity [BCH+14]. Both, T_1 and T_2 relaxation are different for each tissue. By defining the Echo Time (TE) (time between RF pulse and data acquisition) and the Repetition Time (TR) (time between two RF pulses) as a fraction of T1 and T2 relaxation, the resulting contrast is weighted. For instance, water and Cerebrospinal Fluid (CSF) have long (2200ms) T2 values, and thus they appear bright on T2-weighted images, while fat has a short (60ms) T2 value and appears dark on T2-weighted images [PK12] (see Figure 2.3).
Finally, the total detectable signal $s(t)$ can be defined as

$$s(t) = N_{tot}\, P\, \gamma\, B\, sin(\omega t)\, cos(\omega t), \tag{2.1}$$

where N_{tot} is the total number of nuclei, P the polarization, γ the gyromagnetic ratio, B the magnetic field strength, and ω the flip angle.

2.2.2 MRI Spatial Encoding

In contrast to Computed Tomography (CT), Positron Emission Tomography (PET), or X-ray, where the spatial relation is inherent in the measured signal, MRI relies on additional encoding techniques. The following three techniques allow spatial encoding of the total measured signal $s(t)$ (cf. Eq. 2.1), each in another dimension:

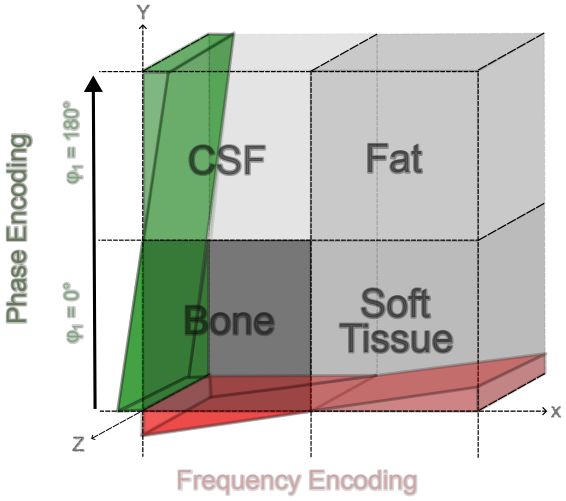


Figure 2.5: Spatial encoding in MRI: The signal of different tissues, such as bone, CSF, fat, and soft tissue are spatially encoded by slice selection, frequency encoding, and phase encoding.

Slice selecting gradient (z-axis): An additional small ($B_1 \approx 1/1e4 \times B_0$ [PK12]) magnetic field gradient denoted as B_1, is applied along the z-axis. The total magnetic field B_{tot} can be expressed by $B_{tot} = B_0 + z \cdot B_1$, with $B_1 \ll B_0$ and $z \in [-1, 1]$. According to the definition of the Larmor frequency, given by $f_L = \gamma \cdot B_{tot}$, the introduction of a magnetic gradient linearly alters the resonance frequency along the z-axis, while staying constant in the x-y plane (see Figure 2.5). This capability enables the excitation of specific areas that match the unique resonance frequencies and later correlates the measured signal back to the source location [PK12].
Frequency-encoding gradient (x-axis): When the magnetic field gradient is active, the spins resonate at different frequencies. Although the received signal is the superposition of all individual signals, the signal can be divided into its components using a Fourier transform [PK12]. Hereby, each of these oscillations is associated with its infinitesimal complex amplitude $S(f)\,df$. Summing of them leads to the inverse Fourier transform expressed in Eq. 2.2.

$$s(t) = \int_{-\infty}^{\infty} S(f) e^{j2\pi ft} d\omega \tag{2.2}$$

Phase-encoding gradient (y-axis): The G_y gradient is activated in a pulsed manner by introducing an additional magnetic field in the transverse plane. This causes protons at one end of the magnetic field to spin slightly faster than those at the other end. Consequently, when the G_y gradient is deactivated, all protons within the slice resume spinning at the same frequency but now possess different phase angles [BCH+14].

Since the general objective of the nonuniform magnetic fields is to ensure unique spatial encoding, the aforementioned concepts are not bound to a specific axis. However, since the phase encoding has to be unique for each slice, it is typically applied on the shorter

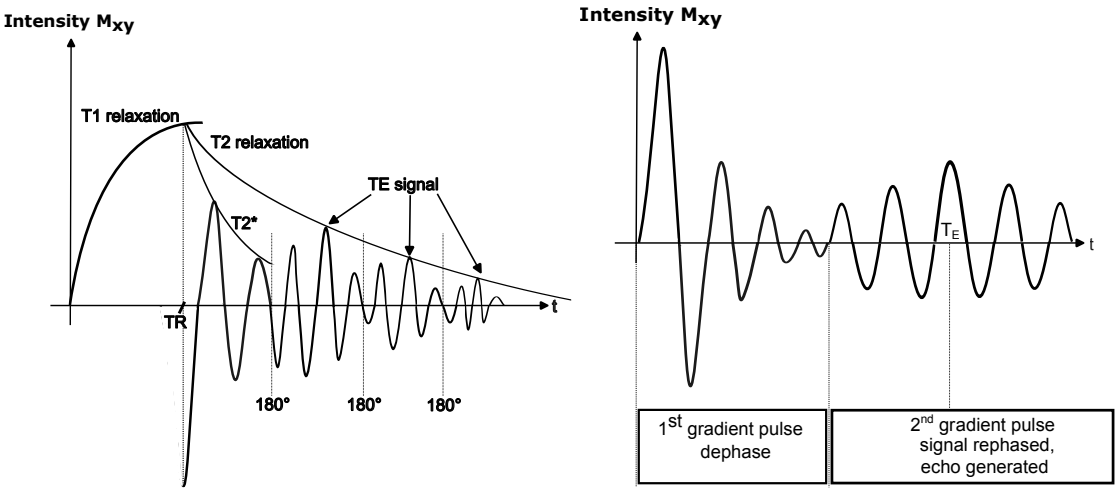


Figure 2.6: Signal intensity curve of the turbo spin echo sequence. The excitation pulse denoted as T_R results in the free induction decay, denoted as T_2^*. However, the 180° refocusing pulse leads to the "echoing" of the intensity signal, reaching its maximum at T_E. Hereby, the T_2 relaxation is prolonged. Figures adapted from [CHC+12].

axis, which allows for larger step sizes between discrete values (see Figure 2.5). The combination of all three techniques enables slice-by-slice image reconstruction from a single measured signal (cf. Eq. 2.1), with each voxel containing unique spatial data based on its frequency and phase [PK12].

2.2.3 Common fetal MRI Sequences

As elaborated in Chapter 1, MRI involves challenges associated with fetal movement and maternal breathing. To address this need, fast imaging techniques were developed and can be categorized into the following three classes: 1.) steady-state imaging, 2.) turbo-sequences, and 3.) gradient-recalled echo methods [Cle11]. The variety of imaging strategies that have been proposed is too extensive to cover in full detail in this work. Thus, we will briefly outline the concepts behind the image sequences and direct readers to the cited references for further insights [Cle11, JG19].

In **steady-state imaging**, the T_1 and T_2 relaxation (see Section 2.2.1) are interrupted by very rapid RF pulses [MTSO11, BS13]. As a result, the magnetization is prevented from returning to thermal equilibrium, and a steady state is established (after approx. $3\times T_1$). Consequently, within a couple of milliseconds, a constant image with high contrast between soft tissue is achieved [MTSO11, BS13]. Recall that with the 90° RF pulse, the net-magnetization is flipped to the transverse plane (see Figure 2.5). After the excitation, the magnetization will lose phase coherence by spin-spin interactions and magnetic field inhomogeneity [BCH+14]. To overcome this, the 90° RF excitation pulse is followed by a 180° refocusing pulse (see Figure 2.6). As a sound wave bounces back at a mountain, creating an echo, the MRI signal gets redirected by the 180° RF pulse. This pulse flips

the direction of each spin vector, so that the spins that were initially ahead in phase are now behind, and vice versa [AG16]. After the 180° RF pulse, the spins continue to evolve in the same magnetic environment as before, but their phase evolution is now reversed. Because of this reversal, the faster spins start to "catch up" with the slower ones. Over time, this refocusing causes all spins to realign [AG16].
In **Turbo Spin Echo (TSE)** sequences, this process is repeated multiple times (see Figure 2.6 b), speeding up the imaging process (30-40s) [JG19]. One limiting factor is the Specific Absorption Rate (SAR) of the fetus [Cle11], since applied RF pulses lead to heating of the underlying tissue (max 2W/kg [PMDC$^+$23]).

In **Gradient Recalled Echo (GRE)** sequences (20-25 s), an echo signal is created by an initial excitation RF pulse (flip angle $< 90°$) in combination with a gradient reversal [Els93, JG19]. Application of a gradient pulse after an initial RF pulse causes protons to rapidly dephase along the direction of the gradient, resulting in a rapid decline in the intensity signal (see Figure 2.6 a). This loss of phase coherence can be reversed by applying a second magnetic field gradient with a slope of equal amplitude but in the opposite direction to the first. As a result, protons move back into phase and return a signal called GRE [CHC$^+$12]. The advantages of GRE are the higher repetition time T_R (due to the lower flip angle) and the high sensitivity to susceptibility effects, which is valuable in detecting brain hemorrhages, microbleeds, or calcification [Els93].

2.3 History of Fetal Brain Atlases

The first spatiotemporal fetal brain atlas was developed by [HKCD$^+$10, HKR$^+$10] through polynomial fitting and non-rigid group-wise registration of 20 manually segmented neurotypical fetal brains. The resulting fetal atlas is defined for 21 to 24 GW and consists of age-specific T2-weighted atlas reference images and tissue probability label maps of the developing cGM, the developing WM, the germinal matrix, and the Lateral Ventricles (LV).
[SAB$^+$12] used free-form deformation models (see Section 3.2) as non-rigid registration in space with adaptive kernel regression in age, allowing interpolation between the subjects for heterogeneous datasets. A total of 80 reconstructed fetuses of 23-37 GW were used to build age-specific T2w atlas reference images and tissue probability label maps of five different brain structures (cerebellum, cGM, WM, brain stem, cortex).
[GRVA$^+$17] further refined this approach on fetal MRI scans and employed a semi-automatic segmentation, by fusing the labels from the neonate atlas ALBERT ([GHC$^+$13]) to the fetal MRI scans. The constructed atlas covers the age range from 23 to 38 GW.
[XSS$^+$22] created a spatiotemporal brain atlas of 90 healthy fetuses covering the GW 23 to 38. A major contribution of this paper is the special focus on the Chinese population, as well as the proposed novelties in comparison with existing atlases, including developmental trajectories and tensor-based morphometry analysis.
Most recent [UKG$^+$23] introduced a multi-channel, spatio-temporal atlas covering a consistent space between 21 to 36 GWs. The atlas was constructed based on the optimized pipeline for neonatal dHCP atlas construction [SMR$^+$18]. Therefore, 187 neurotypical

fetal MRIs from the dHCP dataset were used. The atlas impresses with a high number of multi-modal - T1w, T2w, FA, MD, RD, average DWI, ODF - channels, tissue labels of 19 different ROIs [UKM+23] separated between the left and right hemisphere, and a sharp appearance.

2.3.1 Deep Learning in Fetal Brain Atlases

Recent advancements in deep learning-based registration have demonstrated significant improvements in the quality and speed of constructing atlases ([LSM+21]). For instance, [DRGS19] introduced a novel approach, resulting in conditional, sharp atlases. Subsequently, this atlas is jointly optimized with a registration network to align individual subjects across a shared attribute and single subjects with the atlas. To ensure an unbiased atlas and improve spatial smoothness in the resulting deformation fields, abrupt changes in deformation are penalized while minimal average deformation across the entire dataset is encouraged. The work of [DRGS19] inspired further adaptation of the framework.
For instance, [LSM+21] further adapted this framework by introducing an automatic segmentation framework and application on fetal data. As in the baseline model, the deformation field is optimized by the registration network. In addition to that, the deformation field is leveraged to warp the labels from the atlas space into the subject space. The resulting atlas is constructed from 274 T2 fetal MRI scans ranging from 20.6 to 38.2 GW and covers four different intervals (<25, 26-28, 39-32, >33 GW). The segmentation of CAS-Net showed an improvement in the accuracy of small or complex structures (affected by motion or partial volume effects).
[DRDG21] included a framework based on principles of Generative Adversarial Networks (GANs). In this context, the discriminator network is tasked with distinguishing between the warped image (atlas) and the fixed image (individual scan). When the discriminator classifies the two images correctly, the generator is given another opportunity to attempt to deceive the discriminator once more. The iterative process continues until the discriminator is no longer capable of distinguishing between the synthetically generated image and the original image. This offers the advantages of accommodating datasets with significant variability, alleviating the need for an explicit similarity measure, and producing more accurate and well-defined templates [DRDG21]. However, the proposed approach was solely performed on adult brains with or without Huntington's disease [DRDG21].

Most recently, [ZHZ+25] proposed a Transformer-based architecture incorporating two submodels: 1.) An image registration framework with an Convolutional Neural Network (CNN) to extract local features, and the Vision Transformer (ViT) [CHF+21] to capture global context. 2.) The atlas construction framework, which firstly encodes all group images (sharing the same GW) in the feature space and secondly, during training, fuses the extracted features with the optimized weights. The total loss function comprises two components: the cosine distance between multi-scale features of the images and atlas, and the image loss, which is a weighted combination of the Structural Similarity

Index Measure (SSIM) and DSC of the tissue maps. Specifically, at the beginning of the training, the image loss is determined by SSIM only. Once the SSIM on the validation set reaches a threshold of 20, the contribution of the DSC loss is gradually increased. This ensures an initial intensity-based modeling with an incremental increase in the importance of morphology. The atlas was trained and evaluated on multi-centric data and achieved higher peak SNR and SSIM than pairwise or groupwise registration over all GWs.

In [CWW+22, DKCG+24], implicit neural representations are employed to learn a continuous function that captures the evolving structure of the fetal brain atlas over time. In [CWW+22], the dataset is divided into two subsets, each encompassing different discrete time points of the same time range. Although both subsets aim to learn the same developmental trajectory, they tend to overfit image noise during training, leading to different functions. In this framework, images are represented by a 4D continuous function $\hat{I}_1 = f_\theta(x, y, z, t)$, where the parameters θ are learned by Mean Squared Error (MSE) loss computation between the original and predicted (atlas) voxel intensity. The definition of a continuous function allows for generating images of novel time points ($t \notin T_{total}$). Finally, through averaging the two independently learned functions $f_{\theta_{mean}}(x, y, z, t) = \frac{1}{2}\hat{I}_1 + \frac{1}{2}\hat{I}_2$ a longitudinally-consistent atlas is achieved.
In [DKCG+24], additional anatomical labels are incorporated into the implicit neural representation, enabling conditioning on anatomical features such as ventricular volume and cortical folding. This allows for adaptation to pathologies like ventriculomegaly. Overall, the method achieved a DSC of 0.83 ± 0.04 for neurotypical brains and 0.70 ± 0.18 for ventriculomegaly.

2.4 Summary

Fetal brain development is a complex process encompassing molecular and genetic signaling (see Section 2.1). Due to the innovation of MRI and the construction of fetal brain atlases, it is possible to observe individual brain growth as well as to identify differences and similarities shared between a population. Therefore, we covered the principles of MRI and had a glimpse at the most common fast imaging sequences.
We learned that in the case of the fetal brain, an atlas constructed based on a single subject can not represent the diverse human anatomy. Thus, a spatiotemporal fetal brain atlas has to be constructed. Furthermore, we showcased significant developmental stages all observable with MRI, such as the neuronal migration, building of cortical layers around the 18^{th} PCW, or the increasing morphological changes predominant in the last trimester.
Finally, we explored contemporary literature on fetal brain atlas construction, covering both traditional methods [SAB+12, GRVA+17, UKM+23] and emerging deep learning approaches [DRGS19, DRDG21, CWW+22, DKCG+24]. Two main strategies were discussed: (1) registration-based frameworks [DRGS19], which align multiple subjects to a common space, and (2) implicit neural representations [CWW+22, DKCG+24], which model anatomy directly from data. Registration frameworks, composed of modular

components, such as the extension by segmentation maps [LSM+21], a discriminator [DRDG21], or ViT [ZHZ+25], have proven to be effective in adult brain atlas construction and are adapted for fetal applications.

CHAPTER 3

Medical Image Registration

The objective of image registration is to identify the optimum spatial transformation of two images and map the geometric correspondence, which differs in time, space, modality, or subject. By combining both images, the information aligns, allowing physicians to longitudinally compare subjects or groups. For example, the progression of Alzheimer's disease over time compared to a healthy control group [SBL+14] or the image fusion of the high contrast MRI and the high-resolution CT, guiding clinical diagnostics and decision making [ZGSQ22].
To register two images, one is typically denoted as the fixed image I_f, while the other — the moving image I_m — is transformed to align with it. In traditional image registration, the alignment process is optimized by minimizing an energy function E [CLW+25] expressed in Eq. 3.1, where $I_m \circ \Phi$ describes the transformation of I_m by the deformation field Φ. The first term quantifies the difference between the moving and the fixed image. Commonly, similarity metrics used are normalized Mutual Information (MI) [MBKB99] [SAB+12], SSIM [LRH+22], and Normalized Cross Correlation (NCC) [DRGS19][DRDG21]. More details are provided in Section 3.2.2.
The second term introduces the regularization of the deformation model, enforcing physically plausible folding, topological consistency, and anatomical constraints [CLW+25]. The regularization coefficient λ is used to balance image similarity and regularity of the deformation field.

$$\hat{\Phi} = arg\,min_{\Phi} E(I_f, I_m \circ \Phi) + \lambda R(\Phi) \tag{3.1}$$

Applied transformations can be classified into three groups. In *rigid transformations*, basic alignment is achieved through translation and rotation (6 Degrees Of Freedom (DOF)), while *affine transformations* extend this by incorporating scaling and shearing (12 DOF), making them applicable in a broader range of scenarios [CLW+25]. However, as discussed in Section 2.1, the complex anatomical and tissue composition changes in the fetal brain cannot be accurately modeled using affine transformations alone. Instead,

non-affine or *deformable transformations*, provide a more suitable approach for capturing these variations [ZGSQ22].

In deformable transformation, local areas of two images are smoothly adjusted by stretching, bending, or twisting to align image features (curves, edges, or textures) with each other. Common applications are facial expression [ASS20], organs [WDO+05], or in our case, brain structures [WDO+05, KMAS+11, FVM+21].
Traditional approaches to deformable registration formulate the problem as a mathematical optimization task, grounded in either parametric or nonparametric transformations [CLW+25]. In nonparametric models, the non-rigid deformation is simulated by a physical process such as elasticity, fluid flow, or diffusion [DMVS03]. In contrast, in parametric transformations, the transformation is determined by a linear combination of basis functions defined by control points [Boo89]. The optimization is guided by metrics, which compare the intensity of the voxels and thus determine the similarity between the images. More details are provided in Section 3.2.
With the rise of deep learning, deformable registration techniques have undergone significant transformation [BZS+18, SE19, LLC+21]. Unlike traditional methods, which solve an optimization problem for each image pair, deep learning-based approaches aim to learn a parameterized function that predicts the deformation field directly from the input images. Once trained, these models enable fast inference without the need for iterative optimization, making them well-suited for large-scale and real-time applications [DRDG21]. Further details on deep learning-based methods are provided in Section 3.3.2.

3.1 Mathematical Notation

This section introduces the notation for the mathematical variables used throughout the chapter, along with brief descriptions. Additional context is provided within each equation where these variables appear. Variables not defined here are explained at the point of use.

I_f ..... Fixed image
I_m ..... Moving image
v ..... Velocity field
ϕ ..... Deformation field
u ..... Spatial displacement
F ..... External force
μ_x ..... Mean of x
σ_x ..... Variance of x
σ_{xy} ..... Covariance of xy

3.2 Traditional Approaches in Medical Image Registration

Traditional medical image registration consists of a three-step iterative optimization process [BMPP22]. In the first step, the degree of resemblance between the fixed and moving image is evaluated by the similarity metric of choice (see Section 3.2.2). This evaluation provides an initial assessment of registration quality. In the second step, a transformation model is applied with initial parameters. The selection of the transformation method depends on the extracted features and is based on physical principles, derived from interpolation theory, or formulated from geometric models (see Section 3.2.1). In the final step, an optimization algorithm iteratively updates the transformation parameters. This process is repeated until convergence [BMPP22]. Ultimately, the algorithm yields either the final transformation parameters or a fused image with improved alignment.

3.2.1 Transformation model

The free-form deformation maps any image position from a moving image into the fixed image. The transformation can be separated into two operations: one global, the rigid or affine alignment of the two images, and one local component mapping the non-rigid deformations. The original contribution by [SP86] was further refined by [RSH+99] into the **B-spline** approach.
B-spline utilizes four points b_0, b_1, b_2, b_3 out of a subset of control points, which are uniformly distributed over the fixed image's voxel grid (see Figure 3.1). As a result, the fixed image is partitioned into equal-sized tiles. The number of required control points is $(k+1)^n$, where k describes the order of the employed spine curves and n the dimension of the image (i.e., 64 control points for a 3D image and a cubic B-spline basis function) [SKS13].
The spatial displacement u in any direction of the control point P at the position (l, m, n) can be described as (here in x-direction)

$$u_x(\vec{x}) = \sum_{i=0}^{3}\sum_{j=0}^{3}\sum_{k=0}^{3} \beta_i(a)\beta_j(b)\beta_k(c)P_x(l,m,n) \tag{3.2}$$

with the cubic spline basis function $\beta(a)$, a normalized between [0, 1], $\mathrm{l} = \frac{x}{N_x} - 1 + i$, $\mathrm{m} = \frac{y}{N_y} - 1 + j$, $\mathrm{n} = \frac{z}{N_z} - 1 + k$, and $N_{x,y,z}$ is the spacing between the control points [SKS13].

$$\beta(a) = \frac{1}{6}\begin{bmatrix} a^3 & a^2 & a & 1 \end{bmatrix}\begin{bmatrix} -1 & 3 & -3 & 1 \\ 3 & -6 & 3 & 0 \\ -3 & 0 & 3 & 0 \\ 1 & 4 & 1 & 0 \end{bmatrix}\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{3.3}$$

The B-spline is a common approach used in neonatal [KMAS+11] and fetal [SAB+12] [FVM+21] brain atlas implementations. To achieve a coarse-to-fine registration, [KMAS+11] adjusted the grid dimensions in three steps based on the subject's complexity and age. [FVM+21] defined seven anatomical landmarks that are consistently recognizable in MRI

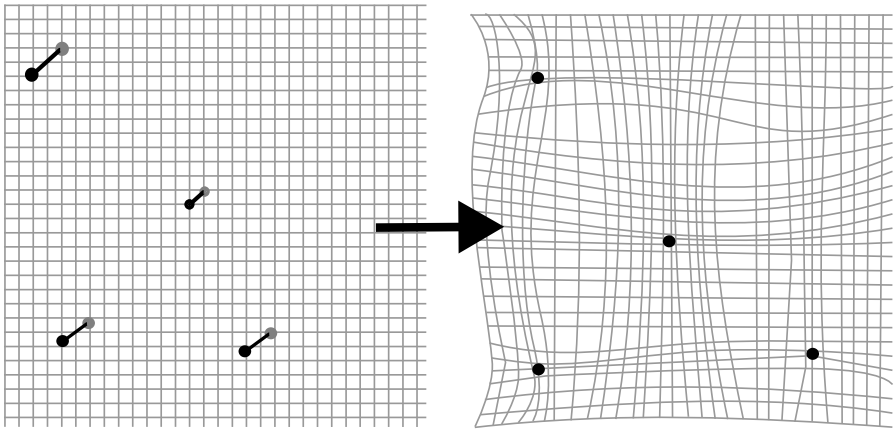


Figure 3.1: Exemplary representation of TPS image registration: **left:** Landmarks (gray points) are placed on corresponding regions in both the fixed and moving images. To align the moving image with the fixed image, the landmarks are warped (gray to black points). **right:** The resulting deformation grid represents the transformation applied to the moving image.

and across different fetal brain tissues. They further refined the registration process by incorporating the Euclidean distance between these landmarks.
The **Thin Plate Splines (TPS)** deformation [Boo89] is inspired by the physical analogy of bending thin metal. A predefined set of at least five [MBK+97] control points (landmarks) in both the fixed and moving image serves as a reference for the deformation (see Figure 3.1). The deformation is modeled under the assumption that the landmark displacements occur normally to the plane, rather than within it [Boo89]. By minimizing the bending energy, the smoothest possible deformation is obtained. TPS is a robust approach, allowing due to the control points multi-modal mapping, is diffeomorphic (invertible), and ensures smooth transformations [Boo89]. Both TPS and B-splines can handle global deformation; however, capturing local deformations at the voxel level involves a dense grid of control points, resulting in high computational complexity [DMVS03].

In **Viscous Fluid Flow Models**, the voxels of an image are regarded as viscous fluid particles, where one image gradually transforms into another [DMVS03]. The deformation is defined by the Navier-Stokes equation:

$$\frac{\partial \vec{v}}{\partial t} = \mu \nabla^2 \vec{v} + (\mu + \lambda)\vec{\nabla}(\vec{\nabla} \cdot \vec{v}) + \vec{F}(\vec{x}, \vec{u}) \stackrel{!}{=} 0 \tag{3.4}$$

with λ, μ as viscosity coefficients, the external force $\vec{F}$, $\vec{v}$ represents the velocity vector, the term $\mu\nabla^2\vec{v}$ indicates constant volume or incompressible viscous flow and the term $(\mu + \lambda)\vec{\nabla}(\vec{\nabla} \cdot \vec{v})$ controls expansion or contraction during the registration process [CT14]. Under the assumption of $\lambda = 0$ and $\mu = 1$, the deformation velocity $\vec{v}$ experienced by a particle at position x can be described by the non-linear relationship to the deformation

field $\vec{\phi}$.

$$\vec{v} = \frac{d\vec{\phi}}{dt} = \frac{\partial\vec{\phi}}{\partial t} + \sum_{i=1}^{3} \vec{v_i} \frac{\partial\vec{\phi}}{\partial x_i} \tag{3.5}$$

[DMVS03] derived an expression for a constant external force $\vec{F}(\vec{x}, \vec{\phi})$ that drives the viscous fluid flow to maximize the MI between the images. This approach enables Equation 3.4 to be solved iteratively as a sequence of linear systems [DMVS03]. The final spatial deformation u is extracted by integration over time. [CT14] examined the registration performance of the viscous fluid model using MI similarity metrics on both synthetic and real MRI data. Their results showed a high correspondence between the moving and fixed images, even for large deformations. However, the numerical solution of the model can end in a singularity, requiring re-initialization. Additionally, they reported a computational time per subject of over 1 minute. In contrast, [DMVS03] reported a registration time of 20 to 90 minutes per subject.

Diffusion Models, or *Demons Algorithm*, is inspired by Maxwell's demons. In this approach, it is assumed that a mix of two different particles is separated by a semi-permeable membrane. The membrane contains a set of "demons" that can distinguish between the two sets of particles and allow one type to diffuse. This results in a clear separation of the two sets [Thi98].
In the context of image registration, the image I_m should be registered with the image I_f. To achieve this, demons are placed along the boundaries/contours of the image I_f. Each voxel of the image I_m is labeled "inside" or "outside" based on whether it has passed through the "membrane". The demon permits voxels labeled as "outside" to move, guiding the diffusion of I_m into alignment with I_f [Thi98]. The external force F, here velocity $\vec{v}$, which drives the image registration, is defined as

$$\vec{v} = \frac{(I_{m_i} - I_{f_i})\vec{\nabla} f}{(\vec{\nabla} f)^2 + (m - f)^2} \tag{3.6}$$

where $I_{m_i} - I_{f_i}$ is the intensity difference between the images I_m and I_f and $\vec{\nabla} f$ is the gradient of the fixed image I_f [Thi98]. The external force F is close to zero when the two images are aligned ($m - f \approx 0$) or the deformation is unstable due to a missing gradient ($\vec{\nabla} f \approx 0$). [WDO+05] quantitatively assessed the registration performance of their diffusion model implementation on CT volumes from the same patients, acquired 30 days apart during radiation therapy, and compared the results with those of the TPS. The algorithm took approximately 6 minutes to align two CT volumes with dimensions of 256×256×61.

3.2.2 Optimization using Similarity Metrics

Similarity metrics in image registration are crucial for quantitatively assessing how well two images align or match after transforming one or both of them. In medical image registration with intensity-based optimization, the similarity metric provides direct feedback regarding the registration performance [HBHH01].

- Mean Squared Error is commonly used due to its probabilistic interpretation based on the Gaussian likelihood approximation [BZS+18][DRGS19]. However, by averaging the intensity values across all voxels in the image, MSE is highly sensitive to local intensity variations, such as image artifacts or contrast agents [HBHH01]. Since it works best for images varying by Gaussian noise, MSE is mainly applied in mono-modal cases [HBHH01].

$$MSE(I_m, I_f) = \frac{1}{N}\sum_{i=1}^{N}(I_{m_i} - I_{f_i})^2$$

- Normalized Cross Correlation captures the statistical correlation between images, making it robust to variations in intensity [DBGS19][LSM+21]. A higher value indicates a better correspondence.

$$NCC(I_m, I_f) = \frac{\sum(I_{m_i} - \mu_m)(I_{f_i} - \mu_f)}{\sqrt{\sum(I_{m_i} - \mu_m)^2 \sum(I_{f_i} - \mu_f)^2}}$$

- Structural Similarity Index Measure is a reliable mono-modal loss function that considers the brightness, contrast, and structure of an image. Larger SSIM indicates the higher similarity between the two images. However, like NCC, it has higher computational costs compared to MSE [CLW+25].

$$SSIM(I_m, I_f) = \frac{(2\mu_m\mu_f)(2\sigma_{mf})}{(\mu_m^2 + \mu_f^2)(\sigma_m^2 + \sigma_f^2)}$$

  where μ_m and μ_f are the mean intensities, σ_m^2 and σ_f^2 are the variances, and σ_{mf} is the covariance of images I_m and I_f.

- Mutual Information quantifies the amount of information shared between two images, moving image I_m and fixed image I_f [TH15]. It leverages entropy, which in the context of images, measures the complexity or variability within an image. The goal of maximizing MI is to ensure the images share useful information (low joint entropy) while preserving their unique individual details (high individual entropies). The MI between two images can be expressed as:

$$MI(I_m, I_f) = H(I_m) + H(I_f) - H(I_m, I_f) = \sum_a \sum_b p(I_m, I_f) \log_2 \left(\frac{p(I_m, I_f)}{p(I_m)p(I_f)} \right) \tag{3.7}$$

  Here, $H(I_m)$ and $H(I_f)$ represent the individual entropies of the moving and fixed images, respectively, while $H(I_m, I_f)$ is the joint entropy [HBHH01]. The variables a and b refer to the voxel indices in I_m and I_f, or to the brain regions in the case of segmentation labels. MI is commonly used in multi-modal imaging applications [MBKB99] [DMVS03]. For instance, it helps to align MRI and PET scans by ensuring they share critical information, such as structural details from MRI and metabolic activity from PET.

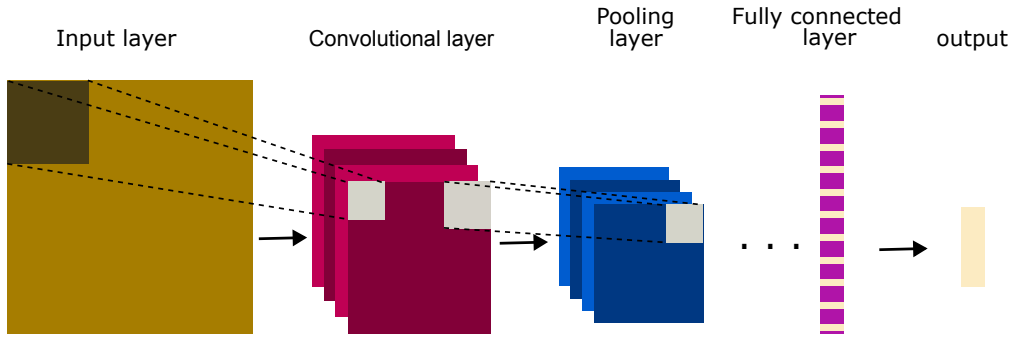


Figure 3.2: Generic representation of a Convolutional Neural Network

3.3 Deep Learning in Medical Image Registration

To better understand recent advancements in medical image registration, the fundamentals of CNNs are examined, with a focus on their composition and functionality. The most prominent contemporary network adaptations are then introduced, along with their recent applications in medical image registration. The interested reader is referred to [Sch15] for a comprehensive overview of key historical milestones in neural network development throughout the 20th century and early 2000s.

3.3.1 Convolutional Neural Networks CNNs

The core structure of a CNN, or ConvNet, is specified by three modules: convolutional layer, pooling layer, and fully connected layer (see Figure 3.2) [KSH12]. As a first step, the input image is passed to a convolutional layer. In these layers, key features of the image, such as edges, gradients, structures, or patterns, are independently detected by multiple convolutional filters. After that, the image is downsampled by pooling layers, facilitating the extraction of additional information with lower computational complexity. By repeatedly applying convolution and pooling, the process captures features that become progressively more complex and abstract. At the end of the network, the feature maps are flattened and handed over to Fully Connected (FC) layers. Here, the learned features and information are combined, leading to a final decision, a prediction, or classification [IGMA18]. The most commonly used image feature extraction models and corresponding adaptations are the LeNet, AlexNet, ResNet, and GoogleNet architectures. [KSH12].

Convolutional layer

A convolutional layer is a fundamental component of CNNs, guiding the process of extracting spatial patterns and hierarchical feature maps. In a convolutional layer, only a fraction of the input neurons is connected to the output neurons. This sparse connectivity minimizes memory requirements and makes computations more efficient compared to fully connected networks [IGMA18].

Feature maps are obtained through a set of convolutional filters (kernels) that slide horizontally or vertically across the input by a predefined step size (stride), computing dot products between the filter values and the corresponding input regions [IGMA18]. This process, known as the convolution operation, continues until no further sliding is possible. Hereby, the important local features are retained, including edges, textures, and shapes, while reducing the amount of raw data [IGMA18]. Additionally, stacking multiple convolutional layers enables a network to learn increasingly abstract and meaningful representations [IGMA18].

Pooling layer

The pooling layer typically follows the convolutional layer and downsamples the feature maps output. This significantly reduces the number of trainable parameters and once a feature is detected, the exact location of the feature becomes less important [IGMA18]. A pooling operation consists of a window selecting a part of the convolutional features, where inside of the window average pooling (average of all elements) or a max pooling (max value of all elements) operation is applied [IGMA18].

Activation function

Activation functions, such as Rectifier Linear Unit (ReLU) or Tanh, are applied after the convolutional layer and commonly used to introduce a non-linearity, allowing to learn complex correlations [IGMA18].
The Sigmoid function (cf. Eq. 3.8) maps the input value to the interval 0 to 1. Its main application is binary classification.

$$s(x) = 1/(1 + e^{-x}) \tag{3.8}$$

Tanh (cf. Eq. 3.9) is a hyperbolic tangent function, which converges faster than the sigmoid function. However, both the sigmoid and Tanh functions share a common drawback that their gradients approach zero. This occurs because they exhibit a steep increase around zero, but remain nearly constant elsewhere.

$$t(x) = (1 - e^{-2x})(1 + e^{-2x}) \tag{3.9}$$

ReLU is zero when the input x is less than zero, and a linear function for x is greater than zero. The constant gradient in ReLU allows faster training in comparison to saturating functions like the Sigmoid or Tanh [KSH12]. To prevent the *Dead* ReLU problem, where a neuron becomes inactive due to a zero gradient and no weight updates are performed, a small gradient ($\alpha = 0.1$) is applied to values less than zero, a technique known as leaky ReLU [IGMA18].

$$r(x) = max(\alpha x, x) \tag{3.10}$$

The Softmax function is the multiclass extension of the sigmoid function.

$$m(x_i) = e^{x_i}/\Sigma_j\, e^{x_j} \tag{3.11}$$

Fully connected layer

In Fully Connected layers, each neuron is connected to every neuron in the next layer. By accumulating the convolution results of the previous layers, a final prediction is formed. In other words, the dot product between the weight vector and the input vector reduces the high-dimensional features to a single prediction [IGMA18]. Multiple FC layers allow to learn non-linear, abstract interactions between features. According to [YMD+15], the FC layers account for more than 90% of the total parameter volume.

Learning Algorithm

In gradient-based optimization algorithms, the training error is minimized by iteratively updating the model's parameters at each epoch. To achieve this, the gradient (slope) of the objective function is computed during training, which guides the adjustment of the model's parameters. The updated parameters are then propagated backward from each neuron to all neurons in the preceding layer [IGMA18].

$$w_{ij_t} = w_{ij_{t-1}} - \Delta w_{ij_t}, \quad \text{where} \quad \Delta w_{ij_t} = \eta \frac{\delta E}{\delta w_{ij}} \tag{3.12}$$

where $w_{ij\,t}$ represents the weight at the current epoch, and $w_{ij\,t-1}$ is the weight from the previous epoch. The term $\Delta w_{ij\,t}$ denotes the weight update, where η is the learning rate and E is the prediction error.

Commonly used techniques include (stochastic) Gradient Descent (GD) and the Adaptive Moment Estimation (ADAM) optimization [KB15]. In comparison to GD, ADAM is more computationally efficient and outperforms large datasets [KB15]. By leveraging the first two moments, the learning rate is adaptively calculated for each parameter in the model, further enhancing accuracy and training speed [IGMA18].

3.3.2 State-of-the-art of Medical Image Registration

Traditional registration methods are based on iteratively solving an optimization process, by minimizing the geometrical distance between the fixed and moving image, while maximizing the similarity (see Section 3.2). These methods are effective, yet limited in clinical practice due to intensive and time-consuming computations. With the introduction of CNNs [KSH12] (see Section 3.3), deep learning-based image registration methods propose a fast and reliable alternative to traditional methods. The U-shaped network (U-Net) architecture was developed by [RFB15] and consists of two parts, an encoder and a decoder. The encoder extracts high-level feature representations from the input image, while the decoder reconstructs a high-resolution segmentation map from the latent space. Skip connections between the encoder and decoder ensure that the most important features are retained and directly passed to the decoder. The output of the U-Net is a segmentation map, where each pixel is classified into a specific category.
The most commonly used model adaptations are 3D U-Net, ResU-Net, V-Net, nnUnet, and Attention U-Net.

With the introduction of **CNNs**, direct registration networks emerged, enabling the direct prediction of the deformation vector field [dVBV+17, BZS+18, MR23]. [SDVB+17] was among the first to integrate an encoder-decoder architecture within a supervised registration framework. [dVBV+17] introduced the first CNN-based and Spatial Transformer Network (STN)-based image registration framework DLIR, presenting an unsupervised, end-to-end approach that achieved accuracy comparable to conventional registration methods while significantly reducing computation time. [BZS+18] proposed VoxelMorph, an unsupervised U-Net-based [RFB15] model that learns a displacement map end-to-end from image pairs. [DBGS19] extended their original framework to VoxelMorph-diff by integrating segmentation maps and surface points into the registration process, providing additional structural guidance. More recently, [MR23] extended this approach by estimating region-specific measurements such as cortical thickness.

In multistage registration networks, direct registration is performed multiple times by subdividing the process into multiple tasks. For instance, an initial CNN conducts affine registration, followed by multiple CNNs for coarse-to-fine registration, where each network progressively refines local deformations. [dVBV+19] adapted this approach for DLIR by varying the grid space and image resolution of B-spline transformation parameters. Although cascading networks enhance accuracy, their high parameter counts significantly increase computational cost and processing time.

To address this need for reduced parameter counts, [ZPC+20] proposed a feature pyramidal network architecture. In their approach, a residual deformation field between the moving and fixed images is estimated by the convolutional output of their features, where each network refines the upsampled deformation field of the previous resolution. The final deformation field is then obtained by summing the residual deformation fields, with each field weighted accordingly. This method was incorporated into VoxelMorph, resulting in improved registration accuracy compared to the standard VoxelMorph model [ZPC+20].

Generative Adversarial Networks [GPAM+14] are an unsupervised generative approach, designed to model the representation of an image based on the distribution of the training dataset. They utilize an adversarial learning framework, consisting of a generator and one or more discriminator networks. The objective of the generator is to generate realistic-looking images that are considered to be classified as real by the discriminator, while the discriminator's aim is to distinguish between generated and real images as accurately as possible. The min-max objective for GANs can be formulated as follows:

$$\min_G \max_D V_{\text{GAN}}(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]. \tag{3.13}$$

The generator G takes a random input $x \sim p_x(x)$, where $p(x)$ represents a (Gaussian) probability distribution. The generator uses this input to create a sample that resembles one drawn from the data distribution $p_{data}(x)$ [GPAM+14]. The discriminator D tries to identify the real images from the generated ones.
Therefore, in adversarial learning, an explicit image similarity metric is not strictly

necessary [ZPIE20]. GANs improve medical image registration through image-to-image (pix2pix) translations, allowing to align and translate anatomical maps between different modalities [AJF+20]. Furthermore, GANs ensure consistency in bidirectional transformations by symmetric learning. CycleGAN [ZPIE20] and further adaptations [ZSJ+22] enable image translation between domains while preserving structural integrity. Therefore, enforcing inverse consistency in the deformation field. Finally, GANs enhance registration with adversarial strategies, refining alignment using adversarial loss, and joint training, where registration is learned alongside tasks such as segmentation.
Despite its promising applications, lots of research is targeted at addressing the challenges of adversarial learning, including training instabilities [ACB17][MLX+17], model collapse, where the generator fails to capture data diversity, high sensitivity to hyperparameters [DJM22], and higher image quality [MLX+17][MK18].

Similar to GANs are **Diffusion models**, generative models as well, and can be classified into two groups [SE19, HJA20]. The variational approach primarily uses variational inference to approximate the target data distribution. Denoising Diffusion Probabilistic Models (DDPMs), introduced by [HJA20], gradually add noise to data through a Markov chain and then learn to reverse this process via a denoising network. In contrast, the score-based approach, introduced by [SE19], leverages Noise Conditional Score Networks (NCSNs) and Stochastic Differential Equations (SDEs) to model data distributions.
DiffuseMorph is an adaptation of DDPMs [KHY22]. Like other deep learning architectures [dVBV+17, DRGS19], it combines two key components: a generative network, which in this case is a diffusion model and a deformation network. Both components are jointly trained end-to-end, enabling the generated image to be mapped to the subject space via an STN. In brain MRI image registration, DiffuseMorph achieves performance comparable to state-of-the-art frameworks [KHY22]. FSDiffReg [QL23] introduced further improvements to the model by utilizing the intermediate features of the diffusion model to enhance the registration process. Additionally, the diffusion model's score function is employed to preserve deformation topology [QL23].

The introduction of the self-attention mechanism in **Transformers** [VSP+17] for natural language processing (NLP) allows the model to identify and capture long-range dependencies within sequential data. Therefore, the input data is transformed to a Query-Key-Value (QKV) model [VSP+17]. The dot product of Queries Q and Keys K are used to determine an attention score AS, which measures the importance of a given data point or feature, while values V are weighted according to these scores [VSP+17].

$$AS(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{3.14}$$

where d_k defines the dimension of the keys. With the first adaptation to computer vision (Vanilla-ViT) [DBK+21] improvements, were proposed, including shifted Window-Transformer (SwinT) [LLC+21], Image Processing Transformer (IPT), Efficient Transformer (ResT), and Graph-based Transformer (GraphT) [REYT24].
Unlike traditional transformer architectures, SwinT processes the input image as a set of

patches, enabling localized and parallel processing while ensuring computational efficiency and scalability for larger image sizes [LLC+21]. TransMorph is a hybrid approach of image registration combining SwinT in the decoder and CNNs in the encoder [CFH+22]. The feature maps extracted from SwinT are initially upsampled and then combined with feature maps from skip connections before being processed by convolutional layers. Their approach achieved significantly better performance than other CNN-based registration models [CFH+22].

For a comprehensive review of recent deep learning approaches in medical image registration, including ViT, neural Ordinary Differential Equations (ODEs), and Implicit Neural Representation (INR), the interested reader is referred to [CLW+25] and [REYT24].

3.4 Summary

This chapter provided a concise overview of medical image registration, emphasizing both traditional optimization-driven methods and deep learning approaches. It briefly explained the underlying physical principles behind traditional techniques. For deep learning, the chapter offered an introduction to the key modules involved in neural networks. The chapter concluded by discussing the current state-of-the-art deep learning applications for medical image registration, highlighting its use in fetal brain atlas construction.

In summary, the chapter highlighted the significant difference in computational resources and time between optimization-based methods (e.g., viscous fluid flow requiring 20-90 minutes, and diffusion models taking about 6 minutes) and the real-time performance of deep learning approaches. Given the vast variety of deep learning-based image registration methods, the CNN-based framework, with its modular extensions as discussed in Section 2.3, allows for flexible adaptation to the specific requirements of fetal brain atlas construction.

CHAPTER 4

Baseline Model Implementation and Preliminary Experiments

This chapter presents the preliminary experiments conducted as part of this thesis, with a particular emphasis on the implementation and evaluation of baseline image registration algorithms. Two state-of-the-art approaches are explored: 1) VoxelMorph [DRGS19], a deep learning-based image registration framework (see Section 3.3.2), and 2) ANTs [AYP+10], a traditional optimization-driven method. The goal of implementing and evaluating these frameworks is to gain a comprehensive understanding of their performance, establish a benchmark for further analysis, and explore potential adaptations for conditional fetal brain atlas construction.

To achieve this, multiple experiments are conducted, each addressing a specific research question and focusing on key aspects of the baseline approaches. The first set of experiments primarily investigates the registration subnet of the VoxelMorph framework. Therefore, the impact of different image similarity metrics on the registration performance is evaluated. Subsequently, an additional experiment examines the regularization components of the loss, introducing anatomical constraints within the generation and registration framework.
The chapter concludes with an analysis of the cpu-based, conventional image registration framework ANTs. Here, its applicability for fetal brain atlas construction is evaluated, along with its effectiveness in segmenting previously unknown subjects.

To conduct these experiments, a preliminary dataset was designed. First, the fetal brain scans undergo preprocessing, including Neuroimaging Informatics Technology Initiative (NifTI) conversion, bias field correction, and SRR using NeSVoR [XMG+23], ensuring an isotropic voxel spacing of 1 mm. A detailed description of the preprocessing steps can be found in Section 5.4.1. After preprocessing, the images are normalized to a range of [0,1] and cropped to a fixed size of 128×128×96, based on the largest brain present in

the dataset. Finally, the scans from 21 to 36 GWs are categorized into either four age groups (21-24, 25-28, 29-32, 33-36) or into eight groups (21-22, 23-24, 25-26, 27-28, 29-30, 31-32, 33-34, 35-36) respectively.

4.1 VoxelMorph

As discussed in Chapter 3, VoxelMorph is a widely used deep-learning-based algorithm for image registration. A notable application of this method is the atlas constructed by [DRGS19], which utilizes the VoxelMorph registration framework and adult brain data to generate sharp brain templates for ages 15 to 90 years. While these findings are promising, applying the same methodology to fetal brain scans introduces new challenges. Differences in anatomical structure, rapid developmental changes (see Section 2.1), and varying image quality necessitate a careful evaluation of VoxelMorph's suitability in this context. This section explores the key considerations and potential adaptations required to extend VoxelMorph's capabilities from adult to fetal brain imaging.

VoxelMorph consists of two main components: an atlas generation network and a registration network. The network takes as input the subject-specific SRR along with the GA. The atlas generation network includes a decoder that, conditioned on the GA, predicts an age-appropriate brain template. To guide this prediction, the decoder's output is added to a linear average (average of 100 training samples). The registration network then takes the predicted template and the SRR (corresponding to the same GA) as input. It estimates the spatial alignment between the generated template and the fixed image, effectively registering the two.
Since the entire architecture is trained end-to-end, both the atlas generation and registration networks are optimized simultaneously. While the loss function addresses the registration task by encouraging accurate alignment, the atlas generation network is indirectly optimized through constraints on the deformation fields (see Section 4.1.2 for more details). By limiting the extent of allowed deformations, the model is forced to produce templates that already closely resemble the target image, thereby guiding the atlas generation process.

4.1.1 Experiment 1: Image Similarity

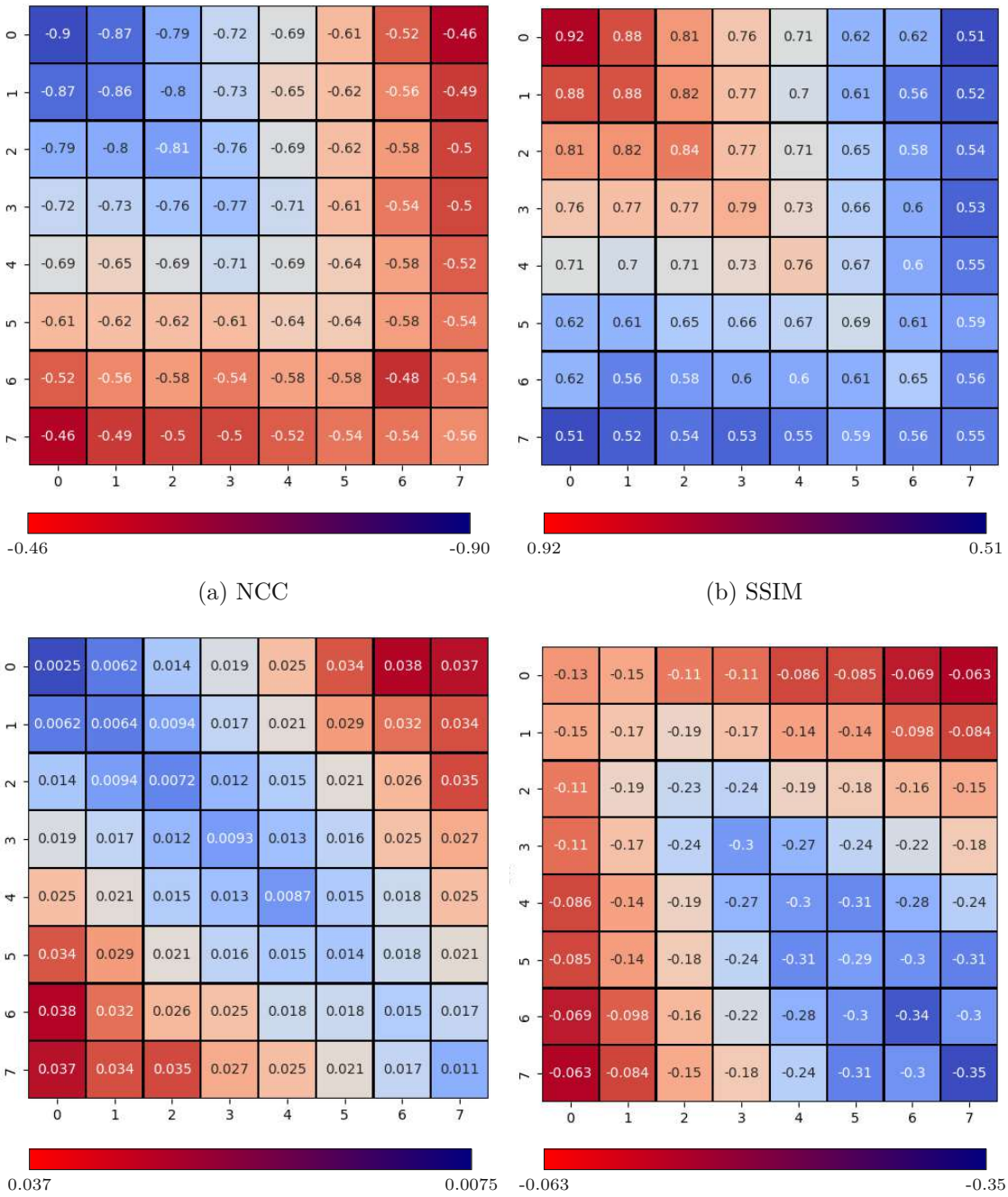
In VoxelMorph, the image loss $\mathcal{L}_{sim}$ is calculated between the framework generated template t, warped into the subject space by the deformation field ϕ, denoted as $g_{t,\theta_t}(a_i) \circ \phi_{v_i}$, and the individual MRI scan I_i [DRGS19]. Hereby, the loss function measures the similarity between the warped image and the individual scan, providing feedback on the registration performance. It should effectively capture both similarity and dissimilarity, ensuring a strong gradient between brains sharing different attributes. In this preliminary analysis, we assess the performance of image similarity metrics for fetal brains, which will later guide our final framework implementation.

Research Question:
"Which loss function most accurately captures and distinguishes structural similarities in fetal brain development across GWs?"

Method:
Four different loss functions are defined and evaluated based on their performance to distinguish between in-class and out-of-class cases. Fetal brain scans between 21 and 36 GW are clustered in eight classes and compared to each other by calculating the loss function. Therefore, four loss functions, namely NCC, SSIM, MI, and MSE, are selected. Detailed information about the loss functions is elaborated in Section 3.2.2.
For each iteration, two samples of the dataset are randomly selected, and the loss function is calculated. To reduce the impact of the heterogeneous sample size of each class and the varying image quality, the experiment is conducted 1000 times. The results are then averaged over all classes and visualized using a heat map.

Results:
The results of the four image similarity metrics are visualized in Figure 4.1. The diagonal entries of the matrix represent comparisons between cases within the same GW range, while the entries immediately adjacent to the diagonal correspond to cases with neighboring GW values. To preserve the original scale of the loss function, the values have not been normalized between zero and one. Nevertheless, the color scheme remains consistent, with red indicating the highest values and blue the lowest.
Both MSE and NCC exhibit a pronounced gradient in performance, with a decline when transitioning from directly matched classes to neighboring ones. In contrast, MI shows limited differentiation between closely related cases, and SSIM displays a small gradient (0.51 to 0.65), particularly for subjects older than 33 GW (see Figure 4.1 age classes 6 and 7). For NCC, image loss performs best in younger subjects (age classes 0 to 3). For instance, NCC reaches -0.9 when GW = 0 but increases to -0.77 for slightly higher values, while older classes yield values between -0.56 and -0.69. In contrast, MI and SSIM perform best in older GW ranges.
Among the evaluated loss functions, MSE consistently achieves the best results when comparing directly matched classes, exhibiting the strongest gradient.

Discussion:
The computed loss in earlier (21 to 26 GW) intervals is predominantly influenced by the background. In later GW periods, the nearly exponential increase in brain volume shifts the focus of the loss calculation toward the brain itself, thereby capturing the developmental processes such as lamination and gyrification (see Figure 4.1b). This shift, combined with the lack of registration among brain images — which results in considerable variability — leads to a reduction of the loss function.
More important than a small loss value within the same age group is a strong gradient between age classes, such as the blue diagonal indicated by MSE. Given the superior performance of the MSE and NCC loss functions for fetal brain imaging, both loss functions are integrated into the selected framework.



(a) NCC (b) SSIM

(c) MSE (d) MI

Figure 4.1: Heat map displaying mean loss values for fetal brain scans (21–36 GW) across eight classes (21–22, 23–24, 25–26, 27–28, 29–30, 31–32, 33–34, 35–36), using a.) NCC, b.) SSIM, c.) MSE, and d.) MI. Values were averaged over 1,000 iterations to account for class size variability and image quality differences.

The limitations of this experiment entail that random samples of the same class are compared with each other, neglecting any intermediate steps, such as registration. The experiments focused on simulating an initial training stage of the VoxelMorph framework (see Section 4.1), where age-appropriate templates are created, and the registration framework has not been able to deform the template. In all stages of the training, the image similarity metric should be able to distinguish between in-class and out-of-class image features. In other words, the objective of the loss function is to differentiate between various brain images while ensuring sufficient accuracy to guide the framework.

4.1.2 Experiment 2: Regularization

A loss term in medical image registration balances two objectives: 1) the image similarity term, which maximizes the correspondence between images and anatomical plausibility, and 2) the regularization term, which enforces realistic or anatomically valid deformations (see Chapter 3).
In VoxelMorph, three regularization terms are employed that constrain the deformation field, penalizing abrupt or irregular changes. This creates a twofold optimization: while the similarity loss prioritizes precise alignment (potentially encouraging overly aggressive deformations), the regularizers limit deformation magnitude and enforce smoothness. In other words, the balance ensures high alignment accuracy without compromising anatomical validity, as deformations remain small and physiologically coherent.
In the case of fetal brain atlas learning, this process is crucial, since the atlas generation and the registration/segmentation of an unknown subject are trained simultaneously. A loss function with only a single component fails to achieve both objectives, leading to either age-appropriate templates or accurate registration/segmentation of previously unseen subjects, but not both. In addition, the different regularizers are weighted by (sensitive) hyperparameters. Therefore, this experiment is twofold. First, the capabilities of the proposed regularizers are explored, and second, the optimal balance between the regularizers is determined.

Research Question:
What is the influence of the regularization terms on image registration, and how can these be used to introduce anatomical plausibility? Which regularization terms are suitable in the case of fetal brain atlas learning?

Method:
The aforementioned regularization terms are defined by: 1.) $\mathcal{L}_{SMOOTH}$, enforcing smoothness by discouraging abrupt local changes. 2.) $\mathcal{L}_{DEF}$, preventing excessively large displacements by constraining the deformation magnitude, and 3.) $\mathcal{L}_{MOV}$ constraining the mean deformation magnitude, ensuring global coherence and stability [DRGS19]. Their mathematical representation is given by

$$\mathcal{L}_{DEF} = \frac{d}{2}\lambda_d \Sigma_i ||u_i||^2 \tag{4.1}$$

$$\mathcal{L}_{MOV} = \lambda_c ||\overline{u}||^2 \tag{4.2}$$

$$\mathcal{L}_{SMOOTH} = \frac{\lambda_a}{2}\Sigma_i ||\nabla u_i||^2 \tag{4.3}$$

where u is the spatial displacement, λ_d, λ_c, and λ_a are the hyperparameters, and i is the index of the voxel.
The experiments are conducted with NCC as image similarity loss (see Section 4.1.1), learning rate of 1e-4, batch size of 8, and for a total of 200 epochs, which took approximately one day on a NVIDIA A100 Tensor Core GPU (40GB). The dataset includes four age groups and is described in Chapter 4.

Results:
In Figure 4.2, the generated atlas for the age group 25-28 GW is visualized, illustrating the deformation effect of different hyperparameter choices for the regularization terms of the loss function, namely $\mathcal{L}_{DEF}$ and $\mathcal{L}_{MOV}$ to train VoxelMorph's atlas generation module.
While variations in $\mathcal{L}_{DEF}$ result in only minor changes in the template's appearance (see Figure 4.2, first row), the choice of $\mathcal{L}_{MOV}$ has a significant impact. Small values ($\mathcal{L}_{MOV} \leq 0.01$) produce sharp templates, whereas increasing the hyperparameter by a factor of ten leads to the emergence of two overlapping brain structures, one larger in the background and a smaller, more profound one in the foreground (see Figure 4.2, $\mathcal{L}_{MOV} = 0.1$). This effect becomes even more pronounced as the hyperparameter increases further (see Figure 4.2, bottom row).

To further analyze the impact of $\mathcal{L}_{DEF}$, the models are retrained an additional 200 epochs (in total 400) while setting the hyperparameter $\mathcal{L}_{MOV}$ to 0.0. Experiments for two values of $\mathcal{L}_{DEF}$ were conducted: 0.01 and 1.0. The results are presented in Figure 4.3.
The predicted atlases shown in the first column of Figure 4.3 exhibit similar features (like lamination or cortical folding) for both hyperparameter settings. However, the brain appears smaller for $\mathcal{L}_{DEF} = 1.0$. Furthermore, differences in registration performance are evident when the template is warped to the subject space to match an unknown subject, as shown in the middle column of Figure 4.3. The lateral sulcus is defined more distinctly for $\mathcal{L}_{DEF} = 0.01$, as highlighted by the small red circle in Figure 4.3 top row. In contrast, the right posterior ventricle is larger for $\mathcal{L}_{DEF} = 1.0$ (see large red circle in Figure 4.3 bottom row), making it more closely resemble the fixed (target) image.

Figure 4.4 illustrates the generated templates warped into the subject space, showing the effect of varying the hyperparameter $\mathcal{L}_{SMOOTH}$. Since the hyperparameters for $\mathcal{L}_{DEF}$ and $\mathcal{L}_{MOV}$ were selected according to the results of the previous experiment, they are fixed at 0.01 and 0.0, respectively, in this analysis. As in previous experiments, the weight of the loss term $\mathcal{L}_{SMOOTH}$ is varied three times by a factor of 10, starting at 0.01 and

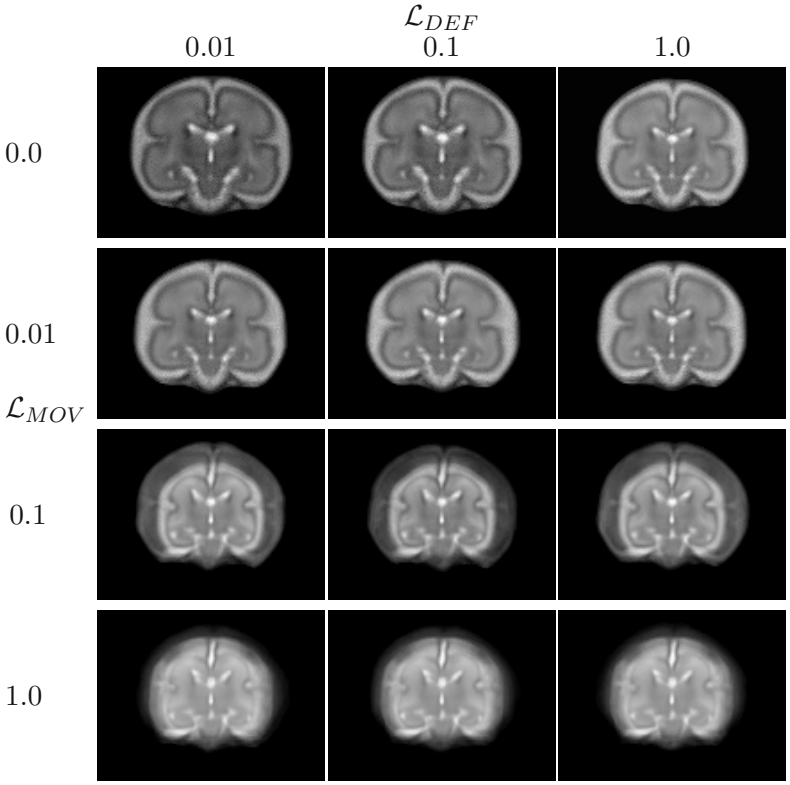


Figure 4.2: Generated atlas (age class: 25 to 28 GW) with varying hyperparameter of the regularization terms L_{MOV}: $\lambda_c = [0.0, \ldots, 1.0]$ and L_{DEF}: $\lambda_d = [0.01, \ldots, 1.0]$.

ending at 1.0. The impact of $\mathcal{L}_{SMOOTH}$ is examined across three different age groups. In the youngest case, 21 to 24 GW (see Figure 4.4, left column), the cortical surface is the smoothest, with folding limited to three lateral sulci. In older age groups, the cGM exhibits more complex folding patterns.
Comparing the same subjects with different hyperparameter settings reveals more pronounced edges of the cGM and an artifact-free visualization. For instance, in the first age group, when $\mathcal{L}_{MOV}$=0.01 (see Figure 4.4 top left corner), the area near the lateral sulci appears blurry. However, it remains sharp when $\mathcal{L}_{MOV}$ is given a higher weight.

Discussion:
In the implementation of VoxelMorph, a global linear average is added to the decoder output (see Section 4.1). The objective of the linear average is to provide a prior to both the template generation and registration network. In the case of [DRGS19], where the development of the adult brain from 18 to 90 years is investigated, the training is initialized by including a template in their model. Since the brain images were all aligned in an affine way, all brains appear in a similar size and thus a single template fulfills its purpose in guiding the prediction. However, in the case of fetal brain data, which are only rigidly aligned, one single template can not depict the variation of the dataset

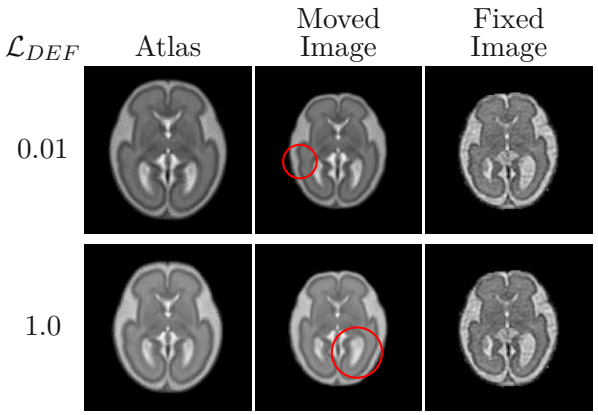


Figure 4.3: Impact of the hyperparameter of $\mathcal{L}_{DEF}$ ($\lambda_d = [0.01, 1.0]$) on the generated atlas (first column) and the template warped into the subject space (middle column). The right column displays the fixed image (GT of the image registration).

(especially the brain size and cortical folding). This can be observed in Figure 4.2, where the network is unable to fully learn the representation of the fetal brain, as the linear average is not learned by the network, but treated as an additional input. Consequently, the network's learning process is predominantly focused on addressing the consequences of the linear average, rather than on learning the true variability of the data. When the generated template differs significantly from the linear average, the model adds the generated brain on top of the linear average, hindering its integration into a meaningful representation. This results in the appearance of two brains, a smaller one generated by the decoder and a larger one in the background (see Figure 4.2 bottom row).

Beyond the linear average, the choice of hyperparameters for the regularization terms significantly influences the generated templates. As illustrated in Figure 4.2, the loss function $\mathcal{L}_{MOV}$ plays a crucial role in this process. The $\mathcal{L}_{MOV}$ computes the deformation field based on the average of the last 100 subjects sharing the same attribute. This approach offers two key advantages: (1) the deformation field is smoothed by averaging over 100 samples, reducing individual variability, and (2) in datasets with a large number of samples, the inclusion of a new subject can still influence the deformation field, ensuring adaptability over training time.
While this method proves beneficial in adult brain atlas construction, its effectiveness is limited in fetal atlas construction due to the substantially higher in-class variance. In particular, the inherent uncertainty in determining gestational age, typically ± 7 days, introduces significant variability. This increased variation complicates the use of a single averaged deformation field, potentially reducing the accuracy and reliability of the generated fetal brain templates.
The second regularization term, $\mathcal{L}_{DEF}$, controls the balance between atlas generation and the registration network. Its weight determines the permissible degree of deformation. Specifically, higher values of this loss term reduce deformation between the template and

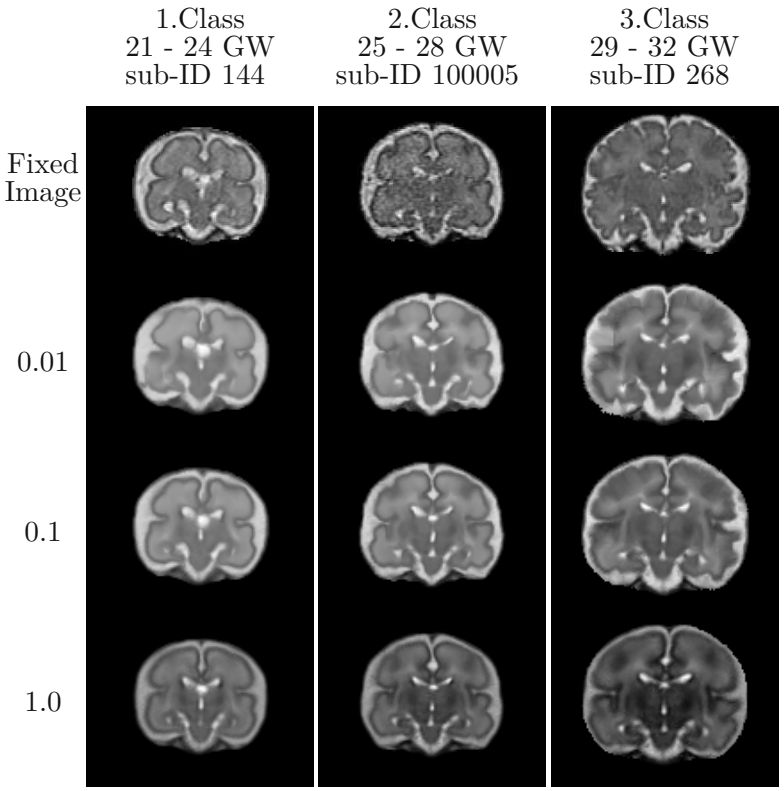


Figure 4.4: Grid figure of generated templates warped into the subject space across three different age groups in coronal orientation, illustrating the effect of the hyperparameter λ_a of $\mathcal{L}_{SMOOTH}$. The top row begins with $\mathcal{L}_{SMOOTH}$ =0.01, and each subsequent row increases the value by a factor of 10.

the warped subjects, resulting in a template that more closely resembles the appearance of the current age group. In contrast, lower values allow for larger deformations, leading to a template that does not need to be as distinct, since the registration network can compensate for variations between the atlas and the individual subject. Consequently, the loss term $\mathcal{L}_{DEF}$ must be carefully chosen to achieve both objectives of the current framework: generating unique, age-appropriate templates and enabling atlas-based segmentation that accommodates highly individualized shapes and appearances.

The smoothness of the cGM is controlled by the hyperparameter $\mathcal{L}_{SMOOTH}$, which regulates the balance between anatomically plausible folding and the suppression of irregular surface deformations. Ideally, this ensures that all gyri and sulci remain visible while preventing excessive, unnatural folding.
This effect is particularly noticeable in age group three (29–32 GWs). At $\mathcal{L}_{SMOOTH}$=0.01, the cranial cortical surface appears almost fibrous, with pronounced lateral edges. In contrast, increasing $\mathcal{L}_{SMOOTH}$ results in a significantly smoother surface, reducing extreme irregularities.

For optimal results, the final choice of $\mathcal{L}_{SMOOTH}$ should fall between 0.1 and 1.0, preferably closer to 1.0. Since a value of 1.0 leads to the disappearance of gyri and sulci, while a value of 0.1 retains excessive sharpness, particularly in the right lateral sulcus.

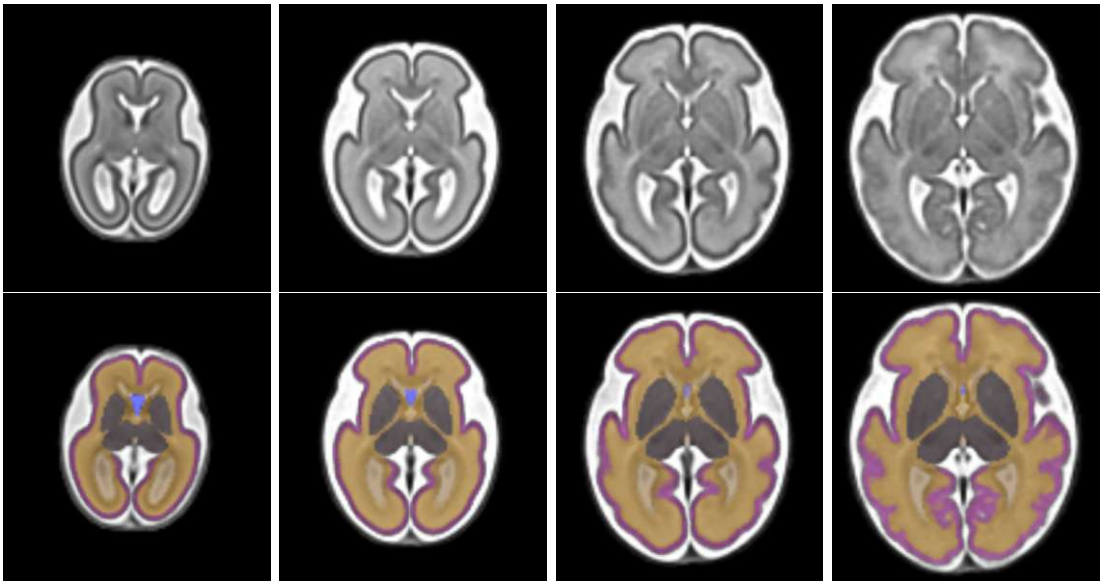
4.2 ANTs SyN

ANTs is a state-of-the-art framework where the atlas is constructed via traditional groupwise image registration. However, due to the data definition in these preliminary experiments, where multiple subjects of different ages are grouped into a single class, each group exhibits significant variations in size and appearance. This experiment aims to evaluate whether ANTs can serve as a baseline approach, suitable for fetal brain atlas construction.

Research Question:
To what extent can the multivariate template construction method in ANTs be applied to fetal brain atlas creation, considering individual variability, developmental stages, and image intensity variability? What are the limitations of this approach?

Method:
The atlas construction is achieved by antsMultivariateTemplateConstruction2.sh, part of the ANTs framework (version 2.4.4) for (brain) image registration [AYP+10]. The dataset includes four age groups and is described in the introduction of this chapter (see Chapter 4). The process follows an iterative coarse-to-fine approach, progressively enhancing alignment accuracy through downsampling and Gaussian smoothing across multiple resolution levels. Precise non-linear registration is performed using the Symmetric Normalization (SyN) transformation model. Cross-correlation with a radius of four serves as the similarity metric, effectively guiding the optimization process. Most parameters are set to their default values to ensure compatibility with standard template construction practices. Since the data had already been preprocessed, no bias field correction or rigid alignment was applied. The predefined initial template, which serves as the starting point, is the linear average of training samples sharing the same attribute (cf. Section 4.1).
The segmentations of the atlas are created by fusing all the segmentation maps of the subject sharing the same attribute using antsJointFusion [AYP+10]. Finally, test cases not included in the template creation can be segmented by registering the template with the test subject. The registration, performed using antsRegistrationSyN and antsRegistrationSyNQuick, employs the same parameters as those used during template creation.

Results:
In Figure 4.5, the constructed atlases of four different age classes are shown. Differences in size and appearance of the fetal brain are captured, characterized by the smooth cGM in younger cases (see Figure 4.5a), the neurotypical asymmetric development of the two hemispheres, particularly evident in the right lateral sulci (see Figure 4.5b), and the cortical lamination or the increased cortical folding in older cases (cf. Section 2.1).



(a) 21-24 GW (b) 25-28 GW (c) 29-32 GW (d) 33-36 GW

Figure 4.5: Fetal brain atlas constructed with ANTs for four different GWs (21–24, 25–28, 29–32, 33–36). The top and bottom rows show the T2-weighted structural templates and their corresponding segmentation maps, respectively.

The fused segmentation maps are sharp and valid, following the cGM (violet segmentation in Figure 4.5a bottom row), even with increasing age and complexity.

Discussion:
Although combining the entire neurotypical dataset spanning 16 different GWs into four classes, ANTs successfully produced sharp templates, demonstrating their ability to handle significant variations. This effectiveness is attributed to the multi-step process employed by ANTs, which involves an initial affine alignment of all subjects within an age group, followed by deformations handled through a coarse-to-fine approach and a pyramidal strategy.
The method proves suitable as a baseline model, enabling the comparisons of traditional image registration algorithms with the proposed deep-learning approach developed in the course of this master's thesis. An independent test dataset has already been established to facilitate the future evaluation of both the deep learning framework and the ANTs. Due to the incorporation of fused segmentation maps, the registration and segmentation performance between the two approaches can be evaluated.
While the deep learning framework achieves registration and segmentation through a single feed-forward operation, ANTs employs optimized parameters through affine and deformable registration. This approach has several limitations. Due to the fusion of the segmentation maps, an uncertainty is introduced, which will be further propagated to the test subject. Furthermore, the test subject needs to be registered to the corresponding age-specific template, which takes a couple of hours for a single case. An additional

limitation is the computational cost to learn the atlas, as its construction requires 45 hours on 16 CPU cores.

4.3 Summary

In this chapter, preliminary experiments were conducted as a baseline for the development of a fetal brain atlas, using a simplified dataset consisting of four age classes. The focus was on VoxelMorph, particularly its loss function implementation for image similarity and registration regularization.
VoxelMorph has demonstrated its suitability as a baseline approach for fetal atlas construction. However, to transform the existing framework to fetal brains, several adjustments are necessary:

1. The deformation regularizer, $\mathcal{L}_{MOV}$, was found to be incompatible in the context of fetal brain atlas learning and will be excluded from future analyses. In contrast, the loss terms $\mathcal{L}_{DEF}$ and $\mathcal{L}_{SMOOTH}$ showed promising results and are incorporated into the proposed framework, with an initial hyperparameter range already identified.

2. The linear average should be either replaced by age-specific priors or excluded. Given our goal of minimizing preprocessing and prior use, the linear average is excluded from the model input.

Lastly, ANTs, representing traditional approaches, generate sharp templates and will serve as a baseline for future evaluation and comparison.

CHAPTER 5

Conditional Fetal Brain Atlas Learning

The fetal brain undergoes complex neurodevelopmental changes in inherent tissue characteristics, including the emergence and disappearance of brain structures, and the substantial increase in brain size and morphology, also outlined in Chapter 2 in more detail. Consequently, a single, static template cannot capture the dynamic and heterogeneous progression of the fetal brain.
For this reason, a contemporary literature review of deep learning–based image registration algorithms was conducted (see Chapter 3) to identify methods suitable for continuous template construction and atlas-based fetal brain tissue segmentation. Hereby, the probabilistic registration framework VoxelMorph proposed by [BZS+18], along with its extension for adult brain templates spanning from 15 to 90 years [DRGS19] as well as its label prediction capability [DBGS19], inspiring this work to extend it for fetal brains. In Chapter 4, the evaluation of the VoxelMorph's components within the context of fetal brain atlas construction is presented. The preliminary experiments focused on the impact of different image similarity loss functions, the introduction of anatomical constraints with additional regularization terms in the loss function, and the role of the linear average in template creation. In summary, the analysis highlighted the need for a solution that minimizes both the preprocessing requirements of the input data and the incorporation of strong anatomical priors into the model.

To facilitate fetal brain atlas learning, in this master's thesis, two deep learning-based models for **C**onditional **A**tlas **L**earning (CAL) are proposed.
CAL-REG: A framework consisting of a generative submodule coupled with a direct registration network (see Section 3.3.2), which predicts a deformation field (see Figure 5.1). The condition-based generated template is iteratively optimized by enhancing the performance of the registration framework. Due to the regularization of the registration, which penalizes large deformations, the generation framework is forced to improve and

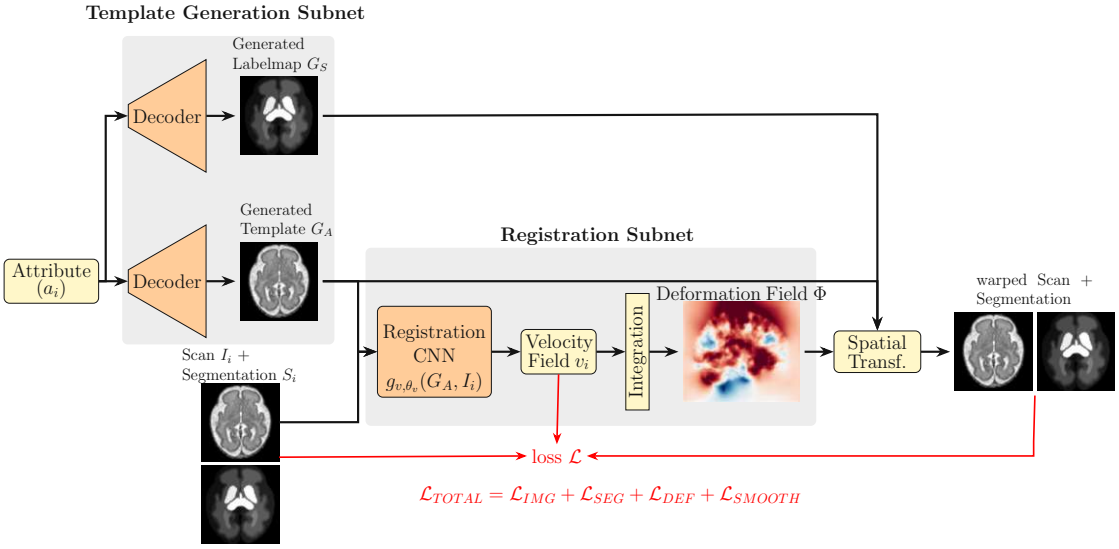


Figure 5.1: Overview of the *CAL-REG* model architecture for conditional fetal brain template generation and brain tissue segmentation. The network can be structured into two sub-nets denoted as the template generation and the diffeomorphic registration net (gray background). The total loss $\mathcal{L}_{total}$ is a superposition of the image similarity loss $\mathcal{L}_{IMG}$, the segmentation loss $\mathcal{L}_{SEG}$, the two regularization terms $\mathcal{L}_{DEF}$ and $\mathcal{L}_{SMOOTH}$.

create attribute-specific images close to the provided subjects.
CAL-GAN: This model extended the CAL-REG approach by incorporating an additional discriminator network that compares the generated output with real dataset images (see Figure 5.2). Here, optimization is driven by an adversarial training approach, where the template generation framework competes against the discriminator to improve its output. The following section provides a more detailed explanation of the methodology and the implementation of the two approaches proposed.

5.1 CAL-REG

The weakly-supervised framework makes concise usage of the VoxelMorph framework [DRGS19] with additional adaptations for fetal imaging data. It consists of two sub-networks: A template generation and a diffeomorphic registration network (see Figure 5.1). The input of the model is a paired MRI scan volume I_i and its segmentation label map S_i, along with associated attributes a_i (e.g., gestational age, sex), where i denotes the subject of the dataset. The segmentation comprises six different labels, including eCSF, cGM, deep GM, total WM, ventricles and "other" (combining cavum, brainstem, thalamus, basal ganglia, and cerebellum). I_i are of the size $128 \times 128 \times 96 \times 1$ and S_i are of size $128 \times 128 \times 96 \times 6$, where each channel hosts an anatomical label as a one-hot encoded representation (see Section 5.4.1).

5.1.1 Template Generation Network

The first step involves upsampling the attribute a_i to align with the image shape. Therefore, a 3D dense layer of 128×128×96 with 8 features is utilized. After that, the conditional input is handed over to the decoder, consisting of fully connected layers, one block of upsampling, four convolutional layers, and ReLu activation layers. The output of the decoder is the generated template. In comparison to VoxelMorph, no linear average is implemented (see Section 4.1.2). The different brain sizes and the large morphological inhomogeneity do not allow for small deviations from the linear average. Since the template is registered with the individual scan I_i the generated atlas is normalized between zero and one.

5.1.2 Diffeomorphic Registration Network

The objective of the unsupervised registration network is to register the generated template G_A to the individual scans I_i. This is achieved using diffeomorphic deformations, which are differentiable, invertible, and thus preserve topology (see Chapter 3) [DBGS19]. Thus, the two images (G_A and I_i) serve as inputs to a U-Net [RFB15] based registration framework, which consists of a convolutional layer with 32 filters, followed by four down-sampling layers with 64 convolutional filters and a stride of two, and three up-sampling convolutional layers with 64 filters. Three-dimensional convolutions in the encoder and decoder using a kernel size of 3, and a stride of 2. Each convolution is followed by a LeakyReLU layer with parameter 0.2 (see Section 3.3). Although training is conducted on three-dimensional volumes, the velocity and displacement fields are initially estimated at half resolution (with three up-sampling steps) and subsequently scaled up linearly during training. This allows the faster calculation of the displacement-guided loss terms.

The output of the registration U-Net is the velocity field v_i. By integrating the stationary velocity field v_i the final deformation field Φ can be obtained by Eq. 5.1 [Ash07].

$$\Phi^{(t=1)} = \Phi^{(t=0)} + \int_{t=0}^{t=1} v(\Phi^t)dt \tag{5.1}$$

where $\Phi^{(t=0)}$ is the initial condition. After that, the deformation field is applied to the generated atlas template by using an STN [JSZK15]. An STN is differentiable, can be trained using backpropagation, can be integrated into any CNN model, and transforms a feature map without supervision. It is composed of three modules. The first module, a localization network, predicts the transformation parameters. The second component, a grid generator, applies the estimated transformation to the input feature map. Finally, the sampler module acts as an interpolator, which warps the generated template into the subject space [REYT24].
Since the deformation field represents how to deform the generated template to achieve the original input image, the same deformation field can be used to warp the segmentation map into the subject space.
Given the invertibility of the deformation field, it can be inverted and employed to deform

the GT image I_i from the subject space into the template space, thereby enabling a bidirectional approach that introduces an additional loss term to enhance performance, without extensive computations.

5.1.3 Loss and Regularization

The loss term of the network proposed is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{IMG} + \mathcal{L}_{SEG} + \mathcal{L}_{DEF} + \mathcal{L}_{SMOOTH} \tag{5.2}$$

where, according to Eq. 3.1, $\mathcal{L}_{IMG}$ quantifies the difference between the moving image G_A and the fixed image I_i, while all the other terms act as regularization, enforcing anatomical plausibility or restrictions (see Chapter 3 and Section 4.1.2).
Based on preliminary experiments, $\mathcal{L}_{IMG}$ is chosen to incorporate a localized NCC loss function L_{sim}, ensuring the similarity between the scan I_i and the generated warped template G_A.

$$\mathcal{L}_{sim} = NCC_{local}(G_A, I_i) = \frac{1}{n}\Sigma_j \frac{\Sigma_{j\in\delta_k}(G_{A_j} - \overline{G_{A_k}})(I_{i_j} - \overline{I_{i_k}})}{\sqrt{\Sigma_{j\in\delta_k}(G_{A_j} - \overline{G_{A_k}})^2 \Sigma_{j\in\delta_k}(I_{i_j} - \overline{I_{i_k}})^2}} \tag{5.3}$$

with the window size n, G_{A_j} and I_{i_j} refer to the j^{th} voxel in the subject images and warped atlases. And $\overline{G_{A_k}}$ and $\overline{I_{i_k}}$ are the average image intensity values over window δ_k, which is centered at the k^{th} voxel. The image loss is computed bidirectionally (see Eq. 5.4), meaning image similarity is assessed twice: first, between the warped template and the fixed image, and second, between the warped fixed image using the inverse deformation field and the generated template. The loss weights sum to one, with the forward calculation contributing the most.

$$\mathcal{L}_{IMG} = (1-\gamma)\mathcal{L}_{sim}(I_i, g_{G_A,\theta_{G_A}}(a_i) \circ \phi_{v_i}) + \gamma\mathcal{L}_{sim}(G_A, g_{I,\theta_{G_A}}(a_i) \circ \phi_{v_i}^{-1}) \tag{5.4}$$

with $\gamma \in [0,1]$, $g_{G_A,\theta_{G_A}}(a_i)$ denotes the generated template based on the condition a_i, and ϕ_{v_i} describes the deformation field.
The second term of Eq. 5.2 represents the MSE loss applied to the one-hot encoded segmentation maps. The generated warped segmentation map is evaluated against the GT label map by measuring the overlap of labels across corresponding channels. By incorporating additional features in the form of segmentation labels within the network while utilizing the same deformation field, the registration outcome can be enhanced while simultaneously minimizing the number of learnable parameters.

$$\mathcal{L}_{SEG} = \lambda_s \Sigma_i (S_i - g_{G_A,\theta_{G_A}}(a_i) \circ \phi_{v_i})^2 \tag{5.5}$$

The third term of Eq. 5.2 accounts for the sum of the spatial displacement u, while the fourth term involves the gradient of the spatial displacement u. Both are regularization terms, where $\mathcal{L}_{DEF}$ penalizes large deformations from the template and $\mathcal{L}_{SMOOTH}$ enforces smoothness. Detailed evaluation of the regularization terms can be found in Section 4.1.2.

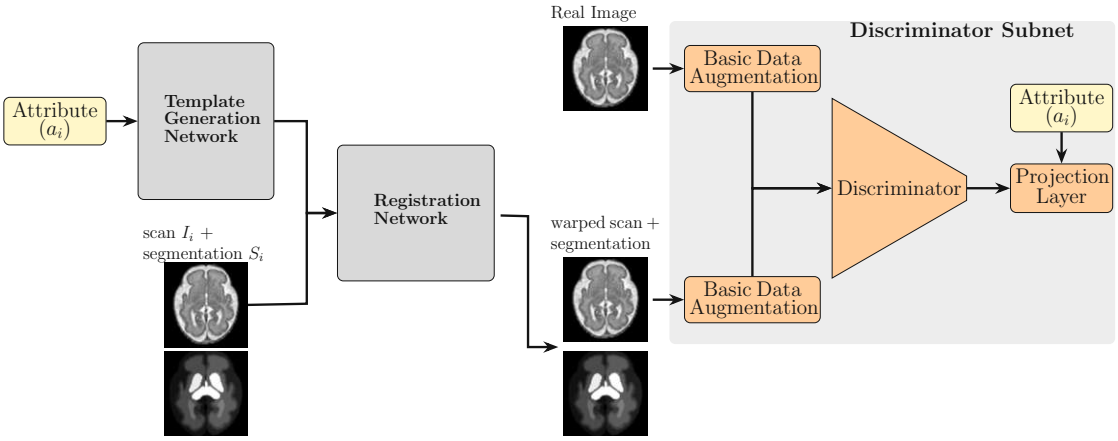


Figure 5.2: Overview of the CAL-GAN model architecture. The CAL-REG architecture is extended by incorporating a discriminator network, which operates on augmented patches from the generated warped atlas and a randomly selected training sample. Additionally, the model includes efficient embedding of the conditional attribute.

$$\mathcal{L}_{DEF} = \frac{d}{2}\lambda_d \Sigma_i ||u_i||^2 \tag{5.6}$$

$$\mathcal{L}_{SMOOTH} = \frac{\lambda_a}{2}\Sigma_i ||\nabla u_i||^2 \tag{5.7}$$

λ_s, λ_d and λ_a are the hyperparameters.

5.2 CAL-GAN

The model presented in Section 5.1 is further enhanced by incorporating a discriminator framework (see Figure 5.2). The generator and discriminator framework are trained simultaneously. The training is two-split: the generator acts as described in Section 5.1. The generator improves through the direct feedback of the similarity metrics and the regularization. The discriminator is trained by comparing the generated template, which is warped into the subject space, with a real image from the dataset. Both the warped template and the real image share the same attributes. Through continuous evaluation by the discriminator, the generation of realistic templates is reinforced. This network adaptation is expected to place particular emphasis on intensity values and their distribution.

5.2.1 Template Generation Subnet

The model architecture closely resembles that of CAL-REG. However, the conditional embedding is incorporated through a Feature-wise Linear Modulation (FiLM) layer,

which applies a feature-wise affine transformation to the intermediate features of a CNN [PSdV+18]. The operation is mathematically described by

$$FiLM(\mathbf{F}_{i,c}|\gamma_{i,c},\beta_{i,c}) = \gamma_{i,c}\mathbf{F}_{i,c} + \beta_{i,c} \tag{5.8}$$

where γ and β are scale and transformation parameters of the feature map $\mathbf{F}$ from the i-th layer and c-th feature, learned by the condition a.
In contrast, in CAL-REG, where the parameters scale with increasing image size, the conditional embedding using FiLM requires only two parameters for each feature map, leading to reduced computational costs. In addition, FiLM generalizes and learns complex connections from little data, allowing rapid adaptations during training [PSdV+18].

5.2.2 Discriminator Network

During training, the discriminator evaluates two types of inputs: (1) a random sample from the real dataset and (2) a generated, warped template produced by the generator. By comparing these inputs, the network learns to identify subtle discrepancies between authentic data distributions and synthesized outputs.
The discriminator employs a five-layer fully convolutional network architecture, adopting a PatchGAN structure [DU18]. This design processes the input image at a patch level, emphasizing local changes, while retaining artifact-free high image quality with fewer parameters (faster processing) [DU18].
The conditional information, such as sex or age, is integrated into the discriminator using the projection method [MK18]. Unlike traditional conditional GANs (cGANs), where conditional vectors are concatenated with the input features, this approach projects the condition onto the intermediate feature space through an inner product operation [MK18]. Specifically, the condition is embedded into a learnable vector, which is then combined with spatial feature maps via a dot product. This method preserves spatial information while leading to higher image quality and sharper outputs [MK18].
The discriminator training loop incorporates data augmentation by applying random flips and translations to both real and fake samples before feeding the transformed data into the discriminator. Training the discriminator on augmented samples avoids overfitting, hence enhancing its generalization capability [ZLL+20]. This improvement arises from increasing feature invariance with respect to real and fake samples under specific transformations, as well as incorporating a regularization term based on the variance of the augmented representations to constrain model complexity.
The discriminator's loss term is computed directly from its output logits. The use of logits (instead of post-activation probabilities) ensures that the model retains gradient information across the entire input range, mitigating issues like vanishing gradients and providing a more stable training signal. The adversarial terms used to train the generator and discriminator networks correspond to the least-squares GANs objective, encouraging more stable and higher image quality [MLX+17].

5.3 Evaluation Strategy

Evaluation metrics are essential for quantitatively assessing results. However, validating the correct determination of deformable transformation parameters is a non-trivial task, as it requires dense landmark correspondences, which are often unavailable [CLW+25]. In addition to accuracy measures that evaluate registration errors using corresponding landmarks, regularity measures analyze the folding behavior of the deformation field to assess its smoothness [CLW+25].
Inspired by the literature [TH15] and the Fetal Tissue Annotation (FeTA) Challenge [PLdD+23], standard evaluation metrics are used to evaluate the segmentation performance. To further quantify the performance of 1.) the generated templates and 2.) the registration quality, the evaluation metrics proposed by [DRDG21] are followed.

The first metric is the **mean Jacobian determinant** $|J_\phi(x)|$, which is defined by

$$|J_\phi| = |\nabla\phi(x)| = \begin{vmatrix} \frac{\partial\phi_1}{\partial x_1} & \frac{\partial\phi_1}{\partial x_2} & \frac{\partial\phi_1}{\partial x_3} \\ \frac{\partial\phi_2}{\partial x_1} & \frac{\partial\phi_2}{\partial x_2} & \frac{\partial\phi_2}{\partial x_3} \\ \frac{\partial\phi_3}{\partial x_1} & \frac{\partial\phi_3}{\partial x_2} & \frac{\partial\phi_3}{\partial x_3} \end{vmatrix} \tag{5.9}$$

where x delineates the voxel location in the deformation field $\phi(x)$. During interference, all test cases are warped into the subject space. The average of the Jacobian determinant is then computed for the deformation field. A Jacobian determinant $|J_\phi|$ close to 1 indicates a smooth local deformation, while values close to zero indicate local folding or inversion of the deformation field [DRDG21, CLW+25].
The second metric is the **average deformation norm** $\overline{||u||}$ a measure for smoothness and the regularity of deformations [DRDG21]. The deformation norm is defined by

$$||u|| = \frac{1}{\Omega}\int_\Omega ||u||dx \tag{5.10}$$

where a small $\overline{||u||}$ indicates minimal displacement from the template to the warped template, while a large $\overline{||u||}$ suggests significant structural changes. In the context of conditional atlas learning, it serves as an age-specific distance measurement—indicating the level of deformation from the generated template to the resulting warped image.
The final metric is the **Entropy Focus Criteria** [AHS+97, DRDG21], which quantifies image quality by measuring the Shannon entropy of voxel intensities. While it was originally introduced as an indicator of blurring caused by fetal motion, it also serves as a general image quality measure and is therefore applicable for assessing the generated atlas and the warped test images. It is defined by:

$$\mathrm{E} = -\sum_{j=1}^{N} \frac{x_j}{x_{\max}} \ln\left[\frac{x_j}{x_{\max}}\right] \tag{5.11}$$

where x_j is the intensity of voxel j, N the number of voxels, and $x_{\max} = \sqrt{\sum_{j=1}^{N} x_j^2}$. A lower EFC (approaching zero) indicates that the energy is highly concentrated in a single voxel, leading to minimal blurring.

For the label maps, three different evaluation metrics are selected. The most commonly used metric to evaluate segmentation performance is the **Dice Similarity Coefficient** or F1-score focuses on the overlap of two label maps.

$$DICE = \frac{2|A_i \cap B_i|}{|A_i| + |B_i|} \tag{5.12}$$

where A denotes the voxels of the GT segmentation (see Chapter 2) and B the voxels of the warped segmentation, respectively. $|A_i|$ defines the total number of voxels of segmentation A. The DSC ranges from 0 to 1, where 0 is no overlap between the two segmentation maps and 1 denotes that the segmentations are identical [TH15].

In addition to the DSC, the **95th Percentile Hausdorff Distance** emphasizes the differences at the boundaries between the predicted and GT segmentations [TH15].

$$\begin{aligned} H(A,B) &= \max\left(h(A,B), h(B,A)\right) \\ h(A,B) &= \max_{a \in A}(\min_{b \in B} d(a,b)) \\ h(B,A) &= \max_{b \in B}(\min_{a \in A} d(b,a)) \end{aligned} \tag{5.13}$$

where $d(a,b)$ is the Euclidean distance between voxels a and b. Essentially, $h(A,B)$ determines the closest point in B for each point in A, then selects the maximum of these distances, identifying the most mismatched point in A. As a result, the degree of discrepancy between the two sets is measured by the distance of the point in A that is farthest from any point in B, and vice versa.

Finally, the **Volumetric Similarity**, which evaluates the similarity between the volume of GT labels V_{GT} and predicted labels V_{pred} [TH15].

$$VS = \frac{2 * (V_{\mathsf{pred}} - V_{\mathsf{GT}})}{(V_{\mathsf{pred}} + V_{\mathsf{GT}})} \tag{5.14}$$

The VS ranges from -1 to 1, where negative VS indicates underestimation, with the predicted volume less than the volume of GT. Positive VS indicates overestimation, respectively, with the predicted volume greater than the GT volume.

5.4 Neurotypical Dataset

Data used in this thesis were obtained from *in utero* fetal MRI scans conducted as part of a prospective fetal MRI research study (EK Nr. 2032/2022, entitled: segmentation of fetal brains with callosal agenesis") at Vienna General Hospital. Participants were part of routine examinations. The examination was performed within 45 min, and neither MRI contrast medium nor sedation was applied. Inclusion criteria were neurotypical pregnancies between 18 and 39 weeks of GA.
In vivo fetal MRI was performed according to the ISUOG Practice Guidelines [PMDC+23] by using 1.5 T (Philips Ingenia/Intera, Best, The Netherlands) or a 3 T (Philips Achieva)

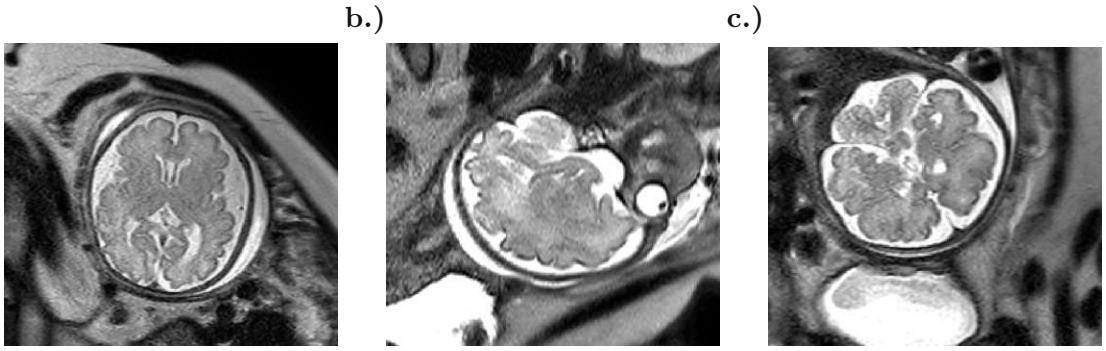


Figure 5.3: During clinical routine, fetal MRI scans are acquired in three anatomical planes: sagittal, coronal, and axial. **a.)** The axial plane provides cross-sectional views of brain morphology, allowing for detailed assessment of structural features. **b.)** The sagittal plane evaluates midline brain structures, offering insights into the central organization and alignment of the brain. **c.)** The coronal plane assesses brain symmetry and cortical maturation. Image courtesy of the Medical University of Vienna.

scanner with a 32-channel body matrix and spine coils. A 3 T scanner provides higher SNR compared to a 1.5 T scanner, which can improve image resolution and contrast while retaining a comparable rate of energy deposition on tissue [PMDC+23]. However, 3 T scanners are also more susceptible to artifacts, such as those caused by maternal and fetal motion or magnetic susceptibility effects, which can sometimes offset the benefits. The field of view and number of slices varied based on maternal and fetal dimensions. Multiple T2-weighted Half-Fourier Single-Shot TSE sequences of the fetal brain were acquired in orthogonal planes (see Figure 5.3).

The acquisition parameters were TR of 1400–2000 ms, TE of 100–120 ms, in-plane resolution of 0.9–1.1 mm, 2-4.5 mm slice thickness, and an acquisition matrix size of 256x204, 256x256, or 320x320 with 2- or 4-slice interleaved acquisition (see Section 2.2).

Prior to the selection of the neurotypical cohort, missing attributes related to GW, sex, and field strength were identified and completed in the research database of the Medical University of Vienna. To achieve this, over 1,000 cases acquired between 2010 and 2022 were reviewed, including their corresponding fetal MRI reports, associated US reports, and birth protocol, to retrieve the necessary information. Subsequently, a medical expert with over three years of experience in fetal imaging identified up to 10 neurotypical cases per GW. Since fetal MRI is mainly conducted between the 20 and 25 GW, fewer cases were identified for late pregnancies ($>$ 35 GW). The selection was based on the diagnoses stated in radiological reports.

After selection, the neurotypical cases underwent pseudonymization by a trusted, separate department of the Medical University of Vienna. Pseudonymization of medical Digital Imaging and Communications in Medicine (DICOM) images involves replacing personally identifiable information with pseudonyms to protect patient privacy while maintaining data utility for analysis. In some cases, this process also includes completely removing DICOM tags to further anonymize the images and eliminate any potential for (future)

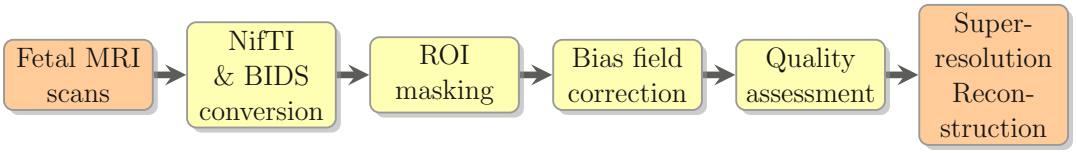


Figure 5.4: Flowchart of the first part of the MRI preprocessing pipeline. The acquired fetal MRI scans undergo several processing steps, leading to the motion-corrected SRR.

re-identification.

5.4.1 Pre-processing

Fetal MRI has inherent characteristics such as short acquisition time and thick slices. These result in high in-plane resolution but poor through-plane resolution (see Section 2.2). In order to assess and quantify the development of the fetal brain, the raw image acquisitions in three anatomical orthogonal planes are transformed into a single, isotropic, high-resolution SRR.

Furthermore, during clinical routine, several MRI protocols are employed to achieve optimal assessment of both fetal and maternal conditions. The following section elaborates on the preprocessing steps to achieve the objectives of a three-dimensional SRR with isotropic resolution of the fetal brain.

During the examination, the MRI sequences are changed multiple times, resulting in different slice thicknesses, field of view, and imaging parameters. Commonly used image sequences include T2-weighted scans from the spine, abdomen, and head (see Section 2.2), Diffusion-Weighted Imaging (DWI) and FLAIR T2-weighted sequences for functional imaging, and Merged Fast Field Echo (M-FFE) for musculoskeletal and cervical spine imaging [PKK+06].
The first step involves identifying the full brain representation in sagittal, coronal, and axial orientations. To facilitate this, multiple steps are needed (see Figure 5.4). First, DICOM images are converted into NifTI, and subsequently organized according to the Brain Imaging Data Structure (BIDS) [GAC+16] standard. While DICOM gathers demographic information about the patient and metadata in a standardized series of tags about a single scan, NifTI aggregates related scans into a cohesive image dataset, complemented by a JSON file that contains metadata [CAB+04]. Second, the NifTI files are restructured according to the international BIDS standard. This standard allows for customization, enabling the extraction of only the T2-weighted images that match specific imaging protocols. In practice, this means that from all the images acquired during clinical routine, only the T2-weighted scans of the head are retained, while all other images are excluded.

Following, the chosen NifTI stacks are further processed by NeSVor (Docker container

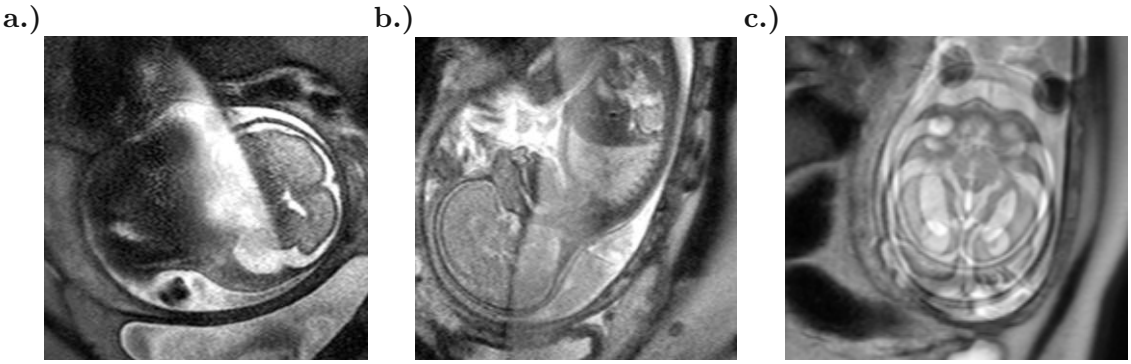


Figure 5.5: T2-weighted fetal MRI images demonstrating various image artifacts: **a.)** brain partially covered by surrounding tissue, **b.)** bias field artifact caused by magnetic field inhomogeneities, and **c.)** in-slice fetal motion leading to image coverage issues, intensity variations, or blurring. Image courtesy of the Medical University of Vienna.

image version 0.5.0 [1]) [XMG+23], which automatically brain masks all the slices to extract the brain from the mother tissue. Since MRIs are often compromised by bias field corruption [LSM+21, UKG+23], a low-frequency variation in the acquired signal caused by RF coil design, gradient Eddy currents, local flip angle variations, and inhomogeneous excitations (see Section 2.2), the bias field correction is conducted.
The final step before the SRR is the assessment of the fetal stacks based on [XLG+20]. Therefore, image artifacts such as signal voids, blurring (e.g., caused by motion), aliasing, and instances where the brain is partially or entirely outside the field of view are detected (see Figure 5.5). This results in a quality score on a scale from zero to one [XLG+20]. Based on these evaluations, the five highest-quality stacks, along with their respective slice thickness, are selected for the SRR process. The final result is a three-dimensional brain volume with an isotropic voxel spacing of 1 mm. The resolution of 1 mm was chosen based on recommendations of [DRGS19] and the increased computational time with increased resolution (17× higher for 0.4 mm) [XMG+23].

In contrast to recent literature [DRGS19, DRDG21, PCZ+21], the resulting SRRs in this study are not spatially normalized. Spatial normalization typically involves affine registration, which aligns all samples to an external template using transformations such as rotation, translation, shearing, and scaling (see Chapter 3 for details). While affine registration preserves the original characteristics of the sample, the scaling component restricts the data to a limited range of sizes—specifically, those of the templates used. To fully capture the spatiotemporal development of the fetal brain, only rigid registration is performed between all SRR volumes. Rigid registration preserves the original scale and proportions of the data, allowing for a more accurate representation of the fetal brain's growth and development over time.

In addition, the segmentation is performed for each SRR, which will later act as GT (see Figure 5.6). Therefore, each case is processed with BOUNTI [UKM+23]. BOUNTI is

[1] `https://hub.docker.com/r/junshenxu/nesvor`; last accessed on 10-APR-2025

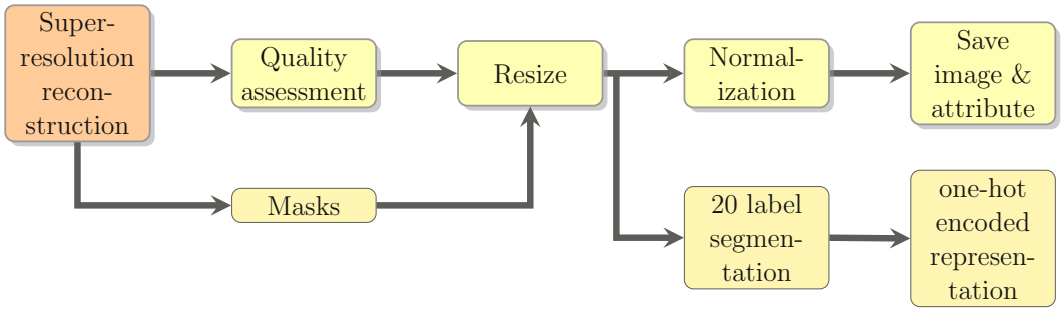


Figure 5.6: Flowchart of the second part of the MRI preprocessing pipeline. The quality of the SRR is assessed. The best quality volumes are affine aligned, resized, and normalized. Finally, the output volume is saved together with their attributes (e.g., age, sex).

based on a semi-supervised, deep-learning segmentation pipeline, trained on the publicly available data of the developing Human Connectome Project (dHCP). The BOUNTI tools generate 19 major brain tissue labels, one of the following labels for each brain hemisphere, including cGM, deep GM, eCSF, WM, cerebellum, thalamus, basal ganglia, LV, third and fourth ventricles, cavum, and brainstem [UKM+23]. After that, the labels of the two hemispheres are merged into six labels, resulting in eCSF, cGM, deep GM, total WM, ventricles (combining LV + third and fourth ventricles), and other (combining cavum, brainstem, thalamus, basal ganglia, and cerebellum) (see Appendix Figure 1). BOUNTI achieved a DSC higher than 0.9 for total WM, deep GM, cerebellum, and brainstem, a DSC higher than 0.85 for eCSF, cGM, LV, and cavum. The lowest DSC is reported for the 3rd and 4th ventricles.

Inspired by [SDG+21], the quality of the SRR is assessed and assigned to one of the levels: poor quality, average quality, and good quality (see Figure 5.4). The inclusion criteria focus on the presence of sharp edges, particularly those defined by the cortex, while exclusion criteria involve artifacts arising from image acquisition or motion (see Figure 5.5). If the SRR fails or results in poor image quality, the pipeline is repeated using manually selected, motion-free scans.
Since the output dimensions of the SRR vary for each scan, they are standardized to a uniform size of (128, 128, 96). To achieve this, masks are created for each brain volume [UKM+23], and the largest brain volume within the dataset is identified. Based on this reference, the dimensions of the entire dataset are harmonized, with the constraint that they must be multiples of 16—a requirement for the subsequent network architecture. Following this, the intensity values of the images are normalized to a range between 0 and 1. As a final step, the processed SRR images are saved in a compressed NumPy array format, accompanied by relevant attributes such as sex and age. The age classes range from 0 (21 GW) to 17 (37 GW). This ensures that the data is stored efficiently and is readily accessible for further analysis.

5.5 Summary

In this section, two frameworks for conditional fetal brain atlas learning are proposed, including CAL-REG and CAL-GAN, with a particular emphasis on network implementation. CAL-REG features a diffeomorphic registration network designed to predict both the structural representation of the brain and the corresponding segmentation maps.
In addition, CAL-GAN incorporates a discriminator network that compares patches of the generated brain images with real patches from a neurotypical dataset. Its implementation includes data augmentation and FiLM, a parameter-efficient layer that enables dynamic parameter adaptation based on conditional input.
Furthermore, evaluation metrics are presented to allow for an objective assessment of the generated templates and their corresponding segmentation maps.
The chapter concludes with a description of the curated neurotypical dataset, with particular emphasis on the preprocessing steps.

CHAPTER 6

Results

This chapter reports the results of the approach proposed in the course of this thesis. Therefore, the chapter starts with a description of the experimental setup, including frameworks used, libraries, and implementation details. Subsequently, the evaluation results of the atlas template construction and its registration performance to test data are presented. Segmentation accuracy is then examined using standard metrics across anatomical labels and age groups (see Section 5.3). The chapter concludes with a volumetric analysis of the segmentation outputs to investigate neurodevelopmental trajectories.

6.1 Experimental Setup

The frameworks CAL-REG and CAL-GAN are implemented using TensorFlow (Version 2.2) with CUDA acceleration, running on a NVIDIA A100 Tensor Core GPU (40GB) within a 64-bit Linux operating system. The model implementation is based on the VoxelMorph architecture (version 2.0) [1] [DRGS19] and its extended work incorporating a discriminator [DRDG21] [2].

Given the codependency of TensorFlow on specific Python and NVIDIA CUDA versions, and its sensitivity to version changes, all required packages have been containerized to ensure reproducibility. The containerized environment is publicly accessible via Docker Hub [3].

[1] `https://github.com/voxelmorph/voxelmorph`; last accessed on 06-MAR-2025.
[2] `https://github.com/neel-dey/Atlas-GAN`; last accessed on 06-MAR-2025.
[3] `https://hub.docker.com/r/jtischer29/fetal_atlas`; last accessed on 06-MAR-2025.

6.1.1 CAL Implementation Details

With the environment standardized, the next step involved hyperparameter tuning for optimal model performance. To optimize hyperparameters for improved conditional atlas learning of CAL-REG and CAL-GAN, a grid search approach is employed. Specifically, the weighting factors of the deformation loss $\mathcal{L}_{DEF}$ and the smoothness loss $\mathcal{L}_{SMOOTH}$ are systematically varied within the range of 0.1 to 1, while the weights of the image similarity loss $\mathcal{L}_{IMG}$ and segmentation loss $\mathcal{L}_{SEG}$ remain fixed at 1. The model undergoes training for 400 epochs (approx. 12 hours) for each hyperparameter combination, using a batch size of 1 for all experiments. The framework's performance is assessed based on mean DSC and mean deformation (see Section 5.3). The best-performing hyperparameter is found by the minimum of the mean deformation, indicating that the predicted template is close to the individual subject, and the highest DSC, reflecting accurate anatomical alignment. The result of the hyperparameter grid search is delineated in Appendix Table 1.
Based on the grid search, the optimal hyperparameters were determined as $\mathcal{L}_{DEF} = 0.1$ and $\mathcal{L}_{SMOOTH} = 1.0$ for CAL-REG. For CAL-GAN, an additional adversarial loss weight of $\mathcal{L}_{GAN} = 0.5$ was included. The weights for image similarity $\mathcal{L}_{IMG}$ and segmentation loss $\mathcal{L}_{SEG}$ were fixed at one throughout all experiments. With these configurations, both models were retrained for a total of 1000 epochs. The best-performing epoch, identified by the lowest loss, is selected for inference on the test dataset. As CAL-GAN introduces a generative adversarial component, additional considerations in training dynamics are required. Training GANs is highly dependent on a balanced training of generator and discriminator. The conducted experiments of CAL-REG implied a fast and precise generation framework (see Section 4.1.2). For this reason, an unbalanced training of the discriminator and generator is implemented, where the generator is trained once, while the discriminator is trained five times.
Normalization of the input attributes within the range of 0 (corresponding to 21 GW) to 1 (corresponding to 37 GW) further improved the stability of GAN training. Without normalization, unintended artifacts—particularly in the early subjects—appeared in the generated templates.

6.1.2 ANTs Implementation Details

To compare our proposed model with traditional methods, ANTs (version 2.4.4) were implemented as a baseline approach. The ANTs fetal brain atlas, covering gestational ages from 21 to 37 weeks, is used along with its corresponding segmentation, following the procedure outlined in Section 4.2.

6.1.3 Training and Test Dataset Configuration

A total of 308 neurotypical MRI examinations were analyzed, comprising 167 males (54.2%), 112 females (36.4%), and 29 cases with unknown sex (9.4%). The acquisitions span a timeline from January 2012 to September 2022. Over this period, changes in

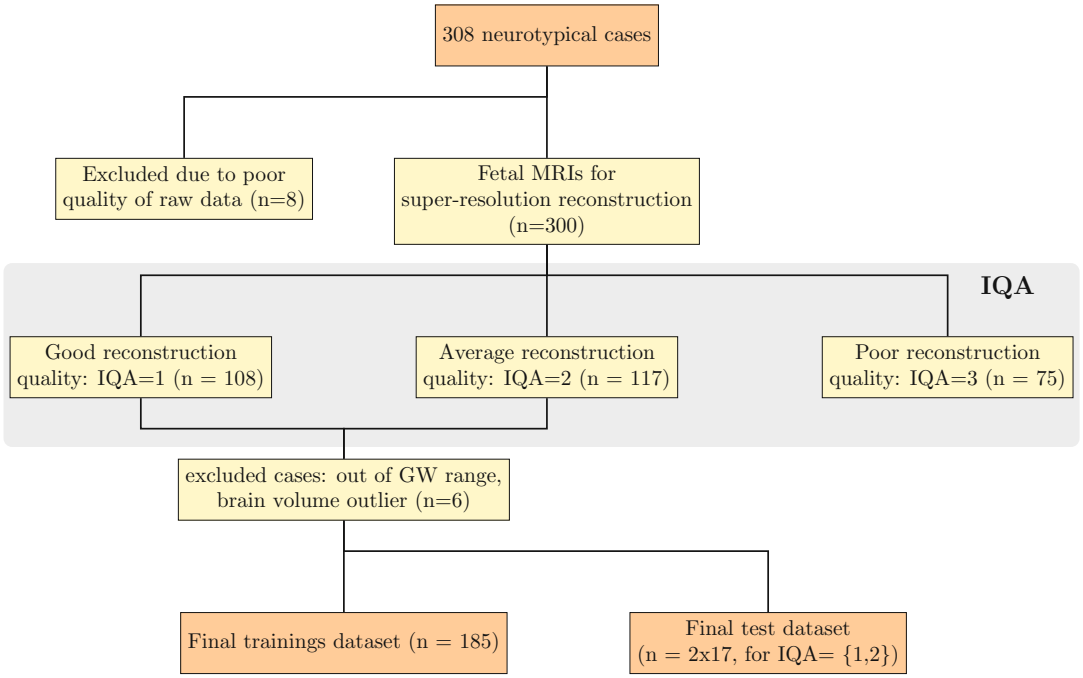


Figure 6.1: Flowchart illustrating the inclusion and exclusion criteria. Starting with the raw dataset, followed by the super-resolution reconstruction and the IQA. Finally, the training dataset of 21 to 37 GWs is defined. The final test dataset includes two cases per GW, respectively

imaging protocols and the use of different scanners resulted in variations in image quality and contrast. Of the total scans, 276 were acquired using 1.5 T scanners, while the remaining 10% (32 scans) were acquired with 3 T scanners. Of the 308 fetal examinations, 8 were excluded due to poor raw data quality or missing T2-weighted acquisitions (see Figure 6.1). The remaining 300 scans were reconstructed using NesVor (docker container image version 0.5.0) [XMG+23]. Image Quality Assessment (IQA)) was performed on all SRRs, with IQA scores ranging from 1 (indicating good reconstruction quality) to 3 (indicating poor reconstruction quality) (see Figure 6.1). For the construction of the training dataset, only samples rated as good or average were included. Age groups with fewer than three subjects (<21 and >37 years) were excluded due to insufficient samples for training and inference. This resulted in a training dataset (n=185) covering a continuous GA range of 21 to 37 GW. The distribution of GA within the training dataset is shown in Figure 6.2.
To address the imbalanced age distribution in the training dataset, specifically the low sample size in the last trimester (see Figure 6.2), the pick rate of underrepresented classes is increased by raising their selection probability. This adjustment ensures that each attribute is presented at equal frequencies during training.

The test dataset is two cases per GW, leading to 34 (2×17) test subjects. One for each

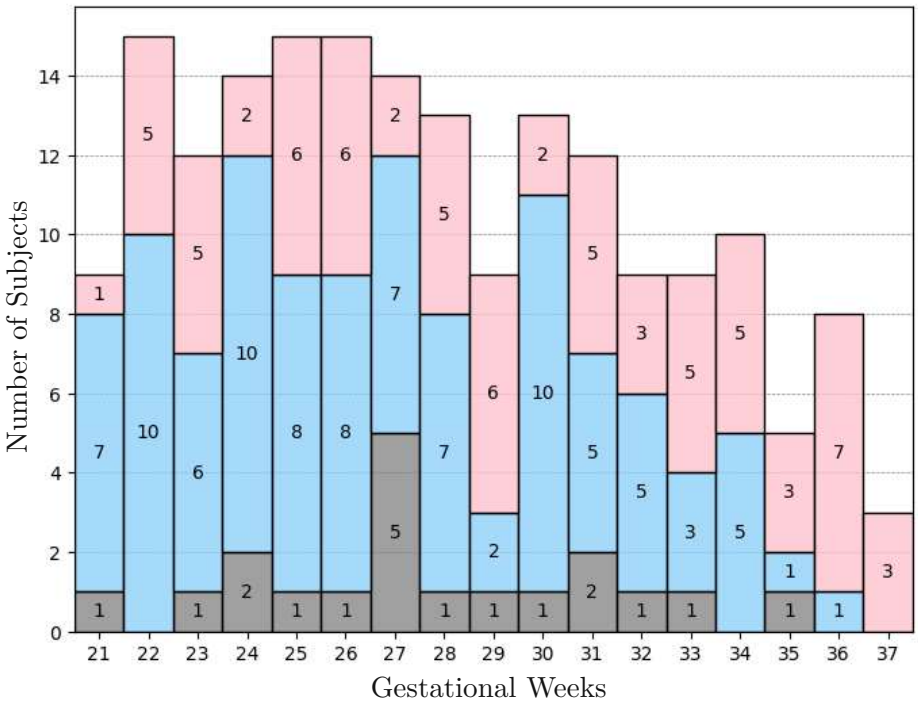


Figure 6.2: Histogram of the training dataset ranges from 21 to 37 GWs. The stacked bins show the number of cases for each GW split by sex: male ( ), female ( ), and unknown ( ).

GW with good reconstruction quality (IQA=1) and one with average reconstruction quality (IQA=2) respectively.

6.2 Atlas Generation and Registration Performance

This section places special emphasis on template generation and registration performance. Qualitative results are first presented based on the templates generated for three different GAs. The interested reader is referred to the Appendix, where detailed quantitative results and figures for each GA are provided (see Chapter 7.6). The section concludes with a more in-depth analysis of the best-performing model proposed in this thesis.

6.2.1 Comparison to Baseline Implementations

The quantitative results of all model implementations are given in Table 6.1, including the average Jacobian determinant, average deformation, and EFC of the test dataset (see Section 5.3). In addition, an exemplary selection of the generated templates is shown in Figure 6.3 for three age classes. The selected age classes capture key neurodevelopmental changes across different stages, reflecting variations in brain size, cortical folding, lamination, and the emergence or disappearance of brain structures (see Section 2.1).

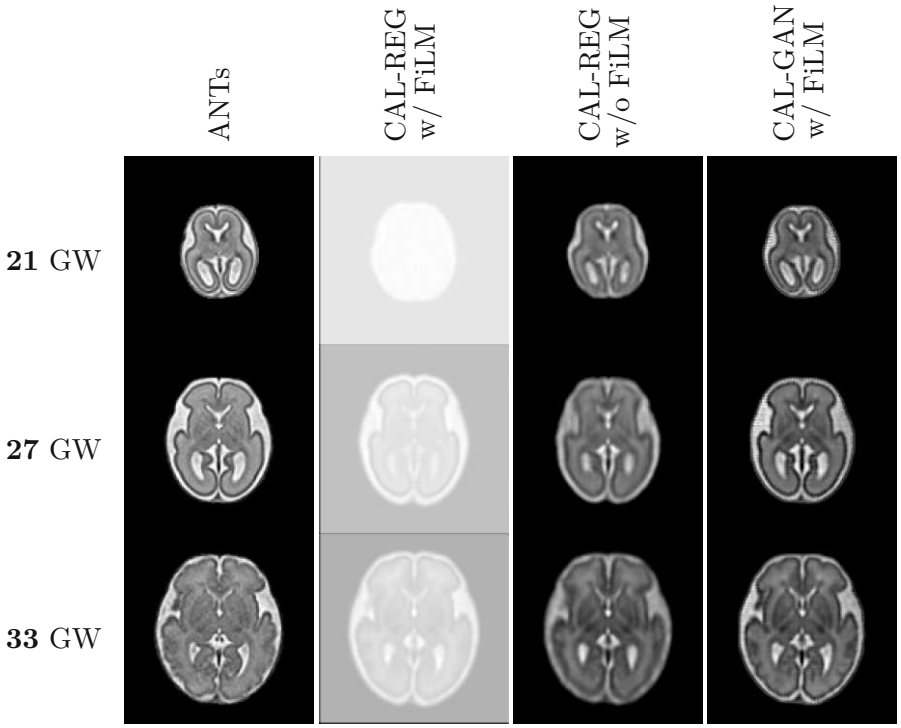


Figure 6.3: Resulting templates of all models (columns) for three age groups (rows).

Recall that ANTs SyN and ANTs SyNQuick differ only in their registration algorithms for test cases, the template generation process remains identical (see Section 4.2). Consequently, in Figure 6.3, a single template is displayed for ANTs.

Shared attributes across the atlases of the same age group are that all templates exhibit the same brain size and the same degree of cortical folding. The cGM is smooth for young fetuses, with complexity increasing as age progresses. For instance, the lateral sulcus, one of the first sulci typically emerging around the 25 GW (see Section 2.1), is visible across all templates of the 27 GW. In the oldest age group, the cortical morphology is complex, indicated by the increased presence of sulcal and gyral regions (see Figure 6.3 bottom row). Additionally, all methods produce smooth local deformations without irregular folding avg. $|J_\phi| \approx 1.000$ (see Table 6.1).

Several differences arise among the models regarding template appearance and image quality. The average deformation between the generated template and the test dataset was lowest for ANTs SyNQuick (692) and highest for ANTs SyN (908) across all models (see Table 6.1). Furthermore, the EFC of 0.35 reflects the high image sharpness (see Figure 6.3). A more detailed representation of the age-specific templates, registration performance of ANTs SyN to the test dataset, and their corresponding segmentations are shown in the Appendix Figures 3 and 4.
In CAL-REG with FiLM, the voxel intensity distribution within the brain region is concentrated close to one, resulting in a bright appearance. Additionally, the background,

Table 6.1: Quantitative results of all model implementations, showing the mean values across all gestational ages. The table compares ANTs, representing traditional optimization approaches, with our deep learning models, CAL-REG and CAL-GAN. The deep learning methods are evaluated with (w) and without (w/o) the additional FiLM implementation.

	Avg $\lvert J_\phi \rvert$ (↑)	Avg $\lvert\lvert \text{Def} \rvert\rvert$ (↓)	EFC (↓)
ANTs SyNQuick	**1.000 ± 0.000**	**691.9 ± 233.6**	0.345
ANTs SyN	**1.000 ± 0.000**	908.5 ± 290.2	0.347
CAL-REG w/ FiLM	0.998 ± 0.001	806.5 ± 147.7	0.995
CAL-REG w/o FiLM	**1.000 ± 0.000**	798.2 ± 284.5	0.344
CAL-GAN w/ FiLM	0.995 ± 0.000	775.2 ± 239.8	**0.326**

typically delineated by a value of zero, exhibits an intensity offset, leading to a generally very blurry image of the templates. While the brain structure in younger cases appears nearly as a binary brain mask, different anatomical regions become more distinguishable in later cases (see Figure 6.3). This progression is also reflected in the high EFC value of 0.995.
In contrast, templates produced by CAL-REG without FiLM appear darker than those from ANTs, but without the intensity offset seen with the FiLM implementation. In this case, the background is correctly set to zero, and the brain intensities vary according to tissue composition, ranging between zero and one (see Section 2.2). Finally, the template generated by CAL-GAN with FiLM exhibits a particularly sharp appearance, with distinct contrast in the dark edges of the cGM.

6.2.2 Best-Performing CAL Model

As shown in Figure 6.3, CAL-GAN with FiLM produces sharp templates with age-appropriate brain size, pronounced definition of the cGM. This is also supported by the quantitative metrics, given in Table 6.1. Here, CAL-GAN with FiLM achieved an EFC of 0.326, and the lowest average deformation among all deep learning-based methods, with 775.2 ± 239.8. Given these results, both the atlas generation and registration performance of CAL-GAN are examined in greater detail.

The constructed spatiotemporal atlas of the fetal brain, covering GW 21 to 37, is presented in the left column of Figure 6.4. Alongside the atlas, the fixed image from the test dataset and its warped representation are displayed. The subsequent two columns show the segmentation maps of the fixed image and the warped representation, with total WM as a representative example.
The final column illustrates the deformation field in the axial orientation. To ensure visually comparable results, the deformation field values are normalized between 0 (black), 0.5 (yellow), and 1 (red).

The generated atlas exhibited a progressive increase in brain size with advancing GA, accompanied by morphological complexity. While the cGM is mostly smooth from 21 to

24 GWs, the cortical folding evolves from 26 GW onward. Interhemispheric differences in cortical folding are shown between 26 and 31 GW. At 26 GW, the lateral sulci emerged with greater prominence in the right hemisphere compared to the left. This asymmetry vanishes by 31 GW, with sulcal patterns becoming bilaterally comparable in subsequent templates. The first occurrence of complex, multi-branched folding in the template was observed at 35 GW.
In terms of appearance, the atlas predicted by CAL-GAN provided homogeneous intensity distributions and sharp anatomical boundaries. In contrast, individual subjects exhibited intensity inhomogeneities, appearing as blurry or noisy regions. These inhomogeneities were unified through the process of warping the atlas to the subject space, resulting in more uniform intensity profiles (see Figure 6.4, 29 GW). However, minor artifacts emerged during the generation process, appearing as thin stripes within the eCSF. These artifacts, which were most prominent in younger subjects and generally located near the brain's peripheral edges (see Figure 6.4, 21–23 GWs), were also projected and warped into the subject space.

The correspondence between the atlas and individual subjects was evaluated by examining alignment quality. Minor discrepancies were observed at 26 GW, where the cortex of the template remained smooth, while individual subjects displayed small lateral gyri and sulci. More pronounced discrepancies were evident at 28 GW, where the template only displayed a small lateral sulcus in the right hemisphere, whereas subjects exhibited well-defined gyri and sulci. Despite these differences, the registration process demonstrated the ability to reconstruct early sulcal folds in subjects at 28 GW, which were absent in the age-matched template (see Figure 6.4).
However, as the complexity of the brain increased with advancing GA, the warped atlas increasingly failed to accurately capture subject-specific features. Gyri and sulci appeared loosely defined or lacked the clarity observed at earlier stages. This limitation was most pronounced at 37 GW, where nearly all anatomical features were predominantly inherited from the template rather than accurately reflecting the subject's unique structural characteristics.

Figure 6.4: Grid figure visualizing the different steps of the CAL-GAN for different GW. Age-specific generated template, fixed image, moved image, fixed segmentation, moved segmentation, deformation field in x-direction (columns from left to right). The test image quality of the SRR is IQA=1.

Table 6.2: Quantitative results of all model implementations, showing the mean values across all gestational ages. The table compares ANTs, representing traditional optimization approaches, with deep learning approaches, including VoxelMorph and our models, CAL-REG and CAL-GAN. The deep learning methods are evaluated both with and without the additional FiLM implementation. t_N represents the time needed for atlas construction (training), while t_i indicates the time required to process each test sample.

	t_N	t_i	DSC (↑)	HD95 (↓)	VS (→ 0 ←)
ANTs SyNQuick	21.5h	0.3h	0.805 ± 0.103	1.66 ± 0.29	-0.018 ± 0.041
ANTs SyN	21.5h	16.3h	**0.872 ± 0.061**	**1.25 ± 0.17**	**-0.003 ± 0.040**
CAL-REG w/ FiLM	67h	34s	0.841 ± 0.019	1.71 ± 0.23	0.040 ± 0.111
CAL-REG w/o FiLM	72h	34s	0.840 ± 0.060	1.99 ± 0.24	0.034 ± 0.128
CAL-GAN w/ FiLM	67h	34s	0.863 ± 0.052	1.60 ± 0.20	0.013 ± 0.033

6.3 Segmentation Performance

In Table 6.2, the results of the segmentation metrics averaged over all labels are given (see Section 5.3). In addition, the metrics of the six segmentation labels, namely eCSF, cGM, deep GM, total WM, ventricles (combining LV + third and fourth ventricles), and other (combining cavum, brainstem, thalamus, basal ganglia, and cerebellum) are displayed in the Appendix Table 2.

In general, similar segmentation results are achieved for the deep learning models with a variance of 2.3 % in DSC, and 0.38 mm of HD95 between the worst and best-performing model. The conventional approaches by ANTs show a broader variance of 6.7 % in DSC and 0.409 mm of HD95.
Over all models, cGM shows the lowest DSC, followed by the ventricles. In contrast, the highest accuracy is observed for total WM and miscellaneous brain structures (see Table 2).

The best segmentation performance is achieved by ANTs SyN, with an average DSC of 0.872, HD95 of 1.25 mm, and VS of -0.003. The best deep learning model, CAL-GAN, follows with an average DSC of 0.863 (see Table 6.2).

The conventional approach ANTs involves atlas construction and fusion of the segmentation maps, taking approximately 21.5 hours. In contrast, training the deep learning models takes about three days, but inference time is significantly shorter: a single test case requires only one forward pass through the model, which includes predicting the warped subject image, segmentation maps, and deformation fields. This forward pass takes approximately one second. For comparison, ANTs SyN requires approximately half an hour, while the faster method, SyNQuick, completes the inference of a single case in about 30 seconds (see Table 6.2).

6.4 Volumetric Analysis of Fetal Brains

The three-dimensional segmentation of brain regions and the calculation of the resulting volume for hundreds of cases spanning across different GWs enables the visualization of brain growth patterns during development. The volumetry was constructed based on the guidelines proposed by [RW98].
Therefore, the volumes of six different brain regions are derived from the segmentation maps (voxel count × voxel spacing) of the training data (n=185). The resulting volumes are plotted as a scatterplot with the GA on the x-axis and the volume in cm^3 on the y-axis (see Figure 6.5). After that, the point cloud is approximated using a second-degree polynomial fit (see solid black line in Figure 6.5).

The total WM volume follows a nearly linear growth pattern, while deep GM exhibits slight exponential growth, and cGM, ventricles, and miscellaneous structures display a stronger exponential trend. In contrast, eCSF initially increases linearly between 20 and 30 GW, then levels off after the 30 GW.
The goodness of the polynomial fit is evaluated by the coefficient of determination denoted as R^2:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{6.1}$$

where SS_{res} is the residual sum of squares, and SS_{tot} the total sum of squares. In the best case, the fit matches the data, resulting in $SS_{res} = 0$ and thus R^2 of 1.
A R^2 value exceeding 0.9 is achieved for total WM, cGM, and other regions. In contrast, lower R^2 values are recorded for deep GM (0.83), eCSF (0.72), and ventricles (0.40) (see Figure 6.5).

The volume calculation is repeated post-training for 1.) two test cases per GW and 2.) the predicted age-specific template.

1. In case of the **test dataset**, GT volumes are represented by triangular markers, while model predictions of the test data are indicated by circular markers (see Figure 6.5). Recall that the test cases were selected based on the quality of the SRR (see Section 5.4.1), where IQA=1 (orange points) indicates good quality and IQA=2 (green points) average quality. In most of the time points, the GT markers of IQA=1 lie closer to the fit than the GT markers of IQA=2 (see Figure 6.5).

2. The volume of the **predicted templates** is represented by filled black squares (see Figure 6.5). For all structures, the predicted template volume closely follows the trajectory of the training dataset, meaning the predictions fall within one standard deviation of the training data. However, for later GWs (>30), the model tends to overestimate deep GM and total WM, while underestimating cGM (see Figure 6.5).

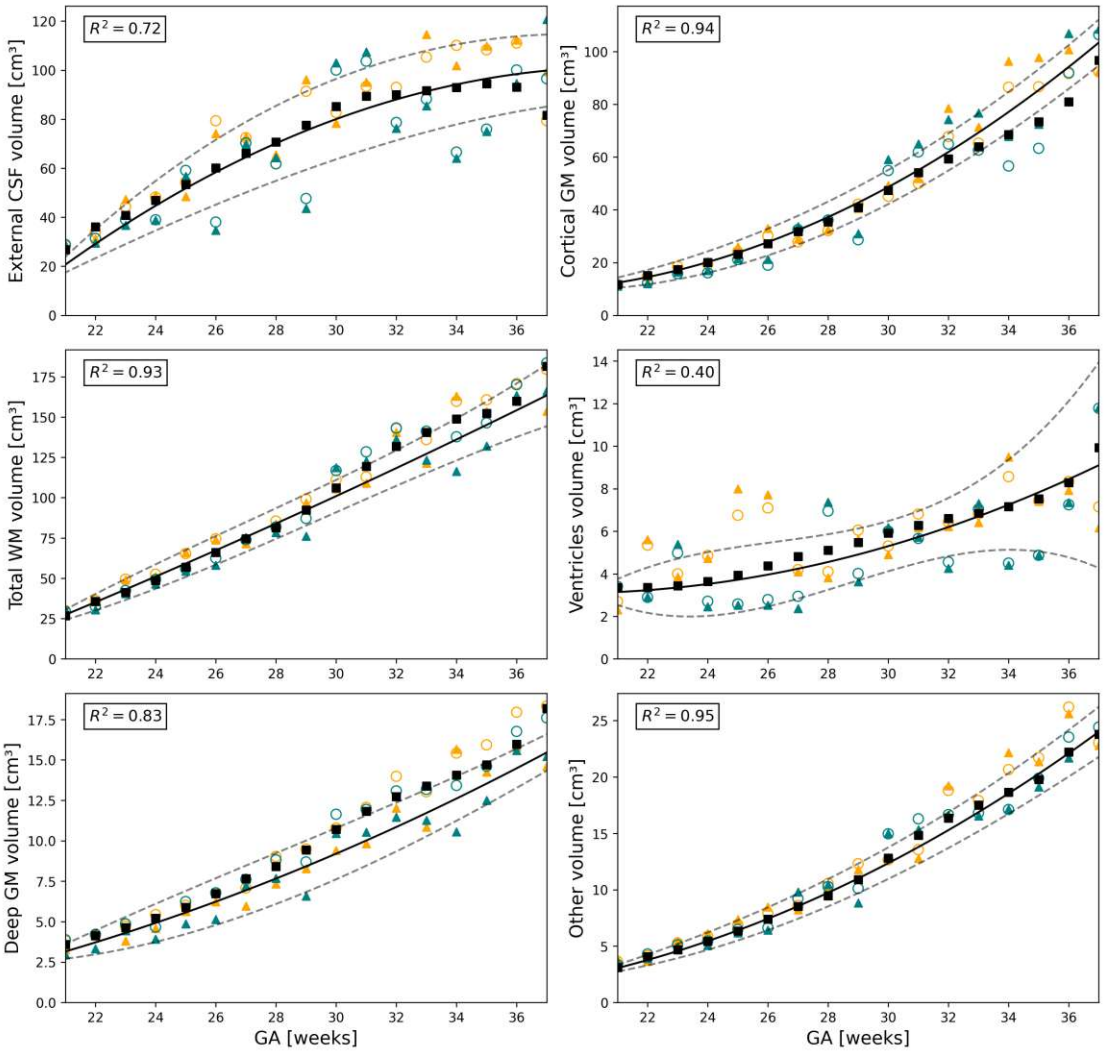


Figure 6.5: Volumetric trajectories of eCSF, cGM, total WM, ventricles, deep GM, and miscellaneous (from top left to bottom right) from 20 to 37 GW. The trajectory of the training data, including its standard deviation, is modeled using a polynomial fit. The brain region-specific volumes of the test data are shown for the GT (△) and the predicted values (◯). The color indicates the quality of the SRR, where orange points correspond to $IQA = 1$ (good SRR) and green points to $IQA = 2$ (average SRR), respectively.

CHAPTER 7

Discussion and Conclusion

This chapter examines the results presented in the previous section, following the same structural outline. The findings are discussed in relation to existing literature and evaluated in comparison to baseline models. Each section discusses the implications, strengths, and limitations of the corresponding results. The chapter concludes with a reflection on the limitations of the approach proposed, suggestions for future work, and a summary of the main insights drawn from the study.

7.1 Atlas Generation and Registration Performance

In Figure 6.3, the CAL-REG with FiLM model is characterized by a narrow intensity distribution, resulting in bright templates. In contrast, both CAL-REG without FiLM and CAL-GAN with FiLM generate templates with a normalized intensity distribution. The following section provides a brief discussion of these observations.

With the integration of a discriminator, the predicted templates became noticeably sharper (see templates of CAL-GAN w/ FiLM in Figure 6.3). The discriminator provides feedback regarding the intensity distribution of the images. Since the discriminator constantly compares the generated fetal brain images to real ones, the generator is forced to produce images with an intensity distribution, similar to the input data, between 0 and 1. Additionally, normalizing both the input image and the generated template improves the precision of the NCC loss, as this metric assumes normalized input images. In CAL-GAN, this representation is learned by the adversarial game between the generator and discriminator. In contrast, CAL-REG with FiLM fails to accurately capture the intensity profile of the input data (see Figure 6.3). To meet the normalization requirement, a normalization layer was implemented that scales the generated atlas between 0 and 1. However, the framework became unstable after approximately 100 epochs. At that point, a small region exhibited high-intensity values (0.02–1.0), while most of the brain remained within a much lower intensity range (0–0.02), as shown in Figure 7.1.

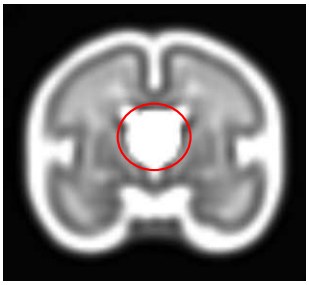


Figure 7.1: Created template of CAL-REG with normalization layer in axial orientation. The model gets unstable during training, resulting in an image artifact (red circle). The red circle accompanies the voxel intensities from 0.02 to 1, and the brain has the voxel intensities of 0 to 0.02.

In this context, the linear average approach implemented in VoxelMorph proves beneficial (see Section 4.1 and discussion in Section 4.1.2). Besides providing an initial estimate of brain shape and size, it offers a rough but consistent approximation of the intensity distribution across the dataset. For completeness, in CAL-REG without FiLM, the predicted intensity distribution aligns with the input images.

7.1.1 Comparison with Baseline Implementation

The average deformation norm of ANTs SyNQuick with 691.9 ± 233.6 suggests that its templates more closely resemble the anatomical characteristics of the respective age classes compared to competing models. This is indicated by the reduced degree of deformation required to achieve spatial correspondence with the test cases.
Nevertheless, both ANTs SyNQuick and ANTs SyN warp the same atlas to the subject space and have similar (low) average deformations. Contrary to this expectation, ANTs SyN exhibits 908.5 ± 290.2 the highest average deformation norm among all evaluated models. In addition, ANTs SyN achieves the highest DSC (0.872 for ANTs SyN vs. 0.805 for ANTs SyNQuick) and qualitatively aligns more closely with the anatomical structures of the test subjects (see Figure 3). These findings indicate that the average deformation norm has to be evaluated under consideration of the segmentation metrics.
In the case of CAL-GAN, where atlas generation and registration are trained simultaneously (end-to-end), the balance between the two networks results in the DSC of 0.863 (Δ=-0.009 to ANTs SyN) and the low average deformation norm of 775.2 ± 239.8 (Δ=+83.3 to ANTs SyNQuick).

7.1.2 Best-Performing CAL Model

The atlas predicted by CAL-GAN with FiLM demonstrated limitations in accurately representing cortical folding between 25 and 31 GW. While the templates during this period displayed predominantly smooth cortical surfaces, individual subjects exhibited pronounced folding patterns. This discrepancy may be attributed to two factors: 1) The

deformation field applied during registration is constrained by regularization in the loss function, which enforces anatomical plausibility. If this restriction is balanced correctly, the framework would be forced to generate templates that more closely represent the corresponding subjects, thereby introducing cortical folding. The current restriction may not be stringent enough, and 2) since cortical folding is subject-specific and variable, it may not be generalizable through template creation alone and requires subject-specific registration for accurate representation.

7.1.3 CAL-GAN and existing Fetal Brain Atlases

[LSM+21] introduced CAS-NET, a framework for fetal brain atlas construction that adopts a similar approach to CAL-GAN, particularly through the use of a diffeomorphic registration subnet. Both methods were developed using datasets of comparable size. CAS-NET was trained on 202 subjects, with 54 used for testing, while CAL-GAN utilized 185 training subjects and 34 for testing. Their performances were also similar, as CAS-NET achieved a mean DSC of 85.2%, which is close to CAL-GAN's DSC of 86.3%. Despite these similarities, several key differences distinguish the two approaches. CAS-NET incorporated a linear average of both the segmentation maps and the structural images, whereas CAL-GAN did not. In addition, CAS-NET used only nine anatomical labels and divided the data into four gestational age groups: less than 25 weeks, 26 to 28 weeks, 29 to 32 weeks, and more than 32 weeks. Another important distinction is that the atlases generated by CAS-NET were of identical size across all age groups. However, these atlases exhibited limited anatomical detail. For instance, the atlas corresponding to the age group of more than 32 gestational weeks showed less cortical folding and appeared less sharp than the actual subject data of that age. Furthermore, CAS-NET was evaluated using only a single performance metric, which restricts the ability to fully assess its registration and atlas generation quality. One notable advantage of CAS-NET is its short training time, requiring only two hours for 500 training epochs (NVIDIA GTX 3080 GPU 10GB). In contrast, our deep learning models required nearly three days to train (NVIDIA A100 Tensor Core GPU 40GB). This discrepancy may be partly explained by CAS-NET's use of fewer classes, which likely reduced the overall computational complexity of the task.

In contemporary brain atlas construction, one of the preprocessing steps includes the affine alignment of the input SRR volumes [AYP+10, DRGS19, DRDG21]. Recall that affine alignment is defined by translation, rotation, shearing, and scaling (see Section 3) and will not alter the individual shape, intensity composition, or morphology of the brain. In the case of adult brains [DRGS19, DRDG21], where size changes are negligible, the affine alignment will set the focus on the deformations and tissue composition itself. Therefore, the whole dataset is registered to a single reference (Talairich see Section 2.1). However, in state-of-the-art fetal brain atlas construction [PCZ+21, LSM+21, ZHZ+25] an affine alignment is conducted by registering the age-grouped volumes to multiple templates and highly influences the conditional atlas learning:

1. Due to the scaling properties of affine alignment, individual size information is entirely lost, as each volume inherits the size of the template.

2. Volumes are grouped by age and registered to reference image(s). On the one hand, this simplifies the conditional atlas learning by eliminating the ambiguity in GW determination. On the other hand, this approach weakens the atlas generation framework, since it involves time-consuming preprocessing [DRDG21] or the cooperation of already published atlases [ZHZ+25].

3. During inference, the framework can only handle discrete sizes of test subjects, as the registration network does not account for variability in brain size. This limitation requires test subjects to be aligned to a pre-defined size.

In contrast, the training and test data of our approach were rigidly (translation and rotation) aligned, meaning the variance in brain size, as well as wrongly determined GWs, are still present. The bigger variability of the training data makes the model more robust, but at the same time also more versatile and generalizable to other data.
The size of the ANTs template is the product of the affine alignment of all subjects sharing the same age group. In other words, the brain size is the average brain size of all subjects sharing the same class. For this reason, the constructed atlas of ANTs has a high image quality and sharp edges (see Figure 6.3). In contrast, in CAL-REG and CAL-GAN, the templates are generated without any prior knowledge. Meaning the final brain size is the result of the balance between image similarity and anatomical constraints in the loss function. This results in the blurry template prediction of CAL-REG without FiLM (see Figure 6.3). With the introduction of the discriminator in CAL-GAN, the EFC further decreased by $\Delta\text{EFC} = 0.018$ to 0.326, indicating enhanced image quality.

7.2 Segmentation Performance

The developed models within this thesis successfully integrate anatomical labels into the framework and are able to predict thin layers with high accuracy. While the DSC scores for CAL-REG with FiLM and without FiLM are similar, the HD95 is 0.28 mm lower for CAL-REG without FiLM, indicating more precise boundary predictions. Among all deep learning-based methods, CAL-GAN with FiLM achieved the highest segmentation accuracy. The reasons for this superior performance are discussed in Section 7.1. Overall, the traditional method, ANTs SyN, outperforms deep learning approaches for four out of six anatomical labels. However, it requires more than 16 hours to process the test dataset due to the need for additional registration. In contrast, the faster ANTs SyNQuick algorithm ran in only 1 hour and outperformed the deep learning frameworks in only one label (see Table 6.2). For comparison, the deep learning framework proposed segments all test cases in real-time, and processing a single subject takes less than one second.

In Figure 7.2, the segmentation accuracy of CAL-GAN is presented in a radar plot, where each axis corresponds to a specific label, and the distance from the origin indicates the

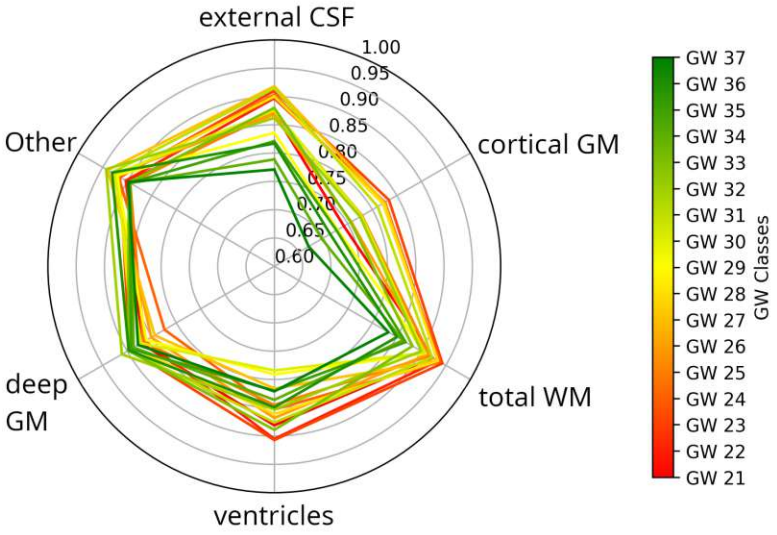


Figure 7.2: Radar plot depicting segmentation performance of CAL-GAN, with segmentation labels as the axes and the DSC representing the distance from the origin. The origin corresponds to a DSC of 0.6, while the outermost point represents a DSC of 1. The colors indicate the GW, with red representing the youngest and green the oldest.

DSC. The center of the radar plot represents a DSC of 0.6, while the outermost circle corresponds to a DSC of 1.0. The underlying data, along with additional segmentation metrics across GA, are provided in Table 3.
As discussed in Section 6.3, the cGM shows the lowest segmentation accuracy, with a mean DSC of 0.768. This may be attributed to the structure's thin and complex morphology, making it more difficult to predict accurately, as well as its high intra-class variability. Across different GAs, the cGM also displays the widest range in segmentation performance, with DSC values ranging from 0.67 at 37 GW to 0.84 at 23 GW. While the cortical surface appears relatively smooth at earlier gestational stages, its complexity increases significantly with advancing age.
This morphological evolution affects three neighboring brain regions: total WM, cGM, and eCSF (see Appendix Figure 1). Consequently, similar trends are observed across these three segmentation maps, with lower overlap of the segmentation maps for older subjects and higher accuracy for the youngest ones (see Figure 7.2). Although, the overall segmentation accuracy is higher for eCSF and total WM, this is likely due to their larger spatial extent. No clear age-related trend can be identified for the remaining labels.

7.3 Volumetric Analysis of Fetal Brains

Figure 6.5 illustrates the trajectory of fetal brain volume in 185 neurotypical fetuses. The polynomial fit closely aligns with findings from recent studies [SAB+12] [AdPM+17] [UKM+23].

The ideal scenario occurs when the atlas prediction closely aligns with the training fit, meaning the generated template accurately reflects the training data. However, individual cases, represented by triangular markers, may fall outside the standard deviation of the training data. In these instances, the goal is for the prediction (the warped template in subject space) to primarily align with the individual case while disregarding the majority of the training data. Representative examples are shown at 26 GW of the eCSF, where both test cases lie outside the standard deviation yet are nearly correctly predicted by the registration network (see Figure 6.5). In contrast, the opposite behavior is observed for the deep GM, where nearly all test cases fall within the standard deviation of the training data, but all predictions follow the overestimation of the template prediction (see Figure 6.5).

Several neurodevelopmental processes, described in Section 2.1, are reflected in these trajectories. The cGM exhibits exponential growth after 25 GW, driven by increased gyrification. Similarly, deep GM and total WM expand progressively due to continuous neuronal migration.

The R^2 values indicate a strong correlation between the training data and the polynomial fit for total WM, cGM, deep GM, and other structures. However, lower correlations are observed for eCSF and ventricular volumes. This discrepancy may stem from the limited sample size in the third trimester, large standard deviations throughout gestation (ventricles), or increased variability specifically in the third trimester (eCSF).

These factors also contribute to the model's difficulty in accurately predicting volumes, particularly in the final trimester (32–37 GW). This is most evident in cGM and ventricular volumes, where growth is underestimated after 32 GW.

7.4 Limitations

The term GW is ambiguous since three different starting points are possible: last menstrual period, ovulation and/or fertilization, and implantation [Jud11]. Hence, we have to work with the more vague and imprecise definition of GW, which is only in a few cases corrected by the US retrospectively. This leads to fetal brains being classified in the same category as those that accurately belong to it, despite being developmentally older or younger than their assigned label. At critical time points, such as the emergence of the first gyri and sulci around 25 GW (see Section 2.1), the modeling may be inaccurate. Consequently, constructing and modeling an age-specific atlas becomes more challenging, as these attributes introduce inherent uncertainty that cannot be fully resolved during training.
Another limitation arises from the selection process of neurotypical cases (see Section 5.4). These cases were included solely based on the radiological report. However, neurotypical development does not necessarily imply age-appropriate development, but rather only the absence of pathologies. By additional sighting and evaluating the imaging data, fetuses can be included/excluded by the appearance of the brain.

As stated in the introduction of the neurotypical dataset (see Section 5.4), the GT segmentations used in this thesis were automatically generated [UKM+23], which depending on the age of the fetus and the quality of the SRR varies in quality (BOUNTI has an average DSC between 0.85 and 0.95). Since the segmentations are also used for the test dataset, they do not reflect the underlying GT, limiting the objective evaluation of our model. This can be visually confirmed by the total WM exemplary displayed in Figure 6.4. While this is negligible in early subjects, due to the smooth surface of the cortex it is getting more prominent in older cases, where inter-individual deviations have to be considered. In addition, the thin segmentation of the cortex automatically affects the neighboring segmentations of the eCSF and the total WM. Furthermore, this not only affects the segmentation performance but also the registration performance, since the segmentation loss is included in the training process.
In addition, due to the vulnerability of the fetus, only low field strengths of the MRI scanner can be applied, leading to low resolution, partial volume effects, and finally to no existing GT. However, with expert-annotated labels, a higher accuracy in generating segmentation and morphological processes can be expected.
Another limitation of this study is that the data used comes from a single center with fixed scanning parameters. Additionally, it can be assumed that the majority of the dataset's population is Caucasian.

7.5 Future work

- Data - The limitations should be addressed in future work, with a primary focus on enhancing the available GT segmentation maps and images. Specifically, segmentation maps need to be manually refined by medical experts. Implementing these improvements is expected to lead to better training outcomes and a more accurate evaluation of the predicted templates and segmentation. In addition, one might focus on improved model prediction in the last trimester (especially 32 - 37 GW). Therefore, a higher sample size of late pregnancies can be incorporated, as well as additional topological information guiding the current shortcomings in morphological changes in the cGM.

- Framework - Various extensions can be applied to the framework. [DBGS19] introduced a surface-based extension of the Voxelmorph framework by bidirectionally applying the deformation field to surface points. [MR23] incorporated the thickness of the cortex to further improve the registration accuracy. Many improvements in registration frameworks were proposed. For instance, the implementation of a pyramidal coarse-to-fine-registration model with affine to local deformations can further improve the registration and segmentation accuracy, while simultaneously reducing the trainable parameters.
Finally, since the establishment of GANs, improvements regarding training stability and image quality have been proposed. The model's performance exhibited high sensitivity to the loss function's hyperparameters. As an extension of this work, rather than relying on a computationally and time-intensive grid search, a gradient-based hyperparameter optimization approach could be employed.

- Applications - The model is versatile and can be applied to various imaging modalities, including US and CT. A co-registration or fusion of MRI and US enables the integration of complementary information from both modalities. This involves aligning features between US and MRI to improve image registration and analysis. Additionally, MRI can support US image segmentation by providing anatomical guidance, enhancing accuracy in clinical applications.
Another application is the implementation of a normative atlas. Since we have already quantified neurotypical development based on volumetric analysis of anatomical labels and polynomial fits, we can now evaluate the volumetry of a new, previously unseen subject. However, the model's ability to assess pathologies must be evaluated beforehand.

7.6 Conclusion

This thesis proposed two conditional fetal brain atlas generation frameworks constructed using a curated dataset of 308 neurotypical T2-weighted MRI scans spanning 21 to 37 GWs. At first, a comprehensive literature review was conducted, encompassing medical image registration algorithm and their application in fetal brain atlas construction. Based on this literature review, two models were developed and implemented. In the first framework, CAL-REG, the template creation is optimized by solving an image registration problem. Based on the GA of a subject, an age-appropriate brain template is created, which is registered in a U-Net-based CNN with the subject's SRR. After registration, the deformation field is used to propagate the created template and its corresponding anatomical labels into the subject space. The loss calculation between the subject and the warped template, as well as anatomical constraints by restricting the deformation field, is used to further optimize the model. Building on this, the second framework, CAL-GAN, incorporates an additional discriminator network. In an adversarial game, the discriminator tries to distinguish generated fetal brain images from real ones, while simultaneously the generator tries to create more realistic images. In addition to the proposed deep learning-based approaches, a fetal brain atlas was also constructed using the traditional method ANTs. For this, the segmentation maps from the training dataset were fused to create the template. Inference on the test data was then performed using a second registration step, either with ANTs SyN or SyNQuick. The performance of the proposed conditional atlas learning frameworks was evaluated using a set of quantitative metrics. The average deformation norm was used to measure the spatial distance between the generated template and the warped subject image. The Jacobian determinant was calculated to assess the smoothness and regularity of the deformation fields. Additionally, the EFC was applied to evaluate the image quality of the generated templates. Segmentation accuracy was evaluated using the DSC for spatial overlap, the HD95 distance for boundary precision, and VS to detect over- or under-segmentation. The proposed pipelines introduce a minimal preprocessing approach, requiring only rigid alignment of the SRR scans and no prior assumptions within the framework. In addition to providing a structural representation of the fetal brain, the models enable atlas-based segmentation of six key brain structures: eCSF, cGM, total WM, deep GM, ventricles, and miscellaneous brain regions. This approach achieves state-of-the-art segmentation performance in real-time. Finally, by analyzing the volumetric data of the anatomical labels from the dataset, as well as from the predicted templates, trajectories of neurotypical brain development were extracted. These trajectories highlight the complex, non-linear growth patterns characteristic of the developing fetal brain.

Acronyms

ADAM Adaptive Moment Estimation. 29

BIDS Brain Imaging Data Structure. 54

cGM Cortical Gray Matter. 7, 16, 39, 41–43, 46, 56, 63, 64, 67–69, 75–77, 79, 101, 103, 104

CNN Convolutional Neural Network. 17, 27, 29, 30, 32, 47, 50, 79

CSF Cerebrospinal Fluid. 13, 14

CT Computed Tomography. 13, 21, 78

DDPMs Denoising Diffusion Probabilistic Models. 31

DICOM Digital Imaging and Communications in Medicine. 53, 54

DOF Degrees Of Freedom. 21

DSC Dice Similarity Coefficient. 3, 4, 18, 52, 56, 60, 67, 72–75, 77, 79, 101, 103, 104

eCSF External Cerebrospinal Fluid. 8, 46, 56, 65, 67–69, 75–77, 79, 101, 103, 104

EFC Entropy Focus Criteria. 3, 4, 51, 62–64, 74, 79

FC Fully Connected. 27, 29

FiLM Feature-wise Linear Modulation. 49, 50, 57, 63, 64, 67, 71, 72, 74, 103, 104

GA Gestational Age. 2, 4, 34, 52, 61, 62, 64, 65, 68, 75, 79

GANs Generative Adversarial Networks. 17, 30, 31, 50, 60, 77

GD Gradient Descent. 29

GM Gray Matter. 11, 46, 56, 67–69, 76, 79, 101, 103, 104

GRE Gradient Recalled Echo. 16

GT Ground Truth. 7, 40, 48, 52, 55, 68, 69, 77

GW Gestational Week. 1–3, 9, 11, 12, 16–18, 34–36, 38, 39, 41, 43, 53, 56, 60–66, 68, 69, 72, 74–77, 79, 101, 102, 105, 106

HD95 95th Percentile Hausdorff Distance. 3, 4, 52, 67, 74, 79, 103, 104

IQA Image Quality Assessment. 61, 62, 66, 68, 69, 102

LV Lateral Ventricles. 16, 56, 67

MI Mutual Information. 21, 25, 26, 35, 36

MRI Magnetic Resonance Imaging. 1–4, 7, 9, 11–18, 21, 23, 25, 26, 31, 34, 46, 52–56, 60, 77–79

MSE Mean Squared Error. 18, 26, 35, 36, 48

NCC Normalized Cross Correlation. 21, 26, 35, 36, 38, 48, 71

NifTI Neuroimaging Informatics Technology Initiative. 33, 54

ODEs Ordinary Differential Equations. 32

PCW Post-conception Weeks. 9–11, 18

PET Positron Emission Tomography. 13, 26

ReLU Rectifier Linear Unit. 28

RF Radiofrequency. 12, 13, 15, 16, 55

SAR Specific Absorption Rate. 16

SNR Signal-to-noise Ratio. 1, 18, 53

SRR Super-Resolution Reconstruction. 1, 4, 33, 34, 54–56, 61, 66, 68, 69, 73, 77, 79, 102

SSIM Structural Similarity Index Measure. 17, 18, 21, 26, 35, 36

STN Spatial Transformer Network. 30, 31, 47

TE Echo Time. 13, 53

TPS Thin Plate Splines. 24, 25

TR Repetition Time. 13, 53

TSE Turbo Spin Echo. 16, 53

US Ultrasound. 1, 12, 53, 76, 78

ViT Vision Transformer. 17, 19, 31, 32

VS Volumetric Similarity. 3, 4, 52, 67, 79, 103, 104

WM White Matter. 1, 8, 11, 16, 46, 56, 64, 67–69, 75–77, 79, 101, 103, 104

Bibliography

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017.

[Ack92] Sandra Ackerman. *Discovering the Brain*. Washington (DC): National Academies Press (US), 1992.

[AdPM+17] Nickie 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. *Cerebral cortex (New York, N.Y. : 1991)*, 27, 10 2017.

[AG16] Richard Ansorge and Martin Graves. *The Physics and Mathematics of MRI*. 2053-2571. Morgan & Claypool Publishers, 2016.

[AHS+97] D. Atkinson, D.L.G. Hill, P.N.R. Stoyle, P.E. Summers, and S.F. Keevil. Automatic correction of motion artifacts in magnetic resonance images using an entropy focus criterion. *IEEE Transactions on Medical Imaging*, 16(6):903–910, 1997.

[AJF+20] Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. MedGAN: Medical image translation using GANs. *Computerized Medical Imaging and Graphics*, 79:101684, January 2020.

[APS12] Friedrich Anderhuber, Franz Pera, and Johannes Streicher, editors. *Waldeyer - Anatomie des Menschen*. De Gruyter, Berlin, Boston, 2012.

[Ash07] John Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95–113, 2007.

[ASS20] ShahRukh Athar, Zhixin Shu, and Dimitris Samaras. Self-supervised deformation modeling for facial expression editing. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 294–301, 2020.

[AYP+10] Brian B. Avants, Paul Yushkevich, John Pluta, David Minkoff, Marc Korczykowski, John Detre, and James C. Gee. The optimal template effect in hippocampus studies of diseased populations. *NeuroImage*, 49(3):2457–2466, 2010.

[BCH+14] Robert W. Brown, Yu-Chung N. Cheng, E. Mark Haacke, Michael R. Thompson, and Ramesh Venkatesan. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. John Wiley Sons, Ltd, 2nd edition, 04 2014.

[BMPP22] Subrato Bharati, M. Rubaiyat Hossain Mondal, Prajoy Podder, and V. B. Surya Prasath. Deep learning for medical image registration: A comprehensive review. *International Journal of Computer Information Systems and Industrial Management Applications*, 14:173–190, 2022.

[Boo89] F.L. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989.

[BS13] Oliver Bieri and Klaus Scheffler. Fundamentals of balanced steady state free precession MRI. *Journal of Magnetic Resonance Imaging*, 38(1):2–11, 2013.

[BSK15] Silvia Budday, Paul Steinmann, and Ellen Kuhl. Physical biology of human brain development. *Frontiers in Cellular Neuroscience*, 9, 2015.

[BZS+18] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John V. Guttag, and Adrian V. Dalca. Voxelmorph: A learning Framework for Deformable Medical Image Registration. *CoRR*, abs/1809.05231, 2018.

[CAB+04] R.W. Cox, John Ashburner, Hester Breman, Kate Fissell, C. Haselgrove, C.J. Holmes, J.L. Lancaster, D.E. Rex, S.M. Smith, J.B. Woodward, and Stephen Strother. A (sort of) new image data format standard: Nifti-1. volume 22, 01 2004.

[CCM+24] Tommaso Ciceri, Luca Casartelli, Florian Montano, Stefania Conte, Letizia Squarcina, Alessandra Bertoldo, Nivedita Agarwal, Paolo Brambilla, and Denis Peruzzo. Fetal brain MRI atlases and datasets: A review. *NeuroImage*, 292:120603, 2024.

[CFH+22] Junyu Chen, Eric C. Frey, Yufan He, William P. Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, November 2022.

[CHC+12] Stuart Currie, Nigel Hoggard, Ian Craven, Marios Hadjivassiliou, and Iain Wilkinson. Understanding MRI: Basic MR physics for physicians. *Postgraduate medical journal*, 89, 12 2012.

[CHF+21] Junyu Chen, Yufan He, Eric Frey, Ye Li, and Yong Du. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. In *Medical Imaging with Deep Learning*, 2021.

[Cle11] Matthew Clemence. *How to Shorten MRI Sequences*, pages 19–32. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[CLW+25] Junyu Chen, Yihao Liu, Shuwen Wei, Zhangxing Bian, Shalini Subramanian, Aaron Carass, Jerry L. Prince, and Yong Du. A survey on deep learning in medical image registration: New technologies, uncertainty, evaluation metrics, and beyond. *Medical Image Analysis*, 100:103385, 2025.

[CSG+23] Tommaso Ciceri, Letizia Squarcina, Alice Giubergia, Alessandra Bertoldo, Paolo Brambilla, and Denis Peruzzo. Review on deep learning fetal brain segmentation from magnetic resonance images. *Artificial Intelligence in Medicine*, 143:102608, 2023.

[CT14] Herng-Hua Chang and Chih-Yuan Tsai. Adaptive registration of magnetic resonance images based on a viscous fluid model. *Computer Methods and Programs in Biomedicine*, 117, 11 2014.

[CWW+22] Lixuan Chen, Jiangjie Wu, Qing Wu, Hongjiang Wei, and Yuyao Zhang. Continuous longitudinal fetus brain atlas construction via implicit neural representation. In *Perinatal, Preterm and Paediatric Image Analysis*, pages 38–47, Cham, 2022. Springer Nature Switzerland.

[DBGS19] Adrian V. Dalca, Guha Balakrishnan, John V. Guttag, and Mert R. Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *CoRR*, abs/1903.03545, 2019.

[DBK+21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[DFS99] Anders M. Dale, Bruce Fischl, and Martin I. Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999.

[DGP15] Ivana Despotović, Bart Goossens, and Wilfried Philips. MRI segmentation of the human brain: Challenges, methods, and applications. *Computational and Mathematical Methods in Medicine*, 2015(1):450341, 2015.

[DJM22] Vincent Dumont, Xiangyang Ju, and Juliane Mueller. Hyperparameter optimization of generative adversarial network models for high-energy physics simulations. *arXiv preprint arXiv:2208.07715*, 2022.

[DKCG+24] Maik Dannecker, Vanessa Kyriakopoulou, Lucilio Cordero-Grande, Anthony N. Price, Joseph V. Hajnal, and Daniel Rueckert. Cina: Conditional implicit neural atlas for spatio-temporal representation of fetal brains. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 181–191, Cham, 2024. Springer Nature Switzerland.

[DKK21] Jordan Dekraker, Stefan Köhler, and Ali Khan. Surface-based hippocampal subfield segmentation. *Trends in neurosciences*, 44, 07 2021.

[DMVS03] Emiliano D'Agostino, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. A viscous fluid model for multimodal non-rigid image registration using mutual information. *Medical Image Analysis*, 7(4):565–575, 2003. Medical Image Computing and Computer Assisted Intervention.

[DRDG21] Neel Dey, Mengwei Ren, Adrian V. Dalca, and Guido Gerig. Generative adversarial registration for improved conditional deformable templates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3929–3941, October 2021.

[DRGS19] Adrian V. Dalca, Marianne Rakic, John V. Guttag, and Mert R. Sabuncu. Learning Conditional Deformable Templates with Convolutional Networks. *CoRR*, abs/1908.02738, 2019.

[DU18] Ugur Demir and Gozde Unal. Patch-based image inpainting with generative adversarial networks. *arXiv preprint arXiv:1803.07422*, 2018.

[dVBV+17] Bob D. de Vos, Floris F. Berendsen, Max A. Viergever, Marius Staring, and Ivana Išgum. *End-to-End Unsupervised Deformable Image Registration with a Convolutional Neural Network*, page 204–212. Springer International Publishing, 2017.

[dVBV+19] Bob D. de Vos, Floris F. Berendsen, Max A. Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical Image Analysis*, 52:128–143, February 2019.

[Els93] A D Elster. Gradient-echo MR imaging: techniques and acronyms. *Radiology*, 186(1):1–8, 1993. PMID: 8416546.

[EWL+20] Michael Ebner, Guotai Wang, Wenqi Li, Michael Aertsen, Premal A. Patel, Rosalind Aughwane, Andrew Melbourne, Tom Doel, Steven Dymarkowski, Paolo De Coppi, Anna L. David, Jan Deprest, Sébastien

Ourselin, and Tom Vercauteren. An automated framework for localization, segmentation and super-resolution reconstruction of fetal brain MRI. *NeuroImage*, 206, 2020.

[FVM+21] Lucas Fidon, Elizabeth Viola, Nada Mufti, Anna David, Andrew Melbourne, Philippe Demaerel, Sébastien Ourselin, Tom Vercauteren, Jan Deprest, and Michael Aertsen. A spatio-temporal atlas of the developing fetal brain with spina bifida aperta. *Open Research Europe*, 1:123, 10 2021.

[GAC+16] Krzysztof Gorgolewski, Tibor Auer, Vince Calhoun, Cameron Craddock, Samir Das, Eugene Duff, Guillaume Flandin, Satrajit Ghosh, Tristan Glatard, Yaroslav Halchenko, Daniel Handwerker, Michael Hanke, David Keator, Xiangrui Li, Zachary Michael, Camille Maumet, Boyce Nichols, Thomas Nichols, John Pellman, and Russell Poldrack. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific data*, 3:160044, 06 2016.

[GCB+01] C Garel, E Chantrel, Herve Brisse, Monique Elmaleh-Bergès, D Luton, J Oury, G Sebag, and Maram Hassan. Fetal cerebral Cortex: Normal gestational landmarks identified using prenatal MR imaging. *American Journal of Neuroradiology*, 22, 02 2001.

[GHC+13] Ioannis Gousias, Alexander Hammers, Serena Counsell, Latha Srinivasan, Mary Rutherford, Rolf Heckemann, Jo Hajnal, Daniel Rueckert, and David Edwards. Magnetic resonance imaging of the newborn brain: Automatic segmentation of brain images into 50 anatomical regions. *PloS one*, 8:e59990, 04 2013.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Y. Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3, 06 2014.

[GRVA+17] Ali Gholipour, Caitlin Rollins, Clemente Velasco-Annis, Abdelhakim Ouaalam, Alireza Akhondi-Asl, Onur Afacan, Cynthia Ortinau, Sean Clancy, Catherine Limperopoulos, Edward Yang, Judy Estroff, and Simon Warfield. A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth. *Scientific Reports*, 7, 12 2017.

[GS24] Janos Groh and Mikael Simons. White matter aging and its impact on brain function. *Neuron*, 113, 2024.

[HBHH01] Derek Hill, Philipp Batchelor, Mark Holden, and David Hawkes. Medical image registration. *Physics in medicine and biology*, 46:R1–45, 04 2001.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[HKCD+10] 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. *NeuroImage*, 53(2), 2010.

[HKR+10] Piotr A. Habas, Kio Kim, Francois Rousseau, Orit A. Glenn, A. James Barkovich, and Colin Studholme. Atlas-based segmentation of developing tissues in the human brain with quantitative validation in young fetuses. *Human Brain Mapping*, 31(9), 2010.

[IGMA18] Sakshi Indolia, Anil Kumar Goswami, S.P. Mishra, and Pooja Asopa. Conceptual understanding of convolutional neural network- a deep learning approach. *Procedia Computer Science*, 132:679–688, 2018. International Conference on Computational Intelligence and Data Science.

[IILU+18] Juan Eugenio Iglesias, Ricardo Insausti, Garikoitz Lerma-Usabiaga, Martina Bocchetta, Koen Van Leemput, Douglas N. Greve, Andre van der Kouwe, Bruce Fischl, César Caballero-Gaudes, and Pedro M. Paz-Alonso. A probabilistic atlas of the human thalamic nuclei combining ex vivo MRI and histology. *NeuroImage*, 183:314 – 326, 2018. Cited by: 301; All Open Access, Hybrid Gold Open Access.

[JG19] Deborah A. Jarvis and Paul D. Griffiths. Current state of MRI of the fetal brain in utero. *Journal of Magnetic Resonance Imaging*, 49(3):632–646, 2019.

[JSZK15] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *CoRR*, abs/1506.02025, 2015.

[Jud11] Miloš Judaš. *Prenatal Development of the Human Fetal Telencephalon*, pages 81–146. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[Kab21] Hiroyuki Kabasawa. MR imaging in the 21st century: Technical innovation over the first two decades. *Magnetic Resonance in Medical Sciences*, 21:71 – 82, 2021.

[KB15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[KHY22] Boah Kim, Inhwa Han, and Jong Chul Ye. Diffusemorph: Unsupervised deformable image registration using diffusion model. In *Computer Vision – ECCV 2022*, pages 347–364, Cham, 2022. Springer Nature Switzerland.

[KMAS+11] Maria Kuklisova-Murgasova, Paul Aljabar, Latha Srinivasan, Serena J. Counsell, Valentina Doria, Ahmed Serag, Ioannis S. Gousias, James P. Boardman, Mary A. Rutherford, A. David Edwards, Joseph V. Hajnal, and Daniel Rueckert. A dynamic 4d probabilistic atlas of the developing brain. *NeuroImage*, 54(4):2750–2763, 2011.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[KSJ19] I. Kostović, G. Sedmak, and M. Judaš. Neural histology and neurogenesis of the human fetal and infant brain. *NeuroImage*, 188:743–773, 2019.

[KV18] Hannah C. Kinney and Joseph J. Volpe. Chapter 8 - myelination events. In *Volpe's Neurology of the Newborn (Sixth Edition)*, pages 176–188. Elsevier, sixth edition edition, 2018.

[Lic15] R. Licandro. Longitudinal diffeomorphic fetal brain atlas learning for tissue labeling using geodesic regression and graph cuts. `https://doi.org/10.34726/hss.2015.21890`, 2015. reposiTUm.

[LLC+21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[LLSH22] Z. Leibovitz, Tally Lerman-Sagie, and Leila Haddad. Fetal brain development: Regulating processes and related malformations. *Life*, 12:809, 05 2022.

[LRH+22] Hélène Lajous, Christopher Roy, Tom Hilbert, Priscille de Dumast, Sébastien Tourbier, Yasser Alemán-Gómez, Jérôme Yerly, Thomas Yu, Hamza Kebiri, Kelly Payette, Jean-Baptiste Ledoux, Reto Meuli, Patric Hagmann, András Jakab, Vincent Dunet, Mériam Koob, Tobias Kober, Matthias Stuber, and Meritxell Bach Cuadra. A fetal brain magnetic resonance acquisition numerical phantom (fabian). *Scientific Reports*, 12:8682, 05 2022.

[LSM+21] Liu Li, Matthew Sinclair, Antonios Makropoulos, Joseph V. Hajnal, A. David Edwards, Bernhard Kainz, Daniel Rueckert, and Amir Alansary. Cas-net: Conditional atlas generation and brain segmentation for fetal MRI. In *Uncertainty for Safe Utilization of Machine Learning in Medical*

Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis, pages 221–230, Cham, 2021. Springer International Publishing.

[MBK+97] Charles R. Meyer, Jennifer L. Boes, Boklye Kim, Peyton H. Bland, Kenneth R. Zasadny, Paul V. Kison, Kenneth Koral, Kirk A. Frey, and Richard L. Wahl. Demonstration of accuracy and clinical versatility of mutual information for automatic multimodality image fusion using affine and thin-plate spline warped geometric deformations. *Medical Image Analysis*, 1(3):195–206, 1997.

[MBKB99] C. R. Meyer, J. L. Boes, B. Kim, and P. H. Bland. Probabilistic brain atlas construction: Thin-plate spline warping via maximization of mutual information. In Chris Taylor and Alain Colchester, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI'99*, pages 631–637, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

[MK18] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations*, 2018.

[MLX+17] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least Squares Generative Adversarial Networks . In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821, Los Alamitos, CA, USA, October 2017. IEEE Computer Society.

[MR23] Richard McKinley and Christian Rummel. Cortexmorph: Fast cortical thickness estimation via diffeomorphic registration using voxelmorph. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 730–739, Cham, 2023. Springer Nature Switzerland.

[MSG+10] Maria Murgasova, Latha Srinivasan, Ioannis S. Gousias, Paul Aljabar, Joseph V. Hajnal, A. David Edwards, and Daniel Rueckert. Construction of a dynamic 4d probabilistic atlas for the developing brain. In *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 952–955, 2010.

[MTSO11] K Miller, R Tijssen, N Stikov, and T Okell. Steady-state MRI: Methods for neuroimaging. *Imaging in Medicine*, 3(1):93–105, 2011.

[PCZ+21] Yuchen Pei, Liangjun Chen, Fenqiang Zhao, Zhengwang Wu, Tao Zhong, Ya Wang, Changan Chen, Li Wang, He Zhang, Lisheng Wang, and Gang Li. Learning spatiotemporal probabilistic atlas of fetal brains with anatomically constrained registration network. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 239–248, Cham, 2021. Springer International Publishing.

[PK12] Donald B. Plewes and Walter Kucharczyk. Physics of MRI: A primer. *Journal of Magnetic Resonance Imaging*, 35(5):1038–1054, 2012.

[PKK+06] Daniela Prayer, Gregor Kasprian, Elisabeth Krampl, Barbara Ulm, Linde Witzani, Lucas Prayer, and Peter C. Brugger. MRI of normal fetal brain development. *European Journal of Radiology*, 57(2), 2006. Fetal Imaging.

[PLdD+23] Kelly Payette, Hongwei Bran Li, Priscille de Dumast, Roxane Licandro, Hui Ji, Md Mahfuzur Rahman Siddiquee, Daguang Xu, Andriy Myronenko, Hao Liu, Yuchen Pei, Lisheng Wang, Ying Peng, Juanying Xie, Huiquan Zhang, Guiming Dong, Hao Fu, Guotai Wang, ZunHyan Rieu, Donghyeon Kim, Hyun Gi Kim, Davood Karimi, Ali Gholipour, Helena R. Torres, Bruno Oliveira, João L. Vilaça, Yang Lin, Netanell Avisdris, Ori Ben-Zvi, Dafna Ben Bashat, Lucas Fidon, Michael Aertsen, Tom Vercauteren, Daniel Sobotka, Georg Langs, Mireia Alenyà, Maria Inmaculada Villanueva, Oscar Camara, Bella Specktor Fadida, Leo Joskowicz, Liao Weibin, Lv Yi, Li Xuesong, Moona Mazher, Abdul Qayyum, Domenec Puig, Hamza Kebiri, Zelin Zhang, Xinyi Xu, Dan Wu, Kuanlun Liao, Yixuan Wu, Jintai Chen, Yunzhi Xu, Li Zhao, Lana Vasung, Bjoern Menze, Meritxell Bach Cuadra, and Andras Jakab. Fetal brain tissue annotation and segmentation challenge results. *Medical Image Analysis*, 88:102833, 2023.

[PMDC+23] D. Prayer, G. Malinger, L. De Catte, B. De Keersmaecker, L. F. Gonçalves, G. Kasprian, S. Laifer-Narin, W. Lee, A.-E. Millischer, L. Platt, F. Prayer, D. Pugash, L. J. Salomon, M. Sanz Cortes, F. Stuhr, I. E. Timor-Tritsch, B. Tutschek, D. Twickler, N. Raine-Fenning, and the ISUOG Clinical Standards Committee. ISUOG practice guidelines (updated): performance of fetal magnetic resonance imaging. *Ultrasound in Obstetrics & Gynecology*, 61(2):278–287, 2023.

[PSdV+18] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

[QL23] Yi Qin and Xiaomeng Li. Fsdiffreg: Feature-wise and score-wise diffusion-guided unsupervised deformable image registration for cardiac images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, 2023.

[REYT24] Hiba Ramadan, Dounia El Bourakadi, Ali Yahyaouy, and Hamid Tairi. Medical image registration in the era of transformers: A recent review. *Informatics in Medicine Unlocked*, 49:101540, 2024.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2015*, 2015.

[RRAVTV+20] Manuel Recio Rodríguez, Cristina Andreu-Vázquez, Israel J Thuissard-Vasallo, Raquel Cano Alonso, Carmina Bermejo López, Ines Tamarit Degenhardt, and Pilar Martínez Ten. Real-life Diagnostic Accuracy of MRI in Prenatal Diagnosis. *Radiology research and practice*, 2020, 2020.

[RSH+99] D. Rueckert, Luke Sonoda, C. Hayes, Derek Hill, M.O. Leach, and D.J. Hawkes. Nonrigid registration using free-form deformations: Application to breast MR images. *Medical Imaging, IEEE Transactions on*, 18:712 – 721, 09 1999.

[RW98] P. Royston and E.M. Wright. How to construct 'normal ranges' for fetal variables. *Ultrasound in Obstetrics & Gynecology*, 11(1):30–38, 1998.

[SAB+12] Ahmed Serag, Paul Aljabar, Gareth Ball, Serena Counsell, James Boardman, Mary Rutherford, David Edwards, Joseph Hajnal, and Daniel Rueckert. Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *NeuroImage*, 59:2255–65, 02 2012.

[Sal14] Sahar N. Saleem. Fetal MRI: An approach to practice: A review. *Journal of Advanced Research*, 5(5):507–523, 2014.

[SBL+14] Denis P. Shamonin, Esther E. Bron, Boudewijn P. Lelieveldt, Marion Smits, Stefan Klein, and Marius Staring. Fast parallel image registration on CPU and GPU for diagnostic classification of alzheimer's disease. *Frontiers in Neuroinformatics*, 7, 2014.

[Sch15] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[SDG+21] Ernst Schwartz, Mariana Cardoso Diogo, Sarah Glatter, Rainer Seidl, Peter C Brugger, Gerlinde M Gruber, Herbert Kiss, Karl-Heinz Nenning, IRC5 consortium, Georg Langs, Daniela Prayer, and Gregor Kasprian. The prenatal morphomechanic impact of agenesis of the corpus callosum on human brain structure and asymmetry. *Cerebral Cortex*, 31(9):4024–4037, 04 2021.

[SDVB+17] Hessam Sokooti, Bob De Vos, Floris Berendsen, Boudewijn Lelieveldt, Ivana Išgum, and Marius Staring. Nonrigid image registration using multi-scale 3D convolutional neural networks. pages 232–239, 09 2017.

[SE19] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[SKS13] James Shackleford, Nagarajan Kandasamy, and Gregory Sharp. Chapter 2 - unimodal b-spline registration. In *High Performance Deformable*

Image Registration Algorithms for Manycore Processors, pages 13–43. Morgan Kaufmann, Boston, 2013.

[SMA+08] David W. Shattuck, Mubeena Mirza, Vitria Adisetiyo, Cornelius Hojatkashani, Georges Salamon, Katherine L. Narr, Russell A. Poldrack, Robert M. Bilder, and Arthur W. Toga. Construction of a 3D probabilistic atlas of human cortical structures. *NeuroImage*, 39(3):1064–1080, 2008.

[SMR+18] Andreas Schuh, Antonios Makropoulos, Emma C. Robinson, Lucilio Cordero-Grande, Emer Hughes, Jana Hutter, Anthony N. Price, Maria Murgasova, Rui Pedro A. G. Teixeira, Nora Tusor, Johannes K. Steinweg, Suresh Victor, Mary A. Rutherford, Joseph V. Hajnal, A. David Edwards, and Daniel Rueckert. Unbiased construction of a temporally consistent morphological atlas of neonatal brain development. *bioRxiv*, 2018.

[SP86] Thomas Sederberg and Scott Parry. Free-form deformation of solid geometric models. *ACM Siggraph Computer Graphics*, 20:151–160, 08 1986.

[Sti08] Joan Stiles. *The Fundamentals of Brain Development: Integrating Nature and Nurture*. Harvard University Press, 2008.

[TH15] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, 15, 08 2015.

[Thi98] J.-P. Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical Image Analysis*, 2(3):243–260, 1998.

[TRH+04] Duygu Tosun, Maryam Rettmann, Xiao Han, Xiaodong Tao, Chenyang Xu, Susan Resnick, Dzung Pham, and Jerry Prince. Cortical surface segmentation and mapping. *NeuroImage*, 23 Suppl 1:S108–18, 02 2004.

[TSN+22] Athena Taymourtash, Ernst Schwartz, Karl-Heinz Nenning, Daniel Sobotka, Roxane Licandro, Sarah Glatter, Mariana Cardoso Diogo, Polina Golland, Ellen Grant, Daniela Prayer, Gregor Kasprian, and Georg Langs. Fetal development of functional thalamocortical and cortico–cortical connectivity. *Cerebral Cortex*, 33(9):5613–5624, 12 2022.

[TT88] J. Talairich and P. Tournoux. *Co-Planar Stereotaxic Atlas of the Human Brain 3-Dimensional Proportional System: An Approach to Cerebral Imaging*. Thieme Medical Publisher, New York, 1988.

[UKG+23] A Uus, V Kyriakopoulou, L Cordero Grande, D Christiaens, M Pietsch, A Price, S Wilson, P Patkee, S Karolis, A Schuh, A Gartner, L Williams, E Hughes, T Arichi, J O'Muircheartaigh, J Hutter, E Robinson,

JD Tournier, D Rueckert, S Counsell, M Rutherford, M Deprez, JV Hajnal, and AD Edwards. Multi-channel spatio-temporal MRI atlas of the normal fetal brain development from the developing human connectome project, 2023. G-Node. https://doi.org/10.12751/g-node.qj5hs7.

[UKM+23] Alena U. Uus, Vanessa Kyriakopoulou, Antonios Makropoulos, Abi Fukami-Gartner, Daniel Cromb, Alice Davidson, Lucilio Cordero-Grande, Anthony N. Price, Irina Grigorescu, Logan Z. J. Williams, Emma C. Robinson, David Lloyd, Kuberan Pushparajah, Lisa Story, Jana Hutter, Serena J. Counsell, A. David Edwards, Mary A. Rutherford, Joseph V. Hajnal, and Maria Deprez. BOUNTI: Brain volumetry and automated parcellation for 3D fetal MRI. *bioRxiv*, 2023.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008. Curran Associates, Inc., 2017.

[WDO+05] He Wang, Lei Dong, Jennifer O'Daniel, Radhe Mohan, Adam Garden, K Ang, Deborah Kuban, Mark Bonnen, Joe Chang, and Rex Cheung. Validation of an accelerated "demons" algorithm for deformable image registration in radiation therapy. *Physics in medicine and biology*, 50:2887–905, 07 2005.

[WEA+16] Tuangsit Wataganara, Alaa Ebrashy, Labaran Dayyabu Aliyu, Renato Augusto Moreira de Sa, Ritsuko Pooh, Asim Kurjak, Cihat Sen, Abdallah Adra, and Milan Stanojevic. Fetal magnetic resonance imaging and ultrasound. *Journal of Perinatal Medicine*, 44(5):533–542, 2016.

[XLG+20] Junshen Xu, Sayeri Lala, Borjan Gagoski, Esra Abaci Turk, P Ellen Grant, Polina Golland, and Elfar Adalsteinsson. Semi-supervised learning for fetal brain MRI quality assessment with ROI consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 386–395. Springer, 2020.

[XMG+23] Junshen Xu, Daniel Moyer, Borjan Gagoski, Juan Eugenio Iglesias, P. Ellen Grant, Polina Golland, and Elfar Adalsteinsson. Nesvor: Implicit neural representation for slice-to-volume reconstruction in MRI. *IEEE Transactions on Medical Imaging*, pages 1–1, 2023.

[XSS+22] Xu & Sun, Jiwei Sun, Wen Shi, Yao Shen, Ruoke Zhao, Wanrong Luo, Mingyang Li, Guangbin Wang, and Dan Wu. Spatiotemporal atlas of the fetal brain depicts cortical developmental gradient. *Journal of Neuroscience*, 42(50):9435–9449, 2022.

[YMD+15] Zichao Yang, Marcin Moczulski, Misha Denil, Nando De Freitas, Le Song, and Ziyu Wang. Deep fried convnets. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1476–1483, 2015.

[ZGSQ22] Jing Zou, Bingchen Gao, Youyi Song, and Jing Qin. A review of deep learning-based deformable medical image registration. *Frontiers in Oncology*, 12, 2022.

[ZHZ+25] Kai Zhang, Shijie Huang, Fangmei Zhu, Zhongxiang Ding, Geng Chen, and Dinggang Shen. Rethinking fetal brain atlas construction: A deep learning perspective. In *Perinatal, Preterm and Paediatric Image Analysis*, pages 94–104, Cham, 2025. Springer Nature Switzerland.

[ZLL+20] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[ZPC+20] Yujia Zhou, Shumao Pang, Jun Cheng, Yuhang Sun, Yi Wu, Lei Zhao, Yaqin Liu, Zhentai Lu, Wei Yang, and Qianjin Feng. Unsupervised deformable medical image registration via pyramidal residual deformation fields estimation. *arXiv preprint arXiv:2004.07624*, 2020.

[ZPIE20] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2020.

[ZSJ+22] Yuanjie Zheng, Xiaodan Sui, Yanyun Jiang, Tongtong Che, Shaoting Zhang, Jie Yang, and Hongsheng Li. Symreg-GAN: Symmetric image registration with generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5631–5646, 2022.

[ZWC+18] Yuyao Zhang, Hongjiang Wei, Matthew J. Cronin, Naying He, Fuhua Yan, and Chunlei Liu. Longitudinal atlas for normative human brain development and aging over the lifespan using quantitative susceptibility mapping. *NeuroImage*, 171:176–189, 2018.

Appendix

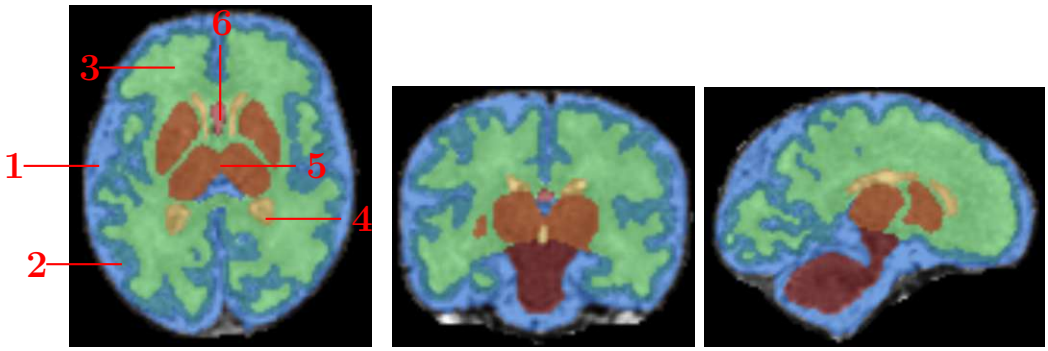


Figure 1: Segmentation of a fetus at 35 GW in axial, coronal and sagittal view (left to right). The segmented tissues are: (1) eCSF (light blue), (2) cGM (dark blue), (3) total WM (green), (4) ventricles (yellow), (5) deep GM (light red), and (6) other (dark red).

Table 1: Grid search of the hyperparameters regularizing the anatomical constraints and smoothness of the deformation field. The best performance is achieved by highest DSC and lowest mean deformation

$\mathcal{L}_{DEF}$	$\mathcal{L}_{SMOOTH}$	Avg DSC (↑)	Avg ||Def|| (↓)
0.1	0.1	0.821	2580.414
0.1	1	0.852	987.864
0.1	10	0.779	695.510
0.25	1	0.825	855.668
0.5	1	0.801	789.178
0.75	1	0.774	716.214
1	0.1	0.789	994.507
1	1	0.768	621.578
1	10	0.667	215.534

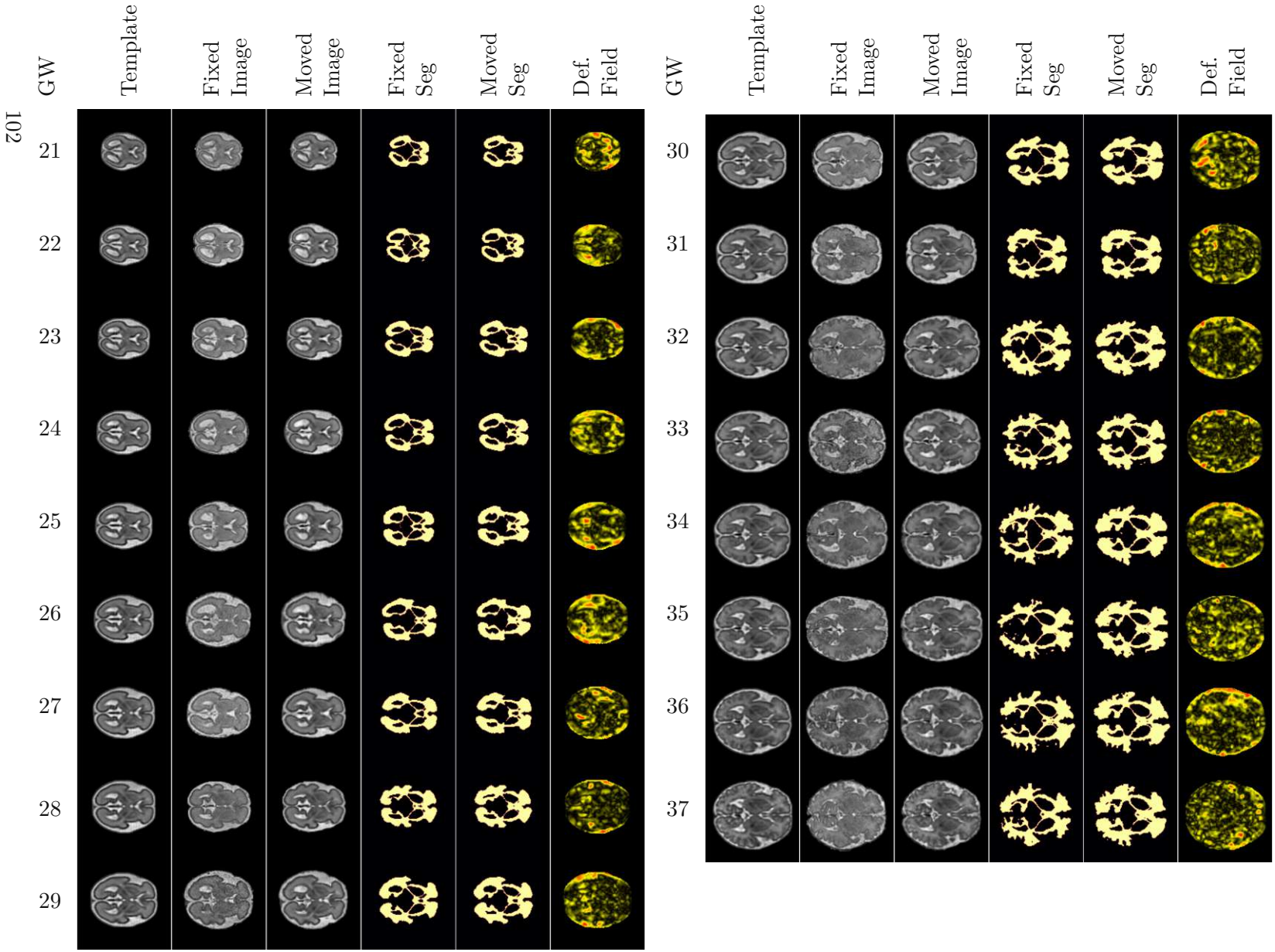


Figure 2: Grid figure visualizing the different steps of the trained model for different GW. Age-specific generated template, fixed image, moved image, fixed segmentation, moved segmentation, deformation field in x-direction (columns from left to right). The test image quality of the SRR is IQA=2.

Table 2: Segmentation evaluation of the proposed deep learning approaches. Average DSC, HD95, and VS for all six labels are reported for the models CAL-REG and CAL-GAN with and without the FiLM layer implementation.

Avg DSC (↑)							
	eCSF	cGM	tWM	Ven	dGM	Other	Total
ANTs SyNQuick	0.803 ± 0.047	0.593 ± 0.054	0.863 ± 0.038	0.776 ± 0.041	0.882 ± 0.032	0.910 ± 0.021	0.805 ± 0.103
ANTs SyN	0.865 ± 0.042	0.750 ± 0.069	0.911 ± 0.039	0.866 ± 0.027	0.909 ± 0.021	0.932 ± 0.013	0.872 ± 0.061
CAL-REG w/ FiLM	0.826 ± 0.062	0.723 ± 0.059	0.890 ± 0.033	0.820 ± 0.040	0.871 ± 0.022	0.913 ± 0.016	0.841 ± 0.019
CAL-REG w/o FiLM	0.845 ± 0.030	0.744 ± 0.041	0.899 ± 0.023	0.797 ± 0.055	0.859 ± 0.031	0.893 ± 0.028	0.840 ± 0.060
CAL-GAN w/ FiLM	0.865 ± 0.044	0.768 ± 0.050	0.904 ± 0.033	0.850 ± 0.036	0.876 ± 0.023	0.915 ± 0.0167	0.863 ± 0.052
Avg HD95 (↓)							
	eCSF	cGM	tWM	Ven	dGM	Other	Total
ANTs SyNQuick	1.608 ± 0.387	1.857 ± 0.444	1.970 ± 0.487	1.882 ± 0.995	1.315 ± 0.351	1.330 ± 0.378	1.660 ± 0.288
ANTs SyN	1.213 ± 0.296	1.361 ± 0.515	1.513 ± 0.574	1.277 ± 1.062	1.073 ± 0.234	1.070 ± 0.176	1.251 ± 0.172
CAL-REG w/ FiLM	1.589 ± 0.580	1.637 ± 0.466	2.018 ± 0.661	1.562 ± 0.419	1.964 ± 0.462	1.483 ± 0.359	1.709 ± 0.225
CAL-REG w/o FiLM	1.555 ± 0.358	1.975 ± 0.722	2.015 ± 0.460	2.266 ± 1.075	2.135 ± 0.482	1.985 ± 0.448	1.989 ± 0.240
CAL-GAN w/ FiLM	1.372 ± 0.384	1.515 ± 0.412	1.890 ± 0.575	1.526 ± 0.480	1.813 ± 0.488	1.492 ± 0.479	1.601 ± 0.203
Avg VS (→ 0 ←)							
	eCSF	cGM	tWM	Ven	dGM	Other	Total
ANTs SyNQuick	0.057 ± 0.114	-0.051 ± 0.072	-0.042 ± 0.068	-0.008 ± 0.273	-0.049 ± 0.080	-0.013 ± 0.072	-0.018 ± 0.041
ANTs SyN	0.055 ± 0.089	-0.035 ± 0.066	-0.018 ± 0.040	0.038 ± 0.078	-0.038 ± 0.048	-0.022 ± 0.045	-0.003 ± 0.040
CAL-REG w/ FiLM	-0.029 ± 0.090	0.187 ± 0.043	0.074 ± 0.056	-0.121 ± 0.072	0.122 ± 0.071	0.007 ± 0.041	0.040 ± 0.111
CAL-REG w/o FiLM	0.113 ± 0.075	0.016 ± 0.100	0.073 ± 0.031	-0.216 ± 0.132	0.125 ± 0.046	0.090 ± 0.051	0.034 ± 0.128
CAL-GAN w/ FiLM	0.007 ± 0.177	0.002 ± 0.142	0.057 ± 0.074	-0.041 ± 0.293	0.024 ± 0.059	0.032 ± 0.076	0.013 ± 0.033

Table 3: Detailed segmentation metrics of CAL-GAN with FiLM across age groups, including DSC, HD95, and VS

	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
DSC (↑)																	
eCSF	.871	.897	.910	.903	.874	.868	.918	.900	.836	.875	.915	.862	.88	.789	.821	.817	.771
cGM	.742	.826	.834	.822	.779	.757	.825	.822	.729	.778	.811	.775	.756	.697	.727	.714	.670
tWM	.923	.934	.943	.938	.917	.912	.933	.934	.908	.906	.922	.900	.882	.862	.868	.850	.833
Ven	.881	.905	.907	.853	.845	.868	.816	.861	.791	.784	.876	.853	.889	.836	.820	.849	.821
dGM	.872	.864	.895	.824	.867	.851	.844	.888	.861	.872	.897	.911	.898	.887	.892	.897	.878
Other	.900	.903	.915	.918	.894	.926	.940	.919	.921	.926	.931	.943	.897	.894	.895	.931	.898
HD95 (↓)																	
eCSF	1.00	1.00	1.00	1.00	1.50	1.21	1.00	1.21	1.62	1.50	1.00	1.41	1.21	1.87	1.71	1.87	2.24
cGM	1.21	1.00	1.00	1.00	1.91	1.50	1.21	1.21	1.62	1.62	1.21	1.41	1.57	2.12	1.83	2.12	2.24
tWM	1.41	1.21	1.00	1.21	1.73	1.93	1.50	1.37	2.00	2.12	1.57	2.12	2.24	2.64	2.53	2.73	2.83
Ven	1.00	1.00	1.00	1.21	1.62	1.21	1.21	1.37	2.00	2.00	1.21	2.29	1.00	1.71	1.97	1.73	2.45
dGM	1.21	1.93	1.21	1.98	1.71	1.83	3.36	1.57	2.12	2.12	1.57	1.37	1.57	1.71	1.71	1.87	1.98
Other	1.00	1.37	1.00	1.21	1.37	1.21	1.00	1.37	1.37	1.41	1.21	1.21	1.98	2.34	2.41	1.57	2.34
VS (→ 0 ←)																	
eCSF	.09	-.07	.03	-.05	.08	-.09	.09	-.11	-.16	.06	.08	-.07	.06	-.09	-.04	.13	.09
cGM	-.04	-.09	.04	-.05	.02	-.07	-.03	-.04	-.19	.06	.02	.11	.03	.02	.01	.18	.03
tWM	.17	.03	.18	.10	.12	.06	.02	.05	.00	.08	.02	.10	.01	.00	.01	.07	-.01
Ven	-.12	.20	.30	-.07	.06	.00	-.30	.03	-.14	-.10	-.03	-.23	.01	-.13	-.23	-.10	-.07
dGM	.08	.04	.07	-.01	.05	.04	-.02	.05	-.06	.06	.02	.06	.00	.03	.05	.10	-.03
Other	.09	.03	.13	.08	.08	.00	.03	.06	-.03	.05	-.03	.07	-.04	.00	.02	.09	-.02

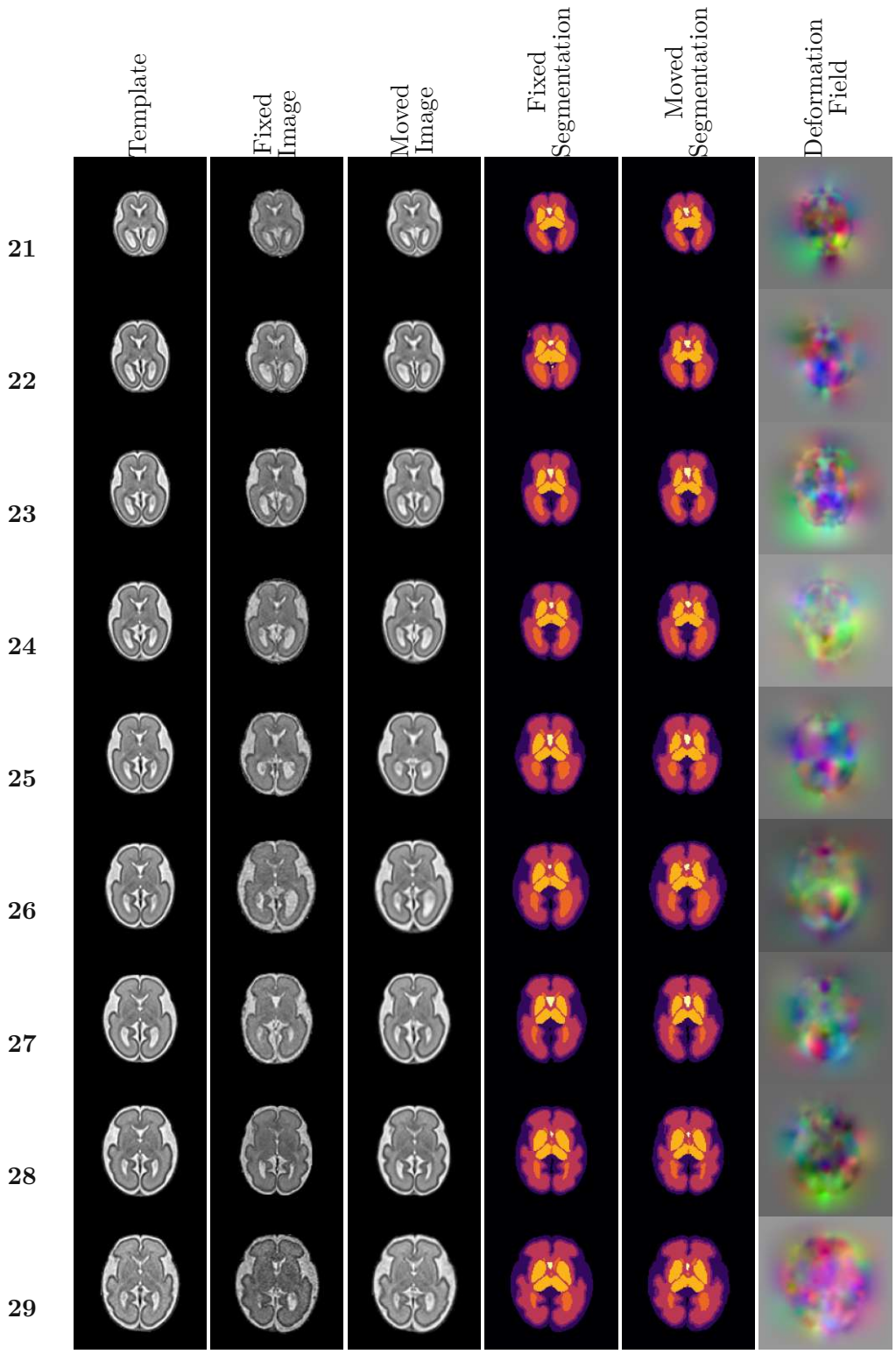


Figure 3: Generated fetal brain atlas with ANTs for the age 21 to 29 GWs. The template and test dataset are registered using the ANTs SyN algorithm.

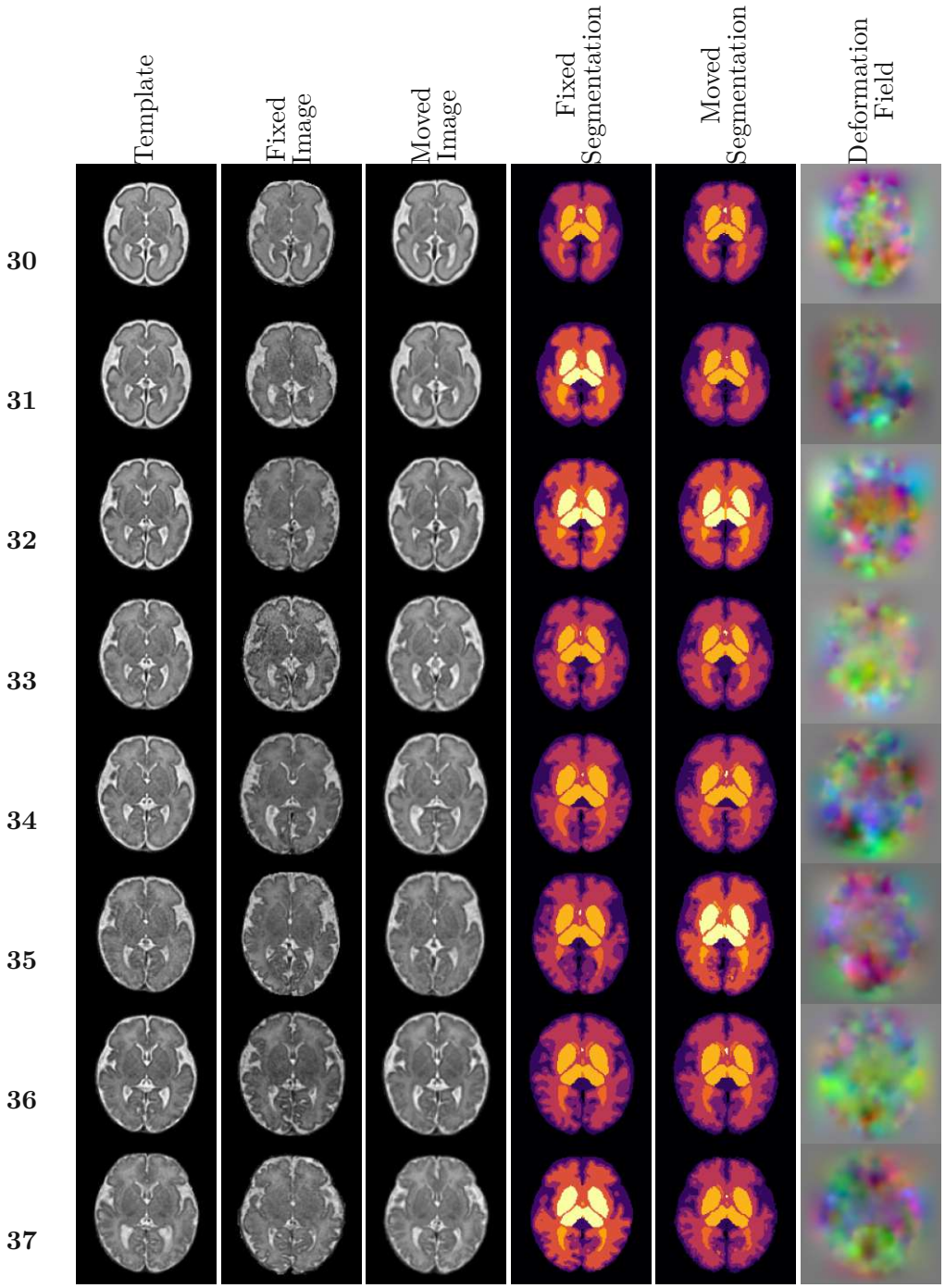


Figure 4: Generated fetal brain atlas with ANTs for the age 30 to 37 GWs. The template and test dataset are registered using the ANTs SyN algorithm.