ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF PHYSICS AND ASTRONOMY "A. RIGHI"

SECOND CYCLE DEGREE
PHYSICS

3D U-NET DOMAIN GENERALIZATION IN FETAL BRAIN MRI SEGMENTATION

Supervisor
Prof. Daniel Remondini

Co-supervisors
Dr. Nico Curti
Dr. Gerard Martí Juan

Defended by
Simone Chiarella

18–19/12/2025

Academic Year 2024/2025

ABSTRACT

This work investigates domain generalization of a 3D U-Net (nnU-Net v2.4.1) for fetal brain MRI tissue segmentation. The central question was whether GIN-IPA, a promising but not widely studied image augmentation method, would improve the network robustness across acquisition shifts. Three datasets with different super-resolution reconstruction (SRR) were employed: Kispi-mial (small size, both healthy and pathological cases), Kispi-irtk (similar to Kispi-mial but higher quality), and dHCP (larger size, high quality, healthy only). Label sets were harmonized to a common 7-tissue scheme.

Three strategies were compared within the nnU-Net training loop: (i) default nnU-Net augmentation, comprising standard geometric and photometric transforms; (ii) GIN-IPA, an appearance perturbation that aims to mimic acquisition shifts; and (iii) the combination of both. Models were trained separately on each dataset and assessed on unseen samples from both the source dataset and external datasets, using typical metrics such as the Dice score and others.

Two conclusions were drawn. First, generalization was primarily determined by data quality and scale: training on dHCP yielded the most stable cross-domain performance, almost independent of augmentation. Second, GIN-IPA was beneficial when source data were limited and differed from the target: training on Kispi-irtk and inferring on dHCP resulted in significant improvements compared to default nnU-Net augmentation. By contrast, stacking the two augmentations was not additive and, at times, detrimental.

Limitations include the small size of public clinical datasets, the need for label-set harmonization, and hardware constraints that prevented a full exploration of the GIN-IPA potential. The results suggest that multi-center, high-quality fetal MRI with standardized SRR should be prioritized. Within constrained single-source regimes, GIN-IPA may represent a pragmatic option for domain generalization, despite the need for further validation.

Agli incroci.

CONTENTS

SUMMARY 1

I INTRODUCTION 3

1 MAGNETIC RESONANCE IMAGING 5
- 1.1 Physical Principles . . . 5
- 1.2 Relaxation and Signal Generation . . . 7
- 1.3 Spatial Encoding . . . 10

2 FETAL BRAIN MRI 13
- 2.1 Fetal Brain Structures . . . 13
- 2.2 Acquisition Protocols . . . 15
- 2.3 Parcellation Protocols . . . 16
- 2.4 Super-Resolution Reconstruction . . . 17
- 2.5 Challenges . . . 18

3 TECHNIQUES FOR FETAL BRAIN SEGMENTATION 21
- 3.1 Domain Generalization . . . 21
- 3.2 Convolutional Neural Networks . . . 23
- 3.3 The FeTA Challenge . . . 26

II MATERIALS AND METHODS 31

4 DATA 33
- 4.1 Kispi . . . 33
- 4.2 dHCP . . . 35
- 4.3 Data Organization and File Format . . . 37

5 METHODS 39
- 5.1 nnU-Net . . . 39
- 5.2 GIN-IPA . . . 40
- 5.3 Performance Metrics . . . 43
- 5.4 Statistical Performance Assessment . . . 44

6 IMPLEMENTATION AND TRAINING STRATEGY 47
6.1 Model Architecture 47
6.2 Training and Validation Strategy 49
6.3 Hyperparameter Tuning 50

III RESULTS AND DISCUSSION 53

7 RESULTS 55
7.1 General Performance 55
7.2 Comparison of Model Performances 59
7.3 Performance by Pathology 66

8 DISCUSSION 69
8.1 Primacy of Data over Augmentation 69
8.2 Efficacy of GIN-IPA 70
8.3 Robustness by Pathology 70
8.4 General Outcomes 70
8.5 Limitations and Future Work 71

CONCLUSIONS 73

ACRONYMS 75

BIBLIOGRAPHY 77

APPENDIX A. SUPPLEMENTARY PLOTS 85
A.1 Low-Quality Kispi-mial Scans 85
A.2 General Performance 85
A.3 Comparison of Model Performances 85

APPENDIX B. SUPPLEMENTARY TABLES 95
B.1 Kispi-dHCP Label Match 95

List of Figures

1.1 Larmor precession of a nuclear spin in a magnetic field. © 2025 Science Info, from [3]. 7
1.2 Transverse relaxation of nuclear spins. © 2025 Informa UK Limited, from [4]. 8
1.3 Bloch equations for the longitudinal (left) and the transverse (right) relaxations in different biological tissues. © 2025 Heinrich-Heine-Universität Düsseldorf, from [5]. 9
1.4 Relationship between the timing parameters TR and TE, and the T1, T2, and proton density weightings. © 2025 AD Elster, from [6]. 9
1.5 Phase and frequency encoding on a brain MRI slice. © 2025 AD Elster, from [6]. 11
1.6 k-space sampling and the corresponding image reconstruction. 11

3.1 Schematic 3D U-Net architecture. © 2011 LMB, University of Freiburg, from [50]. 24

4.1 Top: Kispi cases gestational age distribution, stratified by health condition. Bottom: Quality assessment of Kispi cases, from [37]. 35
4.2 Example of Kispi-mial scan: axial segmentation, axial view, sagittal view, coronal view. Material from: [25, 68]. 36
4.3 Example of Kispi-irtk scan: axial segmentation, axial view, sagittal view, coronal view. Material from: [25, 68]. 36
4.4 Distributions of the gestational age and image shape of dHCP cases. 38
4.5 Example of dHCP scan: axial segmentation, axial view, sagittal view, coronal view. Material from: [69]. 38

5.1 Architecture of the GIN module. © 2022 IEEE, from [42]. 41
5.2 Architecture of the IPA augmentation scheme (A), and construction principles of pseudo-correlation maps (B). © 2022 IEEE, from [42]. 42

7.1 Dice score across datasets and labels for the nnU-Net default DA (baseline model). From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP. 56
7.2 Dice score across datasets and labels for the GIN-IPA DA model. From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP. 57
7.3 Dice score across datasets and labels for the combined DA (default + GIN-IPA) model. From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP. 58

7.4 Baseline vs. GIN-IPA: KDE plots of DSC (left) and HD95 (right) in ventricles, from models trained on Kispi-mial and inferring on Kispi-irtk. . . . 60

7.5 Baseline vs. GIN-IPA: KDE plots of DSC (top) and VS (bottom) in cortical gray matter and deep gray matter, from models trained on Kispi-mial and inferring on dHCP. 60

7.6 Baseline vs. GIN-IPA: KDE plots of DSC across each label and globally, from models trained on Kispi-irtk and inferring on dHCP. 61

7.7 GIN-IPA vs. combined DA: KDE plots of DSC (top) and HD95 (bottom) in cortical gray matter and cerebellum, from models trained on Kispi-irtk and inferring on Kispi-mial. 63

7.8 Comparison of the segmentation performance by pathology, between the baseline and the GIN-IPA augmentation models. From top to bottom: Dice score, volume similarity $\big(\text{cropped, full range is } [-2, 2]\big)$, and Hausdorff distance 95th percentile. Until the bars are brown, the metrics of the two models are equivalent. 67

7.9 Comparison of the segmentation performance by pathology, between the GIN-IPA and the combined augmentation models. From top to bottom: Dice score, volume similarity $\big(\text{cropped, full range is } [-2, 2]\big)$, and Hausdorff distance 95th percentile. Until the bars are brown, the metrics of the two models are equivalent. 68

A.1 Example of low-quality Kispi-mial scans: axial (left) and sagittal (right) views. Material from: [25, 68]. 86

A.2 Volume similarity across datasets and labels for the nnU-Net default DA (baseline model). From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP. 87

A.3 Volume similarity across datasets and labels for the GIN-IPA DA model. From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP. 88

A.4 Volume similarity across datasets and labels for the combined DA (default + GIN-IPA) model. From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP. 89

A.5 Hausdorff distance 95th percentile across datasets and labels for the nnU-Net default DA (baseline model). From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP. 90

A.6 Hausdorff distance 95th percentile across datasets and labels for the GIN-IPA DA model. From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP. 91

A.7 Hausdorff distance 95th percentile across datasets and labels for the combined DA (default + GIN-IPA) model. From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP. 92

A.8 Baseline vs. GIN-IPA: KDE plots of the volume similarity across each label and globally, from models trained on Kispi-irtk and inferring on dHCP. . 93

A.9 Baseline vs. GIN-IPA: KDE plots of the Hausdorff distance 95th percentile across each label and globally, from models trained on Kispi-irtk and inferring on dHCP. 94

List of Tables

4.1 Dataset properties. N_n and N_p respectively indicates the number of neurotypical and pathological cases. *The field strengths respectively refer to the scanners. 33

6.1 Average evaluation metrics across configurations. The best result is colored in green, the second-best in light blue. 51

7.1 Baseline vs. GIN-IPA: mean performance variation and Cohen's $|d|$ across metrics and labels, from models trained on Kispi-irtk and inferring on dHCP. To enhance comprehensibility, the absolute value of VS is shown. The variation of HD95 in WM—marked by the asterisk—is the only one that is not statistically significant (p-value < 0.01). 62

7.2 Baseline vs. GIN-IPA: mean performance variation, Wilcoxon p-value and Cohen's $|d|$ across training/inference datasets and metrics. To improve comprehensibility, the absolute value of VS is shown. Where the p-value and the Cohen's $|d|$ are missing, it is because the performance variation is either negative or negligible. †: p-value < 0.05; ‡: p-value < 0.01. *: $|d| > 0.2$; **: $|d| > 0.5$; ***: $|d| > 0.8$. 64

7.3 GIN-IPA vs. combined aug.: mean performance variation, Wilcoxon p-value and Cohen's $|d|$ across training/inference datasets and metrics. To improve comprehensibility, the absolute value of VS is shown. Where the p-value and the Cohen's $|d|$ are missing, it is because the performance variation is either negative or negligible. †: p-value < 0.05; ‡: p-value < 0.01. *: $|d| > 0.2$; **: $|d| > 0.5$; ***: $|d| > 0.8$. 65

B.1 Conversion table between Kispi and dHCP labels. 95

SUMMARY

This thesis originates from my internship at Universitat Pompeu Fabra (Barcelona, Spain), within the Erasmus+ Traineeship program. Here, I participated in the FeTA Challenge 2024, with the active contribution and support of Dr. Gerard Martí Juan, and under the supervision of Prof. Miguel Ángel González Ballester. Once back to Bologna, I continued my research employing the knowledge and tools acquired during the internship, under the guidance of Prof. Daniel Remondini and Dr. Nico Curti.

Accurate segmentation of fetal brain tissues from magnetic resonance imaging is essential to study early neurodevelopment. However, this task remains challenging due to the large variability introduced by different acquisition settings, scanner hardware, and super-resolution reconstruction pipelines. These differences produce distinct imaging domains, which can lead segmentation models to perform inconsistently when applied to data that differ from those used during training. This thesis investigates whether a 3D U-Net can be made more robust to such variability through the use of an augmentation method, GIN-IPA, originally designed for single-source domain generalization.

The study considers three fetal MRI datasets—Kispi-mial, Kispi-irtk, and dHCP—which differ in size, reconstruction method, and image quality. All data were harmonized to a common seven-tissue parcellation to allow consistent training and evaluation. As baseline architecture, the nnU-Net framework was employed in its 3D full-resolution configuration. Three data augmentation strategies were compared: (i) the default nnU-Net augmentation pipeline, (ii) GIN-IPA used as an augmentation method, and (iii) a combined strategy applying both. Each model was trained separately on each dataset and evaluated both within the same domain and on the remaining datasets. Performance assessment relied on the metrics used in the FeTA Challenge.

The results indicate that the characteristics of the training dataset play a central role in determining the stability of the segmentation across domains. Models trained on the larger and more homogeneous dHCP dataset show consistent performance on all test domains, with minimal dependence on the augmentation strategy. In contrast, models trained on the smaller Kispi datasets exhibit a stronger sensitivity to domain changes. In these cases, GIN-IPA leads to visibly improved robustness compared to the default augmentation. However, combining GIN-IPA with the standard augmentation pipeline does not yield further benefits and may reduce stability in some settings.

Overall, the findings show that high-quality training data remain the most effective means to ensure reliable cross-domain performance in fetal brain MRI segmentation. Within more restricted training scenarios, GIN-IPA represents a practical augmentation option that can improve robustness to acquisition-related variability.

The work is organized in three parts and eight chapters:

Part I. Introduction Comprises Chapters 1-3, establishing the theoretical and technical background of the thesis. It introduces the physical principles underlying magnetic resonance imaging, outlines the characteristics of fetal brain imaging, and reviews the state-of-the-art deep learning segmentation methods with a particular focus on domain generalization. The Part concludes by presenting the context of the FeTA Challenge, which serves as the field benchmark.

Part II. Materials and Methods Comprises Chapters 4-6, presenting the datasets, the methodological framework, and the experimental workflow. Specifically, the acquisition and reconstruction features of the Kispi and dHCP datasets are detailed, as well as the nnU-Net architecture and the GIN-IPA augmentation strategy, and the training configuration. It also presents the evaluation metrics and statistical tools employed.

Part III. Results and Discussion Comprises Chapters 7 and 8, which synthesizes the experimental results and discusses their broader implications. It highlights the dominant role of data quality and scale in determining cross-domain performance, and evaluates the effectiveness of GIN-IPA. Finally, limitations of the present study and directions for future research are mentioned.

PART I

INTRODUCTION

Chapter 1. Introduces the physical foundations of MRI, including nuclear spin dynamics, relaxation phenomena, and spatial encoding principles that underpin modern image formation.

Chapter 2. Describes the biological and technical aspects of fetal brain MRI, summarizing relevant anatomy, acquisition and parcellation protocols, super-resolution reconstruction methods, and the main challenges affecting image quality and segmentation.

Chapter 3. Reviews existing deep learning approaches to segmentation, emphasizing domain generalization strategies and summarizing insights from the FeTA Challenge as the field benchmark.

1 MAGNETIC RESONANCE IMAGING

Magnetic resonance imaging (MRI) is one of the most powerful and versatile techniques in modern medical imaging. Its non-invasive nature and the absence of ionizing radiation have established it as a cornerstone in both clinical diagnostics and biomedical research. MRI exploits the fundamental physical interactions between nuclear spins and magnetic fields, providing not only high-resolution structural images, but also access to functional, metabolic, and microstructural information.

This chapter goes through the fundamental concepts of MRI. First, it introduces the underlying physical principles of nuclear magnetic resonance, including spin dynamics, relaxation processes, and signal generation. Second, it describes how spatial encoding is achieved through magnetic field gradients, which allow the reconstruction of three-dimensional images.

1.1 PHYSICAL PRINCIPLES

Magnetic resonance imaging (MRI) is based on the fundamental physical phenomenon of nuclear magnetic resonance (NMR), first demonstrated experimentally by I. Rabi in 1938 and later independently verified by Purcell and Bloch in 1946 [1]. NMR describes the resonant interaction between the intrinsic magnetic moment of atomic nuclei and external magnetic fields. This effect enables the observation of nuclear spin behavior through the emission of electromagnetic radiation, providing the physical foundation for MRI, a technique capable of generating high-resolution, non-invasive images of biological tissues.

If a nucleus possesses an intrinsic angular momentum, or spin $\mathbf{I}$, the associated magnetic dipole moment is:

$$\boldsymbol{\mu} = \gamma \mathbf{I}, \tag{1.1}$$

where γ is the gyromagnetic ratio, a constant depending on the nucleus species. Only nuclei with a nonzero spin quantum number can exhibit magnetic resonance. Actually, nuclei with a integer spins are not able to produce detectable signals in MRI, because their gyromagnetic ratios are much smaller than those of half-integer spins. Among them, in biological tissues, the hydrogen nucleus (^{1}H) is the most suitable for MRI because of its high abundance in water and lipids and its relatively large γ, which yields strong detectable signals.

The result of an observation of the z-component of the angular momentum $\mathbf{I}$ of a single nucleus in its ground state is an integer or half-integer number m ranging from $-I$ to $+I$, in steps of 1. Thus, m can assume $2I+1$ distinct values, which implies $2I+1$ possible values for the measurement of the z-component of the magnetic moment

$$\mu_z = \gamma\hbar m, \tag{1.2}$$

where $\hbar$ is the reduced Planck constant.

Each value of m corresponds to a specific energy level of the nucleus. When a magnetic field $\mathbf{B_0}$ is applied, these energy levels split due to the Zeeman effect, resulting in a set of discrete energy states. The energy associated with each state is given by

$$E_m = -\gamma\hbar m B_0. \tag{1.3}$$

In the case of ^{1}H, which has $I = \frac{1}{2}$, the two possible values of the magnetic moment correspond to $m = +\frac{1}{2}$ and $m = -\frac{1}{2}$. This results, in turn, in two distinct energy states separated by $\Delta E = \gamma\hbar B_0$. The presence of discrete energy levels generates an observable quantity named magnetization $\mathbf{M}$, which is defined as

$$\mathbf{M} = N\gamma\hbar \left(\langle I_x\rangle \mathbf{i} + \langle I_y\rangle \mathbf{j} + \langle I_z\rangle \mathbf{k}\right), \tag{1.4}$$

where N is the number of nuclei with nonzero spin, and $\langle\cdot\rangle$ denotes the expectation values of the nuclear spin components along the axes. At relatively low temperature T (even room temperature) and high magnetic field $\mathbf{B_0}$, the magnetization obeys Curie's law [2]:

$$\mathbf{M} = N\frac{\gamma^2\hbar^2 I\,(I+1)}{3\mathrm{k_B}T}\mathbf{B_0}, \tag{1.5}$$

where $\mathrm{k_B}$ is the Boltzmann constant. It is important to notice that stronger magnetic fields or lower temperatures increase the detectable magnetization.

The last fundamental principle of NMR is a phenomenon, known as Larmor precession, which originates from the application of an external magnetic field to nuclei with nonzero spin. The magnetic moments $\boldsymbol{\mu}$ of the nuclei will not just align with the magnetic field, but they will describe a conical motion around $\mathbf{B_0}$ (Fig. 1.1), at the Larmor frequency:

$$\nu_\mathrm{L} = \frac{\omega_\mathrm{L}}{2\pi} = \frac{\gamma B_0}{2\pi}. \tag{1.6}$$

To perturb the equilibrium state and generate a measurable signal, a second oscillating magnetic field $\mathbf{B_1}$, also called radiofrequency (RF) pulse, is applied perpendicularly to $\mathbf{B_0}$. If the oscillation frequency of $\mathbf{B_1}$ matches ω_L, resonance occurs and the nuclei absorb energy, undergoing transitions between the energy levels. As a consequence, a nutation angle appears, which increases with the pulse duration, causing the magnetization $\mathbf{M}$ to tilt away from the z-axis by a flip angle, defined as

$$\alpha = \gamma B_1 t, \tag{1.7}$$

where t is the duration of the RF pulse. Typical pulses of 90° or 180° rotate the magnetization fully into the transverse plane or invert it, respectively.

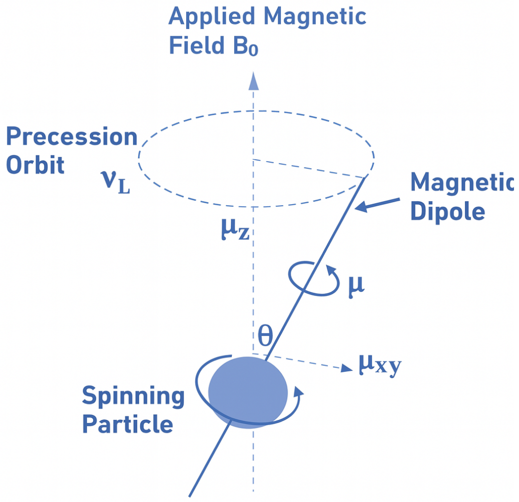

Figure 1.1: Larmor precession of a nuclear spin in a magnetic field. © 2025 Science Info, from [3].

1.2 RELAXATION AND SIGNAL GENERATION

Once the RF pulse has ended, $\mathbf{M}$ returns to the previous equilibrium state around $\mathbf{B_0}$; this induces a voltage in the receiver coil, which is the measurable NMR signal. The return to the equilibrium is not instantaneous, but it occurs through two distinct relaxation processes, characterized by different time constants and both governed by the Bloch equations. These processes are crucial for determining the contrast in MRI images, as they affect the temporal evolution of the magnetization components.

TRANSVERSE RELAXATION. Also known as spin-spin relaxation or T2 relaxation, the transverse relaxation is an entropic process that denotes the progressive loss of phase coherence among spins. It manifests as an attenuation of the observable transverse magnetization. Following the excitation imparted with the RF field $\mathbf{B_1}$, the transverse component of the magnetization M_{xy} decays because relative phases disperse over time, as illustrated in Fig. 1.2. Two principal mechanisms are implicated:

Spin-spin interactions Stochastic dipolar couplings and molecular motion induce local frequency perturbations that irreversibly randomize relative phases, giving rise to the intrinsic decay constant T_2.

Static field inhomogeneities Spatial variations in magnetic susceptibility within the specimen perturb the applied field $\mathbf{B_0}$. These distortions—whose magnitude depends on tissue composition and geometry—generate a rapid, position-dependent dephasing of nucleus spins. Refocusing methods—e.g., spin echoes—can largely reverse the effects of this dephasing factor.

Collectively, these effects produce a distribution of phases across groups of nuclei: individual spins may continue to precess near ω_0, yet their vector sum—i.e., the measurable transverse magnetization—decays toward zero.

The Bloch equation for the transverse magnetization component M_{xy} is

$$M_{xy}(t) = M_{xy}(0)\,\mathrm{e}^{-\frac{t}{T_2}}, \tag{1.8}$$

where $M_{xy}(0)$ is the transverse magnetization immediately after the end of the RF pulse. The time constant T_2 characterizes the rate of decay due to spin-spin interactions. Here the effects of static field inhomogeneities are not included; when they are, the effective decay constant is T_2^*, defined by

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_{\text{inh}}}, \tag{1.9}$$

where T_{inh} accounts for the inhomogeneity effects.

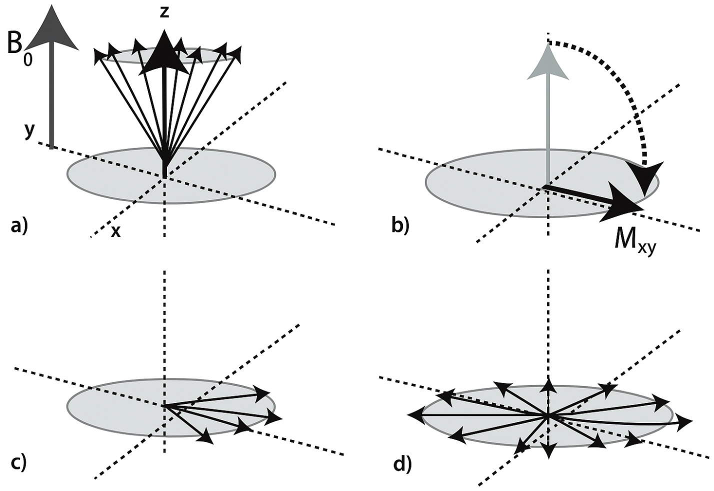

Figure 1.2: Transverse relaxation of nuclear spins. © 2025 Informa UK Limited, from [4].

Longitudinal Relaxation. Also known as spin-lattice relaxation or T1 relaxation, the longitudinal relaxation describes the recovery of the longitudinal magnetization component M_z as energy is exchanged between the spin system and its surrounding molecular environment, that is, the lattice. Once RF excitation ended, M_z reappears, returning toward its thermal equilibrium value M_0. This process is governed by the Bloch equation for the longitudinal magnetization:

$$M_z(t) = M_z(0)\,\mathrm{e}^{-\frac{t}{T_1}} + M_0\left(1 - \mathrm{e}^{-\frac{t}{T_1}}\right), \tag{1.10}$$

where $M_z(0)$ is the longitudinal magnetization immediately after the end of the RF pulse. T_1 depends on field strength, viscosity and temperature, and typically satisfies $T_1 > T_2$ in biological tissues.

The relaxation times T_1 and T_2 vary significantly among different tissue types, providing the basis for image contrast in MRI (see Fig. 1.3). By selecting appropriate timing parameters for the RF pulses and signal acquisition—e.g., the repetition time (TR) and echo time (TE)—the image contrast can be manipulated to highlight specific tissues. As a general rule, short TR and TE values yield T1-weighted images, emphasizing differences in longitudinal relaxation, while long TR and TE values produce T2-weighted images, dominated by transverse relaxation. When TR is long and TE is short, the resulting images are proton-density-weighted, reflecting variations in hydrogen concentration. The aforementioned weighting techniques are schematically illustrated in Fig. 1.4.

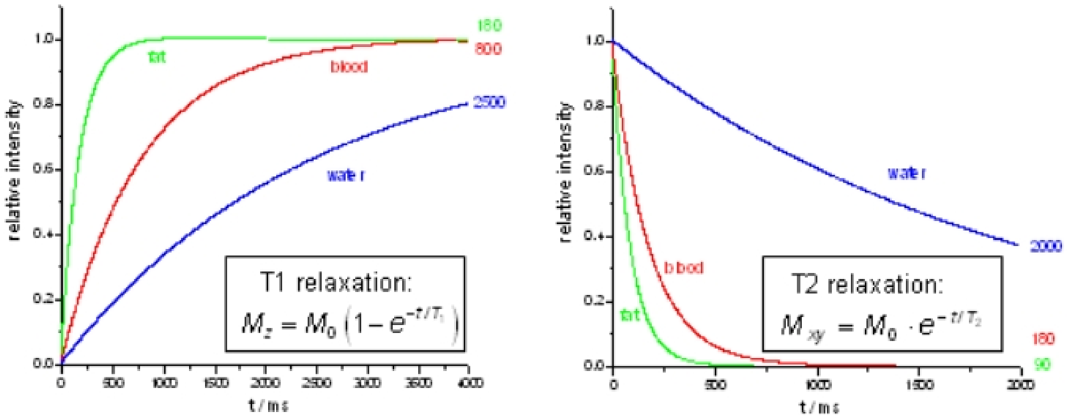

Figure 1.3: Bloch equations for the longitudinal (left) and the transverse (right) relaxations in different biological tissues. © 2025 Heinrich-Heine-Universität Düsseldorf, from [5].

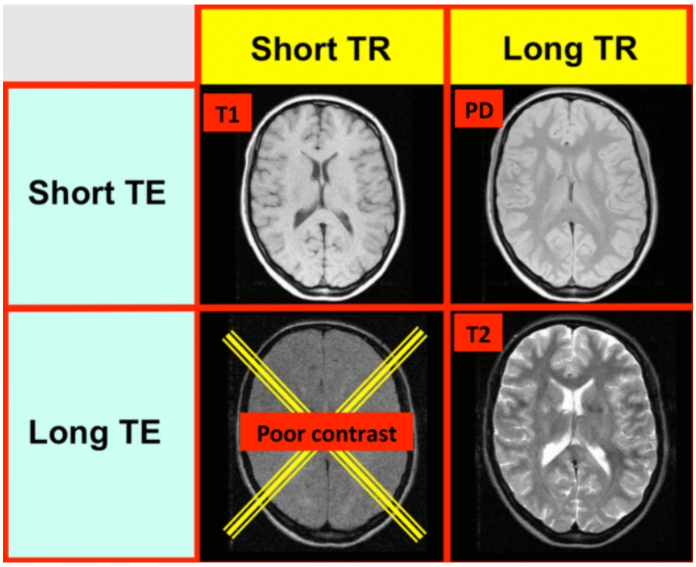

Figure 1.4: Relationship between the timing parameters TR and TE, and the T1, T2, and proton density weightings. © 2025 AD Elster, from [6].

1.3 SPATIAL ENCODING

Spatial encoding assigns a unique phase-frequency signature to each spatial location by applying linear magnetic field gradients. Given the field $\mathbf{B}$ directed along z:

$$\mathbf{G} = \frac{\partial B_z}{\partial x}\,\mathbf{i} + \frac{\partial B_z}{\partial y}\,\mathbf{j} + \frac{\partial B_z}{\partial z}\,\mathbf{k} = G_x\,\mathbf{i} + G_y\,\mathbf{j} + G_z\,\mathbf{k}, \tag{1.11}$$

the angular Larmor frequency at position $\mathbf{r}$ is

$$\omega(\mathbf{r}) = \gamma\,(B_0 + \mathbf{G}\cdot\mathbf{r}), \tag{1.12}$$

and the accumulated phase becomes

$$\phi(\mathbf{r}, t) = \int_0^t \omega(\mathbf{r}, \tau)\,\mathrm{d}\tau = \omega_0 t + \gamma\,(\mathbf{G}\cdot\mathbf{r})\,t. \tag{1.13}$$

Hence, the applied gradient encode spatial position as position-dependent phase, which is subsequently recovered by Fourier inversion. Similarly, frequency encoding can be achieved by applying a gradient during signal acquisition, causing spins at different locations to precess at different frequencies.

In three-dimensional mapping, three orthogonal gradients are applied to uniquely identify each voxel: one for slice selection (e.g., along z), one for phase encoding and one for frequency encoding (respectively along y and x). Each voxel (x, y, z) is thus identified by a specific fingerprint made up of a combination of its frequency and phase information, and its position along the z-axis, as shown in Fig. 1.5.

Finally, to describe how the acquired signal is translated into image pixels, it is convenient to introduce the wavevector

$$\mathbf{k} = \gamma\mathbf{G}t. \tag{1.14}$$

Experimentally, to obtain a 2D mapping one varies $k_x = \gamma G_x t$ and $k_y = \gamma G_y t$, since the spatial distribution of image intensities is

$$\rho(x, y) \approx \iint S(k_x, k_y)\,\mathrm{e}^{i(k_x x + k_y y)}\,\mathrm{d}k_x\,\mathrm{d}k_y, \tag{1.15}$$

where $S(k_x, k_y)$ denotes the measured signal. Therefore, it is necessary to vary k_x in M distinct ways and k_y in N distinct ways, thereby sampling the k-space at $M \times N$ points. There are many ways to sample the k-space, but the simplest is to acquire a "square" line by line. Practically, for each setting of the gradient G_y, an entire line of k-space is acquired along the x-axis; after N such acquisitions, each with a different G_y, the k-space is fully populated and the image is complete. The total acquisition time is $T_R \times N$. The inverse Fourier transformation then reconstructs the spatial image from the k-space data, as illustrated in Fig. 1.6.

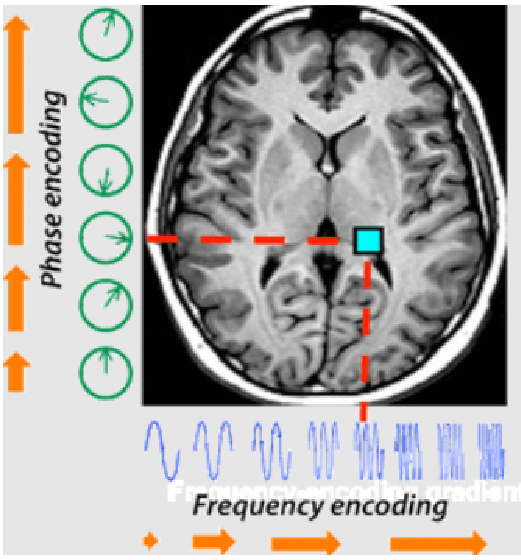
Phase encoding
Frequency encoding
Frequency-encoding gradient

Figure 1.5: Phase and frequency encoding on a brain MRI slice. © 2025 AD Elster, from [6].

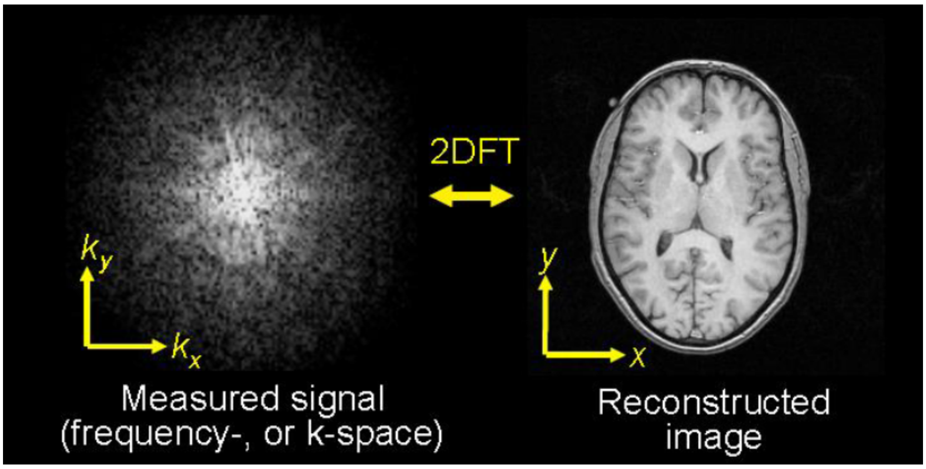
2DFT
k_y
k_x
y
x
Measured signal
(frequency-, or k-space)
Reconstructed
image

Figure 1.6: k-space sampling and the corresponding image reconstruction.

2 FETAL BRAIN MRI

MRI is an indispensable tool for the study of the developing human brain, thanks to its non-invasive nature. In the prenatal context, MRI enables the assessment of brain structures that are critical for monitoring neurodevelopment and detecting abnormalities, making fetal brain MRI a cornerstone in research settings and, when malformations are suspected, in clinical practice.

The developing brain presents features that evolve rapidly throughout gestation, requiring imaging protocols tailored to capture fine structural details while minimizing motion-related artifacts. Standardized acquisition protocols, combined with advanced post-processing techniques such as super-resolution reconstruction, are essential to obtaining images of sufficient quality for clinical diagnosis and quantitative analysis. Additionally, adequate parcellation of fetal brain structures is fundamental for studying developmental trajectories and for training automated segmentation models.

Despite these advances, fetal brain MRI remains technically challenging. Factors such as spontaneous fetal motion, maternal respiration, and variability in scanner hardware or acquisition settings introduce significant heterogeneity in the acquired data. This variability has implications not only for clinical interpretation but also for the development of automated tools.

The present chapter provides an overview of the aforementioned key aspects, presenting the main brain structures of interest during fetal development, alongside typical acquisition and parcellation protocols. Then, the most popular and effective super-resolution reconstruction algorithms are discussed. Finally, it summarizes the current challenges in the field of automated fetal brain segmentation.

2.1 FETAL BRAIN STRUCTURES

Here the fetal brain structures considered in this study are described, focusing on their function and developmental trajectory during gestation.

Cerebrospinal fluid (CSF) CSF mechanically protects the brain, mediates solute transport, and provides a regulated chemical environment for neurodevelopment. In the fetus, the choroid plexus is the dominant CSF-secretory epithelium. It forms early within the ventricular system; directional CSF flow from the lateral ventricles through third and fourth ventricles is established during mid-gestation. The blood-CSF barrier properties of the plexus epithelia mature prenatally, and CSF composition changes over gestation as neurogenesis declines after the 27th gestational week (GW) [7].

Cortical gray matter (cGM) cGM contains the neuronal bodies and local circuits that will support cortical computation. At fetal stages it is a forming laminar sheet, called cortical plate. From the 15th GW the plate thickens and transitions toward a recognizable six-layer pattern across the preterm window—from the 26th to the 36th GW. These structural reorganizations drive age-dependent MRI contrast in cGM and its interfaces with WM and CSF.

White matter (WM) White matter aggregates developing long-range axonal pathways that support inter-areal communication. In the fetus, the WM corresponds largely to the intermediate zone and periventricular crossroads, where major tracts traverse before compact myelination. Substantial myelination is minimal in utero and accelerates around the 10th fetal month and postnatally, which explains the limited T1/T2 shortening of WM in fetal scans [8, 9].

Ventricles The lateral, third, and fourth ventricles are the ducts of CSF and serve as robust anatomical landmarks for orienting fetal images. Their size and shape are clinically relevant, for instance in ventriculomegaly. The ventricular system is established early as the telencephalon expands; during mid-gestation the lateral ventricles are proportionally prominent, with progressive reduction of relative size as parenchyma (cortical plate and WM) expands. Choroid plexus maturation within the ventricles advances side by side with CSF functional maturation [10, 7].

Cerebellum The cerebellum coordinates motor processing. In the fetus it undergoes rapid volumetric growth and lobulation that are visible on T2-weighted MRI and provide robust posterior fossa landmarks. Cerebellar primordia arise early; the volume growth is more consistent after the 20th GW, with accelerated growth in the third trimester. After the 30th GW, cerebellar growth outpaces brainstem growth. These dynamics are crucial for segmentation because the cerebellar cortex exhibits layered signal changes across late gestation [11].

Deep gray matter (dGM) dGM includes thalamus and basal ganglia, which convey cortical information. Their maturation influences the timing of the cortical input to cGM. Thalamus differentiates early and sends axons to the cortical plate around the late second trimester [10].

Brainstem (BS) The brainstem houses vital autonomic and motor functions and ascending/descending tracts. It anchors long-range connectivity to the forebrain and the cerebellum. Quantitative *in vivo* fetal MRI shows specific spatiotemporal growth, with relatively faster expansion before the 30th GW and slower changes thereafter, in contrast to the cerebellum. These coordinated but offset trajectories shape the relative contrast and morphology of the posterior fossa labels in fetal MRI [12].

From the development of the aforementioned structures, three generic trends can be recognized, that impact label separability on T2-weighted fetal MRI [8]:

- The laminar reorganization of the telencephalic wall—including the cortical plate thickening—alters cGM-WM boundaries.

- The limited prenatal myelination maintains WM relatively T2-hyperintense compared to postnatal scans.

- The posterior fossa growth asynchrony—brainstem vs. cerebellum—changes the local curvature and partial-volume patterns.

2.2 Acquisition Protocols

First of all, it must be said that fetal MRI is not a screening tool, but rather a powerful, case-specific examination, when malformations are suspected. It complements ultrasonography and should be tailored to a focused diagnostic question. Protocols must optimize contrast, spatial resolution, and temporal efficiency under strict safety constraints and frequent fetal motion.

The magnetic field strength is a crucial determinant of both signal-to-noise ratio (SNR) and artifact behavior. Historically, most fetal MRI studies have been performed at $1.5\,\mathrm{T}$, which ensures stable image quality with limited dielectric and susceptibility effects. Recent technical advances, however, have enabled the transition toward $3\,\mathrm{T}$ scanners, providing an increase in SNR that can be traded for higher spatial resolution or shorter acquisition times [13, 14].

At $3\,\mathrm{T}$, dielectric artifacts, field inhomogeneities, and chemical shift distortions become more pronounced, especially in large maternal abdomens with high amniotic-fluid content. Specific absorption rate (SAR) also increases with the square of the field strength, imposing strict limits on radiofrequency power deposition. Nevertheless, multiple investigations have confirmed the absence of fetal growth retardation or auditory damage under clinically approved exposure conditions [15, 16]. The European Society of Paediatric Radiology recommends preferential use of $3\,\mathrm{T}$ for neurological and small-structure indications, particularly those involving the posterior fossa or parenchymal lesions, while $1.5\,\mathrm{T}$ remains preferable in cases of polyhydramnios or when fluid-related effects significantly degrade image quality [13, 17].

Recent technological developments have revived the use of low-field MRI, particularly at $0.55\,\mathrm{T}$, as a promising alternative for fetal imaging. Compared to conventional $1.5\,\mathrm{T}$ and $3\,\mathrm{T}$ systems, low-field scanners provide markedly improved magnetic field and radiofrequency pulse homogeneity, longer T_2^* relaxation times, and substantially reduced SAR and acoustic noise, thereby improving maternal comfort [18]. The lower field strength also reduces the occurrence of artifacts and enables the use of higher flip angles and longer echo trains without exceeding SAR limits. For instance, HASTE sequences can employ 180° refocusing pulses and bSSFP acquisitions can adopt contrast-optimal flip angles of approximately 120°, while it can reach only 60° in a $1.5\,\mathrm{T}$ setting [18].

Despite the lower polarization and signal, these limitations are largely compensated by optimized sequence design and the intrinsic relaxometric properties of tissues at low field. The larger bore diameter—70-80 cm—and reduced infrastructure requirements further extend accessibility to smaller clinical centers, potentially democratizing fetal MRI worldwide [18, 19]. Nevertheless, the reduced SNR often necessitates thicker slices or increased averaging, and advanced reconstruction or AI-based denoising strategies are being explored to overcome this constraint [19]. In [20], the authors have demonstrated the clinical feasibility and reliability of 0.55 T fetal MRI across gestational ages between 17 and 39 weeks, even though clear limitations remain about the small population sample and its representativeness of the clinical population.

Modern fetal MRI protocols are dominated by ultrafast T2-weighted and steady-state sequences capable of minimizing the effects of fetal motion. Typical sequence classes include single-shot fast spin-echo (SSFSE or HASTE), balanced steady-state free precession (bSSFP), and diffusion-weighted or intravoxel incoherent motion (IVIM) imaging.

SSFSE/HASTE It is a free-breathing sequence, often the technique of choice in fetal MRI. Each slice is acquired within a single repetition interval, making it highly robust to bulk motion while providing strong T2-weighted contrast that delineates cerebrospinal fluid, cortical gray matter, and white matter. At 3 T, dielectric shading can be mitigated by prescan filters, flip-angle adjustment, or strategic placement of saturation bands [13, 21].

bSSFP It provides high SNR and mixed T1/T2 weighting, enhancing visualization of vascular and fluid-filled structures, including the fetal heart and umbilical cord. At higher fields, frequency-offset scouting is used to shift banding artifacts outside the region of interest. These sequences are often acquired in free-breathing mode, because temporal resolution is good enough to avoid respiratory artifacts [13].

Diffusion-weighted imaging This class of sequences and the IVIM protocol provide complementary information on fetal microstructure and perfusion. They enable simultaneous estimation of diffusion and perfusion parameters, that have proven valuable in characterizing placental and fetal-brain development, distinguishing between normal and growth-restricted conditions [22, 23].

Besides the sequence, the design of the imaging geometry is critical for maximizing diagnostic yield and data quality for subsequent computational analysis. Fetal motion remains the major source of image degradation, despite employing ultrafast acquisitions that minimize motion sensitivity.

2.3 PARCELLATION PROTOCOLS

Currently, there is no established consensus in the existing fetal brain manual annotation protocols [24]. In this field, atlases, scientific papers and segmentation tools usually adopt three major schemes: the FeTA 7-label scheme, the dHCP atlas-based protocols (9-, 17- and 91-label schemes), and the BOUNTI pipeline with 19 labels.

The FeTA protocol segments the fetal brain into seven broad anatomical labels: cerebrospinal fluid, cortical grey matter, white matter, ventricles, cerebellum, deep grey matter and brainstem [25, 26]. This scheme was inspired by neonatal segmentation in the dHCP project, and targets tissue compartments that are important for detecting developmental abnormalities [25]. Its simplicity enhances robustness across varying MRI conditions and gestational ages. It is suitable for benchmarking segmentation models and clinical volumetry tasks. However, it lacks anatomical granularity and does not delineate subcortical structures individually.

The developing Human Connectome Project (dHCP) provides fetal parcellations of 9, 17 and 91 labels. The main strength is anatomical coverage combined with publicly available age-specific templates and segmentation maps. The 17-label protocol is more detailed than FeTA while still feasible for segmentation. For many structures a left-right distinction is made. However, extremely small or transient structures are not included.

BOUNTI (Brain vOlumetry and aUtomated parcellatioN for 3D feTal MRI) [24] is a deep learning pipeline trained on dHCP-derived manual labels, producing 19-region segmentations. It builds on the dHCP atlas and refines it with clinical feedback. Labels were selected for their anatomical relevance, MRI visibility, and clinical importance [24]. Its 19-label definition allows fine-grained volumetric analysis. However, it lacks subdivisions into standard anatomical regions—e.g., the frontal lobe.

As already mentioned, no single protocol has become the universal standard for fetal brain segmentation. FeTA parcellation is popular for algorithm development and benchmarking, thanks to the public datasets released in conjunction with the editions of the FeTA Challenge. dHCP and BOUNTI are favored for anatomical accuracy, with dHCP also providing a large, high-quality public dataset.

2.4 Super-Resolution Reconstruction

Super-resolution reconstruction (SRR) addresses the intrinsic anisotropy and motion corruption that characterize in-utero T2-weighted single-shot acquisitions by fusing multiple, misaligned 2D stacks into a motion-corrected, isotropic 3D volume. Fetal SRR is commonly formulated as an inverse problem, which involves finding the original scene that generates the acquired images under the imaging conditions [27]. A slice acquisition model is needed, that should be robust to motion-corrupted and mis-registered slices, and noise. Early robust estimators for bias field handling established the methodological backbone. Subsequent developments delivered total variation and edge-preserving regularization, to prevent amplification of noise and registration error [28].

Given multiple stacks of thick slices acquired in approximately orthogonal planes, SRR models each observed slice as a blurred and resampled version of an unknown high-resolution volume. Only three modern toolkits provide end-to-end functionality—brain localization, robust SRR, and standardized-space alignment—and constitute the framework for fetal brain SRR [29]: NiftyMIC [30], MIALSRTK [31], and SIMPLE IRTK [28].

An intensity-matching SRR with slice-wise bias handling and with rigid SVR was originally implemented in the SIMPLE IRTK framework [32]. It demonstrated high-quality reconstructions across challenging gestational ages and limited data, while highlighting the necessity of robust outlier rejection [28].

Building on the above, Tourbier and colleagues [31] implemented a total variation-regularized SRR, which has been extensively investigated due to its capacity to preserve tissue interfaces while controlling noise. The pipeline employs a fast convex optimization technique for SRR with adaptive regularization. The resulting MIALSRTK toolbox [33] established as a practical choice for fetal reconstructions, thanks to its resilience to motion and residual misregistrations.

NiftyMIC provides a fully automated pipeline coupling deep-learning-based brain localization and segmentation with an outlier-robust SRR and standardized template-space alignment. Controlled experiments recommend the acquisition of at least three approximately orthogonal stacks—preferably five stacks across three orientations—to ensure sufficient angular sampling and partial-volume recovery [30]. The framework has been further adapted to fetal fMRI, with Huber L2 regularization and reference-volume motion correction [34].

Recent studies have assessed SRR reliability and method-specific artifacts across the main reconstruction pipelines. A multi-rater quality assessment study [35] reported consistent quality scores and analyzed typical reconstruction artifacts among different toolkits. The results show that “excellent reliability can be achieved for global quality scoring across three raters, with good reliability on specific criteria relating to the contrast across tissues and noise levels” [35]. A complementary multi-centric biometry and volumetry study [36] reconstructed each case with multiple SRR methods to test the consistency of biometric and volumetric measurements. The paper confirms that SRR methods don’t alter the studied measurements, which would remain consistent across sites. However, the authors warn about potential intensity alterations and biases across centers and scanners.

2.5 CHALLENGES

MRI, and in particular fetal brain MRI, presents some characteristics that contribute to significant domain shifts, compromising the performance of deep learning models on unseen data distributions [26, 37].

On the biological aspect, the fetal brain is very challenging *per se*, since it undergoes rapid and complex changes throughout gestation:

Rapid Morphological Evolution The brain structure and appearance reorganize dramatically during prenatal development [26, 38]. Defining and consistently identifying different brain structures across varying gestational ages is difficult due to ongoing neuronal migration, gyrification, and sulcation patterns. This means that images acquired at different GAs constitute distinct sub-domains, making a single model challenging to generalize across the entire gestational spectrum [39].

Low Tissue Contrast The fetal brain shows different tissue contrast compared to postnatal brains. Specifically, the intensity difference between white matter and gray matter is reduced due to the absence of myelin in the fetal WM. This causes WM to appear brighter than GM on T2w images [38]. Tissue contrast changes are also significantly influenced by gestational age [26].

Pathological Heterogeneity Congenital disorders introduce further significant morphological variations. Training segmentation algorithms exclusively on neurotypical samples can reduce their robustness when encountering altered morphologies typical of pathological cases (e.g., spina bifida, ventriculomegaly, corpus callosum malformations). The rarity and wide variability of fetal pathologies contribute to the challenge of obtaining sufficient data for exhaustive datasets [26, 37].

On the technical side, the nature of *in-utero* MRI introduces further limitations [38]:

Motion Artifacts Spontaneous fetal movement and maternal breathing during MRI acquisition lead to various artifacts, such as in-plane image blur, slice crosstalk and incongruence in slice location, significantly degrading the quality of individual slices and the reconstructed volume. While ultrafast 2D sequences (e.g., SSFSE, HASTE) are commonly employed to minimize these effects by acquiring each slice rapidly, they cannot entirely eliminate motion-related issues.

Low Contrast-to-Noise Ratio The small size of the fetal brain and the need for shorter scanning periods to minimize motion contribute to a low CNR. This, coupled with acquisition limits such as thick slices—to achieve good SNR— makes it difficult to distinguish fine anatomical details.

Intensity Inhomogeneities Radiofrequency field inhomogeneity, non-uniform reception coil sensitivity, eddy currents driven by field gradients, and electromagnetic interactions with the body can cause intensity inhomogeneities. Such variations produce intensity bias fields across the image space, making consistent tissue intensity representation challenging for automated methods.

Partial Volume Effects Due to thick slices, a single voxel may contain signals from multiple tissue types. This “mixing” of signals leads to ambiguous boundaries and can result in mislabeled segmentations, especially at tissue interfaces. For instance, the mixing of the CSF and cortical GM boundary leads to intensities similar to the intensity profile of the WM.

Ultimately, the scarcity of large, standardized, and annotated datasets further jeopardizes the development of robust segmentation algorithms. [26, 40, 38] This is due to:

Small and Heterogeneous Cohorts Fetal MRI studies often require collaboration from specialized clinical centers due to the small and vulnerable patient populations. This leads to small datasets at individual institutions, making it difficult to create uniform in-house single-center datasets.

Variability in Acquisition Protocols and Hardware Data from different clinical centers tend to exhibit considerable variability in image acquisition parameters, MRI scanner hardware, and manufacturer specifications. The most important differences use to be in magnetic field strengths (e.g., $0.55\,\mathrm{T}$, $1.5\,\mathrm{T}$ and $3\,\mathrm{T}$) and specific sequence parameters (e.g., TR/TE values, flip angles, FOV, slice thickness). The associated acquisition shift cause deep learning models trained on one domain to perform poorly on data from another.

SR Reconstruction Variability Fetal MR images are typically acquired as orthogonal stacks of 2D slices to mitigate motion, which then require SR reconstruction algorithms to generate high-resolution 3D volumes. The use of different algorithms (e.g., MIALSRTK [31, 33], IRTK [28, 32], NiftyMIC [30], SVRTK [41]) introduces an additional layer of domain shift.

Manual Annotation Variability Creating high-quality, manually labeled data is time-consuming—often taking several hours per case—and requires expert anatomical knowledge. This process is also prone to human error and significant inter-rater variability.

In conclusion, the multifaceted nature of these challenges makes robust and generalizable fetal brain MRI segmentation an exceptionally difficult task.

3 TECHNIQUES FOR FETAL BRAIN SEGMENTATION

Automated segmentation of the fetal brain is a central task in the analysis of prenatal MRI, furnishing an anatomical framework for studying brain development and enabling quantitative measurements. Manual annotation, while considered the reference standard, is time-consuming, requires expert knowledge, and suffers from inter-observer variability. Consequently, a wide range of computational approaches have been developed to achieve accurate, reproducible, and efficient segmentation of fetal brain structures.

Over the years, segmentation methods have evolved from traditional atlas-based strategies to modern deep learning approaches. Convolutional neural networks have emerged as the dominant paradigm, outperforming classical approaches while remaining sensitive to differences in acquisition protocols, scanners, and populations. Domain generalization addresses the challenge of adapting models to these variations, ensuring robust performance across diverse imaging conditions.

This chapter reviews the main deep learning methods for fetal brain segmentation, introducing the concept of domain generalization and presenting the state of the art. Ultimately, it outlines the most recent results of the FeTA Challenge, which provides a benchmark for evaluating segmentation methods in a standardized setting.

3.1 DOMAIN GENERALIZATION

In medical image segmentation, a *domain* encapsulates both the feature space and the underlying marginal distribution of imaging data. Domain shifts arise when models trained on one domain are tested on data from another dataset. In medical imaging, the most prevalent sources of domain shift arise from variations in image acquisition processes, encompassing differences in imaging modalities, scanning protocols, and device manufacturers—a more exhaustive description of the potential sources can be found in Section 2.5. This is frequently referred to as "acquisition shift". Unlike data generalization, where test samples share the same distribution as training data, domain generalization addresses the scenario where test data arise from a distinct, unseen domain [42].

Domain generalization (DG) methods can be categorized into three main groups, which are often complementary and can be combined to achieve higher performance [43].

DATA MANIPULATION. This group of techniques focuses on manipulating input data to increase the diversity and quantity of existing training data.

Data Augmentation-Based DG Involves applying various transformations to training data to simulate different domain characteristics and reduce overfitting. Domain randomization is a common technique where new data is generated by introducing random variations in parameters such as object location, texture, shape, number, illumination, camera view, and noise. An example in medical imaging is SynthSeg [44], which generates synthetic images with randomized contrasts based on Gaussian mixture models (GMMs), along with geometric deformations and resampling to simulate different image resolutions. FetalSynthSeg [44] further extends this by incorporating intensity clustering and meta-classes for fetal brain MRI. Another approach is adversarial data augmentation, which generates image perturbations that are specifically designed to easily flip the predictions of classifiers, thereby making the model more robust to such changes.

Data Generation-Based DG Consists in creating diverse and rich synthetic data to enhance the model generalization capabilities. Techniques often leverage generative models such as variational auto-encoders (VAEs) and generative adversarial networks (GANs), or strategies like Mixup [45]. These methods can be complex due to their computational demands and the careful design required for the generative models.

Representation Learning. This category aims to learn feature representations that are robust to domain shifts. The core idea is to decompose the prediction function into a feature extraction (representation learning) function and a classifier function, focusing on making the feature extraction robust.

Domain-Invariant Representation Learning Seeks to reduce the discrepancies between feature distributions across multiple source domains. The underlying principle is that if feature representations remain invariant to different domains, they are inherently more general and transferable to unseen domains. This involves methods like kernel-based methods [46], domain adversarial learning—e.g., domain-adversarial neural networks, DANN [47, 48]—and invariant risk minimization [49].

Feature Disentanglement-Based DG Aims to separate learned features into domain-shared features—which are common across domains and useful for the task—and domain-specific features—which capture domain-specific variations. In this category, causality-inspired methods aim to ensure that the learned representations capture the "true cause" of the labels (e.g., object shape) and are therefore unaffected by correlated but irrelevant features like background, color, or style. Causality-inspired methods have been used to achieve single-source domain generalization through data augmentation. The idea is to simulate interventions on irrelevant features, thereby exposing the network to synthetic acquisition-shifted examples.

Single-Source Domain Generalization (ssdg) This is a specific and highly challenging setting within domain generalization where only a single source domain is available for training. This scenario is particularly common in medical imaging applications due to the high cost of data collection, privacy concerns, and scarcity of diverse datasets [42]. The key to this problem is to generate novel domains using data generation techniques to increase the diversity and informativeness of training data [43]. The performance degradation that deep learning models face due to domain shift can be attributed to [42]:

Shifted Domain-Dependent Features Image appearance, such as intensity and texture, is inherently domain-dependent. Deep networks are susceptible to shifts in these features; in contrast, human annotators can readily identify anatomical structures across different domains by focusing on domain-invariant shape information, which is intuitively causal to segmentation masks, unlike intensity or texture.

Shifted-Correlation Effect Due to confounding variables, objects in the background of an image may be spuriously correlated with the objects of interest, rather than causally related. For example, a network might learn that a certain background artifact (which, in the case of fetal MRI, would be the skull and the maternal tissues) is correlated with a specific anatomical structure within a source domain. If this artifact is absent or appears differently in an unseen target domain, the network's reliance on this spurious correlation can lead to failure.

The goal of SSDG is to mitigate these effects by steering the network towards learning domain-invariant features, such as shape information, and immunizing the segmentation model against the shifted-correlation effect by removing confounders during training. This is exactly the idea behind GIN-IPA [42], that is discussed in greater detail in Section 2.5.

3.2 CONVOLUTIONAL NEURAL NETWORKS

Convolutional neural networks (CNNs) represent a cornerstone among deep learning models for medical image segmentation. Their design is inherently suited to the spatial structure of imaging data: instead of operating on an entire image at once, CNNs employ local filters that slide across the image volume, detecting meaningful spatial patterns such as edges, boundaries, textures, and intensity gradients.

A CNN is constructed as a hierarchy of processing stages. Early layers extract lower-level features, capturing fine structural elements such as tissue interfaces and local contrast changes. As depth increases, subsequent layers integrate these features into increasingly abstract representations that encode global spatial context.

Besides the global context, prediction tasks like tissue segmentation require methods that also preserve spatial detail. Encoder-decoder architectures—such as in U-Net—are designed to meet this requirement. In the encoder, feature maps are progressively downsampled through a sequence of convolutions that reduce spatial resolution while increasing the number of feature channels. However, this hierarchical compression inevitably comes at the cost of fine-grained spatial detail. Downsampling blurs or discards the subtle boundaries that separate closely adjacent tissues. To recovering this high-frequency information, the decoder mirrors the encoder with a sequence of upsampling stages that progressively restore the spatial resolution lost during encoding. At each resolution level, the upsampled feature maps are enriched with context learned at coarser scales, allowing the model to reconstruct tissue boundaries in a globally coherent manner. The U-Net architecture introduces skip connections that link each encoder stage to its corresponding decoder stage. Skip connections transmit the high-resolution spatial information extracted early in the encoder directly to the decoder. This operation allows the decoder to leverage both global context and local detail simultaneously, preserving sharp boundaries. In fetal MRI, skip connections are particularly beneficial for preserving delicate interfaces such as the outer contour of the cortical plate, the borders of the deep gray matter, and the thin structure of the brainstem. In fetal MRI, where tissue contrast may be low, motion artefacts frequent, and partial-volume effects prominent, U-Net's ability to integrate global and local information is essential.

Three-dimensional U-Nets (illustrated in Fig 3.1) extend this scheme to volumetric data by applying convolutions and feature extraction operations in three spatial dimensions. Instead of treating each slice independently, the network processes the entire volume (or patches of it), enabling it to capture inter-slice continuity and structural coherence.

The U-Net architecture [50] has thus become the reference model for biomedical image segmentation and forms the backbone of nnU-Net, investigated in this thesis. Although 3D models impose greater computational demands, frameworks such as nnU-Net automatically adapt patch size, depth, and preprocessing to the characteristics of each dataset, providing a strong and reproducible baseline across diverse clinical and research settings [51].

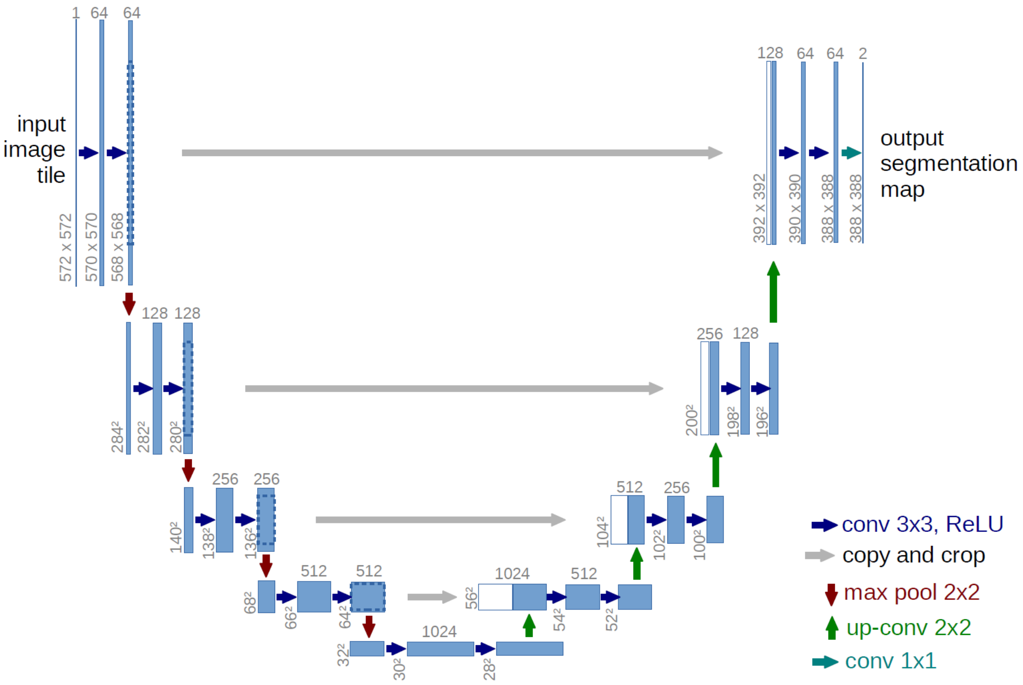
1
64
64
input image tile
572 x 572
570 x 570
568 x 568
128
128
284²
282²
280²
256
256
140²
138²
136²
512
512
68²
66²
64²
1024
32²
30²
28²
1024
512
56²
54²
52²
512
256
104²
102²
100²
256
128
200²
198²
196²
128
64
64
2
392 x 392
390 x 390
388 x 388
388 x 388
output segmentation map
conv 3x3, ReLU
copy and crop
max pool 2x2
up-conv 2x2
conv 1x1

Figure 3.1: Schematic 3D U-Net architecture. © 2011 LMB, University of Freiburg, from [50].

Building on the generic description of U-Net architectures, recent work on fetal brain MRI segmentation has converged toward increasingly sophisticated convolutional architectures that explicitly target the specific challenges of in-utero imaging. These challenges include heterogeneous image quality and acquisition protocols, rapid gestational changes in anatomy and tissue contrast, and the presence of structural abnormalities and motion artefacts.

Early work by Khalili and colleagues [52] demonstrated the feasibility of automatic multi-tissue segmentation in reconstructed fetal brain MRI using convolutional neural networks, showing clear gains over traditional atlas-based and intensity-driven methods but still struggling with generalization across gestational ages and atypical anatomies. Subsequent contributions have focused on alleviating annotation bottlenecks and improving robustness. Fetit and colleagues [53] proposed a deep learning framework for cortical grey-matter segmentation that explicitly addresses the scarcity of high-quality manual labels [53]. Their system operates on volumetric T2-weighted reconstructions from the dHCP fetal cohort and adopts a 3D multi-pathway CNN with parallel convolutional streams operating at different resolutions. Instead of relying on fully manual 3D annotations, they exploit segmentations generated by the neonatal Draw-EM pipeline to train an initial network, and then refine the model by incorporating corrections on fewer than 300 carefully selected 2D slices. This human-in-the-loop strategy dramatically reduces expert annotation workload [53].

Beyond single-tissue or cortical-only approaches, more recent architectures explicitly aim at comprehensive multi-tissue parcellation. CAS-Net (Conditional Atlas Segmentation Network) couples a 3D U-Net-style structure with a conditional atlas branch [54]. The network predicts tissue label maps and a subject-specific atlas. The atlas enables the model to learn anatomical priors without depending solely on the intensity values of the input image. This can improve the segmentation performance especially if there is no gold standard label for training due to the poor image quality. On a nine-label fetal brain parcellation, CAS-Net achieves an overall Dice similarity coefficient of approximately 85 %, highlighting the benefits of embedding a learned anatomical prior directly into the CNN architecture [54].

Another line of work tackles the intrinsic anisotropy and plane-dependent contrast of in-utero acquisitions by exploiting multi-view information. IRMMNET (Inception Residual Multi-view Multi-tissue Network) processes axial, coronal, and sagittal slices through parallel 2D encoder-decoder streams, and fuses the resulting features for joint segmentation and gestational age prediction [55]. The segmentation branch targets multi-tissue fetal brain parcellation on the FeTA dataset [25], while a regression head estimates gestational age from the shared encoder representations. Multi-view fusion allows the network to combine complementary information across orthogonal planes, mitigating ambiguities that arise when segmenting individual 2D stacks independently. IRMMNET demonstrates that multi-view 2D CNNs can compete with fully 3D architectures while being computationally lighter and more flexible with respect to slice thickness and coverage [55].

While the architectures above focus primarily on improving average segmentation accuracy, Fidon and colleagues [56] explicitly address robustness to rare but clinically critical cases. Using 3D nnU-Net as a backbone, they investigate the problem of hidden stratification in multi-tissue fetal brain segmentation—i.e., the tendency of models trained to maximize average performance to fail on under-represented subpopulations, such as fetuses with open spina bifida or other severe malformations. Their dataset consists of reconstructed T2-weighted 3D fetal brain volumes from the FeTA dataset and a private dataset [56]. Standard nnU-Net training yields high mean Dice scores but exhibits catastrophic failures for some abnormal cases, particularly in structures like cerebellum and white matter. To mitigate this, they replace empirical risk minimization—the default in nnU-Net—with a distributionally robust optimization objective that emphasizes hard examples via hardness-weighted sampling during training. The resulting nnU-Net-DRO significantly improves the lower percentiles of the Dice score distribution—for instance, increasing the worst-case performance on cerebellar segmentation in spina bifida—without degrading performance [56].

More recently, CasUNeXt introduced a cascaded 2D CNN tailored to the multi-view and multi-site nature of fetal MRI [57]. The framework comprises a global localization network (Loc-Net) and a fine segmentation network (Seg-Net), both based on encoder-decoder structures but enhanced with separable convolutions and attention gates. The Loc-Net first identifies the brain region in full-field 2D slices across axial, coronal, and sagittal views; the input volume is then cropped around the localized brain and fed to Seg-Net for high-resolution tissue segmentation. Then, the separable convolutions reduce the computational cost, while attention mechanisms selectively emphasize informative features and suppress background or artefact-related responses. On multi-view datasets comprising normal and abnormal fetal brains, CasUNeXt consistently outperforms the U-Net baseline in both localization and segmentation [57]. Notably, qualitative experiments show that CasUNeXt maintains accurate delineations in the presence of severe motion artefacts and maternal tissue interference where standard U-Net architectures generate substantial false negatives and false positives [57].

Taken together, these CNN-based approaches delineate the current state of the art in fetal brain MRI segmentation. Architecturally, most methods are rooted in U-Net-like encoder-decoder designs, either fully 3D or multi-view 2D, but they incorporate additional mechanisms: learned anatomical priors (CAS-Net), multi-task learning (IRMMNET), cascaded localization-segmentation (CasUNeXt), and robustness-oriented training objectives (nnU-Net-DRO).

3.3 THE FETA CHALLENGE

The Fetal Tissue Annotation Challenge (FeTA) [58] was born in 2020, and joined the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) [59] in 2021. Up to now, four editions have been organized (in 2020, 2021, 2022, and 2024), with increasing participation and interest from the medical imaging community. The main contributions of the FeTA Challenge are the creation of a benchmark dataset for fetal brain MRI segmentation and biometry, and the promotion of the development of algorithms for the automatic segmentation of fetal brain tissues.

The main task in FeTA is the segmentation of brain tissues in fetal MRI, which is a challenging problem due to the low contrast between tissues, the presence of noise, and the variability in the shape and size of the fetal brain. The dataset used in the challenge is composed of 3D super-resolution (SR) reconstructions of 2D fetal brain MRI images. Participants are asked to segment the fetal brain into seven tissues: external cerebrospinal fluid (CSF), cortical gray matter (cGM), white matter (WM), ventricles (including cavum), cerebellum, deep gray matter (dGM), and brainstem (BS). The performance is evaluated using three metrics: the Dice similarity coefficient (DSC), the volume similarity (VS), and the Hausdorff 95 distance (HD95). The use of three metrics helps to reduce the reliance on any one metric, which may be misleading in the evaluation of the algorithms [26].

The first edition of the FeTA Challenge was organized in 2020, by Payette et al. [25]. The challenge consisted in segmenting fetal brain MRI T2w images. The initial FeTA dataset comprised 40 super-resolution (SR) reconstructions with manual segmentations for training and 10 SR reconstructions without manual segmentation for validation, encompassing both pathological and non-pathological cases. The gestational age (GA) range spanned from 20 to 33 weeks. This dataset established a standard in fetal brain tissue parcellation—according to the seven-tissues protocol previously introduced in [60]—that would be used in all the following FeTA editions. Four research groups participated, submitting a total of ten algorithms. All the algorithms had more or less the same issues in segmenting the CSF—especially for the pathological cases, because of not clear tissue boundaries—and the GM, because of its rapidly changing structure. The dataset used in the first FeTA edition had important limitations:

- Manual segmentations were based on a single segmentation due to time and resource limitations, without consensus delineation.
- The data were from one single center, the University Children's Hospital Zurich (Kispi), thus limiting the generalizability of the results.
- The images had varying quality grades, with younger GAs and pathological cases often having lower quality.

The 2021 edition of the FeTA Challenge [40, 61] was the first to join the MICCAI conference. The dataset—hereinafter referred to as Kispi dataset—was expanded to 120 scans from the same institution, with GAs ranging from 20 to 35 weeks. The acquisition was carried out at 1.5 T for a subset of cases, and at 3 T for another subset of cases. 60 scans were reconstructed with the MIALSRTK method [31, 33], while the other 60 cases with the SIMPLE IRTK method [28, 32]. For each reconstruction method, 40 cases were included in the training dataset available to the challenge participants (for a total of 80 cases), and 20 cases were included in testing dataset not available to the participants (for a total of 40 cases). 21 algorithms were submitted, of which 19 were U-Nets, with no major differences in the architecture. Overall, the most challenging labels to segment were cortical and deep GM—due to limited image resolution and annotation uncertainty—and brainstem—especially in the pathological cases. The results of the image quality and SR reconstruction methods are related to each other, as the majority of the low quality images were done with the MIALSRTK method, and the excellent quality brain volumes included were reconstructed with the SIMPLE IRTK method.

FeTA 2022 [62, 63] introduced a multi-center dataset to address the generalizability of algorithms, which was one of the main limitations of the previous editions. In addition to Kispi, data from Medical University of Vienna was incorporated into both the training and testing datasets. Data from two further centers were included in the testing dataset—University Hospital Lausanne (CHUV), and Benioff Children's Hospital (UC San Francisco, UCSF)—for a total of four centers.

17 algorithms were submitted, among which nnU-Net was the most used and effective tool. Overall, the median performance metrics in the OOD setting remained equivalent to the in-domain, but for some labels—ventricles, GM and WM—an important drop in performance was observed. The most challenging labels to segment remained cortical and deep GM, and BS. Notably, some algorithms performed better in the OOD setting than in the in-domain setting. This can be explained with the better quality of the images from the CHUV and UCSF centers, which were included only in the test set. Style and photometric augmentations (contrast, blur, sharpness, etc.) turned out to be effective in improving the generalization of the models. However, "the optimum choice of augmentation techniques remains unclear" [62], standing as a critical factor in achieving domain generalization. The paper traces a path for future research, highlighting that it should focus on enhancing the generalizability of the methods and that "conducting a more comprehensive evaluation of the impact of data augmentation and possible biases due to super-resolution reconstruction methods would be very valuable" [62].

Finally, in FeTA 2024 [37, 58] 20 new scans were added to the test set, in order to have more results on the OOD performance. These scans were acquired at St. Thomas Hospital (King's College London, KCL), with field strength of 0.55 T (Siemens MAGNETOM Free.Max) [26]. This decision follows the recent rise in popularity of low-cost low-field MRI systems [20], which are particularly suitable for fetal imaging due to their lower SAR and acoustic noise [18]. 16 algorithms were submitted, of which nine were based on nnU-Net [64, 51]. The top team used an nnU-Net with a denoising autoencoder, generating ensemble predictions from different models. The second top team used a custom U-Net variant, training it on real and synthetic data (SynthSeg [44]), and applying post-processing to discard non-brain tissues from the predictions. All top-three teams applied extensive data augmentation, combinations of standard augmentations, and model ensembling.

Across the leading methods, average DSC plateaued between 0.80 and 0.82, likely due to the quality of both SRR algorithms and manual segmentations [62]. No statistically significant improvement was observed in overall segmentation performance metrics over the last three editions, suggesting that a performance plateau has been reached despite the increasing sophistication of methodologies and dataset diversity. These results indicate that merely architectural modifications are unlikely to produce significant improvements, consistent with observations from other challenges in which U-Net-based methods frequently outperform more complex designs [65, 37]. Notably, the low-field MRI dataset from KCL achieved the highest segmentation performance among all sites. Conversely, the Kispi dataset, in spite of being an in-domain dataset, exhibited the lowest performance. Although domain shifts are widely recognized as a key challenge for deep learning methods in medical imaging [66], the sources of these shifts are rarely disentangled. Analysis of domain shifts revealed that image quality was the most influential factor affecting model generalization, leading to Dice score differences of up to 0.10 between low- and high-quality scans. The choice of the SRR pipeline also exerted a substantial impact on segmentation performance, highlighting the need for better modeling of artifacts specific to fetal brain SR pipelines [35]. Other factors, such as gestational age, pathology, and acquisition site, contributed marginally to performance variability.

PART II

MATERIALS AND METHODS

Chapter 4. Details the datasets employed, highlighting their acquisition parameters, reconstruction pipelines, parcellation schemes, and differences in quality and pathology composition.

Chapter 5. Presents the methodological framework, describing the nnU-Net baseline, the GIN-IPA augmentation technique, and the quantitative metrics and statistical tools used for performance assessment.

Chapter 6. Explains the architecture, training configuration, and data augmentation setups of the compared models, outlining the experimental design, validation strategy, and hyperparameter optimization.

4 DATA

In this study, data from two datasets were employed: Kispi—comprising two subcohorts, Kispi-mial and Kispi-irtk—and dHCP. While details are provided in the respective sections, an overview is presented in Tab. 4.1.

Name	**N_n/N_p**	**GA range (weeks)**	**Field strength**	**Scanner**	**SRR**	**Parcel.**
Kispi-mial	15/25	20.0–32.8	1.5/3 T*	GE Signa Discovery MR450/750*	MIALSRTK [31]	7 labels
Kispi-irtk	16/24	20.01–34.8	1.5/3 T*	GE Signa Discovery MR450/750*	IRTK [28]	7 labels
dHCP	267/0	20.9–38.3	3 T	Philips Achieva	Cordero-Grande [67]	17 labels

Table 4.1: Dataset properties. N_n and N_p respectively indicates the number of neurotypical and pathological cases. *The field strengths respectively refer to the scanners.

4.1 KISPI

The Kispi dataset [25, 68] originates from the University Children's Hospital of Zurich, Switzerland, in the context of the FeTA Challenge. All data were acquired with ethics committee approval from the Canton of Zurich, and informed consent for the use of the data in research was obtained from the mothers. The dataset is open-access and fully anonymized.

The Kispi cohort comprises fetal brain MRI scans from subjects with both normal and pathological neurodevelopment. The pathological cases include a variety of congenital disorders, such as spina bifida and ventriculomegaly, reflecting a clinically relevant population [26, 39]. The dataset spans a gestational age (GA) range of approximately 20 to 35 weeks. For the FeTA 2024 challenge, the Kispi data was partitioned into a training set of 80 volumes and a test set of 40 volumes, but only the training set is publicly available. The training partition consists of 31 neurotypical and 49 pathological cases [37].

All imaging was performed on 1.5 T and 3 T GE Signa Discovery (MR450 and MR750) whole-body scanners. The acquisitions were conducted without the use of maternal or fetal sedation. Depending on the specific case, either an 8-channel cardiac coil or a standard body coil was employed [26]. The acquisition details for the single volumes are not available.

The acquisition protocol consisted of T2-weighted single-shot fast spin-echo (SSFSE) sequences acquired in the axial, coronal, and sagittal planes relative to the fetal brain. Key sequence parameters were maintained as follows [26]:

- **Repetition Time (TR):** 2000–3500 ms.
- **Echo Time (TE):** 120 ms (minimum).
- **Flip Angle:** 90°.
- **Acquisition Resolution:** An in-plane resolution of $0.5 \times 0.5\,\mathrm{mm}^2$ with a slice thickness ranging from 3 to 5 mm was employed.
- **Field of View and Matrix Size:** The FOV (200–240 mm) and image matrix (1.5 T: 256 × 224; 3 T: 320 × 224) were adjusted according to the GA and size of the fetus.

Prior to reconstruction, all acquired images for a given subject underwent a manual quality review to compile a stack of suitable scans, with at least one brain scan in each orientation. Each image stack was then reoriented to a standard anatomical plane, and a semi-automated method was used to generate the masks of the single labels in the fetal brain. Segmentations were then inspected and refined. Following these pre-processing steps, the data were processed using two distinct SR reconstruction pipelines, resulting in two sub-cohorts within the Kispi dataset. The parcellation distinguishes 7 labels: external cerebrospinal fluid (CSF), cortical gray matter (cGM), white matter (WM), ventricles (including cavum), cerebellum, deep gray matter (dGM), and brainstem (BS) [26].

40 image stacks, that is, half of the available Kispi stacks, was processed using the MIAL Super-Resolution Toolkit (MIALSRTK) pipeline [31, 33]. An example of this group of volumes—from now on referred to as the "Kispi-mial" sub-cohort—is shown in Fig. 4.2. The remaining 40 stacks were reconstructed using a pipeline based on the Image Registration Toolkit (IRTK) [28, 32]. For this sub-cohort—from now on referred to as the "Kispi-irtk"—an example is shown in Fig. 4.3.

For both SRR methods the resulting 3D volumes have an isotropic resolution of $0.5 \times 0.5 \times 0.5\,\mathrm{mm}^3$; scans were standardised to $256 \times 256 \times 256$ voxels [40]. The distributions of GA, pathology and image quality of both sub-cohorts are shown in Fig. 4.1. Even though the image quality for each volume is not provided in the open-access data—as it is only available as a plot in [37]—it can be noticed that Kispi-mial generally exhibits lower image quality compared to Kispi-irtk. Moreover, while the image quality is equivalent between the neurotypical and the pathological sub-cohorts in Kispi-irtk, in Kispi-mial the lowest quality scans are predominantly found among the pathological cases. More details about the SR algorithms are given in Section 2.4. Examples of especially low-quality scans are showed in Fig. A.1 (Appendix A).

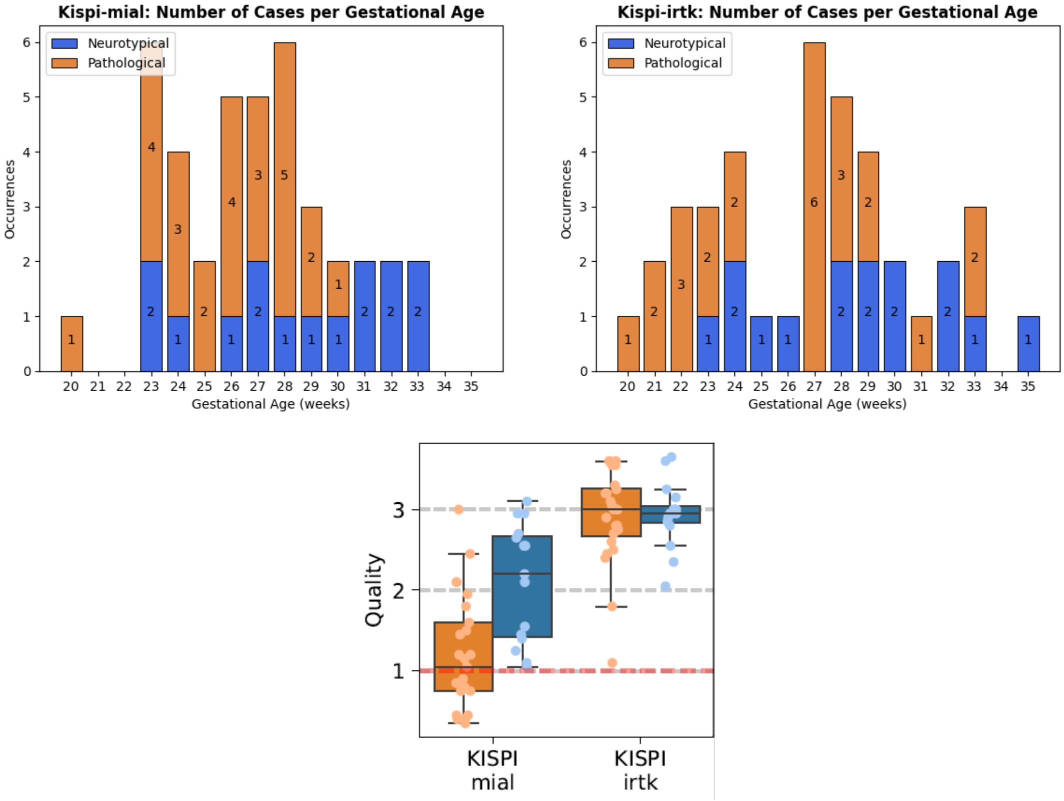

Figure 4.1: Top: Kispi cases gestational age distribution, stratified by health condition. Bottom: Quality assessment of Kispi cases, from [37].

4.2 dHCP

The Developing Human Connectome Project [69] (dHCP) provides an open-access spatio-temporal MRI atlas of normal fetal brain development. The dHCP fetal cohort consists of 296 scans from 272 individuals from St. Thomas' Hospital, London, with inclusion based on healthy pregnancies or those without major brain malformations detectable on screening ultrasound. All participants were reported by a neuroradiologist as showing age-appropriate brain anatomy on T2w anatomical scans, with no clinically significant malformations or lesions. The dataset spans a GA range from 20.86 to 38.29 weeks. All imaging was performed on a Philips Achieva 3 T using a 32-channel cardiac coil. The distributions of GA and image size are shown in Fig. 4.4. No sedation was used [70].

The dHCP fetal atlas is multi-modal. For each subject, a T2-weighted anatomical scan was acquired during the same session as the functional and diffusion scans. While specific parameters for the structural sequences are not detailed—the most complete description of the fetal structural atlas is reported in an article that actually focuses on funtional MRI [70]—the final atlas includes T2w and T1w structural channels.

The final structural atlas was constructed from multiple 2D scans combined into high-resolution 3D volumes. Raw 2D slice data were reconstructed into isotropic 3D volumes using slice-to-volume registration (SVR), following the method illustrated in [67]. The resulting volumes have a high isotropic resolution of $0.5 \times 0.5 \times 0.5\,\mathrm{mm}^3$.

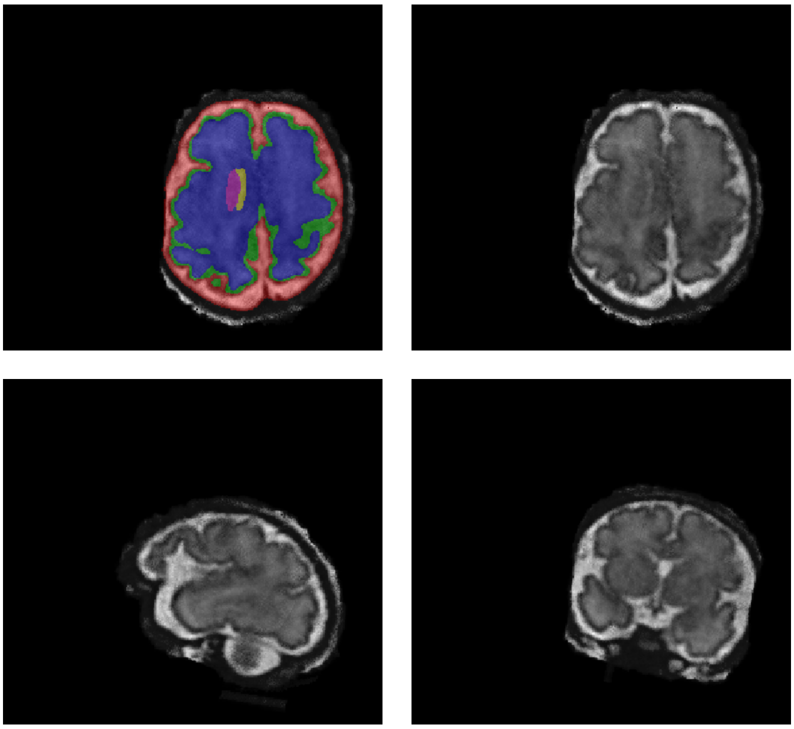

Figure 4.2: Example of Kispi-mial scan: axial segmentation, axial view, sagittal view, coronal view. Material from: [25, 68].

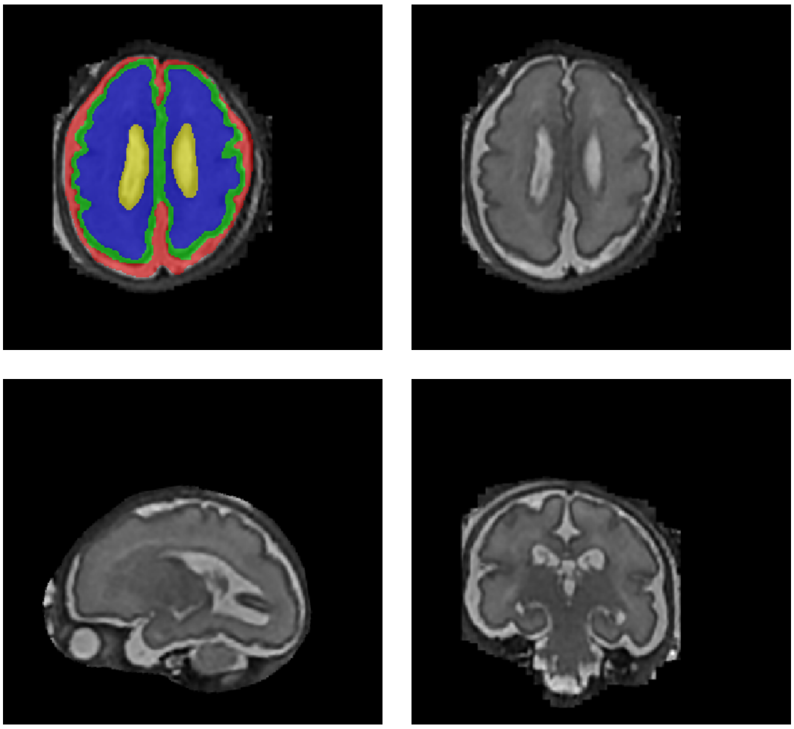

Figure 4.3: Example of Kispi-irtk scan: axial segmentation, axial view, sagittal view, coronal view. Material from: [25, 68].

The atlas features a parcellation of 17 regions of interest, which can be merged to match exactly the 7-label scheme used in the Kispi dataset (see Tab. B.1 in Appendix B). The segmentation is based on the dHCP structural pipeline [71, 72]—an example is in Fig. 4.5. Some ground-truth segmentations are missing, bringing the number of volumes to 267.

4.3 Data Organization and File Format

The two Kispi datasets had the same structure, with a folder for each case containing the reconstructed T2w volume and the corresponding segmentation mask in NIfTI format (`.nii.gz`). A `.json` file provided metadata for each volume, including the scanner manufacturer, the image resolution, and the segmentation labels. Another relevant document was a `participants.tsv` file, listing the subjects' health conditions and gestational ages.

The dHCP data were organized similarly, with each subject's folder containing the reconstructed T2w volume and its three segmentation masks—with 9, 17, and 91 labels—in NIfTI format. In this case, a `.tsv` file for each scan was provided, reporting the gestational age at the time of the scan.

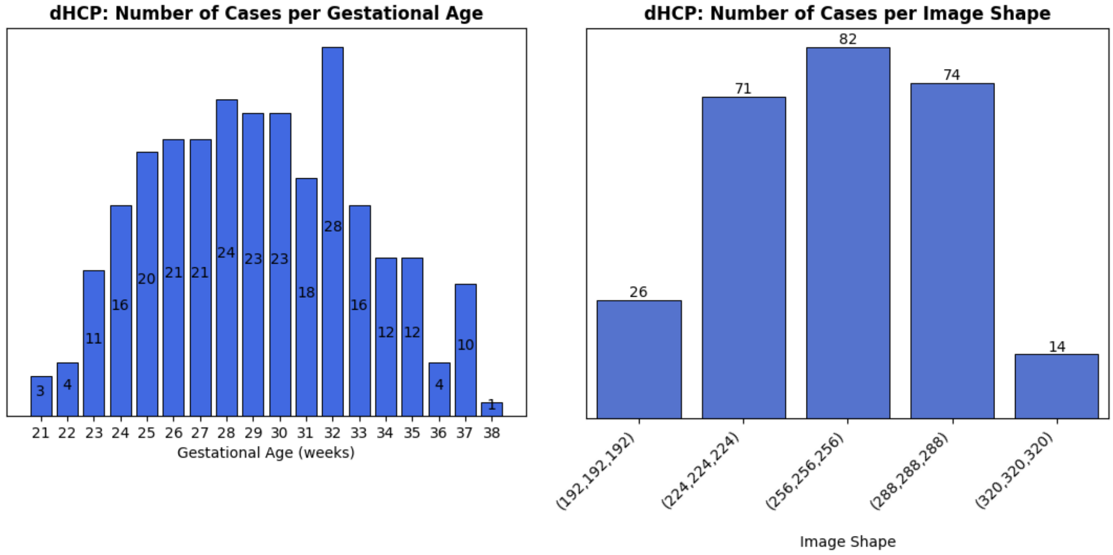

Figure 4.4: Distributions of the gestational age and image shape of dHCP cases.

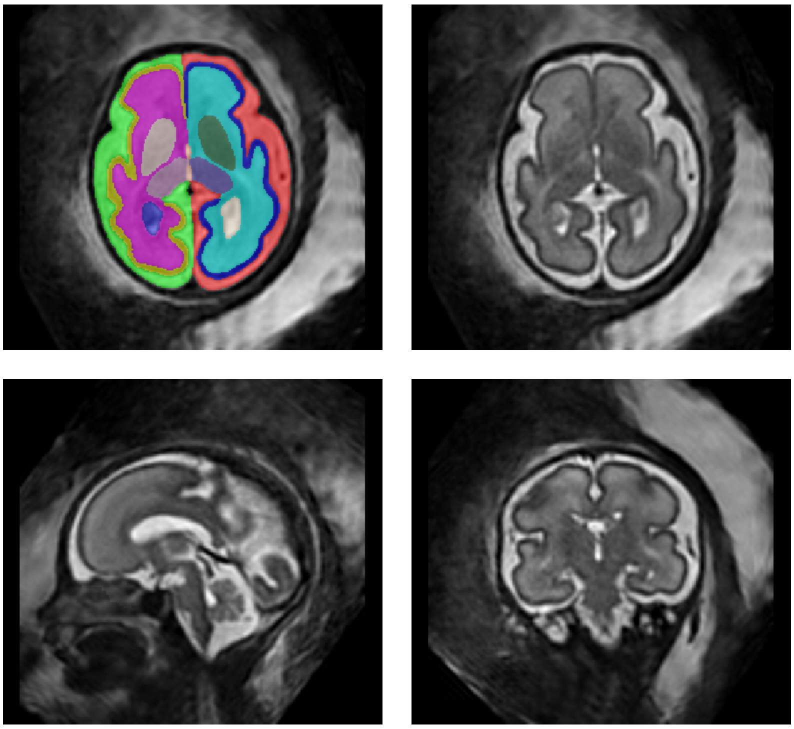

Figure 4.5: Example of dHCP scan: axial segmentation, axial view, sagittal view, coronal view. Material from: [69].

5 METHODS

This chapter describes the methods employed in the present work for fetal brain MRI segmentation and evaluation. The focus is on the domain generalization of deep learning models.

It starts with an overview of nnU-Net, a self-configuring U-Net-based framework that has become a widely adopted baseline in medical image segmentation. Building on this baseline, the method GIN-IPA is introduced as an augmentation strategy to enhance domain generalization. Finally, to objectively assess model performance a set of evaluation metrics is presented, based on segmentation accuracy. These include overlap-based measures such as the Dice coefficient, as well as surface-distance-based metrics that capture geometric fidelity.

5.1 NNU-NET

nnU-Net [51, 64] is an automated pipeline for the tissue segmentation in a task-agnostic configuration, that is, nnU-Net is designed to be applicable to a wide variety of target structures and image properties. Its self-configuring nature allows it to be applied to arbitrary new datasets without manual intervention. nnU-Net does not introduce new elements in the network architecture, but by systematizing the complex process of manual tuning it manages to achieve an automated and efficient setting of parameters in U-Net architecture. This generalized choice of pipeline parameters is carried out based on a "dataset fingerprint", that is "a standardized dataset representation comprising key properties" [51] of the images. The parameter setting strategy is based on distinguishing between three classes of parameters that can be optimized:

Fixed parameters A robust and consistent choice of hyperparameters and design implementation that is suitable for any task and dataset—they can still be manually changed though. Examples are learning rate, loss function, optimizer, training and testing procedure.

Rule-based parameters They are set based on the information taken from the dataset fingerprint, which is acquired during the preprocessing step. This information is fed into rules to automatically make several choices, such as patch and batch size, network topology, intensity normalization, and resampling strategy.

Empirical parameters Limited to the choice of the U-Net configuration (2D, 3D-fullres, 3D-cascade) and post-processing.

The rules established generate the same set of parameters and architecture when the image size and the pixel spacing are the same or similar. Default preprocessing consists in cropping the image around non-zero voxels and performing z-normalization.

Data augmentation in nnU-Net consists of a set of standard transformations: flip, rotation, scaling, Gaussian noise and blur, resolution reduction, brightness and contrast adjustments, and gamma transformation [64]. Each transformation is applied randomly with a predetermined, empirically chosen probability [51].

nnU-Net is currently one of the most popular tools for MRI fetal brain tissue segmentation, representing the state of the art in this field. Nevertheless, like other methods it suffers drops in performance, especially when it comes to inference on out-of-domain (OOD) images and on specific labels, like ventricles, gray matter and white matter.

5.2 GIN-IPA

To improve robustness of segmentation models against acquisition-induced domain shifts, the causality-inspired augmentation pipeline GIN-IPA [42] was adopted. The method, that is aimed at single-source domain generalization (see Section 3.1), treats the image formation process as generated by two independent factors, acquisition A and content C, and seeks to enforce that the segmentation predictor be invariant to interventions on A. Concretely, following the causal formulation, the desired invariance can be expressed as

$$p\Big(Y \mid S,\, \mathrm{do}(A = a_i)\Big) = p\Big(Y \mid S,\, \mathrm{do}(A = a_j)\Big), \quad \forall a_i, a_j, \tag{5.1}$$

where S denotes an ideal, domain-invariant representation determined by content C—i.e., shape information—and Y is the segmentation mask. $p\Big(Y \mid S,\, \mathrm{do}(A = a_i)\Big)$ denotes the distribution that comes from letting images to be generated from a specific acquisition process $A = a_i$—being the same for another acquisition process $A = a_j$.

Because performing real interventions on A is infeasible, GIN-IPA approximates such interventions by sampling photometric transformations $T(\cdot)$ that emulate different acquisition processes, and by enforcing consistency of network predictions across these simulated interventions.

The global intensity non-linear augmentation (GIN) module synthesizes a family of non-linear intensity and texture transforms that preserve anatomical geometry while producing diverse appearances (Fig. 5.1). Each transform is instantiated as a shallow fully-convolutional network whose weights are sampled from isotropic Gaussian priors and whose inter-layer nonlinearity is a Leaky ReLU. The overall operator is expressed as

$$g_\theta(x) = \frac{\alpha\, g_\theta^{\mathrm{Net}}(x) + (1-\alpha)\, x}{\|\alpha\, g_\theta^{\mathrm{Net}}(x) + (1-\alpha)\, x\|_{\mathrm{F}}}\, \|x\|_{\mathrm{F}}, \tag{5.2}$$

where $g_\theta^{\mathrm{Net}}(\cdot)$ denotes the pure network output for random weights θ, $\alpha \sim \mathcal{U}(0,1)$ is an interpolation coefficient and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. The normalization constrains the global energy of the augmented image, preventing global brightness or contrast changes from dominating the image. In other words, it ensures that anatomical shapes remain discernible while still allowing local texture and intensity variations to be introduced.

Key design choices include small receptive fields in the random convolutions (to avoid oversmoothing), linear interpolation with the original image (to retain semantics), and shallow depth (for computational efficiency and to avoid irrealistic results).

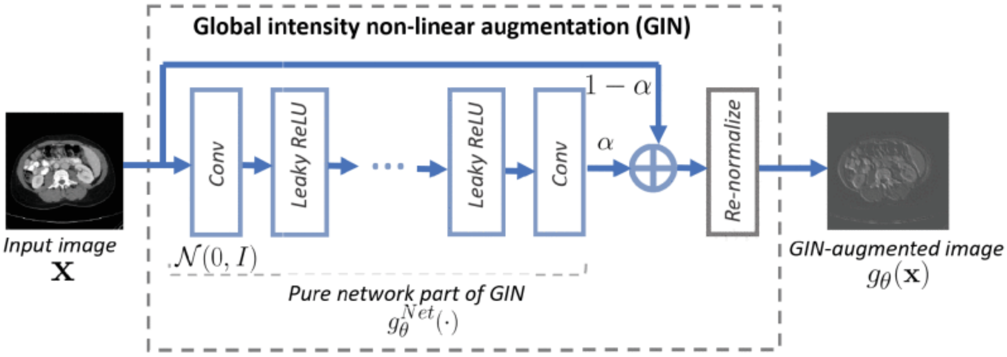

Figure 5.1: Architecture of the GIN module. © 2022 IEEE, from [42].

Then, the interventional pseudo-correlation augmentation (IPA) addresses the shifted-correlation effect, whereby background unlabeled structures X_b become spuriously correlated with objects of interest X_f due to the acquisition process. IPA approximates the causal intervention $\mathrm{do}(X_\mathrm{f} = x_\mathrm{f})$—i.e., it removes the effects of the acquisition factor A on X_f—by resampling background appearances independently of the foreground through spatially-varying assignments of appearance transforms.

Let g_{θ_1} and g_{θ_2} be two independent GIN transforms sampled for the same input image, with number of channels C, height H and width W. IPA creates a low-frequency pseudo-correlation map $b \in [0,1]$ with shape $(C \times H \times W)$ and blends the two GIN outputs as

$$T_1\Big(x; \theta_1, \theta_2, b\Big) = g_{\theta_1}(x) \odot b + g_{\theta_2}(x) \odot (1-b), \tag{5.3}$$

where $\odot$ denotes the Hadamard product, that is, element-wise multiplication. A complementary view $T_2(\cdot)$ is obtained by swapping b and $1-b$. Pseudo-correlation maps are generated by interpolating a lattice of randomly-valued control points with cubic B-splines; in practice the control-point spacing is set to a fraction of the image dimension $\left(\text{e.g., } \frac{1}{4}\right)$ to maintain low spatial frequency and avoid shape distortion (Fig. 5.2).

By applying IPA to the entire image—both background and foreground—the pipeline avoids label-induced shortcuts and approximates sampling from independent background appearance distributions. This operation can be interpreted as assigning different photometric transformations to different spatial regions so that background and foreground appearances are de-correlated across training iterations.

In [42] the training loop samples two GIN transforms and one pseudo-correlation map per image per iteration, generates the two IPA-augmented views $T_1(x)$ and $T_2(x)$, and computes the training loss. Alternative designs to B-spline in pseudo-correlation maps—e.g., random superpixels—were explored but found to be less effective. Both GIN and IPA

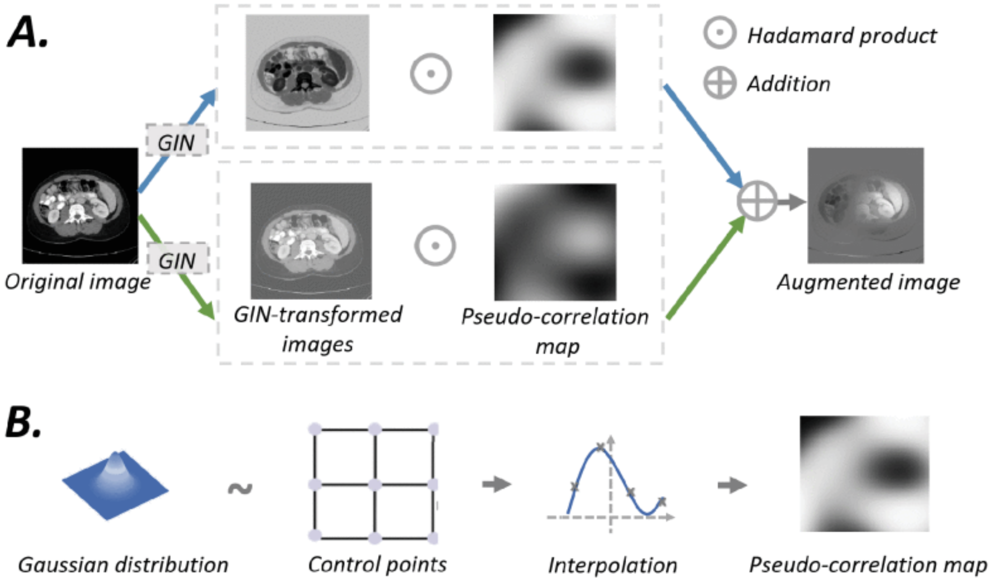
A.
GIN
GIN
Original image
GIN-transformed images
Pseudo-correlation map
Hadamard product
Addition
Augmented image
B.
Gaussian distribution
Control points
Interpolation
Pseudo-correlation map

Figure 5.2: Architecture of the IPA augmentation scheme (A), and construction principles of pseudo-correlation maps (B). © 2022 IEEE, from [42].

steps are ruled by hyperparameters that must be set empirically. Among them, the most relevant (followed by the value that was used in [42], as a reference) are:

- GIN
    - Number of layers of the shallow network: 4.
    - Number of intermediate channels of the shallow network: 2.
    - Interpolation coefficient α: sampled from a uniform distribution.
- IPA
    - Control point spacing in kernel matrix: $\frac{1}{4}$ of the image length; it must be small enough in order to construct low-frequency pseudo-correlation maps.
    - Interpolation order in kernel matrix: 2, it serves as smoothing factor between the control points in the pseudo-correlation map.
    - Control point values: sampled from a Gaussian distribution.
    - Downscale of the image: 2, reduce the computational load.

Default values were set by the authors of the paper by the combination of hyperparameter tuning, heuristic choices and hardware limitations.

Actually, practical optimizations include controlling the depth and width of the GIN networks. In [42], Ouyang and colleagues conducted hyperparameter tuning on the number of layers and channels in the GIN architecture, and on the sampling distribution of the interpolation coefficient α. An ablation study on the contribution of the IPA step to the global performance was also conducted, showing improvements in all tested scenarios with respect to GIN-only training.

The proposed approach consistently yields superior performance gains compared with other methods—MixStyle [73], AdvBias [74], RandConv [75] and others—when tested on unseen domains across three cross-domain segmentation scenarios:

- Cross-modality abdominal image segmentation (CT-MRI).
- Cross-sequence cardiac MRI segmentation (bSSFP-LGE).
- Cross-site prostate MRI segmentation.

Finally, Ouyang and colleagues also participated in the FeTA 2022 challenge [62]. Their approach was based on training different nnU-Net models with varying data augmentation methods. GIN-IPA was employed in the augmentation pipeline for one of these models [76]. The models were subsequently ensembled, and the resulting segmentation method achieved the highest performance in the challenge, ranking first. GIN augmentation and GIN-IPA were also used in three models in the FeTA 2024 challenge [37].

5.3 Performance Metrics

To assess the performance of the segmentation models, three evaluation metrics were employed: the Dice similarity coefficient (DSC), the volume similarity (VS), and the 95th percentile Hausdorff Distance (HD95). These metrics were chosen for consistency with the evaluation protocol adopted in the FeTA Challenge [40].

The Dice similarity coefficient [77, 40] measures the amount of overlap between the manual, ground-truth segmentation S_g label and the predicted segmentation S_p generated by the algorithm. It is defined as

$$\mathrm{DSC}(S_\mathrm{p}, S_\mathrm{g}) = \frac{2|S_\mathrm{p} \cap S_\mathrm{g}|}{|S_\mathrm{p}| + |S_\mathrm{g}|}, \tag{5.4}$$

where $|\cdot|$ denotes the cardinality of a set. The DSC ranges from 0 (no overlap) to 1 (perfect overlap) and is widely used in medical image segmentation tasks.

The volume similarity measures the relative agreement between the predicted volume V_p and ground-truth volume V_g. It is defined as

$$\mathrm{VS}(S_\mathrm{p}, S_\mathrm{g}) = \frac{2(V_\mathrm{p} - V_\mathrm{g})}{V_\mathrm{p} + V_\mathrm{g}}, \tag{5.5}$$

with values closer to 0—both positive and negative—indicating better agreement. Positive values indicate overestimation of the volume of the predicted label, while negative values indicate underestimation. Unlike DSC, VS is sensitive to volumetric discrepancies, regardless of spatial alignment.

The Hausdorff distance [78, 40] evaluates the geometric similarity between two surfaces by measuring the distance between their boundaries. It helps evaluating the contours of segmentations as well as the spatial positions of the voxels. To reduce sensitivity to outliers, the 95th percentile of all pairwise boundary distances is typically reported:

$$\mathrm{HD}_{95}(S_\mathrm{p}, S_\mathrm{g}) = \max\Big\{h_{95}(S_\mathrm{p}, S_\mathrm{g}),\ h_{95}(S_\mathrm{g}, S_\mathrm{p})\Big\}, \tag{5.6}$$

where $h_{95}(S_\mathrm{p}, S_\mathrm{g})$ is the 95th percentile of the distances from each boundary point in S_p to the closest point in S_g.

Together, these three metrics capture complementary aspects of segmentation quality: spatial overlap, volumetric agreement, and boundary accuracy. The metrics were computed using the scikit-image library [79, 80].

5.4 STATISTICAL PERFORMANCE ASSESSMENT

Beyond evaluating segmentation quality with DSC, VS, and HD95, it is essential to assess whether the observed differences between models are statistically significant and to quantify their magnitude. To this end, two complementary statistical tools were employed: the Wilcoxon signed-rank test and Cohen's d.

The Wilcoxon signed-rank test [81] is a non-parametric alternative to the paired t-test, used to compare two related samples without assuming normality of their differences. In our case, the test was applied to the paired metric values obtained for the same anatomical structures across the two segmentation methods. The null hypothesis states that the two models have equal performance, while the one-sided alternative tests whether the proposed method achieves superior results. The p-value was computed with the function `scipy.stats.wilcoxon` [82].

While statistical significance establishes whether a difference is unlikely to have occurred by chance, it does not provide information on the magnitude of the effect. For this reason, Cohen's d effect size [83] was used, adapted for paired samples. In this formulation, the standardized mean difference is computed on the pairwise differences of the metrics between the metric values of the proposed method and those of the baseline [84]:

$$d = \frac{\mu_1 - \mu_2}{\sigma}, \tag{5.7}$$

where μ_1 and μ_2 are the means of the proposed method and baseline, respectively, and σ is the pooled standard deviation, that for paired samples ($N_1 = N_2$) is:

$$\sigma = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}. \tag{5.8}$$

Positive values indicate a benefit of the proposed method over the baseline. Cohen's d values around 0.2, 0.5 and 0.8 are conventionally interpreted as small, medium, and large effects, respectively.

In addition to numerical statistics, kernel density estimation (KDE) plots make it possible to visualize the distribution of performance values across cases and labels. KDE is a non-parametric method for estimating the probability density function of a random variable, obtained by placing a Gaussian kernel on each data point and summing the contributions. This results in a continuous curve that approximates the underlying distribution of the observed values. Beyond offering a compact visualization, KDE plots reveal whether performance improvements are uniform across the dataset or driven by specific subsets of cases. The shape of the distribution further conveys valuable information: a narrow, unimodal density suggests consistent behavior across cases, while broader or multimodal distributions highlight variability or sensitivity to specific anatomical labels.

To summarize, the Wilcoxon signed-rank test and Cohen's d enable a comprehensive comparison: the former addresses the statistical reliability of the observed differences, while the latter quantifies their practical relevance. This dual approach ensures that improvements are not only statistically significant, but also meaningful in terms of segmentation performance. KDE plots complement those statistical methods by providing a case-sensitive view of the results.

6 IMPLEMENTATION AND TRAINING STRATEGY

The present chapter addresses the practical aspects of implementing the methods introduced earlier, focusing on model design choices, training procedures, and hyperparameter selection.

It begins with technical details of the nnU-Net architecture. The discussion then moves to the data augmentation strategies, which are the main distinguishing factor among the three models compared. Next, the training and validation strategy is described, followed by the hyperparameter tuning process, which identifies the optimal configuration for the shallow network used in GIN-IPA.

6.1 MODEL ARCHITECTURE

The model architecture is based on `nnU-Net v2.4.1`, equipped with the residual encoder preset `ResEncM` [85]. The configuration chosen is `3d-fullres`. The architecture involves 6 resolution stages for the encoder and the decoder, with a number of computational blocks of [1, 3, 4, 6, 6, 6] per stage, respectively. Three layers are present in each computational block:

- One convolution layer (`torch.nn.modules.conv.Conv3d`) with a kernel size of [3, 3, 3] and a stride of [1, 1, 1].
- One normalization layer (`torch.nn.InstanceNorm3d`).
- One non-linear activation function (`torch.nn.LeakyReLU`).

Between each encoder stage there is a downsampling layer, and between each decoder stage there is an upsampling layer. One convolution is performed between the output of each decoder stage and the skip connection from the corresponding stage in the encoder.

Further implementation details are:

- **Initialization:** Kaiming uniform distribution (`torch.nn.init.kaiming_uniform`).
- **Optimizer:** Stochastic gradient descent (`torch.optim.SGD`).

- **Learning Rate:** Initial learning rate set to 0.01, following a polynomial schedule (`torch.optim.lr_scheduler.PolynomialLR`).
- **Epochs:** Total of 1000 epochs, with 250 training iterations per epoch.
- **Loss Function:** `nnunetv2.training.loss.compound_losses.DC_and_CE_loss`, a weighted sum of Dice loss and `torch.nn.CrossEntropyLoss`. Dice loss optimizes the evaluation metric directly, but due to the patch based training, in practice merely approximates it. Combining the Dice loss with a cross-entropy loss improves training stability and segmentation accuracy [51].

Moreover, the following rule-based parameters were established by nnU-Net:

- **Batch Size:** 2
- **Patch Size:** $[128, 128, 128]$
- **Features per Stage:** $[32, 64, 128, 256, 320, 320]$

The dimensionality of both the input and output volumes is 3. The default nnU-Net preprocessing was applied (non-zero cropping, z-normalization).

In addition to the default nnU-Net data augmentation (see Section 5.1), GIN-IPA was integrated into the training pipeline, in order to assess its contribution to robustness against domain shifts. Although the experiment described in [42] is carried out with a 3D implementation, the method available online only supports 2D images. Therefore, starting from the 2D implementation, the code was adapted to work with 3D images, and integrated into the nnU-Net training pipeline. The version is publicly available at the GitHub repository `simi-99/nnUNet-ginipa` [86]. However, because of hardware limitations, it was not possible to apply the IPA step in its 3D version—the background generation of the cubic B-splines required too much RAM memory. As a compromise, GIN transformation was applied to the whole volumes, while IPA was implemented through the product between a 2D pseudo-correlation map and each of the slices of one volume, along a randomly chosen axis.

Since the purpose of this work is to validate the effectiveness of GIN-IPA as an augmentation tool to improve model robustness against domain shifts in fetal MRI segmentation, the following three models were trained, in order to test their performance. p represents the probability of application of a transformation to a volume; all of them were left to the default values defined in the nnU-Net framework.

1. **Default nnU-Net augmentation**
    - **Flip:** independently for each axis, $p = 0.5$.
    - **Rotation:** from $-30°$ to $30°$ independently for each axis, $p = 0.2$.
    - **Scaling:** from $70\,\%$ to $140\,\%$, isotropically in 3D, $p = 0.2$ (zoom in/out).
    - **Gaussian noise:** $\sigma^2 \sim \mathcal{U}(0, 0.1)$, $p = 0.1$.
    - **Gaussian blur:** kernel $\sigma \sim \mathcal{U}(0.5, 1)$, $p = 0.2$.

- **Multiplicative brightness:** from 75 % to 125 %, $p = 0.15$.
- **Contrast:** from 75 % to 125 % (preserving the global intensity range), $p = 0.15$.
- **Lower resolution:** downsample to a fraction f of the original image—being $f \sim \mathcal{U}(0.5, 1)$—then resample to the original size, $p = 0.25$.
- **Gamma transform:** $\gamma \sim \mathcal{U}(0.7, 1.5)$, preserving global intensity mean and standard deviation, $p = 0.3$.

2. **GIN-IPA augmentation**
    - **Flip:** independently for each axis, $p = 0.5$.
    - **Rotation:** from $-30°$ to $30°$ independently for each axis, $p = 0.2$.
    - **Scaling:** from 70 % to 140 %, isotropically in 3D, $p = 0.2$ (zoom in/out).
    - **GIN-IPA** ($p = 0.5$):
        - **GIN:**
            - ◇ **Number of layers:** 2 (from hyperparameter tuning, Sec. 6.3).
            - ◇ **Intermediate channels:** 4 (from hyperparameter tuning, Sec. 6.3).
            - ◇ **Interpolation coefficient:** $\alpha \sim \mathcal{U}(0, 1)$.
        - **IPA:**
            - ◇ **Control point spacing:** $(64, 64)$.
            - ◇ **Interpolation order:** 3.
            - ◇ **Control point values:** sampled from a Gaussian distribution.
            - ◇ **Downscale factor:** 1 (no downscale).
3. **Default nnU-Net + GIN-IPA augmentations**
    - Both methods applied with the same probabilities.
    - Flip, rotation, and scaling applied once as first steps.

6.2 TRAINING AND VALIDATION STRATEGY

Each of the three datasets—Kispi-mial, Kispi-irtk and dHCP—was split into training and test sets, with a ratio of 80:20. The split was stratified both by age and pathology, to ensure that all classes were represented in both sets. In-training validation was not performed for two reasons:

- **Limited data availability:** Kispi datasets were already small, and further splitting them would have resulted in insufficient data for training.

- **Focus on generalization:** The primary goal is not to achieve the best performing model on unseen data—albeit it is desirable—but rather to evaluate the model's performance on unseen data, which is better achieved by using a dedicated test set.

For each of the three aforementioned models, three independent trainings were carried out for each of the datasets, for a total of nine models. Then, for each trained model, independent segmentation predictions were made on the three datasets: on the test split of the *in-domain* set, and on the other two *out-of-domain*, unseen datasets. This experimental configuration allows to "isolate" the contribution of each DA method to the global model performance, and to assess the robustness of the model against domain shifts.

6.3 HYPERPARAMETER TUNING

In [42], Ouyang and colleagues performed a hyperparameter tuning to determine the optimal configuration of the shallow network used in the GIN step. They conducted several trainings, varying either the number of layers (2, 4, 8 or 16) or the number of intermediate channels (2, 4, 8 or 16). In this study, their segmentation model was trained on abdominal CT images and tested on abdominal MRI images. The evaluation metric was limited to the Dice score, and the IPA step was not included. Their results indicated that the best configuration was achieved with 4 layers and 2 intermediate channels. The difference in Dice score between the best and worst configurations amounted to 7 when varying the number of layers, and 4 when varying the number of channels.

In the present analysis, adopting the configuration proposed by Ouyang and colleagues would have been bold. Although GIN-IPA is designed to be independent of the acquisition conditions, our focus is on testing it within the MRI domain, while explicitly including the IPA step, since it is a fundamental part of the method. Moreover, our evaluation is not limited to the Dice score, but also considers the volume similarity and the Hausdorff distance.

Hence, a new hyperparameter tuning on the GIN shallow network was conducted, using a grid search approach. The investigated configurations were:

- **Number of layers:** 2, 4, 8.
- **Number of intermediate channels:** 2, 4, 8.

Configurations with 16 layers or 16 channels were not explored, in order to limit training time and because both led to the worst performance in [42].

In summary, the following tables report the average values per metric for nnU-Net—with GIN-IPA as DA technique, see Model 2. in Section 6.1—trained and tested as described in Section 6.2. The best performance is highlighted in green, while the second-best is shown in light blue. Overall, the combination of 2 layers and 4 intermediate channels yielded the best results. This configuration was selected for the final model training, whose performance are analyzed in Chapter 7.

Average DC	2 channels	4 channels	8 channels
2 layers	0.752	0.753	0.748
4 layers	0.753	0.751	0.746
8 layers	0.746	0.746	0.738

Average VS	2 channels	4 channels	8 channels
2 layers	−0.114	−0.103	−0.098
4 layers	−0.134	−0.125	−0.106
8 layers	−0.154	−0.128	−0.104

Average HD95	2 channels	4 channels	8 channels
2 layers	8.34	8.33	8.56
4 layers	9.10	9.00	8.71
8 layers	10.08	9.15	8.87

Table 6.1: Average evaluation metrics across configurations. The best result is colored in green, the second-best in light blue.

Part III

Results and Discussion

Chapter 7. Reports quantitative outcomes across datasets and labels, comparing augmentation strategies through statistical analyses and evaluating model robustness with respect to data domain and pathology.

Chapter 8. Interprets the empirical results, emphasizing the predominant influence of data quality and scale over augmentation, the conditional benefits of GIN-IPA, and the limitations and implications for future research.

7 RESULTS

In this chapter are exposed the main results obtained from the experiments described in Chapter 5. The analysis focuses on comparing the performance of the three investigated models, which differ in their data augmentation (DA) strategies: the nnU-Net default DA (baseline), the GIN-IPA augmentation, and a combination of both.

First, the overall performance is reported of the models across datasets—Kispi-mial, Kispi-irtk and dHCP—and labels—cerebrospinal fluid (CSF), cortical gray matter (cGM), white matter (WM), ventricles, cerebellum, deep gray matter (dGM) and brainstem (BS). Then, a pairwise comparison between models is presented, supported by statistical analyses to assess the significance and magnitude of performance differences. Finally, the robustness of the models is evaluated in a pathology-stratified analysis in the Kispi datasets.

7.1 GENERAL PERFORMANCE

In the plots below is shown the Dice score (DSC) across datasets and labels for the three models. Each model inference is realized on the test set of the same dataset the model was trained on (in-domain), and on the whole set (both train and test) of the other datasets (out-of-domain, OOD). Given that the general performance of the models is well captured with the DSC, here only the plots related to this metric are shown. The plots of the volume similarity (VS) and the Hausdorff distance 95th percentile (HD95) are in Appendix A.

For the baseline model (see Fig. 7.1), the drop in performance between in-domain and OOD is clear in every case, except in the DSC of some labels (CSF, cGM, WM and cerebellum) for the model trained on Kispi-mial. The drop is especially evident for the models trained on the Kispi datasets when applied to dHCP. Ventricles are the most affected, but also dGM and WM. However, the change of domain does not have the same effect on the network trained on Kispi-mial as it has on the other two: for example, we have unusual high DSC values OOD for cerebellum and BS, and even higher values OOD than ID, but limited to the inference on Kispi-irtk for CSF and cGM. This is probably due to the quality of the images in Kispi-mial, which is worse than the others [40].

Although GIN-IPA (see Fig. 7.2) does not cause an increment in DSC in the models trained on Kispi-mial and dHCP, it produces a significant improvement in the model trained on Kispi-irtk when predicting on dHCP. The raise is mainly located in dGM, ventricles and BS. The average Dice passes from 0.34 to 0.54 ($58\,\%$).

Finally, the model that combines the nnU-Net default DA and GIN-IPA is substantially equivalent to the pure GIN-IPA model (see Fig. 7.3).

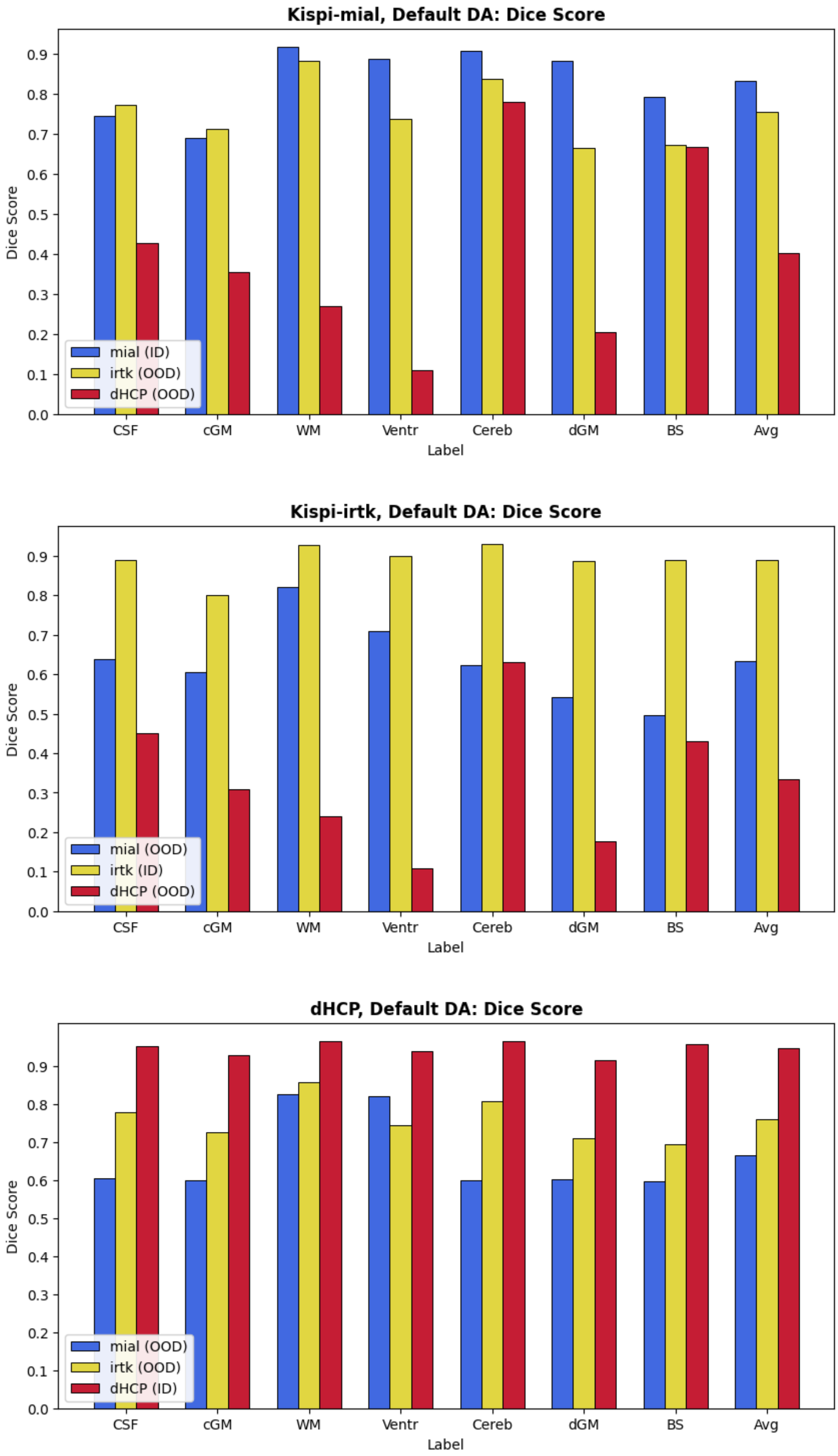

Figure 7.1: Dice score across datasets and labels for the nnU-Net default DA (baseline model). From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP.

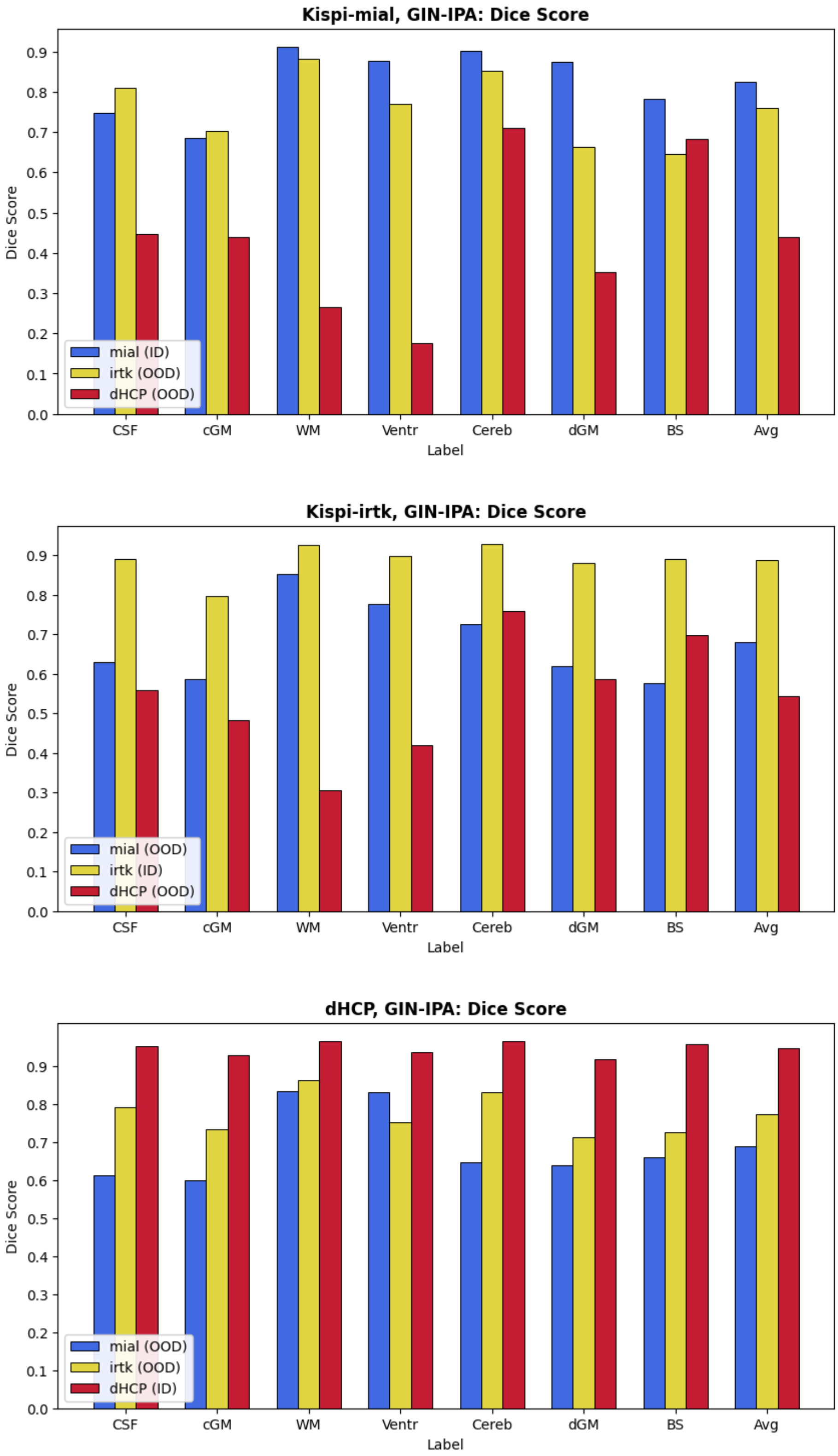

Figure 7.2: Dice score across datasets and labels for the GIN-IPA DA model. From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP.

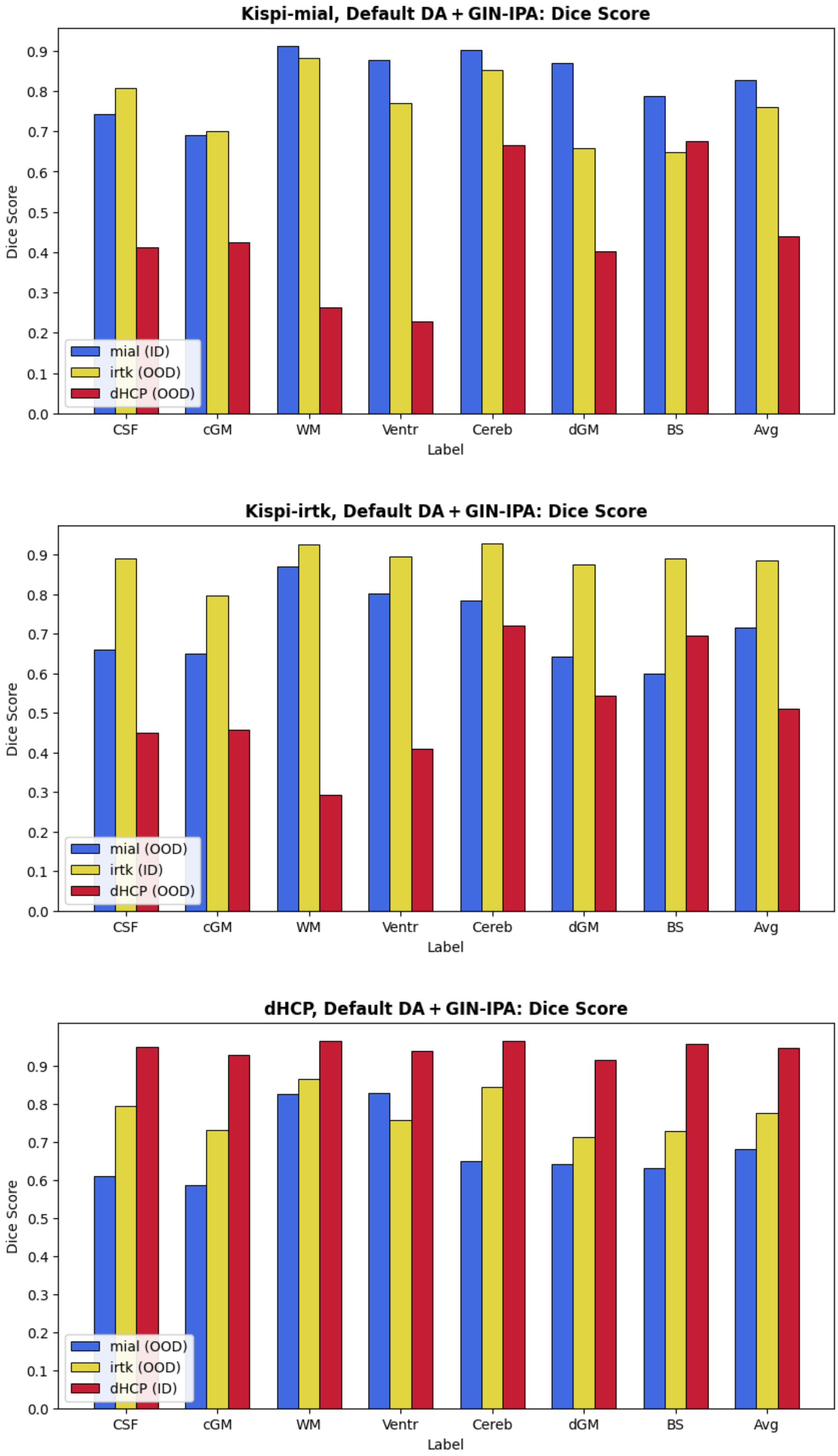

Figure 7.3: Dice score across datasets and labels for the combined DA (default + GIN-IPA) model. From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP.

7.2 COMPARISON OF MODEL PERFORMANCES

In order to assess the relative contribution of the proposed augmentation strategies, two sets of comparisons between the models were designed:

- the nnU-Net baseline with default data augmentation versus the GIN-IPA DA model
- the GIN-IPA DA model versus the combined strategy including both default and GIN-IPA augmentations

The distributions of each of the three evaluation metrics—DSC, VS, and HD95—were analyzed, at the level of individual tissues, and globally, involving all the labels.

Kernel density estimation (KDE) plots were generated for each metric and label, separately for the two model pairs under comparison. Beyond visual inspection, statistical analyses were employed to quantify the significance and magnitude of the observed differences. The Wilcoxon signed-rank test was used to test the null hypothesis of equal paired performance between models. Corresponding p-values were computed to assess whether the proposed augmentation strategy led to statistically significant improvements. Besides, Cohen's d was computed to quantify the magnitude of the improvement. Following conventional thresholds, 0.2, 0.5 and 0.8 correspond to small, medium, and large effects, respectively. See Section 5.4 for more details about the aforementioned tools. Tables with the complete statistical results are reported in Appendix B (Tabs. 7.2–7.3).

BASELINE VS. GIN-IPA

- **Train on Kispi-mial**
  - **Inference on Kispi-mial:** no difference in performance.
  - **Inference on Kispi-irtk:** CSF improves in DSC (from 0.77 to 0.81, $|d| = 0.2$), VS (from -0.17 to -0.2, $|d| = 0.6$), and HD95 (from 3.4 to 2.6, $|d| = 0.2$); ventricles slightly improve in DSC (from 0.74 to 0.77, $|d| = 0.3$) and HD95 (from 3.5 to 1.5, $|d| = 0.4$). These improvements are due to a performance improvement of the GIN-IPA model on volumes that were poorly-segmented by the baseline model (see Fig. 7.4) The same pattern is observed for other tissues (WM, cerebellum).
  - **Inference on dHCP:** significant, strong improvements in DSC and VS for cGM (DSC: from 0.36 to 0.44, VS: from -0.8 to -0.5) and dGM (DSC: from 0.20 to 0.35; VS: from -1.6 to -1.3). The corresponding KDE plots are in Fig. 7.5. Overall, a small gain is observed (DSC: from 0.40 to 0.44).
- **Train on Kispi-irtk**
  - **Inference on Kispi-mial:** significant, moderate improvements across DSC, VS, and HD95 for dGM (DSC: from 0.71 to 0.78; VS: from 0.9 to 0.7) and ventricles (DSC: from 0.36 to 0.44; VS: from -0.5 to -0.2). Small improvement overall (DSC: from 0.63 to 0.68).

  - **Inference on Kispi-irtk:** no difference in performance.
  - **Inference on dHCP:** significant, strong improvements across all metrics and tissues. The mean DSC increases from 0.34 to 0.54 (see Tab. 7.1). The KDE plots of DSC are in Fig. 7.6, while the ones regarding VS and HD95 are reported in Appendix A (Figs. A.8–A.9).

- **Train on dHCP** No differences observed on any inference dataset, metric, or label.

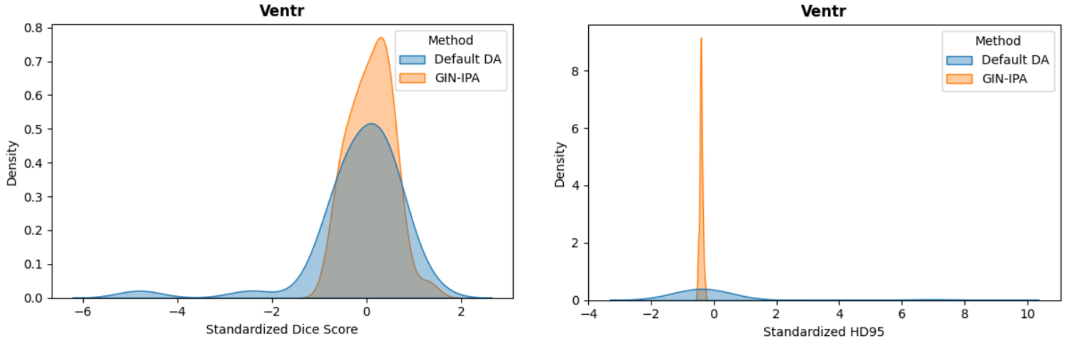

Figure 7.4: Baseline vs. GIN-IPA: KDE plots of DSC (left) and HD95 (right) in ventricles, from models trained on Kispi-mial and inferring on Kispi-irtk.

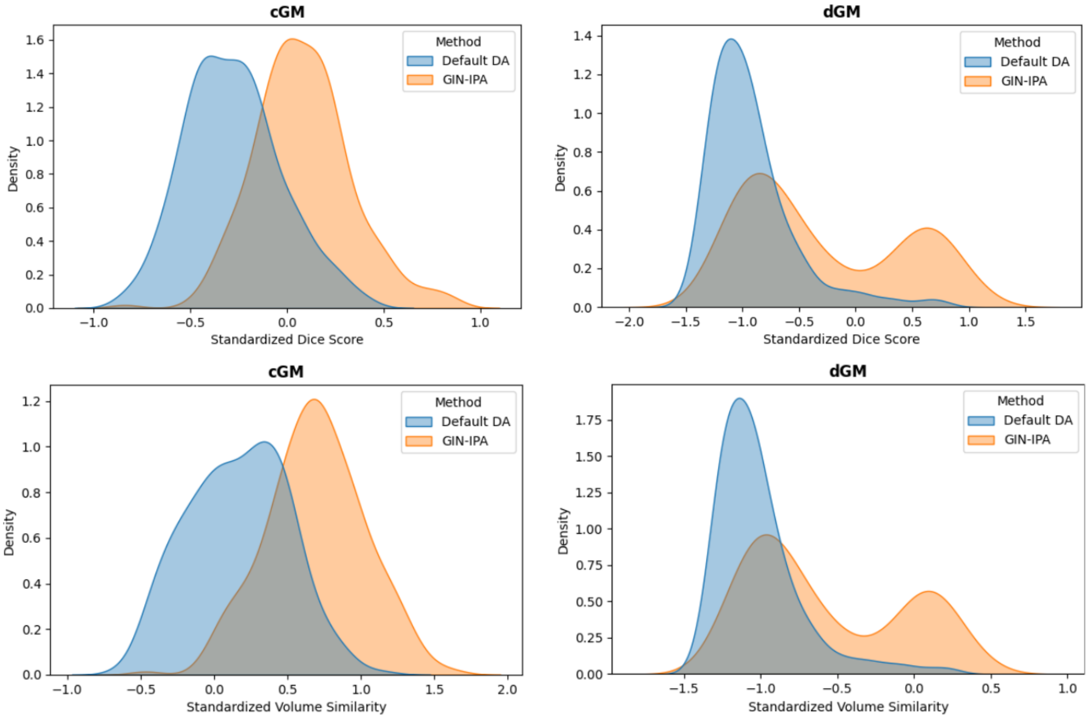

Figure 7.5: Baseline vs. GIN-IPA: KDE plots of DSC (top) and VS (bottom) in cortical gray matter and deep gray matter, from models trained on Kispi-mial and inferring on dHCP.

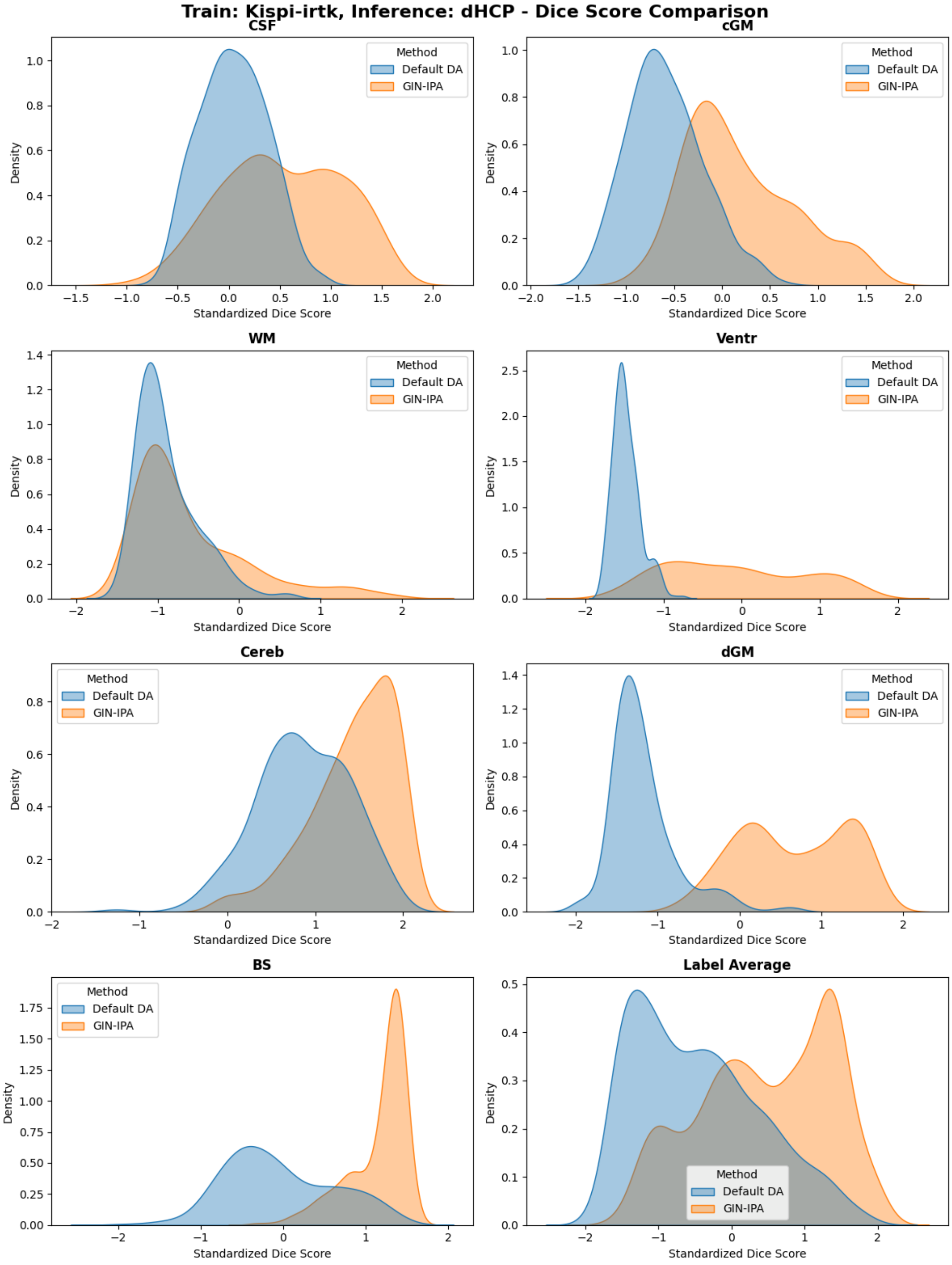

Figure 7.6: Baseline vs. GIN-IPA: KDE plots of DSC across each label and globally, from models trained on Kispi-irtk and inferring on dHCP.

Metric	Label	Mean perf. variation	Cohen's $\lvert d\rvert$
DSC ($\times 10^{-2}$)	CSF	$45 \rightarrow 56$	1.0
	cGM	$31 \rightarrow 48$	1.6
	WM	$24 \rightarrow 30$	0.5
	Ventr.	$11 \rightarrow 42$	2.2
	Cereb.	$63 \rightarrow 76$	1.1
	dGM	$18 \rightarrow 59$	3.4
	BS	$43 \rightarrow 70$	2.3
	Total	$34 \rightarrow 54$	1.1
VS	CSF	$1.0 \rightarrow 0.7$	1.5
	cGM	$1.1 \rightarrow 0.7$	1.6
	WM	$1.4 \rightarrow 1.3$	0.4
	Ventr.	$1.8 \rightarrow 1.1$	2.2
	Cereb.	$0.3 \rightarrow 0.1$	0.5
	dGM	$1.4 \rightarrow 0.6$	2.0
	BS	$0.8 \rightarrow 0.4$	1.4
	Total	$1.1 \rightarrow 0.7$	0.8
HD95	CSF	$35 \rightarrow 32$	0.3
	cGM	$40 \rightarrow 37$	0.3
	WM*	$42 \rightarrow 42$	0.1
	Ventr.	$48 \rightarrow 44$	0.3
	Cereb.	$55 \rightarrow 44$	0.6
	dGM	$57 \rightarrow 37$	1.1
	BS	$64 \rightarrow 30$	1.4
	Total	$49 \rightarrow 38$	0.6

Table 7.1: Baseline vs. GIN-IPA: mean performance variation and Cohen's $|d|$ across metrics and labels, from models trained on Kispi-irtk and inferring on dHCP. To enhance comprehensibility, the absolute value of VS is shown. The variation of HD95 in WM—marked by the asterisk—is the only one that is not statistically significant (p-value < 0.01).

GIN-IPA vs. Combined Augmentation

- **Train on Kispi-mial**
  - **Inference on Kispi-mial:** no difference in performance.
  - **Inference on Kispi-irtk:** no difference in performance.
  - **Inference on dHCP:** small improvements across DSC, VS, and HD95 for ventricles (DSC: from 0.18 to 0.23, $|d| = 0.3$) and dGM (DSC: from 0.35 to 0.40, $|d| = 0.3$).
- **Train on Kispi-irtk**
  - **Inference on Kispi-mial:** moderate improvements (see Fig. 7.7) in DSC and HD95 for cGM (DSC: from 0.59 to 0.65; HD95: from 2.2 to 1.7) and cerebellum (DSC: from 0.73 to 0.78; HD95: from 4.8 to 2.3). Small improvement overall.
  - **Inference on Kispi-irtk:** no difference in performance.
  - **Inference on dHCP:** combined augmentation performs significantly worse than GIN-IPA alone, especially for CSF, cerebellum, and dGM across all the metrics.
- **Train on dHCP** No differences observed on any inference dataset, metric, or label.

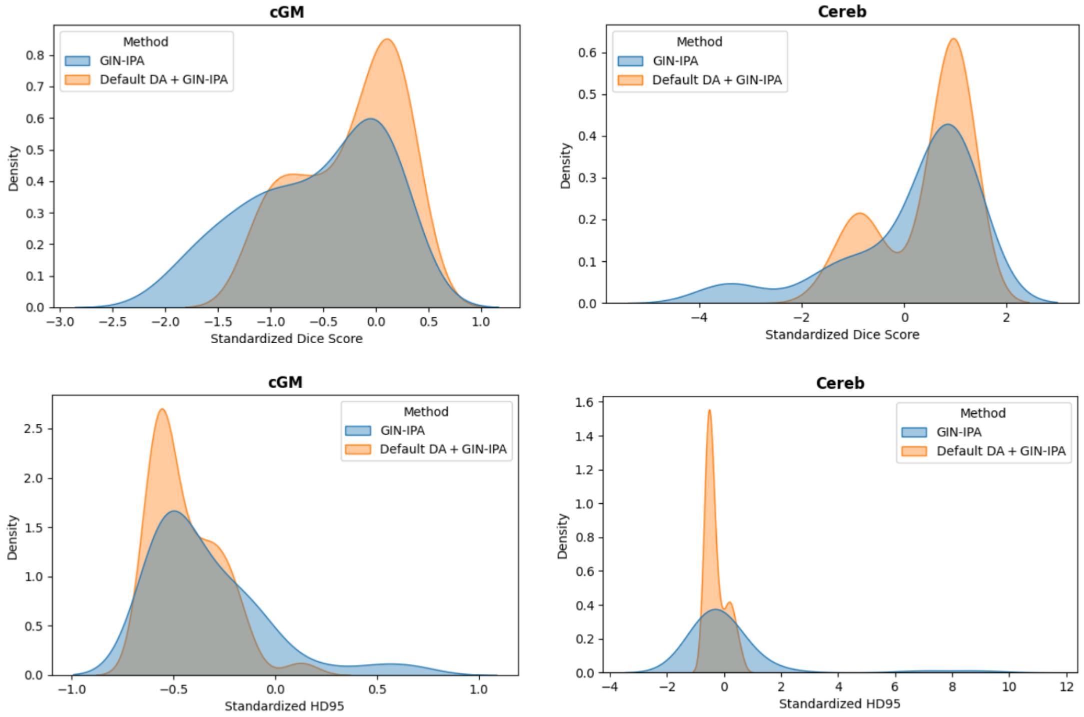

Figure 7.7: GIN-IPA vs. combined DA: KDE plots of DSC (top) and HD95 (bottom) in cortical gray matter and cerebellum, from models trained on Kispi-irtk and inferring on Kispi-mial.

Train DS	**Inference DS (n. of samples)**	**Metric**	**Mean perf. variation**	p**-value** $(\times 10^{-1})$	**Cohen's** $\lvert d \rvert$
Kispi-mial	Kispi-mial ($n = 8$)	DSC ($\times10^{-2}$)	$83 \rightarrow 83$	—	—
		VS ($\times10^{-2}$)	$3.6 \rightarrow 3.6$	—	—
		HD95	$1.6 \rightarrow 1.6$	—	—
	Kispi-irtk ($n = 40$)	DSC ($\times10^{-2}$)	$75 \rightarrow 76$	9.4	< 0.1
		VS ($\times10^{-2}$)	$0.5 \rightarrow 2.7$	—	—
		HD95	$3.4 \rightarrow 2.6$	5.3	0.2
	dHCP ($n = 267$)	DSC ($\times10^{-2}$)	$40 \rightarrow 44$	$\ll 0.1^{\ddagger}$	0.2
		VS ($\times10^{-2}$)	$9.7 \rightarrow 8.7$	$\ll 0.1^{\ddagger}$	0.2
		HD95	$46 \rightarrow 45$	2.5	< 0.1
Kispi-irtk	Kispi-mial ($n = 40$)	DSC ($\times10^{-2}$)	$63 \rightarrow 68$	$\ll 0.1^{\ddagger}$	0.2
		VS ($\times10^{-2}$)	$5.1 \rightarrow 3.1$	$\ll 0.1^{\ddagger}$	0.1
		HD95	$5.0 \rightarrow 4.4$	1.6	0.1
	Kispi-irtk ($n = 8$)	DSC ($\times10^{-2}$)	$89 \rightarrow 89$	—	—
		VS ($\times10^{-2}$)	$0.3 \rightarrow 0.1$	6.9	< 0.1
		HD95	$1.0 \rightarrow 1.0$	—	—
	dHCP ($n = 267$)	DSC ($\times10^{-2}$)	$34 \rightarrow 54$	$\ll 0.1^{\ddagger}$	1.1^{***}
		VS ($\times10^{-2}$)	$112 \rightarrow 71$	$\ll 0.1^{\ddagger}$	0.8^{***}
		HD95	$49 \rightarrow 38$	$\ll 0.1^{\ddagger}$	0.6^{**}
dHCP	Kispi-mial ($n = 40$)	DSC ($\times10^{-2}$)	$67 \rightarrow 69$	$\ll 0.1^{\ddagger}$	< 0.1
		VS ($\times10^{-2}$)	$13 \rightarrow 6$	7.2	0.1
		HD95	$3.6 \rightarrow 3.2$	$0.01^{\ddagger}$	< 0.1
	Kispi-irtk ($n = 40$)	DSC ($\times10^{-2}$)	$76 \rightarrow 77$	$\ll 0.1^{\ddagger}$	0.1
		VS ($\times10^{-2}$)	$7.7 \rightarrow 6.7$	$\ll 0.1^{\ddagger}$	< 0.1
		HD95	$2.7 \rightarrow 2.5$	$< 0.1^{\ddagger}$	< 0.1
	dHCP ($n = 53$)	DSC ($\times10^{-2}$)	$95 \rightarrow 95$	—	—
		VS ($\times10^{-2}$)	$0.2 \rightarrow 0.1$	$0.2^{\dagger}$	< 0.1
		HD95	$7.1 \rightarrow 7.0$	$0.05^{\ddagger}$	< 0.1

Table 7.2: Baseline vs. GIN-IPA: mean performance variation, Wilcoxon p-value and Cohen's $\lvert d \rvert$ across training/inference datasets and metrics. To improve comprehensibility, the absolute value of VS is shown. Where the p-value and the Cohen's $\lvert d \rvert$ are missing, it is because the performance variation is either negative or negligible. †: p-value < 0.05; ‡: p-value < 0.01. *: $\lvert d \rvert > 0.2$; **: $\lvert d \rvert > 0.5$; ***: $\lvert d \rvert > 0.8$.

Train DS	**Inference DS (n. of samples)**	**Metric**	**Mean perf. variation**	**p-value ($\times 10^{-1}$)**	**Cohen's $\lvert d \rvert$**
Kispi-mial	Kispi-mial ($n = 8$)	DSC ($\times 10^{-1}$)	$8.3 \to 8.3$	1.0	< 0.1
		VS ($\times 10^{-2}$)	$3.6 \to 4.1$	9.4	< 0.1
		HD95	$1.6 \to 1.6$	—	—
	Kispi-irtk ($n = 40$)	DSC ($\times 10^{-1}$)	$7.6 \to 7.6$	—	—
		VS ($\times 10^{-2}$)	$2.7 \to 2.6$	9.8	< 0.1
		HD95	$2.6 \to 2.7$	—	—
	dHCP ($n = 267$)	DSC ($\times 10^{-1}$)	$4.4 \to 4.4$	—	—
		VS ($\times 10^{-2}$)	$87 \to 87$	—	—
		HD95	$45 \to 44$	$0.092^{\ddagger}$	< 0.1
Kispi-irtk	Kispi-mial ($n = 40$)	DSC ($\times 10^{-1}$)	$6.8 \to 7.2$	$\ll 0.1^{\ddagger}$	0.2
		VS ($\times 10^{-2}$)	$3.1 \to 6.9$	—	—
		HD95	$4.4 \to 3.3$	$\ll 0.1^{\ddagger}$	0.2^{*}
	Kispi-irtk ($n = 8$)	DSC ($\times 10^{-1}$)	$8.9 \to 8.9$	—	—
		VS ($\times 10^{-2}$)	$0.1 \to 0.1$	—	—
		HD95	$1.0 \to 1.0$	—	—
	dHCP ($n = 267$)	DSC ($\times 10^{-1}$)	$5.4 \to 5.1$	—	—
		VS ($\times 10^{-2}$)	$71 \to 80$	—	—
		HD95	$38 \to 42$	—	—
dHCP	Kispi-mial ($n = 40$)	DSC ($\times 10^{-1}$)	$6.9 \to 6.8$	—	—
		VS ($\times 10^{-2}$)	$5.8 \to 7.1$	—	—
		HD95	$3.2 \to 3.3$	7.2	< 0.1
	Kispi-irtk ($n = 40$)	DSC ($\times 10^{-1}$)	$7.7 \to 7.7$	—	—
		VS ($\times 10^{-2}$)	$6.7 \to 5.7$	$\ll 0.1^{\ddagger}$	< 0.1
		HD95	$2.4 \to 2.4$	—	—
	dHCP ($n = 53$)	DSC ($\times 10^{-1}$)	$9.4 \to 9.4$	—	—
		VS ($\times 10^{-2}$)	$0.1 \to 0.1$	—	—
		HD95	$0.7 \to 0.7$	—	—

Table 7.3: GIN-IPA vs. combined aug.: mean performance variation, Wilcoxon p-value and Cohen's $\lvert d \rvert$ across training/inference datasets and metrics. To improve comprehensibility, the absolute value of VS is shown. Where the p-value and the Cohen's $\lvert d \rvert$ are missing, it is because the performance variation is either negative or negligible. †: p-value < 0.05; ‡: p-value < 0.01. *: $\lvert d \rvert > 0.2$; **: $\lvert d \rvert > 0.5$; ***: $\lvert d \rvert > 0.8$.

7.3 PERFORMANCE BY PATHOLOGY

Since dHCP only includes healthy scans, it can be interesting to investigate how the data augmentation methods behave when stratification on subject health is taken into account. Hence, the performance of the models was analyzed separately for healthy and pathological cases, in order to isolate the effect of brain abnormalities on segmentation accuracy.

The plots below (Fig. 7.8 and Fig. 7.9) are relative to models trained on dHCP and separately tested on healthy and pathological cases in Kispi-mial and Kispi-irtk. Metrics relative to the test set of dHCP are shown as a reference. The difference between neurotypical and pathological cases is evident but not dramatic for all metrics—volume similarity shows a delta which is especially small. Neither GIN-IPA nor the combined DA help at increasing any metric in neurotypical subjects. For the pathological cases, GIN-IPA produces a small improvement in all metrics with respect to the baseline model, while performs equivalent to the combined DA model. It is also interesting to notice that, despite the difference in image quality claimed in [40], the performance of the model trained on dHCP is equivalent when inferring on healthy Kispi-mial and Kispi-irtk. On the other hand, a performance difference is observed when inferring on pathological Kispi-mial and Kispi-irtk cases. This suggests that the low-quality images are mainly among the pathological cases.

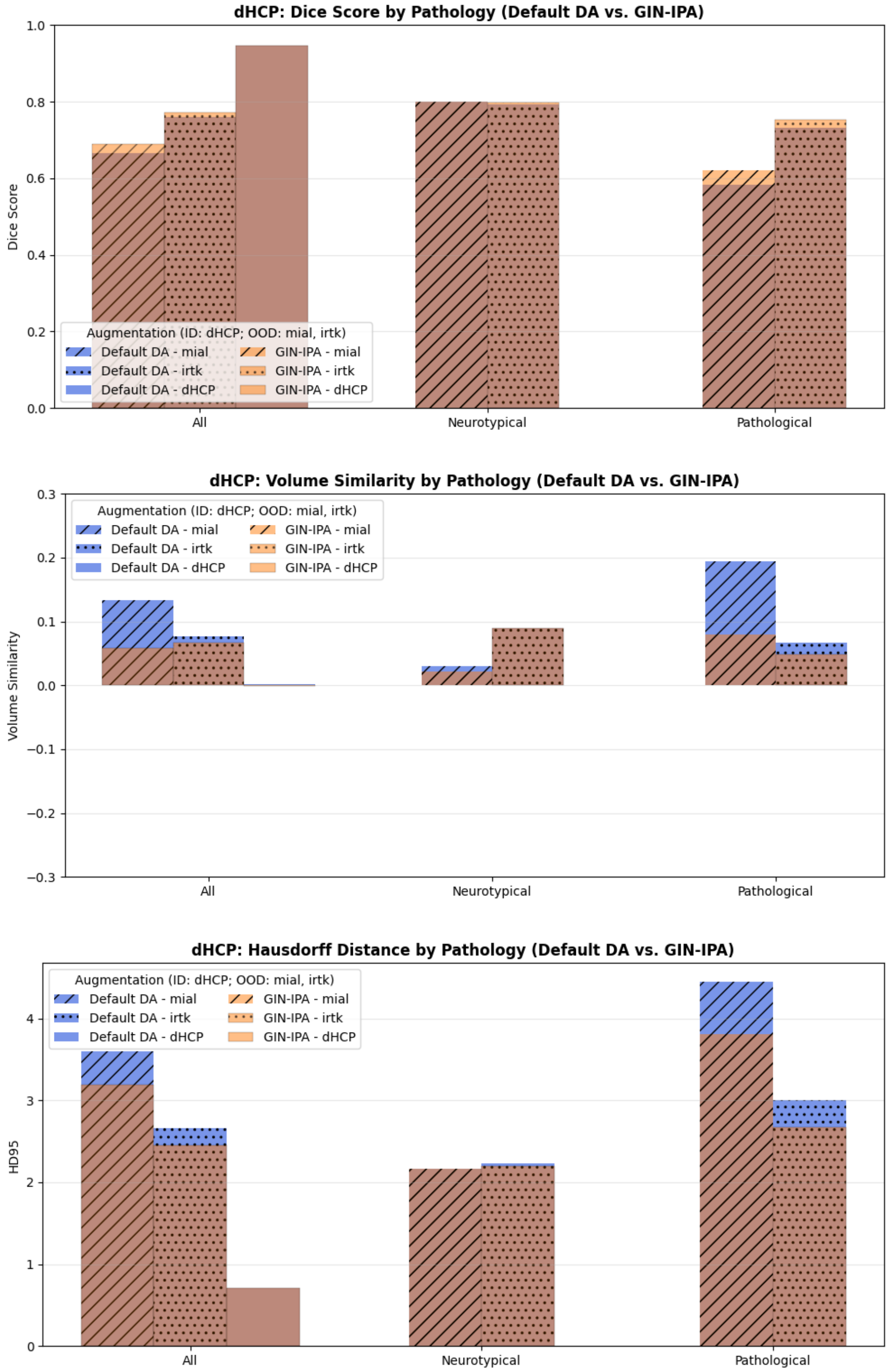

Figure 7.8: Comparison of the segmentation performance by pathology, between the baseline and the GIN-IPA augmentation models. From top to bottom: Dice score, volume similarity (cropped, full range is $[-2, 2]$), and Hausdorff distance 95th percentile. Until the bars are brown, the metrics of the two models are equivalent.

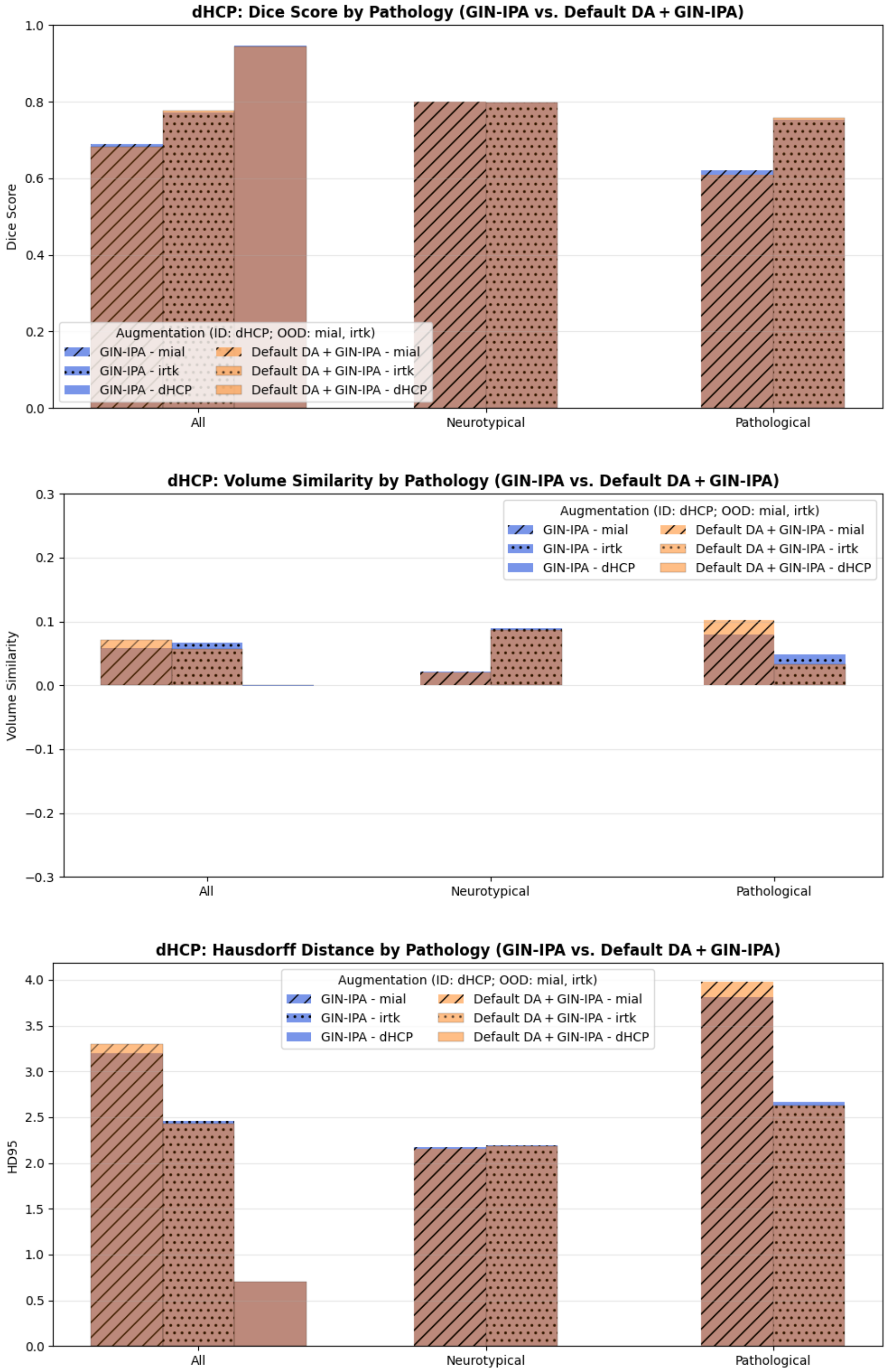

Figure 7.9: Comparison of the segmentation performance by pathology, between the GIN-IPA and the combined augmentation models. From top to bottom: Dice score, volume similarity (cropped, full range is $[-2, 2]$), and Hausdorff distance 95$^{\text{th}}$ percentile. Until the bars are brown, the metrics of the two models are equivalent.

8 DISCUSSION

This chapter discusses the empirical evidence reported in Chapter 7, drawing a unified interpretation across general performance trends across datasets, pairwise model comparisons, and pathology-stratified analyses. The emphasis is on the behavior of the augmentation strategies in realistic in-domain and out-of-domain settings, and on the conditions under which gains are observed.

8.1 PRIMACY OF DATA OVER AUGMENTATION

The first, and most consistent, observation is the primacy of the training dataset over the augmentation recipe. Models trained on dHCP achieve the most stable performance both ID and OOD, and do so *independently* of whether augmentation is the nnU-Net default, GIN-IPA, or their combination. In Section 7.1, this is visible in the DSC histograms (Figs. 7.1–7.3): while Kispi-trained models experience a marked OOD drop—particularly towards dHCP, with ventricles, dGM and WM most affected—the dHCP-trained models exhibit limited degradation across inference domains. Moreover, no DSC gain emerges from switching augmentation when training on dHCP. Together, these findings imply that *dataset quality and scale* dominate generalization in this setting.

A further confirmation of the importance of the data quality emerges from Kispi-mial: the change of domain does not impact the network trained on Kispi-mial in the same manner as the other two sources. Concretely, the performance turns out to be higher OOD than ID for two tissues, and only when inferring on Kispi-irtk. Even though the mechanism is not clear, this is probably attributable to two factors:

- It may be supposed that the lower quality of Kispi-mial [40] prevents the network from learning robust features. This makes it perform better on a dataset—i.e., Kispi-irtk—that is similar to the source, but has a higher image quality, which inherently drives to better segmentations.
- On the other hand, the small size of the ID test set of Kispi-mial—only 8 samples—limits the reliability of the results, which may be strongly influenced by the specific cases included. In other words, the lower ID performance may be due to a few challenging cases in the test set, which are not representative of the whole dataset.

However, the global DSC values still reflect the trends observed in the other two training cases, with ID performance being higher than OOD.

8.2 Efficacy of GIN-IPA

Baseline vs. GIN-IPA. When training on Kispi-irtk and inferring on dHCP, GIN-IPA yields a *substantial improvement* over the baseline across metrics and labels. At the global level, the average DSC increases ($58\,\%$), with concordant improvements in VS ($-36\,\%$) and HD95 ($-22\,\%$). The corresponding paired analysis confirms the statistical significance and a large effect size for DSC in this cross-domain transfer (DSC: from 0.34 to 0.54, $p \ll 0.01$, $|d| = 1.1$; VS: from 1.12 to 0.71, $p \ll 0.01$, $|d| = 0.8$; HD95: from 49 to 38, $p \ll 0.01$, $|d| = 0.6$). The increase is particularly pronounced for dGM (DSC: from 0.18 to 0.59), ventricles (DSC: from 0.11 to 0.42), and BS (DSC: from 0.43 to 0.70). The full set of paired comparisons for all metrics and labels (including VS and HD95) is reported in Section 7.2 (Tabs. 7.2–7.3).

Conversely, training on dHCP shows no material benefit from GIN-IPA over the baseline across any inference domain or label. For Kispi-mial, effects are small and structure-dependent, with modest improvements predominantly in cases where the baseline struggles on poorly segmented volumes.

GIN-IPA vs. Combined Augmentation. Stacking the two augmentation methods does not systematically improve performance over GIN-IPA alone, but rather can make it worse. Actually, when trained on Kispi-irtk and inferred on dHCP, the combined strategy is consistently inferior to pure GIN-IPA for several structures (CSF, cerebellum, dGM) across all metrics. Training on dHCP again yields no differences between the two.

8.3 Robustness by Pathology

The stratified analysis (Section 7.3) indicates that models trained on dHCP generalize equally well to healthy subjects in Kispi-mial and Kispi-irtk, despite the domain shift (different image quality and reconstruction techniques). Performance decreases for pathological cases, reflecting segmentation challenges in the presence of anatomical abnormalities. In this more difficult regime, GIN-IPA achieves very small improvements over the baseline, while remaining broadly equivalent to the combined strategy. Furthermore, these results are explained with the concentration of the lower-quality images in Kispi-mial among the pathological scans; this is the reason for the larger performance gap in that subgroup, once again confirming the relevance of data quality in this segmentation framework.

8.4 General Outcomes

The evidence above supports three general conclusions:

- **Data quality and scale prevail:** When training data are abundant and homogeneous (like in dHCP), the choice of an augmentation pipeline among those tested barely alters the performance, because it is already high for every model—thanks to the inherent robustness of nnU-Net. The converse is also true: with smaller, noisier sources (such as Kispi), OOD degradation is pronounced.

- **GIN-IPA is conditionally beneficial:** Its gains are largest in the domain generalization scenario from Kispi-irtk to dHCP, where it closes a considerable part of the OOD gap. However, the effects shrink or vanish when the source dataset is large enough to cover the target variability, and of high quality.
- **Stacking augmentations is not additive:** The combined strategy often overlaps with, and can even dilute, the benefits of GIN-IPA; it never consistently outperforms GIN-IPA in the examined cross-domain settings. A plausible explanation is that mixing heterogeneous perturbations introduces redundant transformations, which may push samples away from the realistic fetal MRI appearance and weaken the shape-focused invariances that GIN-IPA is designed to ensure.

These patterns are coherent with the intended role of GIN-IPA: by synthesizing intensity and spatial variations, it exposes the network to harder, more diverse views of a limited source, which is most valuable when the source lacks the target domain variability. Once data already cover the relevant distributional modes—as with dHCP—marginal augmentation gains become negligible.

8.5 LIMITATIONS AND FUTURE WORK

Three principal aspects delimit the scope of the present findings. First, the small size of the Kispi cohorts constrained further stratification (e.g., finer gestational age, training by pathology) and limited the precision of subgroup estimates. Generally speaking, the scarcity of public datasets, together with their high intra-variability, jeopardizes the possibility to totally isolate single domains. Second, label-set harmonization between Kispi and dHCP required a mapping (Tab. B.1), which, although carefully defined, introduces an additional layer of variability in label-wise comparisons. Third, because of hardware limitations, the IPA step was performed in its 2D variant, which may be less effective than the full 3D version. Operating in a slice-wise mode prevents the augmentation from modelling cross-plane distortions. Therefore, distortions remain locally consistent within each slice but not necessarily across the reconstructed volume. This mismatch between the 2D augmentation process and the 3D segmentation strategy may limit the ability to mock realistic inter-slice artefacts.

Practically, the results argue for prioritizing *data curation*—larger, multi-center, high-quality fetal MRI with standardized SRR pipelines—since source quality and scale are the dominant predictors of cross-domain success in this task. Within constrained-data regimes, GIN-IPA is a *useful augmentation choice*, particularly for single-source DG from moderately sized, relatively clean sources (e.g., Kispi-irtk). Nonetheless, stacking it with standard nnU-Net augmentation is unnecessary and sometimes counterproductive. For challenging pathological cases, improved acquisition and reconstruction remain central to performance gains.

Besides data, it could be interesting to disentangle the two components of GIN-IPA—possibly employing the 3D IPA variant and broader datasets that allow a more rigorous cross-domain validation—to assess their individual contributions, and to explore other augmentation strategies (e.g., adversarial, style-transfer) in this setting.

CONCLUSIONS

This thesis examined whether a 3D U-Net can sustain cross-domain performance in fetal brain MRI segmentation through causality-inspired data augmentation (GIN-IPA), and whether stacking it with the standard nnU-Net augmentation yields additional gains. The most important result is that the quality and coverage of the training data dominate domain generalization. Models trained on a large, clean, and internally consistent source generalize more stably than models trained on smaller or noisier sources, largely independent of the augmentation recipe. Augmentation matters most when the source domain is limited and distributionally distant from the target.

The central evidence for the validation of GIN-IPA comes from the cross-domain transfer where models trained on Kispi-irtk are evaluated on dHCP. In this setting, GIN-IPA *substantially improves* accuracy and robustness over the baseline, with effects that are both practically relevant and statistically significant across tissues and metrics—especially deep gray matter, ventricles and brainstem. These trends are confirmed by Wilcoxon signed-rank tests with p-values smaller than 0.01. When trained on dHCP, by contrast, augmentation choices have *marginal impact*, underscoring that data fidelity and reconstruction consistency are first-order factors for generalization.

A second conclusion is that stacking heterogeneous transformations is not beneficial. The combined augmentation is not superior to GIN-IPA alone and, in the aforementioned most informative transfer, it is consistently inferior for several structures and metrics. This non-additivity suggests interference between transformation families that may dilute the invariances that GIN-IPA aims to enforce. As expected, more transformations do not automatically translate into better out-of-domain behavior.

As a side result, the analysis stratified by pathology indicates that global domain shifts remain the predominant driver of performance variability. Differences between neurotypical and pathological cases were present but moderate, in part due to the different scan quality of the subcohorts. Gains from GIN-IPA are only noticeable among pathological scans, albeit with small margins, which is consistent with the expectation that morphology variability is only partially addressable by appearance-based perturbations.

The broader implication is methodological and organizational. For prenatal neuroimaging pipelines that must transfer across sites and reconstruction stacks, investment should prioritize data curation, SRR standardization, and label governance before augmentation engineering. Where multi-center high-quality data are unavailable, GIN-IPA offers an effective, lightweight mechanism to mitigate acquisition-driven shifts. Conversely, when ample high-quality data are available, the return on complex augmentation schedules is limited.

In summary, domain generalization in fetal brain MRI is primarily determined by the data pathway, with GIN-IPA yielding substantial benefits exactly when the source domain is narrow or mismatched to the target—although its effectiveness would need further validation. These findings translate into clear guidance for building robust, transferable segmentation systems: prioritize data, deploy targeted augmentation where it closes the domain gap, and evaluate decisions on out-of-domain tests. Future work should disentangle the contributions of the two components within GIN-IPA, and validate it on larger, more diverse datasets.

ACRONYMS

BS	Brainstem
bSSFP	Balanced steady-state free precession
cGM	Cortical gray matter
CHUV	Centre Hospitalier Universitaire Vaudois (Lausanne University Hospital)
CNN	Convolutional neural network
CNR	Contrast-to-noise ratio
CSF	Cerebrospinal fluid
DA	Data augmentation
DG	Domain generalization
dGM	Deep gray matter
DS	Dataset
DSC	Dice similarity coefficient
FeTA	Fetal Tissue Annotation Challenge
FOV	Field of view
GA	Gestational age
GM	Gray matter
GW	Gestational week
HD95	Hausdorff 95 distance
IVIM	Intravoxel incoherent motion
KCL	King's College London (St. Thomas Hospital)
KDE	Kernel Density Estimation
Kispi	Universitäts-Kinderspital Zürich (Zurich University Children's Hospital)
MRI	Magnetic Resonance Imaging
NMR	Nuclear Magnetic Resonance
OOD	Out-of-domain
RF	Radiofrequency
SAR	Specific absorption rate
SNR	Signal-to-noise ratio
SR	Super-resolution
SRR	Super-resolution reconstruction
SSFSE	Single-shot fast spin-echo
TE	Echo time
TR	Repetition time

UCSF	University of California, San Francisco (Benioff Children's Hospital)
VS	Volume similarity
WM	White matter

Bibliography

1. *Nuclear magnetic resonance.* URL: `https://en.wikipedia.org/wiki/Nuclear_magnetic_resonance` (visited on 10/06/2025) (cit. on p. 5).
2. *Curie's law.* URL: `https://en.wikipedia.org/wiki/Curie%27s_law` (visited on 10/06/2025) (cit. on p. 6).
3. *NMR Spectroscopy: Principle, Instrumentation, Applications, Limitation.* URL: `https://scienceinfo.com/nmr-spectroscopy` (visited on 10/11/2025) (cit. on p. 7).
4. C. P. Lanting, E. De Kleine, H. Bartels, and P. Van Dijk. "Functional imaging of unilateral tinnitus using fMRI". *Acta Oto-Laryngologica* 128:4, 2008, pp. 415–421. DOI: `10.1080/00016480701793743` (cit. on p. 8).
5. *Nuclear Magnetic Resonance Theory 3.* URL: `https://www.nmr.hhu.de/main/mtheorie_3_e.html` (visited on 10/12/2025) (cit. on p. 9).
6. *MRI Questions: Image Contrast.* URL: `https://mriquestions.com/image-contrast-trte.html` (visited on 10/12/2025) (cit. on pp. 9, 11).
7. M. P. Lun, E. S. Monuki, and M. K. Lehtinen. "Development and functions of the choroid plexus-cerebrospinal fluid system". *Nature Reviews Neuroscience* 16:8, 2015, pp. 445–457. DOI: `10.1038/nrn3921` (cit. on p. 14).
8. L. Vasung, E. A. Turk, S. L. Ferradal, et al. "Exploring early human brain development with structural and physiological neuroimaging". *Neuroimage* 187, 2019, pp. 226–254. DOI: `10.1016/j.neuroimage.2018.07.041` (cit. on p. 14).
9. S. Wilson, M. Pietsch, L. Cordero-Grande, et al. "Development of human white matter pathways in utero over the second and third trimester". *Proceedings of the National Academy of Sciences* 118:20, 2021. DOI: `10.1073/pnas.2023598118` (cit. on p. 14).
10. I. Kostović, G. Sedmak, and M. Judaš. "Neural histology and neurogenesis of the human fetal and infant brain". *NeuroImage* 188, 2019, pp. 743–773. DOI: `10.1016/j.neuroimage.2018.12.043` (cit. on p. 14).
11. J. A. Scott, K. S. Hamzelou, V. Rajagopalan, et al. "3D Morphometric Analysis of Human Fetal Cerebellar Development". *The Cerebellum* 11, 2012, pp. 761–770. DOI: `10.1007/s12311-011-0338-2` (cit. on p. 14).
12. G. O. Dovjak, V. Schmidbauer, P. C. Brugger, et al. "Normal human brainstem development in vivo: a quantitative fetal MRI study". *Ultrasound in Obstetrics & Gynecology* 58:2, 2021, pp. 254–263. DOI: `10.1002/uog.22162` (cit. on p. 14).

13. L. Manganaro, S. Capuani, M. Gennarini, et al. “Fetal MRI: what’s new? A short review”. *European Radiology Experimental* 7:41, 2023. DOI: `10.1186/s41747-023-00358-5` (cit. on pp. 15–16).
14. T. Victoria, A. M. Johnson, J. C. Edgar, et al. “Comparison between 1.5-T and 3-T MRI for fetal imaging”. *American Journal of Roentgenology* 206, 2016, pp. 195–201. DOI: `10.2214/AJR.14.14205` (cit. on p. 15).
15. A. L. Chartier, M. J. Bouvier, D. R. McPherson, et al. “The safety of maternal and fetal MRI at 3 T”. *American Journal of Roentgenology* 213:5, 2019, pp. 1170–1173. DOI: `10.2214/AJR.19.21400` (cit. on p. 15).
16. C. Jaimes, J. Delgado, M. B. Cunnane, et al. “Does 3-T fetal MRI induce adverse acoustic effects in the neonate?” *Paediatric Radiology* 49, 2019, pp. 37–45. DOI: `10.1007/s00247-018-4261-2` (cit. on p. 15).
17. G. C. Colleran, M. Kyncl, C. Garel, and M. Cassart. “Fetal magnetic resonance imaging at 3 Tesla – the European experience”. *Pediatric Radiology* 52, 2022, pp. 959–970. DOI: `10.1007/s00247-021-05267-6` (cit. on p. 15).
18. S. Ponrartana, H. N. Nguyen, S. X. Cui, et al. “Low-field 0.55 T MRI evaluation of the fetus”. *Pediatric Radiology* 53, 2023, pp. 1469–1475. DOI: `10.1007/s00247-023-05604-x` (cit. on pp. 15–16, 28).
19. S. Neves Silva, S. McElroy, J. Aviles Verdera, et al. “Fully automated planning for anatomical fetal brain MRI on 0.55T”. *Magnetic Resonance in Medicine* 92:3, 2024, pp. 1263–1276. DOI: `10.1002/mrm.30122` (cit. on p. 16).
20. J. Aviles Verdera, L. Story, M. Hall, et al. “Reliability and Feasibility of Low-Field-Strength Fetal MRI at 0.55 T during Pregnancy”. *Radiology* 309:1, 2023. DOI: `10.1148/radiol.223050` (cit. on pp. 16, 28).
21. B. J. Soher, B. L. Dale, and E. M. Merkle. “A Review of MR Physics: 3T versus 1.5T”. *Magnetic Resonance Imaging Clinics of North America* 15:3, 2007, pp. 277–290. DOI: `10.1016/j.mric.2007.06.002` (cit. on p. 16).
22. G. Ercolani, S. Capuani, A. Antonelli, et al. “Intravoxel incoherent motion (IVIM) MRI of fetal lung and kidney: can the perfusion fraction be a marker of normal maturation?” *European Journal of Radiology* 139, 2021, p. 109726. DOI: `10.1016/j.ejrad.2021.109726` (cit. on p. 16).
23. A. Antonelli, S. Capuani, G. Ercolani, et al. “Human placental microperfusion and microstructural assessment by IVIM MRI for discriminating intrauterine growth restriction: a pilot study”. *Journal of Maternal-Fetal and Neonatal Medicine* 35:25, 2022, pp. 9667–9674. DOI: `10.1080/14767058.2022.2050365` (cit. on p. 16).
24. A. U. Uus, V. Kyriakopoulou, A. Makropoulos, et al. “BOUNTI: Brain vOlumetry and aUtomated parcellatioN for 3D feTal MRI”, 2023. DOI: `10.7554/elife.88818.1` (cit. on pp. 16–17).

25. K. Payette, P. de Dumast, H. Kebiri, et al. “An automatic multi-tissue human fetal brain segmentation benchmark using the Fetal Tissue Annotation Dataset”. *Scientific Data* 8:167, 2021. DOI: `10.1038/s41597-021-00946-3` (cit. on pp. 17, 25, 27, 33, 36, 86).
26. M. Bach Cuadra, K. Payette, A. Jakab, et al. *Fetal Tissue Annotation Challenge: Structured description of the challenge design.* 2024. DOI: `10.5281/zenodo.10986046` (cit. on pp. 17–19, 27–28, 33–34).
27. A. Gholipour, J. A. Estroff, and S. K. Warfield. “Robust super-resolution volume reconstruction from slice acquisitions: application to fetal brain MRI”. *IEEE Transactions on Medical Imaging* 29:10, 2010, pp. 1739–1758. DOI: `10.1109/TMI.2010.2051680` (cit. on p. 17).
28. M. Kuklisova-Murgasova, G. Quaghebeur, M. A. Rutherford, et al. “Reconstruction of fetal brain MRI with intensity matching and complete outlier removal”. *Medical Image Analysis* 16, 2012, pp. 1550–1564. DOI: `10.1016/j.media.2012.07.004` (cit. on pp. 17, 20, 27, 33–34).
29. T. Ciceri, L. Squarcina, A. Pigoni, et al. “Geometric reliability of super-resolution reconstructed images from clinical fetal MRI in the second trimester”. *Neuroinformatics* 21:3, 2023, pp. 549–563. DOI: `10.1007/s12021-023-09635-5` (cit. on p. 17).
30. M. Ebner, G. Wang, W. Li, et al. “An automated framework for localization, segmentation and super-resolution reconstruction of fetal brain MRI”. *NeuroImage* 206, 2020. DOI: `10.1016/j.neuroimage.2019.116324` (cit. on pp. 17–18, 20).
31. S. Tourbier, X. Bressonc, P. Hagmann, et al. “An efficient total variation algorithm for super-resolution in fetal brain MRI with adaptive regularization”. *NeuroImage* 118, 2015, pp. 584–597. DOI: `10.1016/j.neuroimage.2015.06.018` (cit. on pp. 17–18, 20, 27, 33–34).
32. M. Deprez. *SIMPLE IRTK*. 2019. URL: `https://gitlab.com/mariadeprez/irtk-simple` (cit. on pp. 17, 20, 27, 34).
33. P. Deman, S. Tourbier, R. Meuli, and M. Bach Cuadra. *meribach/mevislabFetalMRI: MEVISLAB MIAL Super-Resolution Reconstruction of Fetal Brain MRI.* Version 1.0. 2020. DOI: `10.5281/zenodo.3878564` (cit. on pp. 18, 20, 27, 34).
34. D. Sobotka, M. Ebner, E. Schwartz, et al. “Motion correction and volumetric reconstruction for fetal functional magnetic resonance imaging data”. *NeuroImage* 255, 2022. DOI: `10.1016/j.neuroimage.2022.119213` (cit. on p. 18).
35. T. Sánchez, A. Mihailov, G. Martí Juan, et al. “Assessing Data Quality on Fetal Brain MRI Reconstruction: A Multi-site and Multi-rater Study”. In: *Perinatal, Preterm and Paediatric Image Analysis (PIPPI 2024).* Vol. 14747. Lecture Notes in Computer Science. Springer, 2025, pp. 46–56. DOI: `10.1007/978-3-031-73260-7_5` (cit. on pp. 18, 29).

36. T. Sánchez, A. Mihailov, M. Koob, et al. “Biometry and volumetry in multi-centric fetal brain MRI: assessing the bias of super-resolution reconstruction”, 2024. DOI: `10.1101/2024.09.23.24313965` (cit. on p. 18).
37. V. Zalevskyi, T. Sánchez, M. Kaandorp, et al. *Advances in Automated Fetal Brain MRI Segmentation and Biometry: Insights from the FeTA 2024 Challenge.* 2025. arXiv: `2505.02784 [cs.CV]`. URL: `https://arxiv.org/abs/2505.02784` (cit. on pp. 18–19, 28, 33–35, 43).
38. T. Ciceri, L. Squarcina, A. Giubergia, et al. “Review on deep learning fetal brain segmentation from Magnetic Resonance images”. *Artificial Intelligence in Medicine* 143, 2023. DOI: `10.1016/j.artmed.2023.102608` (cit. on pp. 18–19).
39. T. Ciceri, L. Casartelli, F. Montano, et al. “Fetal brain MRI atlases and datasets: A review”. *NeuroImage* 292, 2024. DOI: `10.1016/j.neuroimage.2024.120603` (cit. on pp. 18, 33).
40. K. Payette, B. L. Hongwei, P. de Dumast, et al. “Fetal brain tissue annotation and segmentation challenge results”. *Medical Image Analysis* 88, 2023. DOI: `10.1016/j.media.2023.102833` (cit. on pp. 19, 27, 34, 43, 55, 66, 69).
41. A. U. Uus, I. Grigorescu, M. P. van Poppel, et al. “Automated 3D reconstruction of the fetal thorax in the standard atlas space from motion-corrupted MRI stacks for 21-36 weeks GA range”. *Medical Image Analysis* 80, 2022. DOI: `10.1016/j.media.2022.102484` (cit. on p. 20).
42. C. Ouyang, C. Chen, S. Li, et al. “Causality-Inspired Single-Source Domain Generalization for Medical Image Segmentation”. *IEEE Transactions on Medical Imaging* 42:4, 2023, pp. 1095–1106. DOI: `10.1109/TMI.2022.3224067` (cit. on pp. 21, 23, 40–42, 48, 50).
43. J. Wang, C. Lan, C. Liu, et al. “Generalizing to Unseen Domains: A Survey on Domain Generalization”. *IEEE Transactions on Knowledge and Data Engineering* 35:8, 2023, pp. 8052–8072. DOI: `10.1109/TKDE.2022.3178128` (cit. on pp. 21, 23).
44. B. Billot, D. N. Greve, O. Puonti, et al. “SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining”. *Medical Image Analysis* 86, 2023. DOI: `10.1016/j.media.2023.102789` (cit. on pp. 22, 28).
45. H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. *mixup: Beyond Empirical Risk Minimization.* 2018. arXiv: `1710.09412 [cs.LG]`. URL: `https://arxiv.org/abs/1710.09412` (cit. on p. 22).
46. G. Blanchard, A. A. Deshmukh, Ü. Dogan, et al. “Domain generalization by marginal transfer learning”. *The Journal of Machine Learning Research* 22:1, 2021, pp. 46–100. URL: `https://dl.acm.org/doi/abs/10.5555/3546258.3546260` (cit. on p. 22).

47. Y. Ganin and V. Lempitsky. “Unsupervised Domain Adaptation by Backpropagation”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. PMLR, Lille, France, 2015, pp. 1180–1189. URL: `https://proceedings.mlr.press/v37/ganin15.html` (cit. on p. 22).
48. Y. Ganin, E. Ustinova, H. Ajakan, et al. “Domain-Adversarial Training of Neural Networks”. *Journal of Machine Learning Research* 17:59, 2016, pp. 1–35. DOI: `10.48550/arXiv.1409.7495` (cit. on p. 22).
49. D. Krüger, E. Caballero, J.-H. Jacobsen, et al. “Out-of-distribution generalization via risk extrapolation (REx)”. In: *Proceedings of the 13th International Conference on Machine Learning and Computing*. Association for Computing Machinery, Shenzhen, China, 2021, pp. 5815–5826. DOI: `10.1145/3457682` (cit. on p. 22).
50. O. Ronneberger, P. Fischer, and T. Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: `1505.04597 [cs.CV]`. URL: `https://arxiv.org/abs/1505.04597` (cit. on p. 24).
51. F. Isensee, P. F. Jaeger, S. A. Kohl, et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. *Nature Methods* 18:2, 2021, pp. 203–211. DOI: `10.1038/s41592-020-01008-z` (cit. on pp. 24, 28, 39–40, 48).
52. N. Khalili, N. Lessmann, E. Turk, et al. “Automatic brain tissue segmentation in fetal MRI using convolutional neural networks”. *Magnetic Resonance Imaging* 64, 2019, pp. 77–89. DOI: `10.1016/j.mri.2019.05.020` (cit. on p. 25).
53. A. E. Fetit, A. Alansary, L. Cordero-Grande, et al. “A deep learning approach to segmentation of the developing cortex in fetal brain MRI with minimal manual labeling”. In: *Proceedings of the Third Conference on Medical Imaging with Deep Learning*. Vol. 121. Proceedings of Machine Learning Research. 2020, pp. 241–261. URL: `https://proceedings.mlr.press/v121/fetit20a.html` (cit. on p. 25).
54. L. Li, M. Sinclair, A. Makropoulos, et al. “CAS-Net: Conditional Atlas Generation and Brain Segmentation for Fetal MRI”. In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis*. Ed. by C. H. Sudre, R. Licandro, C. Baumgartner, et al. Springer International Publishing, 2021, pp. 221–230. DOI: `10.1007/978-3-030-87735-4_21` (cit. on p. 25).
55. M. Mazher, A. Qayyum, D. Puig, and M. Abdel-Nasser. “Effective Approaches to Fetal Brain Segmentation in MRI and Gestational Age Estimation by Utilizing a Multiview Deep Inception Residual Network and Radiomics”. *Entropy* 24:12, 2022. DOI: `10.3390/e24121708` (cit. on p. 25).
56. L. Fidon, M. Aertsen, N. Mufti, et al. “Distributionally Robust Segmentation of Abnormal Fetal Brain 3D MRI”. In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis*. Ed. by C. H. Sudre, R. Licandro, C. Baumgartner, et al. Springer International Publishing, 2021, pp. 263–273. DOI: `10.1007/978-3-030-87735-4_25` (cit. on pp. 25–26).

57. C. Zhigao and Z. Xing-Ming. *Enhancing Generalized Fetal Brain MRI Segmentation using A Cascade Network with Depth-wise Separable Convolution and Attention Mechanism.* 2024. arXiv: `2405.15205 [eess.IV]`. URL: `https://arxiv.org/abs/2405.15205` (cit. on p. 26).
58. *FeTA 2024 Challenge.* URL: `https://fetachallenge.github.io/` (visited on 08/09/2025) (cit. on pp. 26, 28).
59. *MICCAI: Medical Image Computing and Computer-Assisted Intervention.* URL: `https://miccai.org/` (visited on 08/09/2025) (cit. on p. 26).
60. K. Payette, R. Kottke, and A. Jakab. "Efficient Multi-class Fetal Brain Segmentation in High Resolution MRI Reconstructions with Noisy Labels". In: *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis.* Ed. by Y. Hu, R. Licandro, J. A. Noble, et al. Springer International Publishing, 2020, pp. 295–304. DOI: `10.1007/978-3-030-60334-2_29` (cit. on p. 27).
61. *FeTA 2021 Challenge.* URL: `https://feta.grand-challenge.org/feta-2021/` (visited on 08/09/2025) (cit. on p. 27).
62. K. Payette, C. Steger, R. Licandro, et al. "Multi-Center Fetal Brain Tissue Annotation (FeTA) Challenge 2022 Results". *IEEE Transactions on Medical Imaging*, 2024. DOI: `10.1109/tmi.2024.3485554` (cit. on pp. 27–28, 43).
63. *FeTA 2022 Challenge.* URL: `https://feta.grand-challenge.org/feta-2022/` (visited on 08/09/2025) (cit. on p. 27).
64. Helmholtz Imaging Applied Computer Vision Lab. *nnU-Net.* Version 2.4.1. 2024. URL: `https://github.com/MIC-DKFZ/nnUNet` (cit. on pp. 28, 39–40).
65. M. Eisenmann, A. Reinke, V. Weru, et al. "Why is the winner the best?" In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2023, pp. 19955–19966. DOI: `10.1109/CVPR52729.2023.01911` (cit. on p. 28).
66. J. Dockès, G. Varoquaux, and J. B. Poline. "Preventing dataset shift from breaking machine-learning biomarkers". *GigaScience* 10:9, 2021, pp. 1–11. DOI: `10.1093/gigascience/giab055` (cit. on p. 28).
67. L. Cordero-Grande, E. J. Hughes, J. Hutter, et al. "Three-dimensional motion corrected sensitivity encoding reconstruction for multi-shot multi-slice MRI: Application to neonatal brain imaging". *Magnetic Resonance in Medicine* 79:3, pp. 1365–1376. DOI: `10.1002/mrm.26796` (cit. on pp. 33, 35).
68. K. Payette and A. Jakab. *Fetal Tissue Annotation Challenge - FeTA MICCAI 2021.* 2021. DOI: `10.7303/SYN25649159` (cit. on pp. 33, 36, 86).
69. *dHCP Fourth Data Release (May 2024).* URL: `https://biomedia.github.io/dHCP-release-notes/index.html` (visited on 08/13/2025) (cit. on pp. 35, 38).
70. V. R. Karolis, L. Cordero-Grande, A. N. Price, et al. "The developing Human Connectome Project fetal functional MRI release: Methods and data structures". *Imaging Neuroscience* 3, 2025. DOI: `10.1162/imag_a_00512` (cit. on p. 35).

71. A. Makropoulos, E. C. Robinson, A. Schuh, et al. "The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction". *NeuroImage* 173, 2018, pp. 88–112. DOI: `10.1016/j.neuroimage.2018.01.054` (cit. on p. 37).

72. A. Makropoulos, A. Schuh, and R. Wright. *dHCP Structural Pipeline*. URL: `https://github.com/BioMedIA/dhcp-structural-pipeline` (cit. on p. 37).

73. K. Zhou, Y. Yang, Y. Qiao, and T. Xiang. *Domain Generalization with MixStyle*. 2021. arXiv: `2104.02008 [cs.CV]`. URL: `https://arxiv.org/abs/2104.02008` (cit. on p. 43).

74. C. Chen, C. Qin, H. Qiu, et al. "Realistic Adversarial Data Augmentation for MR Image Segmentation". In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020*. Ed. by A. L. Martel, P. Abolmaesumim, D. Stoyanov, et al. Springer International Publishing, 2020, pp. 667–677. DOI: `10.1007/978-3-030-59710-8_65` (cit. on p. 43).

75. Z. Xu, D. Liu, J. Yang, et al. "Robust and Generalizable Visual Representation Learning via Random Convolutions". In: *International Conference on Learning Representations*. 2021. DOI: `10.48550/arXiv.2007.13003`. URL: `https://openreview.net/forum?id=BVSM0x3EDK6` (cit. on p. 43).

76. *FeTA 2022 Top Submissions*. URL: `https://feta.grand-challenge.org/feta-2022-top/` (visited on 08/09/2025) (cit. on p. 43).

77. L. R. Dice. "Measures of the Amount of Ecologic Association Between Species". *Ecology* 26:3, 1945, pp. 297–302. DOI: `10.2307/1932409` (cit. on p. 43).

78. F. Hausdorff. *Set Theory*. 4th ed. Chelsea Publishing Company, New York City, NY, USA, 1991. ISBN: 9780828401197. URL: `https://books.google.it/books?id=L9ogL2TNTlIC` (cit. on p. 43).

79. S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, et al. "scikit-image: image processing in Python". *PeerJ* 2, 2014. DOI: `10.7717/peerj.453` (cit. on p. 44).

80. *scikit-image*. Version 0.25.2. 2025. URL: `https://scikit-image.org/` (cit. on p. 44).

81. F. Wilcoxon. "Individual Comparisons by Ranking Methods". *Biometrics Bulletin* 1:6, 1945, pp. 80–83. DOI: `10.2307/3001968` (cit. on p. 44).

82. P. Virtanen, R. Gommers, T. E. Oliphant, et al. *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python*. 2020. DOI: `10.1038/s41592-019-0686-2` (cit. on p. 44).

83. J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Taylor & Francis, 2013. ISBN: 9781134742707. URL: `https://books.google.it/books?id=2v9zDAsLvA0C` (cit. on p. 44).

84. *Effect size*. URL: `https://en.wikipedia.org/wiki/Effect_size` (visited on 09/09/2025) (cit. on p. 44).

85. F. Isensee, T. Wald, C. Ulrich, et al. *nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation.* 2024. arXiv: `2404.09556 [cs.CV]`. URL: `https://arxiv.org/abs/2404.09556` (cit. on p. 47).
86. Helmholtz Imaging Applied Computer Vision Lab. *nnUNet-ginipa.* 2025. URL: `https://github.com/sim1-99/nnUNet-ginipa` (cit. on p. 48).

A Supplementary Plots

A.1 Low-Quality Kispi-mial Scans

In Fig. A.1 are shown some examples of especially low-quality Kispi-mial scans, which exhibit motion artefacts and low SNR. These characteristics make the segmentation task more challenging for the models trained on this dataset.

A.2 General Performance

In the plots below are shown the volume similarity (Figs. A.2–A.4) and Hausdorff distance 95th percentile (Figs. A.5–A.7) across datasets and labels for the three models. The model predictions are realized on the test set of the same dataset the model was trained on (in-domain), and on the whole set (both train and test) of the other datasets (out-of-domain, OOD). The plots regarding the Dice score are in Section 7.1.

A.3 Comparison of Model Performances

In the plots below are shown the KDE plots related to the comparison between the nnU-Net default DA (baseline model) and the GIN-IPA augmentation model. The plots represent the volume similarity (Fig. A.8) and the Hausdorff distance 95th percentile (Fig. A.9) across each label and globally, from models trained on Kispi-irtk and inferring on dHCP. The plots regarding the Dice score are in Section 7.2.

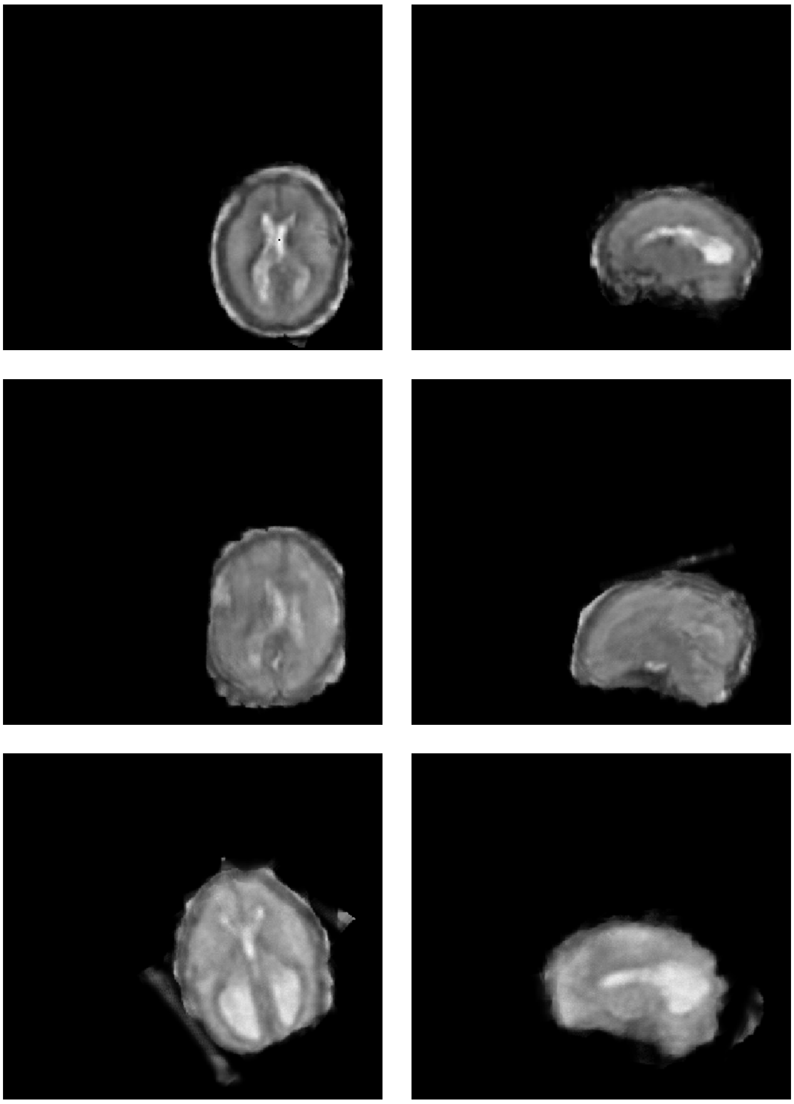

Figure A.1: Example of low-quality Kispi-mial scans: axial (left) and sagittal (right) views. Material from: [25, 68].

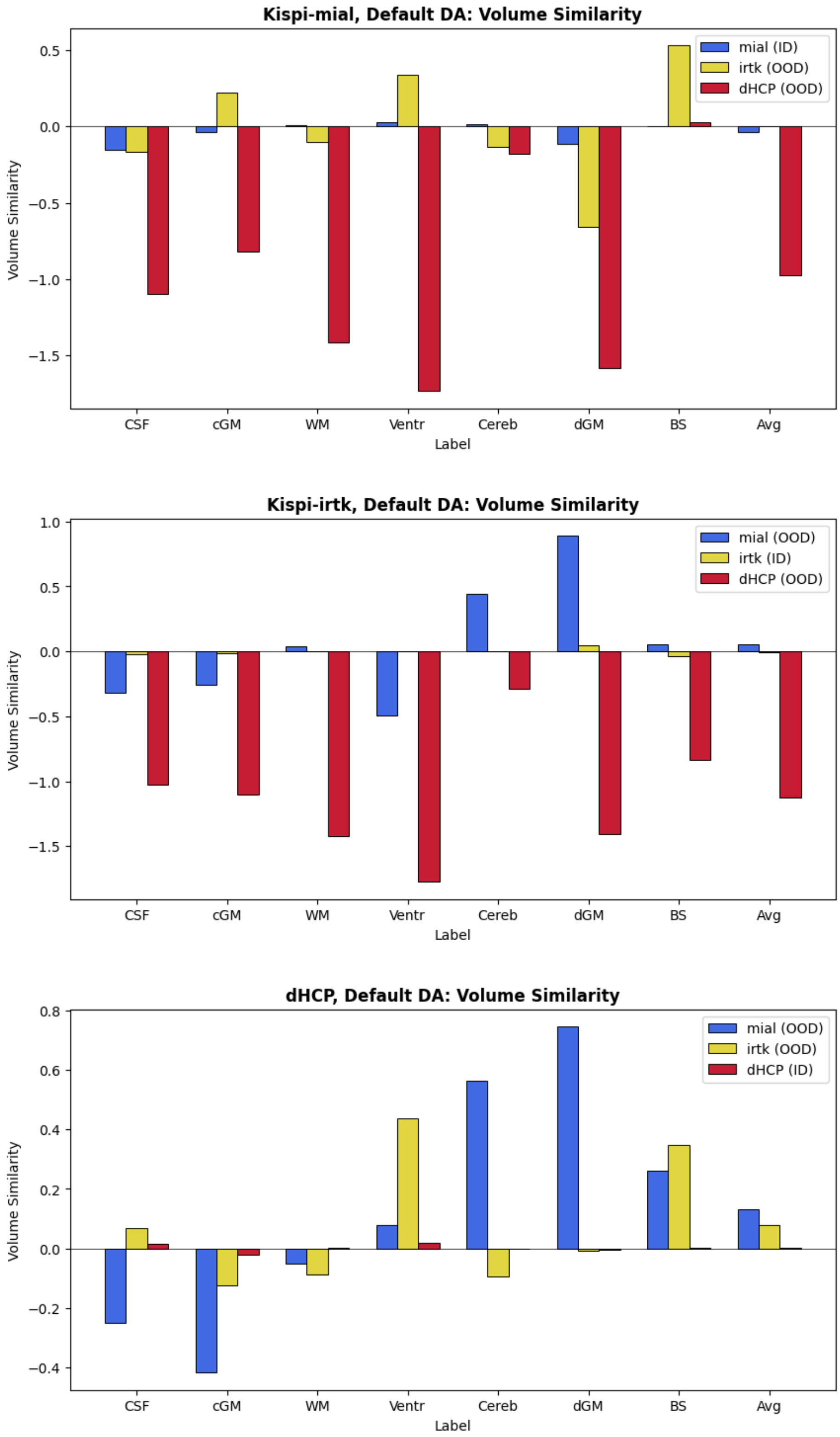

Figure A.2: Volume similarity across datasets and labels for the nnU-Net default DA (baseline model). From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP.

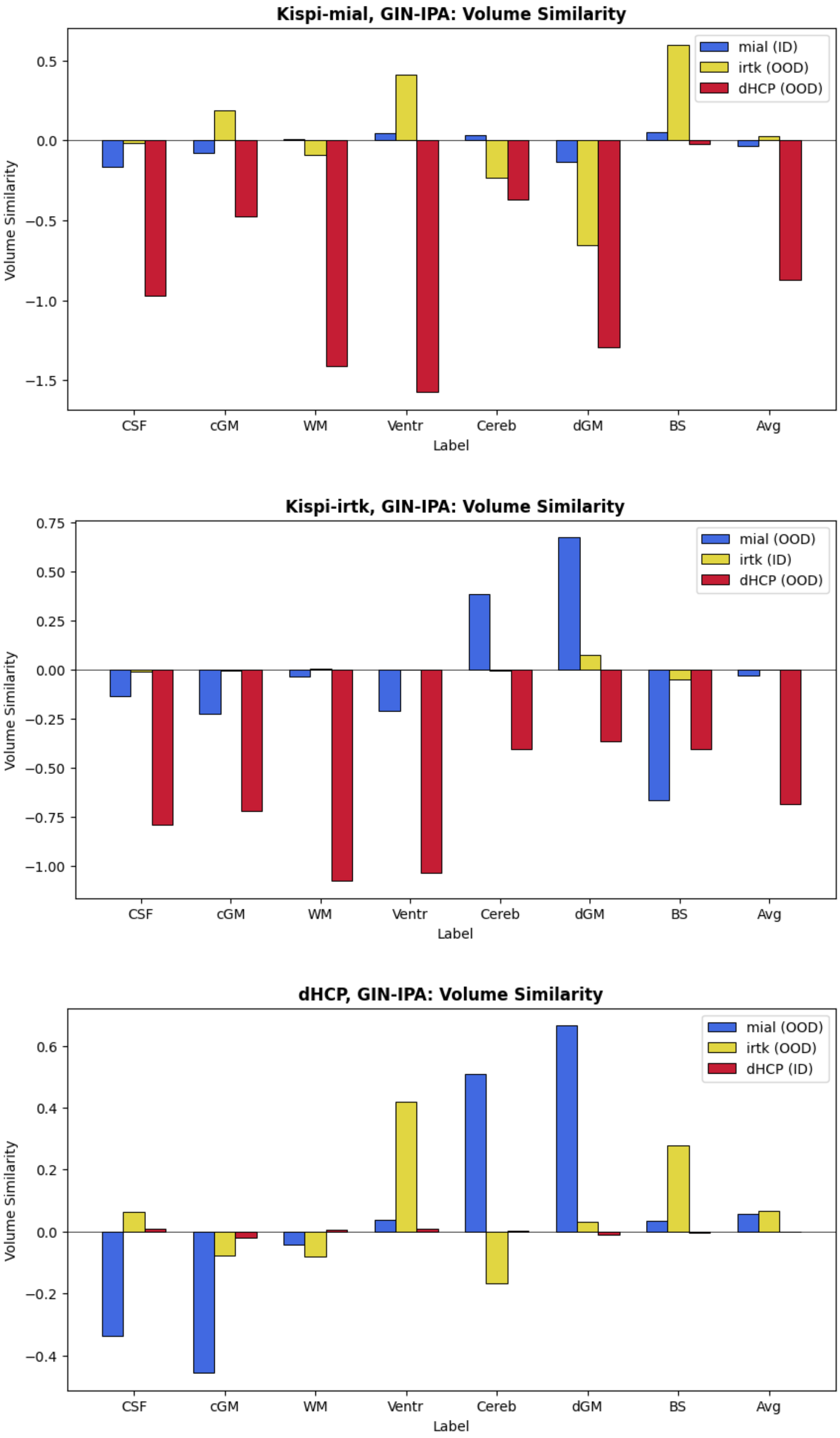

Figure A.3: Volume similarity across datasets and labels for the GIN-IPA DA model. From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP.

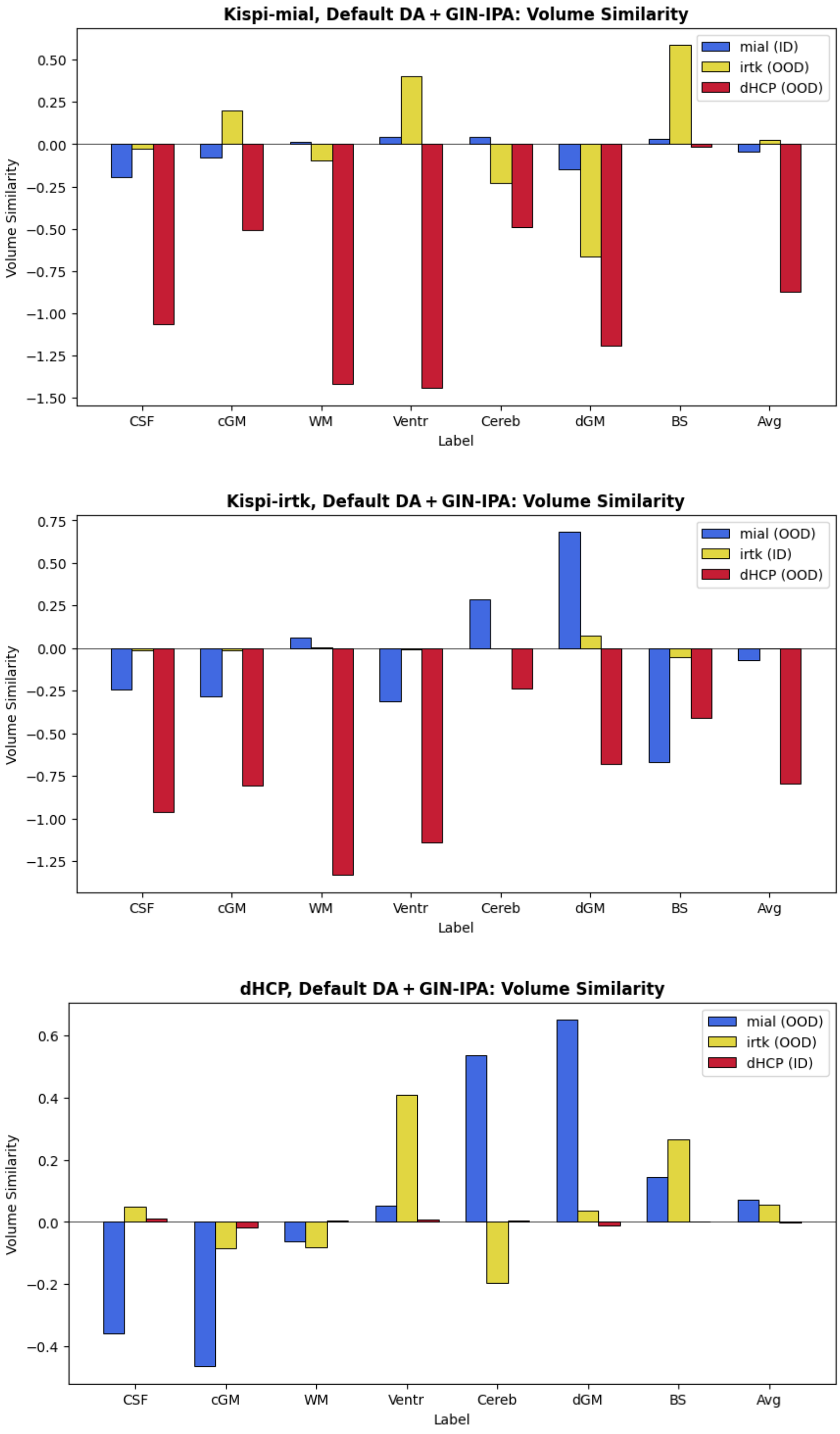

Figure A.4: Volume similarity across datasets and labels for the combined DA (default + GIN-IPA) model. From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP.

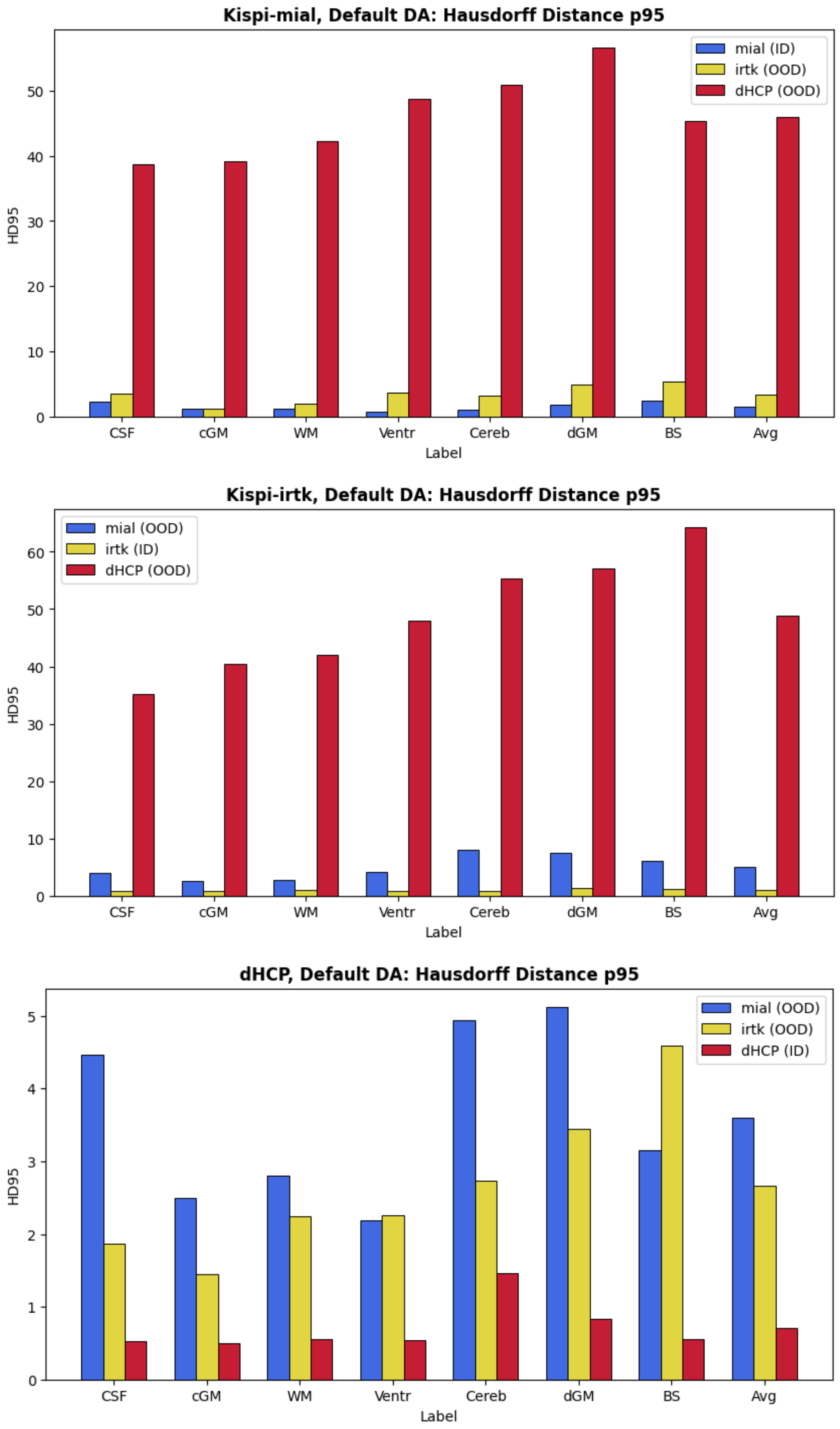

Figure A.5: Hausdorff distance 95th percentile across datasets and labels for the nnU-Net default DA (baseline model). From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP.

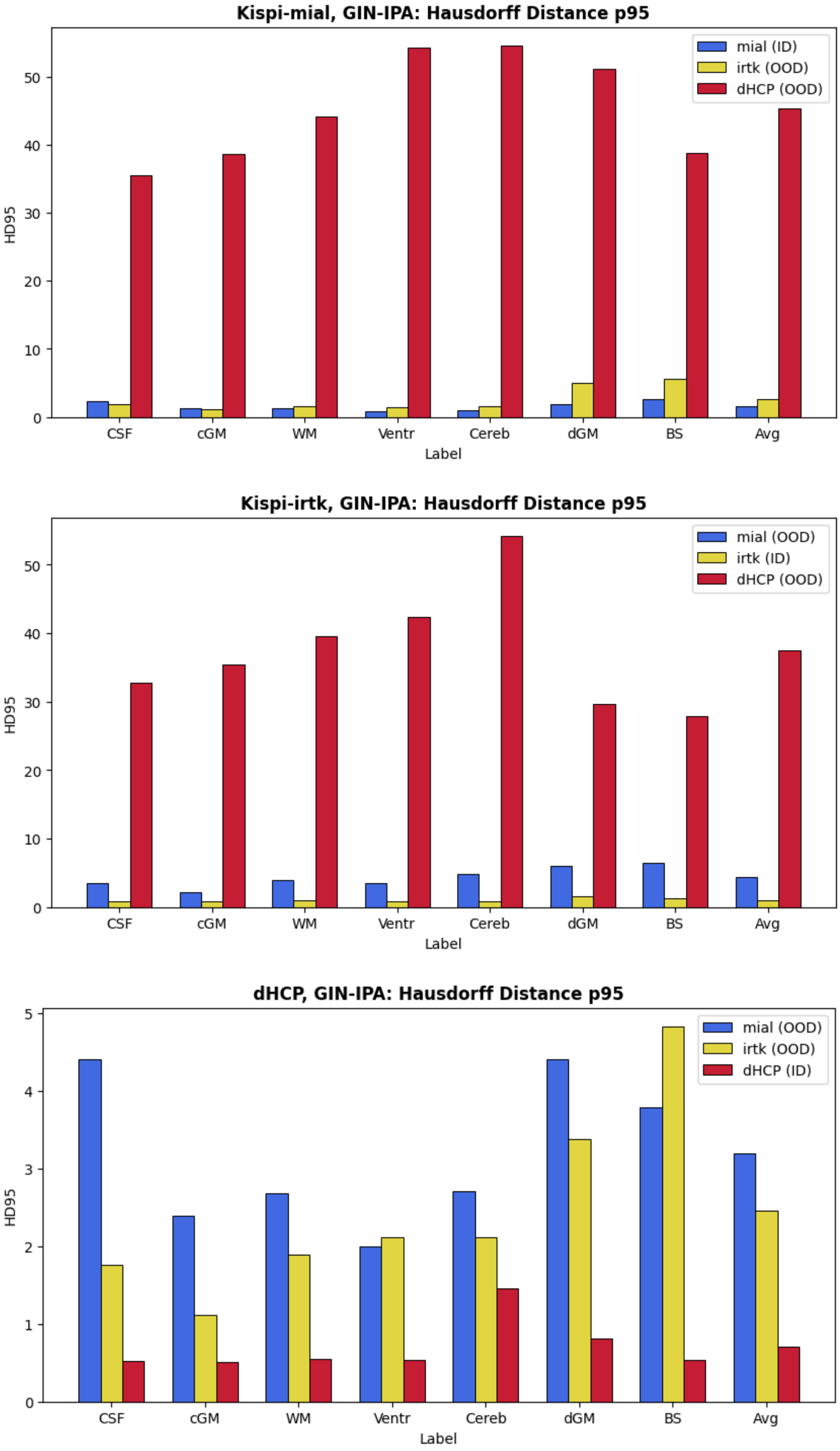

Figure A.6: Hausdorff distance 95th percentile across datasets and labels for the GIN-IPA DA model. From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP.

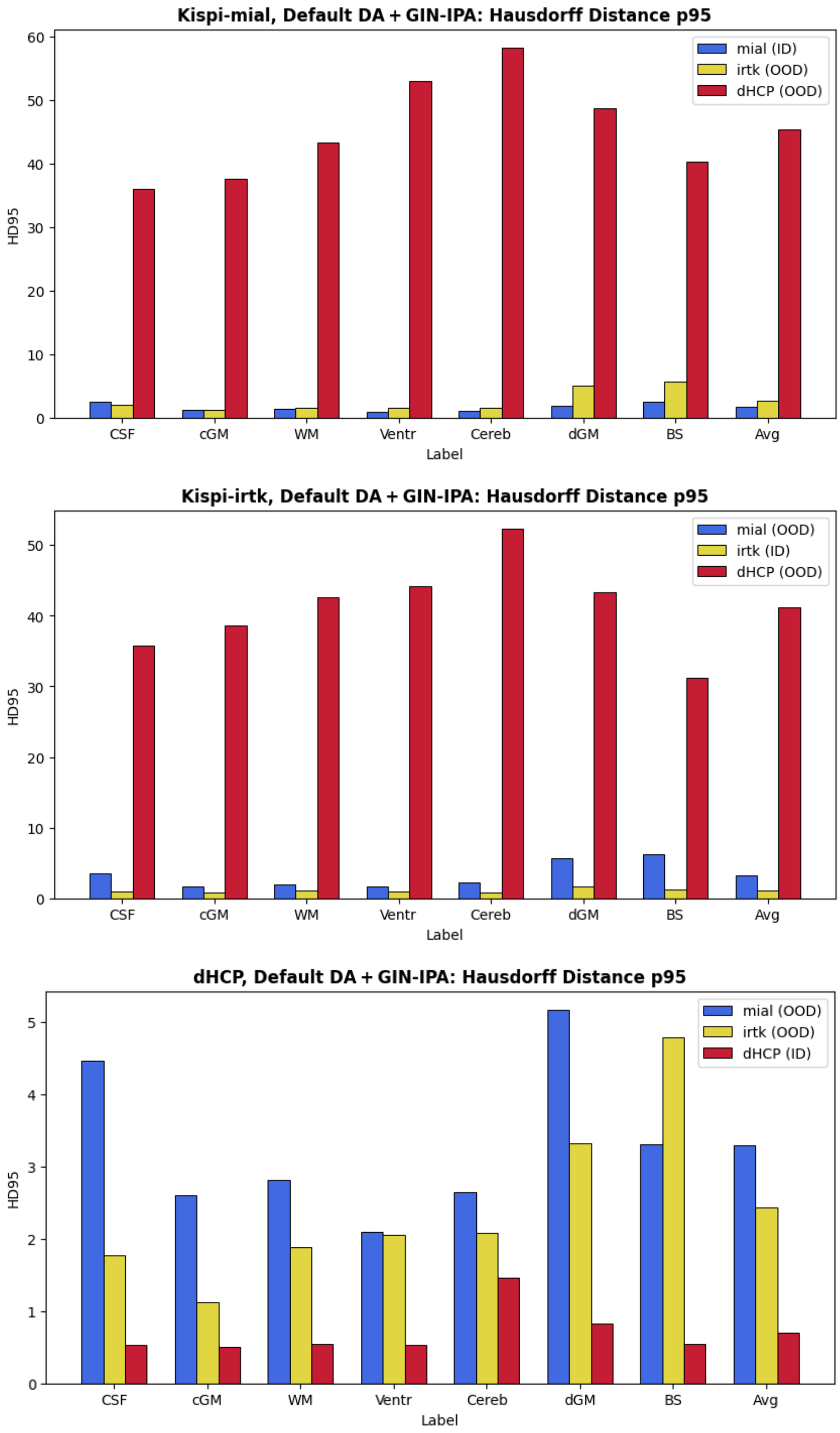

Figure A.7: Hausdorff distance 95$^{\text{th}}$ percentile across datasets and labels for the combined DA (default + GIN-IPA) model. From top to bottom: training on Kispi-mial, on Kispi-irtk, and on dHCP.

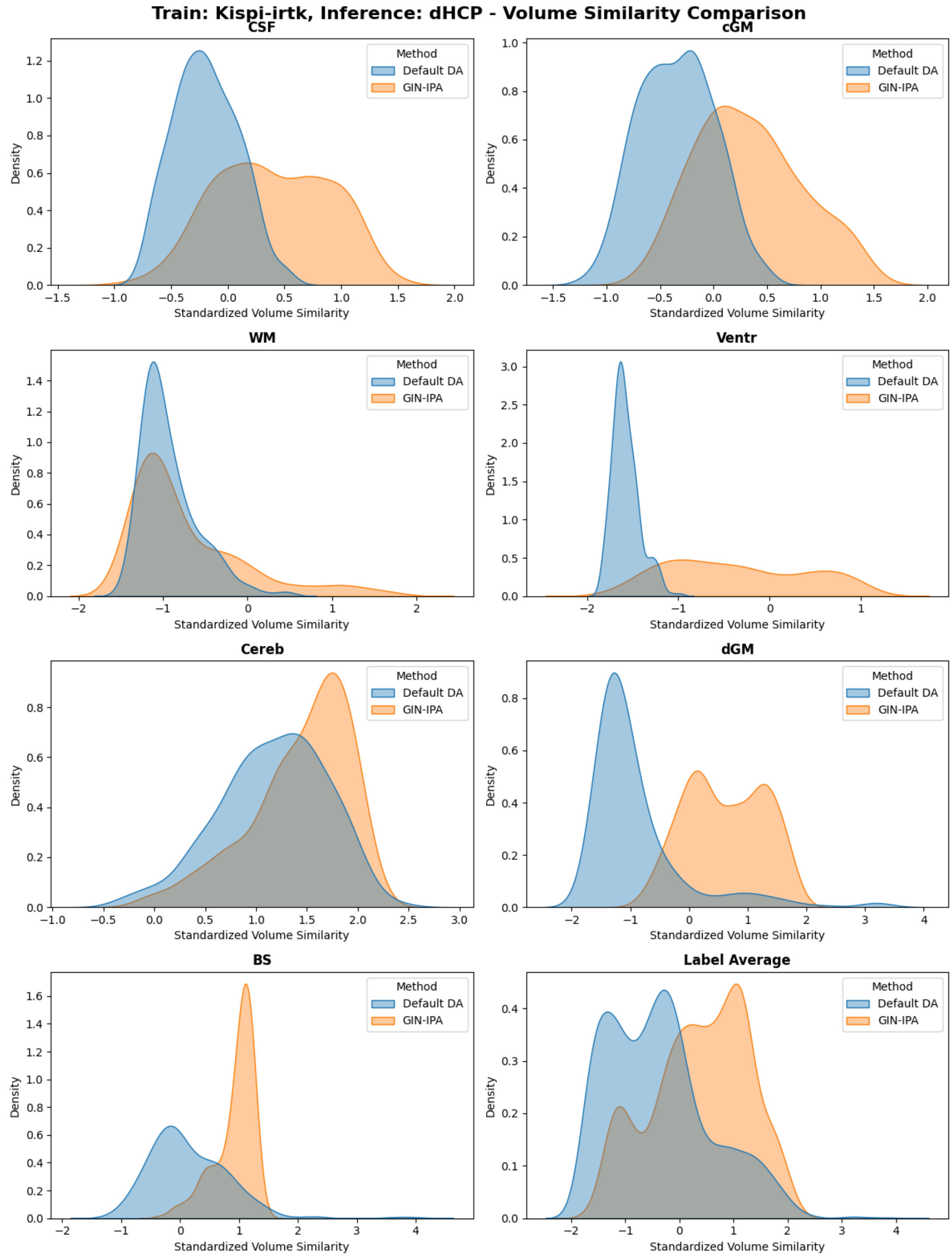

Figure A.8: Baseline vs. GIN-IPA: KDE plots of the volume similarity across each label and globally, from models trained on Kispi-irtk and inferring on dHCP.

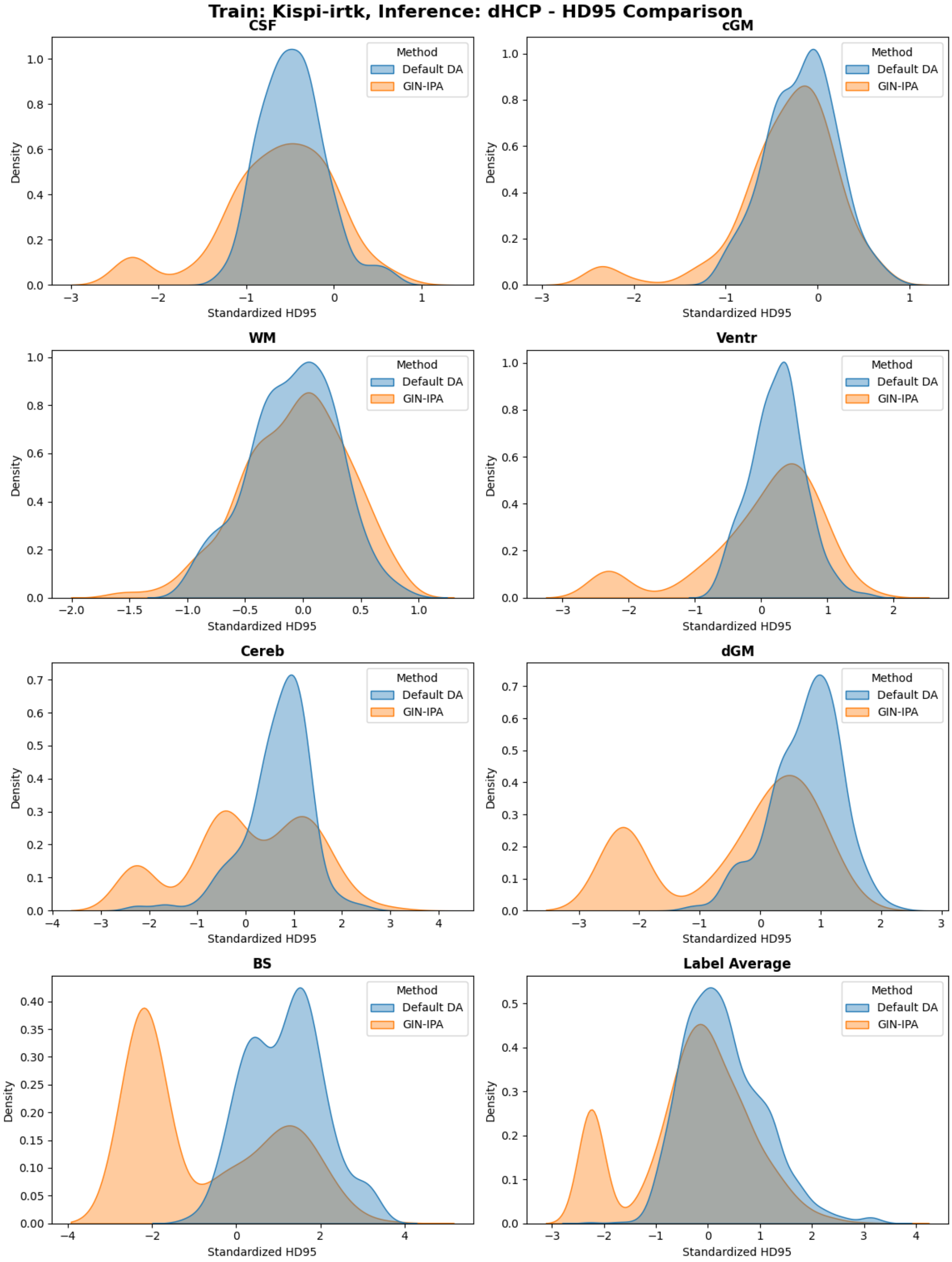

Figure A.9: Baseline vs. GIN-IPA: KDE plots of the Hausdorff distance 95th percentile across each label and globally, from models trained on Kispi-irtk and inferring on dHCP.

B SUPPLEMENTARY TABLES

B.1 KISPI-DHCP LABEL MATCH

Kispi labels	**dHCP labels**
1. CSF	1. Cerebrospinal fluid
2. cGM	2. Cortical gray matter
3. WM	3. White matter 17. Corpus callosum
4. Ventricles	4. Left ventricle 5. Right ventricle 6. Cavum septum 15. Third ventricle 16. Fourth ventricle
5. Cerebellum	8. Left cerebellum 9. Right cerebellum 10. Vermis
6. dGM	11. Left nucleus caudatus and putamen 12. Right nucleus caudatus and putamen 13. Left thalamus 14. Right thalamus
7. BS	7. Brainstem

Table B.1: Conversion table between Kispi and dHCP labels.