Year : 2022

Development and validation of robust MR image reconstruction and segmentation techniques for the quantitative analysis of the fetal brain

Guerrier De Dumast Priscille

UNIL | Université de Lausanne

Faculté de biologie
et de médecine

**Département de Radiodiagnostic et radiologie interventionnelle,
Centre Hospitalier Universitaire Vaudois (CHUV) et Université de Lausanne**

Development and validation of robust MR image reconstruction and segmentation techniques for the quantitative analysis of the fetal brain

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Priscille GUERRIER de DUMAST

Ingénieure diplômée de l'École Supérieure de Chimie, Physique, Electronique de Lyon, France

Jury

Prof. Renaud DU PASQUIER, Président
Dr. Meritxell BACH CUADRA, Directrice de thèse
Dr. Maria DEPREZ, Experte
Dr. Bénédicte MARECHAL, Experte
Prof. Anita TRUTTMANN, Experte

**Lausanne
2023**

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président·e	Monsieur	Prof.	Renaud	**Du Pasquier**
Directeur·trice de thèse	Madame	Dre	Merixtell	**Bach Cuadra**
Expert·e·s	Madame	Prof.	Anita	**Truttmann**
	Madame	Dre	Bénédicte	**Maréchal**
	Madame	Dre	Maria	**Deprez**

le Conseil de Faculté autorise l'impression de la thèse de

Priscille Marie Bénédicte Guerrier de Dumast

Diplôme d'ingénieur - Grade de Master, Ecole supérieure de Chimie Physique Electronique de Lyon, France

intitulée

Development and validation of robust MR image reconstruction and segmentation techniques for the quantitative analysis of the fetal brain

Lausanne, le 31 mars 2023

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Renaud Du Pasquier

C'est par votre persévérance que vous garderez votre vie.

Luc 21, 19

Acknowledgments

First, I would like to thank my advisor Dr. Meritxell Bach Cuadra for giving me the opportunity of pursuing my PhD at the University of Lausanne. I am thankful for her support, guidance and trust throughout these four years.

I would like to express my gratitude to Prof. Renaud du Pasquier, Dr. Maria Deprez, Dr. Bénédicte Maréchal and Prof. Anita Truttmann for being part of my thesis committee and for taking the time to read, evaluate and make suggestions to my work.

My sincere gratitude goes to my colleagues for their help and expertise. Dr. Vincent Dunet, Dr. Mériam Kood and Samuel Lamon, it has been a pleasure to work with you and to learn from you the clinical whys and wherefores of our projects. Hamza and Hélène, we started this fetal project together and Thomas came along. Special thanks to Thomas for being so enthusiastic and supportive over the last months despite my "manipulation experte du sarcasme". I am very glad to have shared this journey with great colleagues at the MIAL and more broadly at RC7 and PET3. Maxence Tigrou, your blindtest skills were the joy of the few months we shared in the same office! Solange, from P24L to Dr. Prout, you have never been too far and I could always rely on your advices, sympathic ear and pauses croissants.

Finally, a special thank goes to my friends for their support, counsel and guidance.

Abstract

Formation and development of the human brain is initiated *in utero* and carries on until young adulthood. During the prenatal period, most significant morphological changes occur, following well-defined spatiotemporal patterns. Eventual disruption occurring during these periods of vulnerability may have major impact later in life. It is, therefore, of the utmost importance to get a better understanding of the fetal brain development.

Magnetic resonance imaging (MRI) is a non-invasive technique that relies on the tissue properties of to generate the image intensities. Towards the quantitative analysis, the fetal brain MRI workflow gathers the image acquisition, the image resolution enhancement through super-resolution (SR) reconstruction and the reduction of image complexity with the tissue segmentation. This thesis focuses on the development and validation of robust automated tools for the quantitative analysis of the fetal brain in MRI. Specifically, we address two key steps that encounter fetal-specific challenges: the SR reconstruction and the tissue segmentation. In practice, both are hindered by the major obstacle of data scarcity of fetal brain MRI.

With the impossibility to acquire motion-free high resolution images, the validation of SR reconstruction is troublesome. Our first contribution is a multi-observer multi-dataset study to validate the practical value of SR reconstruction in a clinical environment. We evidence that SR does not introduce spatial distortions and increases the confidence of the observer. Furthermore, we propose a simulation-based approach for the enhancement of the overall SR-reconstructed image intensity contrast.

Automatic tissue segmentation methods must generalize to be robust to the many sources of variations that may be induced by the gestation-long maturation, the acquisition system or the SR reconstruction method. We propose novel data augmentation strategies in order to increase the heterogeneity of the data. Our methods, either relying on a simulation framework or a multi-reconstruction approach, increases the generalizability of deep-learning (DL) based segmentation models. Finally, a major methodological contribution of this thesis is the topologically-constrained DL framework for the cortical plate segmentation.

Overall, our contributions in image reconstruction and tissue segmentation take a step forward in the accuracy, generalizability and translation of methods. Although

some limitations remain, the combination of these advanced engineering methods set solid grounds for the study of the *in utero* brain development.

Résumé

La formation et le développement du cerveau humain commencent *in utero* et se poursuivent jusqu'au début de l'âge adulte. Au cours de la période prénatale les changements morphologiques les plus importants se produisent selon des schémas spatio-temporels bien définis. Les éventuelles perturbations survenant au cours de ces périodes de vulnérabilité peuvent avoir un impact majeur plus tard dans la vie. Il est donc important de mieux comprendre le développement du cerveau fœtal.

L'imagerie par résonance magnétique (IRM) est une technique non invasive qui s'appuie sur les propriétés des tissus pour générer des variations d'intensité, et construire une image. Les différentes étapes de l'analyse quantitative du cerveau fœtal dérivé de l'IRM comprennent l'acquisition de l'image, l'amélioration de la résolution de l'image par la reconstruction en super-résolution (SR) et la réduction de la complexité de l'image par la segmentation des tissus. Cette thèse se concentre sur le développement et la validation d'outils automatisés robustes pour l'analyse quantitative du cerveau fœtal en IRM. Plus précisément, nous abordons deux étapes clés qui rencontrent actuellement des difficultés spécifiques à l'imagerie fœtale : la reconstruction en SR et la segmentation des tissus. En pratique, ces deux étapes sont freinées et éprouvées par la rareté des données IRM du cerveau fœtal.

Avec l'impossibilité d'acquérir des images haute résolution sans mouvement, la validation de la reconstruction SR est difficile. Notre première contribution est une étude multi-observateurs et multi-dataset pour valider la valeur pratique de la reconstruction SR dans un environnement clinique. Nous démontrons que la SR n'introduit pas de distorsions anatomique et augmente l'assurance de l'observateur. En outre, nous proposons une approche basée sur la simulation pour l'amélioration du contraste global d'intensité de l'image reconstruite par SR.

Les méthodes de segmentation automatique des tissus doivent, pour être robustes, être généralisées aux nombreuses sources de variations qui peuvent être induites par la maturation pendant la gestation, le système d'acquisition ou la méthode de reconstruction SR. Nous proposons de nouvelles stratégies d'augmentation de données afin d'en accroître l'hétérogénéité. Nos méthodes, qui s'appuient soit sur la simulation d'images de synthèse, soit sur une approche de multi-reconstruction, augmentent la généralisation des modèles de segmentation par apprentissage automatique. Finalement, la contribution méthodologique majeure de cette thèse est l'intégration d'une

contrainte topologique dans l'entrainement de méthode de segmentation par apprentissage automatique pour le plaque corticale.

Dans l'ensemble, ce travail permet une avancée majeure en terme de précision, généralisation et implémentation des méthodes quant à la reconstruction et à la segmentation des tissus en IRM. Bien que certaines limites subsistent, la combinaison de ces méthodes d'ingénierie avancées constitue une base solide pour l'étude du développement du cerveau *in utero*.

List of Abbreviations

ASSD	Average symmetric surface distance
BN	Betti number
BNE	Betti number error
CC	Corpus callosum
CCC	Concordance correlation coefficient
cGM	Cortical gray matter
CNS	Central nervous system
CSF	Cerebrospinal fluid
dGM	Deep gray matter
DL	Deep-learning
DSC	Dice similarity coefficient
GA	Gestational age
GT	Ground truth
GW	Gestational weeks
HR	High-resolution
ICC	Intraclass correlation coefficient
LR	Low-resolution
MAS	Multi-atlas segmentation
MR	Magnetic resonance
MRI	Magnetic resonance imaging
NT	Neurotypical
PSNR	Peak signal-to-noise ratio
PT	Pathological
SD	Standard deviation
SNR	Signal-to-noise ratio
SR	Super-resolution
SSIM	Structural similarity index
SST2W	Single-shot T2-weighted
SVR	Slice-to-volume
T2w	T2-weighted
T2WS	T2w sequences
US	Ultrasound
WM	White matter

Contents

List of Figures

List of Tables

1
Introduction

1.1 Early human brain development

Description of *in utero* brain development appeared in the literature in 1816 [140], although it really became of interest from the middle of the XXth century thanks to post-mortem histology [20, 10, 11].

1.1.1 Brain growth and tissue maturation

Formation and development of the human brain is initiated *in utero* about two weeks post-conception and carries on until young adulthood [141]. In the prenatal period, environmental and genetic factors command the development, while after birth maturation processes are additionally experience-driven (e.g. external stimuli) [31].

During the *in utero* development, the human brain undergoes its most significant morphological changes, evolving from a smooth surface to a highly convoluted one [148] and featuring a large increase of the head circumference [31]. Underlying the undeniable drastic macrostructural changes of the *in utero* brain development, microstructural operations occur. Overall, maturation and growth mechanisms follow well defined spatio-temporal patterns [10].

Microstructural changes follow a rigorous timeline of synchronized developmental pathways [93, 86, 108]. However, structure-wise and region-wise development are asynchronous, as each region grows at its own pace. For instance, sensory regions develop early on, while others have delayed development which continues after birth (e.g. frontal area develops until the end of adolescence) [31].

Figure 1.1 shows a summary of the main microstructural processes and their correspondance to macrostructural milestones throughout gestation. The first trimester of gestation features the overall organisation of the central nervous system (CNS) that is
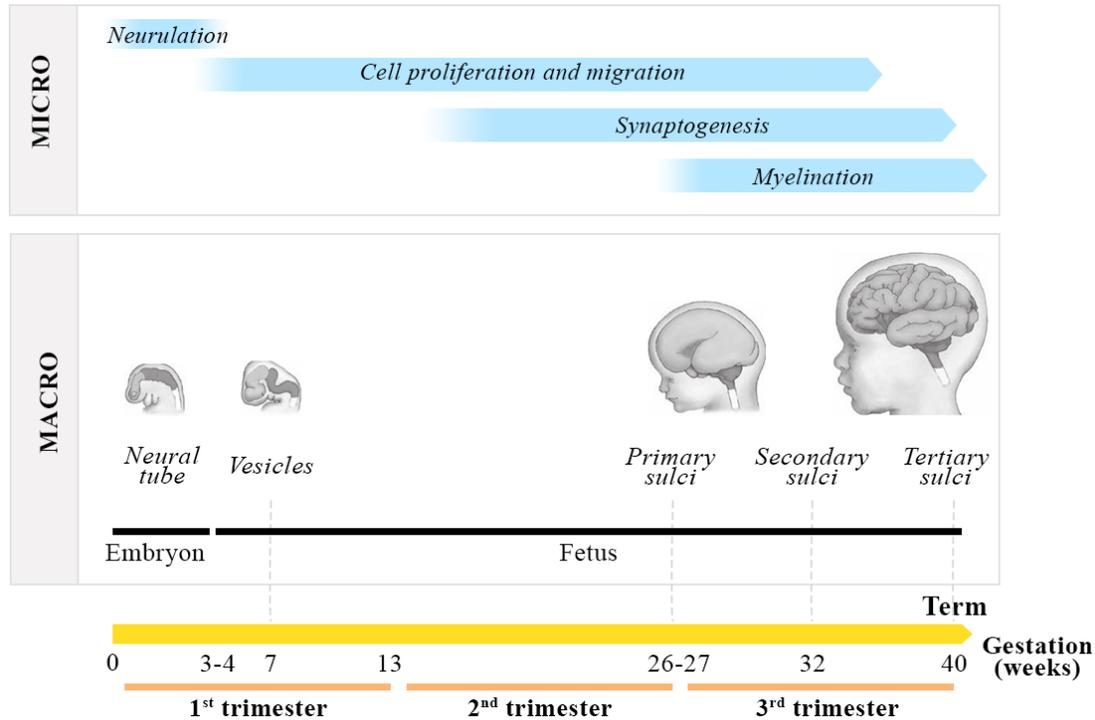
Figure 1.1: Fetal brain maturation throughout gestation. Illustration adapted from [85].

ultimately composed of the brain and the spinal cord. In the second and third trimester, individual portions of CNS further develop, in particular the brain.

The establishment of the CNS is initiated shortly after conception. Within the first couple of weeks of gestation, *neurulation* shapes the embryonic neuroepithelium into a formed neural tube at about four weeks post-conception [10, 93, 78, 108]. In the following week, the neural tube differentiates at one end into the forebrain that will later on fosters the cerebral hemispheres. Subsequent portions will develop into the infratentorial structures (i.e. the hindbrain and midbrain) and the spinal cord. By the end of the first trimester, cerebral hemispheres, cerebellum, pons and medulla are present.

On the forebrain side, the hollow center of the neural tube will foster the ventricles. The surrounding ventricular and subventricular zones form the germinal zones, source location of the *cells proliferation* [93, 86]. The produced neurons and glial cells *migrate* to their finale destination i.e. towards the cortical plate [10, 86]. First cortical neurons are in place at the beginning of the second trimester [78]. From then on, in place neurons assembly into synapses during the *synatogenesis*, hence initiated during the second trimester. *Myelination* of axons, i.e. the nerve fibers, initiated in the late second trimester – between week 25 and 28 – carries on until young adulthood [67, 93, 86].

Concommitantly, these microstructural changes operated through the cell proliferation and migration, the synaptogenesis, and the myelination are closely tied to the brain growth and morphological evolution, one of the major manifestation of which

is the appearance of the cortical foldings [148]. The *morphogenesis* of the cortical folding is yet not fully understood, although multiple hypothesis view their origin in the forces applied by the axons to the cortical plate [121, 44]. Cortical foldings consist in the chronological appearance of sulci: *(i)* primary sulci – the major ones – are in place by week 27, *(ii)* secondary sulci appear at about week 32 and, finally, *(iii)* tertiary sulci are present by the end of gestation. Although all sulci are present at term, important morphological evolution continues within the first two years after birth.

The many underlying complex mechanisms involved in the fetal brain maturation are chronologically well defined and follow clear developmental pathways. Nonetheless, disruption or damages in their completion can occur during these periods of vulnerability.

1.1.2 Abnormal development

Abnormalities of the fetal brain in the prenatal period can either be developmental or acquired [47]. The many complex asynchronous and interconnected mechanisms of the fetal brain development are vulnerable to disruption of the development [87].

Developmental abnormalities are the consequences of defective programming of one or multiple microstructural embryonic or fetal stages presented in Section 1.1.1 [8, 19]. The fetal brain may also incur damages of external origin resulting from punctual or external events (e.g. stroke, injury, infection, toxic or metabolic abnormalities) [25]. Consequences and manifestation of such disruption depend on the developmental stage and location of injury [84, 19]. In general, the earlier the "hit" (either developmental or acquired), the worse the abnormalities (e.g. lissencephaly, cytomegalovirus, alcohol). Derangements occuring later during gestation will induce mild or moderate consequences. Furthermore, early disruptions will be localized in the gray matter (GM) (e.g. basal ganglia and cortex), while later ones will appear in the white matter (WM).

Both acquired and developmental disruption may result into severe conditions [47]. Pathological development may express as congenital disease, i.e. at birth, or later in life. *Congenital malformation* are present at birth and may result into severe condition in early childhood. For instance, polymicrogyria, caused by migrational abnormalities, manifests in developmental delay and/or epilepsy [19]. Moreover, although isolated congenital malformation may occur, associated malformations are frequent, such as in corpus callosum agenesis [59].

Growing evidences of *fetal programming* indicate the importance of *in utero* development as it may influence health later in life [36]. Indeed, many neurological disorders appear to take roots in the prenatal period, as for instance autism, schizophrenia and neurodegenerative diseases (e.g. Alzheimer's disease, Parkinson's disease) [35, 36, 78, 65]. Nevertheless, the manifestation and severity of disorders often remain unexplained (unknown factors).

1.1.3   Anatomical monitoring

Given the importance of the early brain development and the potential severe consequences in case of disruption, it is therefore of the utmost importance to gain knowledge on the many different pathological pathways. Underlying these open questions, clinical motivation ultimately aims at improving the patient care for disorders of fetal development origins.

Although it is clear that many childhood disorders relate to prenatal development, the identification, the characterization and the potential treatment of pathology remain to be further explored [47]. In a clinical routine, the first main objective is the earliest possible identification of defective development, in order to provide adapted pre- and postnatal patient care, plan appropriate treatment, and, provide support for the family [87]. The earlier a pathological development is identified, the better guided will be the therapy [149].

From a research perspective, it is clear that gaining knowledge on the deep developmental mechanisms and their possible deviations is crucial to better address patient care. Understanding the pathological pathways, i.e. the relationship between the brain development and the neurological disorders, can be divided into two intermediate sub-links: *(i)* between incurred disorders and macrostructural abnormalities, and *(ii)* between the macrostructural abnormalities and the underlying microstructural mechanisms. Macrostructural changes hence constitute a turningpoint in the pathological timescale.

In that respect, there is an increasing interest in the **study of morphological and structural evolution throughout gestation** and consequently a compelling need to **gain insights into the fetal brain structures**. The aims are twofold:

1. to **characterize normal development** by understanding its features and variations,

2. to **identify key structural and morphological biomarkers** to differentiate an abnormal development from a normal one.

1.2   Imaging the fetal brain

In light of the above, non-invasive, safe and efficient imaging techniques are necessary for the monitoring and follow-up of the brain development. In this section, we briefly discuss the imaging modalities typically used for fetal brain screening.

1.2.1 Ultrasound: a clinical routine

Ultrasound (US) appeared for the first time for medical imaging use in the early 1950s. Over the following couple of decades, US has made its way in obstetrics and gynecology as it quickly proved useful in the follow-up and diagnosis of fetal and placental abnormalities [139]. In the last decade of the XXth century, US biometric measurements such as the body size and length were related to fetal aging and pathological development [136, 149]. From then on, and thanks to its cost-effectiveness, wide availability and portability, US has established as a routine imaging modality for fetal and placental screening in clinic [128].

US examination is typically performed twice during pregnancy: during the first trimester, fetuses are counted, aged and assessed for viability; from the second trimester, fetal growth is assessed as well as screened for detection of potential anomalies [136]. A third trimester US exam can be performed to follow-up on the fetal growth or any upstream anomaly detected [149, 19].

Intracranial structures are only visible from the late first trimester [149]. Throughout the course of gestation, most fetal brain structures have been characterized by normal value ranges [116]. Common abnormal development typically characterized by out-of-range biometric measurements are diagnosed by comparison to reference charts.

Despite its broad advantages, US presents some limitations. The main US drawback is its sensibility to the surrounding tissues, making the image quality easily altered by the fetal pose, the interference from the fetal skull or the maternal morphology [136, 128]. Further, while US is superior for detecting calcifications, pseudocysts ans other thin membranes, it is less accurate for differentiating some structures (e.g. lipidic structures) [128].

1.2.2 Magnetic resonance imaging: a complementary modality

The first use of magnetic resonance (MR) to image a fetus was performed in 1983 [133]. It is only from the late 1990s [72], with the development of adapted sequences for fetuses (see details in Section 1.3.2), that fetal magnetic resonance imaging (MRI) has become further explored [119].

The realisation of a fetal MR examination is only relevant from the mid-second trimester to the end of gestation, following a US exam and once structures are large enough [118, 19]. We differentiate two main medical indication for fetal brain MRI:

- US cannot be successfully performed: fetal pose or maternal morphology hampers the image quality.

- Additional information are needed: *(i)* inconclusive US leads to equivocal situation where a complementary exam is required to confirm or rule out US findings,

(ii) fetal development may be at risk in case of infection, twin pregnancy or family history of previous brain pathologies [18].

Overall, compared to US, MRI offers a better soft tissue contrast of the developing fetal brain without being affected by the surrounding tissues [118, 83]. Additionally to tissue delineation, intra-tissue maturation can be visualized through the multilayer organization in the WM, indication of the cells migration from the ventricles to the cortical plate. Abnormal multilayer pattern have been related to later cortical pathologies such as polymicrogyria after 32 weeks of gestation [19].

Despite its advantages, MRI remains a complementary-only modality, because of its expensiveness and safety measures, even though no study could prove that contrast-free MR at 1.5T is harmful to the fetus' development [159]. To our knowledge, no equivalent study was performed at higher magnetic field strength [118]. Nonetheless, restrictions on above 4T fetal MR scanning applies [2]. Note that in clinics, the magnetic field strength used is governed by MR scanner availability in each institution, although 3T fetal brain MRI allows better delineation of small structures such as posterior fossa and parenchymal lesions is increasingly used [18].

In most countries such as some Swiss cantons (e.g. Vaud), fetal MR exams are only performed on medical indications as scanning of pregnant volunteers is prohibited for research purposes. Nonetheless, other countries do not share this constraint: for instance in the United Kingdom, the Developing Human Connectome Project (dHCP) aims at getting a broader understanding of the early brain development with the acquisition of longitudinal pre- and postnatal MRI.

1.3   Fetal brain MRI

1.3.1   Principle of structural MRI

Structural, or anatomical, MRI provides information for accurate delineation of the tissues. It is a non-invasive non-ionizing imaging technique that relies on the density and properties of the hydrogen atoms in the tissues to construct the intensities in the image. In a fetal examination clinical set up, the mum is placed into a scanner where a strong static magnetic field is applied. The acquisition of MR images can be decomposed as follows:

1. Creation of an equilibrium state. In the MR scanner, an external magnetic field B0 (usually of 1.5 or 3T) is applied, resulting in the alignment of all nuclei of the body along the direction of this field. A magnetization vector appears along this direction.

2. Perturbation of the equilibrium state. Energy is added to the system through a radiofrequency (RF) pulse. Excited spins are flipped over to a plane that is orthogo-

nal to B0, causing *(i)* a decrease of the longitudinal magnetization component and, *(ii)* the appearance of a transverse magnetization component.

3. Return to equilibrium state. In this relaxation phase, two processes are observed: *(i)* a longitudinal relaxation, characterized by a regrowth of the magnetization in the direction of B0, is described by time constant T1, and *(ii)* the decrease until complete cancellation of the transverse magnetization is characterized by a a time constant T2.

In this latter phase, the induced magnetic field in the plane orthogonal to B0 is measured with receiver coils and leveraged for the reconstruction of the gray scale images. Switchable static field gradients are used to encode the spatial position.

A significant advantage of MRI is the great flexibility offered in the generation of image contrast. With a given acquisition set up, i.e. the sequence of MR excitation, the strength of the signal measured and consequently the image contrast are defined for a specific location. Nonetheless, the same tissue location relates to a different image intensity for different acquisition parameters. Modifying the parameters of an MR sequence changes the tissue contrast in the images by emphasizing different aspects of the tissue composition. For instance, T1-weighted (T1w) imaging shows hyperintensity for fat tissues and, conversely, T2-weighted (T2w) imaging highlights both fat and water.

In fetal MRI, T2w is the contrast of choice for cerebral structure delineation thanks to their high water content, while T1w contrast benefits the visualization of focal components such as haemorrhages and tissue maturation indicated by myelination [47, 128]. Nonetheless, as the cerebral tissues mature throughout gestation, acquisition parameters need to be adapted according to the fetal age for both contrasts [119, 55]. Yet, similarly to US examination, unpredictable fetal motion is a key challenge for fetal MRI.

1.3.2 Fast imaging protocol

Nor isotropic 3D acquisition, nor 2D acquisition with a good through-plane resolution is realistic in fetal brain MRI due to the important unpredictable motion of the patient. Consequently, fast acquisition protocol have been proposed to prevent tampering by stochastic fetal and maternal motion on the acquired images [119]. Single-shot T2-weighted (SST2W) sequences are the most common method for structural imaging of the fetal brain [55]. Following a common principle acquisition scheme, each vendor proposes its own MR sequence design[1]. In addition to being MR vendor-specific, sequences may differ in hardware and acquisition parameters between acquisitions and institutions.

---

[1] As presented in Gholipour et *al.* [55]: **General Electric Medical systems** uses Single shot Fast Spin Echo (SSFSE), **Siemens Healthineers** uses HAlf-fourier Single-shot Turbo spin Echo (HASTE), **Philips** uses Single-Shot half-Fourier Turbo Spin Echo (SShTSE) and **Bruker Instruments** uses Rapid Acquisition with Relaxation Enhancement (RARE) .

In a nutshell, SST2W schemes allow the acquisition of a 2D slice in one "shot", i.e. following a single excitation pulse. A single slice is acquired in about one second, and a *stack* of 2D slices covering the brain can be acquired in between 20 and 30 seconds. Note that the number of slices, and hence the acquisition time, may vary with the size of the brain.

(A) Slice acquisition timeline
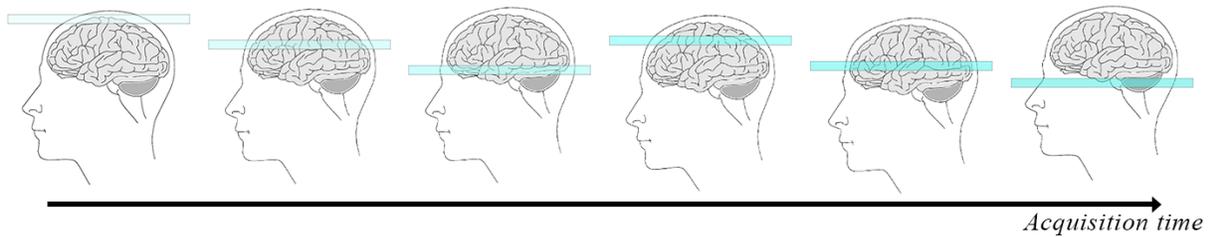
Acquisition time

(B) Theoretical acquisition

(C) Motion-corrupted acquisition
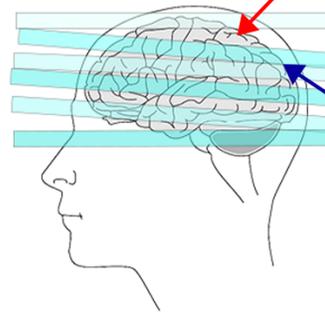
Figure 1.2: Spatial coverage of fetal brain MRI SST2W sequences. (A) Interleaved slice acquisition sequence. (B) Theoretical spatial sampling. (C) Practical motion-corrupted sampling. Red and dark blue arrows respectively indicate under- and over- sampled area.

Nonetheless, the cost of such speeded up acquisition lies in the final image resolution. At 1.5T, although SST2W present a good in-plane resolution of about $1 \times 1mm^2$, slice thickness has to remain higher, from 3 to $5mm$, with a gap of +10 to 15% [55, 118]. A finer out-of-plane resolution would result in a decrease of signal-to-noise ratio (SNR) [119]. Stacks of thick 2D slices are referred to as low-resolution (LR) images.

Another challenge going along with the time constraint is the avoidance of spin history artifacts. The latter occurs in case of successively fast acquisition of adjacent slices. So as to keep a continuous acquisition process of the slices, an interleave approach is often adopted. This acquisition scheme results in the acquisition of two trains of series within one. Figure 1.2 shows the acquisition timeline of slices (A) that results into a full stack (B).

Despite the acquisition speed-up effort, while intra-slice motion is highly minimized, inter-slice and inter-stack motion artifacts remain present in the LR images

(Figure 1.2 (C)). This motion may introduce over- and under- sampled area of the fetal brain, as respectively indicated by the dark blue and the red arrows.

In a clinical examination, acquisitions of multiple shifted stacks in the three orthogonal planes are recommended in order to cover the whole brain volume [118]. In that manner, the brain spatial coverage is hopefully increased or even up to complete. Figure 1.3 shows 2D and 3D representation of the orthogonal LR stacks.

Table 1.1 summarizes the spatial and temporal characteristics of structural T2w fetal brain MR exams performed at the Lausanne University Hospital (CHUV). Note that in clinical routine, typical acquisition protocols feature axial gradient echo T1w and diffusion-weighted imaging or diffusion tensor imaging in some cases. Overall, the whole fetal brain MR exam lasts about 30 minutes.
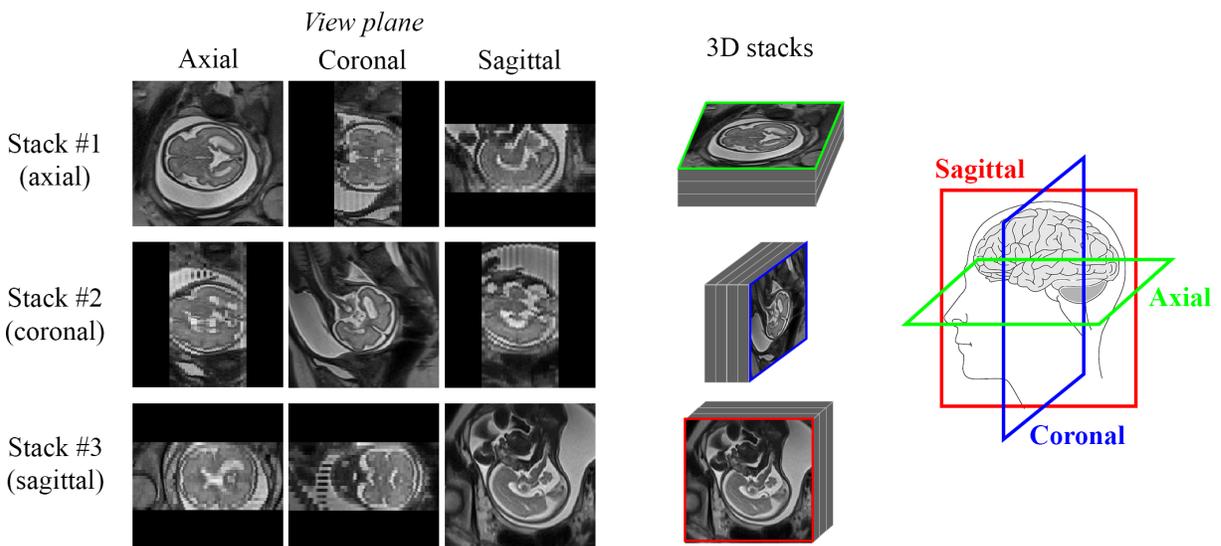
Figure 1.3: Summary of fetal brain MR exam acquisitions.

1.3.3 Limitations of fetal brain MRI

Notwithstanding the unpredictable fetal motion, the good in-plane tissue delineation on fast SST2W acquisition allows accurate measurement of cerebral structures. Biometry, myelination and cortex gyration as a function of gestational ages (GAs) were compiled into textbook atlases [47]. Reference normal biometrics were further characterized for additional infra and supratentorial structures in [142]. To alleviate any confounding, textbook atlases are constructed using MR exam presenting no motion and with orthogonal acquisitions in correct anatomical planes [47, 142]. Reference slices characterized by view planes and anatomical landmarks are well defined for the measurement of each structure [47]. In the atlas construction, structures that could not be reliably measured due to the visualization and alignment criteria were skipped.

Table 1.1: Characteristics of structural fetal brain MR exams at the CHUV on Siemens Healthcare MR scanner with HAlf-fourier Single-shot Turbo spin Echo (HASTE) sequences. Note that the exam time only considers T2w acquisitions.

		Magnetic field strength (T)	
		1.5 T	3 T
SLICE	Acquisition time (s)	1.2	1.1
	Acquisition resolution (mm^2)	1.12×1.12	1.10×1.10
	Image resolution (mm^2)	1.12×1.12	0.55×0.55
	Thickness (mm)	3.0	
STACK	N_{slices}	20 to 30	
	Acquisition time (s)	~ 26	
	Slice gap (mm)	3.3	3.0
	Image resolution (mm^3)	$1.12 \times 1.12 \times 3.3$	$0.55 \times 0.55 \times 3.0$
EXAM	N_{stacks} (target)	6	
	N_{stacks} (practice)	3 - 20	
	Estimated time (min)	< 8	

Similarly, Grossman et *al.* [61] ventured to perform volumetric analysis on axial-only SST2W acquired images. [34] further leveraged cortical plate volumes from multiplanar LR images combined to 2D biometric measurement for the prediction of intrauterine growth restriction. Although these are a good baseline for further studies, yet an important partial volume effect due to the strong anisotropy of the stacks remains. Therefore, volumetric analysis are not reliable.

While an extensive list of structures biometry have been characterized, not all of them are guaranteed to be measurable within a patient MR exam (motion, signal drop), hence leading to incomplete characterization of a fetal brain growth in clinical environment.

Due to the strong anisotropy and to the residual inter-slice motion, three-dimensional (3D) quantitative analysis still cannot be reliably performed. Moreover, despite the definition of reference acquisition planes for each of the anatomical orientation [118], correct alignment remains challenged by fetal motion and thus, even 2D biometric measurements may be unfeasible or corrupted.

1.4 Towards motion-robust 3D quantitative analysis

1.4.1 Super-resolution reconstruction

In order to better model the organ and provide good support for early diagnosis, the stakes of fetal brain MRI lies in the generation of an image with good spatial resolu-

tion while keeping a good quality, such as signal-to-noise-ratio [58]. Nonetheless, we introduced in Section 1.3.2 the constraints that regulate the acquisition process, hence resulting into anisotropic LR images. As a result, the objective is to enhance the image resolution while compensating the acquisition distortions.

In practice, each of the acquired samples, i.e. image slices, represent a different aspect of the fetal brain, thanks to the multiple shifted orthogonal acquisitions (see Section 1.3.2) [118]. Just as a puzzle, a fetal brain MR exam can only be fully appreciated once all pieces are put together. Therefore, post-processing imaging methods called super-resolution (SR) have been proposed to estimate an isotropic volume of enhanced resolution by combining several images of lower resolution [58, 146]. Figure 1.4 illustrates the forward acquisition of LR images and the reverse SR reconstruction.

In order to provide the most accurate volume possible, SR algorithms must comply with the acquisition system constraints [58]. Commonly, for a fetal brain MRI single slice acquisition the following forward linear model is considered:

$$\mathbf{X}_{kl}^{LR} = \mathbf{H}_{kl}^{\theta}.\mathbf{X} + \mathbf{n}_k \qquad (1.1)$$

where $\mathbf{H}_{kl}^{\theta}$ models the distortions occurring during the acquisition of $\mathbf{X}_{kl}^{LR}$, the lth observed slice of the kth LR image of the true target object $\mathbf{X}$, and $\mathbf{n}_k$ is the noise observed in the k_{th} LR image. θ indicates the underlying acquisition system parameters that will need to be determined.
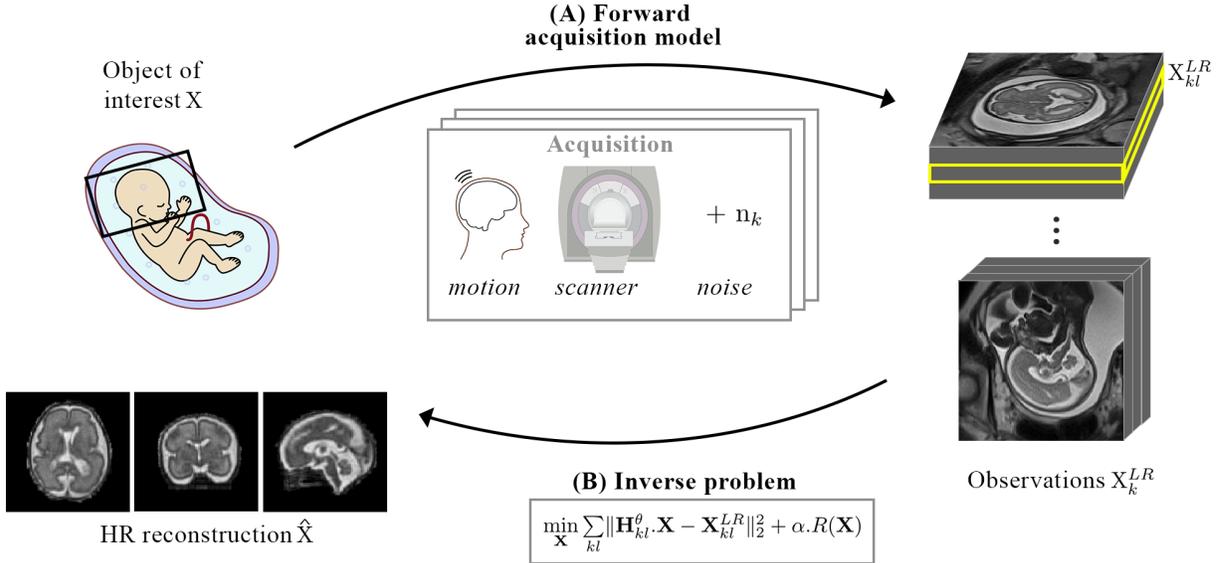
Figure 1.4: Forward model and inverse problem.

The SR reconstruction of a high-resolution (HR) volume $\hat{\mathbf{X}}$ from a set of LR observations is formulated as an inverse problem in which the underlying slice acquisition model $\mathbf{H}_{kl}^{\hat{\theta}}$ is to be determined (see Figure 1.4). Nonetheless, due to the limited availability of observations, this problem is ill-posed and, therefore, no unique solution can

be found. Consequently, regularization must be applied to constrain the problem to having a unique optimal solution:

$$\min_{\mathbf{X}} \underbrace{\sum_{kl} DF(\mathbf{H}_{kl}^{\theta}.\mathbf{X}, \mathbf{X}_{kl}^{LR})}_{\text{Data Fidelity}} \quad + \quad \alpha. \quad \underbrace{R(\mathbf{X})}_{\text{Regularization}} \tag{1.2}$$

where $DF(.)$ is a data fidelity function that aims at measuring the consistency of the observations $\mathbf{X}_{kl}^{LR}$ and the distorted target image $\mathbf{H}_{kl}^{\theta}.\mathbf{X}$, $R(.)$ is a regularization term that enforces constraints on the solution such as the image smoothness, and α weights the contribution of both terms. While DF brings prior information on the acquisition process with $\mathbf{H}_{kl}^{\theta}$, R features prior on the solution.

To solve the inverse problem in Equation 1.2, most – but not all – proposed SR reconstruction methods for fetal brain MRI [125, 76, 52, 82, 124, 88, 144, 33], are decoupled such that (1) the parameters θ of the acquisition model $\mathbf{H}_{kl}^{\theta}$ are estimated in order to (2) be reinjected into the inverse problem. Fogtmann et *al.* [42] proposed a unified motion-reconstruction approach, in which a regularization term for motion estimation is considered. We give here a brief overview of the steps of the two-stages approach.

Step 1: Acquisition model estimation

At scan time, two sources of distortion are identified:

- the acquisition system, i.e. the MR scanner, that can be decomposed into *(i)* the downsampling $\mathbf{D}_{kl}$ and *(ii)* the blurring effect $\mathbf{B}_{kl}$ induced by the slice acquisition process (point spread function and slice profile). The acquisition system-specific distortions $\mathbf{D}_{kl}$ and $\mathbf{B}_{kl}$ can be mathematically modeled as they represent prior knowledge on the common source imaging system [146, 144].

- the target object, i.e. the fetal brain. We note $\mathbf{M}_{kl}^{\theta}$ the unknown 3D rigid displacement, relative to the fixed space of $\mathbf{X}$, of the fetal brain in the lth slice of the kth LR image. The stochastic motion being fetus and acquisition dependent, its parameters must be estimated slice-wise.

Consequently, the forward model of the slice acquisition in Equation 1.1 can be reformulated as:

$$\mathbf{X}_{kl}^{LR} = \underbrace{\mathbf{D}_{kl}.\mathbf{B}_{kl}.\mathbf{M}_{kl}^{\theta}}_{\mathbf{H}_{kl}^{\theta}}.\mathbf{X} + \mathbf{n}_{k} \tag{1.3}$$

Inter-slice inter-stack fetal brain displacement $\mathbf{M}_{kl}^{\theta}$ is most commonly estimated via slice-to-volume registration (SVR) [37], although methods have been proposed based on slice intersection optimization [82] and more recently learning-based methods for pose estimation [70]. In a nutshell, SVR aims at finding the best mapping function θ for each slice to be optimally registered to a target 3D volume $\mathbf{X}$ [37]. An initial volume-to-volume registration is performed to roughly align the

LR stacks in a common space. SVR iteratively registers each 2D slice to a reference 3D HR volume. In the absence of an initial reference volume, a LR stack interpolated onto a HR grid is considered. The reference volume is updated with the interpolation of the current estimation of LR slices position at each iteration. SVR is optimized until convergence based on a distance function or similarity metric.

With the estimation of the optimal motion parameters $\hat{\theta}$ and the approximation of $\mathbf{D}_{kl}$ and $\mathbf{B}_{kl}$, the acquisition system $\mathbf{H}_{kl}^{\hat{\theta}}$ is deemed to be known. We refer to the latter as $\hat{\mathbf{H}}_{kl}$.

Step 2: Image recovery
The recovery of the target image $\hat{\mathbf{X}}$ is obtained by solving the inverse problem Equation 1.2:

$$\hat{\mathbf{X}} = \underset{X}{\operatorname{argmin}} \underbrace{\sum_{kl} DF(\hat{\mathbf{H}}_{kl}.\mathbf{X}, \mathbf{X}_{kl}^{LR})}_{\text{Data Fidelity}} + \alpha. \underbrace{R(\mathbf{X})}_{\text{Regularization}} \tag{1.4}$$

The data fidelity term DF is often formulated as an error, such as the squared norm of the difference $\|\hat{\mathbf{H}}_{kl}.\mathbf{X} - \mathbf{X}_{kl}^{LR}\|_2^2$.

Different regularization functions have been proposed for fetal brain MRI SR reconstruction, such as Charbonnier [88], total variation (TV) [144] or first order Tikhonov [33] regularization. Nonetheless, the absence of true reference HR volume makes the optimization of the regularization parameter difficult.

Although both stages have been widely investigated, the good quality of SR reconstructed volumes is not guaranteed. As a matter of fact, it may occur that SR reconstructed images are altered by remaining artifacts. Commonly, the poor SR reconstruction quality can be explained with the computer science "garbage in, garbage out" concept. In other words, inputs of bad quality result in output of bad quality. In SR reconstruction, the bad input may be *(i)* the LR images for instance due to signal drops or *(ii)* a bad motion estimation that may itself be due to bad LR images. Fortunately, outlier detection and rejection have been proposed in common pipeline, either as a preprocessing step [33] or as part of the reconstruction process [52, 88, 33]

The recent review paper of Uus et *al.* [145] on motion correction in fetal MRI identifies three main open-source toolboxes for the SR reconstruction of the brain: SVRTK [88], MIALSRTK [143] and NiftyMIC [33]. All three compile *(i)* the extraction of the region of interest, *(ii)* the preprocessing of the LR images for intra- and inter- stacks inhomogeneity correction, *(iii)* the estimation of the fetal brain motion and, finally, *(iv)* the recovery of the HR volume. Additionally, an alignment to template space is proposed. Note that amongst these pipelines, substantial inter-SR method variations are introduced due to the broad intensity-based operations induced by the preprocessing and regularization [111]. Thanks to these novel image processing techniques, good HR T2w volumes can be reconstructed with isotropic spatial-resolution that often cor-

responds to the input in-plane resolution, although SVRTK and NiftyMIC offer the possibility to specify the target resolution of the reconstructed volume.

Furthermore, an alternative approach has been proposed at the time of the redaction of this thesis. In contrast to existing methods, *NesVoR* [155] relies on deep-learning (DL) based methods for both steps. The acquisition model – especially the motion – is iteratively estimated using SVoRT [154], a transformers-based approach to perform SVR into a 3D canonical space. The final 3D volume is estimated from a continuous representation relying on an implicit neural model [102], a self-supervised approach.

Together with improved visualization, the availability of a HR isotropic volume allows the clinicians to realign and freely navigate into one single volume for the assessment of brain growth and early pathology diagnosis [115, 89]. Furthermore, reduced blurring due to partial volume effect and motion compensation opens up to the possibility of morphometry- and volumetry-based quantitative analysis of improved accuracy. Nonetheless, the baseline procedure for such quantitative analysis is the MR image segmentation of brain tissues and structures.

1.4.2 Fetal brain tissue segmentation

Three-dimensional quantification of regional volumetry and morphometry requires an accurate delineation of the object of interest in the image. In medical imaging, such delineation is done by proceeding to the image semantic segmentation. The latter consists in the assignment of a label to each voxel, i.e. volumetric pixel, of the image according to what it represents. The outcome of voxel-wise image segmentation is compiled into a labelmap. Although segmentation is performed at a voxel-level and heavily relies on the image intensity, contextual spatial information must be considered.

In fetal brain MRI, segmentation aims at classifying the voxels according to the tissue, structure or region they belong to. The resulting labelmaps offer support for voxel-based morphometry and growth quantification. While morphometry is especially explored for the developing cortical plate (e.g. surface, thickness) [147], volumetric normative ranges were defined for the whole fetal brain [53] as well as tissue-specific [89, 148].

Expert manual annotation is considered as the gold standard. Nevertheless, annotating is a tedious and time-consuming task that cannot be undertaken in a clinical routine nor for large cohort studies. Furthermore, the expert subjective appreciation makes the annotation process prone to intra- and inter-rater variability and thus, poorly reproducible [98, 112]. As a result, manual segmentation is not an enduringly reliable method and hence motivates the development of automatic segmentation methods.

Challenges for automatic methods

In order to be reliable, automatic segmentation methods must be robust to the many variations that may appear in fetal brain MRI. We identify three main sources of variability that result from the different stages of the overall fetal brain MRI workflow:

1. **the target population.** The fetal brain undergoes drastic morphological and microstructural changes throughout gestation (see Section 1.1.1). The latter induce significant variation of image intensity and thus tissue contrast: the older the fetal brain, the darker the WM appears on T2w MR images [98]. Therefore, GA-specific variations appear in the MR images (see Figure 1.5 (A)). Additionally, pathological condition may be the cause of significant inter-patient morphological differences (e.g. ventriculomegaly, polymicrogyria).

2. **the image acquisition system.** The MR scanner introduces variation that can either be hardware- or software-related, although both are closely tied as they may constraint one another (see Section 1.3.2). Additionally to the inhomogeneities introduced during any MR exam, we recall that MR vendor- and institution-specific variations are present in the MR sequence acquisition protocol (Section 1.3.2). One can appreciate in Figure 1.5 (B) the substantial texture variation as well as the image resolution. Acquired image resolution translates into different amount of blurring and hence of partial volume effect in the image.

3. **the postprocessing methods.** The many intensity-based operation performed during SR reconstruction pipeline introduces inter method variation (see Section 1.4.1). Nonetheless, we emphasize that intra-SR method variation also appears according to the availability of LR images and choice of regularization parameters. Figure 1.5 (C) presents inter- and intra- SR method variation.

In comparison, in adult brain MRI, sources of variations mainly come from the acquisition system, as no maturation effect as strong as in fetal development is observed, nor systematic post-processing steps. Additionally to the wide availability of annotated dataset, the automatic segmentation of adult brain MRI is easier to tackle.

Here, while acquisition and post-processing variations are of low influence for expert annotations, the upstream substantial difference they introduce may be critical for computer-assisted segmentation methods. Consequently, **robust** automatic segmentation methods must cope with the downstream management of fetal brain MRI variation.

(A) Target population
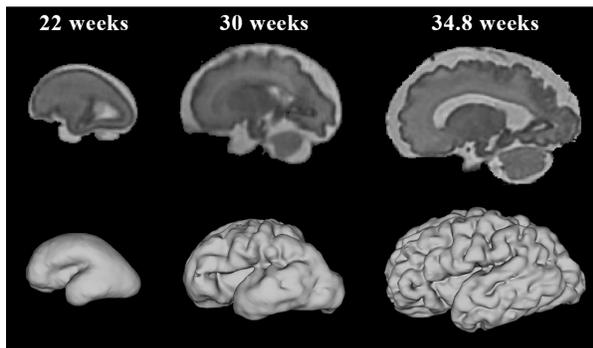
(B) Image acquisition
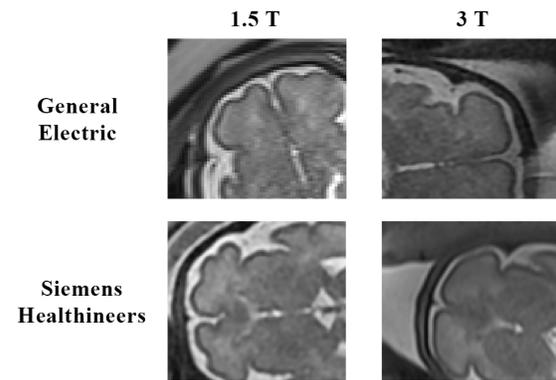
(C) Image post-processing
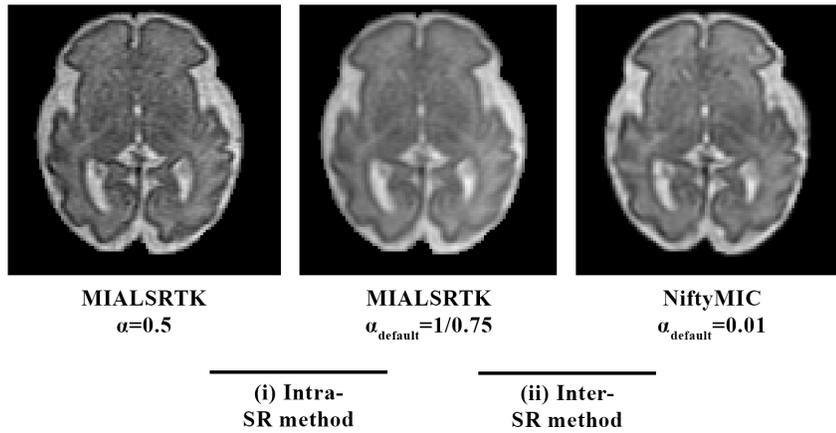
Figure 1.5: Illustration of the fetal brain MRI sources of variation that are either (A) patient, (B) acquisition system, or (C) post-processing related.

Automatic fetal brain MRI tissue segmentation methods

In their 2018 review paper, Makropoulos et *al.* [98] identified two methodology categories for automatic fetal brain MRI tissue segmentation: parametric techniques and deformable models. Since the time of this review, the emergence of digital atlases [56] allowed for registration-based segmentation techniques. In the same vein, the recent release of a publicly available clinical dataset of annotated fetal brain MRI [112] allowed the development of novel DL based methods. We briefly introduce here below the principle of each category. Note that, being mainly used for single-tissue segmentation, deformable models have only be used for ventricles in fetal brain MRI and will not be detailed here [54, 29].

Parametric techniques [98]

In parametric techniques, the probability of a voxel to belong to a tissue is statistically modeled by considering both its spatial location (derived from probabilistic atlases) and intensity value (parameterized with a Gaussian mixture model). All proposed parametric models are fit to the data, i.e. the fetal brain to segment, with an Expectation-Maximization approach, although variations such as Markov Random Field penalty have been proposed [64, 7].

Atlas-based approach

Atlas-based segmentation techniques rely on the image registration processes, which seek to find the best voxel-to-voxel alignment of two images [14]: an atlas from which a labelmap of the structures is available, and the target patients image to be segmented. Once the spatial transformation between the gray-level images is found, the target labelmap is estimated with the propagation of the atlas labels into the target space. In multi-atlas segmentation (MAS), the above-mentioned method is repeated for a set of atlases [73]. Afterwards, label fusion strategies allow for the combination of the propagated label maps, eventually attributing more weights globally or locally to one or another selected atlas. Relying on the quality of the registration step, this method is dependant on the atlas availability, the quality of the reconstructed 3D volumes and the morphological variability of the patients.

Deep-learning models

Over the last couple decades, novel supervised DL based methods have been proposed to address medical image processing, including semantic segmentation [134]. In such data driven approach, DL models are optimized to fit a distribution of annotated data, to which we refer as *training* data, in order to learn task-specific features. In image segmentation tasks, relevant image features are extracted to further discriminate the class to which each voxel belongs. During the training phase, the model is optimized by minimizing a loss function that compares the predicted output to the target output. At inference, segmentation of new unseen image can be predicted with the optimized model. As a result, a prerequisite for DL-based segmentation is the availability of large annotated datasets.

A review of DL-based methods proposed for fetal brain MRI segmentation is presented in the next sub-section.

Automatic methods for MRI tissue segmentation require reference data either for a model optimization (parametric and DL techniques) or for an image-to-template matching (atlas-based methods). In the latter, the more diverse the annotated data, the more likely the target image will be matched to similar atlas. For supervised methods, the more diverse the annotated data, the better parameterized will be the model. Although DL methods are known to be remarkably data hungry, they have been established as the state-of-the-art methods thanks to the different techniques to overcome the data availability limitation.

DL models for fetal brain MRI tissue segmentation

In Khalili et *al.* [79], the first DL-based approach for the segmentation of intracranial tissues in fetal brain MRI was proposed. In this work, seven (7) intracranial tissues were segmented using the now widely used U-Net architecture that was originally proposed for neuronal structure segmentation and cell tracking [122]. This model is composed of an encoding path and a decoding one, with skipped connections. The encoding path, also known as the contracting path, aims at capturing relevant image features. From there, the expanding path reconstructs a predicted labelmap. Although they presented promising results for DL-based fetal brain MRI tissue segmentation, In Khalili et *al.* [79] only considered the segmentation of LR images in the coronal orientation.

In 2021, the public release of the Fetal Tissue Annotation (FeTA) dataset [112] opened up to the possibility of developing automatic segmentation methods for intracranial tissues. Together with the dataset release, a benchmark of automatic segmentation methods was presented, in which all methods but one – ours, a MAS method – are DL-based [112]. In this study, all DL-based methods relied on 2D approaches, i.e. the input samples to the networks were subsets of the image of 2D shape (e.g. an image slice). The overall best performing method adopted a multi-view approach, aggregating predictions from the three orientations. Interestingly, our MAS method outperformed the others specifically on bad quality images, hence being more robust to the image quality than the others. Our MAS approach is presented in Appendix C.

In 2021, the first edition of the FeTA MICCAI challenge was held, relying on an extended FeTA dataset release. All submissions were DL-based methods implementing either the original or a derived version of the U-Net architecture [113]. Submissions differentiated in terms of architecture (e.g. dimensionality, input size, number of trainable parameters), training and optimization strategy (e.g. loss function, optimizer, cross-validation, data augmentation) pre- and post-processing (e.g. ensemble learning, mathematical morphology operation) and data used (e.g. neonatal images). However, top ranking submissions unanimously agree on some parameters, such as the consideration of 3D input samples. Despite the promising results of the automatic

segmentation methods proposed, the scope of this challenge was limited. Indeed, the true variability of fetal brain MRI was not widely represented, as all images – both from the training and testing sets – were all acquired at the same institution and post-processed with two different SR reconstruction methods.

Consequently, the second edition of the FeTA challenge held in 2022 aimed at pushing further the challenge by providing multi-center datasets. To assess the applicability of such methods under real-world constraints, some institutions provided images as pure testing sets.

Open-source annotated dataset and their limitations

Amongst the publicly available datasets of annotated fetal brain HR T2w MRI, we differentiate the digital atlases and the clinical sets, that are respectively presented in Table 1.2 and Table 1.3. By definition, an atlas is the combination of a template intensity image and its segmentation labelmap [14]. Throughout this thesis, we use the term *atlas* to describe normative reference pairs (template, segmentation) that are derived from multiple clinical MR exams. In fetal brain MRI, typical spatiotemporal (4D) atlases compile multiple 3D atlases considering the longitudinal evolution of the fetal brain throughout gestation. As opposed to atlases, clinical dataset compile MR exams and their manual annotation, i.e. one volume corresponds to one patient's brain.

Table 1.2: Summary of the open-source annotated fetal brain MRI atlas datasets. GW: gestational weeks

	Materials			Methods	Outcome	
	Cohort	Acquisition	Post-processing	Construction	Image	Annotations
[129]	N=80 (21.7-38.7 GW) Normal	1.5T Philips	[76]	Serag et al., 2012 [130]	N=15 (23-37 GW) $1.18mm^3$	Tissue probability maps
CRL [56]	N=81 (19-32 GW) Normal	3T Siemens 1.5T Philips	[52]	Gholipour et al., 2017 [56]	N=18 (21-38 GW) 0.8^3mm^3	Parcellation and tissue labelmaps
SBA [41]	N=90 (21-35 GW) Spina Bifida aperta	1.5T Siemens	NiftyMIC [33]	Fidon et al., 2022 [41]	N=15 (21-34 GW) 0.8^3mm^3	Tissue labelmaps
CHN [156]	N=90 (22.5-39 GW) Normal Chinese	3T Siemens	NiftyMIC [33]	Xu et al., 2022 [156]	N=16 (23-38 GW) 0.8^3mm^3	Parcellation and tissue labelmaps

Spatiotemporal atlases provide anatomical references of the fetal brain at each gestational age. However, the construction of a reference template image from multiple clinical exams smoothes the true fetal brain MRI variability. Therefore, annotation on patient-specific images are essential to better represent the true target population.

Table 1.3: Summary of the open-source annotated fetal brain MRI clinical datasets.

	Cohort	Acquisition	Post-processing	Annotations
FeTA [112]	N=80 Pathological and Neurotypical	1.5T or 3T GE Healthcare	MIALSRTK [143] SVRTK [88]	FeTA guidelines [112]

The *first* clinical dataset of annotated fetal brain MRI, the FeTA dataset was released in 2021 [112]. In a nutshell, SR reconstructed clinical fetal brain MRI are provided along with a labelmap of intracranial tissue classified into seven (7) categories: the extra-axial cerebrospinal fluid (CSF), the cortical gray matter (cGM), the WM, the ventricular system (lateral, third and fourth ventricles), the cerebellum, the deep gray matter (dGM) and the brainstem [112]. The original FeTA dataset is under the ethical committee of the Canton of Zurich, Switzerland (Decision numbers: 2017-00885, 2016-01019, 2017-00167).

As represented in Table 1.3, the wide range of possible variation presented in the previous Section 1.3.3 is poorly represented into the available datasets. The limited access to patient data and the expensive cost of annotation yields to an important training data scarcity. Consequently, the limited availability of annotated dataset jeopardizes the development of robust segmentation methods.

In a real case scenario, one might want to segment images from a new target dataset $\mathcal{D}_{\mathbf{T}}$, although only annotated images from a *source* dataset $\mathcal{D}_{\mathbf{S}}$ are available. The upstream acquisition and postprocessing inter-dataset differences may result into different data distribution [81]. We refer to these dataset-specific distribution as *domains*. Consequently, as the core principle of DL is to fit a data distribution, a model trained on a source dataset $\mathcal{D}_{\mathbf{S}}$ may poorly perform on an out-of-source distribution target dataset $\mathcal{D}_{\mathbf{T}}$ [51]. In short, **domain shift** → **performance shift**. Fortunately, DL techniques have been explored to improve the performance of a model in this specific problem known as *domain adaptation*.

Strategies for robust DL models

Domain adaptation falls into the machine learning research fields known as transfer learning (TL). TL aims at leveraging knowledge learned by a model for a *source task*, to achieve good performance on a different but related *target task* [81]. For instance, re-utilizing a fetal brain MRI tissue segmentation model in order to further segment cortical parcellation of fetal brain MRI is within the scope of TL.

Specifically in our example of fetal brain MRI tissue segmentation, the source and target tasks are semantically identical (fetal brain MRI tissue segmentation). However, an intrinsic *shift* is induced by the domains of the source and target images, respectively

$\mathcal{D}_{\mathbf{S}}$ and $\mathcal{D}_{\mathbf{T}}$. In the absence of target domain annotated images, it is therefore necessary to find ways to either adapt or generalize the model to improve the performances.

Data augmentation synthesizes new data from the available ones in order to overcome the lack of data and prevent the model from over-fitting. As a result, the sample size is increased as well as its intrinsic diversity. Conventional data augmentation strategies apply spatial (e.g. rotation, flipping) or intensity-based (e.g. Gaussian noise, intensity inhomogeneities) transformations to the original data. Data augmentation has been broadly adopted in the field to make DL models more robust [57]. In Khalili et al. [79], acquisition-inspired augmentations are applied with the introduction of intensity inhomogeneity in order to make the model robust to acquisition artifact. Nonetheless, while the introduction of small variation to the data distribution has proved efficient for in-domain segmentation enhancement, it may not address larger perturbation as those induced by the domain shift.

Therefore, the challenge lies in the reduction of the inter-domain performance drop without having to annotate more data. Recently, domain-adaptation strategies for MRI segmentation have been proposed [107, 105]. Nevertheless, such methods are all data-driven, hence adjustment to the target task requires optimization efforts. Moreover, while these methods specifically aim at integrating non-linear intensity changes, these are not directly addressing the real domain shifts induced by the acquisition and reconstruction processes.

Assessment of image segmentation

The quantitative evaluation of the performance of a segmentation method is crucial to determine its robustness and reliability. Segmentation evaluation measures, to which we refer to as *metrics*, typically compare the predicted segmentation of an algorithm to a reference segmentation. In medical imaging, reference segmentation are typically obtained from manual expert annotation. Despite the known inter-annotator variability [112], reference segmentation are considered as "ground truth" (GT).

A segmentation labelmap can be (in)accurate in many ways: size, location or shape. Such different aspects of a segmentation quality are commonly quantified by different metrics. According to these aspects, metrics can be classified into three classes: **overlap-, size-, and boundary-distance-based**. Although it is important to consider complementary metrics for the full assessment of a segmentation quality, *Metrics Reloaded* [97] recommends a problem-driven approach for the metrics selection.

In practice, segmentation performances of fetal brain MRI tissue segmentation are usually assessed with overlap-, size- and boundary-distance-based metrics [113]. Nonetheless, we recall that the purpose of fetal brain tissue segmentation addresses both volumetric and morphometric analysis. Voxel-based morphometry studies the shape of the structures and therefore considers not only the spatial localisation but also the connection of the voxels. Therefore, while both volumetry and morphometry

requires good voxel-wise segmentation, morphometry additionally requires topology correctness in the prediction.

1.5   Thesis contributions and outline

The goal of this thesis is the **development and validation** of **robust methods** for **improved 3D quantitative analysis** of fetal brain MRI. As evidenced throughout this chapter, a major obstacle in these tasks is the lack of GT reference data. While the absence of true HR fetal brain MRI hampers the validation and optimization of SR reconstruction algorithms, the development of automatic segmentation methods is held back by the lack of annotated data representative of the true fetal brain MRI variability. Figure 1.6 (A) summarizes the whole fetal brain MRI workflow in either a clinical or research context.
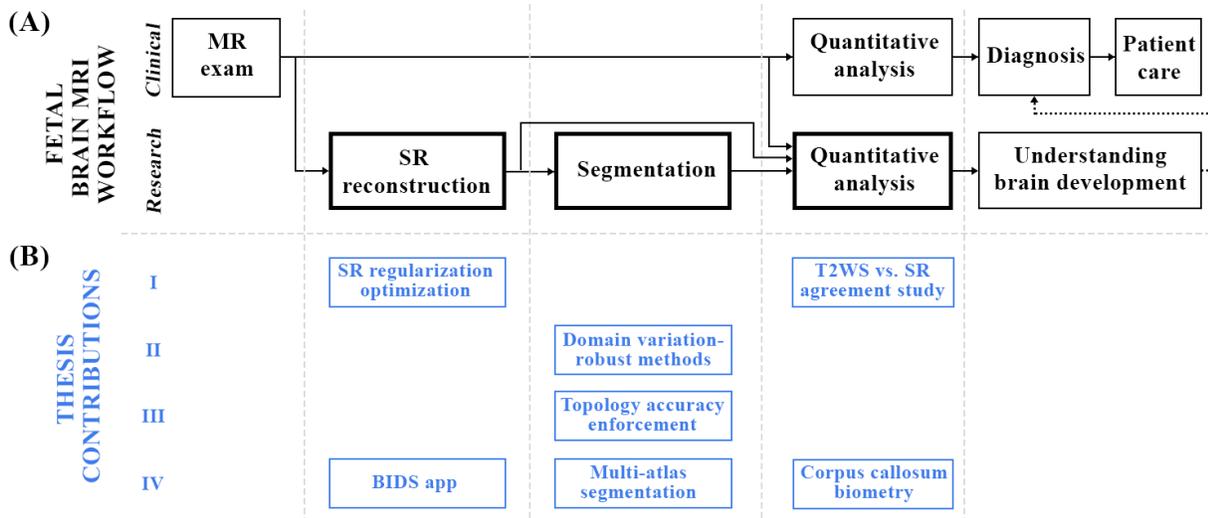
Figure 1.6: Overview of this thesis contributions and outline (B) with respect to the overall fetal brain MRI workflow (A).

Overall, the main contributions of this thesis can be categorized as:

1. Methodological developments

   - **Experimental framework** for the **optimization** of SR reconstruction parameters.
   - **Data augmentation approaches** for the development of **robust** automatic segmentation methods.
   - Optimization **loss function to ensure the topological consistency** of DL based automated segmentation.
   - **Topology-based metric** to quantify the size of the segmentation topological defects.

2. Analysis and methods validation

   - Statistical analysis of **multi-dataset multi-observers biometric measurements**.
   - **Assessment and comparison** of automatic segmentation methods.

Additional contributions are:

- the **co-organization of the FeTA MICCAI challenge**. Beyond the practical organization, I participated to the redaction and revision of proposals and publications, the evaluation and revision of the submissions, and the analysis and interpretation of the results.

- the **software developments** related to our in-house developed MIALSRTK BIDS App for the SR reconstruction of fetal brain MRI. Details on **the maintenance, the improvement of the user-experience and performance, and the development of new features** are briefly presented in Appendix B.

- the **data management** of our institution clinical dataset, following the well defined Brain Imaging Data Structure (BIDS)[2] way of organizing neuroimaging data. These efforts aimed at standardizing the organization of the source images and their derivatives to ease the navigation and communication with image processing softwares. More precisely, it consisted in the **retrieval, anonymization, conversion and organization of clinical data**.

Part I of this thesis combines the works related to SR reconstruction of fetal brain MRI. **Chapter 2** presents a **clinical validation of SR reconstruction algorithms**. Biometric measurements from two observers, one experienced and one junior radiologists, on both LR T2w and SR reconstructed images are compared to **assess the practical value** of SR reconstruction **in a clinical environment**. In this **intra- and inter-observer agreement study**, we evidence the dependence of the observer's confidence to the image observed. An extension of this work for the specific assessment the fetal corpus callosum biometry is presented in Appendix D with the inclusion of US imaging.

Chapter 3 presents a **simulation-based approach to tune the SR reconstruction parameters**, especially the regularization weight. In this study, we propose to leverage a **simulation framework to generate both LR and HR synthetic** T2w fetal brain MR images. In different acquisition settings, the HR simulated images are used as **GT to assess the quality of SR reconstructed LR simulated images**. Results show that default parameters are typically sub-optimally set. Our principled approach for the optimization of regularization parameters translates well to real case scenarios, with **clinical SR reconstructions of qualitatively better quality**.

---

[2]*https://bids.neuroimaging.io*

Part II covers the **generalization of automatic segmentation methods across different domains**, specifically through data augmentation. In both Chapter 4 and Chapter 5, we aim at **increasing the training sample size and variability** represented. **Chapter 4** presents a data augmentation strategy inspired from the sources of variation themselves to overcome the image domain shift. As in Chapter 3, we propose to use a numerical simulator to **generate synthetic fetal brain MRI**. To replicate as faithfully as possible the entire fetal brain MRI generation, LR T2w simulated images are SR-reconstructed. We first prove that **numerically-simulated images can complement clinical images** for data augmentation in a domain shift-free application. Then, we show that a model that was trained in a source domain can be adapted to a target domain through **fine-tuning on SR reconstructed synthetic MR images**.

In Chapter 5, we propose to leverage the **intra-SR reconstruction methods variability**. As we evidence in Chapter 3, variations of the regularization parameter introduce **texture changes in the SR reconstructed volumes**. In this study, **multiple reconstructions of the same fetal brain** with different regularization parameters are used as data augmentation strategy. Overall, our multi-SR reconstruction approach significantly **improves the in-SR domain segmentation performance and the generalization across SR reconstruction methods**.

Part III explores the importance of the **consideration of the topology in segmentation**. In Chapter 6, we **analyse the topological properties** of the top fetal brain MRI multi-tissue segmentation methods. In the context of the MICCAI FeTA challenge, we explore the inclusion of the topology assessment in the comparison of different methods. We further evidence the different topological consideration on a tissue-wise analysis.

Chapter 7 presents a **methodological contribution for the improvement of cortical plate segmentation topological correctness and its validation**. In this work, we propose a DL segmentation framework for automatic and morphologically consistent segmentation. First, we propose a **generalized multi-dimensional topological loss function** in order to enhance the topological accuracy. Second, we introduce hole ratio, a **new topology-based validation measure that quantifies the size of the topological defects** taking into account the size of the structure of interest. Overall, both quantitative and qualitative results support the generalizability and added value of our topology-guided framework for fetal cortical plate segmentation.

Chapter 8 discusses the results of this thesis and the possible future perspectives.

Part I

On super-resolution reconstruction

2

Fetal brain MRI biometric measurements

The work presented in this chapter is originally published as **"Fetal Brain Biometric Measurements on 3D Super-Resolution Reconstructed T2-Weighted MRI: An Intra- and Inter-observer Agreement Study"** with co-authors: P. de Dumast, M. Khawam, P. Deman, H. Kebiri, T. Yu, S. Tourbier, H. Lajous, P. Hagmann, P. Maeder, J.-P. Thiran, R. Meuli, V. Dunet, M. Bach Cuadra and M. Koob, in **Frontiers in Pediatrics, 9:639746, doi: 10.3389/fped.2021.639746**.
P. de Dumast and M. Khawam are equally contributing authors.

2.1 Introduction

Biometric measurements are good markers of fetal brain maturation and growth and are a fundamental basis for the diagnosis of developmental and acquired brain abnormalities [62]. Indeed, an abnormal measurement is often the first warning of disturbed fetal growth that requires further investigation. Quantifying brain development in comparison to reference charts is the first routine step. The accuracy of biometric data is essential for the evaluation of prognosis and prediction of outcome as it may influence both prenatal and postnatal management.

The level of confidence in the process of image analysis [60] is important parameter as it corresponds to the degree of certainty in the correctness of diagnosis and is associated with relevant decisional consequences [13, 104]. An imaging method where measurements would be easier to perform could help in providing accurate biometry with more confidence, particularly on suboptimal images and for less experienced radiologists.

In clinical practice, fetal brain biometric measurements are commonly performed on T2w LR images. Nonetheless, we saw in Section 1.3.3 that LR-based measurements may be corrupted by the strong image anisotropy, the oblique orientation of the image, and the inter-slice motion.

In this study, we aim to compare 2D measurements of fetal brain biometry using orthogonal 2D LR T2w sequences (T2WSs) vs. using one single 3D SR-reconstructed volume for both supratentorial and infratentorial measurements. We evaluate *(i)* the agreement between T2WS and SR measurements, *(ii)* the level of the observers' confidence on both datasets, and *(iii)* the concordance between a junior radiologist and an experienced pediatric radiologist.

2.2 Materials and methods

2.2.1 Dataset

Cohort

We retrospectively collected all consecutive normal fetal brain MRI exams from January 2013 to October 2018 (28 patients in total: 15 males and 13 females) from the MRI database of our institution, the Lausanne University Hospital (CHUV), Lausanne, Switzerland. All MRI scans were conducted on medical indication (ventriculomegaly, suspicion of corpus callosum or posterior fossa anomaly, microcephaly...) within 2 weeks of an expert fetal neurosonographic study and were finally considered normal. Early neonatal neurological clinical evaluation was normal. All images were anonymized prior to further analysis. This retrospective study was part of a larger research protocol at our institution approved by the local ethics committee.

MRI

Clinical MR images were acquired either at 1.5 T (MAGNETOM Aera, Siemens Healthcare, Erlangen, Germany) (88% of the patients) or at 3 T (MAGNETOM Skyrafit, Siemens Healthcare, Erlangen, Germany) (12% of the patients). The fetal brain MRI protocol included SST2W Half-Fourier Acquisition Single-shot Turbo spin Echo (HASTE) sequences in the three orthogonal orientations; usually at least two acquisitions were performed in each orientation, together with axial gradient echo T1-weighted and diffusion-weighted imaging or diffusion tensor imaging in some cases. The coronal plane was parallel to the brain stem, and the axial plane was parallel to the corpus callosum long axis [142]. We excluded twins from the study (n=2). At this point, 26 normal fetal brain MR images were kept for further analysis. Details on the MRI acquisition parameters can be found in Table 2.1.

Table 2.1: MRI acquisition parameters of the T2-weighted Half-Fourier Acquisition Single-shot Turbo spin Echo (HASTE) sequences.

Field strength (Tesla)	Number of exams	Gestational age (weeks)	Number of series	In-plane resolution (mm)		Slice thickness (mm)		Echo time (ms)		Repetition time (ms)	
				min-max	mean ± SD	min-max	mean ± SD	min-max	mean ± SD	min-max	mean ± SD
1.5 T	23	18-34	143	1.125 – 1.172	1.127 ± 0.009	2.42 – 4.0	3.34 ± 0.20	82 – 98	90 ± 1.9	832 – 1200	1185 ± 59
3 T	3	27.4 ± 4.2	14	0.547	-	3.0	-	101	-	1000 – 1100	1090 ± 027

2.2.2 Methodology

3D SR reconstruction

In our study, 3D SR volumes were reconstructed from T2WSs within the PACS station with an in-house *syngo*.via Frontier fetal MRI prototype [26] based on the publicly available MIALSRtoolkit software [144, 143].

All cases were reconstructed by an engineer with 20 years of experience in medical image processing. The selection of the series used for the reconstruction was done based on visual inspection, and T2WSs that exhibited high levels of motion distortion and/or intensity signal dropout (thus, that were not exploitable for radiological reading neither) were excluded from the 3D SR reconstruction process. On average, six series were used per patient for SR reconstruction, with a range from 3 to 11 series (see Figure 2.1). All 3D images were reconstructed with an isotropic spatial resolution matching its input in-plane resolution (in average of around $1.1 \times 1.1 \times 1.1 mm^3$).
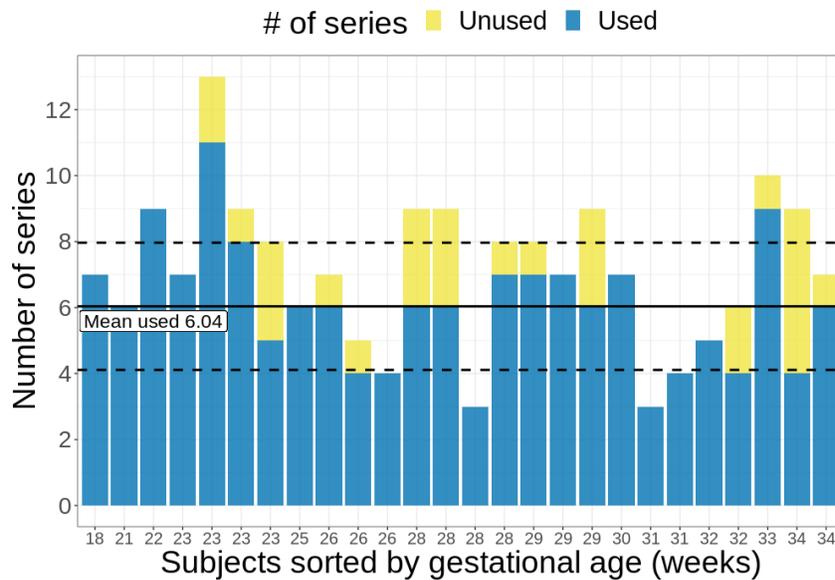
Figure 2.1: Distribution of the number of series used per patient compared to the total available number of series.

(a) Examples of the three different ratings made by the experts.

(b) Quality of SR reconstructions, with the number of cases in each category of GA range.
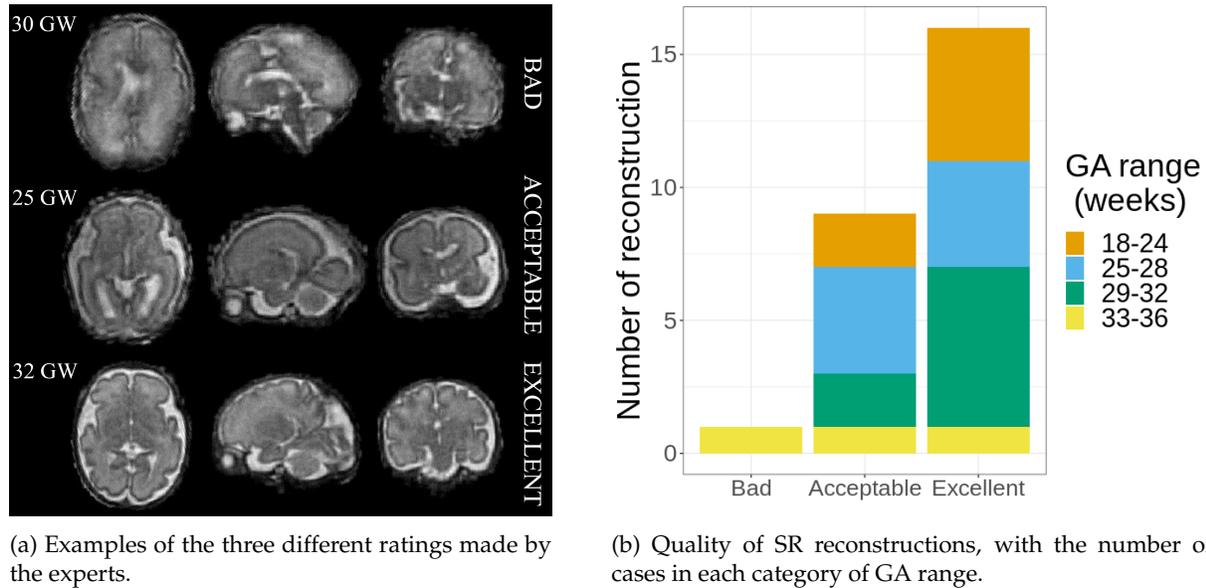
Figure 2.2: Quality of SR reconstructions.

Evaluation of SR quality

The quality of the 3D SR was assessed independently, in a blinded protocol, by three expert raters: one engineer expert in MR image analysis (rater 1, the same who performed the SR reconstructions) and two experienced pediatric radiologists (rater 2 and rater 3, with, respectively, 15 and 9 years of experience in fetal brain MRI). Reconstructed volumes were classified into one of three categories: bad (with remaining motion, very blurred, unusable for diagnosis purposes), acceptable (overall good quality with some blurring but still relevant for diagnosis purposes), or excellent (good quality without any blurring). Examples of SR-reconstructed volumes rated with the three different quality measures are shown in Figure 2.2a.

Biometric measurements

Using standard tools on the PACS station (Carestream Vue PACS©, Version 12.1.6.1005, Carestream Health Inc., NY, USA), two observers, one experienced pediatric neuroradiologist (obs1, 15 years of experience of fetal brain MRI) and one inexperienced radiologist (obs2, without any experience in fetal brain MRI) independently measured 11 fetal brain structures on both datasets (T2WS then SR) during two reading sessions separated by 3 weeks and blinded to the results of the previous measurements.

The following biometric measurements as shown in Figure 2.3 were performed following previously published guidelines for fetal brain MRI biometry [118, 142, 46, 89]:

- The skull coronal biparietal diameter (sBIP_cor) is defined as the greatest transversal diameter between the inner tables of parietal bones on a coronal slice through the temporal horns of the lateral ventricles. The brain coronal biparietal diameter (bBIP_cor) was measured on the same slice.

- The brain axial biparietal diameter (bBIP_ax) is defined as the maximal brain diameter in the transverse plane through the atria. The skull axial biparietal diameter (sBIP_ax) is defined as the inner to inner table maximal skull diameter in the transverse plane through the atria.

- The corpus callosum length (LCC), height of the vermis (HV), and anteroposterior vermis diameter (APDV) were measured in the mid-sagittal plane.

- The transverse cerebellar diameter was measured on a coronal slice (TCD_cor) and on an axial slice (TCD_ax).

- The fronto-occipital diameters [right fronto-occipital diameter (rFOD) and left fronto-occipital diameter (lFOD)] were measured on a sagittal slice between the extreme points of the frontal and occipital cortices.

On T2WS, each observer independently chose the best-quality T2W series for each measurement. SR measurements were performed in orthogonal planes in multiplanar reformations (MPRs) using the tools within the PACS system. Each observer rated the confidence of each of his measurements in both T2WS and SR datasets either as high or low.
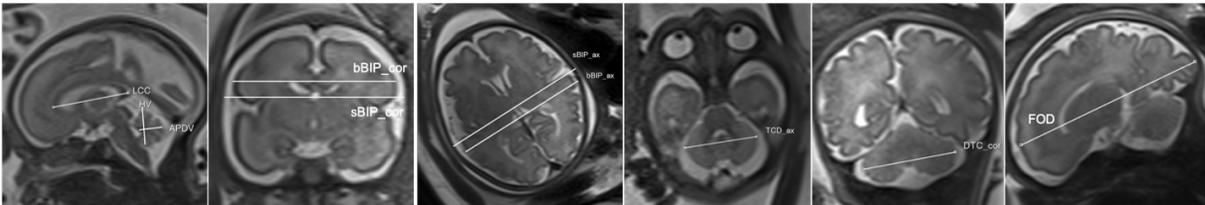
Figure 2.3: Biometric measurements. LCC, length of the corpus callosum; APDV, anteroposterior diameter of the vermis; HV, height of the vermis; bBIP_cor, sBIP_cor, bBIP_ax, sBIP_ax, brain and skull biparietal diameter (coronal and axial); TCD_ax and TCD_cor, transverse cerebellar diameter (axial and coronal); FOD, fronto-occipital diameter.

Statistical analysis

Statistical analysis was conducted with R software (version 3.6.3).

Evaluation of SR quality. The inter-rater reliability was measured using a weighted ordinal Gwet's agreement coefficient (Gwet's AC) and interpreted according to Altman's benchmarking scale [63].

Level of confidence. Chi-square test statistics were used to evaluate the dependence of the level of confidence of the raters on each dataset.

Biometric measurements. The association of paired lFOD and rFOD measurements was tested using Spearman's rho statistic [68], and the difference was tested with the paired Wilcoxon rank sum test. This analysis was performed for each pair of dataset-observer. Biometric measurements were compared statistically to determine interdataset and inter-observer significant difference, respectively, for each observer (obs1 and obs2) or dataset (T2WS and SR), with the paired Wilcoxon rank sum test (without and with Bonferroni multiple comparisons correction). Lin's concordance correlation coefficient (CCC) [95] and intraclass correlation coefficient (ICC) were computed for the agreement on measurements obtained on the two datasets [63]. Bland-Altman plots were used to assess the agreement between the two observers for each of the two datasets and the reliability between the two datasets for each of the two observers. Agreement was rated as follows: poor, <0.5; moderate, 0.5-0.75; good, 0.75-0.9; excellent, >0.90 [63]. Agreement between transverse and coronal measurements was evaluated for brain (bBIP_ax, bBIP_cor) and skull (sBIP_ax, sBIP_cor) biparietal diameters and for transverse cerebellar diameter (TCD_ax, TCD_cor) using ICC, error rate, and paired Wilcoxon rank sum test. The *p*-value level for statistical significance was set at 0.05.

2.3   Results

2.3.1   Evaluation of 3D SR reconstructions

A total of 26 cases were SR-reconstructed through MIALSRTK [26].

The estimated ordinal Gwet's AC between the three raters was 0.85 with a standard error of 0.06. According to Altman's benchmarking scale, this estimated coefficient was considered to be either Good or Very Good with a 0.99 probability.

As the inter-rater agreement was good, we considered the quality of a reconstruction as the averaged consensus between the three raters' assessment. On average, experts rated one case as bad, nine as average, and 16 as excellent Figure 2.2b. No significant differences were found regarding the quality of the SR reconstruction and the GA ranges. The case rated as bad[1] was discarded for further analysis; thus, 25 normal fetal brain MRI (from 18 to 34 weeks of GA, mean ± standard deviation (SD): 27.1 ± 4.2) were considered for the biometric analysis.

2.3.2   Confidence of measurements

On T2WS, some measurements were made with low confidence in three out of 25 fetuses (12%) by obs1, while obs2 reported a low level of confidence in 13 out of 25

---

[1]Bad SR quality was primarily due to only four series were available and brain was cut in two of them.

fetuses (52%). Specifically, obs1 had low confidence in FOD in two fetuses and in axial measurements in DTC, sBIP, and bBIP in another fetus. In contrast, obs2 had low confidence mostly on the LCC (in 11 fetuses) and also in FOD and axial DTC (in two fetuses). On SR, low confidence measures were made only on the LCC in two fetuses by obs2. All the remaining measurements made on SR and T2WS were rated with high confidence. Overall, the level of confidence of obs2 was dependent on the dataset used, either T2WS or SR (p = 0.002), with higher confidence using SR. Conversely, no significant difference was found in the confidence level for obs1 (p = 0.23).

2.3.3 Biometric measurements analysis

Each observer-dataset pair showed a high correlation between right and left FOD (Spearman's rank correlation ρ = 0.975, p < 0.001) and no significant difference (p = 0.8). Hence, rFOD and lFOD were averaged. Agreement between T2WS and SR for each biometric parameter was good for both observers, with minimum Lin's CCC estimated to be 0.86. Both Lin's CCC and ICC were on average 0.997 for both raters (Figure 2.4; Table 2.2).
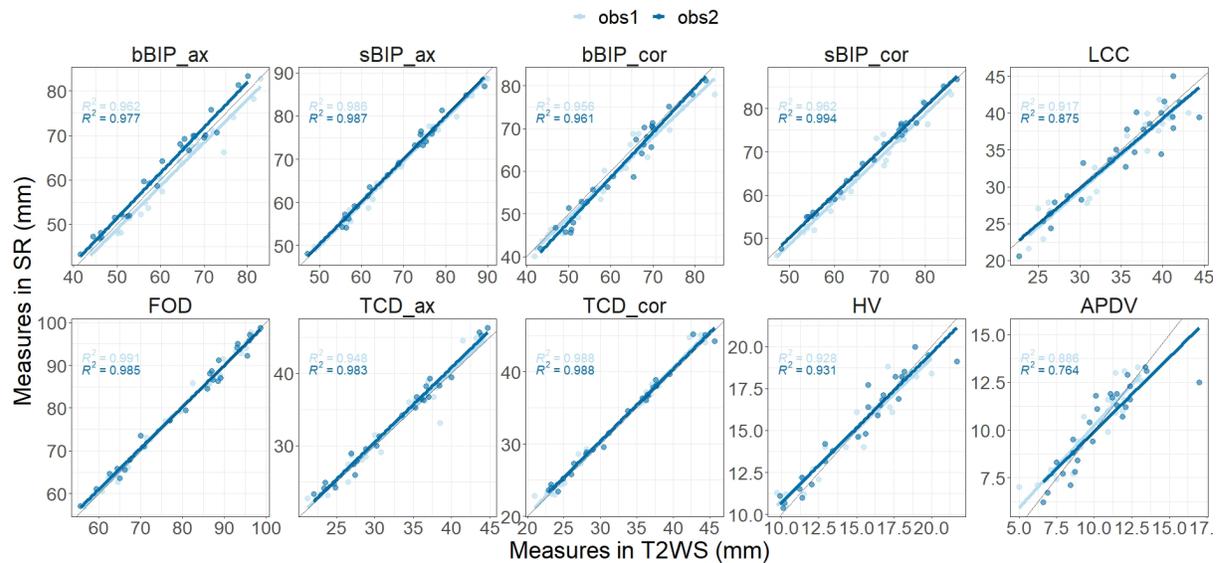
Figure 2.4: Inter-dataset [T2-weighted sequence (T2WS) vs. super-resolution (SR)] agreement for observer 1 and for observer 2 for each biometric measurement. For each biometric, measurements on SR (vertical y-axis) and on T2WS (horizontal x-axis) for each fetal brain MRI examination (point) for observer 1 (light blue) and observer 2 (dark blue). The solid gray line depicts perfect concordance.

Inter-dataset (SR vs. T2WS) intra-observer (obs1 and obs2) comparisons (Figures 2.5 and 2.6) showed statistical differences (p < 0.05) for brain axial and coronal BIP (bBIP_ax and bBIP_cor) for both observers, skull coronal BIP (sBIP_cor) for obs1, and axial and coronal TCD for obs2 (TCD_ax, TCD_cor). After correction for multiple comparisons, only axial brain BIP (bBIP_ax) remained significantly different for

Table 2.2: Lin's concordance correlation coefficient values. Obs1, obs2, observer 1 and observer 2; LCC, length of the corpus callosum; APDV, anteroposterior diameter of the vermis; HV, height of the vermis; bBIP_cor, sBIP_cor, bBIP_ax, sBIP_ax, brain and skull biparietal diameter (coronal and axial); TCD_ax and TCD_cor, transverse cerebellar diameter (axial and coronal); FOD, frontooccipital diameter

	obs1	obs2
bBIP_ax	0.972	0.977
sBIP_ax	0.992	0.994
bBIP_cor	0.969	0.971
sBIP_cor	0.973	0.996
LCC	0.951	0.932
FOD	0.996	0.992
TCD_ax	0.972	0.987
TCD_cor	0.992	0.992
HV	0.956	0.961
APDV	0.930	0.865

both observers, but differences were small ($2.95 \pm 1.73mm$) (Figure 2.6). Overall, the inter-dataset average error rate in our study was 3.3%. Additionally, Supplementary Table A.3 shows a high intra-rater reliability for one observer on a sub-cohort of five fetuses.

Inter-observer intra-dataset comparisons showed no significant differences ($p > 0.05$), except for brain axial biparietal diameter (bBIP_ax) on T2WS and for brain and skull coronal BIP (bBIP_cor, sBIP_cor) and coronal transverse cerebellar diameter (TCD_cor) on SR. Overall, differences remained small (Figure 2.7) and independent of the measured size (Figure 2.8). After correction for multiple comparisons, none of them remained significant.

We computed the agreement between transverse and coronal measurements within each dataset and for both observers (ICC, percentage of error range, mean ± SD of percentage error). Agreement was excellent without statistical differences ($p > 0.05$) for TCD on T2WS (ICC = 0.983, 0.3 – 16%, 2.9 ± 2.6%) and SR (ICC = 0.997, 0 – 1%, 1 ± 1%), for bBIP on T2WS (ICC = 0.951, 0.3 – 14%, 5.2 ± 3.5%) and SR (ICC = 0.970, 0.1 – 15%, 5 ± 3.8%), and for sBIP on T2WS (ICC = 0.972, 0 – 15%, 3.4 ± 3.2%) and SR (ICC = 0.972, 0 – 11%, 3.8 ± 3.4%). Complete results are presented in Supplementary Tables A.1 and A.2.

2.4   Discussion

2.4.1   Summary of contributions

Our study showed a sound agreement for biometric measurements performed on T2WS and on SR. Thus, a good-quality SR volume is as valuable as T2WS for fetal brain biometric assessment. Our results are in line with and complement a previous
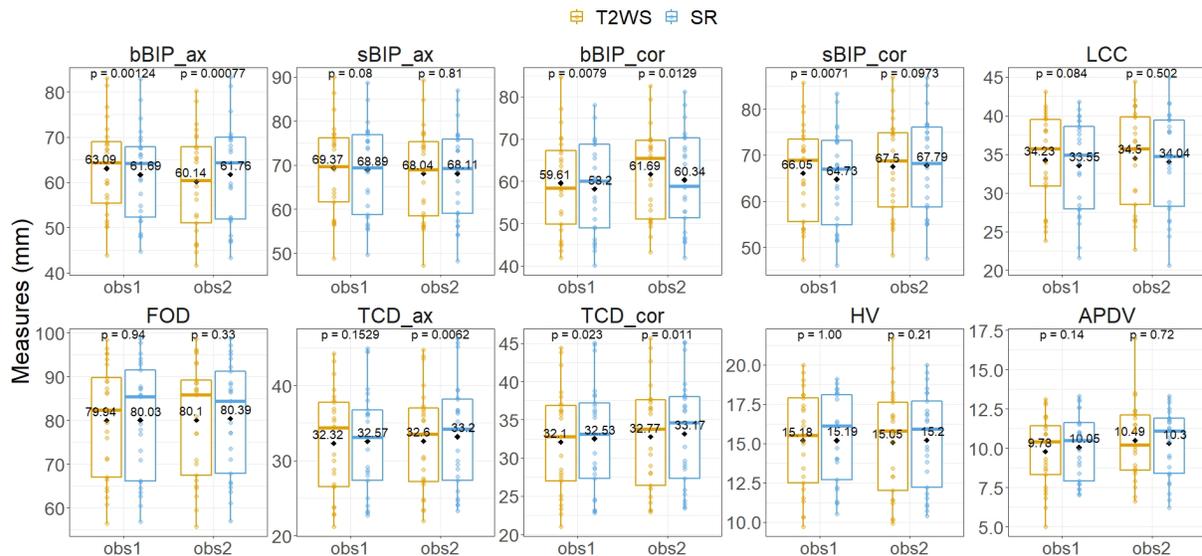
Figure 2.5: Inter-dataset intra-observer comparison for each biometric measurement using the paired Wilcoxon signed rank test. Vertical y-axis, value in mm of each biometric measurement for each fetal brain exam (point) for observer 1 (obs1) and observer 2 (obs2) (horizontal x-axis) for T2-weighted sequence (T2WS; orange) and super-resolution (SR; blue). Solid horizontal lines depict the median; diamonds depict the average of measurement in mm. See Figure 2.3 for measurement abbreviations. $p < 0.05$ is considered significant.

study that has also validated the use of SR for posterior fossa [115]. Our SR biometric measures are also concordant with the normative charts of the fetal brain that have been recently published [89].

The intra-observer agreement in 25 fetuses (both Lin's CCC and ICC were on average 0.997 for both observers) is in line with previous work [89] that also compared T2WS and SR measurements and reported an ICC between 0.95 and 0.99. Furthermore, the error rate is also of the same range (3.3% in average in our study as compared to 0.2 – 2.4%). However, let us note though that in the work by Kyriakopoulou et *al.* [89], the comparison of T2WS vs. SR measurements was performed in a controlled sub-cohort of 10 fetuses specifically selected with symmetrical non-rotated images.

Intra-observer comparisons between T2WS and SR measurements showed significant differences only for axial brain BIP, though differences remained small (2.95 ± 1.73*mm*) and acceptable in clinical practice, as age-specific reference intervals for this biometric parameter are larger [46]. The discrepancies observed in our study for the axial brain BIP may be explained by a different plane used for SR and T2WS and the possibility offered by SR to change the windowing to more precisely identify anatomical landmarks. In contrast to US, there are no true standardized measurement criteria for the BIP on MRI; indeed, the chosen plane, axial or coronal, differs between authors. For instance, measurements of brain and skull BIP reference data were performed in the coronal plane through the temporal horns of the ventricles by some authors [142,
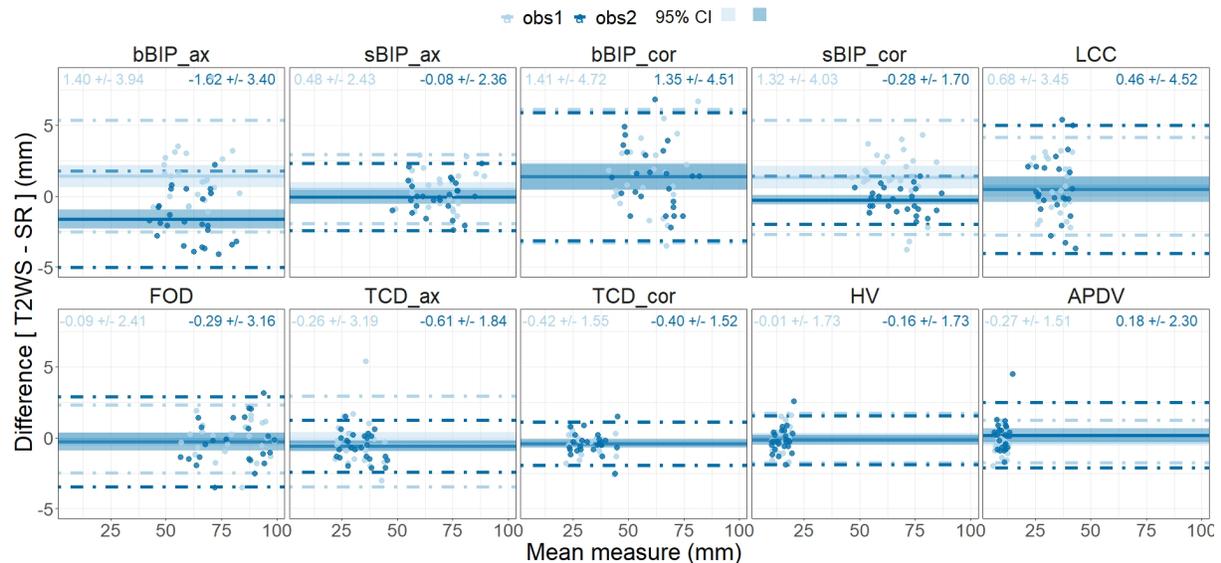
Figure 2.6: Bland-Altman plots of measurement differences between datasets [inter-dataset, T2-weighted sequence (T2WS) vs. super-resolution (SR)] for observer 1 and for observer 2 (intra-observer) for each biometric measurement. Each point corresponds in the x-axis to the mean measure between datasets for each observer (light blue for observer 1, dark blue for observer 2) and on the vertical y-axis to the difference between the two measurements T2WS/SR in mm. Horizontal solid line = mean of all measurement differences. Dashed lines = 95% limits of agreement, and shadow areas correspond to the 95% confidence interval (CI).

45], while it was made in the axial plane by others [89]. We could question if the BIP obtained in the coronal plane really compares to the BIP obtained in the axial plane; however, our results still showed good agreement between transverse and coronal measurements for the BIP. A similar pattern was found for axial vs. coronal TCD, but with a shorter error interval for TCD in SR, similarly to another study [115] that also found superior concordance in SR rather than in T2WS measurement of TCD.

Our study also showed overall good inter-observer agreement for biometric measurements performed on T2WS and on SR, without any statistically significant difference after correction for multiple comparisons. This indicates that experience is not crucial for fetal brain biometric assessment and that SR can be used for biometry even by junior trainee radiologists without extensive expertise in fetal brain MRI. Indeed, obs2 did not report any fetal brain MRI before the study nor received special training. Obs1 showed obs2 how to perform the measurements on one patient and provided schematical guidelines similar to Figure 3. This is to be compared to recent work where the experience of the radiologist (number of previous fetal MRI exams reported) had an influence on the diagnostic error rate, with the less experienced radiologists having higher error rates of 11%, while experts had < 3.8% [9].

In practice, the main additional value of SR is the possibility to reorient the planes in the standard anatomical planes using MPR. In our study, the junior radiologist had more confidence in identifying and measuring the corpus callosum on SR compared
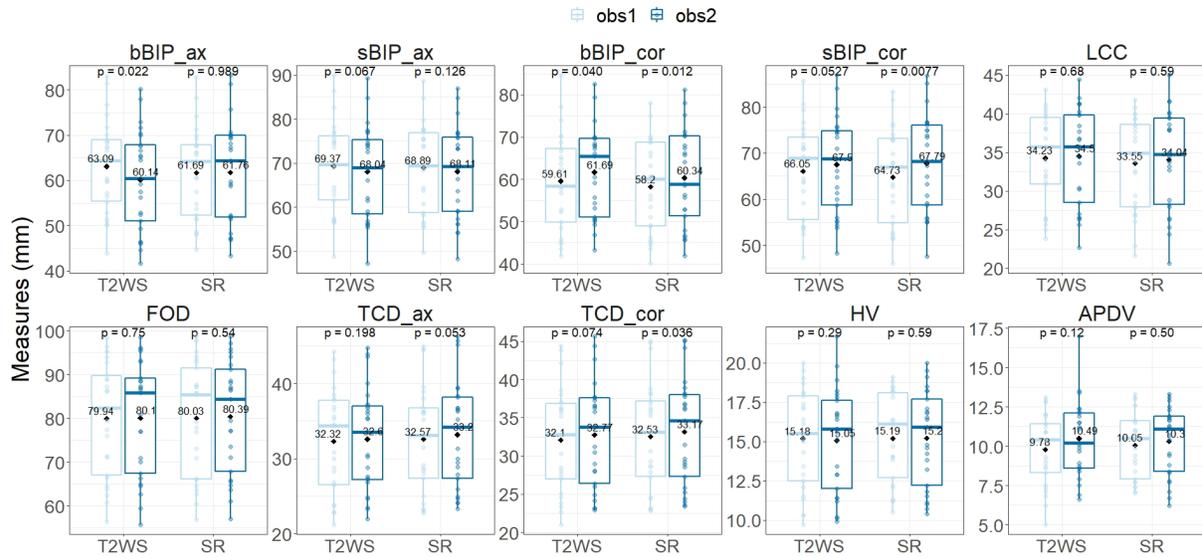
Figure 2.7: Inter-observer intra-dataset comparison for each biometric measurement using the paired Wilcoxon signed rank test. Vertical y-axis, measurement in mm for each biometric measurement for each fetal brain MRI exam (point) for T2-weighted sequence (T2WS) and super-resolution (SR) (horizontal x-axis) for observer 1 (light blue) and observer 2 (dark blue). Solid horizontal lines depict the median; diamonds depict the average of measurement in mm. See Figure 2.3 for measurement abbreviations. $p < 0.05$ is considered significant.

to T2WS. Indeed, it may be difficult to visualize the whole corpus callosum, particularly the rostrum, on conventional LR 2D T2WS, and US is considered to have a better resolution in this case [16, 50]. Our results suggest that SR makes it easier to identify the whole corpus callosum. The potential benefit of SR in this indication should be explored on larger series.

However, the 3D SR reconstruction also has some limitations. Obviously, the quality of SR depends on the quality of the native T2WSs, which are frequently affected by motion. Indeed, due to the reconstruction process itself, small or narrow structures (e.g., optic chiasm or corpus callosum) can appear blurred due to partial volume effects [56]. Nevertheless, in our study, we obtained a good or excellent quality of SR images in 90% of cases, with a good visual rating concordance between three expert raters (one senior engineer and two senior radiologists). The fact that the concordance between T2WS and SR measurements was high indicates that the process of SR reconstruction does not distort the fetal brain anatomy [115]. Our success ratio in SR reconstruction is slightly higher than the one previously reported at 3T [115], where 79% of 62 cases were successfully reconstructed.

SR reconstruction requires a minimum number of orthogonal series to ensure a good reconstruction quality [96, 124]. In our study, six series were used on average (range of 3 – 11) per patient's reconstruction with an average computing time of 1h. Previous works reported a similar number of series and processing time (from 4 to 15, in
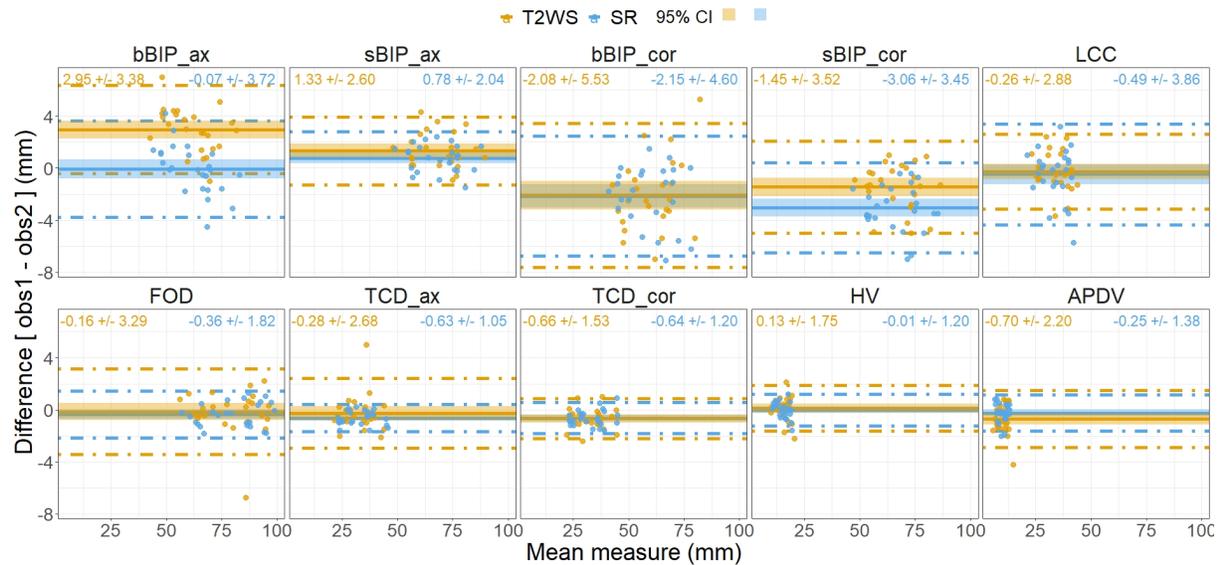
Figure 2.8: Bland-Altman plots of measurement differences between observers (observer 1 vs. observer 2) for each dataset [intra-dataset, T2-weighted sequence (T2WS) and super-resolution (SR)] for each biometric measurement. Each point corresponds in the x-axis to the mean measure between observers and in the vertical y-axis to the difference in mm between the two observers on T2WS (orange point) and SR (blue point). Horizontal solid line = mean of all measurements for T2WS (blue) and SR (red); dashed lines = 95% limits of agreement, and shadow areas correspond to the 95% confidence interval (CI).

average eight series, 1 – 20h in Gholipour et *al.* [56]; and eight series and 40 min in Kyriakopoulou et *al.* [89]). The processing time, the need for user interaction for the series selection or the refinement of brain masks, and the lack of integration into the clinical environment preclude the use of SR volumes for fetal brain exploration in daily practice. However, current developments aim at automatizing and accelerating SR reconstruction [77, 33, 143] to facilitate its integration as a supporting tool in clinical routine and consequently also to be adopted by non-engineer users.

The clinical adoption of SR-reconstructed fetal MRI is thus at its earliest stage [74, 106, 12, 66]. In our center, SR-reconstructed MRI is used only for certain cases, in particular to confirm or reject suspected cortical abnormalities on 2D T2WS. In our opinion, for now, SR could be complementary to but cannot replace low-resolution 2D T2W planes for brain parenchyma analysis, as voxel intensities in SR volumes have undergone many changes. Indeed, beyond the promising value of SR fetal brain biometry, the intensity contrast of SR images still has to be evaluated for its diagnostic value in comparison to native T2WS and eventually improved. Let us recall that voxel intensities in the reconstructed SR volume are computed from the original voxel values within the multiple series throughout many image processing steps; therefore, they are not directly generated from the MRI scanner, and their interpretation as pathological features has to be done with caution.

2.4.2 Strengths and limitations

Our study is the first to compare both supratentorial and infratentorial measurements with SR, while only posterior fossa measurements were made in Pier et *al.* [115]; this latter study included only normal volunteers, while our cases were taken from clinical workflow. We also compared SR measurements between a junior and a senior radiologist to evaluate the influence of experience on biometric results. We have shown that even a radiologist without expertise in fetal brain MRI can perform accurate fetal brain biometry - thanks to SR. In contrast, multiple raters who received extensive training developed normative charts for 2D T2WS and 3D SR fetal brain measurements [89]. In their study, the agreement analysis was performed between measurements on native T2WS and SR in a sub-cohort of only 10 fetuses, with the same rater, and on selected symmetrical nonrotated images. Our study on 25 fetuses comparing two raters provides additional insight into the value of SR techniques for biometry of fetal brain MRI.

A limitation of our study is the small-size cohort, and so the few patients available per gestational week. For this reason, we cannot draw conclusions about gestational age or sex influence on measurement quality. Finally, the reproducibility of our measurements were tested only with one observer on a subcohort of five fetuses.

2.5 Conclusion

This study demonstrates that SR is a valid, reliable, and simple method for biometric measurements. SR measurements are concordant with T2WS measurements, even when conducted by non-expert radiologists. Some biometric measurements like the biparietal diameter show small statistically significant differences, which can be explained by poorly defined and standardized measurement criteria. As soon as full automatization of SR is available in the clinical environment, the use of 3D SR could initially complement conventional T2WSs by faster providing the reference planes and further facilitating biometric measurements. Afterward, SR could provide a new standard of measurements in real orthogonal planes more than mimicking ill-defined or incorrect 2D planes. Future studies on pathological cases will enable the evaluation of other potential benefits of SR in clinical practice.

3

Optimization of SR reconstruction regularization

The work presented in this chapter has been submitted to the 26th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), to be held in October 2023, as **"Simulation-based parameter optimization for fetal brain MRI super-resolution reconstruction"** with co-authors : P. de Dumast, T. Sanchez, H. Lajous and M. Bach Cuadra. P. de Dumast and T. Sanchez contributed equally to this work.

3.1   Introduction

The previous Chapter 2 presented a clinical validation of SR reconstruction practical value. Nonetheless, a key component for high-confidence biometric measurements is not only the enhanced image resolution but also the overall image quality and the ability to identify structures of interest. Consequently, the challenge of SR reconstruction lies in the compromise between the image smoothness and sharpness. In the SR reconstruction formulation presented in Chapter 1, Equation 1.4, the latter is done through the regularization term which contribution is defined by α.

In the context of fetal brain MRI, the absence of HR reference image makes very challenging the optimization of the regularization parameter α. Indeed, while the problem of optimizing regularization parameters has been long studied in inverse problems [43, 3], most existing approaches explicitly rely on reference data. In order to overcome the lack of HR fetal brain MRI, some works [88, 144] proposed to degrade HR newborn MR images to generate LR images and further assess the SR reconstruction against the original image. Alternatively, both works considered a leave-one-out approach, where a LR stack is used to serve as reference for quantitative evaluation [88, 144]. Similarly, Ebner et al. [33] used a volume reconstructed from all available LR series as reference,

to assess the quality of reconstructions using fewer LR series. However, they did not study the influence of the regularization parameter on the quality of reconstruction. Besides, most studies based on SR reconstruction are still carried out using pre-set default values [112, 145] (see Chapter 2).

In this study, we propose to leverage a fetal brain MRI numerical simulations framework [91] to improve the setting of regularization. This numerical phantom generates realistic LR T2w MR images of the fetal brain as well as corresponding HR volumes. We use the pair of LR images with HR reference image to optimize the parameter α in a data-driven manner. Our contribution is twofold. First, using numerically simulated data, we study the sensitivity of the regularization parameter to three common variables in fetal brain MRI: *(i)* the number of series used in the SR reconstruction, *(ii)* the magnetic field strength and *(iii)* the GA. Second, we qualitatively assess the translation of our approach on clinical MR exams showing that optimized regularization parameter α on simulated data echos a substantial improvement of the clinical SR reconstructed image quality. To generalize the validity of our findings, we perform our study using two state-of-the-art SR reconstruction pipelines, namely MIALSRTK [144] and NiftyMIC [33], that apply different regularization terms.

3.2 Materials and methods

3.2.1 Simulated acquisitions

Fetal Brain magnetic resonance Acquisition Numerical phantom (FaBiAN) is a numerical phantom for the simulation of realistic fetal brain MR images [90, 91]. From a fetal brain model, i.e. a tissue segmentation labelmap, acquisition of SST2W sequences are simulated, accounting for the tissue properties (T1 and T2 values) and the acquisition process. Additionally, the level of inter-slice motion can be specified.

In this study, MR images are either simulated at 1.5T, with a resolution of $1.1 \times 1.1 \times 3mm^3$ or 3T at $0.5 \times 0.5 \times 3mm^3$, based on the CRL spatiotemporal atlas [56] (see Section 1.4.2) featuring one subject per week from 21 to 38 weeks of GA. We generate both LR clinical-like images and a reference HR isotropic volume. All acquisition parameters are kept fixed at a given magnetic field strength. We further detail the amplitude of fetal motion and number of simulated series generated in the experimental settings in Sections 3.2.3 and 3.2.4.

3.2.2 Super-resolution reconstruction

Reconstruction frameworks

We use two widely adopted reconstruction pipelines, MIALSRTK [144] and NiftyMIC [33], to reconstruct 3D isotropic HR images of the fetal brain from orthogonal LR series. Both methods feature the preprocessing of the LR images, the estimation of the fetal motion and the finale image reconstruction. In the latter step, pipelines differ by the regularization term applied: while MIALSRTK uses total variation (TV) [144], NiftyMIC uses first-order Tikhonov [33]. For each pipeline, we perform a grid search approach of the regularization parameter space.

Note that, contrary to NiftyMIC [33], MIALSRTK [144] places its regularization parameter λ on the data term. For the sake of consistency, we will only use the formulation of Equation 1.4, with $\alpha = 1/\lambda$ in the case of MIALSRTK.

Quality assessment

Solving Problem 1.4 yields a SR reconstructed image $\hat{\mathbf{X}}$ which quality can be compared against the simulated reference using various metrics. We use two common metrics for SR reconstruction assessment [88, 144, 33], namely the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) [150]. The best regularization parameter α is identified as the one maximizing a given performance metric.

3.2.3 Experiment 1 – Controlled environment

In this first experiment, we study the sensitivity of the parameter α to common variations in the acquisition pipeline.

Dataset

For every CRL subject [56], we simulate nine LR series (three per anatomical orientation) at 1.5T and 3T with little amplitude of stochastic 3D rigid motion[1].

Experimental setting

We define four configurations based on the number of LR series given as input to the SR reconstruction pipeline (three or six series) and the magnetic field strength (1.5T or 3T). Note that the inter-magnetic field difference is particularly captured in the image

---

[1]5% of corrupted slices, translation of [-1,1]mm in every direction, 3D rotation of [-2,2]°

resolutions, with through-plane in-plane ratio of 3.3/1.1 = 3 for 1.5T and 3.3/0.5 = 6.6 for 3T. In each configuration, individual brains are repeatedly reconstructed (n = 3) from a selection of different LR series among the 9 available per subject.

The grid of regularization parameter searched for NiftyMIC consists of 10 values geometrically spaced between 10^{-3} and 2, plus the default parameter α_{def} = 0.01. For MIALSRTK, we use $\alpha \in \{1/0.75, 1/1.0, 1/1.5, 1/2.0, 1/2.5, 1/3.0, 1/3.5, 1/5.0\}$ (8 values including the default parameter α_{def} = 1/0.75). At the end of the experiment, the best parameter, for either of the pipelines, is referred to as α_1^*.

Statistical analysis

The optimal regularization parameters evaluated for the different SR configurations are compared using the Wilcoxon rank sum test. The difference between the metrics performance obtained with default or optimal parameters is tested with a paired Wilcoxon rank sum test. The p-value statistical significance is set to 0.05.

3.2.4   Experiment 2 – Clinical environment

Clinical MR fetal exams are prone to substantial inter-patient variation and heterogeneity. In particular, the number of LR series available for reconstruction greatly varies from patient to patient, as well as the amplitude of fetal motion [80].

Therefore, this second experiment has two purposes. First, we translate our findings from the first experiment to clinical data using the best value α_1^*. Secondly, we study an alternative approach to perform a tailored patient-wise regularization tuning by simulating synthetic data for each patient that mimic the clinical acquisitions available. We refer to the obtained value as α_2^*.

Dataset

Twenty fetal brain MR exams conducted upon medical indication were retrospectively collected from our institution. All brains were finally considered normal. Fetuses were aged between 21 and 34 weeks of GA (mean ± SD: 29.7 ± 3.6) at scan time. At least three orthogonal series were acquired at 1.5T (voxel size: $1.125 \times 1.125 \times 3mm^3$). After inspection, four to nine series (mean ± SD: 6.3 ± 1.5) were considered exploitable for SR reconstruction.

In addition, simulated acquisitions are performed with exam-specific parameters chosen to mimic as closely as possible the clinical acquisitions of each patient. In particular, we match the number, the orientation and the level of motion (from little to moderate) of the LR series, as well as the GA of each patient.

The local ethics committee of the Canton of Vaud, Switzerland (CER-VD 2021-00124) approved the retrospective collection and analysis of MRI data and the prospective studies for the collection and analysis of the MRI data in presence of a signed form of either general or specific consent.

Experimental setting

We consider the same regularization parameter space as in Experiment 1 (Section 3.2.3), and evaluate both clinical and simulated data on this parameter grid.

Statistical analysis

We compare the similarity between the reconstructed images of MIALSRTK and NiftyMIC using both default and optimized parameters. In this experiment, no reference images are available for clinical exams. Statistical significance of the performance difference is tested using a paired Wilcoxon rank sum test ($p < 0.05$ for statistical significance).

3.3 Results

3.3.1 Experiment 1 – Controlled environment

Optimal regularization parameter

Figure 3.1 shows the optimal regularization parameters α_1^* of SR reconstruction through MIALSRTK and NiftyMIC for each configuration. Regardless of the magnetic field strength and the number of LR series used for reconstruction, the optimal regularization parameters that maximize the PSNR and SSIM compared to a synthetic HR volume greatly differ from the default values. For MIALSRTK, $\alpha_{\text{def}} = 1/0.75$, while the optimal range is found between 1/2.25 and 1/4.5. For NiftyMIC $\alpha_{\text{def}} = 0.01$, whereas the optimal range is found between 0.015 and 0.15.

We observe that for both the PSNR and the SSIM, the optimal regularization weight *increases* with the number of series used in the reconstruction, and decreases with the resolution. This is because changing the number of LR series or the magnetic field strength changes the magnitude of the data term with respect to the regularization term. When more series are used for reconstruction, a larger regularization parameter must be used to keep the ratio data fidelity over regularization constant.
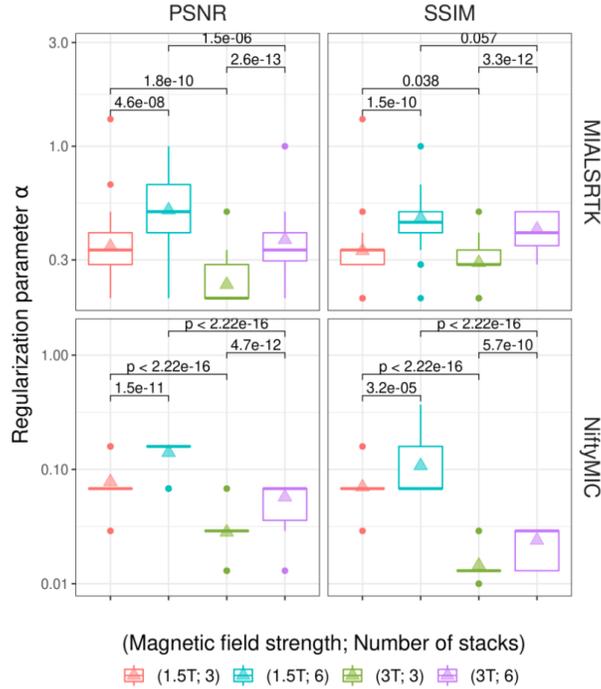
Figure 3.1: Optimal regularization parameters α_1^* using MIALSRTK or NiftyMIC for the four different configurations studied. $\triangle$ indicates the mean optimal parameter. Inter-configurations p-values (Wilcoxon rank sum test, statistical significance: $p < 0.05$) are indicated.

Quality improvement

The corresponding mean PSNR and SSIM values obtained with optimal α_1^* regularization parameters across all subjects are displayed in Table 3.1. Overall, the quality metrics obtained with optimal parameters are significantly improved compared to default values. This suggests that regularization parameters are highly sub-optimally set for both reconstruction pipelines.

GA-based analysis

Since the human brain undergoes drastic morphological changes throughout gestation (see Section 1.1), one could expect to adjust α to GA. However, in our experiments, we do not observe an influence of GA on the value of α that yields to the best reconstruction, which was still in line with the values reported on Figure 3.1.

3.3.2 Experiment 2 – Clinical environment

In this experiment, we compare two differently optimized regularization parameters. First, we use the optimal value α_1^* (from Figure 3.1 at 1.5T, and rounded to the clos-

Table 3.1: Mean metrics (PSNR and SSIM) for SR reconstructions using the default regularization parameters α_{def} or the optimal parameter α_1^* compared to the simulated reference HR volume. Paired Wilcoxon rank sum test statistical significance ($p < 0.05$) is indicated as †.

| | MIALSRTK | | | | NiftyMIC | | | |
| | PSNR (↑) | | SSIM (↑) | | PSNR (↑) | | SSIM (↑) | |
Setting	α_{def}	α_1^*	α_{def}	α_1^*	α_{def}	α_1^*	α_{def}	α_1^*
(1.5T; 3)	18.9	$\mathbf{20.2}^{\dagger}$	0.78	$\mathbf{0.80}^{\dagger}$	17.3	$\mathbf{20.8}^{\dagger}$	0.79	$\mathbf{0.82}^{\dagger}$
(1.5T; 6)	20.1	$\mathbf{20.8}^{\dagger}$	0.82	$\mathbf{0.83}^{\dagger}$	17.0	$\mathbf{21.5}^{\dagger}$	0.80	$\mathbf{0.84}^{\dagger}$
(3T; 3)	19.9	$\mathbf{21.8}^{\dagger}$	0.75	$\mathbf{0.77}^{\dagger}$	20.5	$\mathbf{21.2}^{\dagger}$	0.77	$\mathbf{0.77}^{\dagger}$
(3T; 6)	21.0	$\mathbf{22.2}^{\dagger}$	0.78	$\mathbf{0.80}^{\dagger}$	20.9	$\mathbf{22.0}^{\dagger}$	0.80	$\mathbf{0.80}^{\dagger}$
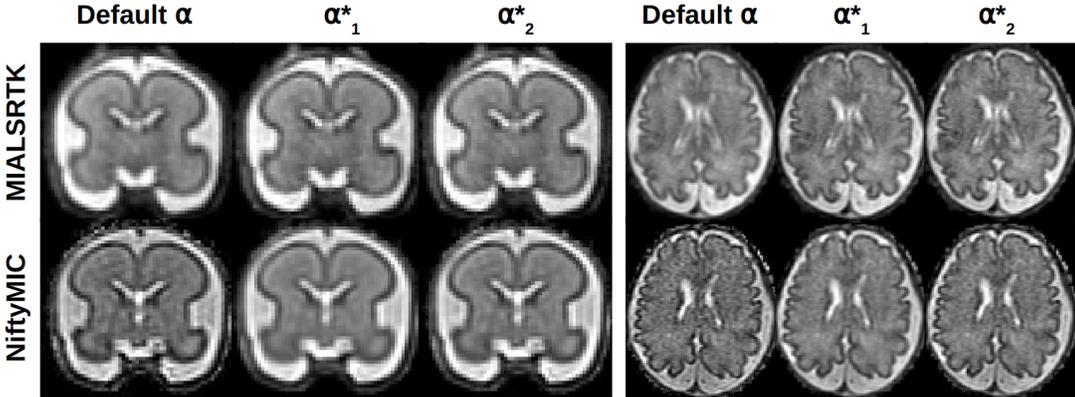
Figure 3.2: Comparison of SR-reconstructed clinical cases using default and optimal parameters from Experiment 1 (α_1^*) and Experiment 2 (α_2^*). Top row: MIALSRTK, bottom row: NiftyMIC. Left: coronal view of a 23 weeks-old fetal brain reconstructed using eight LR series. Right: axial view of a 32 weeks-old fetal brain reconstructed using four LR series.

est value on the searched grid of parameters). Second, we use α_2^* estimated from the patient-specific simulation.

Figure 3.2 shows the SR reconstruction of two patients with default and optimal parameters using both pipelines. We first observe that the qualitative difference between the reconstruction obtained using α_1^* and α_2^* for a given pipeline is subtle. One can also notice that the volumes reconstructed with simulation-based optimized parameters from both SR methods appear more similar to each other, compared to the reconstructions using default parameters. Indeed, the default parameters of MIALSRTK and NiftyMIC promote opposite behaviours: the first towards smooth image (default regularization higher than optimal one), while the second makes it appear noisier (default regularization lower than optimal).

We quantitatively confirm the closeness of the optimized SR images by computing the PSNR and the SSIM between the reconstructed images from both methods – no references are available for the clinical acquisitions – for α_{def}, α_1^* and α_2^*. The results are shown on Figure 3.3, where we notice a net difference between the default parameters
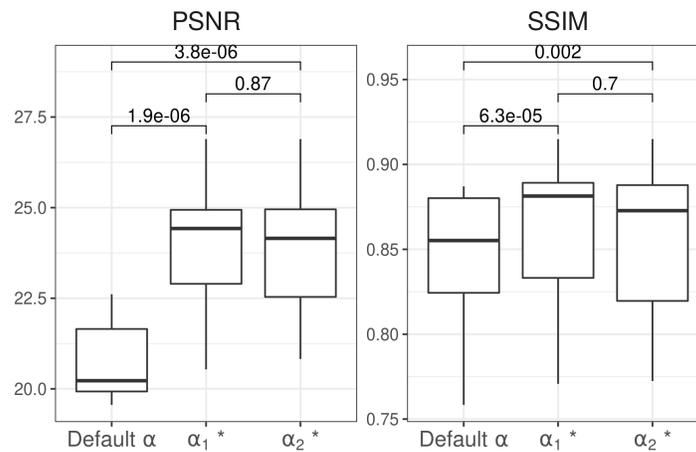
Figure 3.3: Inter-method similarity between MIALSRTK and NiftyMIC, with default and optimal regularization parameters, using all 20 clinical exams. Inter-regularization parameters p-values are indicated (paired Wilcoxon rank sum test, statistical significance: $p < 0.05$).
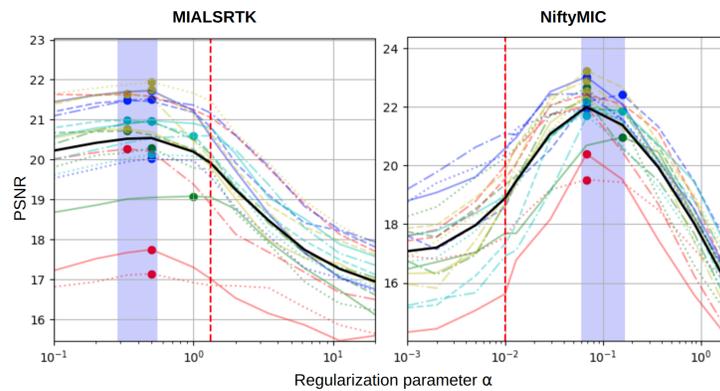
Figure 3.4: PSNR curves for the simulated patients of Experiment 2, reconstructed with MIALSRTK [143] (left) and NiftyMIC [33] (right). For each SR method, light traits represent individual patients and the bold black curve the averaged PSNR across all patients. The vertical red line recalls the default parameter values, and the purple region highlights the range of optimal parameters found in Experiment 1 at 1.5T.

and the optimized ones, which is statistically significant for both PSNR and SSIM. In addition, there is no significant difference between the images reconstructed with the setting-wise optimized parameter α_1^* and the patient-wise optimized one α_2^*. This is expected, as Figure 3.4 confirms that the optimal patient-wise parameters α_2^* lie within the range where the optimal parameters α_1^* are found in Experiment 1.

3.4 Conclusion

In this chapter, we quantitatively address the automated optimization of the regularization weight in the inverse problem of the SR reconstruction of fetal brain MRI. We demonstrate that our principled approach of the regularization parameter optimization on synthetic data successfully transfers into an important qualitative gain in clinical acquisitions reconstruction. Our study shows that it might not be necessary to use *patient-specific* simulations, as optimization at a *setting-specific* level is equivalent and can be ideally estimated beforehand. Nonetheless, the quantitatively-determined optimal value may not be in line with the experts perceptual appreciation. Therefore, we recommend to use our approach as a coarse tuning method for the order of magnitude of α rather than as a true validated optimal value. Thus our proposed method has a high practical value in clinical settings where multi-scanner multi-sequence data are to be dealt with. Nonetheless, extensive analysis across SR reconstruction pipelines and diverse dataset remain to be done to validate the generalizability of our approach. We emphasize that we did not evaluate our method on pathological cases in which the accurate preservation of localized hypo- or hyper-intensities may be crucial.

Moreover, we recall that the quality of a HR volume generated from a whole SR reconstruction pipeline does not only rely on the image restoration step itself, but also on the accuracy of the prior steps (e.g. LR images intensity corrections or motion estimation) and the quality of the input images. Common clinical acquisition artifacts, as for instance some signal drops or surrounding tissues remaining from the gross brain extraction, cannot be generated with the numerical simulator we use, and hence limit its usage.

Nonetheless, we concomitantly show that our optimization approach reduces the variability between the two reconstruction methods. This contributes to reducing the domain-shift due to reconstruction, currently a key challenge in the development of automated tissue segmentation methods [111]. We further address the domain-shift between SR reconstruction methods in Part II.

Part II

Domain variations-robust segmentation methods

4

Synthetic data augmentation for domain adaptation

This chapter compiles three contributions presented as: *(i)* an application experiment in the original publication **"A Fetal Brain magnetic resonance Acquisition Numerical phantom (FaBiAN)"** in **Scientific Reports Vol.12, p8682 (2022)** with co-authors: H. Lajous, C. W. Roy, T. Hilbert, P. de Dumast, S. Tourbier, Y. Alemán Gómez, J. Yerly, T. Yu, H. Kebiri, K. Payette, J.B. Ledoux, R. Meuli, P. Hagmann, A. Jakab, V. Dunet, M. Koob, T. Kober, M. Stuber, M. Bach Cuadra, *(ii)* a conference paper at the IEEE International Symposium on Biomedical Imaging (ISBI) 2022 (oral communication) **"Synthetic magnetic resonance images for domain adaptation: Application to fetal brain tissue segmentation"** with co-authors: P. de Dumast, H. Kebiri, K. Payette, A. Jakab, H. Lajous, M. Bach Cuadra (H. Lajous and M. Bach Cuadra equally contributed to this work), and *(iii)* a conference abstract at the 31st Annual Meeting of the International Society for Magnetic Resonance in Medicine (ISMRM) in 2022 (oral communication) **"On the importance of fetal brain numerical models for domain adaptation strategies in fetal brain MRI tissue segmentation"** with co-authors: P. de Dumast, H. Kebiri, M. Bach Cuadra, H. Lajous. M. Bach Cuadra, H. Lajous equally contributed to this work.

4.1   Introduction

The development of robust DL-based methods for fetal brain tissue segmentation in MRI is hampered by the large variability represented in clinical images and the lack of annotated dataset. As presented in Chapter 1, Section 1.4.2, the first non-fully patient-dependant source of variability in fetal brain MRI is the image acquisition system, followed by the postprocessing operations. Moreover, TL has recently proven to be effective in adapting automatic segmentation methods for fetal brain MRI from different reconstruction methods [111].

In this chapter, we aim to achieve domain adaptation between SR reconstruction methods. Specifically, to overcome the lack of annotated data, we study how synthetic MR images that are processed to mimic the acquisition and postprocessing of our target domain can help to adapt a tissue segmentation network trained from another domain. We hypothesize that SR reconstruction of synthetic MR images can alleviate the lack of true target images in the development of segmentation methods.

Our contribution is threefold. First, we evaluate the practical value of numerically-simulated MR images to serve as data augmentation in the training of automatic fetal brain tissue segmentation methods (see Section 4.2.3). Second, we leverage the synthetic MR images to further replicate the whole target domain image generation, i.e. the acquisition and the reconstruction, in a domain adaptation task (see Section 4.2.4). Finally, we investigate the importance of the fetal brain model used in for the numerical simulations (see Section 4.2.5).

4.2 Materials and Methods

4.2.1 Datasets

Acquisitions

Eighty (80) fetal brains were scanned at the University Children's Hospital Zurich (Kispi) on medical indications. Fetuses were aged from 20 to 34.8 weeks of GA (mean ± SD: 27.0 ± 3.6). MR exams were either performed at 1.5 or 3T on GE Healthcare scanners, resulting to LR images of $0.5 \times 0.5 \times 3mm^3$ resolution. In this cohort, 33 patients were neurotypical and 47 were diagnosed with neuropathological development.

Postprocessing

All patients were SR-reconstructed either with MIALSRTK [143] (N=40) to an isotropic resolution of $0.5mm$ or with SVRTK [88] (N=40) to an isotropic resolution of $0.8594mm$. The extra-axial CSF, the cGM, the WM, the ventricles, the cerebellum, the dGM and the brainstem were manually annotated on the HR volumes following the FeTA annotation guidelines [112]. All SR-reconstruction and their intracranial tissue labelmap were publicly released in the *FeTA* dataset v2.0 [112] (see Section 1.4.2).

The subset of MIALSRTK-reconstructed images was resampled to an isotropic resolution of $0.8mm$ and annotations were refined [39]. An engineer with 20 years of experience in medical image processing assessed the quality of all the 3D SR reconstructions. Quality categories followed the same criteria as in Chapter 2. Only patients with excellent and acceptable SR were considered in the rest of this study.

Furthermore, fifteen (15) clinical fetal brains of the SVRTK reconstructed subset were provided with the LR T2w images (*FeTA-lrsr*). MR exams were either

acquired at 1.5T (N=6) or 3T (N=9). We wrap up in Table 4.1 the three datasets used in this study, based on their reconstruction method and the data available. The provenance of each dataset is detailed in Figure 4.1.

Domains

SR reconstruction techniques apply many intensity-based operation that may lead to different data distribution and thus different domains (see Chapter 1, Section 1.4.2). In this Chapter, we consider the set of clinical fetal brain MRI reconstructed with SVRTK [88] as the **source domain**, and the clinical fetal brain MRI reconstructed with MIALSRTK [143] represent the **target domain** (see Figure 4.1).

Atlas

We consider the tissue labelmaps of the *CRL* spatiotemporal atlas [56] as fetal brain model for the simulations.

4.2.2 Processing

Simulations

We introduced in Chapter 3 FaBiAN [90, 91], a numerical phantom for the simulation of realistic fetal brain MR images. From a fetal brain model, i.e. a segmentation labelmap, this simulation framework generates T2w LR MR images at either 1.5T or 3T.. Additionally, automated segmentation of the fetal brain is propagated in all LR image series from the initial fetal brain model by a nearest-neighbour interpolation. All simulations rely on *FeTA-lrsr* acquisition parameters.

Reconstruction

Reconstruction of numerically simulated images are performed through the same method than the target domain images, i.e. MIALSRTK. The SR reconstruction method is adapted to reconstruct the associated labelmaps.

Segmentation model

All segmentation models follow the same 2D U-Net [122] architecture, as state-of-the art fetal brain tissue segmentation methods all rely on it (see Chapter 1, Section 1.4.2). Networks are trained on 64×64 voxel size overlapping image patches considering intracranial voxels only. Networks are optimized with the minimization of a hybrid loss function defined as the sum of a categorical cross-entropy and a Dice loss.

In the following experiments, configurations where SR reconstructed isotropic images are available (i.e. in experiments 4.2.4 and 4.2.5) are trained using a multi-view approach: patches from the three orientations (axial, coronal and sagittal) are used as input. At test-time, overlapping patches are inferred from the three orientations and final prediction is reconstructed using a majority voting strategy.

Table 4.1: Summary of clinical datasets. NT: Neurotypical; PT: Pathological.

	Number (NT/PT)	GA (weeks) (mean ± SD)	Reconstruction method	Data		Public
				LR	HR	
FeTA-mial	31 (15/16)	20.0-33.4 (27.5 ± 3.2)	MIALSRTK [143]	-	T2w+annotations ($0.80 \times 0.80 \times 0.80 mm^3$)	Yes
FeTA-svrtk	37 (17/20)	20.1-34.8 (27.4 ± 4.0)	SVRTK [88]	-	T2w+annotations ($0.86 \times 0.86 \times 0.86 mm^3$)	Yes
FeTA-lrsr	15 (13/2)	21.0-34.6 (27.5±4.4)	SVRTK [88]	T2w ($0.5 \times 0.5 \times 3 mm^3$)	T2w+annotations ($0.86 \times 0.86 \times 0.86 mm^3$)	No
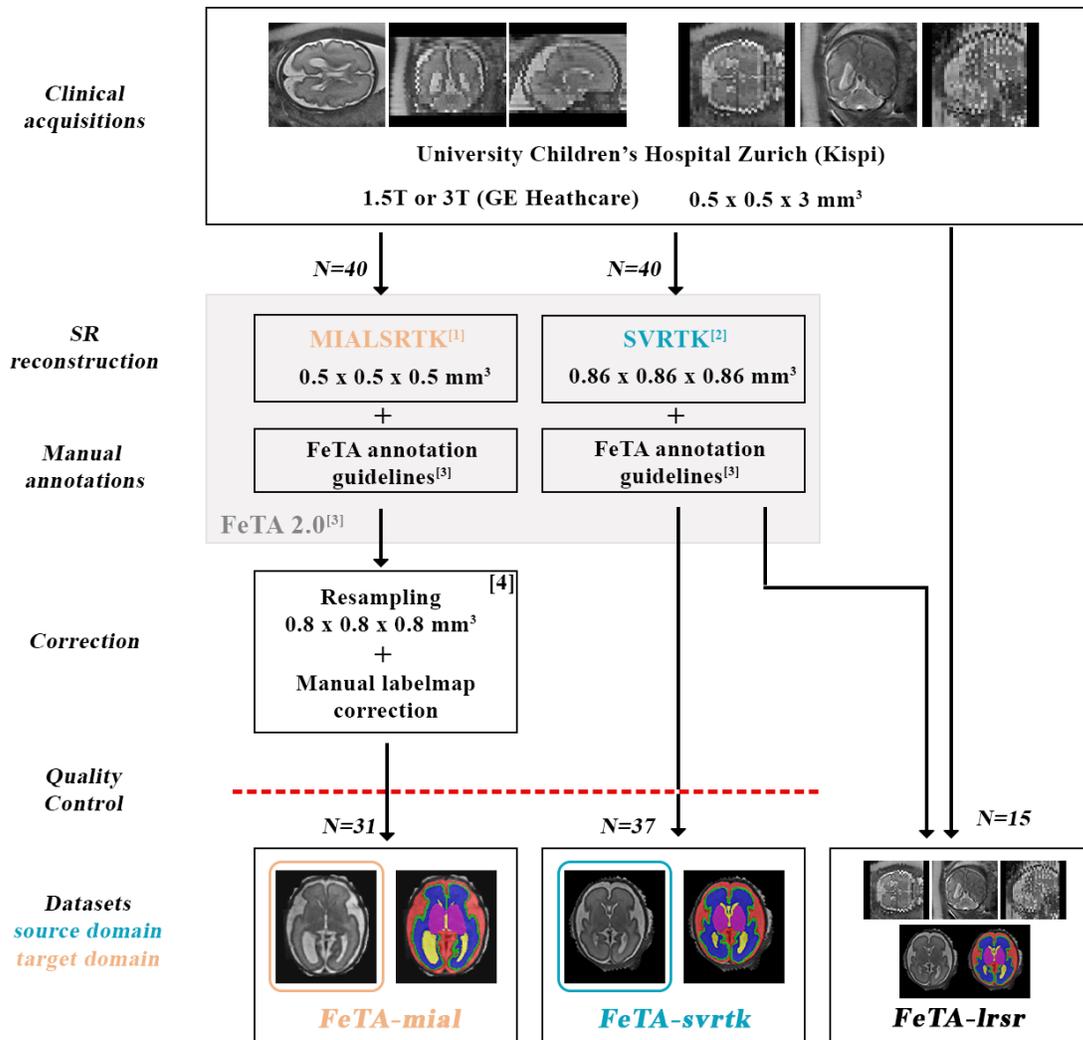
[1] Toubier et *al.*, 2020, [2] Kuklisova-Murgasova et *al.*, 2012,
[3] Payette et *al.*, 2021, [4] Fidon et *al.*, 2021.

Figure 4.1: Summary of clinical datasets. Acquisition, postprocessing and generation of derivatives are detailed for *FeTA-mial*, *FeTA-svrtk* and *FeTA-lrsr* datasets.
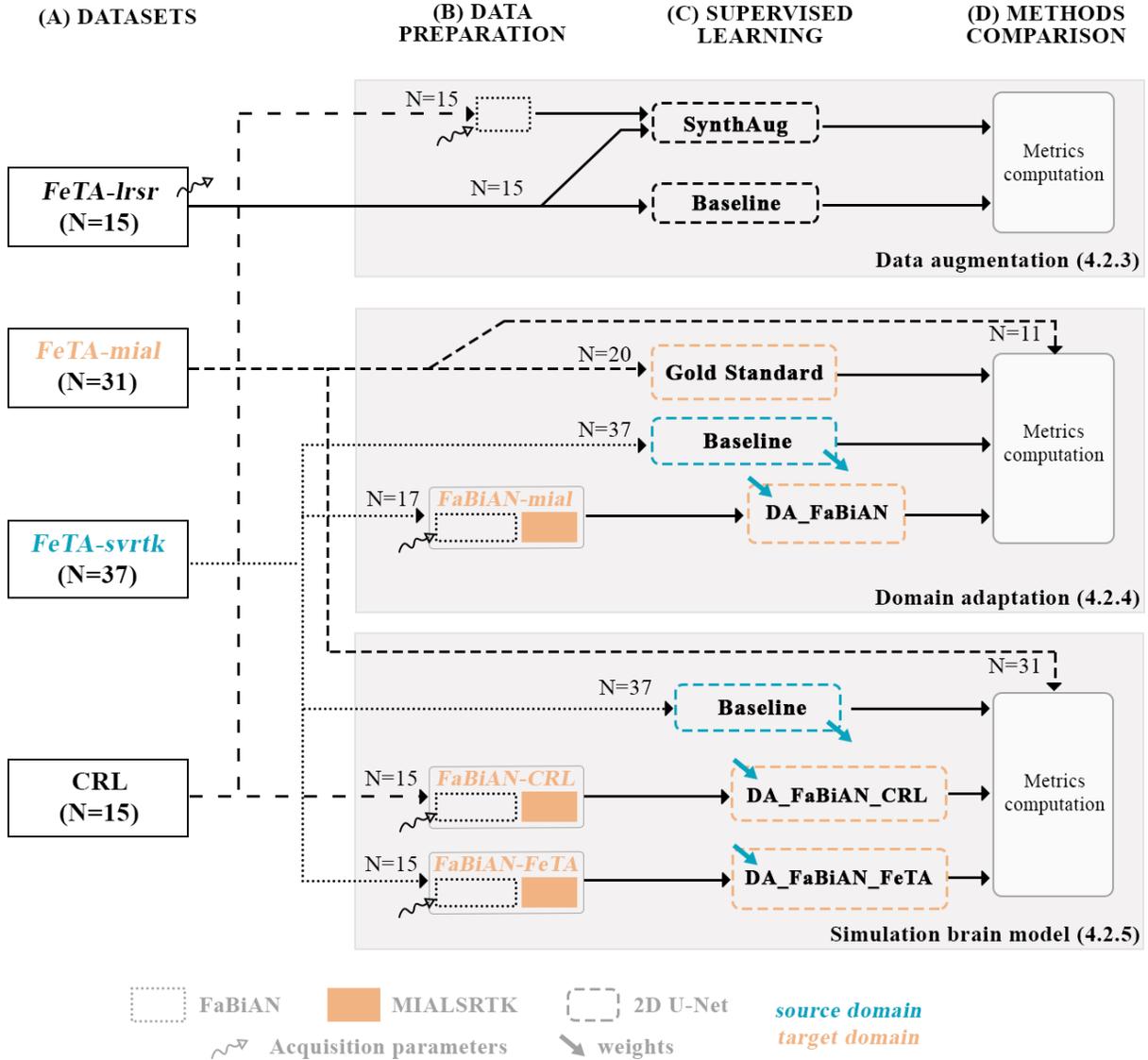
Figure 4.2: Design of synthetic MRI-based domain adaptation experiments. Datasets (A), data preparation (B), supervised learning (C) and methods comparisons (D) are presented for each independent experiment (4.2.3) data augmentation, (4.2.4) domain adaptation, and (4.2.5) simulation brain model.

4.2.3 Data augmentation

The design of the data augmentation experiment is presented in Figure 4.2, top panel.

CRL subjects from 21 to 35 weeks of GA are used as fetal brain model. For each subject, three partially overlapping numerical simulations of SST2W sequences are generated in axial orientation: 2 with low and 1 with moderate motion. Acquisition parameters from *FeTA-lrsr* are replicated. Note that, *CRL* brain model does not differentiate external CSF and ventricles. Consequently, ventricles and CSF are merged

under a unique CSF label in the clinical exams to match the annotation tissue classes of the simulations.

In the training phase of the segmentation frameworks, we consider three categories of augmentation: *(i)* spatial augmentation (random flipping and rotation (n times, by $n \times 90°$, $n \in [0,3]$)), *(ii)* intensity-based augmentation (extensive conventional augmentation strategies used including random Gaussian noise, random gamma, and random bias field [114]), and *(iii)* synthetic augmentation which consists in the upstream generation of synthetic MR images with FaBiAN.

Based on these types of augmentation, we evaluate the practical value of FaBiAN as an augmentation strategy by comparing the following two training configuration: **Baseline** is trained with on-the-fly intensity-based augmentation strategy. **SynthAug** is trained with synthetic images to complement the training clinical dataset. Both Baseline and SynthAug configurations are trained with on-the-fly spatial augmentation. Table 4.2 summarizes the data used in each configuration and the data augmentation performed at train-time. To maintain a similar number of input samples, each patch from configuration SynthAug is repeated once, while patches from configuration Baseline are repeated three times. To avoid any bias coming from the number of training samples, we ensure an equivalent number of 2D patches between a SR-reconstructed clinical case: three shifted axial LR images are used for a given subject. Intensities of all image patches are standardized.

We adopt a five-fold cross-validation (CV) strategy. Due to the small sample size of the *FeTA-lrsr* dataset, no set is kept apart for testing purposes. Consequently, we only present the average CV performances.

Table 4.2: Summary of the data augmentation configurations.

Configuration	Amount of data				On-the-fly augmentation
	Number of clinical subjects	Number of simulated subjects	Total number of subjects	Number of replicates	
Baseline	15	0	15	4	Spatial Intensity-based
SynthAug	15	15	30	2	Spatial

4.2.4 Domain adaptation

In this domain adaptation experiment (Figure 4.2, middle panel), we introduce a new domain to which we refer as target-like domain. Target-like domain images are numerically-simulated fetal brain MRI reconstructed with the same SR method as the target domain images, i.e. MIALSRTK [143].

The manual annotations of the SR reconstructions of 17 clinical cases (9 neurotypical and 8 pathological patients in the GA range of 20.9 to 34.8 weeks) from *FeTA-svrtk* serve

as fetal brain models to generate realistic synthetic images of the fetal brain throughout development with FaBiAN. For every patient, partially-overlapping series of 2D thick slices are simulated in each of the three orthogonal orientations with random little to moderate rigid motion. The synthetic LR images are interpolated to $0.8594 \times 0.8594 mm^2$ in the in-plane direction to match the resolution of the clinical SR reconstructions *FeTA-mial* and *FeTA-svrtk*. We refer to this **target-like domain** as *FaBiAN-mial*. Figure 4.3 shows the original *FeTA-mial* image, a FaBiAN-simulated LR image and the reconstructed target-like domain *FaBiAN-mial* image of a 30.3 weeks of GA fetal brain.
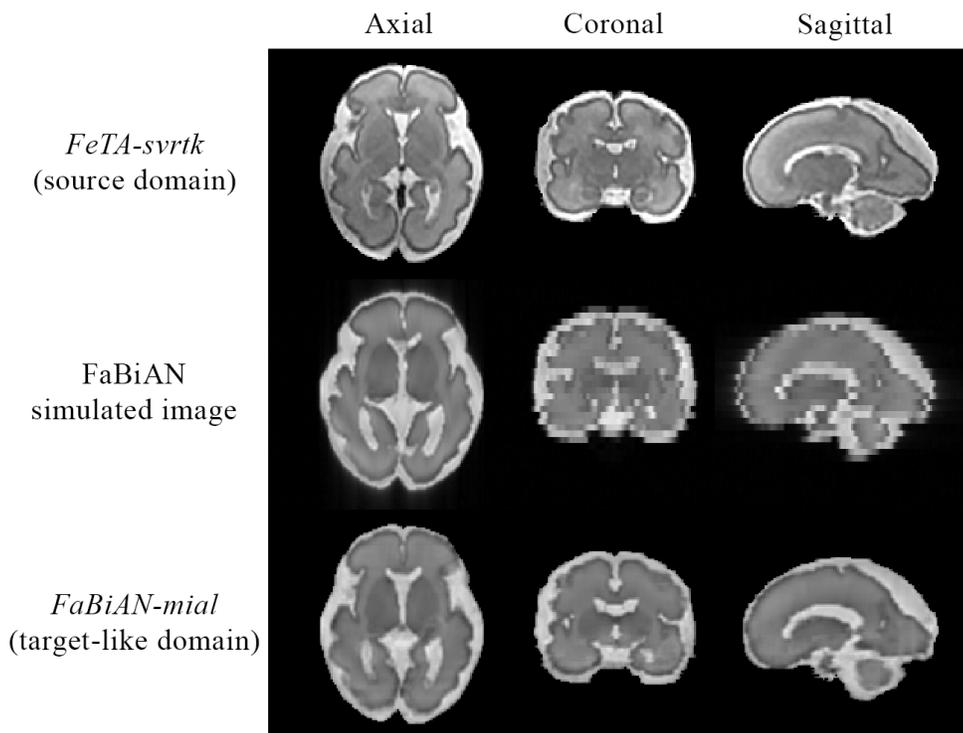
Figure 4.3: Illustration of the source domain *FeTA-svrtk*, a FaBiAN-simulated LR and the target-like domain *FaBiAN-mial* images for a fetus of 30.3 weeks of GA.

Three training configurations are studied in this experiment:

- A **Gold Standard** network is trained using 20 randomly selected patients from *FeTA-mial*. This configuration represents an ideal setting where target domain images are available during the training phase. (N=20, target domain)

- A **Baseline** network is trained on the full *FeTA-svrtk* set. Only images in the source domain are available during the training phase to mimic a more realistic setting. (N=37, source domain)

- **DA_FaBiAN** is initialized with Baseline pre-trained weights to perform TL using the 17 *FaBiAN-mial* subjects. (N=17, target-like domain)

Note that, an additional configuration was originally computed, where the fetal brains used as models for FaBiAN simulations were not used in the training phase. For clarity, we do not present this configuration here. All configuration are trained with spatial and intensity-based augmentation, using a CV strategy.

A pure testing set that consists of the 11 remaining *FeTA-mial* subjects (five neurotypical and six pathological patients) is used to compare the models.

4.2.5 Simulation brain model

In this experiment (see Figure 4.2, bottom panel), we further explore in a domain adaptation task the influence of the fetal brain model used for the FaBiAN simulations. Hence, we generate two sets of target-like domain images (see Table 4.3), using either the clinical *FeTA-svrtk* images or *CRL*. We respectively refer to these two new target-like domain sets as *FaBiAN-FeTA* and *FaBiAN-CRL*. While simulations based on *CRL* rely on an averaged anatomy, using *FeTA-svrtk* patients as models of the fetal brain provides larger anatomical and annotation variability. For every patient, partially-overlapping orthogonal series of 2D thick slices are generated with random little to moderate motion. Similarly to Experiment 4.2.3, ventricles and CSF are merged in FeTA images to match the annotation tissue classes of the CRL images.

Table 4.3: Simulation-based datasets from different fetal brain models.

	Model	Acquisition parameters	Reconstruction method
FaBiAN-FeTA	*FeTA-svrtk* (20.9-34.8 weeks) (8/7)	*FeTA-lrsr*	MIALSRTK [143]
FaBiAN-CRL	*CRL* (21-35 weeks)	*FeTA-lrsr*	MIALSRTK [143]

Three configurations are trained and compared. A **Baseline** network is trained on the entire *FeTA-svrtk* dataset. Similarly to the previous domain adaptation experiment presented in Section 4.2.4, the domain adaptation networks are initialized with the Baseline pre-trained weights to perform TL. **DA_FaBiAN_CRL** and **DA_FaBiAN_- FeTA** are respectively fine-tuned on *FaBiAN-CRL* and *FaBiAN-FeTA*.

All 31 target domain *FeTA-mial* images are used in a pure testing phase.

4.2.6 Evaluation and analysis

The performance of the networks is evaluated with the Dice similarity coefficient (DSC) [28] between the GT manual annotations and the predicted segmentation. In each experiment (data augmentation in Section 4.2.4, domain adaption in Section 4.2.4 and importance of simulation model in Section 4.2.5), a paired Wilcoxon rank-sum test

is performed between each experimental configuration and the experiment-specific "Baseline" network. For individual fetal brain tissues, *p*-values are adjusted for multiple comparisons using Bonferroni correction. Statistical significance level is set to 0.05.

4.3 Results

4.3.1 Data augmentation

Figure 4.4 shows for an axial view slice of a 30.6 weeks-old *FeTA-svrtk* patient the T2w image, and the GT, Baseline and SynthAug segmentation overlaid to the T2w images. SynthAug presents a substantially more accurate delineation of the tissues, especially for the cGM. The model trained with spatial and simulations based augmentations appears more sensitive to the cortical folding than Baseline network that is only trained on clinical images that were spatially and intensity-based augmented.

SR image   Ground truth   Baseline   SynthAug

█ CSF   █ cGM   █ WM   █ dGM
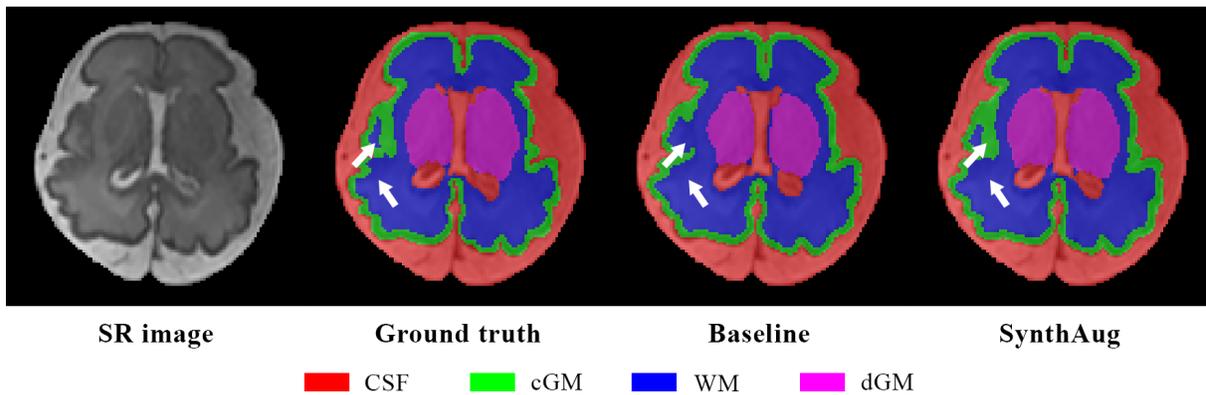
Figure 4.4: Appreciation of synthetic MRI-based data augmentation accuracy of fetal brain tissue segmentation in a patient of 30.6 weeks of GA on an axial slice.

Quantitative results are presented in Table 4.4. Overall, SynthAug results in a significantly improved mean DSC of 0.90 ± 0.05. Using the FaBiAN-based augmentation, all fetal brain structures present a higher DSC, with statistical significance for the CSF, the cGM, the cerebellum, the dGM and the brainstem. FaBiAN augmentation is an efficient data augmentation strategy that competes with classic intensity-based augmentation methods.

4.3.2 Domain adaptation

Figure 4.5 shows qualitative results for a *FeTA-mial* 31.2 weeks of GA fetus. Domain adaptation resulting from the addition of synthetic MR images reconstructed in the target domain (DA_FaBiAN) enhances the accuracy of segmentation compared to the

Table 4.4: Data augmentation quantitative results. DSC (mean ± SD) in the configurations studied (Baseline and SynthAug) for all segmented brain tissues and on average. The best scores between configurations are shown in bold. *p*-values of Wilcoxon rank sum test between configuration for individual fetal brain tissue segmentation are adjusted for multiple comparisons using Bonferroni correction. $p < 0.05$ (*) is considered statistically significant.

	Baseline	SynthAug
CSF	0.93 ± 0.01	**0.95 ± 0.02** (*)
cGM	0.77 ± 0.02	**0.84 ± 0.05** (*)
WM	0.92 ± 0.01	**0.93 ± 0.02**
Cerebellum	0.87 ± 0.06	**0.92 ± 0.04** (*)
dGM	0.85 ± 0.04	**0.90 ± 0.04** (*)
Brainstem	0.85 ± 0.03	**0.88 ± 0.03** (*)
Overall	0.86 ± 0.06	**0.90 ± 0.05** (*)

Baseline network, especially in the cGM, and the infratentorial structures, i.e. the brainstem and the cerebellum.

SR image	**Ground truth**	**Baseline**	**DA_FaBiAN_Baseline**	**Gold Standard**

■ CSF   ■ cGM   ■ WM   ■ Ventricles   ■ Cerebellum   ■ dGM   ■ Brainstem
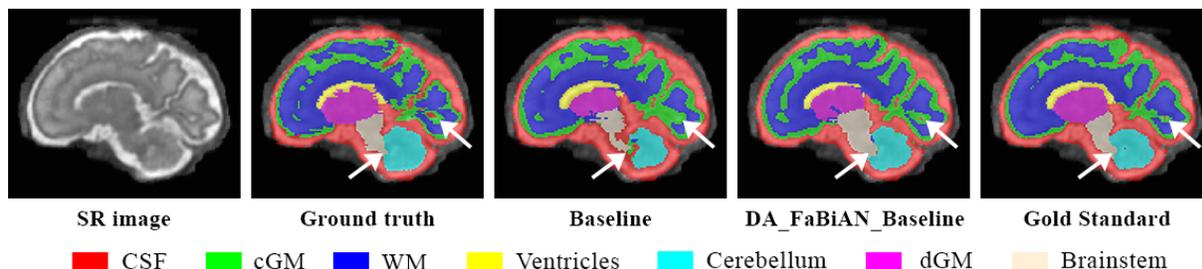
Figure 4.5: Appreciation of synthetic MRI-based domain adaptation accuracy of fetal brain tissue segmentation in a sagittal view of a patient of 31.2 weeks of GA. White arrows show representative areas where domain adaptation resulting from the addition of synthetic MR images in the target domain improves the segmentation accuracy.

These results are further supported by the quantitative results presented in Table 4.5 for every segmented fetal brain tissue in each configuration studied. Overall, the performance of the segmentation algorithm is significantly enhanced when fine-tuning the Baseline network with synthetic, yet realistic fetal brain MR images reconstructed in the target domain (DA_FaBiAN). The accuracy of the segmentation algorithm is slightly decreased in the ventricles only, but without statistical significance. This may be due to erroneous label propagation (after simulation and reconstruction) resulting in neighboring areas of similar intensities at the mid-sagittal plane either segmented as CSF or ventricles. Conversely, the segmentation of the cerebellum is even more accurate when complementing the Baseline dataset with synthetic images (DA_FaBiAN) than when training the network directly with the target *FeTA-mial* dataset (Gold Standard).

Interestingly, our domain adaptation strategy leads to a more accurate segmentation of the CSF, the cGM, the WM, the cerebellum, the dGM and the brainstem in pathological cases than the Baseline, as shown in Figure 4.6. The trend is different in neu-

rotypical patients, where DA_FaBiAN only results in an improved segmentation of the dGM, the brainstem and the cerebellum. The segmentation of the latter is even more accurate than when training the network on clinical data in the target domain (Gold Standard). However, the performance of the segmentation algorithm is stable for the CSF and the cGM, with a slight increase for the WM in the configuration DA_FaBiAN, whereas it decreases for the ventricles.

Table 4.5: Domain adaptation quantitative results. DSC (mean ± SD) of different training configurations. The best scores between domain adaptation (DA_FaBiAN) configurations and the Baseline are shown in bold. The corresponding p-values (paired Wilcoxon rank sum test) are adjusted for multiple comparisons using Bonferroni correction. $p < 0.05$ (*) is considered statistically significant.

	Experimental configurations		Gold Standard
	Baseline	DA_FaBiAN	
CSF	0.75 ± 0.32	**0.77 ± 0.33**	0.81 ± 0.30
cGM	0.57 ± 0.22	**0.64 ± 0.21** (*)	0.68 ± 0.18
WM	0.77 ± 0.22	**0.80 ± 0.21** (*)	0.86 ± 0.19
Ventricles	**0.79 ± 0.20**	0.76 ± 0.23	0.85 ± 0.17
Cerebellum	0.58 ± 0.30	**0.70 ± 0.35** (*)	0.65 ± 0.32
dGM	0.48 ± 0.20	**0.59 ± 0.22** (*)	0.82 ± 0.19
Brainstem	0.53 ± 0.28	**0.65 ± 0.25** (*)	0.76 ± 0.26
Overall	0.64 ± 0.23	**0.70 ± 0.24** (*)	0.77 ± 0.22
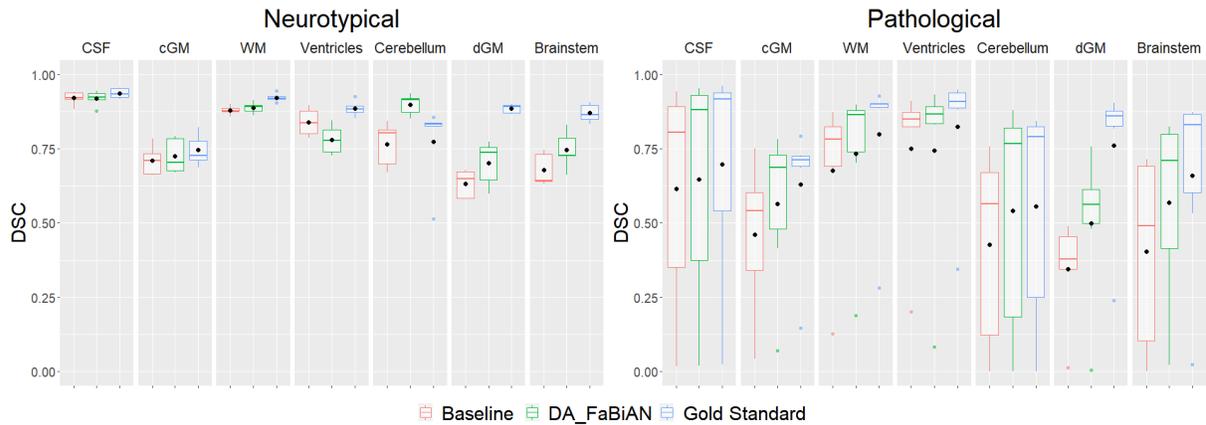
Figure 4.6: Domain adaptation label-wise configuration performances for neurotypical and pathological groups.

4.3.3 Simulation brain model

Figure 4.7 highlights areas of the fetal brain that are annotated in a significantly different way between *CRL* and *FeTA*, leading to two different models. Particularly, the cGM is thicker in the *CRL* model, with a more accurate annotation of the deep sulci.
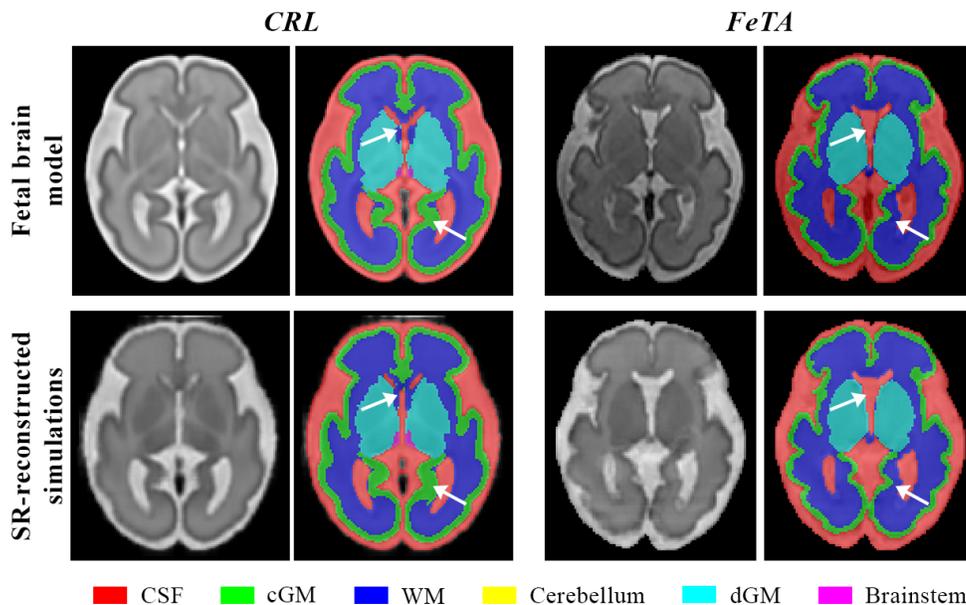
Figure 4.7: Comparison of fetal brain models and SR-reconstructed simulations. Representative T2-weighted MR images (top row) of the two models from which fetal brain MR acquisitions are derived, corresponding SR reconstructions (bottom row) and associated label maps. Left panel: CRL 30 weeks-old subjects. Right panel: *FeTA-svrtk* clinical neurotypical patient of 30.3 weeks of GA. The arrows point out discrepancies in annotations between both models of the fetal brain, mainly in the cGM, the dGM and the CSF.

Figure 4.8 shows an appreciation of the automatic segmentation methods compared to the GT on two representative slices of a 30.3 weeks-old fetus. DA_FaBiAN_FeTA qualitatively presents more accurate tissue delineation. In the DA_FaBiAN_CRL segmentation that was trained on CRL-based SR reconstructed simulation (Figure 4.7), we observe that the cGM is better captured in the sulci depth, although it seems less similar to the GR than DA_FaBiAN_FeTA .

Table 4.6 reports the mean DSC for each tissue in the different configurations studied. Overall, the performance of the segmentation algorithm is significantly improved when supplementing clinical data (Baseline) with synthetic MR images reconstructed in the target domain (DA_FaBiAN_X), especially with a fetal brain model close to the target domain, i.e. DA_FaBiAN_FeTA.

DA_FaBiAN_FeTA provides statistically more accurate segmentation than the Baseline in all the tissues but the CSF. Conversely, DA_FaBiAN_CRL is less accurate in the CSF, the cGM and the WM than the Baseline, whereas it outperforms all other configurations in the dGM. DA_FaBiAN_FeTA outperforms DA_FaBiAN_CRL in the cerebellum, the brainstem, the WM and the cGM.

Figure 4.9 displays the segmentation performance separated into neurotypical and pathological patients. Improved segmentation of WM and cGM is especially observed in pathological patients.

| SR image | Ground truth | Baseline | DA_FaBiAN_CRL | DA_FaBiAN_FeTA |

■ CSF   ■ cGM   ■ WM   ■ Cerebellum   ■ dGM   ■ Brainstem
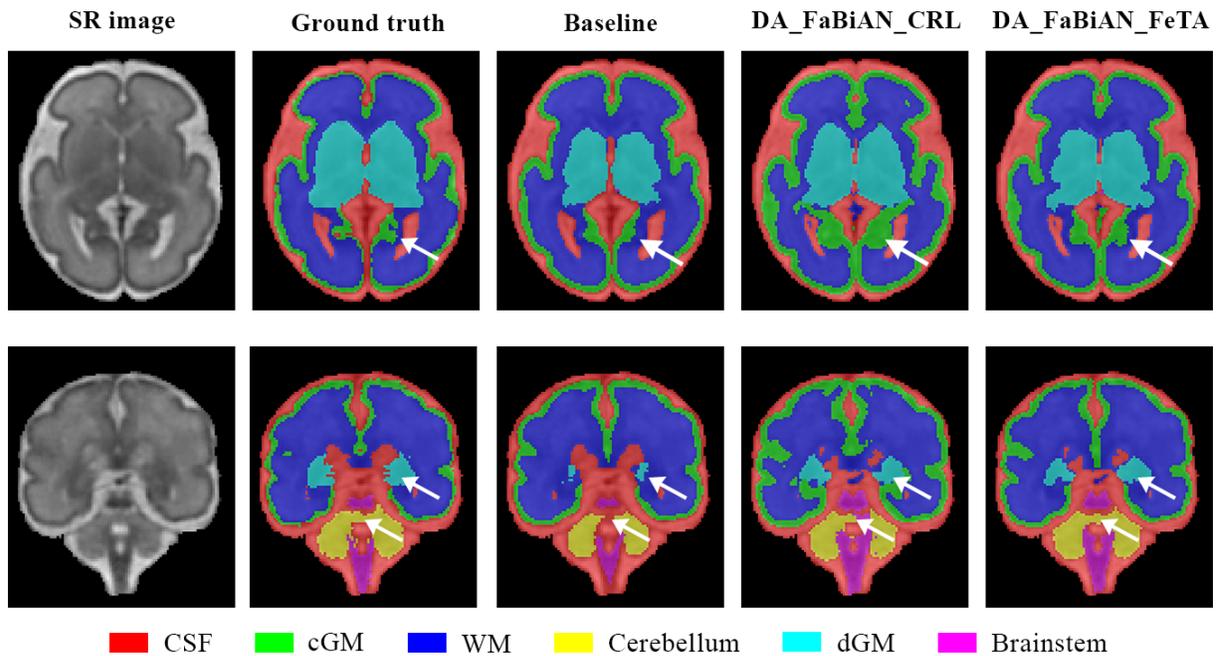
Figure 4.8: Qualitative results in simulation-based domain adaptation from different fetal brain models on an axial (top row) and a sagittal (bottom row) view for a neurotypical fetus of 30.3 weeks of GA. The arrows point out areas that are more accurately segmented in our domain adaptation strategy using *FeTA-svrtk* patients as fetal brain models instead of HR images from *CRL* [56] in comparison with the Baseline.

Table 4.6: Quantitative results in simulation-based domain adaptation from different fetal brain models. DSC (mean ± SD) in the different training configurations. The best scores between configurations are shown in bold. *p*-values of Wilcoxon rank sum test between configuration for individual fetal brain tissue segmentation are adjusted for multiple comparisons using Bonferroni correction. $p < 0.05$ (*) is considered statistically significant.

	Baseline	DA_FaBiAN_CRL	DA_FaBiAN_FeTA
CSF	**0.86 ± 0.17**	0.84 ± 0.18 (*)	**0.86** ± 0.17 (+)
cGM	0.66 ± 0.17	0.61 ± 0.16 (*)	**0.69 ± 0.15** (*,+)
WM	0.85 ± 0.12	0.79 ± 0.18 (*)	**0.86 ± 0.12** (*,+)
Cerebellum	0.61 ± 0.29	0.65 ± 0.30 (*)	**0.73 ± 0.30** (*,+)
dGM	0.65 ± 0.16	**0.74 ± 0.18** (*)	0.72 ± 0.17 (*,+)
Brain stem	0.57 ± 0.23	0.69 ± 0.16 (*)	**0.71 ± 0.18** (*,+)
Overall	0.70 ± 0.18	0.72 ± 0.18 (*)	**0.76 ± 0.17** (*,+)

4.4 Conclusion

In the context of domain adaptation for the tissue segmentation of fetal brain MRI, we propose an augmentation pipeline that relies on a simulation framework for synthetic MR images. We address the inter-dataset domain shift by replicating both the acquisition and postprocessing steps. Specifically, we demonstrated that numerically-
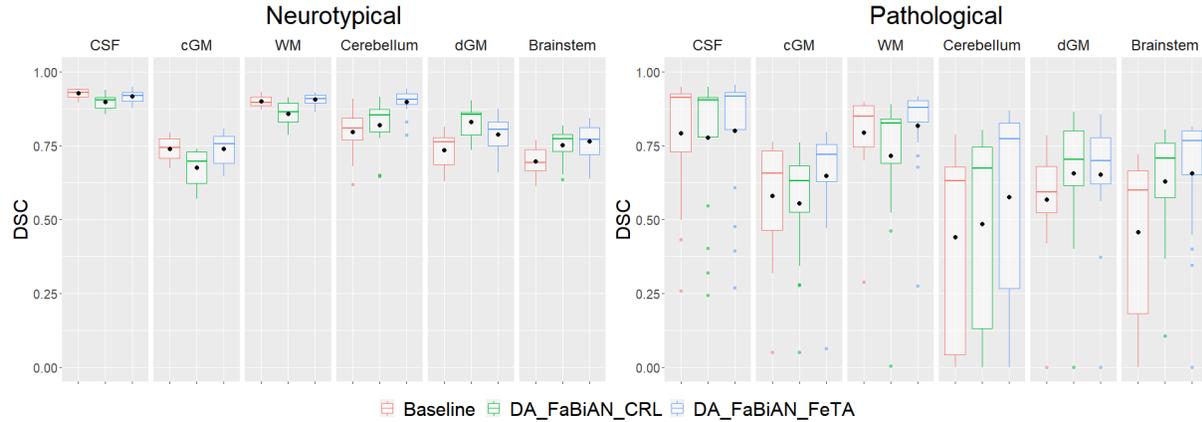
Figure 4.9: Label-wise performance comparison for simulation-based domain adaptation from different fetal brain models for neurotypical (left) and pathological (right) groups.

generated synthetic MR images of the developing fetal brain can complement clinical images in the training of automatic segmentation methods.

From there, we demonstrated the relevance of such synthetic MR images for the generation of target-like domain images. The latter allowed a significant reduction of the domain shift. Our approach that is free of additional expert annotations overcomes the lack of annotated clinical data by generating synthetic images from pre-existing labelmaps. Nonetheless, we evidence the importance of the choice of the latter as it appears to play a key role both from the annotation procedure followed and in the morphological variability presented to the network.

In this chapter, all experiments were conducted using 2D patch-based approaches. Although we use a state-of-the-art U-Net architecture, 3D-based methods are known to be better performing (see Chapter 1, Section 1.4.2). Nonetheless, we address here the DL limitation of the fitting of a specific data distribution and would expect a similar behaviour of our data augmentation-based adaption strategy in a 3D approach.

In the automatic segmentation methods benchmark presented in [112], the only non DL-based method – our MAS approach – appeared to be the less altered by variations in the image quality. We recall that MAS relies on the iterative optimization of an image similarity metric, and not on the fitting of a data distribution. Therefore, similarly to the image quality variation, we hypothesize that the variations induced by the image acquisition and reconstruction may be less troublesome for such MAS approaches. Even though we specifically address in this chapter the domain shift in DL segmentation methods, DL-free methods should be considered to further evaluate the performance and benefit of our approach.

5

SR reconstruction-based domain generalization

The work in this chapter is accepted for presentation at the IEEE International Symposium on Biomedical Imaging (ISBI) 2023 as **"Domain generalization in fetal brain MRI segmentation with multi- reconstruction augmentation"** with co-authors: P. de Dumast and M. Bach Cuadra.

5.1 Introduction

As introduced in Section 1.4.2 and further studied in Chapter 4, the post-processing variability may not only be introduced by the reconstruction method itself but also by its parameters settings. Specifically, the contribution weight α given to the regularization function in the image recovery step (see Equation 1.4) induces substantial image texture changes (see Figure 1.5 (C)). In Chapter 3, we additionally evidenced that, when the regularization weight is well chosen across methods, the inter-SR method reconstruction variability may be reduced, as the image similarity increases.

Hence, after introducing a domain adaptation strategy relying on acquisition domain variations in Chapter 4, we present here an in-SR domain data augmentation approach to increase the segmentation model generalizability. Specifically, our in-SR domain multi-reconstruction approach aims at taking advantage of multiple SR-reconstructed volumes of the same fetal brain obtained with different parameters, in order to increase the intensity variability represented in the training data. We show that our multi-reconstruction approach increases performance on in- and out- of SR target domain datasets.
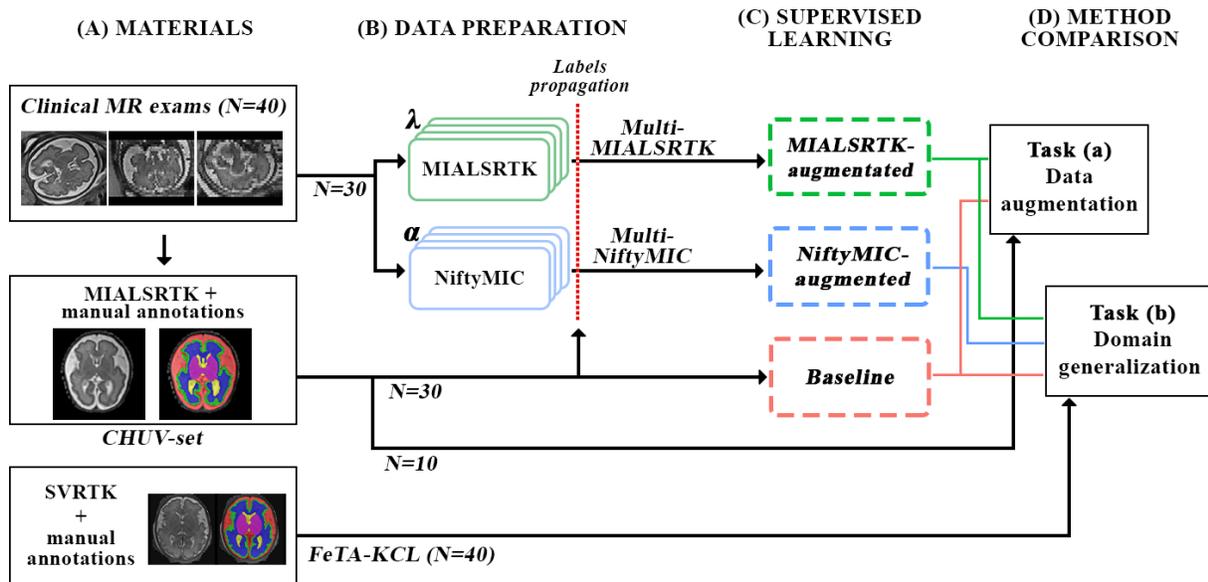
Figure 5.1: Overall framework. Materials (A), data preparation (B) and supervised learning (C) are presented for the data augmentation (a) and domain generalization (b) tasks.

5.2   Materials

5.2.1   Clinical MR exams

Forty (40) clinical fetal brain MR exams from 21 to 35 weeks of GA (mean ± standard deviation (SD): 28.4 ± 4.2) were retrospectively collected from the CHUV. Acquisitions were either performed at 1.5T (N = 37) or 3T (N = 3) on Siemens Healthcare scanners, respectively resulting to 1.125mm and 0.5469mm in-plane isotropic, and 3.3mm and 3mm through-plane resolution.

The local ethics committee of the Canton of Vaud, Switzerland (CER-VD 2021-00124) approved the retrospective collection and analysis of MRI data and the prospective studies for the collection and analysis of the MRI data in presence of a signed form of either general or specific consent.

5.2.2   In-SR domain dataset – *CHUV-set*

All clinical exams (presented in Section 5.2.1) were SR-reconstructed into the patient's space through the MIALSRTK pipeline (with default regularization weight, i.e. λ = 0.75) [143]. Volumes are aligned into a common reference space and resampled to $1.1 \times 1.1 \times 1.1 mm^3$. Following the FeTA annotation guidelines [112], the extra-axial

CSF, the cGM, the WM, the ventricles, the cerebellum, the dGM and the brainstem were manually annotated.

5.2.3 Out-of-SR domain pure testing set – *FeTA-KCL*

Forty (40) clinical fetal brain images from FeTA dataset are used as out-of-domain pure testing set [112]. Patients were aged from 21.2 to 34.8 weeks of GA (27.1 ± 3.9). They were reconstructed with SVRTK [88, 135]. SR volumes were resampled to an isotropic resolution of 0.8*mm*. Tissue annotations of the CSF, the cGM, the WM, the ventricles, the cerebellum, the dGM and the brainstem were manually refined and completed with the additional corpus callosum (CC) label [39, 40]. For the remainder of this study, CC and WM tissue classes are merged in order to match the tissue distribution available in *CHUV-set*.

5.3 Methodology

Our experiment design is shown in Figure 5.1 including datasets (panels A and B), the methods development (panel C) and their assessment (panel D). We propose a single-pipeline multi-reconstruction approach as data augmentation for fetal brain MRI segmentation. To strengthen the generalization of our findings, we assess our multi-reconstruction approach with two different SR pipeline, namely NiftyMIC [33] and MIALSRTK [143]. First, we assess our single-pipeline multi-reconstruction method in a pure data augmentation set up (Task (a), Section 5.3.2). Second, we further evaluate our augmentation approach in an out-of-domain experiment (Task (b), Section 5.3.2).

5.3.1 SR reconstruction-based data augmentation

SR reconstructions

All clinical MR exams in Section 5.2.1 are additionally reconstructed through two independent SR pipelines, such that we have the following two new independent HR *Multi-SR* datasets:

Multi-MIALSRTK fetal brains are reconstructed through the MIALSRTK [143] pipeline with the following regularization weight - including the default one - $\lambda \in \{0.1, 0.75, 1.5, 3.0\}$;

Multi-NiftyMIC fetal brains are reconstructed through the NiftyMIC [33] pipeline with the regularization weight $\alpha \in \{0.01, 0.02, 0.05, 0.1\}$.

Figure 5.2 shows a representative patch of a 32 weeks of GA fetal brain multi-reconstructed through MIALSRTK and NiftyMIC SR pipelines. High λ (respectively low α), offers a better tissue contrast, although the overall image appears more noisy. Conversely, low λ values (respectively high α) increase the overall smoothness of the image. Thus, variation of the regularization weights echoes a *texture* change in the SR image.
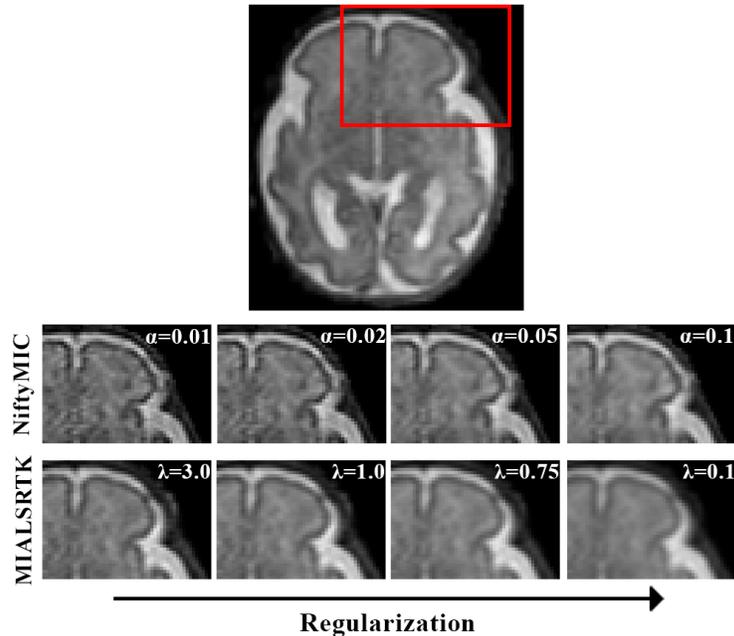
Figure 5.2: Appreciation of the intensity variation depending on the SR reconstruction regularization parameters λ and α for MIALSRTK and NiftyMIC reconstructions on a 32 weeks of GA neurotypical patient.

Weak labelling

All reconstructed brains of *Multi-MIALSRTK* and *Multi-NiftyMIC* have a tissue labelmap in the space of its *CHUV-set* reconstruction. Through a rigid registration-based approach, manual annotations are propagated to the newly generated *Multi-SR* sets.

Training configurations

We define three configurations based on their training data (Table 5.1, Figure 5.1 (C)). *Baseline* is trained on the 30 training patients from *CHUV-set*, and *MIALSRTK-augmented*, respectively *NiftyMIC-augmented* are trained on the 120 *Multi-MIALSRTK*, respectively *Multi-NiftyMIC*, reconstructed volumes of the same 30 fetal brains.

Table 5.1: Summary of training configurations.

	Baseline	*MIALSRTK-augmented*	*NiftyMIC-augmented*
Training set	*CHUV-set*	*Multi-MIALSRTK*	*Multi-NiftyMIC*
# patients	30	30	30
# volumes	30	120	120
Labelling	Manual	Weak	Weak

5.3.2 Evaluation

We compare our multi-reconstruction approach in:

(a) Data augmentation task. In this in-domain experiment, we compare the performances of *Baseline* and *MIALSRTK-augmented* on the 10 left-out patients from *CHUV-set*.

(b) Domain generalization task. We evaluate the performance of *Baseline*, *MIALSRTK-augmented* and *NiftyMIC-augmented* on the 40 out-of-domain *FeTA-KCL* images.

The performance of the *SR-augmented* and *Baseline* configurations are evaluated with the DSC [28] and the average symmetric surface distance (ASSD) [157] between the ground truth manual annotations and the predicted segmentation. A paired Wilcoxon rank-sum test is performed between *SR-augmented* configuration and the *Baseline*. p-values are adjusted for multiple comparisons using Bonferroni correction in the statistical analysis of individual fetal brain tissues. Statistical significance level is set to 0.05.

5.3.3 Model and training strategy

From the MONAI framework [103, 17], we use the popular 3D U-Net architecture that performed remarkably on fetal brain MRI tissue segmentation in the 2021 MICCAI FeTA challenge [113]. Inputs are $96 \times 96 \times 96$ voxel size patches that are randomly sampled from the original SR volumes. At train time, input samples are augmented, based on the most applied and successful transformations used in the FeTA challenge [113]. Spatial (flipping, rotation, resampling) and intensity-based (bias field, Gaussian noise) transformations are randomly applied. Lastly, image patch intensities are normalized. We adopt a 5-folds CV. Networks are trained for 100 epochs minimizing a dice focal loss function, where both losses equally contribute. All configurations adopt the same training strategy.

At test-time, we proceed to an ensemble evaluation of all 5 CV networks on 50% overlapping patches inferred through a sliding window approach.

5.4　Results

5.4.1　Data augmentation

Table 5.2 (a) reports the DSC and ASSD performances for both tasks. Overall, the performance of the segmentation algorithm is significantly enhanced when each fetal brain is reconstructed multiple times through the same pipeline as the target (testing) images, even though the weak labelling process incurred. The benefit of *MIALSRTK-augmented* is further statistically significant for all tissue classes.

Table 5.2: Segmentation performances of the different training configurations in data augmentation (a) and domain generalization (b) tasks. DSC and ASSD (mean ± SD) of the different training configurations in data augmentation (a) and domain generalization (b) tasks. The best scores between *SR-augmented* configurations and *Baseline* are shown in bold. Arrows indicates weither the metric is better maximized (↑) or minimized (↓). The corresponding *p*-values (paired Wilcoxon rank sum test) are adjusted for multiple comparisons using Bonferroni correction. Statistical significance (*) is $p < 0.05$.

		(a) Data augmentation		(b) Domain generalization		
		Baseline	*MIALSRTK-augmented*	*Baseline*	*MIALSRTK-augmented*	*NiftyMIC-augmented*
DSC (↑)	CSF	0.87 ± 0.02	**0.90 ± 0.02** (*)	0.87 ± 0.17	0.90 ± 0.16 (*)	**0.90 ± 0.17** (*)
	cGM	0.75 ± 0.03	**0.78 ± 0.03** (*)	0.76 ± 0.14	**0.81 ± 0.14** (*)	0.81 ± 0.14 (*)
	WM	0.86 ± 0.02	**0.89 ± 0.01** (*)	0.85 ± 0.13	**0.89 ± 0.12** (*)	0.88 ± 0.12 (*)
	Ventricles	0.81 ± 0.05	**0.84 ± 0.04** (*)	0.76 ± 0.11	0.82 ± 0.10 (*)	**0.86 ± 0.10** (*)
	Cerebellum	0.88 ± 0.02	**0.91 ± 0.03** (*)	0.67 ± 0.28	0.79 ± 0.21 (*)	**0.86 ± 0.11** (*)
	dGM	0.88 ± 0.05	**0.89 ± 0.03** (*)	0.79 ± 0.11	0.85 ± 0.08 (*)	**0.86 ± 0.06** (*)
	Brainstem	0.85 ± 0.02	**0.87 ± 0.02** (*)	0.68 ± 0.20	0.76 ± 0.13 (*)	**0.76 ± 0.10**
	Overall	0.84 ± 0.02	**0.87 ± 0.01** (*)	0.77 ± 0.14	0.83 ± 0.11 (*)	**0.85 ± 0.10** (*)
ASSD (↓)	CSF	0.47 ± 0.06	0.35 ± 0.05 (*)	1.0 ± 2.41	0.91 ± 2.5 (*)	0.870 ± 2.45 (*)
	cGM	0.63 ± 0.09	0.55 ± 0.09 (*)	1.0 ± 0.57	0.96± 1.4 (*)	0.754 ± 0.761 (*)
	WM	0.81 ± 0.1	0.64 ± 0.09 (*)	1.5 ± 1.5	1.3 ± 2.1 (*)	1.40 ± 2.23 (*)
	Ventricles	0.88 ± 0.2	0.57 ± 0.09 (*)	1.5 ± 0.70	0.96± 0.521 (*)	0.797 ± 0.505 (*)
	Cerebellum	0.70 ± 0.06	0.56 ± 0.07 (*)	2.5 ± 3.0	1.2 ± 0.998 (*)	0.886 ± 0.508 (*)
	dGM	0.75 ± 0.2	0.66 ± 0.1 (*)	1.8 ± 1.4	1.4 ± 0.951 (*)	1.22 ± 0.729 (*)
	Brainstem	0.70 ± 0.07	0.58 ± 0.07 (*)	2.1 ± 2.8	1.2 ± 0.580	1.27 ± 0.684
	Overall	0.70 ± 0.07	**0.56 ± 0.05** (*)	1.64 ± 1.32	1.14 ± 0.94 (*)	**1.03 ± 0.80** (*)

5.4.2　Domain generalization

SR-augmented qualitatively enhance the accuracy of segmentation compared to *Baseline*, especially in the infratentorial structures, namely the brainstem, the cerebellum and the 4th ventricle (Figure 5.3, white circles). Additionally, the WM tract of the CC are better captured with our multi-reconstruction approach, and even more than compared to the ground truth.

Quantitative results (Table 5.2 (b)) show in the assessment of 10 in-domain to 40 out-of-domain patients a loss of performances of both *Baseline* and *MIALSRTK-augmented* configurations. Nevertheless, while *Baseline* shows a loss of 0.07 and 0.94, respectively in DSC and ASSD *MIALSRTK-augmented* only drops of 0.04 and 0.58. Our multi-
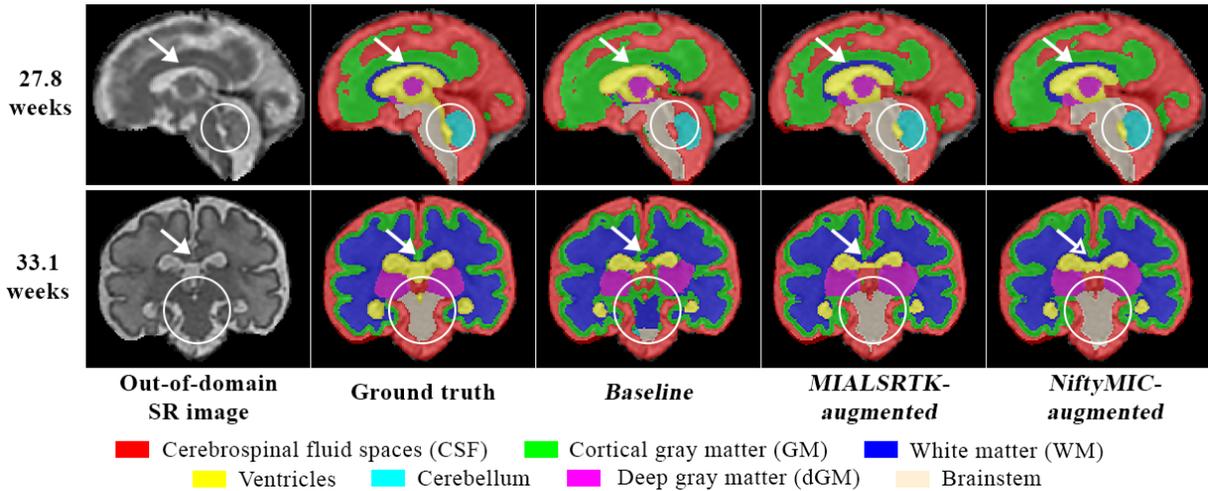
Figure 5.3: Sagittal view of a 27.8 weeks-old (top) and coronal view of the 33.1 weeks-old (bottom) fetal brain tissue segmentation obtained in the different configurations studied. White arrows and circles show representative areas where our multi-reconstruction approach improves the segmentation accuracy.

reconstruction approach hence seems more robust to the inter-SR method domain gap. Regardless of the SR method of the training data, our multi-reconstruction augmentation strategy significantly improves the segmentation performance on out-of-SRR domain images, both in DSC and ASSD. On a tissue-wise analysis, *SR-augmented* configurations are always significantly better performing than *Baseline*, except for the ASSD in the brainstem. The benefit of our multi-reconstruction approach is even more pronounced in the ventricles, the cerebellum, the dGM and the brainstem where the gain in DSC is greater than 0.05. Specifically, the steepest improvement appears in the cerebellum (0.67 for the *Baseline* vs. 0.79 and 0.86, respectively for *MIALSRTK-* and *NiftyMIC-augmented*).
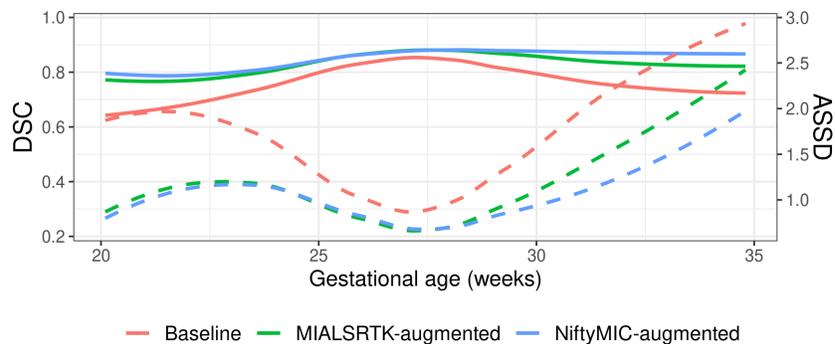
Figure 5.4: Mean DSC (plain) and ASSD (dashed) performance in the domain generalization task as a function of GA in weeks.

From a GA-based analysis (Figure 5.4), we observe that *Baseline* overall performs worse on young and old (< 25 and > 30 weeks of GA) fetuses. On the contrary, al-

though a similar trend is noticeable for *SR-augmented* methods, it is substantially less pronounced. Consequently, in this inter-domain segmentation task, *SR-augmented* configurations seems to better benefit the endpoint of the GA range studied.

5.5   Conclusion

In this chapter, we have demonstrated that having single-pipeline multi-reconstruction of fetal brain MR exams *(i)* is an efficient intensity-based data augmentation strategy and *(ii)* reduces the performance drop in target image domain shift in segmentation task. Our multi-reconstruction approach, combined to conventional data augmentation strategies, increases the representation of fetal brain MRI variability in the training phase of supervised segmentation method. Although we did not investigate multi-pipeline multi-reconstruction augmentation, one can expect an even stronger benefit of our method. Nonetheless, a proper validation of our method on larger and more diverse datasets remains to be explored as well as a comparison against state-of-the-art augmentation methods. In its batch processing approach, our multi-reconstruction strategy is an out-of-the-box easy to adapt method. Future work will investigate this multi-reconstruction augmentation at inference in order to increase the prediction robustness, as test-time augmentation had been proposed in [69].

Part III

On segmentation consistency

6

Importance of topology assessment

The work presented in this chapter is an analysis done in the context of the FeTA 2022 MICCAI challenge. It will be included in the paper in preparation for submission to **IEEE Transactions on Medical Imaging** entitled **"Multi-Center Fetal Brain Tissue Annotation and Segmentation (FeTA) Challenge Results"** with co-authors K. Payette, C. Steger, R. Licandro, P. de Dumast, H. Li, M. Barkovich, L. Li, M. Dannecker, C. Chen, C. Ouyang, N. McConnell, A. Miron, Y. Li, A. Uus, I. Grigorescu, P. Ramirez Gilliland, Md. Mahfuzur R. Siddiquee, D. Xu, A. Myronenko, H. Wang, Z. Huang, J. Ye, M. Alenyà, V. Comte, O. Camara, J.-B. Masson, A. Nilsson, C. Godard, M. Mazher, A. Qayyum, Y. Gao, H. Zhou, S. Gao, J. Fu, G. Dong, G. Wang, Z. Rieu, H. Yang, M. Lee, S. Potka, M K. Grzeszczyk, A. Sitek, L. Vargas Daza, S. Usma, P. Arbelaez, W. Lu, W. Zhang, J. Liang, R. Valabregue, A.A. Joshi, K. N. Nayak, R.M. Leahy, L. Wilhelmi, A. Dändliker, Hui Ji, A. Jakovi, M. Klai, A. Adi, P. Markovi, G. Grabari, M. Rados, L. Vasung, M. Bach Cuadra, A. Jakab.

6.1 Introduction

The FeTA MICCAI challenge[1] aims at encouraging the development of automatic fetal brain MRI tissue segmentation methods. The task the participants have to solve is the automatic segmentation of the fetal brain intracranial tissues into seven categories: CSF, cGM, WM, Ventricles, Cerebellum, dGM and Brainstem. A training dataset is provided to the participants and their submitted algorithms are further assessed on non-released testing data. To identify the better performing algorithms amongst the submissions, complementary metrics are inferred, from which a ranking of the methods is derived [112]. Typically, overlap-, boundary-distance- and size-based metrics are considered. Nonetheless, the topology correctness is usually not considered, even though the morphometry analysis heavily relies on it.

---

[1] *https://feta.grand-challenge.org/*

In this chapter, we propose to further analyze the importance of the topology in the comparison of automatic segmentation methods. Specifically, we propose the integration of a global topology metric in the finale ranking of the 2022 FeTA challenge submissions.

6.2 Materials

6.2.1 Challenge datasets

Four institutions provided data for the 2022 FeTA challenge: University Childrens Hospital Zurich, Zurich, Switzerland (*Kispi*), General Hospital Vienna/Medical University of Vienna, Vienna, Austria (*Vienna*), Lausanne University Hospital, Lausanne, Switzerland (*CHUV*) and University of California San Francisco, CA, USA (*UCSF*).

Table 6.1 presents a summary of the acquisition and postprocessing steps of the images per institution. Specifically regarding the SR reconstruction methods: *Vienna* and *UCSF* use NiftyMIC [33], half of *Kispi* use SVRTK [88] and, finally, *CHUV* and the rest of *Kispi* use MIALSRTK [143] although an additional histogram matching is done in *Kispi*. *Kispi* and *Vienna* provided both training and testing data, while *CHUV* and *UCSF* only provided testing images.

		Kispi	Vienna	CHUV	UCSF
Cohort	# Patients	120	80	40	40
	GA (weeks)	[20-35]			
Acquisition	Field strength	1.5 or 3 T	1.5T	1.5 or 3T	3T
	MR vendor	GE	Philips	Siemens	GE
	Resolution (mm^3)	$0.5 \times 0.5 \times 3$ to 5	$- \times - \times 3$ to 5	$1.13 \times 1.13 \times 3.3$ $0.55 \times 0.55 \times 3$	$\sim 0.5 \times \sim 0.5 \times 3$
Post-processing	SR reconstruction	MIALSRTK [143] SVRTK [88]	NiftyMIC [33]	MIALSRTK [143]	NiftyMIC [33]
	Additional	Atlas alignment Histogram matching	Atlas alignment	Atlas alignment	Atlas alignment
	Resolution (mm^3)	$0.5 \times 0.5 \times 0.5$	$1.0 \times 1.0 \times 1.0$	$1.13 \times 1.13 \times 1.13$	$0.8 \times 0.8 \times 0.8$
Splitting	Train	80	40	-	-
	Test	40	40	40	40

Table 6.1: FeTA 2022 challenge datasets summary.

6.2.2 Challenge results

Sixteen (16) algorithms were submitted to the 2022 edition of the FeTA challenge. Prediction labelmaps were inferred on the 160 testing patients through all submitted algorithms. Similarly to the 1st FeTA challenge edition [113], three complementary metrics

are assessed: the Dice similarity coefficient (DSC), the 95th percentile Hausdorff Coefficient (HD95) and the volume similarity (VS).

Rankings for each metric (DSC, HD95 and VS) over all segmented labels were performed through the challengeR toolkit, specifically designed to calculate and display imaging challenge results [151]. The final FeTA2022 challenge ranking is derived from all metric-wise rankings.

6.3 Methods

6.3.1 Computational topology

Topology defines the properties of an object that are preserved through deformation [123]. In computational topology, local features are derived to conclude on the global properties of an object. Specifically, in a binary image, local information relies on the connectivity of a voxel to its neighbors in an objects. Global features such as the number of connected components or holes can be drawn by generalizing the connectivity information.

Topology structures are defined in the different topological dimensions. The number of topological structures in each k-dimension is counted with the k-dimensional Betti numbers (BNk). In the context of 3D binary images, one can count BN0 connected components, BN1 holes (also called handles or tunnels), and finally BN2 cavities. Figure 6.1 shows three 2D binary image patches and their Betti numbers. Note that for a 2D binary image, BN2 is always 0.

Patch 1	**Patch 2**	**Patch 3**

| BN0=1; BN1=3; BN2=0 | BN0=1; BN1=4; BN2=0 | BN0=1; BN1=3; BN2=0 |
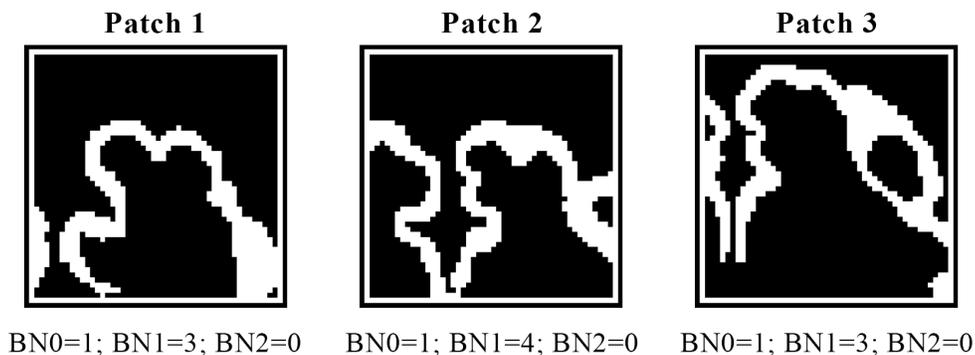
Figure 6.1: Example patches of 2D binary images with the topological properties: number of connected components BN0, number of holes BN1, and number of cavities BN2.

6.3.2 Topology metric

To quantitatively compare the topology of each segmented structure, we introduce the k-dimensional Betti number error (BNEk) as the absolute difference of the GT ex-

pected value and the prediction measure. BNE*k* are difference metrics that must be minimized. A summary of the GT expected BN is presented in Table 6.2.

	BN0	BN1	BN2
CSF	1	0	0
GM	2	0	0
WM	1	0	0
Ventricles	1	0	0
Cerebellum	1	0	0
dGM	1	0	0
Brainstem	1	0	0

Table 6.2: Ground truth *k*-dimensional Betti numbers (BN*k*) per tissue class.

In the absence of segmentation for a tissue in a labelmap predictions, it is attributed twice the value of the worst performing segmentation of the same label over all submissions.

Similarly than in Section 6.2.2, a ranking for each BNE*k* is computed through the challengeR toolkit [151]. We infer a global topology ranking BNE as the ranking of the sum of all three BNE*k* rankings.

6.3.3   Analysis

Algorithms ranking

We first analyze the topology rankings: each BNE*k* and the derived global BNE ranking. Then, we propose a topology-integrative ranking of the methods. Specifically, while the FeTA2022 ranking considers the DSC, the HD95 and the VS, our Topo-aware will additionally consider the global BNE topology ranking.

We further investigate the global BNE ranking in each the seven tissue labels.

Label-wise analysis

Over all submissions, we aim at exploring the topological correctness on a tissue-wise basis. In each of the *k*-dim BNE, we compute the average of the 95th percentile per tissue and algorithm. In this analysis, we discard the missing tissue segmentation that were penalized (see Section 6.3.2).

6.4 Results

6.4.1 Algorithms rankings

Overall

Table 6.3 (a) presents the topology rankings of the submissions for each dimension $k \in \{0, 1, 2\}$ (BNEk) and global (BNE). We observe that the methods seem to rank similarly across the three k-dimensional BNE, with a maximum rank difference of less than three, despite some exceptions. *FMRSK* presents a relatively big delta in its BNE rankings of dimension 1 (rank=4) and 2 (rank=13). We hypothesize that such inter-dimension variation may come from tissue-specific errors. Interestingly, *hilab* that does not perform well in BNE0 (rank=10) and BNE1 (rank=11) is the best performing submission in BNE2. Nonetheless, the good BNE2 performance is not sufficient to pass on to the global BNE ranking.

Changes in the global ranking (see Table 6.3 (b) Topo-aware vs. FeTA2022) are small: maximum one rank difference except for *Blackbean* that goes from rank 5 in FeTA2022 to rank 3 in Topo-aware thanks to its good and steady topology-related performances (ranks 3, 3 and 4). The winner and second submissions remain the same.

	(a) Topology				(b) Global	
	k-dim BNE			BNE	Topo-aware	FeTA2022
	BNE0	BNE1	BNE2			
ajoshiusc	16	15	16	16	17	17
Blackbean	3	3	4	3	**3**	5
BlueBrune	2	2	3	2	**2**	**2**
deepsynth	17	16	17	17	16	16
dolphines	5	7	7	5	8	9
FIT (nnUnet)	1	1	2	1	**1**	**1**
FIT (Transformers)	7	8	5	6	7	7
FMRSK	9	4	13	9	4	**3**
fudan_zmic	8	10	9	10	10	10
hilab	10	11	1	8	11	11
Institut_Pasteur_DBC	11	12	11	12	9	8
Neurophet	14	14	15	14	12	12
NVAUTO	6	6	8	7	5	4
Sano	13	13	12	13	13	13
symsense	4	5	6	4	6	6
Uniandes	12	9	10	11	14	14
xinlab-scut-iai-ahu	15	17	14	15	15	15

Table 6.3: Topology (a) and global (b) rankings of the submissions. (a) Betti number errors (BNE) per dimension and overall. (b) Comparison of the FeTA challenge ranking (FeTA2022) and our topology-integrative ranking (Topo-aware). Top 3 submissions are shown in bold.

Label-wise ranking

Table 6.4 presents the global topology BNE ranking of the submissions per tissue class. While overall BNE*k* rankings (Table 6.3 (a)) remained steady, the inter-tissue topology ranking are relatively inconsistent. For instance, *hilab* that is the better performing submission in the CSF ranks 13 out of 17 in the WM.

Apart from the unquestionable top 2 teams, *FIT (nnUnet)* and *BlueBrune*, only *Blackbean* and *dolphines* manage to rank in the upper part of the table for all tissue class. Specifically, the average tissue topology ranking of *Blackbean* is 3.3, while *FMRSK* ranks on average 9.1. Consequently, *Blackbean* per-tissue steadiness is rewarded with its appearance on the podium with our topology-integrative ranking (see Table 6.3 (b)) at the expense of *FMRSK*.

	Tissues							
	CSF	cGM	WM	Ventricles	Cerebellum	dGM	Brainstem	Average
ajoshiusc	10	15	5	17	16	17	14	13.4
Blackbean	4	4	1	2	4	6	2	3.3
BlueBrune	5	7	2	3	1	1	3	3.1
deepsynth	14	16	7	15	17	16	17	14.6
dolphines	7	8	3	8	2	5	6	5.6
FIT (nnUnet)	2	5	4	1	3	2	1	2.6
FIT (Transformers)	3	2	6	10	8	7	5	5.8
FMRSK	12	10	9	9	6	8	10	9.1
fudan_zmic	9	11	8	11	9	9	8	9.3
hilab	1	6	13	5	11	12	12	8.6
Institut_Pasteur_DBC	15	13	11	12	7	10	11	11.3
Neurophet	16	12	14	14	15	14	16	14.4
NVAUTO	11	1	12	6	5	3	7	6.4
Sano	13	14	15	13	12	13	9	12.7
symsense	8	9	10	4	10	4	4	7.0
Uniandes	6	3	16	7	14	11	15	10.3
xinlab-scut-iai-ahu	17	17	17	16	13	15	13	15.4

Table 6.4: Topology (BNE) ranking of the submissions per tissue class and on average.

6.4.2   Label-wise topology

Table 6.5 presents the average of the 95th percentile *k*-dimensional BNEs. We first observe that the least well segmented structure varies across dimensions: the CSF for BNE0 and BNE2, and the cGM for BNE1. Regardless of the dimension, the CSF, the cGM and the WM are always the worst 3 segmented structures. In dimension 1, we observe the greatest variation across structures (from 0.6 in the brainstem to 136.4 in the cGM), and the worst two structures are by far the cGM and the CSF. Both CSF and cGM are thin structures that are only a few voxels-wide in SR images, and are therefore prone to 1-dimensional topological errors.

	BNE0	BNE1	BNE2
CSF	41.8	106.0	0.8
cGM	23.1	136.4	0.3
WM	37.7	28.0	0.5
Ventricles	5.6	3.8	0.2
Cerebellum	14.5	0.8	0.1
dGM	5.7	1.4	0.1
Brainstem	5.4	0.6	0.1

Table 6.5: Average of the 95th percentile k-dimensional BNE per tissue class. Worst scores are shown in red.

Figure 6.2 presents 95th percentile of BNE1 of the submissions prediction per tissue and institution. The CSF, the cGM and the WM are undeniably and regardless of the dataset the worst segmented structures,

In the cGM, we observe a strong variation of BNE1 across the subsets: the worst results being obtained on *CHUV* images, followed by *Vienna*. We hypothesize this performance drop is due to the image resolution: *CHUV* and *Vienna* images have respectively isotropic resolutions of $1.13mm^3$ and $1mm^3$, while other others have lower resolutions ($0.5mm^3$ for *Kispi* and $0.8mm^3$ for *Vienna*).

The thickness of the cGM increases from 1 to $2mm$ for fetuses between 21 and 40 weeks of GA [147]), Consequently, the low image resolution ($> 1mm$ isotropic) and the strong partial volume effect hamper the accurate segmentation of this thin structure.

6.5 Conclusion

In this Chapter, we evidence the importance of the consideration of the topology in the assessment and comparison of automatic segmentation methods. In the algorithms ranking, the inclusion of a topology-based metric does not drastically change the final results, although minor updates are observed. Specifically, the across-tissue steadiness in topological accuracy is rewarded. Furthermore, over all submissions, we observe that the CSF, the cGM and the WM are always the most affected structures by topological defects, especially the cGM. As we relate the image resolution - and hence partial volume effect - to the segmentation accuracy, we evidence that topology metric must be considered, yet with care, depending on the data properties and structure of interest.
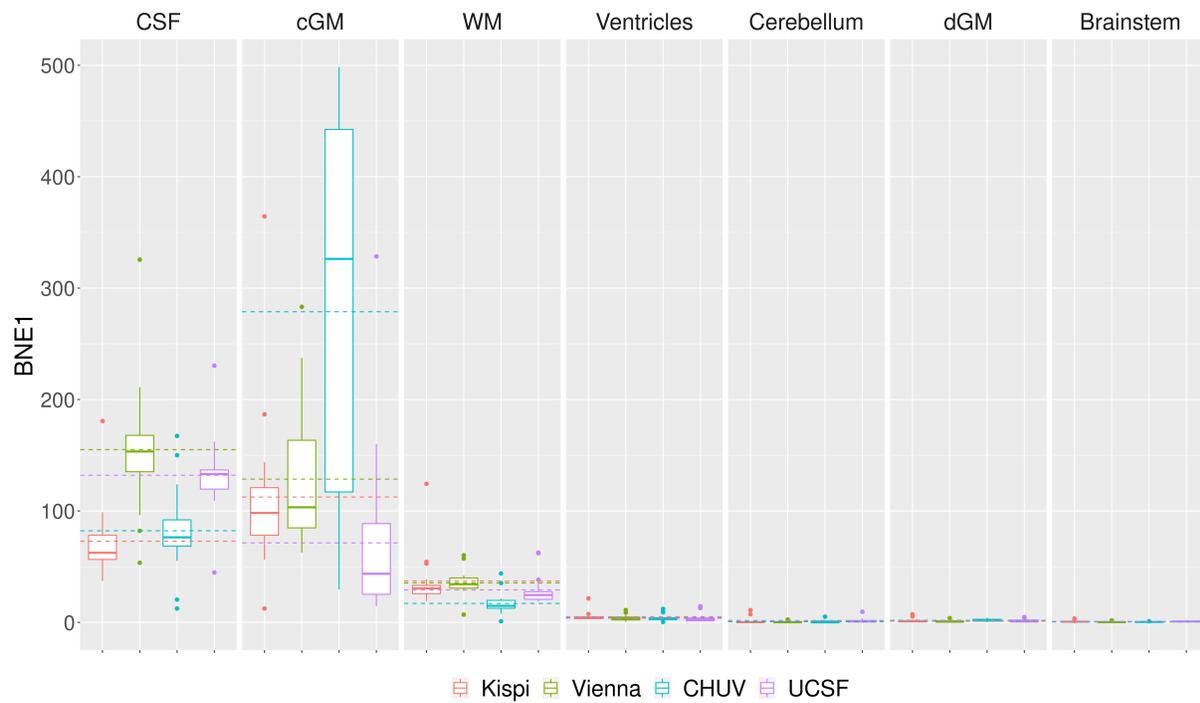
Figure 6.2: Average algorithm BNE1 performances per Institution for each tissue class.

7

Multi-dimensional topological loss

The work presented in this chapter is in preparation for submission to **NeuroImage** as **"Multi-dimensional topological loss for cortical plate segmentation in fetal brain MRI"** with co-authors P. de Dumast, H. Kebiri, V. Dunet, M. Koob and M. Bach Cuadra.

7.1 Introduction

7.1.1 Clinical context

During *in utero* development, the human fetal cortical plate (CP) that is the future cortex undergoes drastic changes (see Chapter 1, Section 1.1.1), in particular with the appearance of the cortical gyrifications (see Chapter 1, Figure 1.5 (A)). Jointly, the surface area and the volume of the future cortex are respectively increased 50 and 40 times during the 2nd and 3rd trimesters of pregnancy [147, 148].

Cortical gyrification is considered to be a relevant marker of fetal brain maturation, as the chronological sequence of appearance of sulci is well known during the fetal period [49] (see Chapter 1, Section 1.1.1). Conversely, abnormal fetal sulcation and gyration indicate disruption of one of the three main fetal stages of normal cortical development (i.e. cell proliferation, migration and cortical organization) [8, 92].

Altered cerebral cortex formation, induced either by genetic mutations, vascular injuries, metabolic diseases, fetal infection or teratogenic causes, may lead to malformations of cortical development. An updated classification of this group of heterogenic disorders has been recently published in a consensus statement [131]. Those rare disorders usually manifest with developmental delay, seizures, and motor and sensory deficits [94]. Given the consequences of abnormal brain gyrification, early diagnosis is crucial, for which the analysis of cortical maturation is an asset.

Quantitative analysis of imaging biomarkers has defined cortical development for the typically developing brain [120, 22, 152, 153], while other works evidenced discrepancies in cortical volumes and sulcal patterns in the pathological brain [21, 34, 75, 138]. Nevertheless, these analyses require prior additional segmentation processing steps that must be efficiently automatized to be continuously performed (see Chapter 1, Section 1.4.2). Nonetheless, in contrast to adult brain segmentation, fetal brain segmentation remains challenging as to provide an age-invariant method [98].

7.1.2   Related works in cortical plate segmentation

Segmentation of the CP is particularly challenging as it undergoes significant changes throughout gestation due to brain growth and maturation, respectively modifying the morphology and the image contrast (see Chapter 1, Figure 1.5 (A)). Furthermore, the cortex being a thin structure (from 1 to 2 mm for fetuses between 21 and 40 weeks of GA [147]) that is strongly affected by partial volume effects, anatomical topology is prone to be incorrectly captured by automatic segmentation methods (see Chapter 6).

In Caldairou *et al.* [15], they introduced the first topological-based segmentation of the fetal cortex based on geometrical constraints along with anatomical and topological priors. However, the sample size in that study was small (i.e. six fetuses) and topological correctness was not evaluated for the provided segmentations. More recently, DL methods have also focused on fetal brain MRI cGM segmentation. Based on a neonatal segmentation framework, a recent study introduced a hybrid segmentation process that minimizes the need for human interaction in the segmentation of the developing cortex [38]. Also with DL models, a multiple-predictions approach with a test-time augmentation to improve the robustness of the method was suggested [69]. Finally, a two-stage segmentation framework with an attention refinement was proposed [30]. Nevertheless, while the segmentation accuracy of these recent DL-based methods is promising, none of these CP segmentation works includes topological constraints nor assesses the topology. Overall, these studies of automatic segmentation frameworks report high similarity, as for overlap (DSC of 0.87 in [30] and 0.90 in [69]), and low difference, as for the boundary distance-based metrics (ASSD of 0.28 *mm* in [30] and mean surface distance of 0.18 *mm* in [69]), compared to the GT segmentation, but illustrated results show a lack of topological consistency with notably discontinuous/broken cortical ribbons (see Fig. 5 in [30], Figures 5 and 6 in [38]).

Here, we propose to integrate a topological constraint in a deep image segmentation framework to overcome the limitation of disjoint CP segmentation in fetal MRI. To our knowledge, only two works previously explored the topological fidelity of the semantic medical image segmentation with DL. First, Hu et *al.* [71] proposed to use a topological loss for neuronal membrane segmentation. Second, a study presented topological constraints for MR cardiac image segmentation [23] based on prior topological knowledge, such as the number of connected components or handles present in the structure of interest. Although theoretical CP topological features are known, such

prior information could only be valid in a whole 3D volume analysis. Therefore, we inspired from [71] that does not share this limitation to build a prior-free framework.

7.1.3 Contributions

In this chapter, we incorporate topological constraints and assess the topology of the CP segmentation in fetal brain MRI. To this end, our first contribution is the generalization of a topological loss function based on persistent homology into a multi-dimensional (in topological spaces) formulation. Using the public *FeTA* dataset [112] along with the spatiotemporal *CRL* atlas [56], we compare our topology-integrative optimization method (*TopoCP*) to the original topological loss function we built-on (*Hu2019*) [71] and other widely used loss functions (*Baseline* and *Hybrid*) [113] with a state-of-the-art U-Net segmentation network. We further assess our proposed method compared to semi-automatic manual annotations. For the first time, the topological correctness of the segmentation is assessed in the evaluation of automatic CP segmentation. In that respect, our second contribution is a new topology-based metric for the quantification of the CP segmentation defects not only in number but also in size. We quantitatively compare automatic and semi-automatic segmentation with complementary metrics on two independent pure testing sets. As a complement, three fetal brain MRI experts further visually compare automatic segmentation on an additional out-of-distribution clinical dataset. The results evidence an overall significant improvement in the segmentation using our proposed topological loss function.

In Section 7.2 we introduce our multi-dimensional topological loss function, the overall segmentation framework and our original topology hole ratio assessment metrics; in Section 7.3, we describe the experimental design, along with materials and description of the experiments performed; in Section 7.4, we describe and discuss the results; and finally, we conclude on this work in Section 7.5.

7.2 Methodology

We propose a topologically-guided DL framework for the CP segmentation of the fetal brain MRI. This is done by including a topological constraint in the optimization of state-of-the-art DL image segmentation strategies. First, we will introduce our custom loss function that we adapted for generalization from Hue *et al.* [71] (see Section 7.2.1). Second, we present the segmentation framework in which our custom loss function is integrated (see Section 7.2.2). Finally, we present topology-based metrics for further assessment of our method (see Section 7.2.3).

7.2.1 Topological loss

In semantic image segmentation, conventional optimization loss functions (e.g. the cross-entropy) often proceed to a pixel-wise comparison of the class-prediction that is summarized in a likelihood function f to the one-hot encoded target vector. In this work, we aim at integrating the analysis of the global shape correctness of the prediction through the study of topology during the model optimization.

Persistent homology

Persistent homology offers a workaround to analyze the topology of a continuous-valued n-dimensional image function. In the context of our semantic image segmentation, we consider the likelihood map $f : \Omega \subset \mathbb{R}^n \to [0, 1]$ of a voxel to belong to the CP that is predicted from a DL-based model (Figure 7.1 (A)). In order to reduce the problem to a binary analysis as presented in the previous chapter (see Chapter 6, Section 6.3.1), persistent homology tracks the topological structures of f through filtration g_γ to different thresholds $\gamma \in [0, 1]$:

$$g_\gamma : [0, 1] \longrightarrow \{0, 1\}$$

$$x \longmapsto g_\gamma(x) = \begin{cases} 1 & \text{if } x \geq \gamma, \\ 0 & \text{otherwise.} \end{cases} \tag{7.1}$$

Snapshots of the topology are reported into a persistence barcode (Figure 7.1 (B)). Each bar corresponds to a topological structure (e.g. connected components, handles) which is characterized by its appearance and disappearance threshold values $(\gamma_{birth}; \gamma_{death})$. The persistence barcode can be filtered based on the structures persistence. The persistence of a structure is defined by its life time $\Delta\gamma = \gamma_{death} - \gamma_{birth}$. In persistent homology, the minimum persistence (mp) is the minimum lifetime accepted in the topological structures filtration. Figure 7.1 (B) shows an example of a likelihood image binarized at different thresholds $\gamma \in [0, 1]$. Two persistence barcodes with $mp = 0.001$ (top) and $mp = 0.1$ (bottom) are presented. With a lower mp, we observe the presence of a considerable amount of irrelevant structures. Indeed, the CP is a thin cerebral tissue (only a few voxels-wide in SR volumes), and is therefore sensitive to broken connections. In the barcode, this turns into the appearance of many connected components (blue) with short life time, i.e. low persistence.

Finally, the structure pairs $(\gamma_{birth}; \gamma_{death})$ extracted from the persistence barcodes are considered as coordinates. These coordinates define critical points, transcribed in the associated persistence diagram (see Figure 7.1 (C)). With low mp (top) we observe many critical points close to the diagonal. This diagonal corresponds to $\gamma_{birth} = \gamma_{death}$, i.e. the structure does not exist.

The choice of the *mp* must be set to avoid noise structures without being too strict. Note that for a binary image, all topological structures have a persistence $\Delta\gamma = 1$, with coordinates $(0; 1)$ in the persistent diagram representation.
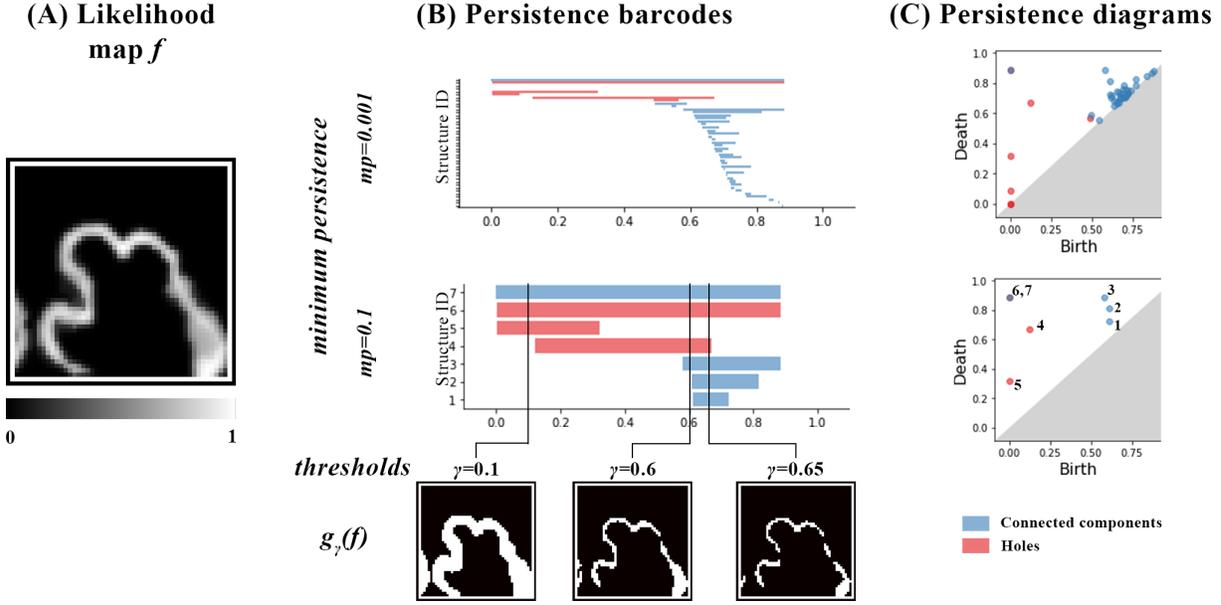
Figure 7.1: Illustration of persistent homology for a CP likelihood map f (A). Panel (B) illustrates the progressive filtration of f with minimum persistence (mp) of 0.001 (top) and 0.1 (bottom). Topological structures are reported into a persistence barcode. Panel (C) shows the persistence diagram for both mp. In the persistence barcode, respectively diagram, blue bars, respectively blue dots, represent the connected components. Similarly, red elements represent the 1-dimensional holes.

In practice, and similarly to [71], prior to the computation of topology, we pad the image patches twice with constant values (see Figures 7.1 (A) and 7.2 (A)). We perform a first padding with the maximal value of the patch in order to work with closed structures. Then, we pad the patch with 0 value voxels to define a background.

Topological loss function for fetal brain MRI

The topological loss function aims at directly comparing the persistent homology of the predicted likelihood map f to the target true topology. We propose a topological loss function $\mathcal{L}_{topo}$ that is adapted from [71]. Our contribution lies in the multi-dimensional approach of the topological loss computation. While our focus in this work is on the 1-dimensional holes, we observed in Chapter 6, Section 6.4.2 the importance to consider all topological dimensions. Additionally, we evidence in Section 7.2.1 the importance the connected components can have in the persistent homology of the fetal CP. Therefore, differently from [71] that only considered 1-dimensional structures (i.e. 2D holes), our topological loss for the fetal CP segmentation will additionally integrate the analysis of 0-dimensional homology structures.

We detail here the process of the computation of $\mathcal{L}_{topo}$ between two 2D image patches, the target segmentation and the predicted likelihood map f (see Figure 7.2 (A)), hence the dimensions involved are $k \in \{0;1\}$. First, persistent homology is computed tracking all k-dimensional structures in both images (see in Figure 7.2 (B), the persistence barcodes partitioned by dimension). Second, the per-dimension persistence diagrams are matched between the GT and the prediction (Figure 7.2 (C)). All k-dimensional structures are matched such that, the N-greater persistence structures are matched to the N true GT elements. Note that a sufficiently accurate likelihood is needed to prevent structures mismatch. Others are matched to the diagonal. Based on the implementation in [71][1], we compute a total squared distance [24], from the matched pairs of critical points in each dimension. The k-dimensional distance is our $\mathcal{L}_{topo-k}$, the k-dimensional topological loss function. Ultimately, $\mathcal{L}_{topo-k}$ are combined such that :

$$\mathcal{L}_{topo} = \sum_{k=0}^{K} \omega_k \mathcal{L}_{topo-k} \tag{7.2}$$

where $\mathcal{L}_{topo-k}$ is the topological loss of the k-dimensional space with a contribution weight of ω_k.

(A)   Patches   (B)   *k*-dimensional   (C) Persistence diagram   (D)   Topological loss
                              persistence barcodes   matching
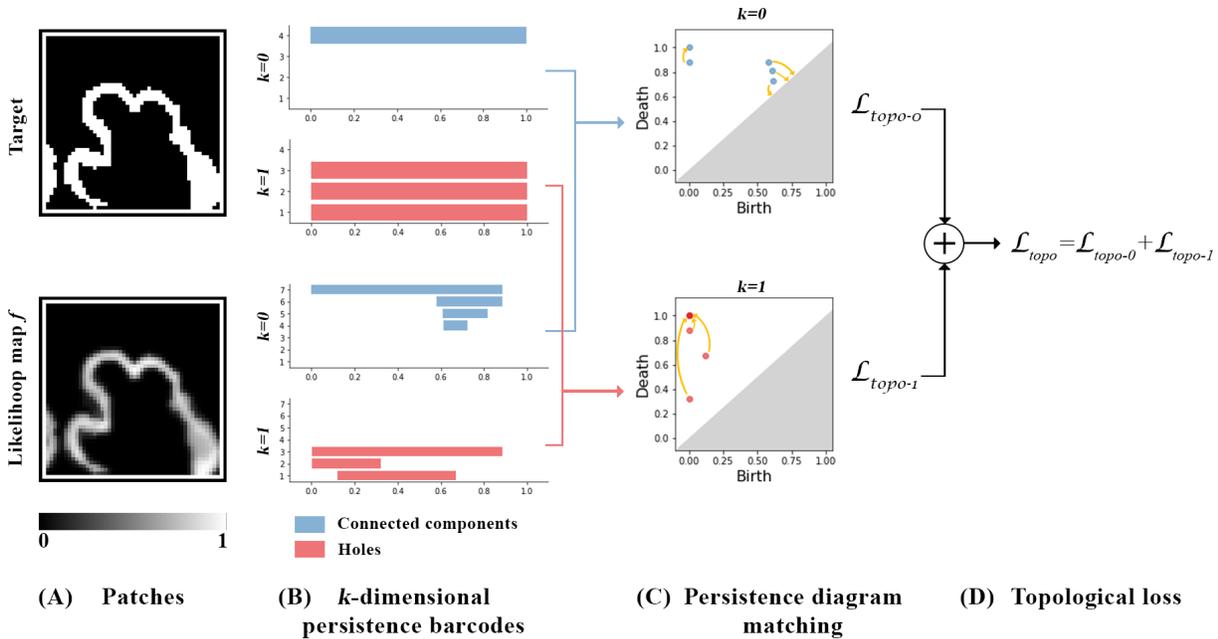
Figure 7.2: Illustration of the topological loss computation process between a GT binary image and a likelihood map f (A). (B) shows their corresponding persistence barcodes in each dimension ($k = 0$ and $k = 1$). (C) illustrates a per-dimension persistence diagram matching. (D) shows how the final $\mathcal{L}_{topo}$ is inferred, with $\omega_0 = \omega_1 = 1$. In the persistence barcode, respectively diagram, blue bars, respectively blue dots, represent the connected components. Similarly, red elements represent the holes.

---

[1]*https://github.com/HuXiaoling/TopoLoss*

7.2.2 Segmentation framework

Our topological loss is an architecture-agnostic optimization function. In other words, it is independent of the deep-learning network architecture. In our segmentation framework, we use a state-of-the-art architecture, U-Net [122], to compare different optimization methods. Three reference loss functions (*Baseline*, *Hybrid* and *Hu2019*) are implemented to evaluate the added value of our topological loss function (*TopoCP*) (see configurations in Section 7.2.2).

Model architecture

The well established U-Net [122] architecture is selected as it has recently proved its good accuracy in fetal brain MRI tissue segmentation [79, 112, 113]. We use a 2D U-Net architecture that is composed of an encoding and a decoding paths with skipped connections. The encoding path in our study is composed of 5 repetitions of the followings: two 3x3 convolutional layers, followed by a rectified linear unit (ReLU) activation function and a 2x2 max-pooling downsampling layer. Feature maps are hence doubled after each block, starting from 32 to 512. In the expanding path, 2x2 upsampled encoded features concatenated with the corresponding encoding path are 3x3 convolved and passed through ReLU. The network prediction is computed with a final 1x1 convolution. The number of network trainable parameters is 7,852,002.

Multiview patch-based approach

First, T2w images are masked in order to only consider intracranial space voxels in the CP segmentation. Second, to alleviate the computational cost due to the topological loss, 64×64 voxel size sub-image patches are extracted from the 3D volume in the three orthogonal planes (axial, coronal and sagittal). Bringing information from the three dimensional orientations, our method thus implements a 2.5D, or multiview, patch-based approach. To increase the number of predictions per voxel, overlapping patches are extracted. Empirically, the sliding window's step size for the patch extraction is set 16 voxels. Finally, intensities of all patches are standardized to have mean 0 and variance 1.

Training and optimization strategies

Input samples are randomly augmented at each epoch of the training phase. Extensive augmentations are performed spatially (flipping and elastic deformation) and intensity-based (bias, blurring, gamma and noise). All augmentation have a probability of occurrence of 0.5, except flipping occuring with a probability of 0.2. Augmentation are performed with the TorchIO python package (v0.18.75) [114].

Two reference segmentation methods (*Baseline* and *Hybrid*) are trained with fetal brain MRI state-of-the-art optimization loss function, in order to compare with our method (*TopoCP*). Thus, we evaluate the following three configurations:

- *Baseline* is trained using the distribution-based binary cross-entropy loss function $\mathcal{L}_{bce}$.

- *Hybrid* is trained with an hybrid loss function combining the dice loss $\mathcal{L}_{dice}$ and the binary cross-entropy loss $\mathcal{L}_{bce}$ such that:

$$\mathcal{L} = \mathcal{L}_{bce} + \mathcal{L}_{dice} \tag{7.3}$$

  Such hybrid loss function proved efficient in multi-tissue fetal brain MRI segmentation, as it has been used by the Top 5 teams of the 2022 edition of the MICCAI *FeTA* challenge [113].

- *Hu2019* is trained with the original topological loss function proposed in [71], such that:

$$\mathcal{L} = \mathcal{L}_{bce} + \lambda_{Hu2019}\mathcal{L}_{Hu2019} \tag{7.4}$$

  Note that $\mathcal{L}_{Hu2019}$ equates to our $\mathcal{L}_{topo-0}$. λ_{Hu2019} is set to 0.0001.

- *TopoCP* is trained with the following loss combination:

$$\mathcal{L} = (1 - \lambda_{topo})\mathcal{L}_{bce} + \lambda_{topo}\mathcal{L}_{topo} \tag{7.5}$$

  where $\mathcal{L}_{bce}$ is the binary cross-entropy loss and $\mathcal{L}_{topo}$ is the topological term presented in Section 7.2.1. λ_{topo} defines the weight of the contribution of $\mathcal{L}_{topo}$ in the final loss.

As the computation of the topological losses is expensive and need sufficiently accurate probability maps to perform a relevant matching of the structures (see Section 7.2.1), we adopt the training strategy presented in [71]: 1) a common warm-up network is trained over 15 epochs using the binary cross-entropy loss $\mathcal{L} = \mathcal{L}_{bce}$; 2) *Baseline*, *Hybrid*, *Hu2019* and *TopoCP* are initialized with the pretrained *warm-up* weights. An early stopping strategy monitors the global validation loss $\mathcal{L}$, respectively the topological validation losses $\mathcal{L}_{Hu2019}$ and $\mathcal{L}_{topo}$, for the *Baseline* and *Hybrid* configurations, respectively for *Hu2019* and our *TopoCP* configuration. All learning rates are set to 0.01. Training and evaluation were performed with Tensorflow v2.7 [1] and a GeForce RTX 2080TI GPU.

A 4-folds cross-validation approach is adopted to assess the learning performances of the different methods. In this way, we will assess multiple λ_{topo} in order to determine an optimal value (see Section 7.3.3).

Ensemble learning

In order to reduce the variance and increase the generalization power of our model, we adopt *ensemble learning*. In the final testing inference phase of each configuration, we

perform a majority voting on the summed likelihoods from all 4 cross-validation networks. Finally, despite cortex would theoretically be two connected components (left and right hemispheres), in practice, partial volume in the mid-sagittal plane most often leads to having one single component. Thus, only the biggest connected component of the **whole cortical volume** ensemble prediction is kept.

TopoCP **parameters setting**

As introduced in Section 7.2.1, persistent homology structures are filtered on a minimum persistence criteria. This criteria tunes the sensitivity of our loss to the noisy structures out of the filtration step. Empirically, we observed that the higher minimum persistence is the tougher filtration of the structures and thus may discard relevant ones. Reducing the persistence threshold leads to having an increasingly large formation of noisy irrelevant structures to be matched. Based on these empirical analysis, we set our minimum persistence to 0.01 for all experiments.

Equation 7.2 presents our global topological loss in which different contributions can be assigned to each dimension. We analyzed the importance of both 0-dimensional and 1-dimensional topological terms on a reduced set of randomly sampled patches. Empirically, we observe that the importance of 0-dimensional and 1-dimensional topological terms is patch-dependent. In some patches, $\mathcal{L}_{topo-0}$ is more affected than $\mathcal{L}_{topo-1}$, and vice versa in others. We therefore decided to give equal contribution to both terms, as they are undoubtedly both important. In other words, all k dimension had the same weighting $\omega_k = 1$.

Finally, the contribution of our global topological loss is valued with the λ_{topo}. We describe in Section 7.3.3 the cross-validation approach implemented to determine an optimal value.

7.2.3 Topological assessment of the fetal CP

Recent works [157, 97] evidenced the importance of considering complementary metrics for the assessment of semantic segmentation methods. Additionally, we observed in Chapter 6 the relevance of considering the topological correctness in the assessment and comparison of automatic segmentation methods.

Similarly to Chapter 6, Section 6.3.2, we consider BNEk to quantitatively compare the topology of the predicted segmentation and the GT expected value features. As the CP segmentation is filtered for its biggest connected component (see Section 7.2.2), BNE0 is incidentally irrelevant to consider. Additionally, this work specifically focuses on the presence of 1-dimensional holes. In this way, we only consider BNE1. Besides, considering that the human cortex is a closed structure with no obstruction, its GT expected BN1 is 0.

While BNE1 focuses on the count of occlusions on the CP surface, this score is not providing any information on the holes themselves. Indeed, we often observe that the number of holes that is a discrete value is not necessarily correlated to the size of the broken connections (i.e. the obstruction) nor to the size of the structure of interest (i.e. the fetal CP).

In that respect, we introduce a new metric that aims at quantifying the size of the hole. The hole ratio (HoR) is the ratio of false negative voxels that are connected to a hole (FN_{holes}) over the true voxels of the region of interest, which are represented by the true positives (TP) and false negatives (FN).

$$HoR = \frac{FN_{holes}}{TP + FN} \tag{7.6}$$

Our in-house implementation[2] illustrated in Figure 7.3 successively:

1. identifies the location of one voxel per 1-dimensional hole, using the cubical complex representation of the open-source GUDHI library[3]

2. propagates these voxels into the mask of holes candidates, i.e. the FN,

3. computes the volume ratio presented in Equation 7.6.

Let us note that this measure strongly relies on the topological correctness of the GT.
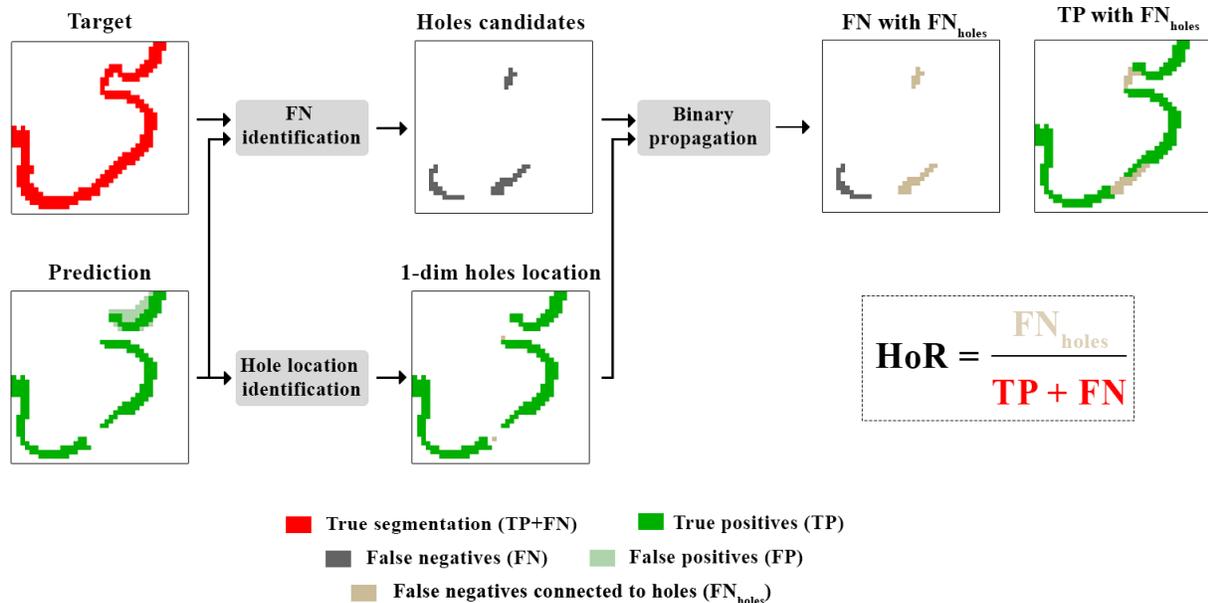
Figure 7.3: Illustration of the workflow for the computation of the Hole Ratio (HoR).

Figure 7.4a shows (B) and (C), two 3D rendering of the same image patch CP segmentation. (A) shows the 3D rendering of the GT, enlighting the region of interest. One can easily observe that the (B) segmentation has a main hole compared to the (C) segmentation that presents multiple medium-size holes. Additionally, quantitative results (see Table 7.4b) confirm the discrepancy between the quantity of holes as a number and the quantity of holes as a ratio of the region of interest.

Albeit BNE1 must be used with caution, it is still a relevant score to assess the CP topology in the absence of topologically accurate GT segmentation. Nonetheless, we promote the use of an additional quantitative metric relative to the size of the occlusions to undertake a robust quantitative analysis of the broken connections in the CP segmentation.

Both topology-based measures (BNE1 and HoR) rely on the cubical complex implementation of the GUDHI library (v3.5.0) [27].
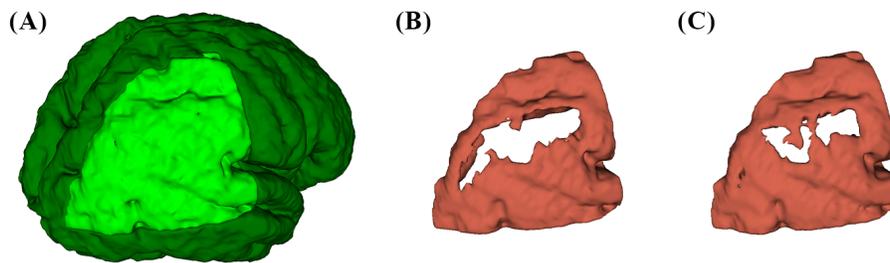
(a) 3D rendering of (A) GT subject CP (dark green) and region of interest (light green), and two different segmentation (B) and (C).

	(B)	(C)
BNE1	1	4
HoR	0.36	0.33

(b) Table of the topology-based metrics (BNE1: 1-dimensional Betti number error; HoR: Hole ratio) for the example cortical plate segmentation (B) and (C) shown in Figure 7.4a.

Figure 7.4: Illustration of the discrepancy between the quantification of holes as a number and its quantification as a ratio over the region of interest.

7.3 Experiment design

The overall experiment design to compare the four segmentation frameworks *Baseline*, *Hybrid*, *Hu2019* and our *TopoCP* is outlined in Figure 7.5.
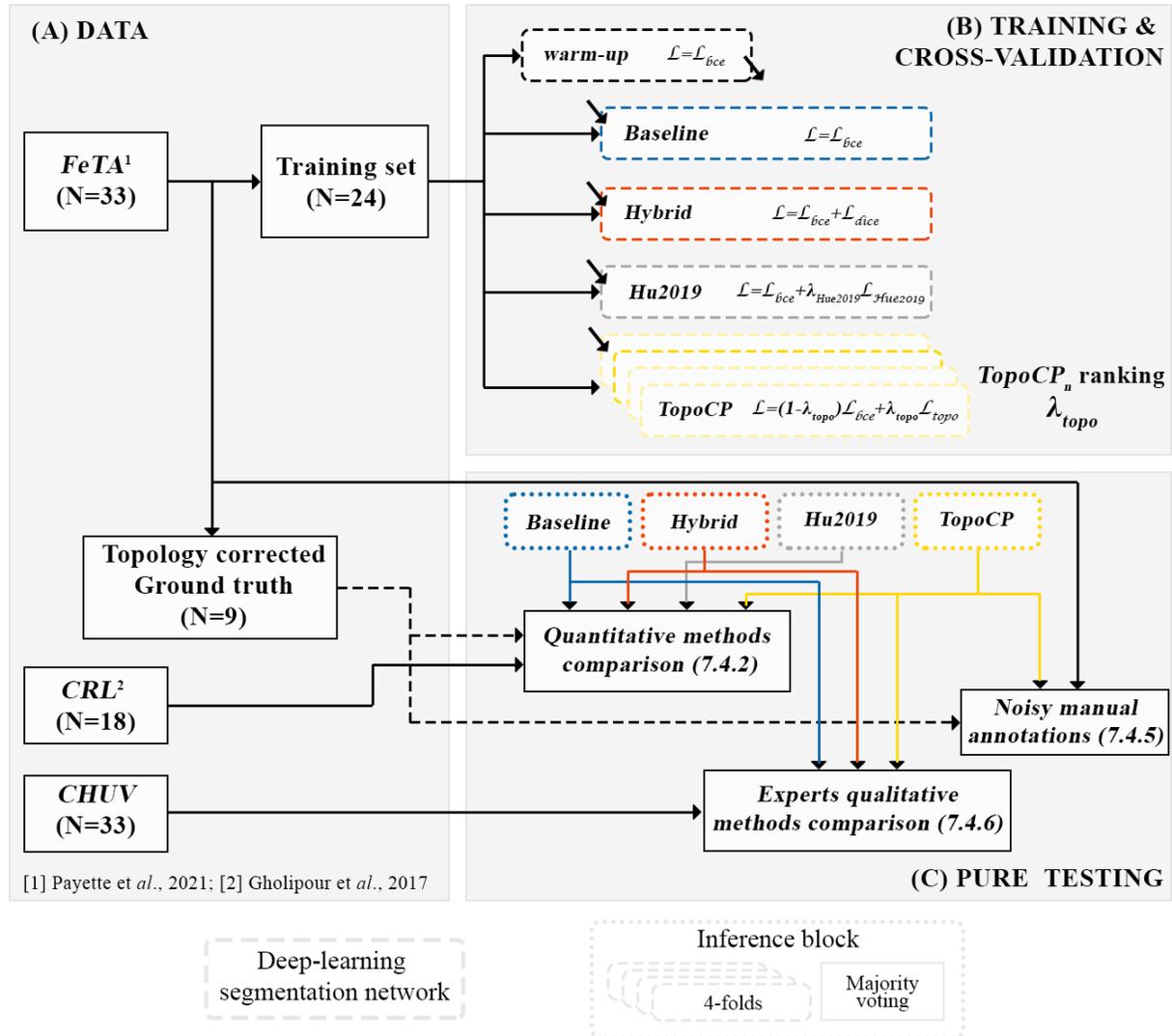
Figure 7.5: Illustration of the overall experiment design. Panel (A) illustrates the different datasets and their splitting for training/testing purposes (see Section 7.3.1). Panel (B) illustrates the training phase. A common warm-up network is trained to initialize the three configurations *Baseline*, *Hybrid*, *Hu2019* and *TopoCP*, each optimized with its own optimization loss function (see Section 7.2.2). A cross-validation approach is used to determine an optimal hyperparameter λ_{topo} (see Section 7.3.3). Panel (C) illustrates the testing phase. Predictions are inferred through the cross-validation networks and combined in a majority voting step. Methods are assessed and compared quantitatively with complementary performance metrics and qualitatively by three experts (see Section 7.3.3).

7.3.1   Datasets

A summary of clinical and atlas datasets is shown in Table 7.1.

	Dataset	Number of subjects (NT/PT)	Gestational age (weeks)	SR reconstruction	Image resolution (mm^3)
TRAINING	*FeTA* [112]	24 (13/11)	[20.9-34.8] (28.2±3.6)	SVRTK [88]	$0.86 \times 0.86 \times 0.86$
TESTING	*FeTA* [112]	9 (4/5)	[22.9-34.8] (27.4±3.6)	SVRTK [88]	$0.86 \times 0.86 \times 0.86$
	CRL [56]	18 (18/0)	[21-38]	Gholipour et. al, 2017 [56]	$0.80 \times 0.80 \times 0.80$
	CHUV	33 (24/9)	[21-35] (29.6±3.6)	MIALSRTK [143]	$0.80 \times 0.80 \times 0.80$

Table 7.1: Summary of the data used for training and quantitative and qualitative evaluation. NT: Neurotypical, PT: Pathological

Clinical dataset: *FeTA*

We use the subset of the public *FeTA* v2.0 Dataset [112] of fetal brains reconstructed with SVRTK [88] at isotropic resolution of 0.86*mm*. After visual inspection of the images, seven volumes were excluded due to bad SR quality (3) and severe pathology (7) (e.g. major ventriculomegaly). The remaining 33 fetal brains were composed of 17 neurotypical and 16 pathological subjects, in the GA range of 20.9 to 34.8 weeks. Twenty-four (24) subjects (13 neurotypical and 11 pathological subjects in the GA range of 20.9 to 34.8 weeks, 28.2±3.6) were randomly selected for the method development and the remaining nine (9) subjects (4 neurotypical and 5 pathological subjects in the GA range of 22.9 to 34.8 weeks, 27.4±3.6) were retained for pure testing purposes. Note that details on the fetal brain pathologies are not disclosed in the dataset information.

In this chapter, we exclusively consider the cGM label – i.e. the CP. Annotations were manually performed following an optimized protocol. Two experts respectively annotated the external border of the cortex cerebri and the external border of the white matter, on every 2nd to 3rd slice of the axial view. Individual structure annotations are post-processed with interpolation and smoothing prior to merging into a final label maps. Ultimately, sparse interpolated annotations result in noisy label maps often showing topological inconsistencies. Figure 7.6 shows the extracted cGM from the final label maps (left) for (A) Subject 1, a 34.8 weeks of GA neurotypical subject and (B) Subject 2, a 28.1 weeks of GA pathological subject. Three-dimensional (3D) rendering evidences the presence of apertures in the final CP annotations.

As motivated in Section 7.2.3, topologically accurate GT segmentation are necessary to perform a valid topological assessment of an automatic method. In this respect, we perform further manual correction of the CP *FeTA* manual annotations. Four engineers refined the CP label maps of the 9 fetal brains of the clinical pure testing set. Editing of the label maps were performed using the ITK-SNAP [158] software with a specific focus on the topological correctness and contour refinement of the label maps. Finally, all

CP manual corrections were checked (and corrected if needed) by a pediatric radiologist with 17 years of experience. Right columns of panel (A) and (B) of Figure 7.6 show the corrected tissue annotations overlaid to the T2w image and their 3D rendering. In our further experiments, we refer to the corrected manual annotations as the GT.
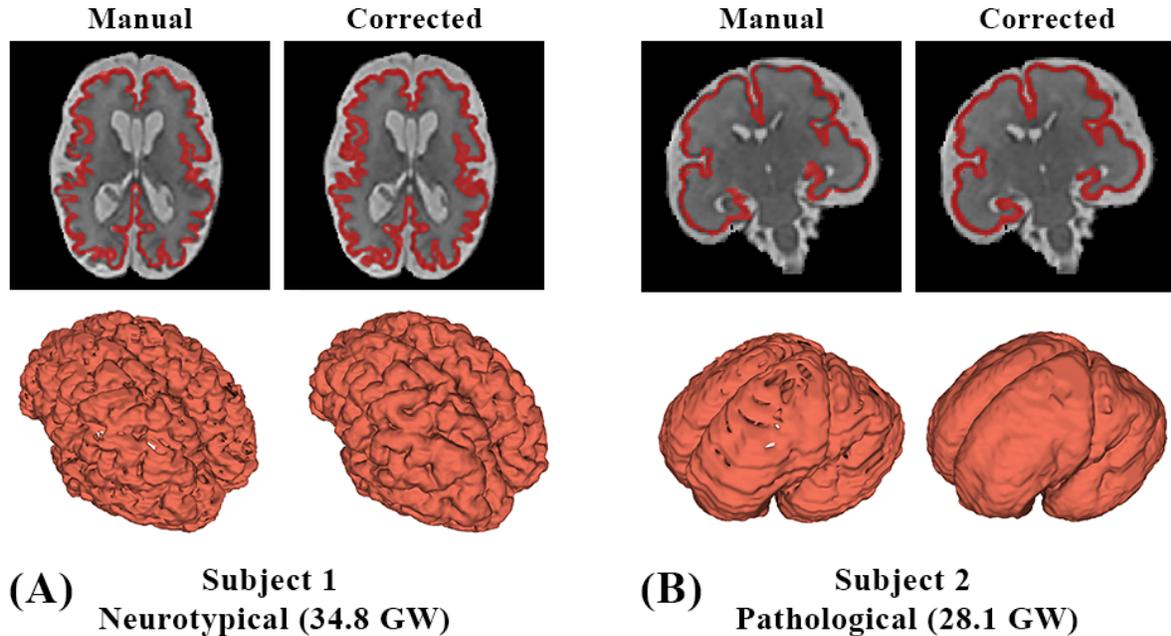
Figure 7.6: Illustration of the *FeTA* original manual and the corrected GT CP segmentation for (A) Subject 1, a 34.8 weeks of GA neurotypical subject and (B) Subject 2, a 28.1 weeks of GA pathological subject. T2w images and CP segmentation overlaid (top) are respectively shown on an axial and coronal view for Subject 1 and Subject 2. 3D rendering is presented for all segmentations (bottom).

Atlas dataset: *CRL*

We use the *CRL* spatiotemporal atlas of the fetal brain [56] that provides 3D high-quality isotropic volumes for all GA between 21 and 38 weeks (see Chapter 1, Section 1.4.2). Each atlas subject is constructed with the contribution of 6 to 23 SR-reconstructed individual fetal brains. The integration of multiple subjects per GA reduces the morphological variability. Therefore, the T2w atlas images appear smoother than clinical acquisitions.

The atlas comes with two label maps, respectively containing the cerebral tissue and structure labels, and a regional cortex parcellation. The initial tissue label maps of more than 50 classes are converted into a 7-tissue label maps, matching those defined in the *FeTA* dataset. Minor adjustments are performed while synchronizing tissue and regional maps. For instance, voxel labelled as cortex in one map and as corpus callosum in the other are dumped in WM class. As opposed to the clinical dataset, atlas labels do not require further manual corrections as they were already manually refined and present decent topology [56].

Out-of-domain clinical dataset: *CHUV*

Thirty-three fetal brain clinical MR examination conducted at the Lausanne University Hospital (*CHUV*), Lausanne, Switzerland, were SR-reconstructed with the MIALSRTK pipeline [144, 143]. SR volumes are further resampled to an isotropic resolution of 0.8*mm* and an engineer coarsely realigned the volumes to the anatomical plane. This clinical set is composed of 24 neurotypical and 9 pathological subjects in the GA range of 21 to 35 weeks (29.6±3.6). The quality of the SR reconstructions was assessed similarly as in Chapter 2 into three categories [80]: bad (non usable, very blurred), average (overall good with remaining partial volume effect/blurring), and excellent (good quality with no blurring). Overall, none of the clinical SR reconstruction is bad, 17 are average 14 are excellent. No reference segmentation are available for this dataset.

The local ethics committee of the Canton of Vaud, Switzerland (CER-VD 2021-00124) approved the retrospective collection and analysis of MRI data and the prospective studies for the collection and analysis of the MRI data in presence of a signed form of either general or specific consent.

7.3.2 Assessment metrics

Although our segmentation framework infers CP segmentation in a 2.5D multi-view strategy (2D image patches from the three orthogonal planes), we proceed to the quantitative evaluation in 3D that is of the whole cortical volume. Automatic medical image segmentation requires the use of complementary metrics for the assessment of different aspects of the segmentation [157, 97]. Most commonly, overlap-based (e.g. DSC, Jaccard similarity coefficient, Intersection over union) and distance-based (e.g. X^{th} percentile Hausdorff distance, ASSD) metrics are reported [157]. However, in this work we aim at assessing the segmentation not only in terms of overlap and distance accuracy to the GT but also in terms of shape correctness. Therefore, we also consider topology-based metrics. Table 7.2 summarizes the metrics used in the training (for learning monitoring) and testing (for final evaluation) phases.

The DSC [28] is an overlap-based similarity metric. Robust to outliers, it is a widely-used metric to assess medical image segmentation accuracy. ASSD is the mean of the directed average surface distances [157]. The latter is defined as the average of the distances of points from one surface to their closest points on the other object boundary. The ASSD is computed using the python MedPy [4] implementation (v0.4.0).

In the case of absence of topologically accurate GT (i.e. see Section 7.3.3 for cross-validation details), we consider BNE1 to quantitatively assess the topology, while our proposed hole ratio HoR is used in the pure testing phase (see details on topology metrics in Section 7.2.3).

---

[4]https://loli.github.io/medpy/

Arrows in Table 7.2 indicate whether each metric is better maximized or minimized. Taking values between 0 and 1, DSC is a similarity metric that is better maximized (↑). Difference metrics (ASSD, BNE1 and HoR) must be minimized (↓).

	Overlap	Boundary-distance	Topology
Training	DSC ↑	ASSD ↓	BNE1 ↓
Testing	DSC ↑	ASSD ↓	HoR ↓

Table 7.2: Summary of the metrics used during the training phase (for learning monitoring) and testing phase (for evaluation). Arrows indicate whether higher ↑ or lower ↓ scores are better.

7.3.3 Experiments

$\mathcal{L}_{topo}$ parameter settings

Our first experiment consists in the setting of the *TopoCP* λ_{topo} parameter that quantifies the contribution of our topological loss. As mentioned in the prior Section 7.2.2, we use a cross-validation approach by means of which we assess multiple λ_{topo} in order to determine an optimal value. The ideal λ_{topo} is a dataset dependant hyperparameter. According to [71], λ_{topo} must be chosen to avoid the risk of over-weighting of $\mathcal{L}_{topo}$ over $\mathcal{L}_{bce}$. Indeed, while $\mathcal{L}_{bce}$ is defined at every voxel of the image, $\mathcal{L}_{topo}$ is only defined at some critical points. The values 0.0002, 0.005, 0.001, 0.05, 0.1 and 0.2 are the N λ_{topo} evaluated in the training phase. *TopoCP$_n$* define the set of 4 networks trained for cross-validation with λ_n. We consider the DSC, the ASSD and the BNE1 for evaluation. The average performances over the folds are computed for each *TopoCP$_n$* networks and ranked for each metric. *TopoCP$_n$* are finally ranked, based on the sum of metric-wise ranking, to elect the optimal λ_{topo}. The latter is then selected in the following experiments.

Methods comparison

We compare our *TopoCP* method to the three reference segmentation methods *Baseline*, *Hybrid* and *Hu2019* on both the clinical and atlas test sets. We assess the three complementary metrics DSC, ASSD and HoR. We perform paired Wilcoxon rank-sum tests to assess the statistical significance between *TopoCP* and the three reference configurations. Significance level is set to 0.05.

GA analysis

The *CRL* set presents a large and steady range of GA with one subject per week from 18 to 38 weeks. We observe the quantitative performances of the DSC, the ASSD and the HoR along gestation, i.e. as a function of the subject GA.

Spatial topological analysis

From the *CRL* set, we group the regional labels into 5 classes corresponding to the brain lobes, namely the frontal lobe, the occipital lobe, the parietal lobe, the temporal lobe, and the insula lobe. Figure 7.7 shows 3D rendering of the final maps of the brain lobes for the subject atlas of 21, 30 and 38 weeks of GA.

We proceed to a lobe-based analysis of the topology HoR metric to analyze if *TopoCP* benefits in one, some, all or none of them. We perform a paired Wilcoxon rank-sum test to assess the statistical significance between *TopoCP* method and the reference configurations. Statistical significance level is set to 0.05.
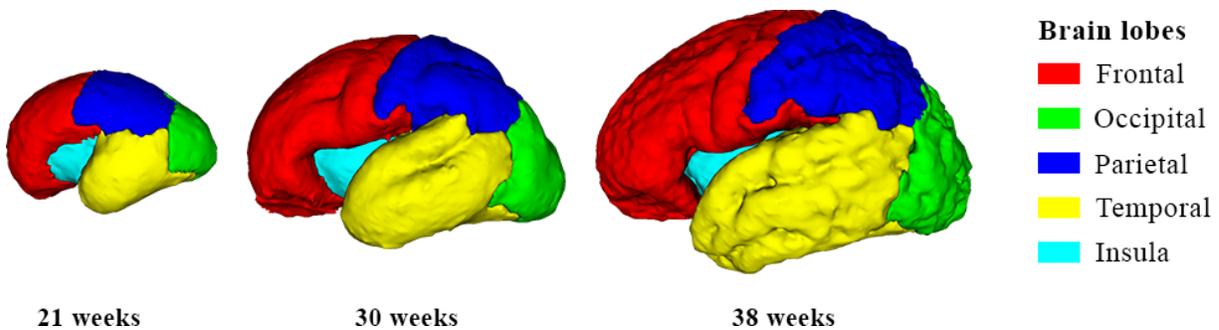
Figure 7.7: 3D rendering of the CP for *CRL* subjects of 21, 30 and 38 weeks of gestation. The cortical volumes are split into the 5 lobes of the brain: the frontal lobe (red), the occipital lobe (green), the parietal lobe (dark blue), the temporal lobe (yellow), and the insula lobe (light blue).

Manual annotations comparison

Ultimately, we evaluate the performances of *TopoCP* compared to the original *FeTA* manual annotations, using the topologically corrected segmentations as GT.

Let us note that these original *FeTA* annotations are sparse and interpolated, hence resulting in *noisy* references. Nevertheless, they are still used for training, as manual topological correction of 24 volumes would not be realistic (time/expertise effort). We quantitatively assess these segmentation with the DSC, the ASSD and the HoR. The segmentation correctness difference was tested with the paired Wilcoxon rank-sum. The *p*-value level for statistical significance was set at 0.05.

Experts evaluation

Three experts in fetal brain MRI (two radiologists and one engineer) perform an independent and blind assessment of the three automatic segmentation methods on the clinical *CHUV* dataset. Each fetal brain MR exam is provided with the SR reconstructed volume, the subject's GA at scan time, the subject's group (i.e. Neurotypical or Patho-

logical) and the three segmentation (from configurations *Baseline*, *Hybrid* and *TopoCP*) that are randomly anonymized with labels *A*, *B* and *C*.

The experts are asked to rank the segmentation masks *A*, *B* and *C* as Best, Medium and Worst. Visualization of the images and their segmentation are done with the open-source ITK-SNAP [158] software. Specifically, binary segmentation are both visualized in 2D, as an overlay to the T2w gray-scale SR images, and in 3D, with the ITK-SNAP integrated 3D viewer.

We assess the inter-rater reliability with the percentage agreement and an ordinal Gwet's agreement coefficient (Gwet's AC) that we interpret according to Altman's benchmarking scale [63]. We further consider a consensus evaluation as the majority voting of the experts' evaluation.

7.4   Results and Discussion

7.4.1 λ_{topo} hyper-parameter tuning

Table 7.3 shows the averaged validation scores of all three configurations, and specifically for each λ_{topo} assessed in the *TopoCP* configuration. Our first observation is that, regardless of the value of the λ_{topo} parameter, *TopoCP* is better performing than all three reference methods, as we reach the state of the art performances in all the three complementary metrics. Overall, all $TopoCP_n$ give similar DSC (mean: 0.76) and ASSD (mean ± standard deviation: 0.27 ± 0.01) performances, although λ_{topo} = 0.01 is of the highest rank for both overlap and boundary-distance based scores. An increased inter-$TopoCP_n$ variability appears for the topology-based metric (BNE1) with mean score from 20.6 to 22.5. Counting the number of bores in the CP segmentation, λ_{topo} = 0.005 is giving the best performances. We observe large BNE1 standard deviation for all $TopoCP_n$. Nonetheless, the finest topology-relative λ_{topo} is not only giving the minimum averaged BNE1, but is also noticeably presenting a smaller BNE1 standard deviation of 7.8 (BNE1 range: from 7.8 to 10). Therefore, λ_{topo} = 0.005 is the most accurate and precise of the λ_{topo} assessed as for the topology fidelity. The substantial fluctuation in the topological metric shows the importance of the choice of the λ_{topo} hyper-parameter.

Finally, our global ranking that is derived from metric-wise rankings evidences the ideal value λ_{topo} = 0.005. We observe that none of the extreme values considered (i.e. 0.0002 and 0.2) are in the Top 3 best performing λ_{topo}. Therefore, we can say that although λ_{topo} = 0.005 might not be the exact optimal λ_{topo}, it certainly falls in a relevant range and in the right order of magnitude.

Configuration		DSC ↑	ASSD ↓	BNE1 ↓	Ranking ↓
Baseline		0.748 ± 0.009	0.292 ± 0.02	29.8 ± 14.5	
Hybrid		0.744 ± 0.004	0.297 ± 0.01	31 ± 13.4	
Hu2019		0.753 ± 0.006	0.282 ± 0.01	26.6 ± 12.1	
TopoCP λ_{topo}	0.0002	0.758 ± 0.007 (5)	0.274 ± 0.01 (5)	22.1 ± 10 (4)	5
	0.001	0.761 ± 0.007 (2)	0.270 ± 0.01 (3)	21.0 ± 8.5 (2)	2
	0.005	0.760 ± 0.007 (3)	0.269 ± 0.01 (2)	**20.6 ± 7.8 (1)**	**1**
	0.01	**0.762 ± 0.007 (1)**	**0.268 ± 0.01 (1)**	22.5 ± 9.0 (5)	2
	0.2	0.760 ± 0.005 (4)	0.272 ± 0.01 (4)	21.6 ± 7.9 (3)	4

Table 7.3: Table of the validation scores (mean ± SD) of the DSC, the ASSD and the BNE1. Arrows indicate whether the metric is better maximized ↑ or minimized ↓. The best scores between all λ_{topo} are shown in bold. A ranking for each metric is shown in parenthesis. The final ranking is formulated from the sum of metric-wise ranking scores.

7.4.2 Methods comparison

Figure 7.8 illustrates the accuracy of the fetal CP segmentation for a pathological subject of 26.6 (Subject 1) and a neurotypical subject of 34.8 (Subject 2) weeks of GA. The topologically corrected GT and the four configurations segmentation with *Baseline*, *Hybrid*, *Hu2019* and *TopoCP*, are compared. Qualitative 2D assessment (top rows) of the segmentation are presented as an overlay on the T2w image on an axial, respectively coronal, view for Subject 1, respectively Subject 2. Additionally, 3D rendering of the CP segmentation are presented in the bottom rows. Overall, we observe that all configurations generates a thinner ribbon than the corrected GT. Specifically, *TopoCP* presents fixed cortical connections that are broken in the *Baseline*, *Hybrid* and *Hu2019* segmentations (white arrows). The CP *TopoCP* segmentation 3D rendering seems to present less bores than the three reference configurations *Baseline*, *Hybrid* and *Hu2019*. In particular, the segmentation appears, equivalently for the young and the old fetuses, more challenging, for all methods, in the lower parts of the frontal and temporal lobes, although *TopoCP* seems to exhibit a more sensitive segmentation in these areas. *TopoCP* appears to be more sensitive to the complexity of the CP morphology in older fetuses. White circles evidence in Subject 2 an improved segmentation in the hippocampal area and the depth of a gyrification. White arrows show area where the topological correctness recovered with *TopoCP*, compared to the *Baseline*. In Subject 1, two connections are fixed in the frontal lobe, although one of them is already fixed in the *Hybrid* configuration.

Quantitative results for both test sets (total of 27 cases), *FeTA* and *CRL*, are respectively presented in Table 7.4a and Table 7.4b. Tables show the mean ± SD of the CP segmentation for each testing metric (DSC, ASSD and HoR) in each configuration. Overall, the performance of the segmentation framework is improved when trained with our optimal *TopoCP* configuration (i.e. λ_{topo} = 0.005) for all metrics in both datasets. *TopoCP* is always performing significantly better than all three reference methods *Baseline*, *Hybrid* and *Hu2019*.
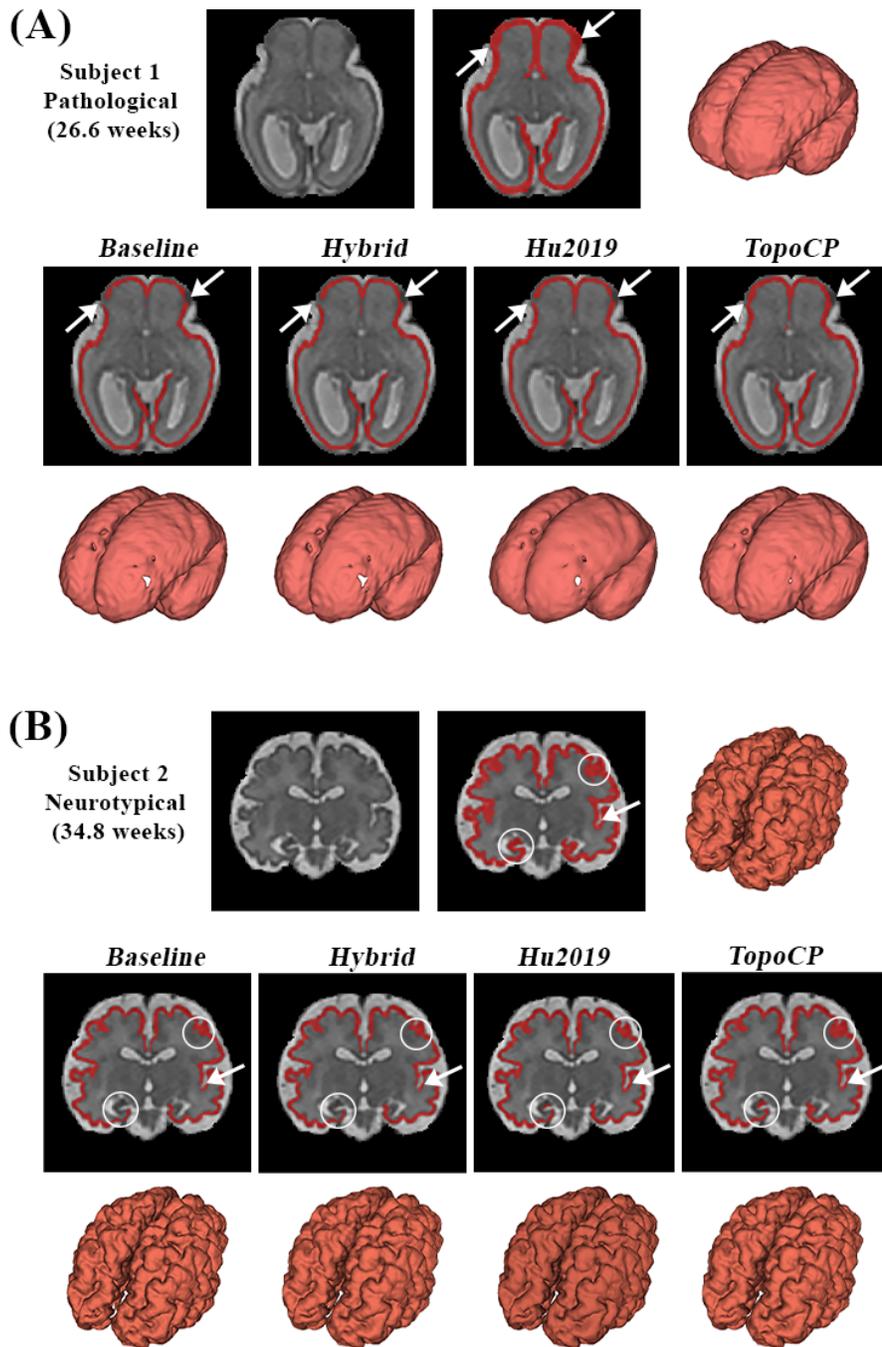
Figure 7.8: Segmentation results and 3D rendering of the CP on *FeTA* subjects: (A) a neurotypical subject of 34.8 weeks of GA, and (B) a pathological subject of 28.1 weeks of GA. Comparison of the manually corrected GT segmentation (top), and the results of automatic segmentation (bottom) with the *Baseline*, *Hybrid*, *Hu2019* and our custom *TopoCP* configurations. White circles show representative area where cortical gyrification have a better in-depth segmentation with *TopoCP* method. White arrows show fixed connections using our *TopoCP* method compared to the reference segmentations.

Furthermore, while all analyzed aspects of the segmentation, namely the overlap, the boundary-distance and the topology are improved with *TopoCP* compared to the reference methods, we observe a drop in the performances between *FeTA* and *CRL* evaluation. Indeed, DSC goes from 0.85 in *FeTA* to 0.79 in *CRL*, ASSD from 0.19 to 0.41 and HoR from 0.06 to 0.23. We believe this is due to the domain shift between *FeTA* and *CRL* images (different reconstruction pipelines, different intensity-based processing, etc) generating an inter-dataset variation in the data distribution. Such domain gap between the two sets of images is not learned from the training data that are only composed of *FeTA* images. Still, DL can generalize to some extent. Further training with multi-dataset images or the use of domain adaptation strategies can partially fit the domain gap. Let us note that we do not address the data distribution generalization in this paper as it is beyond its scope. Furthermore, such performance drop occurs also in *Baseline* and *Hybrid* approach.

Table 7.4: Tables of the metrics computed on the pure testing sets *FeTA* 7.4a and *CRL* 7.4b. Mean ± SD for the DSC, the ASSD and the HoR are presented. Arrows indicate whether the metric is better maximized ↑ or minimized ↓. The best scores between all three configurations are shown in bold. *p*-values of Wilcoxon rank sum test between *TopoCP* and the reference configurations, *Baseline* (*), *Hybrid* (+) and *Hu2019* (⋆), are considered statistically significant for $p < 0.05$.

	DSC ↑	ASSD ↓	HoR ↓		DSC ↑	ASSD ↓	HoR ↓
Baseline	0.82 ± 0.02	0.22 ± 0.05	0.093 ± 0.03	*Baseline*	0.77 ± 0.05	0.42 ± 0.14	0.25 ± 0.10
Hybrid	0.82 ± 0.02	0.23 ± 0.06	0.10 ± 0.04	*Hybrid*	0.77 ± 0.05	0.42 ± 0.15	0.26 ± 0.11
Hu2019	0.83 ± 0.02	0.21 ± 0.05	0.08 ± 0.04	*Hu2019*	0.78 ± 0.05	**0.40 ± 0.015**	0.29 ± 0.14
TopoCP	**0.85 ± 0.01** (*, +, ⋆)	**0.19 ± 0.04** (*, +, ⋆)	**0.06 ± 0.03** (*, +, ⋆)	*TopoCP*	**0.79 ± 0.05** (*, +, ⋆)	0.41 ± 0.18 (*, +,)	**0.23 ± 0.10** (*, +, ⋆)

| (a) *FeTA* | (b) *CRL* |

7.4.3 Segmentation performance over gestation

Taking advantage of the steady GA-distribution in the *CRL* set, we perform an analysis of the metrics throughout gestation. Figure 7.9 shows the performance metrics as a function of the GA for 18 cases from 21 to 37 weeks of GA. Regardless of the configuration, we observe a trend in the performances based on the GA. Indeed, all metrics reach better performances for subjects younger than 30 weeks of GA. From week 23 to 31, all four methods (*Baseline*, *Hybrid*, *Hu2019* and *TopoCP*) seem to give equivalent scores. Outside this range (i.e. GA< 23 and GA> 30 weeks), *TopoCP* is always performing better than the other two methods, except for one outlier subject of 38 weeks of GA.

We visually inspect the *TopoCP* segmentation mask of the *CRL* 38-weeks-old subject to better understand the origin of this outlier. Specifically in the cerebellum, we observe the presence of false positives that are connected to the main cortical segmentation through brainstem false positives. The cerebellum is a "little brain" composed of

white matter encased in the cerebellar cortex. In terms of fetal brain T2w MR contrast and similarly to the CP, the cerebellar cortex expresses as a thin dark ribbon surrounding white matter. Therefore, it is a challenging area to accurately differentiate in the segmentation of the fetal CP at a patch-level. Similar mis-segmentations appear in younger fetuses, nevertheless our post-processing step to keep the biggest connected component filters out most of it. In this oldest subject, cerebellum errors are worsened with brainstem mis-segmentation. We hypothesize such false positive errors in the cerebellum are due to mis-leading contextual information, due to the reduced field of view of the patches. Therefore, we believe that increasing the patch size could help to overcome these mis-segmentation. Nevertheless, while the performances are particularly damaged in the distance metric, *TopoCP* still performs better in terms of DSC and HoR compares to the other configurations.

Overall, it is in the second half of the third trimester GAs (i.e. GA>30 weeks) that we observe an increased benefit from *TopoCP*, compared to other methods. Our topological loss has a stronger positive effect on the topological errors for old subjects with more complex topology, although the whole range of GA consistently presents benefits from *TopoCP*.

We derive two hypothesis on the variation of the performances throughout gestation. First, we recall that the training data present subjects in the range 20.9 to 34.8 weeks of GA with mean 28.2 and standard deviation 3.6. Therefore, variation of young and old fetal brains are less represented in the learning process. Additionally, third trimester subjects present advanced sulcal patterns, resulting in a substantially more complex topology. Therefore, we postulate this accentuate the unstable evolution of segmentation accuracy over gestation.

7.4.4   Topology analysis per brain lobes

Figure 7.10 presents a comparison of HoR at a lobe-level between the configurations. This boxplot evidences the significant benefits ($p < 0.05$) of *TopoCP* in most areas (frontal, occipital, temporal and insula lobes) compared to the *Baseline* configuration. In the parietal lobe, *TopoCP* is on average performing better than the *Baseline* although without statistical significance. Compared to either the *Hybrid* or *Hu2019* configuration, *TopoCP* presents a significantly lower HoR in all brain lobes. Regardless of the configuration, the parietal lobe is always the better segmented lobe in terms of HoR as opposed to the insula lobe.

7.4.5   Robustness to noisy manual annotations

Figure 7.11 (top) shows a comparative T2w axial view of the GT topologically corrected segmentation (A), the original manual annotation provided in *FeTA* (B) and the *TopoCP* predicted segmentation (C). Overall, we observe an improved accuracy with the auto-
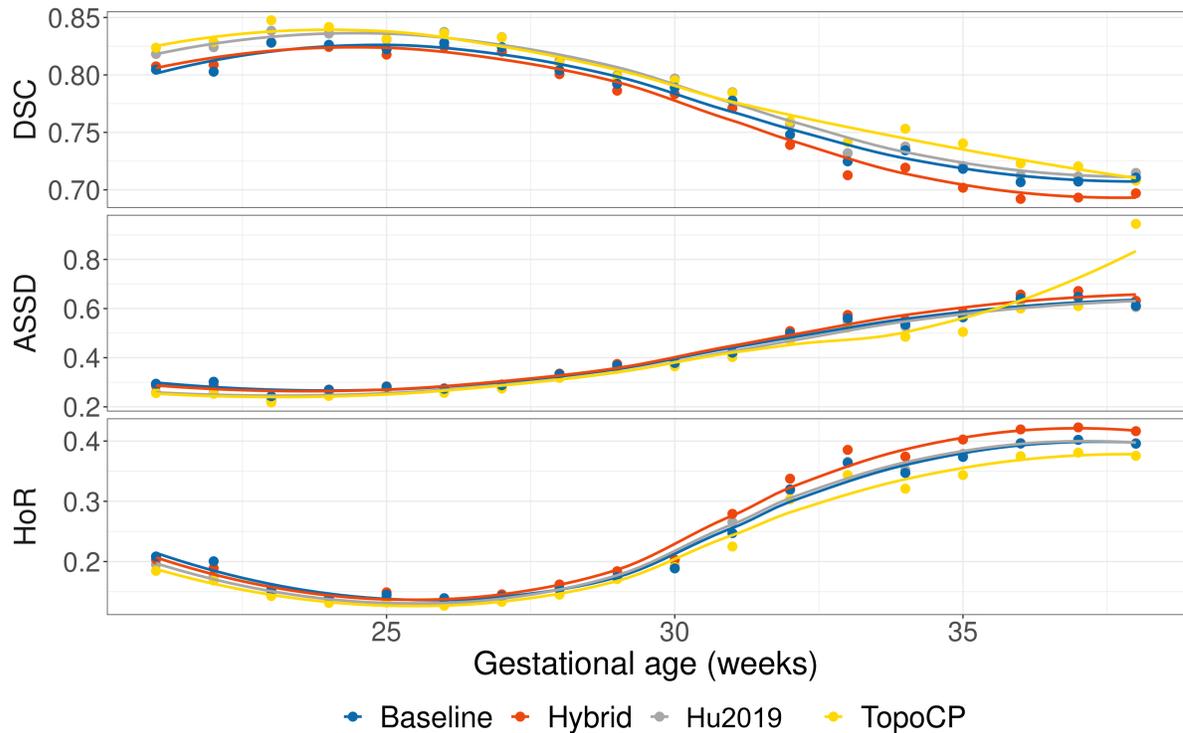
Figure 7.9: Evolution of the performance metrics (DSC, ASSD and HoR) on the *CRL* images as a function of the GA (from 21 to 38 weeks of gestation).

matic method. Specifically, white arrows indicate cortical location where *TopoCP* fixes topological inconsistencies, compared to the original manual annotations. The white circle focuses on the hippocampal area where the manual annotations are confused. Figure 7.11 (bottom) shows 3D rendering of the true cortical volume (green). In the manual and automatic *TopoCP* segmentation, the 1-dimensional holes are evidenced with the false negatives connected to 1-dimensional holes (light camel).

Table 7.5 shows a comparison of the performance metrics of the original *FeTA* manual annotations on 9 subjects and our *TopoCP* method. *TopoCP* is significantly better (*) than the original manual annotations in all metrics (DSC, ASSD and HoR).

The automatic *TopoCP* segmentation method is able to learn segmentation features from noisy annotations. This improvement is conveyed in all three similarity, boundary-distance and topology -based metrics.

7.4.6 Out-of-domain qualitative assessment

Table 7.6 summarizes the classification results of the segmentation masks according to each expert into Worst, Medium or Best. A consensus of the three experts assessment is presented in the bottom row.
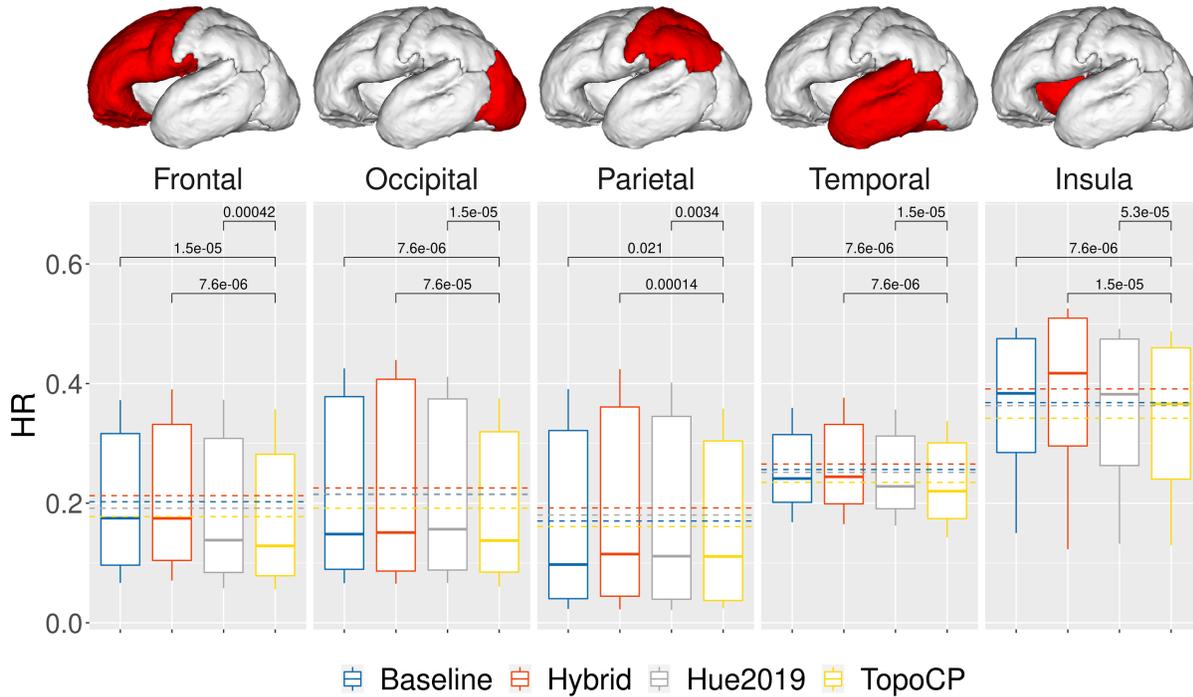
Figure 7.10: Comparison of the hole ratio (HoR) in the fetal brain lobes (frontal, occipital, parietal, temporal and insula) on the *CRL* dataset for all configurations (*Baseline*, *Hybrid*, *Hu2019* and *TopoCP*). Dashed horizontal lines indicate the per-lobe mean HoR for each configuration. *p*-values of paired Wilcoxon rank-sum tests are displayed comparing *TopoCP* to each reference methods.

Table 7.5: Table of the metrics computed on the pure testing sets *FeTA*. Mean ± sd for the DSC, the ASSD and the HoR are presented. Arrows indicate whether the metric is better maximized ↑ or minimized ↓. The best scores between our *TopoCP* method and the original annotations are shown in bold. *p*-values of Wilcoxon rank sum test between *TopoCP* and the original annotations are considered statistically significant (*) for $p < 0.05$.

	DSC ↑		ASSD ↓		HoR ↓	
TopoCP	**0.85 ± 0.01**	(*)	**0.19 ± 0.04**	(*)	**0.062 ± 0.03**	(*)
Manual	0.82 ± 0.02		0.23 ± 0.04		0.23 ± 0.06	

The estimated Gwet's AC between the three experts was 0.68 (standard error (SE): 0.10) for Worst, 0.68 (SE: 0.10) for Medium and 1 for Best segmentation classifications. According to Altman's benchmarking scale, the estimated coefficients for Worst and Medium segmentations are considered to be either *Moderate*, *Good* or *Very Good* with a probability of 0.99. The Best segmentation classification has a *Very good* agreement between the experts with a probability of 1. With a percentage agreement of 100 %, the consensus of experts classifies *TopoCP* as the Best segmentation method in 100% of the cases. Inter-rater discrepancies are observed in the choice of Worst and Medium between *Baseline* and *Hybrid* segmentation.

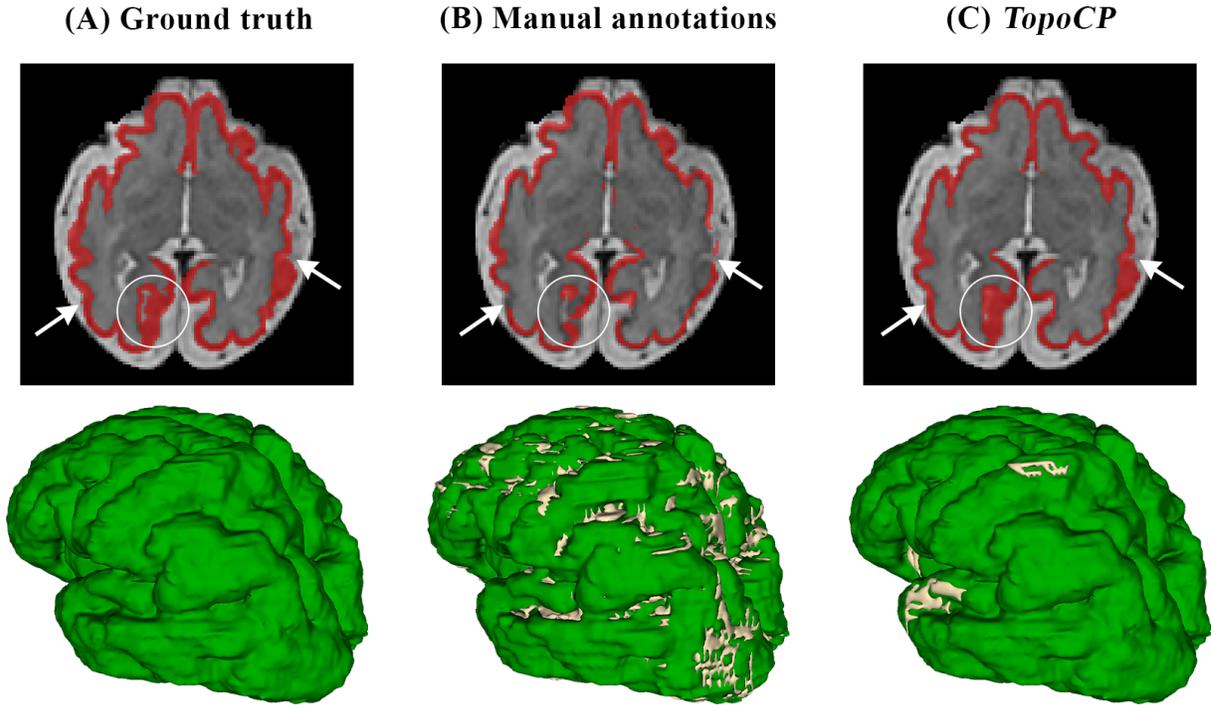**(A) Ground truth**   **(B) Manual annotations**   **(C) *TopoCP***

Figure 7.11: Qualitative 2D and 3D assessment of CP segmentation on a 31.5 weeks of GA neurotypical subject. Comparison of (A) the corrected GT to (B) the original manual annotation and (C) our *TopoCP* automatic segmentation method. Segmentations are overlaid on a T2w axial view (top). Segmentation 3D renderings (bottom) highlight the true positives (green) and false negatives connected to 1-dimensional holes (light red).

Although *TopoCP* is ranked as Best segmentation in all cases, predictions still present many segmentation errors. We emphasize that the distribution of this clinical set differs from the *FeTA* training set as they were generated with different SR methods. Nevertheless, while all three configurations present altered segmentation due to the domain shift, still, *TopoCP* remains the better performing method.

Table 7.6: Table of the three experts (two radiologists and one engineer) qualitative ranking of the three segmentation configurations (*Baseline*, *Hybrid* and *TopoCP*) as Best, Medium and Worst. Percentage agreement between experts and Gwet's AC with standard error (SE) are presented for each ranking category. Finally, a consensus ranking is presented in the bottom row as a majority voting of the experts' evaluation.

	Best			Medium			Worst		
	Baseline	*Hybrid*	*TopoCP*	*Baseline*	*Hybrid*	*TopoCP*	*Baseline*	*Hybrid*	*TopoCP*
Radiologist 1	0	0	31	25	6	0	6	25	0
Radiologist 2	0	0	31	26	5	0	5	26	0
Engineer	0	0	31	23	8	0	8	23	0
Percentage agreement		100 %			78 %			78 %	
Gwet's AC (SE)		1 (-)			0.68 (0.10)			0.68 (0.10)	
Consensus	0	0	**31**	26	5	0	5	26	0

7.5   Conclusion

In this work, we developed a topological loss function for the optimization of DL based segmentation methods of the fetal CP in MRI. Our core contribution lies in the multi-dimensional approach of this generalized loss function. Jointly, we presented an original topology-based metric to quantify the 1-dimensional topological errors both in terms of count and size. We presented extensive quantitative and qualitative validation on a total of 58 fetal brains of a wide range of GA (from 21 to 38 weeks of GA), including both neurotypical and pathological subjects. We compared our *TopoCP* method to (i) state-of-the-art methods and (ii) semi-automatic *noisy* reference segmentation. Experiments have shown that the integration of a topological constraint in the segmentation framework of the CP in fetal brain MRI significantly benefits not only the shape correctness – as it first aims, but also the overlap and distance accuracy. Although our segmentation framework is implemented for 2D image patches, 3D information is integrated thanks to the multi-view pipeline with the extraction of patches from the three orthogonal orientations (axial, coronal and sagittal). While our approach cannot be considered to be 3D, yet the benefit of our multi-dimensional topological loss is conveyed in the 3D metrics, including the topology-based one. Nevertheless, we believe that the adoption of a real 3D-based framework could only improve the overall performances, although we acknowledge that the computational cost of the topological loss in this process is an important shortcoming. Moreover, results evidence that the generalization of the learned topology is not hampered by the *noisiness* of the manual annotations used for training.

This study is the first to address both the specific improvement of the topological correctness of the CP segmentation, and the definition of a topological assessment. The reduced gap in the topological and shape correctness accuracy is ultimately associated with minimal manual refinement needed for further quantitative surface-based analysis. Future work will focus on a wider generalization of our method application. Indeed, while our method is formulated to consider multiple dimensions, we only present a 2D application. The overall framework can be generalized for a 3D-based approach. Similarly, we focused here on a single-tissue, namely the cGM segmentation, segmentation problem, although generalization to a multi-tissue segmentation approach could be applied.

8

Discussion

The success of the fetal brain MRI workflow (see Figure 8.1 (A)) lies in the accurate and robust quantitative analysis as it will ultimately condition the understanding of the brain development and, the clinical diagnosis and patient care. Therefore, each of the upstream steps of this well-defined sequence, i.e. the acquisition, the reconstruction and the segmentation, is crucial as the inaccurate achievement of any of these may propagate errors to all the subsequent ones.

In this context, this thesis addressed the development and validation of robust automated tools for the quantitative analysis of the fetal brain in MRI. Specifically, this work focused on the two main computer-assisted preprocessing steps: the generation of a good image through SR reconstruction, and the automatic semantic segmentation of the fetal brain tissues. However, in practice, the *data scarcity* of fetal brain MRI is a major obstacle for both the development and validation of such algorithms.

In this Chapter, we first summarize and discuss the contributions and limitations of this thesis to the existing pipeline (Figure 8.1 (B)). Second, we evoke some future perspectives of this work (Figure 8.1 (C)).

Summary of contributions

Super-resolution reconstruction

The accuracy of the SR reconstruction cannot be quantified as it is not possible to acquire a motion-free HR reference image. Nonetheless, we presented in Part I two studies that aimed at *(i)* validating that SR reconstruction preserves the brain morphology with no distortion and *(ii)* proposing an approach for optimizing the parameters of the SR reconstruction. In Chapter 2, our intra- and inter-observer agreement study on biometric measurements from both T2WS and SR reconstructed images highlighted that
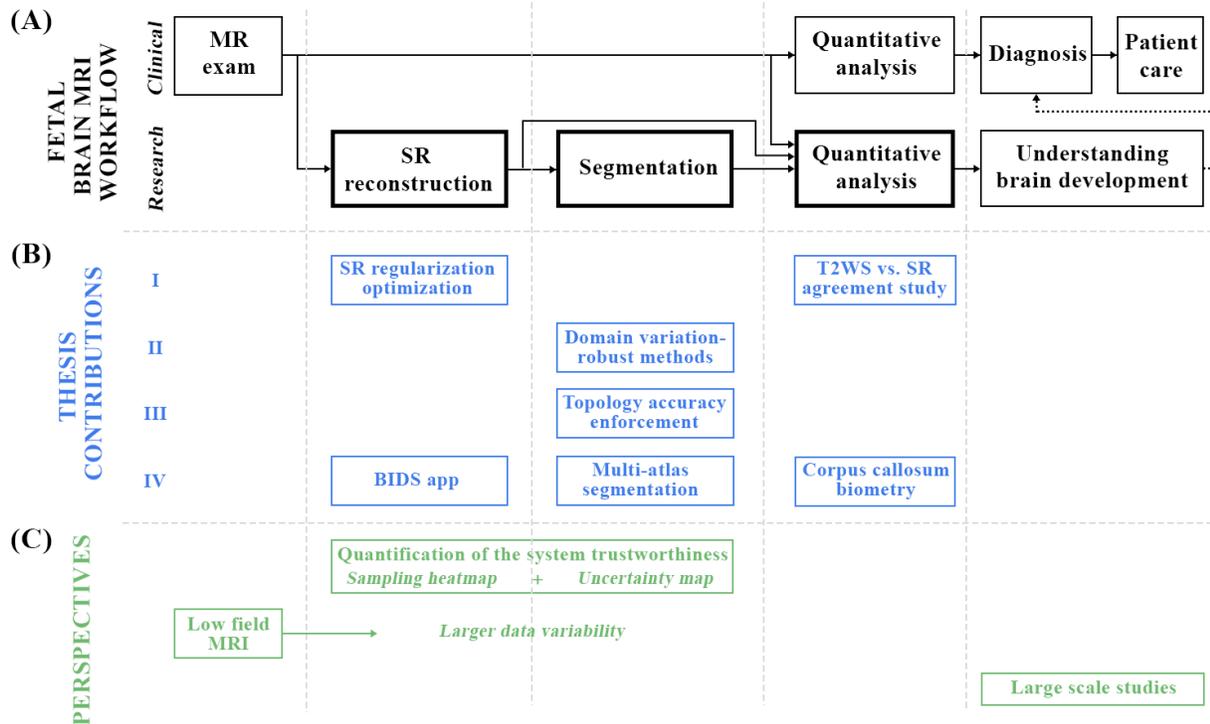
Figure 8.1: Overview of this thesis contributions (B) and future perspectives (C) with respect to the overall fetal brain MRI workflow (A).

SR does not introduce anatomical distortions and increases the confidence of the observer compared to T2WS [80]. However, good tissue contrast and delineation are essential for this task and can be achieved by finding a good compromise between the image smoothness and noisiness. Specifically, this is achieved by balancing the contribution of the data fidelity and the regularization terms in the resolution of the inverse problem (see Chapter 1, Equation 1.4). In Chapter 3, we proposed – in the absence of GT reference image – to adopt **a simulation-based approach for the optimization of the regularization weight**. Our principled approach allows an overall enhancement of the inter-tissue contrast.

Beyond the development, optimization and validation of such image reconstruction tools, translation to either the clinical or research environment is the last, not to underestimate, step. Over the time span of this thesis, we invested efforts towards the release of a BIDS App for the reconstruction of fetal brain MRI (see Appendix B). The establishment of such **open-source portable and reproducible tools** is a solid ground to possible large scale multi-centric analysis.

Fetal brain tissue segmentation

In the development of DL-based automatic tissue segmentation methods, pairs of (SR-reconstructed T2w image, GT annotation) are necessary to train and optimize models. However, neither type of data is widely available nor easy to obtain. FeTA [112] is the unique dataset of SR reconstructed clinical fetal brain MR exams that is publicly available. Additionally, the tediousness and the time consumption of manual labelling are major bottlenecks to the overall data scarcity problem.

As the principle of DL methods is to fit a data distribution, it is of the utmost importance to have a wide representation of the true data variability, in order to develop **variation-robust methods that adapt and generalize well across domains**. In Part II of this thesis, we proposed methods to mitigate two out of three sources of variations as *(i)* in Chapter 4, we replicated the whole imaging pipeline [91, 32] and *(ii)* in Chapter 5 we took advantage of the inherent variability of the last step of the pipeline, the reconstruction. With the generation of multiple HR volumes from a single subject, we evidenced that **the imaging pipeline can be leveraged to minimize the data scarcity limitation and further reduce the influence of the domain shift**. Additionally, we confirmed their data augmentation usage as multiple SR volumes of the same fetal brain can benefit the training of automatic segmentation methods.

Following this "one brain, multiple volumes" principle and extrapolating the usage of the imaging workflow inherent variations, we believe the complexity of fetal brain MRI segmentation could be further reduced. Specifically, without investing much efforts in the acquisition, reconstruction and annotation of new clinical fetal brain MR exams, **many more HR annotated fetal brains could be synthesized** by modifying the inputs and/or parameters of either the numerical simulator [91] (e.g. GA, MR sequences), the SR reconstruction pipeline (e.g. custom preprocessing, SR reconstruction parameters), or both.

As observed in the benchmark of automatic segmentation methods presented in Payette et *al.* [112], a drawback of DL-based methods is the lack of robustness to the input image quality, as all were outperformed by our MAS approach (Appendix C) when input images were considered bad. Similarly to the domain shift robustness that we addressed by increasing the representation of the variations, we propose to generate multiple reconstructions of the same brain using different combination of LR images. **We hypothesize that this larger heterogeneity of the data could benefit the model robustness to the SR image quality**, although this remains to be explored.

Besides the burden of SR reconstructed images diversity for the development of variation-robust methods, the burden of manually annotated images remains the major bottleneck to larger datasets. To overcome this limitation, optimized manual annotation protocols relying on the interpolation of sparse manual tissue annotations are used, even though **reference annotations may become unreliable**. In Part III of this thesis, we specifically focused on the topology correctness of the automatic segmentation methods. In Chapter 6, we showed the importance of considering the topology

correctness in the evaluation of recent segmentation methods that were submitted to the 2nd edition of the FeTA challenge (2022). In Chapter 7, we proposed **a method to enforce the topological correctness of the segmentation of the developing fetal cortical plate (CP)** throughout training. This method successfully yields to **topologically more consistent** segmentations, even when trained on data that presented topologically inconsistent labels due to the optimized procedure. Our proposed method was validated with an original metric *hole ratio (HoR)* that was defined as to **quantify the size of the topological error**.

Limitations

Despite the promising results observed throughout this thesis, some limitations still require further work.

Regarding SR reconstruction, while the assessment of the conservation of the brain morphology and the overall tissue delineation by SR reconstruction has been addressed, there is a lack of assessment of the intensity correctness. Beyond the use of SR reconstruction as a preprocessing for morphometric and volumetric analyses, the **conservation of subtle intensity variations is essential for clinical assessment** as it may correspond to an abnormal development (e.g. diffuse white matter abnormalities). Specifically, the many preprocessing and reconstruction-specific intensity-based operation may alter or even subside inhomogeneities. While the assessment of the intra- and inter-tissue contrast and the conservation of inhomogeneities would be needed to widely use SR in clinical environment, this remains unexplored so far.

On the side of segmentation, despite all the efforts invested by the community towards the development of better performing automatic segmentation methods, performances seem to reach a plateau [113]. Overall, the methods used in this thesis and in the FeTA challenges (2021 and 2022 editions) are direct adaptations of supervised DL strategies from previous biomedical imaging tasks. However, some challenges in fetal brain MRI segmentation are different. Specifically, **methods do not consider** the **evolving morphology and image intensities** throughout gestation, the increased **partial volume effect**, neither the **data scarcity** and the **lack of consistency in the reference annotations** that is more pronounced than in other applications. Additionally, the validation against noisy *ground truth* images hampers the true accurate assessment of the methods.

Topology consistency is also crucial for the study of cortical folding, a key biomarker of the brain development. So far, few works attempted to include topological priors in the tissue segmentation of the fetal brain. In this thesis, we integrated a topological constraint in a DL framework for the segmentation of the developing CP. However, our proposed method remains to be computationally optimized to be suitably extended either to multi-tissue segmentation or to 3D frameworks. Moreover, to fully validate our topology-integrative training strategies, a **clinical validation on subjects known**

to present cortical pathology remain to be explored. We also proposed a relevant new metric HoR for the assessment of the segmentation accuracy. However, its usage is limited by its strong dependency on the GT correctness. Alternatively, the topological defects can be counted, albeit with care.

Perspectives

Throughout the development of decision support tools based on automatic processing methods, the engineering point of view often adopts a top-down approach in which from the data available the aim is to generate results that are as accurate as possible. However, the original clinical motivation that is **the patient care must be kept in sight**. Therefore, beyond the overall accuracy and robustness of such decision support tools, one should ask: "*how much can we trust/rely on their outcomes?*".

In such a complex workflow, the trustworthiness of the system is especially challenging as a failure at any stage would propagate to any subsequent ones. Just as a human, our system can be wrong, but the acknowledgment of its uncertainty is crucial in order to prevent any downstream conclusion unwarranted by data. It is, therefore of decisive importance to **quantify the quality and the trustworthiness** – the (un-)certainty – of all the stages of our workflow. The need of such measures is amplified with the wide adoption of DL strategies.

In SR reconstruction, the quality and the aspect of the generated HR volume can be altered – globally or locally – in different ways. For instance, we know that SR is an ill-posed problem where the solution is recovered from under-sampled data. As illustrated in Figure 1.2, the acquisition scheme and the fetal motion may lead to under- or un-sampled regions. Additionally, cut field-of-view or discarded slices due to signal drops or motion artifact may amplify the phenomenon. Consequently, in the reconstructed volumes under-sampled areas would be estimated from interpolation of neighboring information rather than by consideration true acquired data. Similarly, despite the optional outliers rejection in SR reconstruction pipelines, misregistration may occur in the motion estimation step. In both cases, **uncertainty maps** could be generated as *(i)* a sampling heatmap that could be computed from the amount of data available at each voxel and/or *(ii)* an average registration similarity metric map from the motion estimation step. As a result, even though the overall SR image quality might be good, such maps could inform the clinician on the unreliable area and therefore prevent him from drawing conclusion or diagnosis based on localized and unsure observations.

Over the last years, many efforts have been invested in the field of medical imaging towards the quantification of uncertainty in the image semantic segmentation [101, 99]. Two types of uncertainty can be considered: the uncertainty related to the input data and the uncertainty related to the system. Typically, an estimation of the segmentation **system uncertainty** is provided as a voxel-wise map. In our workflow,

the **data-related uncertainty** of our segmentation block could be quantified from our SR reconstruction definition maps.

As for now, most fetal brain MR exams are performed either at 1.5T or 3T depending on the hardware available at each institution [18]. Nonetheless, **low-field MRI** (<1T) has recently aroused a growing interest thanks to its cost effectiveness and wider accessibility [126, 100]. However, lower field strengths lead to loss of SNR and image resolution (in-plane of about $1.5mm$ isotropic and slice thickness from 4 to $5mm$) [6]. Consequently, a **wider imaging data variability** will progressively appear in the field and, **adaptation and transfer of knowledge from methods at higher field strengths will be necessary**. Furthermore, the democratization of MRI technique will bring additional variations from the target population from a broader horizon, such as from developing countries. Despite the emergence of these new challenges, the wider access to low-field MRI will ease the possibility of **large scale studies** for the **general understanding of the human brain development**.

Overall, the on-going research and development in the field of fetal brain MRI are **promising towards the better understanding of *in utero* development**. Inevitably, ethical questions are raised about these methods and how their outcome may be used. Nonetheless, one should keep in mind that the ultimate goal of these works, notably the one presented in this thesis, is to ultimately **improve the patient care**, and here, **the patient is the baby**.

References

[1] M. Abadi et al. "TensorFlow: A system for large-scale machine learning". In: (2016). Publisher: arXiv Version Number: 2. DOI: *10.48550/ARXIV.1605.08695*.

[2] M. Aertsen et al. "Fetal MRI for dummies: what the fetal medicine specialist should know about acquisitions and sequences". en. In: *Prenatal Diagnosis* 40.1 (Jan. 2020), pp. 6–17. ISSN: 0197-3851, 1097-0223. DOI: *10.1002/pd.5579*.

[3] B. M. Afkham, J. Chung, and M. Chung. "Learning regularization parameters of inverse problems via deep neural networks". In: *Inverse Problems* 37.10 (Oct. 2021), p. 105017. ISSN: 0266-5611, 1361-6420. DOI: *10.1088/1361-6420/ac245d*.

[4] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de-Solorzano. "Combination Strategies in Multi-Atlas Image Segmentation: Application to Brain MR Data". In: *IEEE Transactions on Medical Imaging* 28.8 (Aug. 2009), pp. 1266–1277. ISSN: 0278-0062, 1558-254X. DOI: *10.1109/TMI.2009.2014372*.

[5] B. B. Avants et al. "A reproducible evaluation of ANTs similarity metric performance in brain image registration". eng. In: *NeuroImage* 54.3 (Feb. 2011), pp. 2033–2044. ISSN: 1095-9572. DOI: *10.1016/j.neuroimage.2010.09.025*.

[6] J. Aviles et al. "A Fast Anatomical and Quantitative MRI Fetal Exam at Low Field". en. In: *Perinatal, Preterm and Paediatric Image Analysis*. Ed. by R. Licandro et al. Vol. 13575. Series Title: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022, pp. 13–24. ISBN: 978-3-031-17116-1 978-3-031-17117-8. DOI: *10.1007/978-3-031-17117-8_2*.

[7] M. Bach Cuadra et al. "Brain tissue segmentation of fetal MR images". In: *Workshop on Image Analysis for Developing Brain, in 12th International Conference on Medical Image Computing and Computer Assisted Intervention*. 2009.

[8] A. J. Barkovich et al. "A developmental and genetic classification for malformations of cortical development". en. In: *Neurology* 65.12 (Dec. 2005), pp. 1873–1887. ISSN: 0028-3878, 1526-632X. DOI: *10.1212/01.wnl.0000183747.05269.2d*.

[9] R. Batty et al. "Analysis of errors made on in utero MR studies of the foetal brain in the MERIDIAN study". en. In: *European Radiology* 29.1 (Jan. 2019), pp. 195–201. ISSN: 0938-7994, 1432-1084. DOI: *10.1007/s00330-018-5508-x*.

[10] S. A. Bayer and J. Altman. "The Human Brain During the Third Trimester". In: *Atlas of Human Central Nervous System Development*. 2004. ISBN: 978-0-8493-1421-6.

[11] S. A. Bayer and J. Altman. "The Human Brain During The Second Trimester". In: *Atlas Of Human Central Nervous System Development*. CRC Press, 2005.

[12] O. M. Benkarim et al. "Cortical folding alterations in fetuses with isolated non-severe ventriculomegaly". en. In: *NeuroImage: Clinical* 18 (2018), pp. 103–114. ISSN: 22131582. DOI: *10.1016/j.nicl.2018.01.006*.

[13] C. Benvenuto-Andrade et al. "Level of Confidence in Diagnosis: Clinical Examination Versus Dermoscopy Examination". en. In: *Dermatologic Surgery* 32.5 (May 2006), pp. 738–744. ISSN: 1076-0512. DOI: *10.1097/00042728-200605000-00034*.

[14]   M. Cabezas et al. "A review of atlas-based segmentation for magnetic resonance brain images". en. In: *Computer Methods and Programs in Biomedicine* 104.3 (Dec. 2011), e158–e177. ISSN: 01692607. DOI: *10.1016/j.cmpb.2011.07.015*.

[15]   B. Caldairou et al. "Segmentation of the cortex in fetal MRI using a topological model". In: *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. Chicago, IL, USA: IEEE, Mar. 2011, pp. 2045–2048. ISBN: 978-1-4244-4127-3. DOI: *10.1109/ISBI.2011.5872814*.

[16]   A. M. Cardenas, M. T. Whitehead, and D. I. Bulas. "Fetal Neuroimaging Update". en. In: *Seminars in Pediatric Neurology* 33 (Apr. 2020), p. 100801. ISSN: 10719091. DOI: *10.1016/j.spen.2020.100801*.

[17]   M. J. Cardoso et al. "MONAI: An open-source framework for deep learning in healthcare". In: (2022). Publisher: arXiv Version Number: 1. DOI: *10.48550/ARXIV.2211.02701*.

[18]   M. Cassart and C. Garel. "European overview of current practice of fetal imaging by pediatric radiologists: a new task force is launched". en. In: *Pediatric Radiology* 50.12 (Nov. 2020), pp. 1794–1798. ISSN: 0301-0449, 1432-1998. DOI: *10.1007/s00247-020-04710-4*.

[19]   B. P. Chaudhari and M.-L. Ho. "Congenital Brain Malformations: An Integrated Diagnostic Approach". en. In: *Seminars in Pediatric Neurology* 42 (July 2022), p. 100973. ISSN: 10719091. DOI: *10.1016/j.spen.2022.100973*.

[20]   J. G. Chi, E. C. Dooling, and F. H. Gilles. "Gyral development of the human brain". en. In: *Annals of Neurology* 1.1 (Jan. 1977), pp. 86–93. ISSN: 0364-5134, 1531-8249. DOI: *10.1002/ana.410010109*.

[21]   C. Clouchoux et al. "Delayed Cortical Development in Fetuses with Complex Congenital Heart Disease". en. In: *Cerebral Cortex* 23.12 (Dec. 2013), pp. 2932–2943. ISSN: 1047-3211, 1460-2199. DOI: *10.1093/cercor/bhs281*.

[22]   C. Clouchoux et al. "Normative fetal brain growth by quantitative in vivo magnetic resonance imaging". en. In: *American Journal of Obstetrics and Gynecology* 206.2 (Feb. 2012), 173.e1–173.e8. ISSN: 00029378. DOI: *10.1016/j.ajog.2011.10.002*.

[23]   J. Clough et al. "A Topological Loss Function for Deep-Learning based Image Segmentation using Persistent Homology". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), pp. 1–1. ISSN: 0162-8828, 2160-9292, 1939-3539. DOI: *10.1109/TPAMI.2020.3013679*.

[24]   D. Cohen-Steiner et al. "Lipschitz Functions Have L p -Stable Persistence". en. In: *Foundations of Computational Mathematics* 10.2 (Apr. 2010), pp. 127–139. ISSN: 1615-3375, 1615-3383. DOI: *10.1007/s10208-010-9060-6*.

[25]   D. Connolly et al. "A Review of Fetal Brain Pathology Acquired In Utero". en. In: *Neurographics* 9.1 (Feb. 2019), pp. 79–89. ISSN: 2637-8329. DOI: *10.3174/ng.1800019*.

[26]   P. Deman et al. *meribach/mevislabFetalMRI: MEVISLAB MIAL Super-Resolution Reconstruction of Fetal Brain MRI v1.0*. June 2020. DOI: *10.5281/ZENODO.3878564*.

[27]   G. Developers. *GUDHI, Geometry Understanding in Higher Dimensions*.

[28]   L. R. Dice. "Measures of the Amount of Ecologic Association Between Species". en. In: *Ecology* 26.3 (July 1945), pp. 297–302. ISSN: 00129658. DOI: *10.2307/1932409*.

[29]   E. Dittrich et al. "A spatio-temporal latent atlas for semi-supervised learning of fetal brain segmentations and morphological age estimation". en. In: *Medical Image Analysis* 18.1 (Jan. 2014), pp. 9–21. ISSN: 13618415. DOI: *10.1016/j.media.2013.08.004*.

[30]   H. Dou et al. "A Deep Attentive Convolutional Neural Network for Automatic Cortical Plate Segmentation in Fetal MRI". In: *IEEE Transactions on Medical Imaging* 40.4 (Apr. 2021), pp. 1123–1133. ISSN: 0278-0062, 1558-254X. DOI: *10.1109/TMI.2020.3046579*.

[31]   J. Dubois et al. "The early development of brain white matter: A review of imaging studies in fetuses, newborns and infants". en. In: *Neuroscience* 276 (Sept. 2014), pp. 48–71. ISSN: 03064522. DOI: *10.1016/j.neuroscience.2013.12.044*.

[32] P. de Dumast et al. "Synthetic Magnetic Resonance Images for Domain Adaptation: Application to Fetal Brain Tissue Segmentation". In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. Kolkata, India: IEEE, Mar. 2022, pp. 1–5. ISBN: 978-1-66542-923-8. DOI: *10.1109/ ISBI52829.2022.9761451*.

[33] M. Ebner et al. "An automated framework for localization, segmentation and super-resolution reconstruction of fetal brain MRI". en. In: *NeuroImage* 206 (Feb. 2020), p. 116324. ISSN: 10538119. DOI: *10.1016/j.neuroimage.2019.116324*.

[34] G. Egaña-Ugrinovic et al. "Differences in cortical development assessed by fetal MRI in late-onset intrauterine growth restriction". en. In: *American Journal of Obstetrics and Gynecology* 209.2 (Aug. 2013), 126.e1–126.e8. ISSN: 00029378. DOI: *10.1016/j.ajog.2013.04.008*.

[35] G. Faa et al. "Fetal programming of the human brain: is there a link with insurgence of neurode-generative disorders in adulthood?" eng. In: *Current Medicinal Chemistry* 21.33 (2014), pp. 3854–3876. ISSN: 1875-533X. DOI: *10.2174/0929867321666140601163658*.

[36] G. Faa et al. "Fetal programming of neuropsychiatric disorders: Fetal Programming". en. In: *Birth Defects Research Part C: Embryo Today: Reviews* 108.3 (Sept. 2016), pp. 207–223. ISSN: 1542975X. DOI: *10.1002/bdrc.21139*.

[37] E. Ferrante and N. Paragios. "Slice-to-volume medical image registration: A survey". en. In: *Medical Image Analysis* 39 (July 2017), pp. 101–123. ISSN: 13618415. DOI: *10.1016/j.media.2017.04.010*.

[38] A. Fetit et al. "A deep learning approach to segmentation of the developing cortex in fetal brain MRI with minimal manual labeling." In: PMLR, 2020, pp. 241–261.

[39] L. Fidon et al. "Label-Set Loss Functions for Partial Supervision: Application to Fetal Brain 3D MRI Parcellation". en. In: *Medical Image Computing and Computer Assisted Intervention MICCAI 2021*. Vol. 12902. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 647–657. ISBN: 978-3-030-87195-6 978-3-030-87196-3. DOI: *10.1007/978-3-030-87196-3_60*.

[40] L. Fidon et al. *Label-set Loss Functions for Partial Supervision: Application to Fetal Brain 3D MRI Parcellation*. Type: dataset. July 2021. DOI: *10.5281/ZENODO.7358931*.

[41] L. Fidon et al. "A spatio-temporal atlas of the developing fetal brain with spina bifida aperta". en. In: *Open Research Europe* 1 (Aug. 2022), p. 123. ISSN: 2732-5121. DOI: *10.12688/openreseurope. 13914.2*.

[42] M. Fogtmann et al. "A unified approach for motion estimation and super resolution reconstruction from structural Magnetic Resonance imaging on moving subjects". In: *MICCAI workshop on Perinatal and Paediatric Imaging: PaPI*. 2012.

[43] N. Galatsanos and A. Katsaggelos. "Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation". In: *IEEE Transactions on Image Processing* 1.3 (July 1992), pp. 322–336. ISSN: 10577149. DOI: *10.1109/83.148606*.

[44] K. E. Garcia, C. D. Kroenke, and P. V. Bayly. "Mechanics of cortical folding: stress, growth and stability". en. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1759 (Nov. 2018), p. 20170321. ISSN: 0962-8436, 1471-2970. DOI: *10.1098/rstb.2017.0321*.

[45] C. Garel. "Fetal cerebral biometry: normal parenchymal findings and ventricular size". en. In: *European Radiology* 15.4 (Apr. 2005), pp. 809–813. ISSN: 0938-7994, 1432-1084. DOI: *10.1007/s00330-004-2610-z*.

[46] C. Garel. *Le développement du cerveau foetal: atlas IRM et biométrie*. French. OCLC: 62539824. Montpellier: Sauramps médical, 2000. ISBN: 978-2-84023-245-2.

[47] C. Garel. *MRI of the Fetal Brain*. en. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. ISBN: 978-3-642-62275-5 978-3-642-18747-6. DOI: *10.1007/978-3-642-18747-6*.

[48] C. GAREL, M. CASSART, and H. NAHUM. *Imagerie du foetus au nouveau-né*. fre. OCLC: 1267761165. Cachan: Flammarion medecine-sciences, 2015. ISBN: 978-2-257-70651-5.

[49] C. Garel et al. "Fetal MRI: normal gestational landmarks for cerebral biometry, gyration and myelination". In: *Child's Nervous System* 19.7-8 (Aug. 2003), pp. 422–425. ISSN: 0256-7040, 1433-0350. DOI: *10.1007/s00381-003-0767-4*.

[50] C. Garel and M. Cassart, eds. *Imagerie du ftus au nouveau-né*. fre. OCLC: 1013590399. Paris: Lavoisier médecine sciences, 2016. ISBN: 978-2-257-20651-0.

[51] M. Ghafoorian et al. "Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation". en. In: *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*. Vol. 10435. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 516–524. ISBN: 978-3-319-66178-0 978-3-319-66179-7. DOI: *10.1007/978-3-319-66179-7_59*.

[52] A. Gholipour, J. A. Estroff, and S. K. Warfield. "Robust Super-Resolution Volume Reconstruction From Slice Acquisitions: Application to Fetal Brain MRI". In: *IEEE Transactions on Medical Imaging* 29.10 (Oct. 2010), pp. 1739–1758. ISSN: 0278-0062, 1558-254X. DOI: *10.1109/TMI.2010.2051680*.

[53] A. Gholipour et al. "Fetal brain volumetry through MRI volumetric reconstruction and segmentation". en. In: *International Journal of Computer Assisted Radiology and Surgery* 6.3 (May 2011), pp. 329–339. ISSN: 1861-6410, 1861-6429. DOI: *10.1007/s11548-010-0512-x*.

[54] A. Gholipour et al. "Multi-atlas multi-shape segmentation of fetal brain MRI for volumetric and morphometric analysis of ventriculomegaly". en. In: *NeuroImage* 60.3 (Apr. 2012), pp. 1819–1831. ISSN: 10538119. DOI: *10.1016/j.neuroimage.2012.01.128*.

[55] A. Gholipour et al. "Fetal MRI: A Technical Update with Educational Aspirations". eng. In: *Concepts in Magnetic Resonance. Part A, Bridging Education and Research* 43.6 (Nov. 2014), pp. 237–266. ISSN: 1546-6086. DOI: *10.1002/cmr.a.21321*.

[56] A. Gholipour et al. "A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth". en. In: *Scientific Reports* 7.1 (Dec. 2017), p. 476. ISSN: 2045-2322. DOI: *10.1038/s41598-017-00525-w*.

[57] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. Adaptive computation and machine learning. Cambridge, Massachusetts: The MIT Press, 2016. ISBN: 978-0-262-03561-3.

[58] H. Greenspan. "Super-Resolution in Medical Imaging". en. In: *The Computer Journal* 52.1 (Feb. 2008), pp. 43–63. ISSN: 0010-4620, 1460-2067. DOI: *10.1093/comjnl/bxm075*.

[59] P. D. Griffiths et al. "Anatomical subgroup analysis of the MERIDIAN cohort: failed commissuration". en. In: *Ultrasound in Obstetrics & Gynecology* 50.6 (Dec. 2017), pp. 753–760. ISSN: 09607692. DOI: *10.1002/uog.17502*.

[60] P. Griffiths et al. "Change in diagnostic confidence brought about by using in utero MRI for fetal structural brain pathology: analysis of the MERIDIAN cohort". en. In: *Clinical Radiology* 72.6 (June 2017), pp. 451–457. ISSN: 00099260. DOI: *10.1016/j.crad.2017.01.009*.

[61] R. Grossman et al. "Quantitative MRI measurements of human fetal brain development in utero". en. In: *NeuroImage* 33.2 (Nov. 2006), pp. 463–470. ISSN: 10538119. DOI: *10.1016/j.neuroimage.2006.07.005*.

[62] L. Guibaud and A. Lacalm. "Diagnostic imaging tools to elucidate decreased cephalic biometry and fetal microcephaly: a systematic analysis of the central nervous system: Editorial". en. In: *Ultrasound in Obstetrics & Gynecology* 48.1 (July 2016), pp. 16–25. ISSN: 09607692. DOI: *10.1002/uog.15926*.

[63] K. L. Gwet. *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters*. eng. Fourth edition. Gaithersburg, Md: Advances Analytics, LLC, 2014. ISBN: 978-0-9708062-8-4.

[64] P. A. Habas et al. "Atlas-Based Segmentation of the Germinal Matrix from in Utero Clinical MRI of the Fetal Brain". In: *Medical Image Computing and Computer-Assisted Intervention MICCAI 2008*. Vol. 5241. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 351–358. ISBN: 978-3-540-85987-1 978-3-540-85988-8. DOI: *10.1007/978-3-540-85988-8_42*.

[65] B. J. al-Haddad et al. "The fetal origins of mental illness". en. In: *American Journal of Obstetrics and Gynecology* 221.6 (Dec. 2019), pp. 549–562. ISSN: 00029378. DOI: *10.1016/j.ajog.2019.06.013*.

[66] N. Hahner et al. "Global and Regional Changes in Cortical Development Assessed by MRI in Fetuses with Isolated Nonsevere Ventriculomegaly Correlate with Neonatal Neurobehavior". en. In: *American Journal of Neuroradiology* (Aug. 2019), ajnr, ajnr.A6165v1. ISSN: 0195-6108, 1936-959X. DOI: *10.3174/ajnr.A6165*.

[67] M. Hasegawa et al. "Development of myelination in the human fetal and infant cerebrum: A myelin basic protein immunohistochemical study". en. In: *Brain and Development* 14.1 (Jan. 1992), pp. 1–6. ISSN: 03877604. DOI: *10.1016/S0387-7604(12)80271-3*.

[68] M. Hollander, D. A. Wolfe, and E. Chicken. *Nonparametric statistical methods*. Third edition. Hoboken, New Jersey: John Wiley & Sons, Inc, 2014. ISBN: 978-0-470-38737-5.

[69] J. Hong et al. "Fetal Cortical Plate Segmentation Using Fully Convolutional Networks With Multiple Plane Aggregation". eng. In: *Frontiers in Neuroscience* 14 (2020), p. 591683. ISSN: 1662-4548. DOI: *10.3389/fnins.2020.591683*.

[70] B. Hou et al. "3-D Reconstruction in Canonical Co-Ordinate Space From Arbitrarily Oriented 2-D Images". In: *IEEE Transactions on Medical Imaging* 37.8 (Aug. 2018), pp. 1737–1750. ISSN: 0278-0062, 1558-254X. DOI: *10.1109/TMI.2018.2798801*.

[71] X. Hu et al. "Topology-Preserving Deep Image Segmentation". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.

[72] A. M. Hubbard, M. P. Harty, and L. J. States. "A new tool for prenatal diagnosis: Ultrafast fetal MRI". en. In: *Seminars in Perinatology* 23.6 (Dec. 1999), pp. 437–447. ISSN: 01460005. DOI: *10.1016/S0146-0005(99)80023-8*.

[73] J. E. Iglesias and M. R. Sabuncu. "Multi-atlas segmentation of biomedical images: A survey". en. In: *Medical Image Analysis* 24.1 (Aug. 2015), pp. 205–219. ISSN: 13618415. DOI: *10.1016/j.media.2015.06.012*.

[74] K. Im et al. "Quantitative Folding Pattern Analysis of Early Primary Sulci in Human Fetuses with Brain Abnormalities". en. In: *American Journal of Neuroradiology* 38.7 (July 2017), pp. 1449–1455. ISSN: 0195-6108, 1936-959X. DOI: *10.3174/ajnr.A5217*.

[75] K. Im et al. "Quantification and discrimination of abnormal sulcal patterns in polymicrogyria". eng. In: *Cerebral Cortex (New York, N.Y.: 1991)* 23.12 (Dec. 2013), pp. 3007–3015. ISSN: 1460-2199. DOI: *10.1093/cercor/bhs292*.

[76] S. Jiang et al. "MRI of Moving Subjects Using Multislice Snapshot Images With Volume Reconstruction (SVR): Application to Fetal, Neonatal, and Adult Brain Studies". In: *IEEE Transactions on Medical Imaging* 26.7 (July 2007), pp. 967–980. ISSN: 0278-0062. DOI: *10.1109/TMI.2007.895456*.

[77] B. Kainz et al. "Fast Volume Reconstruction From Motion Corrupted Stacks of 2D Slices". In: *IEEE Transactions on Medical Imaging* 34.9 (Sept. 2015), pp. 1901–1913. ISSN: 0278-0062, 1558-254X. DOI: *10.1109/TMI.2015.2415453*.

[78] K. Keunen, S. J. Counsell, and M. J. Benders. "The emergence of functional architecture during early brain development". en. In: *NeuroImage* 160 (Oct. 2017), pp. 2–14. ISSN: 10538119. DOI: *10.1016/j.neuroimage.2017.01.047*.

[79] N. Khalili et al. "Automatic brain tissue segmentation in fetal MRI using convolutional neural networks". en. In: *Magnetic Resonance Imaging* 64 (Dec. 2019), pp. 77–89. ISSN: 0730725X. DOI: *10.1016/j.mri.2019.05.020*.

[80] M. Khawam et al. "Fetal Brain Biometric Measurements on 3D Super-Resolution Reconstructed T2-Weighted MRI: An Intra- and Inter-observer Agreement Study". In: *Frontiers in Pediatrics* 9 (Aug. 2021), p. 639746. ISSN: 2296-2360. DOI: *10.3389/fped.2021.639746*.

[81] H. E. Kim et al. "Transfer learning for medical image classification: a literature review". en. In: *BMC Medical Imaging* 22.1 (Dec. 2022), p. 69. ISSN: 1471-2342. DOI: *10.1186/s12880-022-00793-7*.

[82] K. Kim et al. "Intersection Based Motion Correction of Multislice MRI for 3-D *in Utero* Fetal Brain Image Formation". In: *IEEE Transactions on Medical Imaging* 29.1 (Jan. 2010), pp. 146–158. ISSN: 0278-0062, 1558-254X. DOI: *10.1109/TMI.2009.2030679*.

[83] B. M. Kline-Fath. "Ultrasound and MR Imaging of the Normal Fetal Brain". en. In: *Neuroimaging Clinics of North America* 29.3 (Aug. 2019), pp. 339–356. ISSN: 10525149. DOI: *10.1016/j.nic.2019.03.001*.

[84] M. S. van der Knaap and J. Valk. "Classification of congenital abnormalities of the CNS". eng. In: *AJNR. American journal of neuroradiology* 9.2 (Apr. 1988), pp. 315–326. ISSN: 0195-6108.

[85] B. Kolb and B. D. Fantie. "Development of the Childs Brain and Behavior". en. In: *Handbook of Clinical Child Neuropsychology*. Ed. by C. R. Reynolds and E. Fletcher-Janzen. Boston, MA: Springer US, 2009, pp. 19–46. ISBN: 978-0-387-70708-2 978-0-387-78867-8. DOI: *10.1007/978-0-387-78867-8_2*.

[86] I. Kostovic, G. Sedmak, and M. Juda. "Neural histology and neurogenesis of the human fetal and infant brain". en. In: *NeuroImage* 188 (Mar. 2019), pp. 743–773. ISSN: 10538119. DOI: *10.1016/j.neuroimage.2018.12.043*.

[87] I. Kostovic and L. Vasung. "Insights From In Vitro Fetal Magnetic Resonance Imaging of Cerebral Development". en. In: *Seminars in Perinatology* 33.4 (Aug. 2009), pp. 220–233. ISSN: 01460005. DOI: *10.1053/j.semperi.2009.04.003*.

[88] M. Kuklisova-Murgasova et al. "Reconstruction of fetal brain MRI with intensity matching and complete outlier removal". en. In: *Medical Image Analysis* 16.8 (Dec. 2012), pp. 1550–1564. ISSN: 13618415. DOI: *10.1016/j.media.2012.07.004*.

[89] V. Kyriakopoulou et al. "Normative biometry of the fetal brain using magnetic resonance imaging". en. In: *Brain Structure and Function* 222.5 (July 2017), pp. 2295–2307. ISSN: 1863-2653, 1863-2661. DOI: *10.1007/s00429-016-1342-6*.

[90] H. Lajous et al. *Medical-Image-Analysis-Laboratory/FaBiAN: FaBiAN v1.1*. Oct. 2021. DOI: *10.5281/ZENODO.5599311*.

[91] H. Lajous et al. "A Fetal Brain magnetic resonance Acquisition Numerical phantom (FaBiAN)". en. In: *Scientific Reports* 12.1 (May 2022), p. 8682. ISSN: 2045-2322. DOI: *10.1038/s41598-022-10335-4*.

[92] Z. Leibovitz, T. Lerman-Sagie, and L. Haddad. "Fetal Brain Development: Regulating Processes and Related Malformations". en. In: *Life* 12.6 (May 2022), p. 809. ISSN: 2075-1729. DOI: *10.3390/life12060809*.

[93] R. K. Lenroot and J. N. Giedd. "Brain development in children and adolescents: Insights from anatomical magnetic resonance imaging". en. In: *Neuroscience & Biobehavioral Reviews* 30.6 (Jan. 2006), pp. 718–729. ISSN: 01497634. DOI: *10.1016/j.neubiorev.2006.06.001*.

[94] R. J. Leventer, R. Guerrini, and W. B. Dobyns. "Malformations of cortical development and epilepsy". eng. In: *Dialogues in Clinical Neuroscience* 10.1 (2008), pp. 47–62. ISSN: 1294-8322.

[95] L. I.-K. Lin. "A Concordance Correlation Coefficient to Evaluate Reproducibility". In: *Biometrics* 45.1 (Mar. 1989), p. 255. ISSN: 0006341X. DOI: *10.2307/2532051*.

[96] Z. Lin and Heung-Yeung Shum. "Fundamental limits of reconstruction-based superresolution algorithms under local translation". en. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.1 (Jan. 2004), pp. 83–97. ISSN: 0162-8828. DOI: *10.1109/TPAMI.2004.1261081*.

[97] L. Maier-Hein et al. "Metrics reloaded: Pitfalls and recommendations for image analysis validation". In: (2022). Publisher: arXiv Version Number: 1. DOI: *10.48550/ARXIV.2206.01653*.

[98] A. Makropoulos, S. J. Counsell, and D. Rueckert. "A review on automatic fetal and neonatal brain MRI segmentation". eng. In: *NeuroImage* 170 (Apr. 2018), pp. 231–248. ISSN: 1095-9572. DOI: *10.1016/j.neuroimage.2017.06.074*.

[99] A. Malinin et al. *Shifts 2.0: Extending The Dataset of Real Distributional Shifts*. arXiv:2206.15407 [cs, stat]. Sept. 2022.

[100] J. P. Marques, F. F. Simonis, and A. G. Webb. "Lowfield MRI: An MR physics perspective". en. In: *Journal of Magnetic Resonance Imaging* 49.6 (June 2019), pp. 1528–1542. ISSN: 1053-1807, 1522-2586. DOI: *10.1002/jmri.26637*.

[101] B. Menze et al. "Quantification of Uncertainties in Biomedical Image Quantification". In: (Mar. 2020). Publisher: Zenodo. DOI: *10.5281/ZENODO.3718911*.

[102] B. Mildenhall et al. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis". en. In: *Computer Vision ECCV 2020*. Ed. by A. Vedaldi et al. Vol. 12346. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 405–421. ISBN: 978-3-030-58451-1 978-3-030-58452-8. DOI: *10.1007/978-3-030-58452-8_24*.

[103] MONAI Consortium. *MONAI: Medical Open Network for AI*. June 2022. DOI: *10.5281/ZENODO.6639453*.

[104] C. Ng and C. Palmer. "Analysis of diagnostic confidence and diagnostic accuracy: a unified framework". en. In: *The British Journal of Radiology* 80.951 (Mar. 2007), pp. 152–160. ISSN: 0007-1285, 1748-880X. DOI: *10.1259/bjr/64096611*.

[105] M. Orbes-Arteaga et al. "Augmentation based unsupervised domain adaptation". In: (2022). Publisher: arXiv Version Number: 1. DOI: *10.48550/ARXIV.2202.11486*.

[106] C. M. Ortinau et al. "Early-Emerging Sulcal Patterns Are Atypical in Fetuses with Congenital Heart Disease". en. In: *Cerebral Cortex* 29.8 (July 2019), pp. 3605–3616. ISSN: 1047-3211, 1460-2199. DOI: *10.1093/cercor/bhy235*.

[107] C. Ouyang et al. "Causality-inspired Single-source Domain Generalization for Medical Image Segmentation". In: *IEEE Transactions on Medical Imaging* (2022), pp. 1–1. ISSN: 0278-0062, 1558-254X. DOI: *10.1109/TMI.2022.3224067*.

[108] M. Ouyang et al. "Delineation of early brain development from fetuses to infants with diffusion MRI and beyond". en. In: *NeuroImage* 185 (Jan. 2019), pp. 836–850. ISSN: 10538119. DOI: *10.1016/j.neuroimage.2018.04.017*.

[109] S. Pashaj, E. Merz, and S. Wellek. "Biometry of the fetal corpus callosum by three-dimensional ultrasound: Biometry of the fetal corpus callosum". en. In: *Ultrasound in Obstetrics & Gynecology* 42.6 (Dec. 2013), pp. 691–698. ISSN: 09607692. DOI: *10.1002/uog.12501*.

[110] L. K. Paul et al. "Agenesis of the corpus callosum: genetic, developmental and functional aspects of connectivity". en. In: *Nature Reviews Neuroscience* 8.4 (Apr. 2007), pp. 287–299. ISSN: 1471-003X, 1471-0048. DOI: *10.1038/nrn2107*.

[111] K. Payette, R. Kottke, and A. Jakab. "Efficient Multi-class Fetal Brain Segmentation in High Resolution MRI Reconstructions with Noisy Labels". en. In: *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*. Vol. 12437. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 295–304. ISBN: 978-3-030-60333-5 978-3-030-60334-2. DOI: *10.1007/978-3-030-60334-2_29*.

[112] K. Payette et al. "An automatic multi-tissue human fetal brain segmentation benchmark using the Fetal Tissue Annotation Dataset". en. In: *Scientific Data* 8.1 (Dec. 2021), p. 167. ISSN: 2052-4463. DOI: *10.1038/s41597-021-00946-3*.

[113] K. Payette et al. "Fetal Brain Tissue Annotation and Segmentation Challenge Results". In: (2022). Publisher: arXiv Version Number: 1. DOI: *10.48550/ARXIV.2204.09573*.

[114] F. Pérez-García, R. Sparks, and S. Ourselin. "TorchIO: A Python library for efficient loading, pre-processing, augmentation and patch-based sampling of medical images in deep learning". en. In: *Computer Methods and Programs in Biomedicine* 208 (Sept. 2021), p. 106236. ISSN: 01692607. DOI: *10.1016/j.cmpb.2021.106236*.

[115] D. B. Pier et al. "3D Super-Resolution Motion-Corrected MRI: Validation of Fetal Posterior Fossa Measurements: Super-Resolution Motion-Corrected MRI of the Fetal Posterior Fossa". en. In: *Journal of Neuroimaging* 26.5 (Sept. 2016), pp. 539–544. ISSN: 10512284. DOI: *10.1111/jon.12342*.

[116] L. R. Pistorius et al. "Fetal Neuroimaging: Ultrasound, MRI, or Both?:" en. In: *Obstetrical & Gynecological Survey* 63.11 (Nov. 2008), pp. 733–745. ISSN: 0029-7828. DOI: *10.1097/OGX.0b013e318186d3ea*.

[117] L. Pomar et al. "Comparison between Two-Dimensional and Three-Dimensional Assessments of the Fetal Corpus Callosum: Reproducibility of Measurements and Acquisition Time". en. In: *Journal of Pediatric Neurology* 19.05 (Oct. 2021), pp. 312–320. ISSN: 1304-2580, 1875-9041. DOI: *10.1055/s-0041-1722854*.

[118] D. Prayer et al. "ISUOG Practice Guidelines: performance of fetal magnetic resonance imaging". en. In: *Ultrasound in Obstetrics & Gynecology* 49.5 (May 2017), pp. 671–680. ISSN: 09607692. DOI: *10.1002/uog.17412*.

[119] D. Prayer, P. Brugger, and L. Prayer. "Fetal MRI: techniques and protocols". en. In: *Pediatric Radiology* 34.9 (Sept. 2004). ISSN: 0301-0449, 1432-1998. DOI: *10.1007/s00247-004-1246-0*.

[120] V. Rajagopalan et al. "Local tissue growth patterns underlying normal fetal human brain gyrification quantified in utero". eng. In: *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 31.8 (Feb. 2011), pp. 2878–2887. ISSN: 1529-2401. DOI: *10.1523/JNEUROSCI.5458-10.2011*.

[121] L. Ronan and P. C. Fletcher. "From genes to folds: a review of cortical gyrification theory". en. In: *Brain Structure and Function* 220.5 (Sept. 2015), pp. 2475–2483. ISSN: 1863-2653, 1863-2661. DOI: *10.1007/s00429-014-0961-z*.

[122] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". en. In: *Medical Image Computing and Computer-Assisted Intervention MICCAI 2015*. Vol. 9351. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24573-7 978-3-319-24574-4. DOI: *10.1007/978-3-319-24574-4_28*.

[123] G. Rote and G. Vegter. "Computational Topology: An Introduction". en. In: *Effective Computational Geometry for Curves and Surfaces*. Ed. by J.-D. Boissonnat and M. Teillaud. Springer Berlin Heidelberg, 2006, pp. 277–312. ISBN: 978-3-540-33258-9. DOI: *10.1007/978-3-540-33259-6_7*.

[124] F. Rousseau et al. "On super-resolution for fetal brain MRI". eng. In: *Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* 13.Pt 2 (2010), pp. 355–362. DOI: *10.1007/978-3-642-15745-5_44*.

[125] F. Rousseau et al. "Registration-Based Approach for Reconstruction of High-Resolution In Utero Fetal MR Brain Images". en. In: *Academic Radiology* 13.9 (Sept. 2006), pp. 1072–1081. ISSN: 10766332. DOI: *10.1016/j.acra.2006.05.003*.

[126] M. Sarracanie et al. "Low-Cost High-Performance MRI". en. In: *Scientific Reports* 5.1 (Oct. 2015), p. 15177. ISSN: 2045-2322. DOI: *10.1038/srep15177*.

[127] H. C. Sauerwein and M. Lassonde. "Cognitive and sensori-motor functioning in the absence of the corpus callosum: Neuropsychological studies in callosal agenesis and callosotomized patients". en. In: *Behavioural Brain Research* 64.1-2 (Oct. 1994), pp. 229–240. ISSN: 01664328. DOI: *10.1016/0166-4328(94)90135-X*.

[128] W. Sepulveda et al. "Fetal magnetic resonance imaging and three-dimensional ultrasound in clinical practice: General aspects". en. In: *Best Practice & Research Clinical Obstetrics & Gynaecology* 26.5 (Oct. 2012), pp. 575–591. ISSN: 15216934. DOI: *10.1016/j.bpobgyn.2012.05.010*.

[129] A. Serag et al. "A Multi-channel 4D Probabilistic Atlas of the Developing Brain: Application to Fetuses and Neonates". In: 2012 (2012), pp. 1–14.

[130] A. Serag et al. "Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression". en. In: *NeuroImage* 59.3 (Feb. 2012), pp. 2255–2265. ISSN: 10538119. DOI: *10.1016/j.neuroimage.2011.09.062*.

[131] M. Severino et al. "Definitions and classification of malformations of cortical development: practical guidelines". en. In: *Brain* 143.10 (Oct. 2020), pp. 2874–2894. ISSN: 0006-8950, 1460-2156. DOI: *10.1093/brain/awaa174*.

[132] V. Siffredi et al. "A Neuropsychological Profile for Agenesis of the Corpus Callosum? Cognitive, Academic, Executive, Social, and Behavioral Functioning in School-Age Children". en. In: *Journal of the International Neuropsychological Society* 24.5 (May 2018), pp. 445–455. ISSN: 1355-6177, 1469-7661. DOI: *10.1017/S1355617717001357*.

[133] F. Smith. "NMR IMAGING IN PREGNANCY". en. In: *The Lancet* 321.8314-8315 (Jan. 1983), pp. 61–62. ISSN: 01406736. DOI: *10.1016/S0140-6736(83)91588-X*.

[134] K. Suzuki. "Overview of deep learning in medical imaging". en. In: *Radiological Physics and Technology* 10.3 (Sept. 2017), pp. 257–273. ISSN: 1865-0333, 1865-0341. DOI: *10.1007/s12194-017-0406-5*.

[135] *SVRTK: SVR reconstruction toolkit for fetal MRI.* 2022.

[136] P. J. Switzer, C. A. James, and M. A. Frettag. "Value and Limitations of Obstetrical Ultrasound: Uncovering abnormalities at earlier stages". eng. In: *Canadian Family Physician Medecin De Famille Canadien* 38 (Jan. 1992), pp. 121–128. ISSN: 0008-350X.

[137] A. A. Taha and A. Hanbury. "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool". en. In: *BMC Medical Imaging* 15.1 (Dec. 2015), p. 29. ISSN: 1471-2342. DOI: *10.1186/s12880-015-0068-x*.

[138] T. Tarui et al. "Disorganized Patterns of Sulcal Position in Fetal Brains with Agenesis of Corpus Callosum". eng. In: *Cerebral Cortex (New York, N.Y.: 1991)* 28.9 (Sept. 2018), pp. 3192–3203. ISSN: 1460-2199. DOI: *10.1093/cercor/bhx191*.

[139] H. E. Thompson. "THE CLINICAL USE OF PULSED ECHO ULTRASOUND IN OBSTETRICS AND GYNECOLOGY:" en. In: *Obstetrical & Gynecological Survey* 23.10 (Oct. 1968), pp. 903–932. ISSN: 0029-7828. DOI: *10.1097/00006254-196810000-00001*.

[140] F. Tiedemann. *Anatomie und Bildungsgeschichte des Gehirns im Foetus des Menschen nebst einer vergleichenden Darstellung des Hirnbaues in den Thieren.* Steinisch, 1816.

[141] A. L. Tierney and C. A. Nelson. "Brain Development and the Role of Experience in the Early Years". eng. In: *Zero to Three* 30.2 (Nov. 2009), pp. 9–13. ISSN: 0736-8038.

[142] B. Tilea et al. "Cerebral biometry in fetal magnetic resonance imaging: new reference data". en. In: *Ultrasound in Obstetrics and Gynecology* 33.2 (Feb. 2009), pp. 173–181. ISSN: 09607692, 14690705. DOI: *10.1002/uog.6276*.

[143] S. Tourbier et al. *Medical-Image-Analysis-Laboratory/mialsuperresolutiontoolkit: MIAL Super-Resolution Toolkit v2.0.1.* Dec. 2020. DOI: *10.5281/ZENODO.4392788*.

[144] S. Tourbier et al. "An efficient total variation algorithm for super-resolution in fetal brain MRI with adaptive regularization". en. In: *NeuroImage* 118 (Sept. 2015), pp. 584–597. ISSN: 10538119. DOI: *10.1016/j.neuroimage.2015.06.018*.

[145] A. U. Uus et al. "Retrospective motion correction in foetal MRI for clinical applications: existing methods, applications and integration into clinical practice". en. In: *The British Journal of Radiology* (Aug. 2022), p. 20220071. ISSN: 0007-1285, 1748-880X. DOI: *10.1259/bjr.20220071*.

[146] E. Van Reeth et al. "Super-resolution in magnetic resonance imaging: A review". en. In: *Concepts in Magnetic Resonance Part A* 40A.6 (Nov. 2012), pp. 306–325. ISSN: 15466086. DOI: *10.1002/cmr.a.21249*.

[147] L. Vasung et al. "Quantitative and Qualitative Analysis of Transient Fetal Compartments during Prenatal Human Brain Development". In: *Frontiers in Neuroanatomy* 10 (Feb. 2016). ISSN: 1662-5129. DOI: *10.3389/fnana.2016.00011*.

[148] L. Vasung et al. "Spatiotemporal Differences in the Regional Cortical Plate and Subplate Volume Growth during Fetal Development". en. In: *Cerebral Cortex* 30.8 (June 2020), pp. 4438–4453. ISSN: 1047-3211, 1460-2199. DOI: *10.1093/cercor/bhaa033*.

[149] R. K. Wagner and B. C. Calhoun. "THE ROUTINE OBSTETRIC ULTRASOUND EXAMINATION". en. In: *Obstetrics and Gynecology Clinics of North America* 25.3 (Sept. 1998), pp. 451–463. ISSN: 08898545. DOI: *10.1016/S0889-8545(05)70021-8*.

[150] Z. Wang et al. "Image Quality Assessment: From Error Visibility to Structural Similarity". en. In: *IEEE Transactions on Image Processing* 13.4 (Apr. 2004), pp. 600–612. ISSN: 1057-7149. DOI: *10.1109/TIP.2003.819861*.

[151] M. Wiesenfarth et al. "Methods and open-source toolkit for analyzing and visualizing challenge results". en. In: *Scientific Reports* 11.1 (Dec. 2021), p. 2369. ISSN: 2045-2322. DOI: *10.1038/s41598-021-82017-6*.

[152] R. Wright et al. "Automatic quantification of normal cortical folding patterns from fetal brain MRI". en. In: *NeuroImage* 91 (May 2014), pp. 21–32. ISSN: 10538119. DOI: *10.1016/j.neuroimage.2014.01.034*.

[153] J. Xia et al. "Fetal cortical surface atlas parcellation based on growth patterns". eng. In: *Human Brain Mapping* 40.13 (Sept. 2019), pp. 3881–3899. ISSN: 1097-0193. DOI: *10.1002/hbm.24637*.

[154] J. Xu et al. "SVoRT: Iterative Transformer for Slice-to-Volume Registration in Fetal Brain MRI". en. In: *Medical Image Computing and Computer Assisted Intervention MICCAI 2022*. Ed. by L. Wang et al. Vol. 13436. Series Title: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022, pp. 3–13. ISBN: 978-3-031-16445-3 978-3-031-16446-0. DOI: *10.1007/978-3-031-16446-0_1*.

[155] J. Xu et al. "NeSVoR: Implicit Neural Representation for Slice-to-Volume Reconstruction in MRI". In: *IEEE Transactions on Medical Imaging* (2023), pp. 1–1. ISSN: 0278-0062, 1558-254X. DOI: *10.1109/TMI.2023.3236216*.

[156] X. Xu et al. "Spatiotemporal atlas of the fetal brain depicts cortical developmental gradient". en. In: *The Journal of Neuroscience* (Nov. 2022), JN–RM–1285–22. ISSN: 0270-6474, 1529-2401. DOI: *10.1523/JNEUROSCI.1285-22.2022*.

[157] V. Yeghiazaryan and I. Voiculescu. "Family of boundary overlap metrics for the evaluation of medical image segmentation". In: *Journal of Medical Imaging* 5.01 (Feb. 2018), p. 1. ISSN: 2329-4302. DOI: *10.1117/1.JMI.5.1.015006*.

[158] P. A. Yushkevich et al. "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability". en. In: *NeuroImage* 31.3 (July 2006), pp. 1116–1128. ISSN: 10538119. DOI: *10.1016/j.neuroimage.2006.01.015*.

[159] E. Zvi et al. "Fetal Exposure to MR Imaging: Long-Term Neurodevelopmental Outcome". en. In: *American Journal of Neuroradiology* 41.11 (Nov. 2020), pp. 1989–1992. ISSN: 0195-6108, 1936-959X. DOI: *10.3174/ajnr.A6771*.

Part IV

Appendices

A

Fetal brain MRI biometric measurements

The following tables were presented as supplementary materials of the paper presented in Chapter 2, **"Fetal Brain Biometric Measurements on 3D Super-Resolution Reconstructed T2-Weighted MRI: An Intra- and Inter-observer Agreement Study"** with co-authors: <u>P. de Dumast</u>, M. Khawam, P. Deman, H. Kebiri, T. Yu, S. Tourbier, H. Lajous, P. Hagmann, P. Maeder, J.-P. Thiran, R. Meuli, V. Dunet, M. Bach Cuadra and M. Koob, in **Frontiers in Pediatrics, 9:639746, doi: 10.3389/fped.2021.639746**.
<u>P. de Dumast</u> and M. Khawam are equally contributing authors.

Metric	Observer	Dataset	Measurements Mean +/- SD (mm)		Reliability Lin's CCC	ICC	Paired Wilcoxon's rank sum test p-value	Bonferroni	Error Mean +/- SD (mm)	Mean +/- SD (%)
APDV	obs1	T2WS	9.918 +/- 2.169	9.784 +/- 2.282	0.930	0.938	0.14	1.000	-0.268 +/- 0.772	6.92 +/- 7
		SR	9.918 +/- 2.169							
	obs2	T2WS	10.4 +/- 2.261	10.49 +/- 2.413	0.865	0.868	0.72	1.000	0.184 +/- 1.17	7.43 +/- 6.36
		SR	10.4 +/- 2.261							
bBIP_ax	obs1	T2WS	62.39 +/- 10.21	63.09 +/- 10.35	0.972	0.981	**0.0012**	0.025	1.4 +/- 2.01	2.79 +/- 2.68
		SR	62.39 +/- 10.21							
	obs2	T2WS	60.95 +/- 11.15	60.14 +/- 11.07	0.977	0.988	**0.00077**	0.015	-1.62 +/- 1.74	3.25 +/- 1.91
		SR	60.95 +/- 11.15							
bBIP_cor	obs1	T2WS	58.9 +/- 11.16	59.61 +/- 11.48	0.969	0.977	**0.0079**	0.160	1.41 +/- 2.41	3.54 +/- 2.56
		SR	58.9 +/- 11.16							
	obs2	T2WS	61.02 +/- 10.95	61.69 +/- 10.67	0.971	0.978	**0.013**	0.260	1.35 +/- 2.3	3.59 +/- 3.1
		SR	61.02 +/- 10.95							
FOD	obs1	T2WS	79.99 +/- 12.91	79.94 +/- 13.04	0.996	0.996	0.94	1.000	-0.088 +/- 1.23	1.19 +/- 0.969
		SR	79.99 +/- 12.91							
	obs2	T2WS	80.24 +/- 12.9	80.1 +/- 13.19	0.992	0.992	0.33	1.000	-0.288 +/- 1.61	1.7 +/- 1.17
		SR	80.24 +/- 12.9							
HV	obs1	T2WS	15.19 +/- 2.958	15.18 +/- 3.166	0.956	0.956	1.0	1.000	-0.012 +/- 0.884	4.79 +/- 3.81
		SR	15.19 +/- 2.958							
	obs2	T2WS	15.13 +/- 3.154	15.05 +/- 3.31	0.961	0.962	0.21	1.000	-0.156 +/- 0.881	4.38 +/- 3.47
		SR	15.13 +/- 3.154							
LCC	obs1	T2WS	33.89 +/- 5.988	34.23 +/- 5.989	0.951	0.957	0.084	1.000	0.684 +/- 1.76	4.87 +/- 3.43
		SR	33.89 +/- 5.988							
	obs2	T2WS	34.27 +/- 6.329	34.5 +/- 6.324	0.932	0.935	0.50	1.000	0.46 +/- 2.3	4.91 +/- 4.23
		SR	34.27 +/- 6.329							
sBIP_ax	obs1	T2WS	69.13 +/- 10.33	69.37 +/- 10.47	0.992	0.993	0.080	1.000	0.48 +/- 1.24	1.63 +/- 1.09
		SR	69.13 +/- 10.33							
	obs2	T2WS	68.07 +/- 10.5	68.04 +/- 10.65	0.994	0.994	0.81	1.000	-0.076 +/- 1.2	1.36 +/- 1.1
		SR	68.07 +/- 10.5							
sBIP_cor	obs1	T2WS	65.39 +/- 10.36	66.05 +/- 10.32	0.973	0.981	**0.0071**	0.140	1.32 +/- 2.06	3.29 +/- 1.64
		SR	65.39 +/- 10.36							
	obs2	T2WS	67.65 +/- 10.53	67.5 +/- 10.57	0.996	0.997	0.097	1.000	-0.284 +/- 0.868	1.11 +/- 0.675
		SR	67.65 +/- 10.53							
TCD_ax	obs1	T2WS	32.44 +/- 6.94	32.32 +/- 6.903	0.972	0.973	0.15	1.000	-0.256 +/- 1.63	3.55 +/- 3.25
		SR	32.44 +/- 6.94							
	obs2	T2WS	32.9 +/- 6.914	32.6 +/- 6.847	0.987	0.991	**0.0062**	0.120	-0.608 +/- 0.939	2.73 +/- 2.11
		SR	32.9 +/- 6.914							
TCD_cor	obs1	T2WS	32.32 +/- 6.953	32.1 +/- 6.928	0.992	0.994	**0.023**	0.450	-0.424 +/- 0.793	1.89 +/- 2.19
		SR	32.32 +/- 6.953							
	obs2	T2WS	32.97 +/- 6.9	32.77 +/- 6.949	0.992	0.994	**0.011**	0.210	-0.4 +/- 0.777	2.13 +/- 1.45
		SR	32.97 +/- 6.9							

Table A.1: Summary table of inter-dataset intra-observer analysis for each biometric measurements: mean and standard deviation, agreement analysis (Lin's concordance Correlation Coefficient and Intraclass Correlation Coefficient), comparison using paired Wilcoxon's rank sum test (without correction (bold: significant *p*-values) and with Bonferroni multiple comparisons correction (red: significant *p*-values after correction)) and measurement error (in both*mm* and in percentage).

Metric	Dataset	Observer	Measurements Mean +/- SD (mm)	Reliability Lin's CCC	ICC	Paired Wilcoxon's rank sum test p-value	Bonferroni	Error Mean +/- SD (mm)	Mean +/- SD (%)
APDV	SR	obs1 / obs2	10.05 +/- 2.089	0.722	0.728	0.50	1.00	-0.252 +/- 0.706	11 +/- 9.53
	T2WS	obs1 / obs2	9.784 +/- 2.282	0.659	0.690	0.12	1.00	-0.704 +/- 1.12	13.4 +/- 11.7
bBIP_ax	SR	obs1 / obs2	61.69 +/- 10.24	0.752	0.752	0.99	1.00	-0.072 +/- 1.9	6.94 +/- 8.19
	T2WS	obs1 / obs2	63.09 +/- 10.35	0.729	0.758	**0.022**	0.44	2.95 +/- 1.73	9.27 +/- 7.42
bBIP_cor	SR	obs1 / obs2	58.2 +/- 11.02	0.741	0.755	**0.012**	0.25	-2.15 +/- 2.35	8.15 +/- 8.99
	T2WS	obs1 / obs2	59.61 +/- 11.48	0.751	0.765	**0.040**	0.79	-2.08 +/- 2.82	8.59 +/- 7.7
FOD	SR	obs1 / obs2	80.03 +/- 13.05	0.849	0.850	0.54	1.00	-0.356 +/- 0.927	4.48 +/- 6.73
	T2WS	obs1 / obs2	79.94 +/- 13.04	0.852	0.852	0.75	1.00	-0.156 +/- 1.68	4.61 +/- 6.68
HV	SR	obs1 / obs2	15.19 +/- 2.801	0.851	0.851	0.59	1.00	-0.012 +/- 0.613	6.27 +/- 7.2
	T2WS	obs1 / obs2	15.18 +/- 3.166	0.836	0.837	0.29	1.00	0.132 +/- 0.891	7.5 +/- 7.29
LCC	SR	obs1 / obs2	33.55 +/- 6.089	0.864	0.866	0.59	1.00	-0.488 +/- 1.97	5.81 +/- 6.68
	T2WS	obs1 / obs2	34.23 +/- 5.989	0.914	0.915	0.68	1.00	-0.264 +/- 1.47	4.31 +/- 5.39
sBIP_ax	SR	obs1 / obs2	68.89 +/- 10.4	0.799	0.802	0.13	1.00	0.776 +/- 1.04	5.33 +/- 6.71
	T2WS	obs1 / obs2	69.37 +/- 10.47	0.755	0.761	0.067	1.00	1.33 +/- 1.33	6.29 +/- 7.15
sBIP_cor	SR	obs1 / obs2	64.73 +/- 10.57	0.742	0.774	**0.0077**	0.15	-3.06 +/- 1.76	8.12 +/- 6.79
	T2WS	obs1 / obs2	66.05 +/- 10.32	0.774	0.782	0.053	1.00	-1.45 +/- 1.8	6.12 +/- 7.09
TCD_ax	SR	obs1 / obs2	32.57 +/- 7.117	0.643	0.646	0.053	1.00	-0.632 +/- 0.536	9.28 +/- 13.2
	T2WS	obs1 / obs2	32.32 +/- 6.903	0.679	0.680	0.20	1.00	-0.28 +/- 1.37	9.28 +/- 11.9
TCD_cor	SR	obs1 / obs2	32.53 +/- 7.114	0.667	0.670	**0.036**	0.71	-0.64 +/- 0.61	9.19 +/- 12.6
	T2WS	obs1 / obs2	32.1 +/- 6.928	0.705	0.708	0.074	1.00	-0.664 +/- 0.783	8.61 +/- 12

Table A.2: Summary table of inter-observer intra-dataset analysis for each biometric measurements: mean and standard deviation, agreement analysis (Lin's concordance Correlation Coefficient and Intraclass Correlation Coefficient), comparison using paired Wilcoxon's rank sum test (without correction (bold: significant *p*-values) and with Bonferroni multiple comparisons correction (red: significant *p*-values after correction)) and measurement error (in both *mm* and in percentage).

	LCC	HV	APDV	DTC_cor	DTC_ax	sBIP_cor	sBIP_ax	bBIP_cor	bBIP_ax	DFO	ALL
T2WS	0.83	0.87	0.77	0.97	0.97	0.87	0.81	0.91	0.88	0.88	0.97
SR	0.85	0.88	0.85	0.96	0.96	0.85	0.85	0.90	0.86	0.88	0.97
Overall	0.84	0.87	0.79	0.96	0.96	0.86	0.83	0.90	0.87	0.88	0.97

Table A.3: **Intra-rater analysis.** To ensure the reliability of the measures, we have performed a complementary intra-rater variability analysis with obs1. This additional analysis was performed in a subcohort of five fetuses (gestational ages: 18, 22, 26, 30 and 34 weeks). For each subject, obs1 repeated 3 times the measurement of all the biometrics (measurements 2 and 3 were done one week apart, approximatively a year after measurement 1), on both T2WS and SR datasets. We present in this table the intra-class correlation (ICC) of the repeated measures by obs1 for each biometric and dataset. Overall, the intra-rater reliability is excellent with ICC = 0.97. Similarly for each dataset, ICC= 0.97 and 0.97 respectively for T2WS and SR dataset. For SR, this is in line with (Kyriakopoulou et. *al* [89]) that presents an ICC of 0.99.

B

MIALSRTK BIDS App

MIALSRTK [1] is originally a library of C++ and Python algorithms, that are successively called from a script to generate an end-to-end SR reconstruction pipeline. A *syngo*.via Frontier prototype of the MIALSRTK pipeline was further implemented in order to ease the user interaction (e.g. LR images selection, brain extraction) [26]. This version allowed the reconstruction of clinical exam within the PACS system and was used in the study presented in Chapter 2. Nonetheless, as mentioned in Chapter 2, the necessity for repeated user interaction, the means of dissemination and the computation time of our MIALSRTK tool hamper a wider use of our reconstruction pipeline.

Over my PhD thesis, we, in a joint work with Dr Sébastien Tourbier [2], Hamza Kebiri [3] and Dr Thomas Sanchez [3], invested efforts in the reorganizaion of MIALSRTK and integration of new features.

In this appendix, we first present the software development related to the refactoring into an interaction-reduced and self-contained application being compliant with the well dened Brain Imaging Data Structure (BIDS) [4], a community effort that describes a way of organizing neuroimaging data. Second, we briefly introduce the developments that aimed at optimizing the performance, improving the user experience and further integrating new features.

---

[1] *https://github.com/Medical-Image-Analysis-Laboratory/mialsuperresolutiontoolkit*

[2] Connectomics Lab, Department of Radiology, Lausanne University Hospital (CHUV) and University of Lausanne (UNIL), Lausanne, Switzerland

[3] MIAL, Department of Radiology, Lausanne University Hospital (CHUV) and University of Lausanne (UNIL), Lausanne, Switzerland

[4] *https://bids.neuroimaging.io*

B.1 Towards large scale studies

B.1.1 A modular pipeline

The Neuroimaging in Python Pipelines and Interfaces (Nipype)[5] dataflow offers the possibility to encapsulate each algorithm into a well defined interface. As a result, each step of the processing is being instantiated as a module which are connected in a pipeline graph. This design architecture offers, among others: 1) **flexibility** as connection can dynamically be set, allowing to add/skip processing steps; 2) **modularity** as interaction are not restricted to our own software library, e.g. one could imagine replacing one or another processing steps by the one of another toolbox; and 3) an improved **management of the resources**, allowing to fasten the process by running steps in parallel.

From this, we designed and built the graph of a whole end-to-end SR reconstruction pipeline with options to either run or not specific steps. One should note that this new version of our toolbox makes it easier for a user to design his own workflow using, not exclusively, our modules, as long as they are wrapped into Nipype node.

B.1.2 A fully automatized BIDS App

A frequent conflict in neuroimaging is the way to store data, which, in turn, is often different between research groups. Consequently, an inherent issue is the waste of time in either rearranging data to meet specific software requirement or in rewriting scripts to make the interface between data and software. The lack of well defined standard makes the development of robust image processing software to any data organization scheme difficult. Regrettably, it might additionally prevent groups from using tools due to extensive efforts needed for rearrangement.

The BIDS architecture provides guidelines in the way of organizing neuroimaging data. A multi-level organization and a well defined nomenclature makes this new standard easy to use and easy to share. Along with the emergence of this new standard, have also been defined the *BIDS Apps*, defined as "portable neuroimaging pipelines that understand BIDS datasets". These BIDS Apps are built such that they are self-contained, easy to install and compliant with BIDS data formatting while still user-friendly. The MIALSRTK pipeline refactoring described in the previous Section has been containerized such that it is now disseminated as a BIDS App. Both Singularity and Docker encapsulations[6] are publicly available, following the Open Science guidelines of Swiss National Science Foundation, allowing for other research groups to perform large-scale studies with a fully automatized pipeline.

---

[5]*https://nipype.readthedocs.io*
[6]https://mialsrtk.readthedocs.io/en/latest/

B.2 Features development

B.2.1 User interaction reduction

In the *syngo*.via Frontiers prototype version of MIALSRTK, user-interaction were necessary for *(i)* the choice of LR stacks, *(ii)* the choice of the reference stack (needed in the motion estimation step), and *(iii)* the localization of the brain. In order to minimize the need for user interaction, two new modules have been developed and integrated:

`BrainExtraction` is a DL-based automatic segmentation method that outputs a binary mask of the fetal brain. This development of this module was lead by Hamza Kebiri and presented in the European Congress of Magnetic Resonance in Neuropediatrics (ECMRN) in 2020, as "Automated fetal brain segmentation of 2D magnetic resonance images: transfer learning and 3D topology correction" with co-authors H. Kebiri, P. de Dumast, T. Yu, H. Lajous, J.-Ph. Thiran, R. Meuli, M. Koob and M. Bach Cuadra.

`StacksOrdering` ranks the LR stacks based on a motion index computed from the brain masks. The less motion-corrupted stack is defined as the stack of reference. The development of the motion index computation was lead by Thomas Yu[7] and was presented at ECMRN 2020 as "Translating fetal brain magnetic resonance image super-resolution reconstruction into the clinical environment", with co-authors: P. de Dumast, P. Deman, M. Khawam, T. Yu, S. Tourbier, H. Lajous, P. Hagmann, P. Maeder, J.-Ph. Thiran, R. Meuli, V. Dunet, M. Koob and M. Bach Cuadra.

Note that, although these two modules are aimed to eliminate the user interaction from the MIALSRTK pipeline, validation or correction may be necessary. Indeed, binary masks may need refinement either for *(i)* intra-slice correction (due to over- or under-segmentation) or *(ii)* full-slice rejection (e.g. brain does not appear, poor quality due to motion or signal drop). Additionally, while the `StacksOrdering` module ranks the LR stacks, it is only used for the selection of the reference. Therefore, by default all LR stacks available are used. The optional pre-selection of the LR input images remains at the user discretion.

B.2.2 Performance optimization

In MIALSRTK, the most computationally expensive step is the motion estimation, followed by the image recovery. The motion estimation that is performed though SVR registration is a voxel-based operation. Consequently, its complexity and hence computation time heavily depends on the number and size of LR images.

---

[7]Signal Processing Laboratory (LTS5), Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

To address this limitation, we developed the **ReduceFOV** preprocessing module that aims at cropping the LR images around the region-of-interest. The latter is determine with the binary mask of the brain that is either provided as input or computed with the previously introduce `BrainExtraction` module.

B.2.3   New MIALSRTK features

Anatomical alignment
By default, the SR reconstruction of a fetal brain MRI exam is done in the subject's space. We integrated into our pipeline a module for the realignment into the anatomical planes (axial, coronal, sagittal) based on the NiftyMIC principal brain axis (PBA) implementation[8] [33]. The computation and application of the alignment transformation are applied in the postprocessing stage **postproc_stage**. Figure B.1 shows the SR reconstruction of a 29.3 weeks of GA subject pre- and post-alignment to the template's space.

Labelmap HR reconstruction
The **recon_labelmap_stage** allows the optional reconstruction of LR labelmaps when available. One HR reconstruction is computed per label and per-label reconstructed maps are combined through a majority voting step. This feature was particularly used in the work presented in Chapter 4.

SR reconstruction assessment
The **srr_assessment_stage** allows the optional computation of quality metrics against a HR reference image when available. This feature was used in the work presented in Chapter 3.
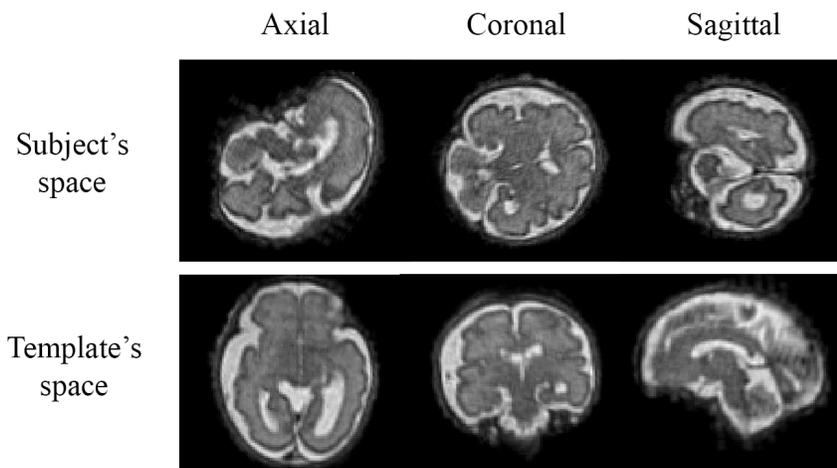

Figure B.1: SR reconstruction of a 29.3 weeks of GA neurotypical subject in the subject's space (top) and in a template space (bottom).

---

[8]https://github.com/gift-surg/NiftyMIC

Figure B.2 shows a simplified diagram of the overall MIALSRTK reconstruction pipeline.
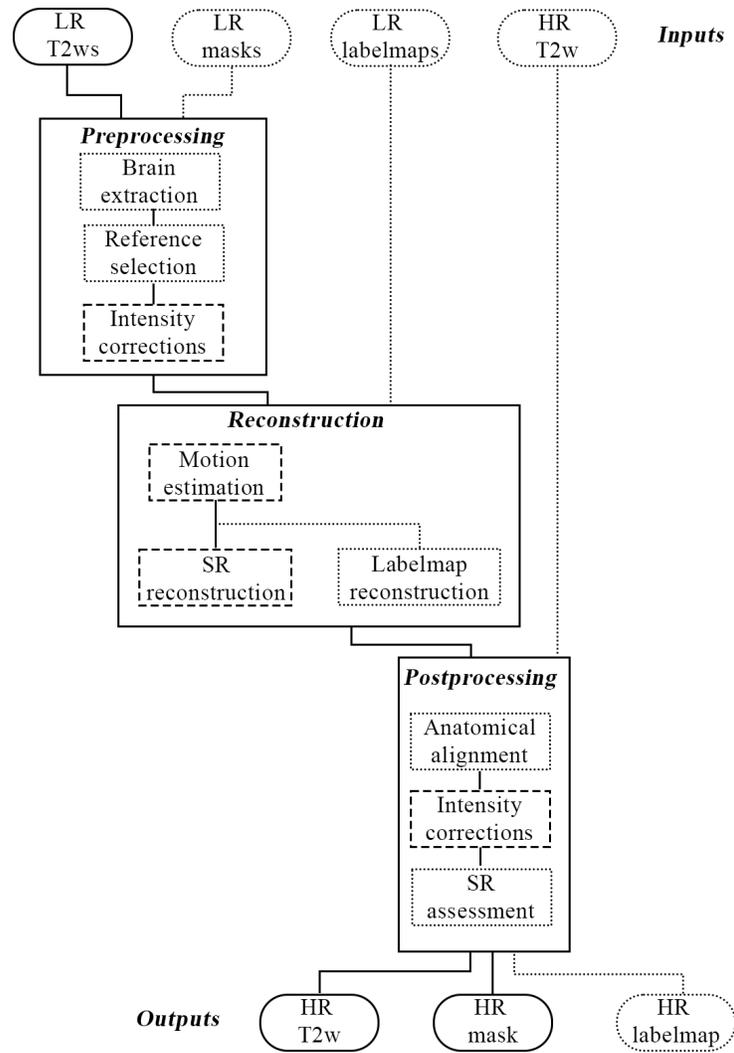

Figure B.2: Simplified representation of the overall MIALSRTK reconstruction pipeline. Dotted elements are optional.

C

Multi-atlas segmentation

The work presented in this chapter is part of the originally published work **"An auto-matic multi-tissue human fetal brain segmentation benchmark using the Fetal tissue annotation Dataset"** with co-authors: K. Payette, P. de Dumast, H. Kebiri, I. Ezhov, J.C. Paetzold, S. Shit, A. Iqbal, R. Khan, R. Kottke, P. Grehten, H. Ji1, L. Lanczi, M. Nagy, M. Beresova, T. D. Nguyen, G. Natalucci, T. Karayannis, B. Menze, M. Bach Cuadra and A. Jakab, in **Scientific Data, doi: 10.1038/s41597-021-00946-3**.

On the occasion of the release of the Fetal Tissue Annotation and Segmentation Dataset (FeTA) v1.2, I participated to the benchmark of automatic multi-tissue fetal brain MRI segmentation.

C.1   Dataset

The 1st Fetal Tissue Annotation and Segmentation Dataset (FeTA) [1] release was composed of fifty (50) SR-reconstructed volumes of T2w images, splitted into two groups: training (40) and testing (10). The cohort is composed of 18 neurotypical and 32 pathological pre- or post-operative spina bifida subjects. Gestational age repartition of subjects into the subgroups is presented in Figure C.1.

Training subjects are provided along with manual annotation of the following tissues: extra-axial cerebrospinal fluid (CSF) spaces, gray matter (GM), white matter (WM), ventricles, cerebellum, deep gray matter (deep GM) and brainstem.
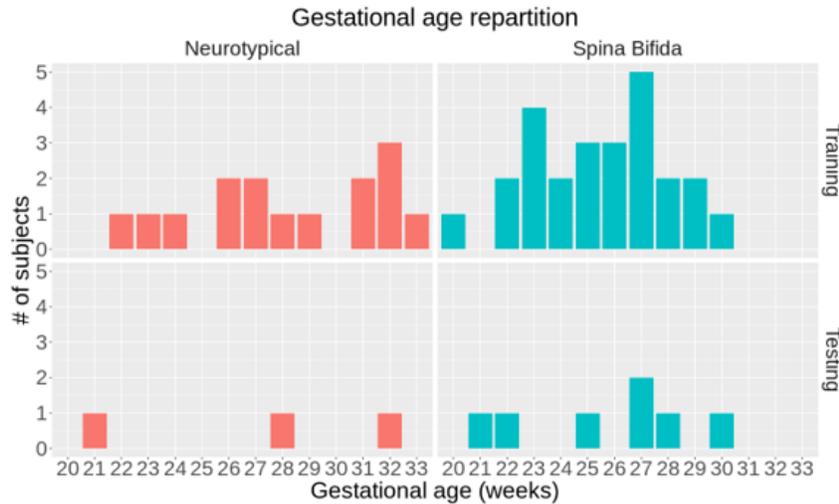
---

[1]DOI: 10.5281/zenodo.4541606

Figure C.1: Subjects repartition into dataset subgroups.

C.2   Evaluation

Evaluation are performed with the EvaluateSegmentation tool [137] for the measurement of the Dice similarity coefficient (DSC), the 95th percentile of the Hausdorff distance (Hausdorff$_{95}$) and the volume similarity (VS).

C.3   Multi-atlas segmentation

As introduced in Section 1.4.2, registration-based segmentation techniques aim to estimate the best image-to-template transformation in order to further propagate the labelmap. In case of multi-atlas segmentation, propagated labelmaps from multiple atlases are combined.

In our approach, all 40 training subjects - hence with manual annotations - are initially considered as atlases, although a more specific criteria-based selection of atlas candidates is performed. Each of the atlas' SR image is classified into one of the three categories: bad, acceptable or excellent and those estimated as bad were considered not good enough for registration purposes and thus discarded. In fetal brain segmentation, the selection of appropriate atlases is challenging due to the major morphological development of the fetal brain that is happening along the course of gestation. Intuitively, atlas candidates that are selected are the ones more likely to present similar morphology than the target subject. Therefore, in our approach only those within two weeks younger or older were registered to the subject's space. Finally, the limitation on the number of available atlases did not allow for pathology-based discrimination criteria, hence both normal developing and spina bifida subjects were used as atlases (see Figure C.2 for an overview of atlases used in label fusion).
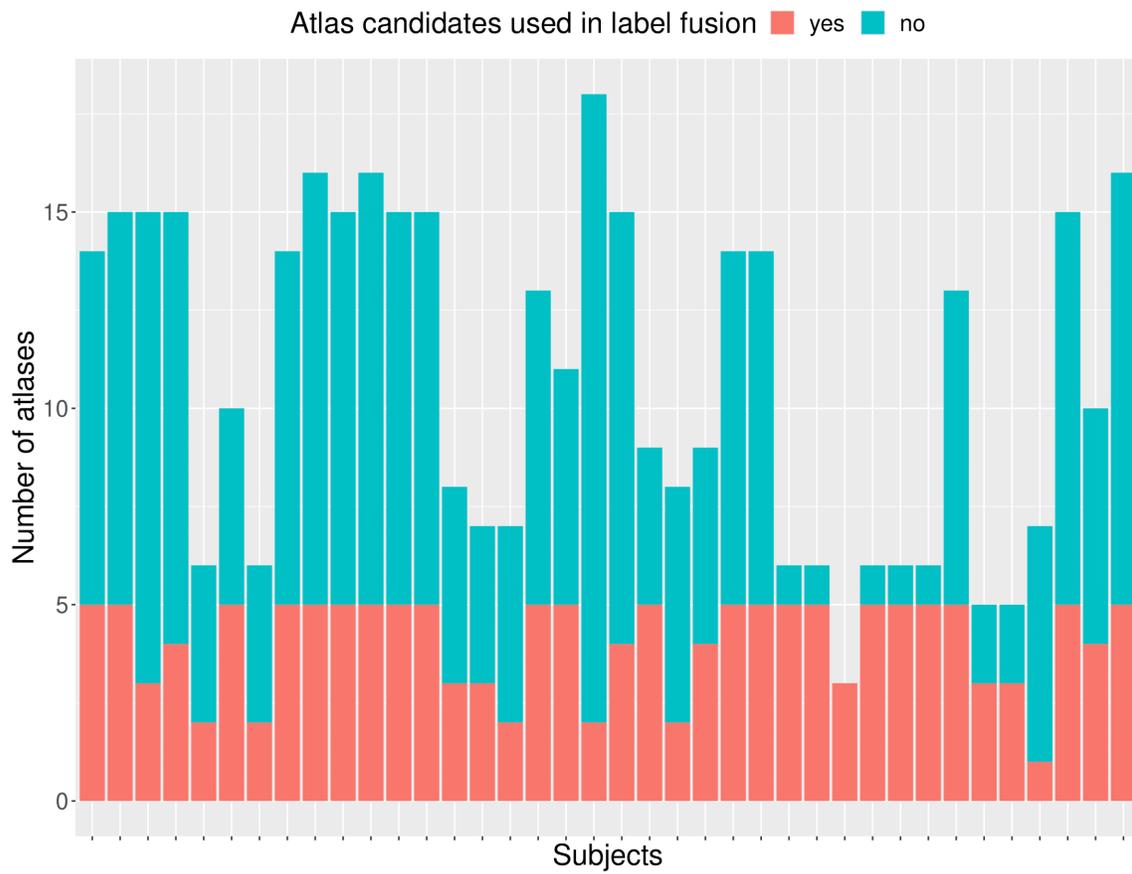
Figure C.2: Number of atlas candidates vs. number of atlases (5 best ranked NCC and NCC $\geq$ 0.8) used in the label fusion step, for the training dataset, using a leave-one-out approach..
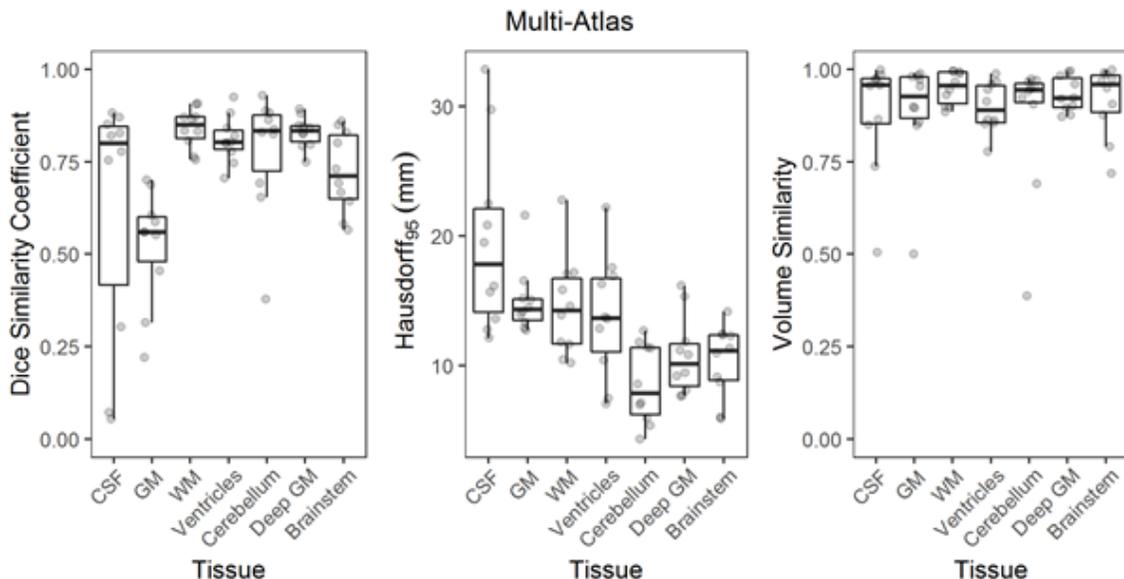
Figure C.3: Overview of the Multi-Atlas Segmentation performance.

To proceed from global alignment of MR volumes to local deformation, a three-level registration is performed, using the Advanced Normalization Tools (ANTS)[5], in the following order, increasing the freedom of the transformation: rigid, affine and finally non-linear symmetric diffeomorphic. We used the intensity-based normalized cross-correlation (NCC) as similarity registration metric. Consequently, we define registration as being sufficiently good for being part of the label fusion when its NCC is equal to or higher than 0.8 and the atlases that do not reach this NCC value are discarded. We use the NiftySeg implementation for NCC computation[2]. We limit the maximum number of atlases used for the label fusion to 5 (having the highest NCC). A local weighted voting algorithm is used for the final label fusion step [4] in which weighting is based on the NCC computed on a Gaussian kernel of size 10 voxels.

C.4   Results

We present partial results of our MAS approach from [112]. Methods were ranked based on the following three metrics: the DSC, the HD95 and the VS. Figure C.3 presents the performance of our method for each label. Overall, our MAS method ranked 4th out of 10 benchmarked.

---

[2]*https://github.com/KCL-BMEIS/NiftySeg*

D

Multi-modal assessment of the corpus callosum biometry

The work presented in this appendix gathers the preliminary results of the original paper in preparation for submission to *Frontiers in Medicine*, as **"Assessment of fetal corpus callosum and sub-segments biometry by 3D super-resolution reconstruction of brain MRI"** with co-authors: S. Lamon, P. de Dumast, V. Dunet, L. Pomar, Y. Vial, M. Koob and M. Bach Cuadra.

D.1   Introduction

The corpus callosum (CC) is the largest white matter commissure structure that allows interhemispheric connections [110]. Agenesis of the CC (ACC) is the developmental absence, either partial or complete, of the CC. ACC is amongst the most common brain malformation, with a prevalence of more than 1 per 1000 individuals [110]. Such malformation may be isolated of associated with other congenital anomalies [132]. ACC may cause neuropsychological disorders, however, cognitive functions are not necessarily hindered [127, 132]. Given such heterogenous prognosis, the clinical assessment of ACC is challenging and a better understanding of its development is necessary.

Similarly to other cerebral structures, the CC benefits from the imaging techniques, namely US and MRI, to quantify its biometry. However, CC assessment is the greatest source of discrepancies between US and MRI [59], and US remains the gold standard.

In this study, we focus on the biometric analysis of the fetal corpus callosum (CC) in a multi-modal multi-dataset analysis. We aim to assess SR reconstruction of fetal brain MR exams against US, compared to T2WS.

D.2   Materials and methods

D.2.1   Cohort

We retrospectively collected fetal brain MR exams from the Lausanne University Hospital (CHUV). All MR exams were performed on medical indication. Fifty-seven (57) fetal brain MRI exams were selected: 46 normal subjects – including those with minor ventriculomegaly (<12mm), and 11 subjects with a partial ACC. To minimize the time difference between exams, we only consider US exams performed within maximum two weeks difference from the MR one.

The local ethics committee of the Canton of Vaud, Switzerland (CER-VD 2021-00124) approved the retrospective collection and analysis of MRI and US data.

D.2.2   Datasets

Ultrasound imaging: *US*
US images were acquired on General Electric (ZIPF, Austria) Voluson 730, E8, E10 machines. The brain volumes containing the CC were acquired from the midsagittal plane, using the anterior fontanelle and the sagittal suture as an acoustic window [109]. Of all the US image used, the majority of them were 3D-US (93%) and only 7% were 2D-US.

MR acquisitions: *T2WS*
Clinical MR images were either acquired at 1.5T or 3T on Siemens Healthcare (Erlangen, Germany) scanners. The fetal brain MRI protocol included T2w Half-Fourier Acquisition Single-shot Turbo spin Echo (HASTE) sequences in three orthogonal orientation.

Super-resolution reconstruction: *SR*
All MR exams were SR reconstructed by an engineer using the MIALSRTK [144] BIDS App (see Appendix B). In a first round, all exams were reconstructed using a fully automatic pipeline (selection of the reference, brain extraction). After a visual inspection, a second round of reconstruction was done if necessary after a manual refinement of the masks, and a finer selection of the reference and stacks used. All SR images were reconstructed with an isotropic spatial resolution matching its input in-plane resolution (in average of around $1.1mm$ for 1.5T and $0.5mm$ for 3T exams).

Figure D.1 illustrates the three dataset images on a subject of about 32 weeks of GA.
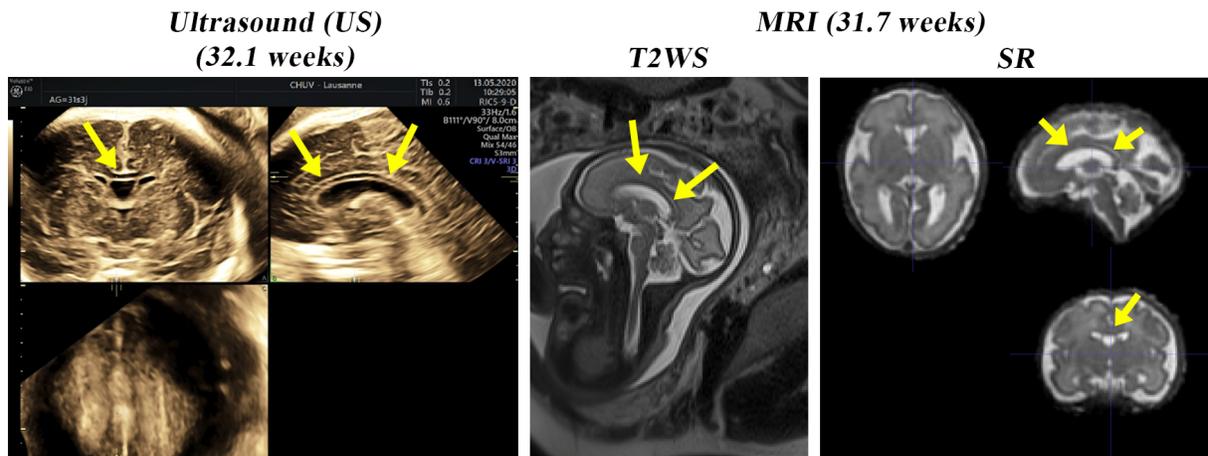
Figure D.1: Visualization of the corpus callosum of a 32 weeks old subject on *US*, *T2WS* and *SR* images.

D.2.3 Biometric measurements

Two observers, one junior (obs1) with little experience in fetal brain MRI (<6 months) and one experienced pediatric neuroradiologist (obs2) performed CC biometric measurements. Obs1 performed measurements on all three datasets (*US*, *T2WS* and *SR*). US measurements were made under supervision of a midwife specialized in fetal brain sonography with 10 years of experience and a skilled obstetrician with 35 years of experience. Both supervisors reviewed obs1's measures at multiple time and validated them. Obs2 only measured on MR datasets (T2WS and SR).

Measurements of the outer-outer **length**, the heights of the **rostrum**, the **genu**, the **body** and the **splenium** were made, all in triplicates and averaged, following the guidelines defined by Pashaj et *al.* [109]:

- the **rostrum** is the most anterior part which is oriented postero-inferiorly,

- the **genu** is defined as the segment situated anteriorly to a line passing through the anterior fornix and parallel to another line passing through the posterior fornix and the quadrigeminal plate,

- the **body** is situated between the splenium and the genu, and

- the **splenium** corresponds to the posterior 20% of the CC,

Figure D.2 schematically illustrates the measurements performed.

For all subjects, CC measurements were performed on the best midsagittal image available. On *US*, measurements were directly performed on the US device, which allowed the observer to travel in the planes to find the best mid-sagittal view. On *T2WS*, the free-resource ITK-SNAP software was used. The best acquisition for each
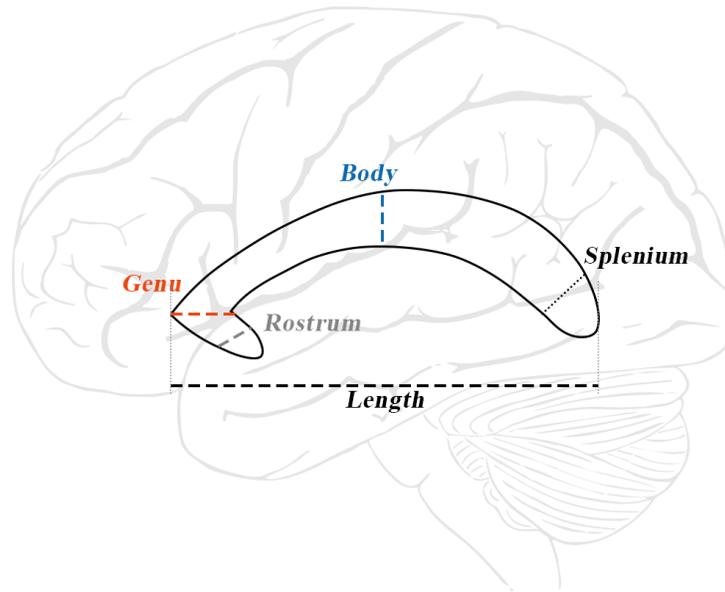
Figure D.2: Schematic view of the corpus callosum biometric measurements: the length, the heights of the rostrum, the genu, the body and the splenium.

orthogonal plane was chosen after a visual examination. On *SR*, the reconstructed volumes were re-oriented to fit in the anatomical orthogonal axis and reach an accurate mid-sagittal view.

D.2.4   Image quality

Quality scores are given to each dataset exam, based on the precise criteria:

- On *US*, five items were considered, as defined in [117]: *(i)* strict sagittal plane with clear visibility of the cerebellar vermis, the brain stem, the fourth ventricle and the CC, *(ii)* the four parts of the CC can be seen, *(iii)* zoom is sufficient, *(iv)* CC from the cavum septum pellucidum is clearly distinguable, and *(v)* the calipers are correctly placed.

- On *T2WS*, seven items are considered: *(i)* quality of the visualization of the CC, *(ii)* the whole CC can be visualized on the same cut, *(iii)* blurring of the CC, *(iv)* stack is oblique, *(v)* quality of the stack, *(vi)* motion-corruption of the stack, and *(vii)* blurring of the stack.

- On *SR*, five of the seven *T2WS* are considered: *(i)* quality of the visualization of the CC, *(ii)* the whole CC can be visualized on the same cut, *(iii)* blurring of the CC, *(iv)* quality of the stack, and *(v)* blurring of the stack.

On *US*, each item was rated as 1=yes, 0=no and the sum of all is considered as the final quality score. On *T2WS* and *SR* datasets, each item is rated from 0 to 3 (0=un-

usable, 1=bad, 2=average, 3=good). The round mean of all items is considered as the final quality score.

D.2.5 Statistical analysis

Gestational age

Regression analysis of each measure according to the GA were performed. For quality control, we compare them to published reference charts. ACC subjects are not included in this analysis. *US* measurements are compared to Pashaj et *al*. [109] for all biometric measurements. *T2WS* length is compared to Tilea et *al*. [142]. To our knowledge, nor *T2WS* sub-segments, nor *SR* reference measurements exist.

Inter-subset agreement

We evaluate the discrepancies between the subsets of the measurements performed:

- We assess the discrepancies between the inter-dataset (*US*, *T2WS*, and *SR*) measurements of obs1.

- We further compare the inter-observer agreement between obs1 – junior – and obs2 – experienced – on the *SR* dataset.

- To remove the confounding of the experience, we assess the inter-modality agreement (*US* vs. *T2WS*, and *US* vs. *SR*), in which *US* measurements are those performed by obs1 that are validated by two experts, and MR measurements (*T2WS* and *SR*) are from obs2.

For each of these subset comparisons, a paired Wilcoxon rank sum test is performed, with statistical significance set to $p < 0.05$. p-values are adjusted for multiple comparisons using Bonferroni correction.

D.3 Results

D.3.1 Dataset analysis

Amongst the 57 fetal brain exams considered (*T2WS*), 43 had a *US* exam available with a maximum time difference of two weeks. Finally, *SR* reconstruction was successful for 51 subjects.

Figure D.3 shows for each subject the images available, the measurements that could or could not be performed and the quality score D.3. Overall, it appears that *SR* dataset presents less missing value that others (only 2 in rostrum). In *T2WS*, missing values seem to be mostly localized in the anterior segments (mainly rostrum, and genu), while in US they span over the CC. While *T2WS* only present acceptable and excellent exams

according to the quality score, it does not seem to ensure the possibility to measure the different sub-segments of the CC, especially the rostrum.

Table D.1 summarizes the frequency of each of the measurements per dataset. The length could be measured on all the images available for each subject. Although *US* is considered to be the gold standard for CC analysis, it is in this dataset that we observe the more missing values, for each of the sub-segments. We recall that this study is retrospective. Therefore, measurements were made on volumes that were not expressly acquired for the analysis on CC sub-segments.

Overall, regardless of the dataset, the rostrum that is the thinner sub-segment (see Figure D.2) has the more missing values.
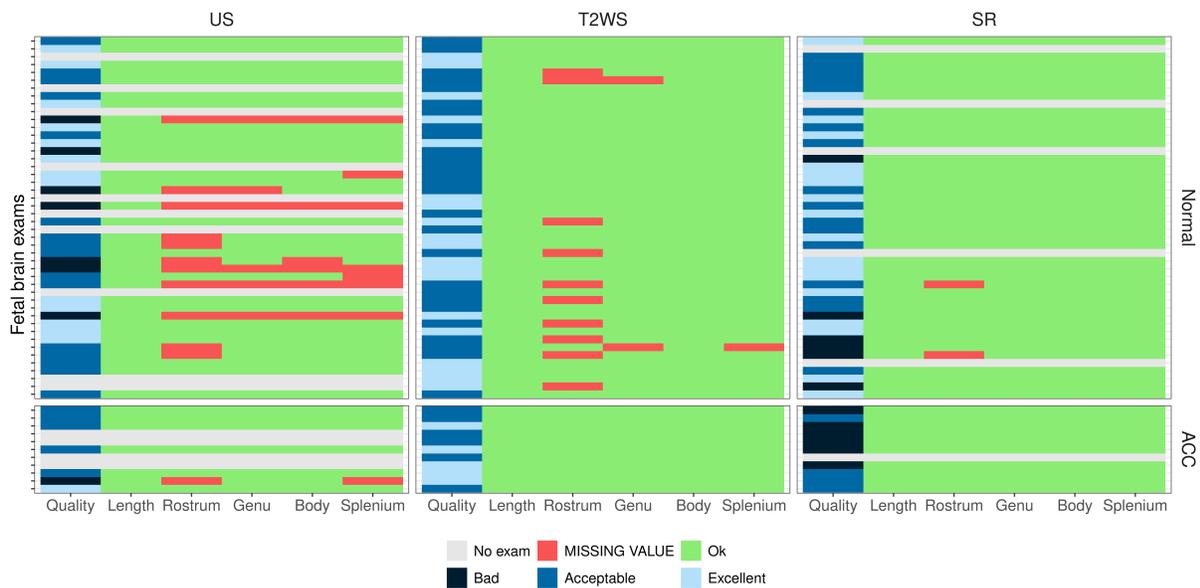
Figure D.3: Summary of image quality, image availability and measurements performed for each subject.

Table D.1: Frequency of measurements per biometric measure.

	LCC	Rostrum	Genu	Body	Splenium
US	100%	72%	86%	86%	81%
T2WS	100%	82%	96%	100%	98%
SR	100%	96%	100%	100%	100%

D.3.2   Gestational age analysis

Figure D.4 presents a GA-based regression of the length of the CC for each dataset, in which *US* and *T2WS* are respectively compared to Pashaj et *al*. [109] and Tilea et *al*. [142]. We observe that the regression of *SR* measurements is closer than the *T2WS* measurements to the *US* regression that is considered the gold standard.

Figure D.4: Regression according to the gestational age of the length of the corpus callosum for each dataset compared to literature. *US* reference charts: Pashaj *et al*. [109]. *T2WS* reference charts: Tilea et *al*. [141].

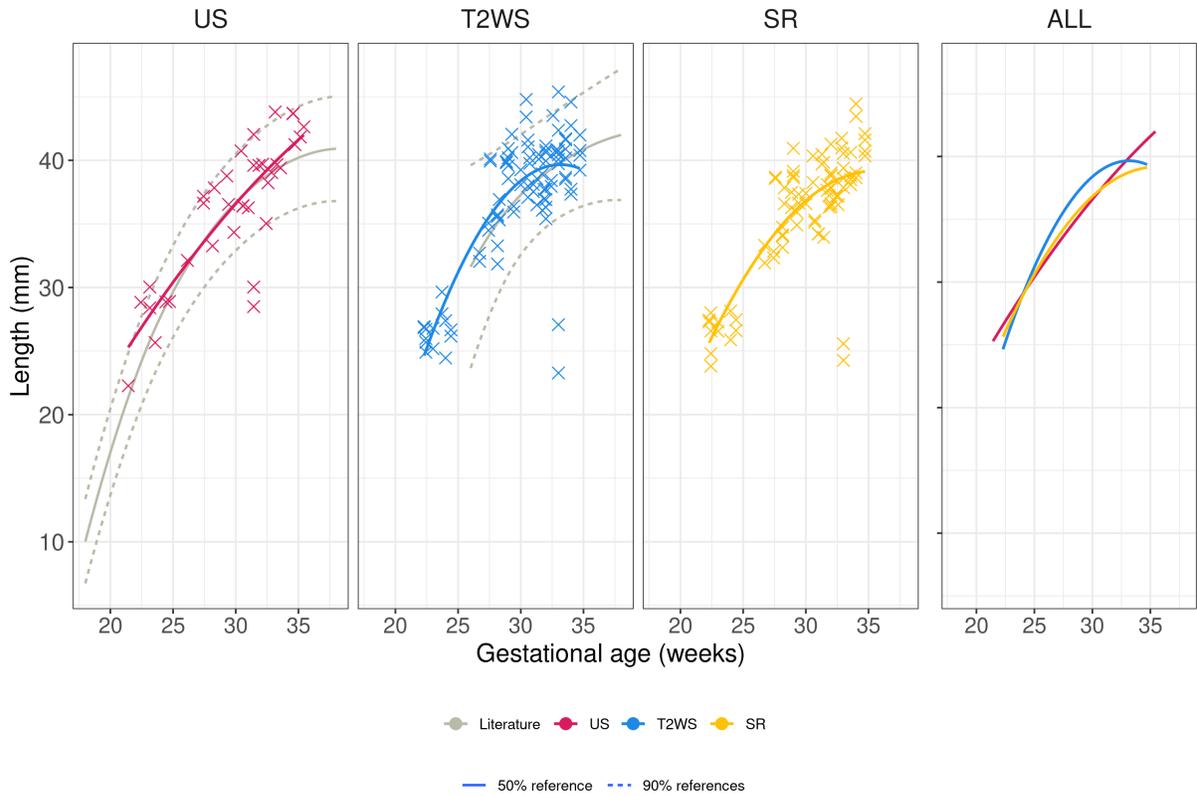Similarly, Figure D.5 presents a GA-based regression for each sub-segment of the CC in each dataset. All datasets are compared to the reference *US* measurements from Pashaj et *al*. [109]. On *US*, we observe that all measurements seem to be in the lower half of the reference range.

D.3.3 Inter-dataset intra-observer

Table D.2 shows the inter-dataset statistical comparison for the measurements of obs1 along with the number of exams considered. Significant differences are observed for the genu and the splenium between *US* and *T2WS*, for the rostrum and the genu between *T2WS* and *SR*, and only for the genu between US and *SR*.

We observe that the genu is a source of disagreement in all inter-dataset analysis. We know that US and MR exams were not performed on the same day. Moreover, we observe in the reference charts from [109] that the genu has the biggest growth throughout gestation, while the rostrum and splenium seem steady from about 25 weeks (see
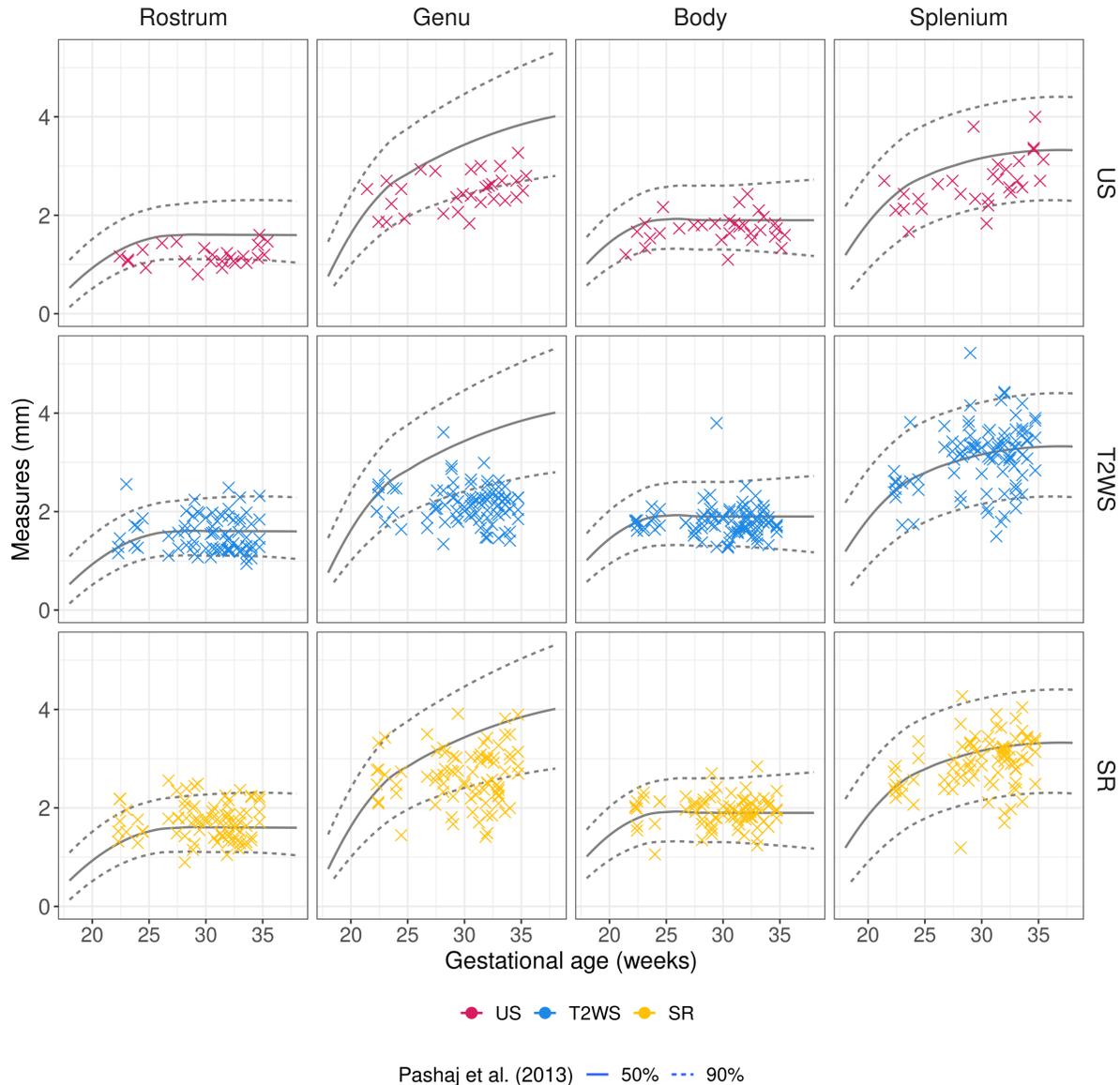
Figure D.5: Regression according to the gestational age of each sub-segment of the corpus callosum for each dataset compared to *US* reference charts from Pashaj *et al.* [109].

Figure D.5). Therefore, we may explain the discrepancies in the genu between *US* and MR datasets by the influence of the GA.

Additionally, *T2WS* and *SR* show a disagreement in the rostrum. From Figure D.3 and Table D.1, we recall that this segment seemed to be a challenging one to measure and is the shorter of all, hence being more prone to partial volume effect alteration.

Table D.2: Inter-dataset comparison using the Wilcoxon rank sum test adjusted for multiple comparisons using Bonferroni correction for obs1. $p < 0.05$ (*) is considered statistically significant.

	Length	Rostrum	Genu	Body	Splenium
US vs. *T2WS*	1.0 (N=43)	1.0 (N=25)	**6.3e-03*** (N=35)	1.0 (N=37)	**0.0067*** (N=34)
T2WS vs. *SR*	0.25 (N=51)	**0.027*** (N=43)	**1.1e-07*** (N=49)	0.071 (N=51)	0.41 (N=50)
US vs. *SR*	0.12 (N=38)	0.30 (N=26)	**1.8e-02*** (N=32)	0.086 (N=32)	0.14 (N=30)

D.3.4 Inter-observer intra-dataset (*SR*)

Table D.3 shows inter-observer agreement on the *SR* dataset. Statistical significance is observed for the heights of the genu and the rostrum.

Table D.3: Inter-observer measurements comparison using the Wilcoxon rank sum test adjusted for multiple comparisons using Bonferroni correction on the *SR* dataset. $p < 0.05$ (*) is considered statistically significant.

	Length	Rostrum	Genu	Body	Splenium
SR	1 (N=51)	**9e-05*** (N=48)	**0.011*** (N=50)	0.094 (N=51)	0.086 (N=51)

D.3.5 Experts inter-modality

Table D.4 shows the inter-modality comparison (*US* vs. MR datasets), considering for each dataset the expert's measurements (i.e. obs1 in *US*, and obs2 in *T2WS* and *SR*). In both *US* vs *T2WS* and *US* vs. *SR* comparisons, significant difference is only observed for the rostrum.

Table D.4: Experts inter-dataset comparison using the Wilcoxon rank sum test adjusted for multiple comparisons using Bonferroni correction. $p < 0.05$ (*) is considered statistically significant.

Experts inter-dataset	Length	Rostrum	Genu	Body	Splenium
US vs. T2WS	1.0 (N=38)	**1.6e-04*** (N=28)	0.11 (N=34)	1.0 (N=35)	0.23 (N=34)
US vs. SR	0.16 (N=38)	**9.1e-05*** (N=25)	1.0 (N=31)	0.066 (N=32)	1.0 (N=30)

D.4   Discussion

The results presented in this appendix are preliminary and will require further analysis to conclude on the potential added value of *SR* as compared to *T2WS* in the biometric analysis of CC. Nonetheless, *SR* seems to help in the visualization of the entire CC, which is in line with our previous findings presented in Chapter 2 [80]. Again, we hypothesize that the possibility to freely navigate in the HR volumes plays a key role. Moreover, the rostrum stands out of all subset paired comparison. Although we do not have an explanation at the moment, such observation had already been made [48, 16, 80].

Priscille DE DUMAST

WORK EXPERIENCE

APRIL 2019 MARCH 2023	PhD candidate at the Center for Biomedical Imaging (CIBM), University of Lausanne , Lausanne, Switzerland
MARCH 2018 DEC 2018	Research intern – master project – at Advanced Clinical Imaging Technology (ACIT), Siemens Healthineers, Lausanne, Switzerland
JULY 2016 JUNE 2017	Research assistant at the Dental and Craniofacial Bionetwork for Image Analysis (DCBIA) lab, Department for Orthodontics and Pediatric Dentistry, University of Michigan, Ann Arbor, MI, USA

EDUCATION

APR 2019 MARCH 2023	PhD in Life Sciences at University of Lausanne, Lausanne, Switzerland	
	Thesis title	Development and validation of robust MR image reconstruction and segmentation techniques for the quantitative analysis of the fetal brain
	Supervisor	Dr. Bach Cuadra
SEP 2015 SEP 2018	Engineering degree in Digital sciences at Lyon School of Chemistry, Physics and Electronics (CPE Lyon), Lyon, France	
	Specialization	Image, modeling and computing
	Sep 2017–Feb 2018	International exchange at Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
SEP 2012 JUNE 2015	Bachelor Degree in Digital sciences at CPE Lyon, Lyon, France	
	Main program	Electronics, programming, signal processing
JUNE 2012	Scientific Baccalaureate at Externat Sainte-Marie, Lyon, France	

EXTRACURRICULAR

2019–2021	Assistant scout leader, Scoutisme Européen Suisse
2015–2016	Tutoring in Maths, physics and biology for middle school and high school students, Lyon, France

LANGUAGES

FRENCH	Native speaker
ENGLISH	Advanced, C1
SPANISH	Basic knowledge

ICT

LANGUAGES	Python, C++, Matlab, C, R
OSS	MacOS, Linux, Windows
LIBRARIES	Tensorflow, MONAI, ITK, VTK
TOOLS	git/Github, CMake
SOFTWARES	MeVisLab, 3DSlicer, ITK-SNAP

Publications

Journal papers

H. Kebiri, E.J. Canales-Rodríguez, H. Lajous, P. de Dumast, G. Girard, Y. Alemán-Gómez, M. Koob, A. Jakab and M. Bach Cuadra, "Through-plane super-resolution with autoencoders in diffusion magnetic resonance imaging of the developing human brain", in *Frontiers in Neurology*, 2022, doi:10.3389/fneur.2022.827816.

H. Lajous, C. W. Roy, T. Hilbert, P. de Dumast, S. Tourbier, Y. Alemán Gómez, J. Yerly, T. Yu, H. Kebiri, K. Payette, J. B. Ledoux, R. Meuli, P. Hagmann, A. Jakab, V. Dunet, M. Koob, T. Kober, M. Stuber and M. Bach Cuadra, "A Fetal Brain magnetic resonance Acquisition Numerical phantom (FaBiAN)", in *Scientific Reports*, 2022, doi:10.1038/s41598-022-10335-4.

K. Payette, P. de Dumast, H. Kebiri, I. Ezhov, J.C. Paetzold, S. Shit, A. Iqbal, R. Khan, R. Kottke, P. Grehten, H. Ji1, L. Lanczi, M. Nagy, M. Beresova, T. D. Nguyen, G. Natalucci, T. Karayannis, B. Menze, M. Bach Cuadra and A. Jakab, "An automatic multi-tissue human fetal brain segmentation benchmark using the Fetal tissue annotation Datase", in *Scientific Data*, 2021, doi:10.1038/s41597-021-00946-3.

P. de Dumast*, M. Khawam*, P. Deman, H. Kebiri, T. Yu, S. Tourbier, H. Lajous, P. Hagmann, P. Maeder, J.-P. Thiran, R. Meuli, V. Dunet, M. Bach Cuadra and M. Koob, "Fetal Brain Biometric Measurements on 3D Super-Resolution Reconstructed T2-Weighted MRI: An Intra- and Inter-observer Agreement Study", in *Frontiers in Pediatrics*, 2021 doi:10.3389/fped.2021.639746.

P. de Dumast, C. Mirabel, L. Cevidanes, A. Ruellas, M Yatabe, M. Ioshida, N. Tubau Ribera, L. Michoud, L. Gomes, C. Huang, H. Zhu, L. Muniz, B. Shoukri, B. Paniagua, M. Styner, S. Pieper, F. Budin, J.-B. Vimort, L. Pascal and J.C. Prieto, "A Web-based System For Neural Network based Classification in Temporomandibular Joint Osteoarthritis", in *Computerized Medical Imaging and Graphics*, 2018, doi:10.1016/j.compmedimag.2018.04.009.

M.L. Pinheiro, M. Yatabe, M. Ioshida, L. Orlandi, P. de Dumast and I.K. Trindade-Suedam, "Volumetric reconstruction and determination of minimum crosssectional area of the pharynx in patients with cleft lip and palate: comparison between two different softwares", in *Journal of Applied Oral Science*, 2018, doi:10.1590/1678-7757-2017-0282.

Pre-prints

P. de Dumast, H. Kebiri, V. Dunet, M. Koob and M. Bach Cuadra, "Multidimensional topological loss for cortical plate segmentation in fetal brain MRI", doi: 10.48550/arXiv.2208.07566.

K. Payette, H. Li, P. de Dumast, R Licandro, H. Ji, M.M. Rahman Siddiquee, D. Xu, A. Myronenko, H. Liu, Y. Pei, L. Wang, Y. Peng, J. Xie, H. Zhang, G. Dong, H. Fu, G. Wang, ZH. Rieu, D. Kim, H. Gi Kim, D. Karimi, A. Gholipour, H.R. Torres, B. Oliveira, João L. Vilaça, Yang Lin, Netanell Avisdris, O. Ben-Zvi, D. Ben Bashat, L. Fidon, M. Aertsen, T. Vercauteren, D. Sobotka, G. Langs, M. Alenyà, M.I. Villanueva, O. Camara, B. Specktor Fadida, L. Joskowicz, L. Weibin, L. Yi, L. Xuesong, M. Mazher, A. Qayyum, D. Puig, H. Kebiri, Z. Zhang, X. Xu, D. Wu, K. Liao, YX. Wu, JT. Chen, Y. Xu, L. Zhao, L. Vasung, B. Menze, M. Bach Cuadra and A. Jakab, "Fetal Brain Tissue Annotation and Segmentation Challenge Results", doi: 10.48550/arXiv.2204.09573.

P. de Dumast, T. Sanchez, H. Lajous and M. Bach Cuadra, "Simulation-based parameter optimization for fetal brain MRI super-resolution reconstruction", doi: 10.48550/arXiv.2211.1427.

IN PREP.

S. Lamon, P. de Dumast, V. Dunet, L. Pomar, Y. Vial, M. Koob and M. Bach Cuadra, "Assessment of fetal corpus callosum biometry by 3D super-resolution reconstructed T2-weighted MRI ".

K. Payette, C. Steger, R. Licandro, P. de Dumast, H. Li, M. Barkovich, L. Li, M. Dannecker, C. Chen, C. Ouyang, N. McConnell, A. Miron, Y. Li, A. Uus, I. Grigorescu, P. Ramirez Gilliland, M.M. Rahman Siddiquee, D. Xu, A. Myronenko, H. Wang, Z. Huang, J. Ye, M. Alenyà, V. Comte, O. Camara, J.-B. Masson, A. Nilsson, C. Godard, M. Mazher, A. Qayyum, Y. Gao, H. Zhou, S. Gao, J. Fu, G. Dong, G. Wang, Z. Rieu, H. Yang, M. Lee, S. P   otka, M.K. Grzeszczyk, A. Sitek, L. Vargas Daza, S. Usma, P. Arbelaez, W. Lu, W. Zhang, J. Liang, R. Valabregue, A.A. Joshi, K.N. Nayak, R.M. Leahy, L. Wilhelmi, A. Dändliker, H. Ji, A. Jakov i , M. Klai , A. Ad i , P. Markovi , G. Grabari , G. Kasprian, G. Dovjak, M. Rados, L. Vasung, M. Bach Cuadra and A. Jakab, "Multi-Center Fetal Brain Tissue Annotation and Segmentation (FeTA) Challenge Results".

P. Zanchi, E. Mullier, E. Fornari, P. de Dumast, Y. Alemán-Gómez, Jean-B. Ledoux, R. Beaty, P. Hagmann and S. Denervaud, "Impact of school pedagogy on dynamic brain network development in students".

N. Rimorini, N. Bourdillon, A. Rey, S. Urben, C. Besson, J.-B. Ledoux, P. de Dumast, Y. Alemán-Gómez, E. Fornari and S. Denervaud, ""Follow your heart": Heart-to-Brain-driven interplay relates to self-congruency".

CONFERENCE PAPERS

P. de Dumast and M. Bach Cuadra, "Domain generalization in fetal brain MRI segmentation with multi- reconstruction augmentation", at the *IEEE International Symposium on Biomedical Imaging* (ISBI) 2023.

P. de Dumast, H. Kebiri, K. Payette, A. Jakab, H. Lajous*, M. Bach Cuadra*, "Synthetic magnetic resonance images for domain adaptation: Application to fetal brain tissue segmentation", at the *IEEE International Symposium on Biomedical Imaging* (ISBI) 2022, doi:10.1109/ISBI52829.2022.9761451.

P. de Dumast, H. Kebiri, C. Atat, V. Dunet, M. Koob and M. Bach Cuadra, "Segmentation of the cortical plate in fetal brain MRI with a topological loss", at the 6th International Workshop *Perinatal, Preterm and Paediatric Image Analysis* (PIPPI) held in cunjunction with the 24th International Conference on *Medical Image Computing and Computer Assisted Intervention* (MICCAI) 2021, doi:202110.1007/978-3-030-87735-4_19.

H. Lajous, T. Hilbert, C.W. Roy, S. Tourbier, P. de Dumast, Y. Alemán-Gómez, T. Yu, H. Kebiri, J.-B. Ledoux, P. Hagmann, R. Meuli, V. Dunet, M. Koob, M. Stuber, T. Kober and M. Bach Cuadra , "Simulated Half-Fourier Acquisitions Single-shot Turbo Spin Echo (HASTE) of the Fetal Brain: Application to Super-Resolution Reconstruction", at the 6th International Workshop *Perinatal, Preterm and Paediatric Image Analysis* (PIPPI) held in cunjunction with the 24th International Conference on *Medical Image Computing and Computer Assisted Intervention* (MICCAI) 2021, doi:10.1007/978-3-030-87735-4_15.

P. de Dumast, C. Mirabel, B. Paniagua, M. Yatabe, A. Ruellas, N. Tubau, M. Styner, L. Cevidanes and J.C. Prieto, "SVA: shape variation analyzer", at the *SPIE: medical imaging* in 2018, doi:10.1117/12.2295631.

H. Lajous, P. Gordaliza, A. le Boeuf Fló, P. de Dumast, O. Esteban, M. Bach Cuadra, "Minimizing the need for clinical MRI data in supervised fetal brain tissue segmentation", at the 2023 conference of the *Organization for Human Brain Mapping* (OHBM) in 2023.

P. de Dumast, H. Kebiri, M. Bach Cuadra, H. Lajous. M. Bach Cuadra*, H. Lajous*, "On the importance of fetal brain numerical models for domain adaptation strategies in fetal brain MRI tissue segmentation", at the 31st Annual Meeting of the *International Society for Magnetic Resonance in Medicine* (ISMRM) in 2022 (oral communication).

H. Kebiri, E. Canales-Rodríguez, P. de Dumast, A. Taymourtash, H. Lajous, Y. Alemán-Gómez, G. Langs and M. Bach Cuadra, "Through-plane diffusion MRI super-resolution with autoencoders: validation on outlier replacement scheme for pre-term baby brains", at the 31st Annual Meeting of the *International Society for Magnetic Resonance in Medicine* (ISMRM) in 2022.

P. de Dumast, P. Deman, M. Khawam, T. Yu, S. Tourbier, H. Lajous, P. Hagmann, P. Maeder, J.-Ph. Thiran, R. Meuli, V. Dunet, M. Koob and M. Bach Cuadra, "Translating fetal brain magnetic resonance image super-resolution reconstruction into the clinical environment", at the *European Congress of Magnetic Resonance in Neuropediatrics* (ECMRN) in 2020 (oral communication).

M. Khawam, P. de Dumast, P. Deman, H. Kebiri, T. Yu, S. Tourbier, H. Lajous, P. Hagmann, P. Maeder, J.-Ph. Thiran, R. Meuli, V. Dunet, M. Bach Cuadra and M. Koob, "Fetal brain biometrics: comparison of 2D T2-weighted and 3D volumetric super-resolution magnetic resonance imaging", at the *European Congress of Magnetic Resonance in Neuropediatrics* (ECMRN) in 2020 (oral communication).

H. Kebiri, P. de Dumast, T. Yu, H. Lajous, J.-Ph. Thiran, R. Meuli, M. Koob and M. Bach Cuadra, "Automated fetal brain segmentation of 2D magnetic resonance images: transfer learning and 3D topology correction", at the *European Congress of Magnetic Resonance in Neuropediatrics* (ECMRN) in 2020.