Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Universität Zürich[UZH]

# Domain adaptation for multi-centric fetal and infant MRI segmentation

## MASTER THESIS

supervised by

MSc. Céline Steger
Prof. Dr. András Jakab
Prof. Dr. Sebastian Kozerke

submitted at the

ETH Zurich
Dept. of Information Technology and Electrical Engineering

by

Charles Moatti
Matriculation number 19-952-449
Stapferstrasse 61
8006 Zürich
Schweiz
Zürich, October 2022

This page is intentionally left blank.

# Abstract

MRI has become an important clinical tool for the diagnosis of congenital and acquired brain abnormalities in fetuses and infants. In the research context, the quantitative analysis of the shape and volume of anatomical structures requires segmenting the brain into these tissues. Automatic methods have been previously applied in the field of medical image segmentation, with deep learning techniques being the state-of-the-art. While these methods perform well when the MR images to segment originate from the same probability distribution (termed *domain*) as the ones used in training, they tend to fail to generalize to new domains. This is a critical issue in multi-centric MRI datasets, which are commonly used in research collaborations, or when multiple sources have to be pooled to achieve good statistical power. The goal of this thesis is to select accurate and robust segmentation methods from the current state-of-the-art in medical image segmentation, implement them on our multi-centric neonatal and fetal data and evaluate their performance and capacity at domain adaptation. The implemented techniques are the 2d U-Net, the nnU-Net (2d and 3d) and the swin-UNETR. The 3d nnU-Net is the most accurate of our selected methods, outperforming the method previously used at the University Children's Hospital Zürich - the 2D U-Net - by 5.1% on our neonatal dataset. Our findings further indicate that it is the most generalizable method, surpassing the other implemented methods in our domain adaptation experiments. It has a performance drop of 0.58% when segmenting data from a new center, unseen at training time, 6.6% when segmenting data of a new structure (fetal), and 11.0% when segmenting synthetic MR images. Our results indicate that a method with strong generalization capability exists. Such a method has great potential to be used in large-scale clinical studies and on multi-centric datasets.

# Acknowledgements

I would like to thank Prof. Dr. András Jakab for his supervision and his review of the manuscript. I am grateful for his guidance, involvement and support in every step of the research. Without him, most plots of this thesis would have been lost unindexed on one of Kispi's Linux machine.

I would like to thank MSc. Céline Steger for her supervision, her review of the manuscript and her constant availability. I am grateful for her insights, dedication and support throughout. She has greatly helped me to bridge the gap in medical knowledge I had during the thesis.

I would like to thank Prof. Dr. Sebastian Kozerke for his supervision, his review of the manuscript and his flexibility.

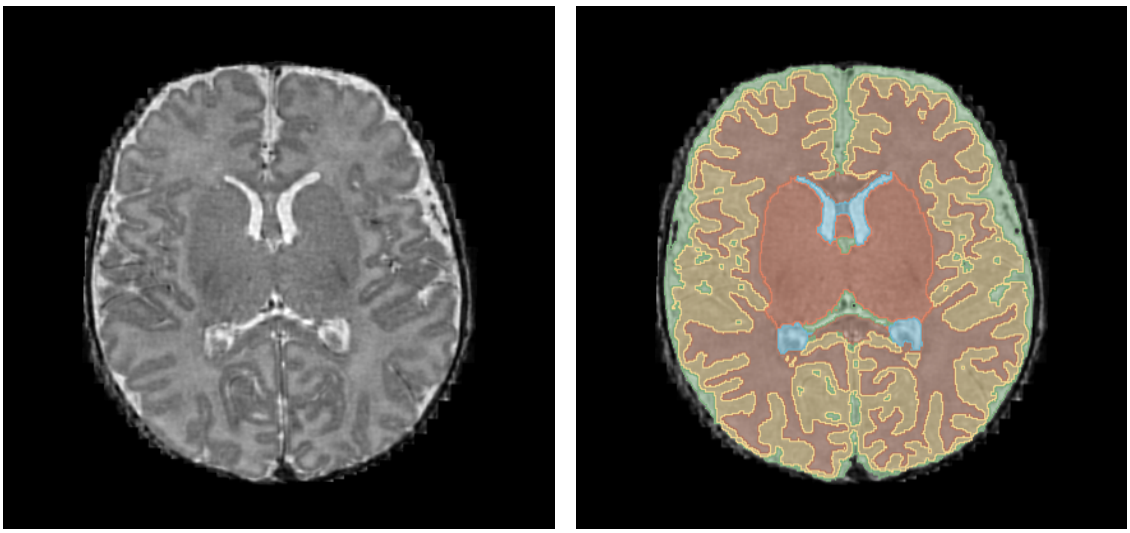
I would also like to thank the whole FII research group, past and present, as my work has built on top of the work of many others. I am grateful for the Research Group Heart and Brain at the University Children's Hospital and the study participants for providing valuable MRI data for my research.

# Contents

# 1 Introduction and Motivation

Due to recent technical improvements that led to shorter imaging time in MRI, fetal and infant MRI became a clinical reality. They can image the developing human brain at an unprecedented resolution and depict various tissue types and anatomical structures. This qualitative information is valuable for clinicians to observe if brain development fits into a normal trajectory. Furthermore, due to its high sensitivity, fetal and infant cerebral MRI is useful for diagnosing congenital disorders affecting the brain, such as spina bifida, corpus callosum agenesis or congenital heart defects. Infant MRI is valuable for the assessment of brain tissue impairments after premature birth or for the follow-up of infants suffering from congenital disorders. Such disorders lead to different shapes and volumes of different fetal brain tissues and these abnormal structures can be observed through a brain MRI. The quantitative analysis of brain development requires the measurement of size and shape of neuroanatomical structures (brain morphology).

Brain segmentation (shown in figure 1.1) is the image analysis method that outlines certain anatomical structures. It is therefore an important emerging tool for the diagnosis and assessment of the course of certain diseases. In addition, it is a research objective in and of itself as it allows evaluating how well segmentation algorithms perform in the context of real medical tasks, thereby driving scientific discovery.



(a) Before segmentation. (b) After segmentation.

Figure 1.1: Brain segmentation. Segmentation by the draw-EM algorithm of a dHCP axial slice into different tissues: the Cerebrospinal fluid (green), grey matter (yellow), white matter (red), ventricles (blue) and deep grey matter (orange).

Fetal MRI has gained popularity in the last ten years, due to technical advancements in MRI acquisition techniques. Single-shot fast spin echo, turbo spin echo or similar ultra-fast MRI sequences are used, yielding excellent soft tissue contrast and reduced motion artifacts [1]. More recently, it has become possible to reconstruct the single-shot 2D ("low resolution") images into a high-resolution image, which step also combats fetal or infant motion and other artifacts. The 2D images are then reconstructed into a 3D volume using slice-to-volume reconstruction methods: rigid or deformable e.g. SV-RTK or NiftyMIC algorithm [2, 3]. The importance of these 3D super-resolution datasets (3D-SR) is that they have isotropic image resolution, while the original fetal or infant MRI usually consist of thick-slice images to accelerate imaging. They are optimal for image segmentation tasks. Generally, fetal and infant MRI are vastly different from adult images in terms of the image contrast – as the brain is still in development – but also because they are reconstructed images and not primarily acquired 3D T2. Therefore, specialized tools are necessary for the segmentation and further analysis of these images.

To this end, the FeTA challenge (at the MICCAI Conference 2021) was organized to promote the development of automatic, reliable, valid and reproducible segmentation methods for fetal brain segmentation on the FeTA dataset [4]. The challenge used a dataset consisting of fetal 3D-SR MRI images, with each 3D image manually labelled into seven anatomical structures [5]. Twenty international teams participated in the challenge; each one using a deep learning algorithm to accurately segment the 3D reconstructed MR images into seven different tissues.

Manual segmentation has the disadvantages of being time-consuming, error-prone and non-objective, depending on the human annotator. On the other hand, a trained deep learning approach will segment a new 3D brain volume in a matter of seconds, whereas it takes several hours for a specialist to do the same. In order for the AI approach to replace the human in the field of medical image segmentation, the AI needs to be on par with human segmentation performance. This is not always easy to achieve, especially when one tries to develop algorithms to work across different modalities, imaging centers, or patient age groups. The MR scanner, acquisition protocol, reconstruction methods and study cohort used all impact the 3D-reconstructed image's contrast and it can be difficult to segment new images when an algorithm has not been trained on such images [6].

Generalization in machine learning (ML) is the ability of an algorithm to generalize: extend its predictive power to test data, unseen at training time. When this testing data originates from a different probability distribution (called domain) than the training data, it is termed domain adaptation [7]. Many ML methods are very biased towards the domain of the training data, tend to overfit and fail to generalize to new domains [6].

Our goal in this thesis is to develop methods (or implement accurate existing ones) with strong generalization capabilities, in the domain of both fetal and infant MRI segmentation. For this purpose, we will work primarily on neonatal MRI data acquired at three different study centers (Zürich, London using the dHCP and Giessen), fetal MRI data of the FeTA dataset and synthetic MRI data. Pooling data across study centers will also allow us to reach good statistical power when evaluating our methods.

Another goal of this thesis is to deliver useful tools that the University Children's Hospital's research team can use in the future for its volumetric assessment of fetal and neonate brain MRI. This is particularly important since the Hospital is currently

conducting studies where fetal and infant MRI is pooled across study centers across Europe. For this purpose, my goal was to produce well-documented and reproducible methods.

In chapter 2, we will do a survey of the state-of-the-art methods in the field of medical image segmentation and select some of them for implementation and evaluation on our multi-domain data, described in chapter 3. In chapter 3, we will delve into the architectures of the selected methods and present our experiments and metrics used for evaluation. In chapter 4, we will evaluate their performance on our data, in a traditional train/test split, in a domain-adaptation setting, on fetal data and on synthetic MRI data. In chapter 5, we will discuss our results overall, explain how they fit into the field and analyse the limitations of our approach. The results will be summarized in chapter 6 and the future of the field will be explored.

# 2 Current state-of-the-art of the field

The problem of image segmentation is a long-standing problem in computer vision and digital image processing [8]. Some key differences between generic image segmentation and medical image segmentation are the variability of image modalities (MRI, CT, mammography, ultrasound, ...), image quality and image availability. While it is common in computer vision to have million of samples to train on, it is rather rare in the medical imaging field. There are several approaches to medical image segmentation; we can divide them into classical computer vision techniques (e.g. thresholding, region growing, deformable models, atlas-guided approaches, ...) and AI based techniques (e.g. classifiers, clustering, bayesian, deep learning, ...). Deep learning techniques have had great success in the recent years and will be our focus in this thesis.

In 2012 the first major breakthrough of deep learning in computer vision took place. In this year, Krizhevsky *et al.* achieved state-of-the-art performance, by a large margin, on the ImageNet dataset in image classification [9]. They used a convolutional neural network (CNN), an architecture developed by the ML pioneers since the 1980s [10, 11], with several millions of parameters and trained it on a million of classified images: this was the biggest CNN to date and largely surpassed the previous classification methods. Since then, CNN have been thoroughly studied, extended and applied to various vision tasks with great success [12]. In 2015, Long *et al.* proposed the fully convolutional network [13], extending the CNN architecture to work for image segmentation as well, and achieving state-of-the-art performance on various non-medical segmentation benchmarks.

In 2015, Ronneberger *et al.* proposed the U-Net [14], a powerful segmentation architecture built on top of the fully convolutional neural networks [13]. It is a U-shaped encoder-decoder network with skip connections, assigning every pixel of an input 2D image to its most probable tissue structure. The network architecture will be described in detail in Chapter 3. It had the advantage of requiring less input data than the previous methods for a better performance [14]. The authors showed their improvement over the current state-of-the art by winning the ISBI challenge for segmentation of neuronal structures in electron microscopic tasks [15] and the ISBI cell tracking challenge 2015 in two categories [16]. These successes in diverse tasks showed that their network is not tailored to perform well on a specific task, but has the ability to generalize to new tasks. In 2016, the same team of researchers released the 3D U-Net which extended their architecture to perform volumetric segmentation without the need to partition 3D data into 2D slices. This proved to be more accurate than the 2D U-Net in the context of segmenting the *Xenopus* kidney, a complex 3D structure [17].

The FeTA challenge at MICAII 2021 was organized by KISPI in 2021 with the aim of promoting the development of automatic algorithms in the segmentation of the developing fetal brain. The key takeaways of this challenge is that U-Net is the dominant underlying architecture of all solutions. Moreover, two architectures stood out: nnU-Net and

SegResNet. Because of its well-documented use, we have placed our focus on the nnU-Net.

In 2017, the nnU-Net ("no-new-Net") was introduced by researchers at the DKFZ [18]. It is a self-configuring method which automatically configures the preprocessing, network architecture, training and post-processing of several U-Net architectures (2D and 3D) based on the properties of the dataset. Instead of proposing a novel U-Net like architecture with slight architectural modifications as many researchers were doing in this period (e.g. UNet++ [19], Attention U-Net [20], ResUNet [21]), the authors focused on systematizing the process of configuration of the U-Net. The nnU-Net won the 2018 Medical Segmentation Decathlon challenge (a challenge comprised of 10 different datasets and 14 different segmentations tasks of radiological data ranging from brain tumor segmentation to prostate or kidney), staying on top of the live leaderboard in the next two years following the challenge [22]. In addition, during these two years, nnU-Net was tested on 53 different segmentation tasks and achieved the first place in 33 of them, including the famous BraTS challenge 2020 and the KiTS challenge 2019 [18]. This confirmed the idea of the MSD challenge's organizers that a network that would generalize well to different tasks, would become the new state-of-the-art of the field, even overperfoming task-specific designed methods; this is what nn-Unet achieved.

In late 2020, some new algorithm took the first place of the MSD challenge; the DiNTS [23]. It is part of a broader class of methods using Neural Architecture Search (NAS) [24]. This is an empirical approach where the network's architecture itself is being optimized; several architectures are searched within an architectural space and selected automatically based on its predicting performance on a given task. In addition to being very computationally expensive and slow (much more than nnU-Net which is based on carefully designed heuristics), this method showed only marginal improvements over nnU-Net. In addition, it did not stay long at the top of the MSD leaderboard.

In late 2021, the Swin-UNETR (Shifted Window U-Net Transformers) became top of the leaderboard of the MSD challenge [25]. Swin-UNETR combines a hierarchical shifted window (Swin) transformer at the encoder with a fully convolutional neural network (FCNN) decoder [26]. It is part of a broader class of methods using transformers in biomedical image segmentation. Transformers have been the state-of-the-art in NLP since their development in 2017 [27], used in models such as BERT or GPT. In 2020, they were successfully applied to computer vision in the field of image classification - the Vision Transformers (ViT) proposed by Dosovitskiy *et al.* [28] outperformed state-of-the-art CNNs on various image classification baselines - and have gained popularity ever since. For a good review paper on the topic, see Shamshad *et al.* [29].

While some researchers suggest that transformers will make CNN obsolete in computer vision - and therefore medical imaging - there is still a long way to go and the state-of-the-art methods currently combine both approaches [26]. One disadvantage of CNN is that, because of their localized field-of-view in the convolutions, they fail to capture long-range interactions in an input image[1]. On the other hand, self-attention - the building block of transformers - can better model long-range interactions due to the tokenization of image patches [29].

While the swin-UNETR was reported to show performance improvements over the

[1] This locality is also what makes them efficient.

nnU-Net, at present it is not as well documented as the the nnU-Net. The next years will tell if this method turns out to stay the state-of-the art, in which case the authors should work on the usability of their code for implementation on new data. They did a step forward during the middle of my thesis by releasing a GitHub in June 2022.

In this thesis, we hypothesise that a network that generalizes to different biomedical imaging segmentation tasks such as the MSD challenge will also perform well on our domain adaptation problem. We therefore select three of the state-of-the-art methods mentioned above - U-Net, nnU-NET and swin-UNETR - for implementation, evaluation and comparison on our multi-centric neonate and fetal data acquired at the University Children's Hospital Zürich, Giessen and London through the developing human connectome project (dHCP) [30]. This evaluation will be performed in a classical setting first - where each method sees all domains at training time - and then in a domain adaptation setting - a domain is absent from training data, only present at inference time. Next we will have a look at the algorithm's performance on different sub-classes of data. By synthesising T2 data using a Generative Adversarial Network (GAN) as described in section 3, we will create another domain and see how the state-of-the-art methods perform on synthetic data.

# 3 Methods

## 3.1 MRI data sources and data heterogeneity

To resolve our domain adaptation problem, we will use neonatal MR scans acquired in 3 different centers: dHCP, Giessen and Zürich. We make a further division between the Zürich data before a scanner software upgrade in 2014 and after the upgrade. We hypothesise that the data looks different enough to constitute two separate domains. This gives us four distinct data domains to analyze. Additionally, fetal MRI data acquired at Zürich [5] will also be used and constitute another domain; in this case, the domain shift will be larger as fetal brains are still in development, and the reconstructed images have varying data quality [31].

1. **developing Human Connectome Project (dHCP) data**

   We have used dHCP data from the dHCP neonatal structural pipeline's second release [30]. The dataset consists of T1-weighted and T2-weighted MRI scans acquired in 558 sessions over 505 neonate subjects. The gestational age spread of the neonates is between 24 and 45 weeks with 75% between 37 and 42 weeks. The scanner used is a Phillips Achieva of 3T with a 32 channel phased-array head coil. The T2w and T1w multi-slice fast spin-echo were acquired in sagittal and axial slice stacks. The in-plane resolution of each slice is $0.8 \times 0.8$ mm$^2$ and the slice width is 1.6mm with a 0.8mm overlap between each consecutive slices. The repetition time (TR) and echo time (TE) are respectively TR/TE = 12000/156ms for T2w images and 4795/8.7ms for T1w images.

   The data then goes through the dHCP reconstruction and structural pipeline as described in [32]. Motion corrected volumetric reconstruction of multi-slice T2w images is performed [33]. This accounts for the fact that most neonates move during their sleep and are not sedated during the scan. Simple IRTK (old name of SV-RTK), a super resolution reconstruction is then applied [2]. The 3D data is then registered, bias corrected and segmented with the draw-EM algorithm, a semi-automatic segmentation method which uses manually-labeled atlases as structure priors. The resulting T2w images have voxel size of $0.5 \times 0.5 \times 0.5$ mm$^3$. The segmentation is composed of eight labeled tissues: cerebrospinal fluid, grey matter, white matter, ventricles, cerebellum, deep grey matter, brainstem and hippocampus. More details on the full structural pipeline can be found at [32].

   We selected a subset of 80 dHCP images in order to have a comparable number of cases across our centers. Among these 80, 52 are placed in the training set, while 28 are placed in the testing set. We note that, for most subjects, T1 images are also available in the same space as the T2 images. We will use these T1-T2 pairs

to train a Generative Adversarial Network (GAN) for synthetizing T2 images as described in section 3.5 and evaluate our methods on this new synthetic domain.

2. **Zürich**

   The data we used originated from neonates with congenital heart disease (CHD) and a control group, it was collected in the context of the Zürich Heart and Brain study [34]. The scans were performed on a GE Signa MR750: a 3.0T MRI scanner with a 8-channel head coil. 2D fast spin-echo T2w sequences were acquired in the axial, coronal and sagittal plane. The following sequence parameters were used, TR/TE: 59000/97ms, in-plane voxel dimensions: $0.7 \times 0.7$ mm$^2$ resampled to $0.35 \times 0.35$ mm$^2$, slice thickness: 2.5mm and slice gap: 2 mm.

   The data were run through the dHCP reconstruction and structural pipeline [32, 33]. It is therefore 3D reconstructed using Simple IRTK [2], bias corrected and automatically segmented into seven tissues by the draw-EM algorithm [35]. To ensure the validity of the segmentation, it was manually checked and re-run with different settings or manual masks in failure cases. If that still did not help, these cases were excluded.

   Infants with CHD have been observed to have smaller brain volumes for each tissue [34]; in addition, they are more prone to brain injuries. Given that the study cohort (CHD vs healthy), the scanner and the scanner parameters are different between the Zürich and dHCP data; this gives us two distinct domains of data to test our segmentation methods on.

   Moreover, we hypothesise that the software scanner upgrade performed in 2014 caused the data within the Zürich dataset to look different and keep these data separated as two domains in the rest of our study. We will refer to these two domains as *Zürich pre-upgrade* and *Zürich post-upgrade* in the rest of this thesis. The number of data samples used in total is 172 with 67 acquired before the scanner upgrade and 105 acquired after.

3. **Giessen**

   The University Hospital in Giessen was part of an ongoing multi-centric study on CHD. The scanner used is the Siemens Verio 3.0T. The TR/TE is variable across scans and even within one patient's scan. Data is collected over the three planes. It was then run through the dHCP pipeline [32, 33] where it gets 3D reconstructed using Simple IRTK [2], bias corrected and segmented with draw EM [35].

   The number of data samples used in our study is 8.

4. **Fetal data**

   Lastly, we will use fetal MRI data from the University Children's Hospital (FeTA dataset [5]). The scanners used are the Signa MR450 and MR750 with a 8-channel body coil, TR is set between 2000 and 3500ms with TE set to 200ms. Half of the cases are SR reconstructed using MIAL [36, 37] and the other half with Simple IRTK [2]. The resulting SR images have resolution $0.5 \times 0.5 \times 0.5$ mm$^3$, shape

$256 \times 256 \times 256$ voxels. The segmentation into seven tissues was done manually by experienced individuals [5]. More details can be found here [4].

We use only a subset of the FeTA dataset: neurotypical subjects with gestational age (GA) more than 28 weeks. Such a restriction is applied to evaluate our methods, which will be trained on neonate data, in the fairest of way - they will already be exposed to new structure of data, we want to limit the amplitude of domain shift in our domain adaptation study. This leaves us with 25 samples with a GA range of 28 to 35 weeks.

The details are summarized in the following tables.

| Center | Field | Coils | TR/TE | Recon. | Ground Truth | GA (weeks) |
|---|---|---|---|---|---|---|
| dHCP | 3T | 32 | 12000/156 | IRTK | draw-EM | 31-44 |
| Zürich | 3T | 8 | 59000/97 | IRTK | draw-EM + checks | 37-48 |
| Giessen | 3T | 32 | variable | IRTK | draw-EM | 36-38 |
| Fetal | 3T | 8 | 2750/200 | MIAL/IRTK | manual | 28-35 |

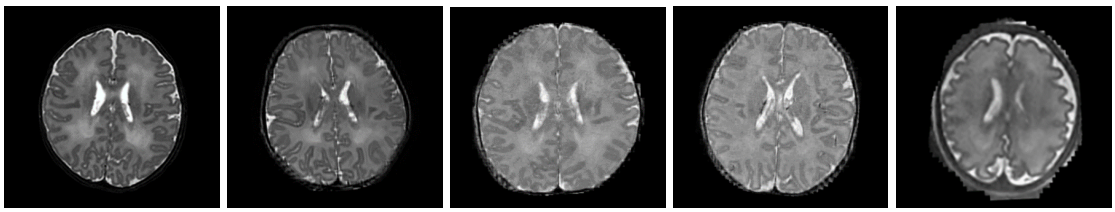
Table 3.1: Summary of the data domains used in our study: from a scanner, pipeline and cohort's perspective. GA refers to corrected gestational age in postnatal datasets.

| Center | Resolution | Median shape | Labels | # of Samples | Training size | Test size |
|---|---|---|---|---|---|---|
| dHCP | $0.5^3$ | (283, 308, 212) | 8 | 80 | 58 | 22 |
| Zürich | $0.4^3$ | (284, 236, 254) | 8 | 173 | 115 | 57 |
| Giessen | $0.4^3$ | (277, 243, 287) | 8 | 8 | 5 | 3 |
| Fetal | $0.5^3$ | (256,256,256) | 7 | 25 | 0 | 25 |

Table 3.2: Summary of the data domains used in our study: details of the reconstructed images, segmentation labels and data size.

An example axial slice of a subject MRI for each center is shown in figure 3.1. For a fair comparison between neonatal centers, we selected subjects to be about 40 gestational weeks. With an eye inspection, we observe that Zürich data has less grey matter - white matter contrast than dHCP or Giessen, that Zürich pre and post-upgrade appear similar and the same for dHCP and Giessen. This is also confirmed quantitatively when we compare the intensity distribution of voxels (3D pixels) across domains. To have a fair comparison between our domains, we perform a min-max scaling (to 0-1) of the data for each center.[1] We then select 400 000 voxels at random within each center's MRI data to get an intensity distribution. The resulting intensity distribution for each domain is displayed in a violin plot (see figure 3.2).

[1] A z-score normalization, as done subsequently, would be preferable but it is not viable here. The relative differences between domains would be lost.



(a) dHCP data (b) Giessen (c) Zürich pre-upgrade (d) Zürich post-upgrade (e) Fetal

Figure 3.1: Inter-center domain shift of our data: example MRI axial slice for each domain. One can observe the differences in image contrast across domains.

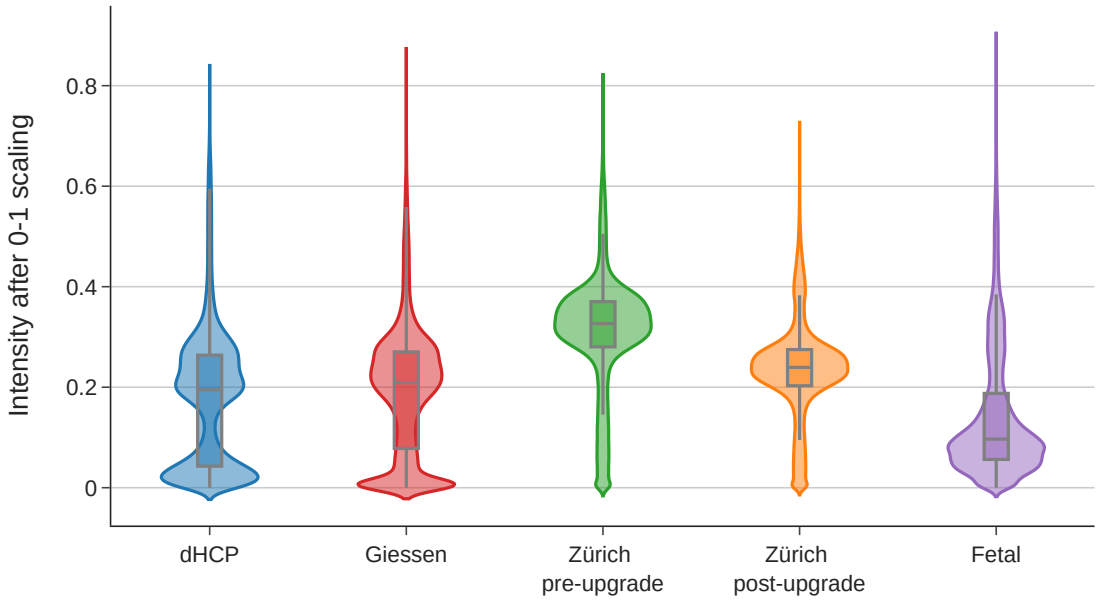



Figure 3.2: Inter-center domain shift of our data: voxel intensity distribution for each domain. dHCP and Giessen data both have a bimodal intensity distribution with medians 0.18 and 0.21 respectively. Zürich pre and post-upgrade have a unimodal intensity distribution with medians 0.29 and 0.25.

To extend our domain shift analysis, we also look at per-label intensity distribution over all samples of a given center and compare across centers. Before comparing the label intensities of different centers, we perform a z-score normalization per sample ($X = \frac{X-\mu}{\sigma}$) to standardize the absolute intensity range across centers. This z-score normalization is also performed in the pre-processing of our segmentation methods; the resulting intensity values are thus the same as at the input layer of our networks. The resulting intensity distribution per label within each center is shown in figure 3.3.

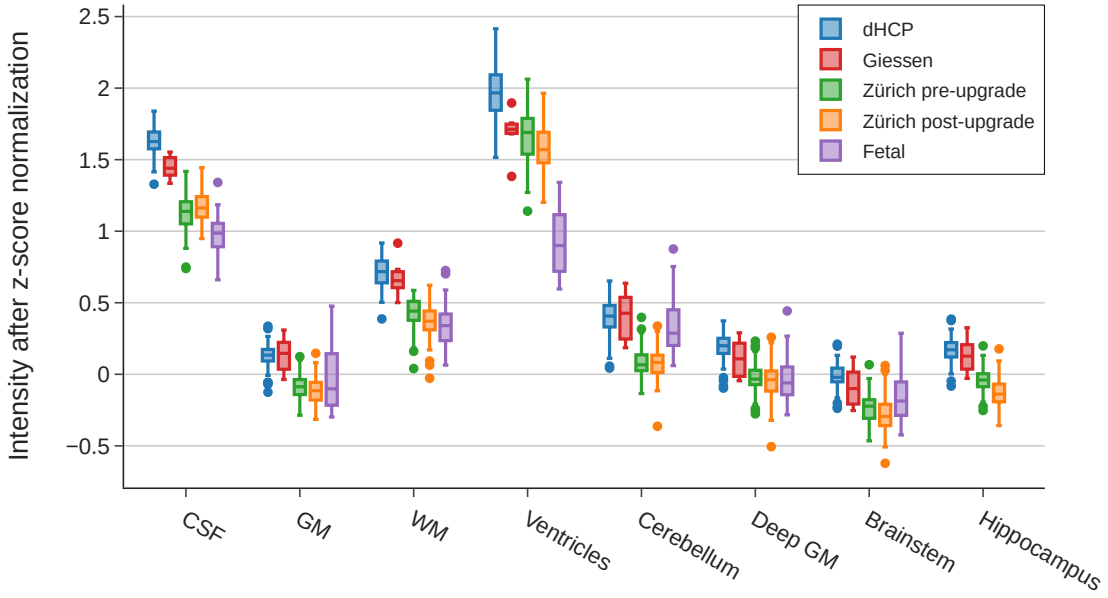


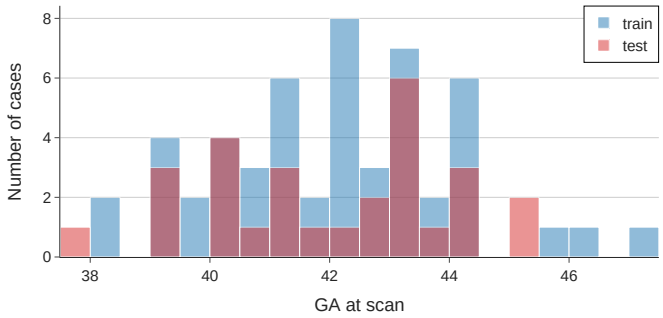

Figure 3.3: Grouped plot of intensity distribution after z-score normalization for each anatomical structure's label and center. CSF: cerebrospinal fluid, GM: grey matter, WM: white matter.

Zürich pre-upgrade and Zürich post-upgrade data share the same range of intensity values for some labels. For each label, we make the null hypothesis that these two domains share the same mean intensities on that label (after z-score normalization per sample). We perform an unpaired two-sample t-test between the two distributions and cannot reject the null hypothesis for the cerebellum ($p = 0.97$), the grey matter ($p = 0.09$) and the deep grey matter ($p = 0.41$). For the other labels, the null hypothesis is rejected with $p<0.05$.

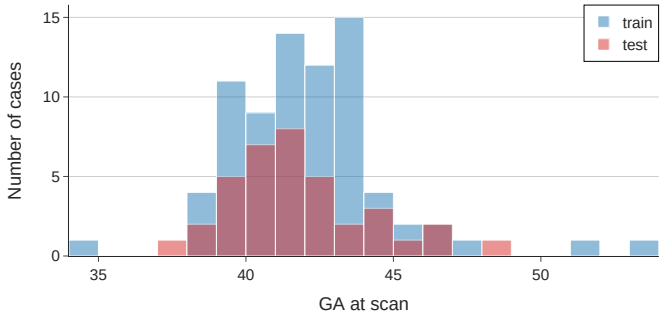
We can also observe similarities between dHCP and Giessen data and apply the same procedure (t-test within each label to test if the mean intensity is the same). The null hypothesis cannot be rejected for the grey matter ($p = 0.88$), the white matter ($p = 0.2695$), the cerebellum ($p = 0.95$) and the hippocampus ($p = 0.25$).

The fetal data shows major difference in the intensity of the MRI within the ventricles label. It appears substantially darker than the same label in other domains. The inter-quartile range (IQR) of the grey matter is also much larger for the fetal data; this could be explained by the constantly changing cortical surface of the fetus throughout the gestation - due to neuronal migration and gyrification - causing contrast variations across ages [4].

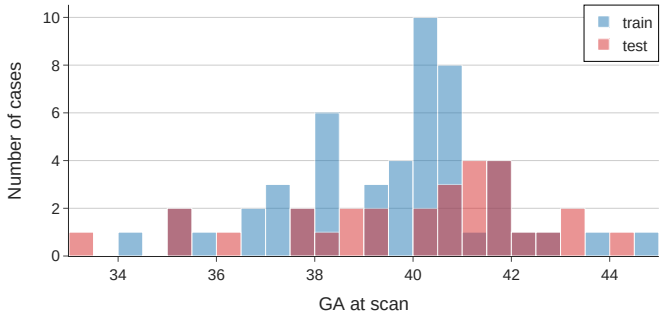
A 66/33% train test split is performed on each domain independently (except fetal data where all data is put in the test set). We have 172 samples in the training set - 58 from dHCP, 5 from Giessen, 44 from Zürich pre-upgrade and 71 from Zürich post-upgrade - and 114 samples in the test set - 28 from dHCP, 3 from Giessen, 24 from Zürich pre-upgrade, 34 from Zürich post-upgrade and 25 from fetal data. We performed the split while ensuring matching gestational age (GA) distribution in train and test split for each domain. This distribution is shown in figure 3.4.




(a) Zürich pre-upgrade




(b) Zürich post-upgrade




(c) dHCP

Figure 3.4: Age distribution of the train/test split for every center. Giessen is excluded because it only has eight samples. GA refers to corrected gestational age for postnatal samples.

*The pregnant women, parents or legal guardians of infants gave written informed consent for the further use of their health-related data in research. The Zurich CHD dataset was created in a prospective observational study 'Heart and Brain', approved by the ethical committee of the Canton of Zürich. The ethical committee of the Canton of Zurich approved the retrospective studies that collected fetal and neonatal MRI data (KEK decision numbers: ID 2016-01019 and ID 2022-01157). The dHCP dataset was publicly released under the approval of the National Ethics Committee of the United Kingdom.*

## 3.2 U-Net

The U-Net architecture was developed in 2015 by Ronneberger *et al.* [14]. It is a U-shaped fully convolutional neural network taking as input a 2D image and outputting a label map which assigns each pixel to its most probable class.

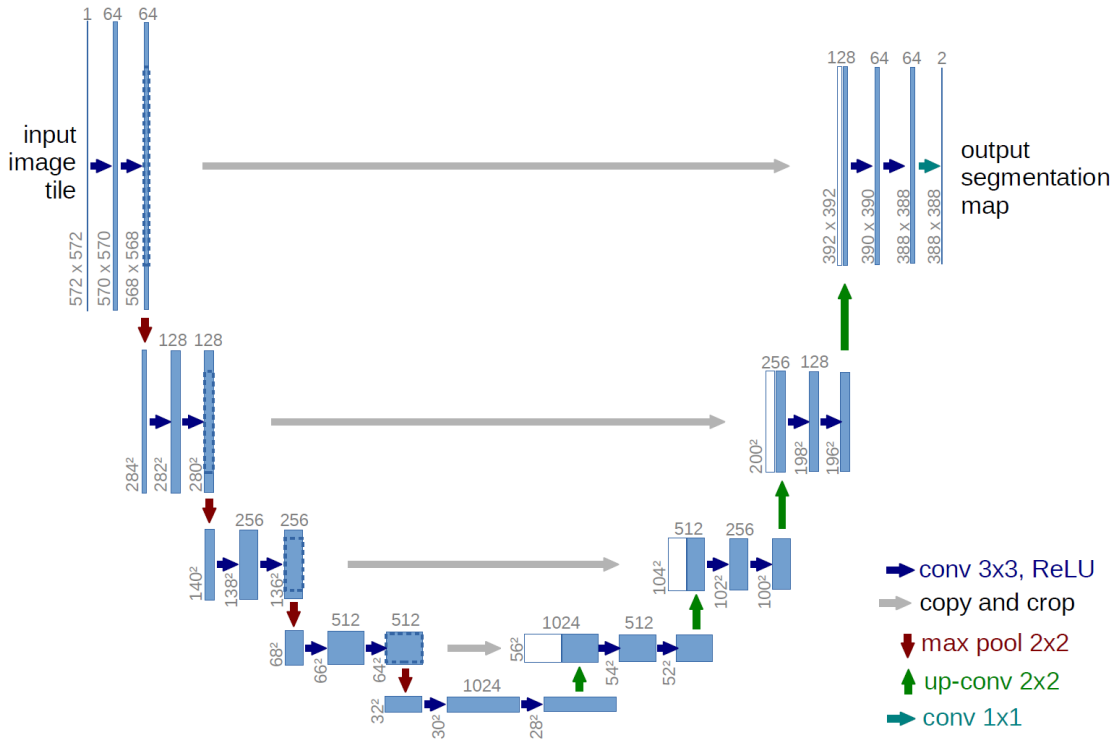



Figure 3.5: The original U-Net architecture with a depth of 4 proposed by Ronneberger *et al.* in 2015 [14]. The network consists of a contracting, encoding path in the left half and an expanding, decoding path in the right half of the "U" with skip connections. Reproduced from the original paper [14] with permission from *Springer*.

The contracting path is composed of a variable number of contracting blocks (4 in the original paper). Each of these contracting blocks is formed by two successive $3 \times 3$ convolutions with ReLU activation and a $2 \times 2$ max-pooling layer (with stride of 2). After passing into a contracting block, the image is shrunk by a factor of 2 in both spatial dimensions while the number of feature channels (depth of the image) is multiplied by 2. Intuitively, what the encoder does is to compress the data into a low spatial resolution but with high feature information; it trades spatial resolution for context information.

Similarly, the expanding path is composed of a variable number of expanding blocks (4 as well in the original paper). Each of these blocks is formed by two successive $3 \times 3$ convolutions with ReLU activation, a $2 \times 2$ up-convolution and concatenation with a cropped feature map from the contracting path. After passing into an expanding block, the spatial dimensions of an image are multiplied by 2, while the number of feature

channels is divided by 2. Intuitively, what the decoder does is to recover the original spatial dimensions of the image while keeping the context information of every pixel. The skip connections were the key addition to previous methods such as the F-CNN [13] and allowed more precise segmentations by incorporating raw information, which would otherwise be lost in the downsampling, in the expanding path.

A final $1 \times 1$ convolution is applied to get as output, an image with the same spatial dimensions as the input and a depth equal to the number of classes. When we apply a softmax function to the last layer of the network, one obtains the probability for each pixel that it belongs to a given class. The network is optimized by optimizing the cross-entropy loss via stochastic gradient descent. Data augmentation techniques used include random elastic deformations of the input data; this allows the network to perform well on test data even with few annotated training images.

In 2016, the U-Net was extended to a 3D version, replacing all 2D operations by their 3 counterparts [17]. This allowed greater context information and improved performance on 3D images, at the expense of more parameters to optimize in the model.

## 3.3 nnU-Net

The nnU-Net method was developed in 2017 at the German Cancer Research Center (DKFZ) by Isensee *et al.* [18]. Built on top of the U-Net architecture, it is "a deep learning based segmentation method that automatically configures itself including preprocessing, network architecture, training and post-processing for any new task in the biomedical domain" [18]. With a novel approach that is neither purely data-driven such as AutoML or NAS approaches [24, 38] and that neither requires expert knowledge for the design of a task-specific solution, the authors achieved state-of-the-art performance in 53 different segmentation tasks of the biomedical domain.

The method works as follows. The space of design choices is separated into three parameter groups: the fixed, rule-based and empirical parameters. The fixed parameters comprise the design choices that are not dataset-dependent according to the authors: the U-net like architecture template, the training schedule (optimizer, loss function, number of epochs, learning rate), the data augmentation techniques and the inference procedure. The rule-based parameters are configured on the fly based on the fingerprint of the dataset at hand; they are the dataset-dependent design choices. With well-designed heuristic rules, key properties of the network are configured in an almost-instantaneous manner. These include the network topology, the patch size, the batch size, the image resampling, the target spacing and the intensity normalization. Finally, the empirical parameters are configured empirically after training while evaluating validation performance. These encompass the model selection - between a 2D U-Net, a 3D U-Net, a 3D cascaded U-Net or an ensemble of them - and the postprocessing.

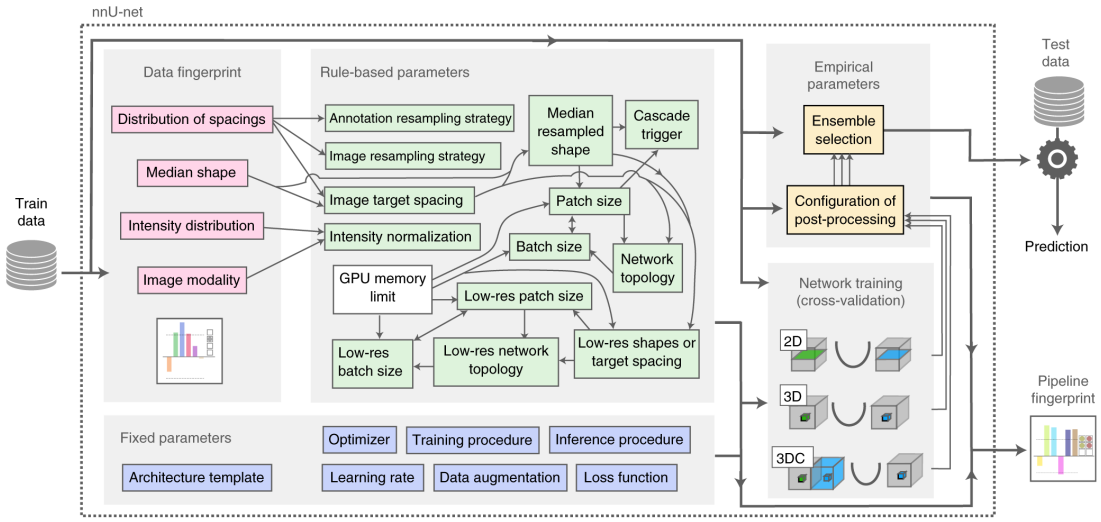



Figure 3.6: The **nnU-Net** : an automatic self-configuring method for out-of-the-box biomedical image segmentation. Reproduced from the original paper [18] with permission from *Springer*.

On the architectural level, we note that the nnU-Net uses the 2D and 3D U-Net architectures with a leaky ReLU replacing the ReLU after each convolution, instance normalization instead of batch normalization and strided convolutions instead of max-pooling for the downsampling layers. They also designed a 3D cascaded U-Net which gets trained when the patch size is too small compared to the full image size (less than 12.5%) to capture more context information; this was not necessary in our case, because our dataset has median size of $217 \times 256 \times 256$, 2 patches of size $128 \times 128 \times 128$ easily fitted in the GPU's memory.

## 3.4 Swin-UNETR and transformer networks

The Swin-UNETR method was developed in early 2022 by Hatazamidah *et al.* [26]. It is a hybrid architecture using building blocks from both Transformers models [28] and U-Net like methods [14].

### 3.4.1 Transformers in computer vision

Initially developed by Xawari *et al.* in 2017 [27], transformers quickly became the state-of-the art on natural language processing tasks. They were were then successfully applied in computer vision with the Vision Transformer (ViT) and have been popular in medical imaging ever since [28]. The most important building block of transformers is the multi-head self attention module which allows the network to capture long range interactions in the data [26]. The architecture of the ViT and its building blocks are shown in figure 3.7.

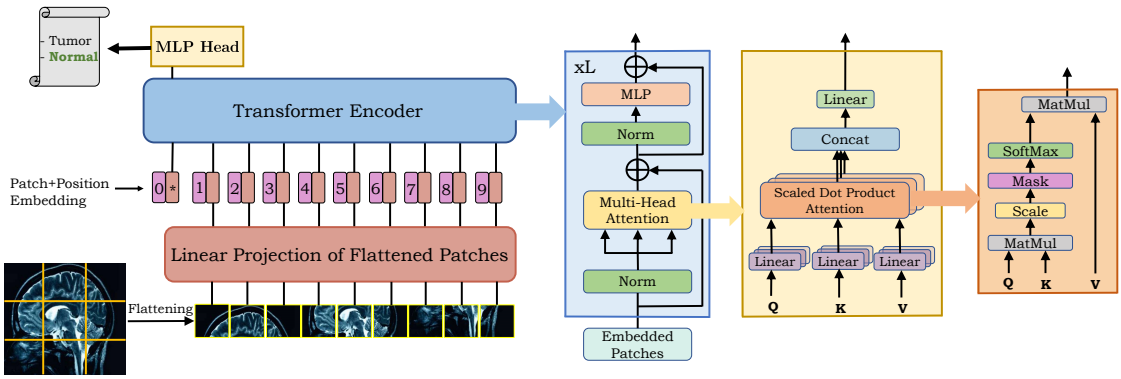



Figure 3.7: Architecture of the Vision Transformer and details of its core component: the multi-head self attention. The image is divided into patches which are fed into the transformer encoder. There, the relative importance of each patch with respect to one another is determined in the self attention module. Here a classification output is produced but one can also do segmentation if we reconstruct the original dimensions e.g. using up-convolutions. Reproduced from the original paper [29], under a Creative Commons Attribution-ShareAlike 4.0 International license.

The Swin Transformer, introduced by Liu *et al.*, is a variation of the ViT [28]. It proposes two key changes: introducing hierarchical feature maps instead of fixed scale patch tokens and replacing the self attention blocks by shifted window self attention. The Swin Transformer reduces the computational complexity of the ViT from quadratic to linear in the image size, allowing it to serve as an efficient backbone for dense prediction tasks such as image segmentation.

### 3.4.2 Swin UNETR

The Swin transformer is therefore used as the backbone encoder in the Swin UNETR[2]. The authors combine it with a fully convolutional network in the decoder using skip connections. The architecture is shown in figure 3.8.

[2] A previous method also proposed by Hatazamidah *et al.* that used the ViT in the encoder instead of the Swin Transformer is the UNETR. [39]

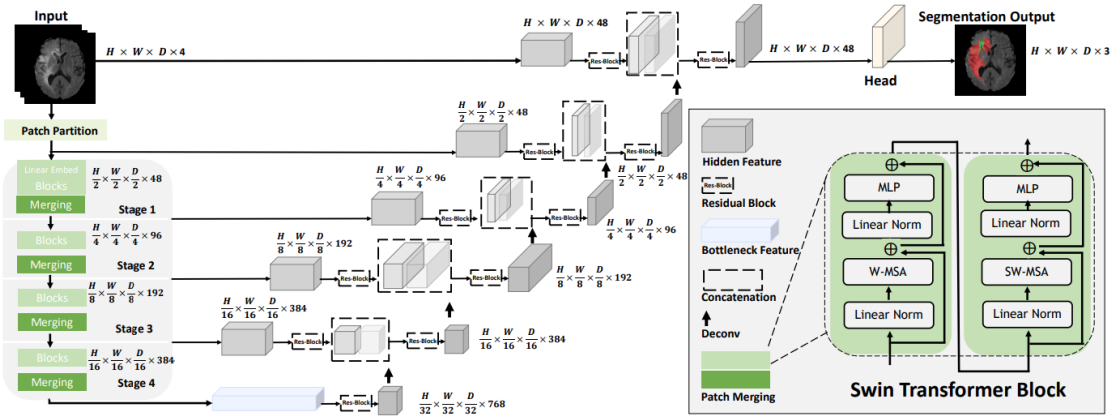



Figure 3.8: The **Swin-UNETR** Architecture [26]. It combines a Swin Transformer as the encoder connected to a fully convolutional decoder via skip connections. Reproduced from the original paper [26], under a Creative Commons Attribution Non-commercial Non-Derivative 4.0 International license.

Similarly to the U-Net, we get this encoder-decoder structure with skip-connections. Moreover, the image is first divided into patches and the encoder shrinks the spatial dimensions of the patch while increasing the feature information. By contrast with the U-Net, the patch is itself further divided into window tokens and the contraction is now done with Swin Transformer blocks instead of convolutions and pooling. The relative importance of patch windows is calculated in these blocks to model both close and long-range context information. The decoder part is almost identical to a U-Net decoder with concatenation with feature maps from the encoder path, up-convolutions and a fully convolutional head. The only difference is the use of residual blocks, each composed of two $3 \times 3 \times 3$ convolutional layers with instance normalization, instead of convolutions.

## 3.5 CycleGAN from T1 to T2

CycleGAN was developed in 2017 by Zhu *et al.* [40]. It addressed the problem of using a GAN for image-to-image translation with unpaired training data, an unsupervised approach. It differs from other approaches such as pix2pix which are supervised approaches to the problem [41]. A comparison between different GAN models for T1 and T2-weighted MR images translation has been done by Welander *et al.* [42]. They compared the performance of unsupervised methods such as UNIT [43] and CycleGAN with a supervised version of CycleGAN [40]. Their main takeaway is that all three models can synthesise realistic MR images (incorrectly labeled as real by an expert) with minor differences. Their other takeaway is that mean absolute error, average mutual information, and peak signal to noise ratio are insufficient indicators of the quality of a synthetic image - their model with the top scores produced overly smooth, unrealistic images that were easily labeled as synthetic by an expert.

We have used the CycleGAN implementation of Welander *et al.* in a supervised manner to generate T2 images from T1 MR images and vice-versa [42]. We trained it using 40

dHCP T1-T2 pairs. In order to train the network which is 2D, we sliced our 3D images into 2D axial slices. We cropped them from their original size to $256 \times 256$ as required by the input layer of the GAN. The sample size of the training set is 8087 image pairs (2D slices), and the size of the test set is 6007 images. After training, one has two generative models: a T1 to T2 translator and a T2 to T1 translator. One can use either of them to generate T2 slices (resp. T1 slices) from T1 slices (resp. T2 slices); one can then stack the generated slices to construct a 3D T2 image (resp. T1 image). Having been trained on dHCP data only, we expect it to generate realistic synthetic data only on the dHCP.

## 3.6 Experiments and evaluation metrics

We trained a vanilla U-Net, a 2D nnU-Net, a 3D nnU-Net and a swin-UNETR on our training set (training set = 58 dHCP, 80 Zürich pre-upgrade, 70 Zürich post-upgrade, 6 Giessen). The implementation details of our methods are shown in table 3.3.

Table 3.3: Overview of the methods selected for evaluation.

| Method | 2D/3D | Loss function | Patch size | Optimizer | Batch size | Epochs |
|---|---|---|---|---|---|---|
| U-Net | 3D | Dice | $256 \times 256$ | Adam | 10 | 200 w. early stopping |
| 2D nnU-Net | 2D | Dice and CE | $320 \times 256$ | Nesterov | 38 | 1000 |
| 3D nnU-Net | 3D | Dice and CE | $128 \times 128 \times 128$ | Nesterov | 2 | 1000 |
| swin-UNETR | 3D | Dice and CE | $96 \times 96 \times 96$ | AdamW | 2 | 1500 |

| Method | Learning Rate | # of Parameters | Data augmentation | Post-processing |
|---|---|---|---|---|
| U-Net | 0.00001 w. decay | 8M | Elastic | None |
| 2D nnU-Net | 0.01 w. decay | 30M | Rotations, Scaling, Gaussian noise and blur, brightness, contrast, mirroring | Skipped |
| 3D nnU-Net | 0.01 w. decay | 31M | ————-”————- | Skipped |
| swin-UNETR | 0.0008 | 63M | Intensity shift and scaling, flipping | Skull stripping |

For evaluation, we make predictions of each model without performing ensembling. Cross-validation is natively implemented in the nnU-Net to determine both the post-processing procedure and to perform ensembling at inference time. We decide to skip this step as it would have yielded an unfair comparison to the other models, where ensembling was not implemented. As we observed on other data, this should only reduce the nnU-Net's performance by less than a percent, while reducing the training (and inference time) by a factor of 5.

The evaluation metrics that we will use throughout our evaluation of the four selected methods are, the Sørensen-Dice coefficient (also known as *Dice score* or *Dice*) and the Hausdorff distance 95 (HD95). The Dice score measures the overlap between two sets, here the ground truth segmentation ($Y$) and the predicted segmentation ($\hat{Y}$).

$$\text{Dice score} = \frac{2|\hat{Y} \cap Y|}{|\hat{Y}| + |Y|} = \frac{2\sum_{i=1}^{N} \hat{y}_i y_i}{\sum_{i=1}^{N} \hat{y}_i + \sum_{i=1}^{N} y_i} \tag{3.1}$$

where $y_i$, $\hat{y}_i$ denote the individual voxels of the ground truth and predicted segmentation, seen as binary vectors of length N.

The Hausdorff Distance calculates the distance between two subsets, here $Y$ and $\hat{Y}$.

$$\text{Hausdorff Distance}(Y, \hat{Y}) = \max\{\sup_{\hat{y}} \inf_{y} d(y, \hat{y}), \sup_{y} \inf_{\hat{y}} d(y, \hat{y})\} \tag{3.2}$$

where *sup* is the supremum (least upper bound), $inf$ the infimum (greatest lower bound), $d(y, \hat{y})$ denotes the distance between two points $y$ and $\hat{y}$ of the ground truth set $Y$ and predicted set $\hat{Y}$ respectively. The HD95, which calculates the $95^{\text{th}}$ percentile of the Hausdorff Distance is usually used as it is more robust against outliers.

Moreover, we define the *Dice performance loss* as the relative error in average Dice score on a given test domain between a reference algorithm which has been trained on all the training data and a test algorithm which has not been exposed to this particular test domain at training time.

$$\text{Dice performance loss (\%)} = \frac{\overline{\text{Dice}}(\text{reference algo}) - \overline{\text{Dice}}(\text{test algo})}{\overline{\text{Dice}}(\text{reference algo})} \tag{3.3}$$

where the bar denotes an average Dice over all samples of the test domain. We expect the Dice performance loss to always be positive, with smaller values indicating stronger generalizability of a method. A negative value would indicate that training an algorithm with a given domain may deteriorate its performance on this same domain at test time; this would be questionable.

The following experiments have been designed to evaluate and compare the methods. First, we will compare their performance on our neonatal data in a traditional train/test split with all centers (dHCP, Giessen, Zürich) both in the training and in the testing data.

We then study how our methods perform in three different domain adaptation scenarios: predicting on an unseen (at training) center, an unseen structure (fetal) and on synthetic data. By iteratively removing a center from the training set, training a new model and evaluating on that particular center, we will obtain the Dice performance loss for each of our centers. This will be averaged and constitute our center adaptation metric. We will then move on and investigate how our methods extend their predictive power to fetal data, a new domain of test that was absent from training. Finally, we will evaluate the performance of our methods when exposed to synthetic T2 images, generated using our trained CycleGAN by translating T1w images from the dHCP dataset.

We will then perform further evaluations on the best performing model - the 3D nnU-Net - in more detail. We perform a per-label evaluation. We split our testing data in GA ranges of 2.5 weeks to see if our network's performance is impacted by age. We expect this to be relevant because of the fast-developing brain in both fetuses and neonates. Because different brain tissues develop at a different pace, we also look at differences in the evaluation metrics across labels within each age range.

# 4 Results

## 4.1 Evaluation of our selected algorithms in a regular setting

We evaluate our different methods when training on all the training neonatal data (Zürich, dHCP, Giessen) and evaluating on all the test neonatal data (Zürich, dHCP, Giessen), excluding fetal data from the testing set. This is an evaluation in a "regular" setting, meaning that each method is exposed to all test domains during training.

The Dice and HD95 is calculated, for each label, between the segmentations and the ground truth. For each sample, we average the metric over all labels to get a score per sample. A violin plot of the resulting metric distribution over all test samples is shown in figure 4.1 for the Dice and in figure 4.2 for the HD95.

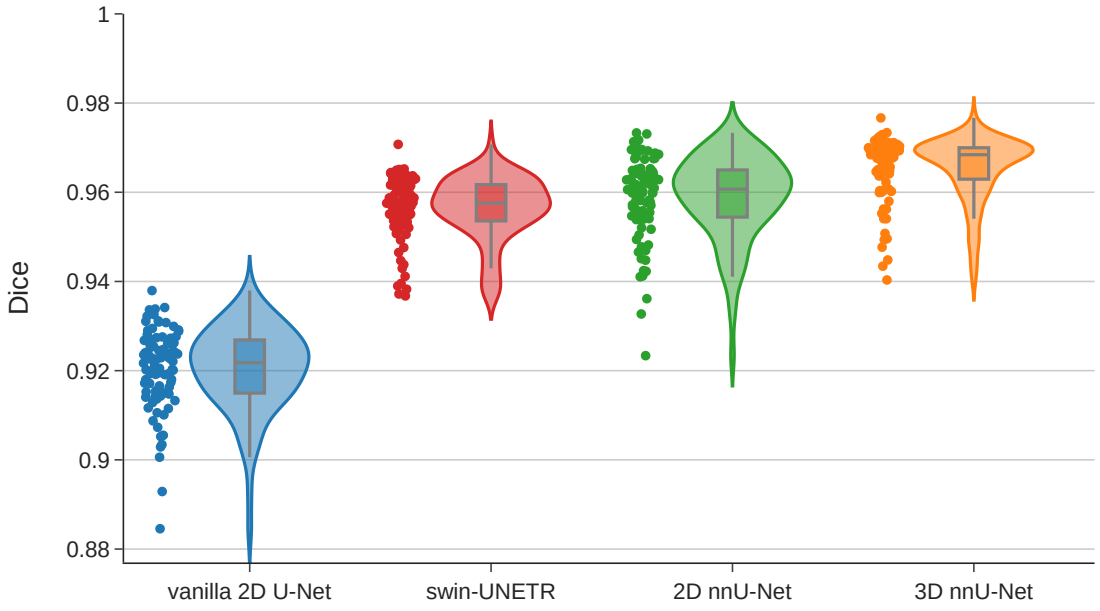



Figure 4.1: Evaluation of our selected methods on all neonatal testing data. Violin plot of Dice on all testing images of our selected methods. Median (over all testing samples) Dice per model - vanilla 2D U-Net: 0.92, swin-UNETR: 0.95, 2D nnU-Net: 0.96, 3D nnU-Net: 0.97.

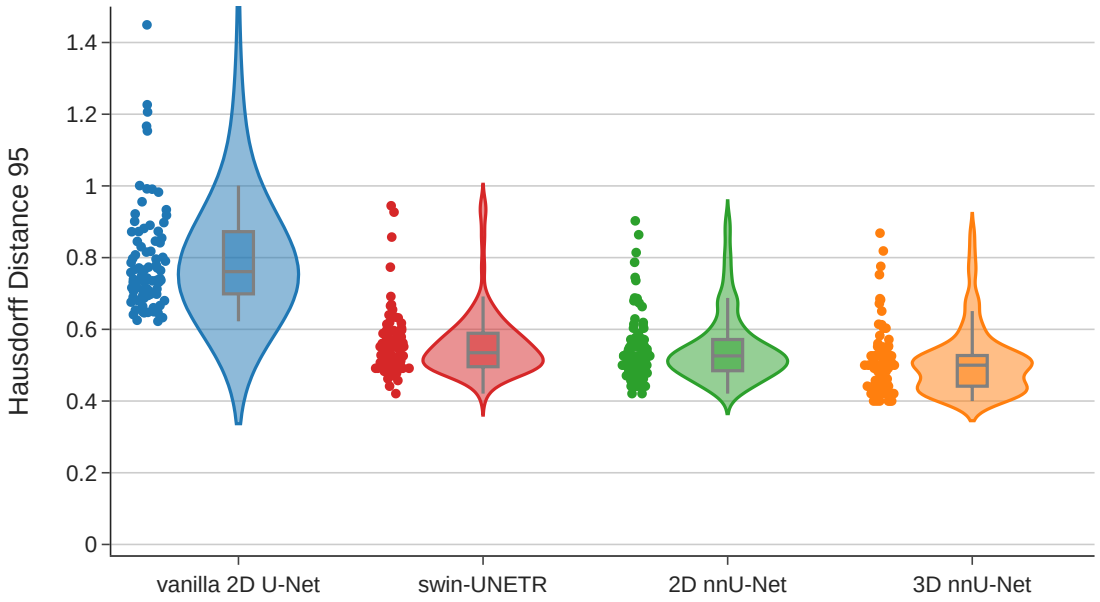



Figure 4.2: Violin plot of HD95 on all testing images of our selected methods. Median (over all testing samples) HD95 per model - vanilla 2D U-Net: 0.76, swin-UNETR: 0.54, 2D nnU-Net: 0.53, 3D nnU-Net: 0.5.

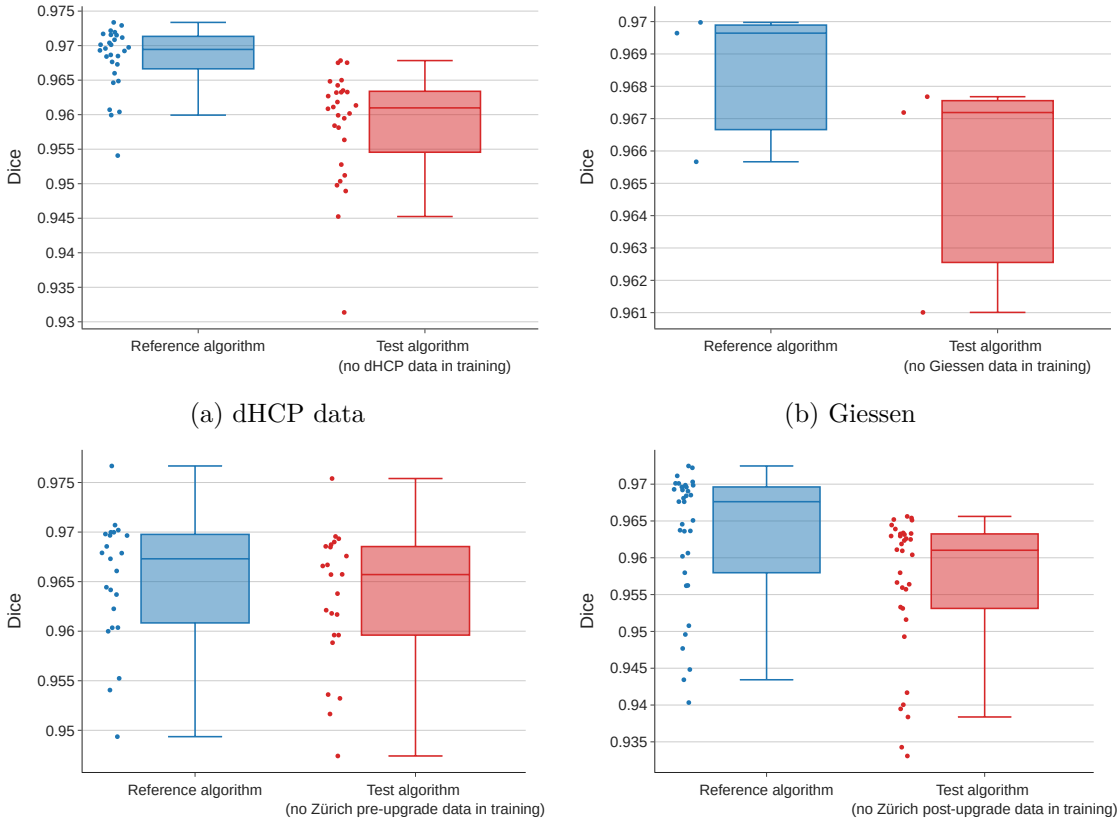
The top-performing model in a "regular" setting on multi-centric neonatal data is the 3D nnU-Net with a median Dice of 0.97 and a median HD95 of 0.5mm. Comes next the 2D nnU-Net (Dice: 0.96, HD95: 0.53), the swin-UNETR (Dice: 0.95, HD95: 0.54) and a 2D vanilla U-Net (Dice: 0.92, HD95: 0.76), method previously used for segmentation in the MR research group.

The 3D nnU-Net yields a 5.1% improvement (in Dice) over the previously used method in addition to being faster at inference time. We can see that all our selected methods perform very well on our data in a traditional train/test split with median Dice above 0.95 and HD95 less than 0.54mm. We want to emphasize that such a low HD95 represents a 95th percentile error of about a voxel (0.4-0.5mm) on average, for each label segmentation. These results are promising for the ability of our methods to generalize to unseen domains.

## 4.2 Domain adaptation capability of our selected algorithms

### 4.2.1 Adaptation to an unseen center

We now evaluate our methods in a domain-adaptation setting, when segmenting data from a domain unseen during training. As described in section 3.6, we iteratively remove a domain from the training set, train a new model and evaluate on this particular domain. For each of our methods, we obtain the Dice performance loss on each of our domains. The process is shown in figure 4.3 for the 3D nnU-Net. Our multi-centric domain adaptation results for all our methods are compiled in table 4.1.




(a) dHCP data

(b) Giessen

(c) Zürich pre-upgrade

(d) Zürich post-upgrade

Figure 4.3: Generalization performance of the 3D nnU-Net on the different centers, unseen at training time. The Dice performance loss is respectively 0.87% for dHCP, 0.25% for Giessen, 0.17% for Zürich pre-upgrade and 0.68% for Zürich post-upgrade.

| Dice performance loss(%) | dHCP | Giessen | Zürich pre-upgrade | Zürich post-upgrade | **Avg.**[1] |
|---|---|---|---|---|---|
| swin-UNETR | 2.16 | 0.82 | 0.37 | 1.21 | 1.26 |
| 2D nnU-Net | 1.57 | 0.66 | 0.31 | 1.18 | 0.99 |
| 3D nnU-Net | 0.98 | 0.32 | 0.20 | 0.69 | **0.58** |

Table 4.1: Dice performance loss of our implemented methods on each of our domains.

We obtain that the 3D nnU-Net is the most generalizable method with an average Dice performance loss of 0.58% followed by the 2D nnU-Net and the swin-UNETR. We would like to see how the 3D nnU-Net performs when exposed to a new structure at test time. We therefore extend our domain adaptation setting to also include fetal data.

[1]To compensate the fact that Zürich pre-upgrade and post-upgrade are not two distinct centers, the average is weighted with $\lambda = 0.5$ for both Zürich domains and $\lambda = 1$ for the remaining two.

### 4.2.2 Adaptation to an unseen structure: fetal data

We now have a look on how our 3D nnU-Net (the "test" algorithm), trained on neonatal data, performs on fetal data. This is more than center adaptation but also structural adaptation, the algorithm is exposed to fetal data at testing time while it has been trained on neonatal data only. We compare its performance to a reference algorithm, a 3D nnU-Net trained on the FeTA Challenge training set and evaluated on the test set cases [4]. To avoid too big a domain shift for the "test" algorithm, we evaluate both of our algorithms on fetuses with gestational age greater than 28 weeks and neurotypical only.

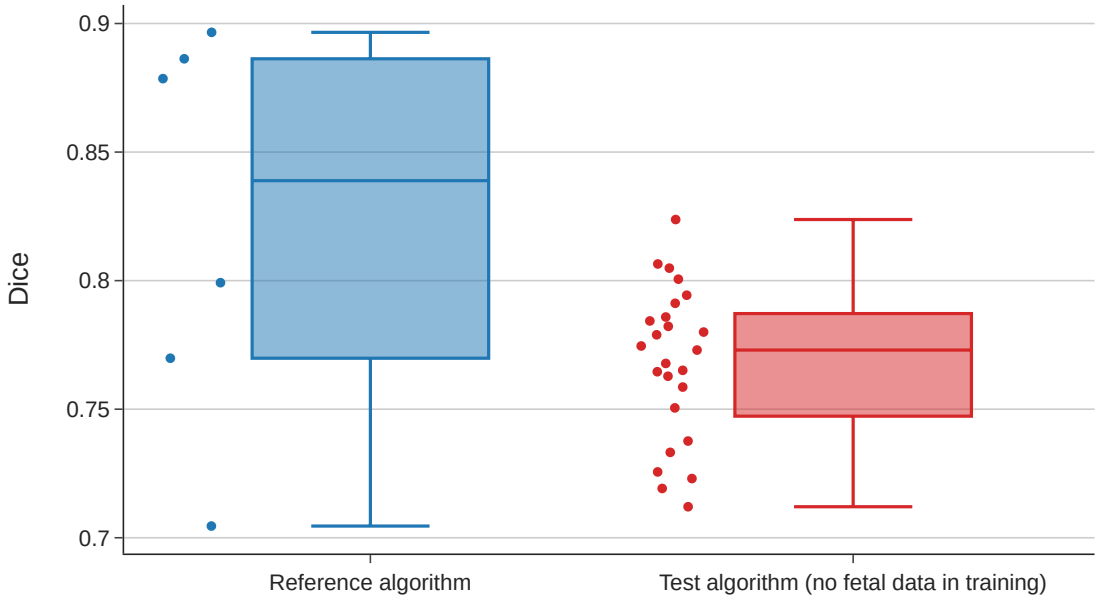



Figure 4.4: Generalization capacity of the 3D nnU-Net to a new structure: fetal data. The median Dice scores are respectively 0.84 for the reference algorithm and 0.77 for the "test" algorithm. The Dice performance loss is 6.6%. We note that our reference algorithm is evaluated on 6 data samples, and our "test" algorithm on 25. This discrepancy comes from the fact that the reference algorithm had some of these cases (19) in its training set so we removed them from evaluation.

The 3D nnU-Net is the best performing method at generalizing to the fetal domain with a Dice performance loss of 6.6%. Next comes the 2D nnU-Net with a Dice performance loss of 12.3% and the swin-UNETR with 19.4%.

We want to emphasize that our reference algorithm, a 3D nnU-Net trained on the training set of the FeTA Challenge 2021, is on par with the top performing method of the challenge (the official winner SegResNet achieves 0.786 of average Dice in the FeTA ranking, we get 0.773) [4]. Hence, the low Dice performance loss of the nnU-Net (6.6%) without any exposure to fetal data during training is even more impressive.

### 4.2.3 Adaptation to synthetic T2-weighted MRI data generated using CycleGAN

The idea of generating T2-weighted MRI (referred to T2) images from T1-weighted MRI (referred to T1) emerged when we tested our T2 trained methods on T1 data: a completely different domain. As anticipated, the results were not optimal (Median Dice of 0.1 over 31 T1 images, see figure 4.7) and this is understandable given that T1 images have a completely different tissue contrast. While water containing tissues are bright in T2 images, fat containing tissues are bright in T1 images [44]. The CSF for example is the brightest on T2 images but dark on T1 images as it is mainly composed of water. The ventricles, where the CSF is produced and circulates, are also bright on T2 and dark on T1. Grey matter, on the other hand, is brighter on T1 images and dark on T2 images. A side-by-side illustration of the differences is shown in Figure 4.5.

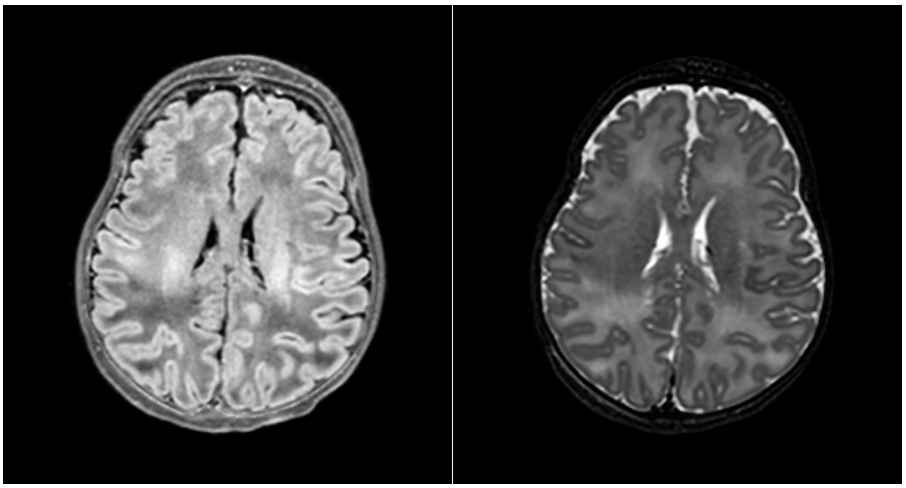


Figure 4.5: T1 vs T2 slice from the dHCP neonatal dataset; CSF and ventricles are bright in the T2, dark in the T1. Grey matter is bright in the T1, dark in the T2. White matter is largely non-myelinated in newborns, therefore the WM remains darker than the GM.

A way to overcome this problem, e.g. if T1 is not available in the training dataset, is to synthesize T2 images out of T1 using generative adversarial networks (GANs). Using our trained CycleGAN implementation, as described in Section 3, we generated T2 images from T1 on the dHCP data. An example synthetic axial slice is shown in figure 4.6 along the T1w image used for generation, and the real T2w image. The structure of the tissues is well preserved after translation; one can easily recognize the ventricles, cerebellum and brainstem in the synthetic image. The grey matter/white matter boundary is also well recognizable after synthesis, although it gets slightly blurred in the translation process. We can notice that blood vessels (black dots in the CSF and ventricles) get lost in the T1 to T2 translation; these are however irrelevant in our segmentation work. More examples of T1 to T2 generation (and vice-versa) on the dHCP data are shown in Appendix B.

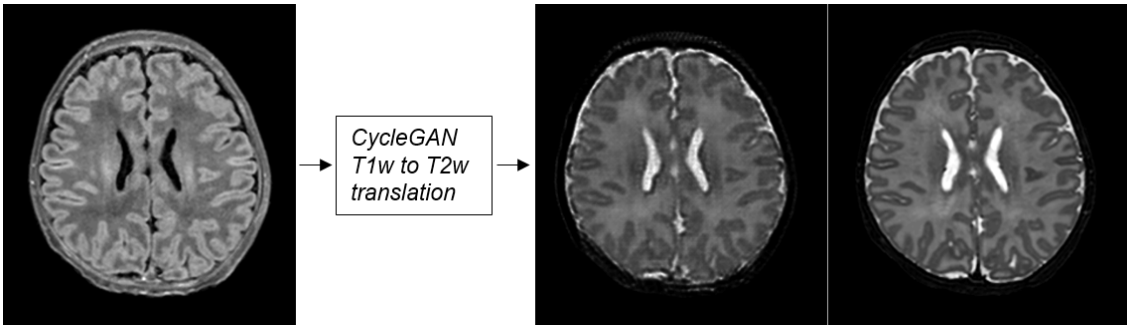



Figure 4.6: Translating T1w to T2w with CycleGAN. T1 (first image) vs. synthetic T2 (second image) vs. real T2 (third image) after passing the T1 image through CycleGAN. See appendix B for more synthetic vs. ground truth comparison.

3D T1 and T2 images are not always available - some study centers may decide to acquire only one of it, or only one of it is available in good quality. In this case, one can still use our segmentation methods - nnU-Net, swin-UNETR, ... - only trained on one type of data (in our case, T2) by first generating a synthetic T2 image from the T1. We evaluated our segmentation methods on 30 T1 images, the corresponding 30 synthetic T2 images after translation and the 30 real T2 images. The Dice score distribution for the 3D nnU-Net is shown in figure 4.7.

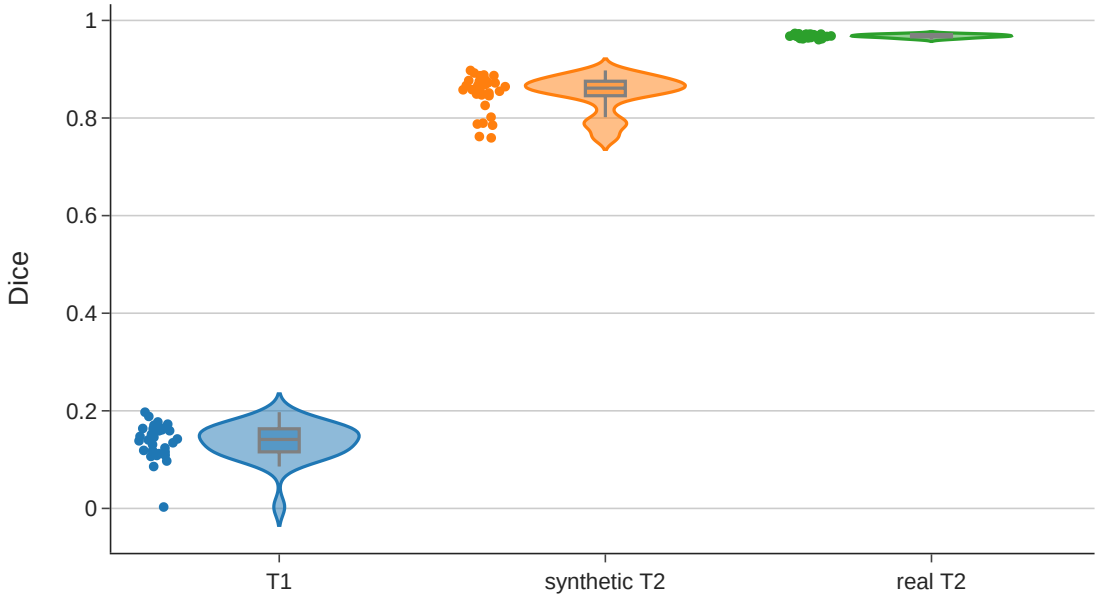



Figure 4.7: Generalizability of the 3D nnU-Net to synthetic MRI data. Median Dice scores are respectively 0.14 on T1 images, 0.86 on synthesised T2 images and 0.97 on real T2 images.

With a Dice performance loss of only 11.0% when segmenting synthetic versus real T2

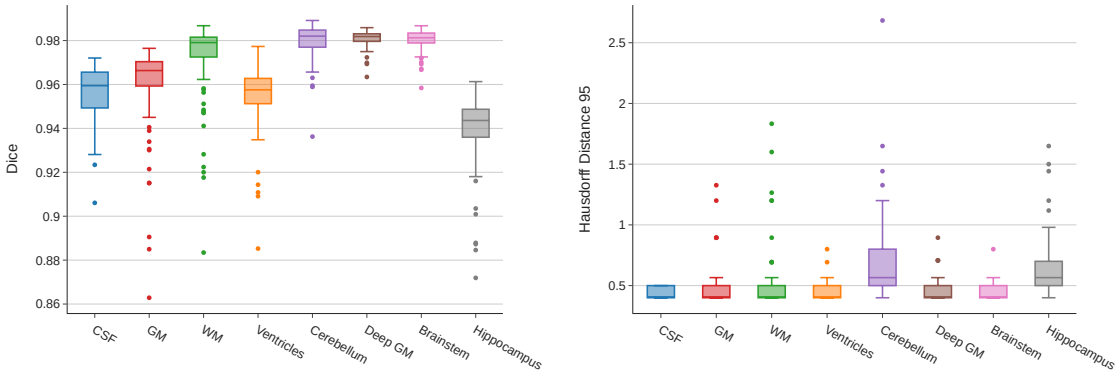
data, the generalization capacity of the 3D nnU-Net is remarkable. This confirms the fact that it is the most robust of our tested methods, better than the swin-UNETR (12.8% dice performance loss on synthetic data) and the 2D nnU-Net (13.7% dice performance loss on synthetic data). Some additional usages of the trained T1/T2 GAN network will be discussed in section 5.

## 4.3 Detailed evaluation of the 3D nnU-Net on different sub-classes of data

In this section, we have a detailed look into our best performing method (the 3D nnU-Net) trained on all our training neonate data (back to the setting of section 4.1). We evaluate its performance on the testing set on different sub-classes of data: per label and per age range. Because there is fetal data in our testing set, with substantially worse results than neonatal data, we perform two separate evaluations: on neonatal data only and on fetal data only.

### 4.3.1 Detailed evaluation on neonatal data

We start by evaluating our predictions on neonatal data for each label. The Dice score and HD95 distributions of results per label over all neonate samples are shown in figure 4.8.




(a) Median Dice - CSF: 0.96, GM: 0.97, WM: 0.98, Ventricles: 0.96, Cerebellum: 0.98, Deep GM: 0.98, Brainstem: 0.98, Hippocampus: 0.94.

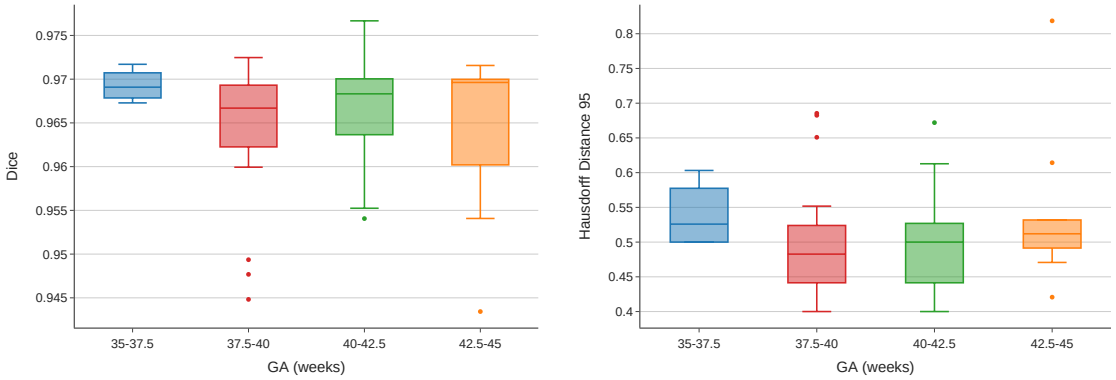
(b) Median HD95 - CSF: 0.40, GM: 0.40, WM: 0.40, Ventricles: 0.40, Cerebellum: 0.57, Deep GM: 0.40, Brainstem: 0.40, Hippocampus: 0.57.

Figure 4.8: Performance of the 3D nnU-Net on all our neonatal data for each label.

The nnU-Net performs best in segmenting white matter, deep grey matter and brainstem with median Dice score around 0.98. The cerebrospinal fluid, grey matter and ventricles have Dice scores around 0.96. The worst performing label is the hippocampus with a median Dice of 0.94.

Most labels have the least possible median HD95 of 0.4 which corresponds to an error of one voxel. The worst labels are the Cerebellum and Hippocampus with a median HD95 of 0.57 over all samples. This corresponds to a distance of one diagonal voxel in an image with voxel spacing 0.4mm ($\sqrt{0.4^2 + 0.4^2} = 0.57$). All the other labels have median HD95 of 0.4, the least possible error. These results show that the 3D nnU-Net learnt to label all tissues with great accuracy; the HD95 is so small that it would be an impossible task for a human to differentiate between the segmentation and the ground truth in most cases.

We now split the results by age ranges (with ranges of length 2.5 weeks) and average over labels. We make an age cutoff at less than 35 GA weeks and more than 45 GA weeks as there are not enough samples there to get meaningful statistics. The Dice score and HD95 distributions of results per age range over all neonate samples are shown in figure 4.9.




(a) Median Dice - GA 35-37.5: 0.97, GA 37.5-40: 0.97, GA 40-42.5: 0.97, GA42.5-45: 0.97.

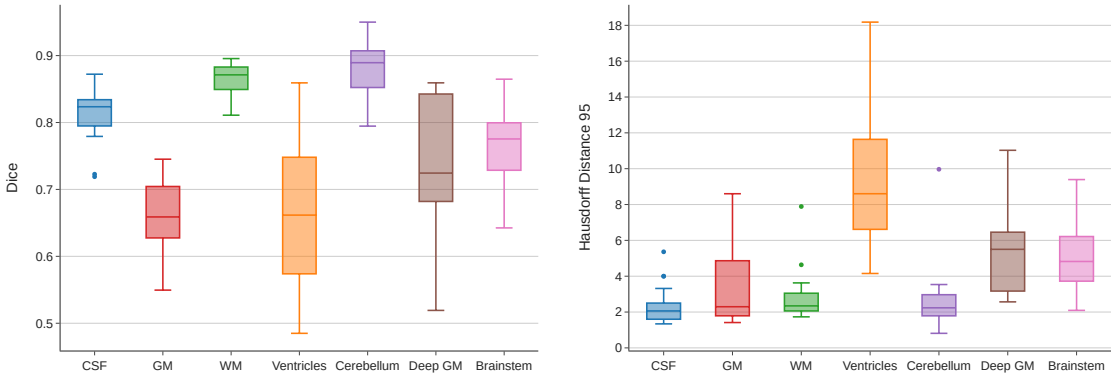
(b) Median HD95 - GA 35-37.5: 0.5, GA 37.5-40: 0.4, GA 40-42.5: 0.4, GA 42.5-45: 0.5.

Figure 4.9: Performance of the 3D nnU-Net on all our neonatal data for each GA range.

There are no obvious differences in segmentation results observed for 2.5 weeks age range with similar Dices and HD95 across GA ranges 35-37.5, 37.5-40, 40-42.5 and 42.5-45 weeks. We therefore stop our detailed evaluation on neonatal data and move on to fetal data analysis.

### 4.3.2 Detailed evaluation on fetal data

We start by evaluating our predictions on fetal data for each label. The Dice score and HD95 distributions of results per label over all fetal samples are shown in figure 4.10.




(a) Median Dice - CSF: 0.82, GM: 0.66, WM: 0.88, Ventricles: 0.66, Cerebellum: 0.89, Deep GM: 0.72, Brainstem: 0.78.

(b) Median HD95 - CSF: 2.1, GM: 2.3, WM: 2.3 Ventricles: 8.6, Cerebellum: 2.2, Deep GM: 5.5, Brainstem: 4.8.

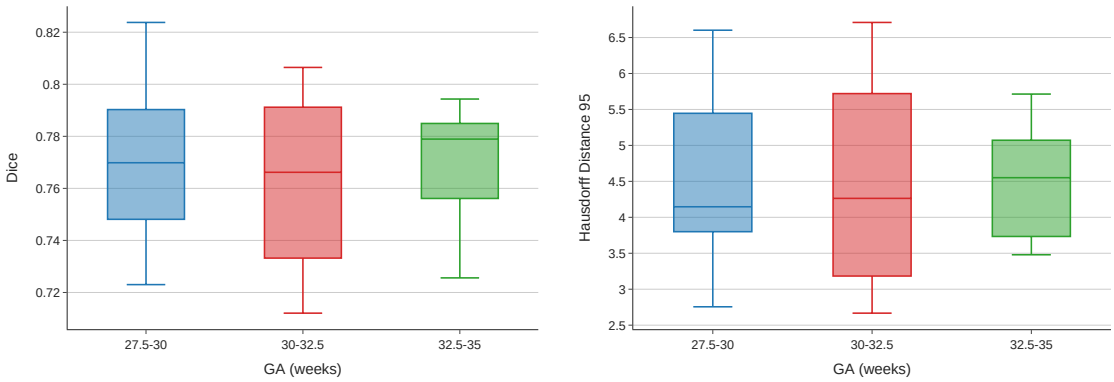
Figure 4.10: Performance of the 3D nnU-Net (trained on neonates) on the fetal data for each label.

The best labels according to the Dice score are the cerebellum (Dice: 0.9) and white matter (Dice: 0.87). Next comes the CSF (Dice: 0.82) and brainstem (Dice: 0.76). Deep grey matter has a Dice of 0.72 but is has a high standard deviation in the Dice distribution with a large IQR of 0.16. Grey matter and ventricles are the worst performing labels, both with median Dice 0.64, and a large spread across cases for the ventricles label (IQR = 0.18).

The HD95 box plots reinforce that the ventricles is the least well segmented tissue on the fetal data. Surprisingly, grey matter has a low HD95 score, on par with top performing labels (cerebellum, CSF and white matter) but given the shape of this structure with lots of fine details, the Dice score may be a better metric than Hausdorff Distance in this case.

It is interesting to put these results in perspective with what we observed in the intensity distributions of the fetal tissues (figure 3.2). Indeed, we saw that the ventricles appeared much darker in fetal data than in neonatal data. It was the only tissue to exhibit such a significant discrepancy and this explains the low segmentation performance of our method, trained only on neonatal data. The grey matter also was shown to have a large intensity IQR across cases when compared to neonatal data, this may explain its low performance. Additionally, we note that the CSF and ventricle spaces undergo significant changes from fetal to neonatal brain. In the fetal brain, the lateral ventricles are more dilated than in the neonatal brain, which might cause the low segmentation performance of a model that was trained on infant datasets only.

We now split our segmentation results by age range (averaging over all labels) and evaluate accordingly. The results are shown in figure 4.11.



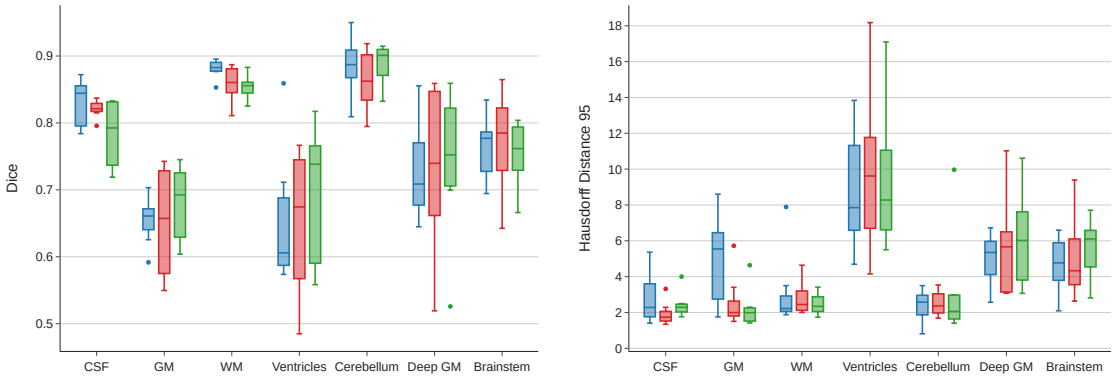

(a) Median Dice - GA 27.5-30: 0.78, GA 30-32.5: 0.78, GA 32.5-35: 0.77.

(b) Median HD95 - GA 27.5-30: 4.2, GA 30-32.5: 3, GA 32.5-35: 2.8.

Figure 4.11: Performance of the 3D nnU-Net (trained on neonates) on the fetal data for each GA range.

On one hand, the Dice results do not show any significant difference in performances across the selected age ranges. On the other hand, the HD95 box plots seems to indicate that the GA 27.5-30 is worse compared to the two older age ranges. We can interpret this as follows; although the number of errors (reflected by the Dice) is approximately the same, the magnitude of the errors is larger in the younger fetuses than in the older ones (reflected by the HD95).

We now split our results per age range within each label. The results are shown in figure 4.12.




(a) Dice score.

(b) HD95.

Figure 4.12: Performance of the 3D nnU-Net (trained on neonates) on the fetal data for each label and GA range. The colors blue, red and green correspond respectively to gestational weeks 27.5-30, 30-32.5 and 32.5-35.

Looking at performances across age range on each label individually, we observe that for the ventricles label, age seems to play a role on the Dice score. The older the fetus,

the better the ventricles segmentation is. The Dice goes from a low median score of 0.6 for the 27.5-30 age range to a score of 0.74 for the 32.5-35 age range. This result is to take with caution because of the 7 samples in the 32.5-35 age range but intuitively would make sense given that the younger age range (27.5-30) is completely absent from the neonatal training data while the older age range (32.5-35) still has some samples in the neonatal training data (moderate pre-term). Similarly, the grey matter segmentation is substantially worse (as evaluated by HD95) in younger fetuses (27.5-30). This can be explained by the different appearance of this tissue in younger fetuses because of the cortical maturation not being fully finished [4].

# 5 Discussion

In our study, we evaluated various deep learning based methods for infant and fetal brain segmentation. The evaluations were done in various settings: first by testing them on a complete multi-centric dataset across developmental stages, and then by testing them in various domain adaptation experiments.

Our results show that the 3D nnU-Net yields a substantial improvement - 5.1% in Dice - over a Vanilla 2D U-Net used in a recent publication for fetal brain MRI segmentation. It is also substantially faster as the pre-processing takes a few seconds per sample 3D image compared to a few minutes for the Vanilla 2D U-Net. The reason for this is that the Vanilla U-Net requires resampling, cropping or padding to a fixed size and registration to an atlas for improved performance. On the other hand, registering the image and resizing is not required by the 3D nnU-Net which processes the data in patches, independent of total image size. This greatly accelerates the pre-processing and thus the inference time. It is also easier to use out-of-the box by a non-expert as all the pre-processing steps are fully integrated within the segmentation pipeline.

The ability of the 3D nnU-Net to generalize to new domains in the form of new centers but also new structures and synthetic data unseen at training time has been proven. It exhibits a loss in performance of only 6.6% (as evaluated by Dice score) when segmenting fetal data (with only exposure to neonatal data at training) compared to the state-of-the-art on the FeTA challenge 2021 [4]. This shows that having one trained method for both neonate and fetal MRI segmentation could be feasible in the near future. Likewise, the loss in performance of merely 11.0% when segmenting synthetic T2 data (versus real T2 data) also shows the robustness of our implementation towards the artifacts and blurriness introduced by using the GAN. This also indicates that using synthetic data in the training could be a reliable way of training a method when lacking input data.

Contrary to what has been observed in the live MSD challenge and in the BraTS 2022 [25, 26], swin-UNETR is not the top-performing method on our data. We postulate that, given more time to optimize task-specific design choices (pre-processing, data augmentation used, hyper-parameter tuning), swin-UNETR could achieve a better performance. We have evaluated different fine-tuning approaches by changing the number of epochs, batch size, applying more data augmentation (random flipping, random intensity shift and scaling) and using pre-trained weights from the winning BraTS model [26]; it still underperformed the nnU-Net in all cases. On the other hand, the 3D nnU-Net performed the best out-of-the-box, without the need for us to apply any fine-tuning on the original method design. We attribute this to the well-designed heuristics of nnU-Net for setting the hyperparameters, thus limiting the number of design decisions impacting the final segmentation's performance and robustness.

Our implementation of CycleGAN for T1w to T2w (and vice-versa) translation has shown to be useful in the context of segmenting data when real T2 (or T1) images are not

available. We would like to emphasize that it has much more potential relevance in other applications, for example in applying MRI post-processing pipelines such as FreeSurfer that only work on one image modality. This is a common case encountered in research studies at the University Children's Hospital and it could be solved by first applying a modality translation with our GAN implementation. The GAN implementation still needs to be tested on other domains than the dHCP, its training domain, to see if it generalises well.

One limitation of our approach is the low number of different medical centers - three - used in evaluating our methods. The Dice performance loss in the order of 1-2% on each of our center also indicates that our centers have a reasonably low domain shift. We have tried to overcome this practical limitation by including fetal data and synthetic T2 data to challenge our methods with more domains at test time.

Moreover, the distinction we have made between Zürich's data pre and post scanner upgrade as two separate domains would need more investigation. Indeed, if we look at figure 3.2, one can see that the intensity distribution between the two domains is very close for a few labels. The null hypothesis that these two domains have the same mean intensity (after z-score normalization) for some labels cannot be rejected for the cerebellum ($p=0.97$ after a paired t-test between the two distributions), the grey matter ($p=0.09$) and the deep grey matter ($p=0.41$). This may have biased our generalization metric on Zürich's center overall.

Another limitation of our approach is that the ground truth segmentation used for training and evaluating our methods is automatically generated via the draw-EM algorithm (except the FeTA dataset which is manually annotated) [35]. Although it was coupled with manually annotated atlases in the dHCP data, and manually checked in the case of Zürich data, it is still an imperfect ground truth. As we are trying to replace this method for a faster, more robust segmentation technique; one can raise the question of the validity of using such a ground truth for both training and evaluation. This limitation is mitigated by the fact that one is usually interested in bulk metrics, such as the shape or volume of the segmented tissues, which are robust to errors of a few voxels in the segmentation.

# 6 Conclusion

Our study indicates that the 3D nnU-Net is the best automatic segmentation method on our multi-centric infant MRI data in terms of performance, generalizability and usability. The average Dice performance loss of 0.58% when evaluated on an unseen center, 6.6% when evaluated on an unseen structure (fetal data) and 11.0% when evaluated on synthetic T2 data demonstrate the domain adaptation capacity of the nnU-Net. This performance was achieved using the nnU-Net directly out-of-the-box, without the need for manual fine-tuning, hence making the results even more reliable.

While the Swin-UNETR shows good performances, its need for a task-specific manual optimization makes it a less usable technique than the nnU-Net. With the rise of transformers-based architectures in computer vision, one can expect more and more of them to be developed for medical image segmentation over the coming years. With the technical advances of hardware technology, one can also expect more and more AutoML methods such as Neural Architecture Search to remove the need for manual fine-tuning. By monitoring the Medical Segmentation Decathlon live leaderboard, one can then stay up-to-date with the most accurate and robust segmentation methods of this rapidly evolving field.

The future of automatic segmentation and its practical adoption in the clinic as a tool for diagnosis, assesment and monitoring of diseases relies on robust methods that can handle atypical cases, multi-centric data. Ideally, one would like to have one method that can segment brain MRI of both fetal and neonatal data across gestational ages and clinics, particularly in longitudinal studies. Our study which evaluated state-of-the-art methods on different domains in fetal and infant MR images is thus one crucial step towards this goal.

# Bibliography


[1] A. Gholipour, J. A. Estroff, C. E. Barnewolt, R. L. Robertson, *et al.*, "Fetal mri: A technical update with educational aspirations," *Concepts in Magnetic Resonance Part A*, vol. 43, no. 6, pp. 237–266, 2014. DOI: https://doi.org/10.1002/cmr.a.21321.

[2] M. Kuklisova-Murgasova, G. Quaghebeur, M. A. Rutherford, J. V. Hajnal, and J. A. Schnabel, "Reconstruction of fetal brain mri with intensity matching and complete outlier removal," *Medical Image Analysis*, vol. 16, no. 8, pp. 1550–1564, 2012. DOI: https://doi.org/10.1016/j.media.2012.07.004.

[3] M. Ebner, G. Wang, W. Li, M. Aertsen, *et al.*, "An automated framework for localization, segmentation and super-resolution reconstruction of fetal brain mri," *NeuroImage*, vol. 206, p. 116 324, 2020. DOI: https://doi.org/10.1016/j.neuroimage.2019.116324.

[4] K. Payette, H. Li, P. de Dumast, R. Licandro, *et al.*, *Fetal brain tissue annotation and segmentation challenge results*, Manuscript under review, 2022. DOI: 10.48550/ARXIV.2204.09573.

[5] K. Payette, P. de Dumast, H. Kebiri, I. Ezhov, *et al.*, "An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset," *Scientific Data*, vol. 8, no. 1, p. 167, Jul. 2021. DOI: 10.1038/s41597-021-00946-3.

[6] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, 2022. DOI: 10.1109/TBME.2021.3117407.

[7] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. Mit Press, 2009.

[8] N. R. Pal and S. K. Pal, "A review on image segmentation techniques," *Pattern Recognition*, vol. 26, no. 9, pp. 1277–1294, 1993. DOI: https://doi.org/10.1016/0031-3203(93)90135-J.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25, Curran Associates, Inc., 2012. DOI: https://doi.org/10.1145/3065386.

[10] K. Fukushima and S. Miyake, "Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position," *Pattern recognition*, vol. 15, no. 6, pp. 455–469, 1982.

[11] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. DOI: 10.1109/5.726791.

[12] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020. DOI: 10.1007/s10462-020-09825-6.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440. DOI: 10.1109/CVPR.2015.7298965.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241. DOI: https://doi.org/10.1007/978-3-319-24574-4_28.

[15] I. Arganda-Carreras, S. Seung, A. Cardona, and J. Schindelin. "Segmentation of neuronal structures in em stacks challenge - isbi 2012." Accessed: 21-09-2022. (2012), [Online]. Available: https://imagej.net/events/isbi-2012-segmentation-challenge.

[16] V. Ulman, M. Maška, K. E. G. Magnusson, O. Ronneberger, *et al.*, "An objective comparison of cell-tracking algorithms," *Nature Methods*, vol. 14, no. 12, pp. 1141–1152, Dec. 2017. DOI: https://doi.org/10.1038/nmeth.4473.

[17] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds., Cham: Springer International Publishing, 2016, pp. 424–432. DOI: https://doi.org/10.1007/978-3-319-46723-8_49.

[18] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Nnu-net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021. DOI: 10.1038/s41592-020-01008-z.

[19] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, *et al.*, Eds., Cham: Springer International Publishing, 2018, pp. 3–11. DOI: https://doi.org/10.1007/978-3-030-00889-5_1.

[20] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, *et al.*, "Attention u-net: Learning where to look for the pancreas," in *Medical Imaging with Deep Learning*, 2018. DOI: https://doi.org/10.48550/arXiv.1804.03999.

[21] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020. DOI: https://doi.org/10.1016/j.isprsjprs.2020.01.013.

[22] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, *et al.*, "The medical segmentation decathlon," *Nature Communications*, vol. 13, no. 1, p. 4128, Jul. 2022. DOI: 10.1038/s41467-022-30695-9.

[23] Y. He, D. Yang, H. Roth, C. Zhao, and D. Xu, "Dints: Differentiable neural network topology search for 3d medical image segmentation," in *2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 5841–5850. DOI: 10.1109/CVPR46437.2021.00578.

[24] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *Journal of Machine Learning Research*, vol. 20, no. 55, pp. 1–21, 2019.

[25] Y. Tang, D. Yang, W. Li, H. R. Roth, *et al.*, "Self-supervised pre-training of swin transformers for 3d medical image analysis," in *2022 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 20 730–20 740.

[26] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, *et al.*, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi and S. Bakas, Eds., Cham: Springer International Publishing, 2022, pp. 272–284.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[29] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, *et al.*, *Transformers in medical imaging: A survey*, 2022. DOI: 10.48550/ARXIV.2201.09873.

[30] E. J. Hughes, T. Winchman, F. Padormo, R. Teixeira, *et al.*, "A dedicated neonatal brain imaging system," *Magnetic resonance in medicine*, vol. 78, no. 2, pp. 794–804, Aug. 2017. DOI: 10.1002/mrm.26462.

[31] O. A. Glenn, "MR imaging of the fetal brain," *Pediatric Radiology*, vol. 40, no. 1, pp. 68–81, Nov. 2009. DOI: 10.1007/s00247-009-1459-3.

[32] A. Makropoulos, E. C. Robinson, A. Schuh, R. Wright, *et al.*, "The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction," English, *NeuroImage*, vol. 173, pp. 88–112, Jun. 2018. DOI: 10.1016/j.neuroimage.2018.01.054.

[33] L. Cordero-Grande, R. P. A. G. Teixeira, E. J. Hughes, J. Hutter, *et al.*, "Sensitivity encoding for aligned multishot magnetic resonance reconstruction," *IEEE Transactions on Computational Imaging*, vol. 2, no. 3, pp. 266–280, 2016. DOI: 10.1109/TCI.2016.2557069.

[34] E. Meuwly, M. Feldmann, W. Knirsch, M. von Rhein, *et al.*, "Postoperative brain volumes are associated with one-year neurodevelopmental outcome in children with severe congenital heart disease," *Scientific Reports*, vol. 9, no. 1, p. 10 885, Jul. 2019. DOI: 10.1038/s41598-019-47328-9.

[35] A. Makropoulos, I. S. Gousias, C. Ledig, P. Aljabar, *et al.*, "Automatic whole brain mri segmentation of the developing neonatal brain," *IEEE Transactions on Medical Imaging*, vol. 33, no. 9, pp. 1818–1831, 2014. DOI: 10.1109/TMI.2014.2322280.

[36] S. Tourbier, C. Velasco-Annis, V. Taimouri, P. Hagmann, *et al.*, "Automated template-based brain localization and extraction for fetal brain mri reconstruction," *NeuroImage*, vol. 155, pp. 460–472, 2017. DOI: https://doi.org/10.1016/j.neuroimage.2017.04.004.

[37] S. Tourbier, X. Bresson, P. Hagmann, J.-P. Thiran, *et al.*, "An efficient total variation algorithm for super-resolution in fetal brain mri with adaptive regularization," *NeuroImage*, vol. 118, pp. 584–597, 2015. DOI: https://doi.org/10.1016/j.neuroimage.2015.06.018.

[38] J. V. Frank Hutter Lars Kotthoff, *Automated Machine Learning, Methods, Systems, Challenges*. Springer Cham, 2019. DOI: https://doi.org/10.1007/978-3-030-05318-5.

[39] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, *et al.*, "Unetr: Transformers for 3d medical image segmentation," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 1748–1758. DOI: 10.1109/WACV51458.2022.00181.

[40] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[41] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

[42] P. Welander, S. Karlsson, and A. Eklund, *Generative adversarial networks for image-to-image translation on multi-contrast mr images - a comparison of cyclegan and unit*, 2018. DOI: 10.4550/ARXIV.1806.07777.

[43] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.

[44] D. G. Lloyd-Jones. "Radiology masterclass, MRI interpretation." Accessed: 21-09-2022. (2017), [Online]. Available: https://www.radiologymasterclass.co.uk/tutorials/mri/t1_and_t2_images.

[45] L. Wang. "MICCAI grand challenge on multi-domain cross-time-point infant cerebellum MRI segmentation 2022 (cseg-2022)." Accessed: 27-09-2022. (2022), [Online]. Available: https://tarheels.live/cseg2022/.

# A Participation in cSeg Challenge

In the context of this thesis, we have participated to the cSeg challenge at MICAII 2022 [45], on multi-domain, cross-time point infant cerebellum segmentation. Indeed we thought that it fits closely the theme of our study, and proposed our evaluated best-performing method: the 3d nn-UNet. We trained it on their multi-centric training data and performed ensembling of our cross-validation models. The 2$^{nd}$ place was obtained, this shows once more its strong generalization capacity.

| RANK | TEAM | SUBMISSION DATE | DSC [CSF] | MHD [CSF] | ASD [CSF] | DSC [GM] | MHD [GM] | ASD [GM] | DSC [WM] | MHD [WM] | ASD [WM] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | KispiToSee | 07/11/2022 | 0.935 | 7.091 | 0.093 | 0.935 | 5.998 | 0.136 | 0.934 | 5.929 | 0.132 | Details |
| | IBIME | 07/18/2022 | 0.935 | 5.948 | 0.091 | 0.935 | 5.511 | 0.133 | 0.933 | 5.683 | 0.132 | Details |
| | DIKU-SMM | 07/22/2022 | 0.909 | 8.390 | 0.136 | 0.916 | 5.864 | 0.175 | 0.916 | 6.466 | 0.167 | Details |
| | Dolphins | 07/29/2022 | 0.933 | 6.927 | 0.096 | 0.934 | 6.121 | 0.138 | 0.933 | 5.979 | 0.133 | Details |
| | nic_vicorob | 07/31/2022 | 0.856 | 7.326 | 0.212 | 0.854 | 5.862 | 0.262 | 0.844 | 5.616 | 0.286 | Details |

Figure A.1: The 2$^{nd}$ place was obtained by our team, KispiToSee, using the 3d nn-UNet.

# B Example images of T1 to T2 generation (and reverse) with CycleGAN

CycleGAN allows us to generate T2 images from T1 MR images and vice-versa. The network was found online at `https://github.com/simontomaskarlsson/GAN-MRI`, the GitHub implementation of [42], and trained using 40 dHCP T1-T2 pairs. In order to train the network which is 2d, we sliced our 3d images into 2d axial slices. We cropped them from their original size to 256x256 as required by the input layer of the GAN. The sample size of the training set is 8087 images (2d slices), and the size of the test set is 6007 images. After training, one has two generator models, a T1 to T2 generator and a T2 to T1 generator. One can use either of them to generate T2 slices (resp. T1 slices) from T1 slices (resp. T2 slices); one can then stack the generated slices to construct a 3D T2 image (resp. T1 image).

To judge the accuracy of our GAN results, we look at two metrics; the correlation coefficient and the mutual information between the synthetised image and the ground truth image. In the following, we show some example images of generated T1 slices and T2 slices alongside the respective ground truth.

We start with some good examples:

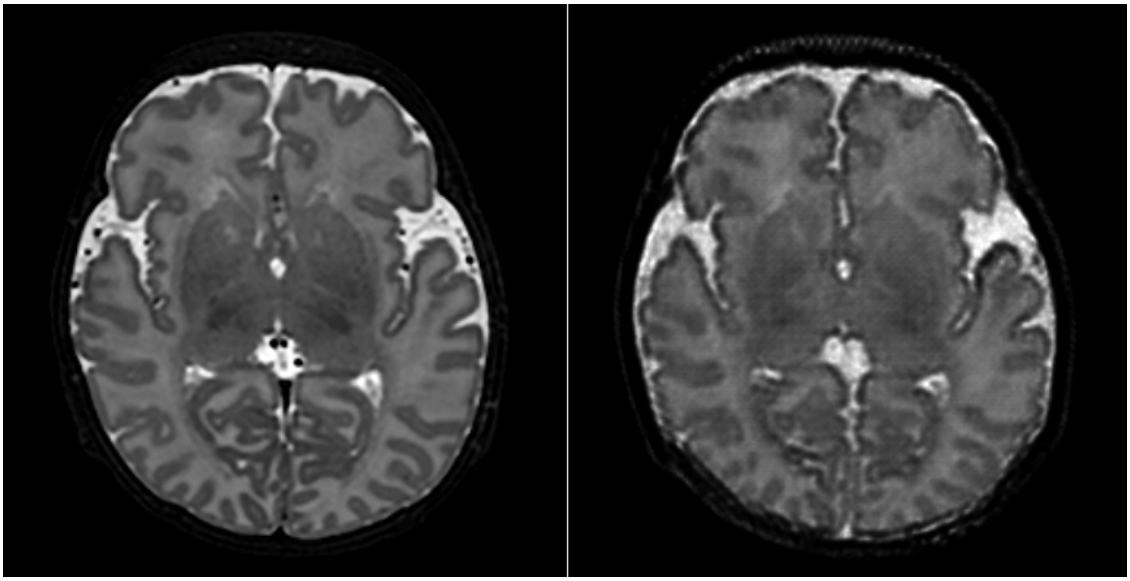


Figure B.1: 101$^{st}$ Axial slice of sub-CC0174, dHCP data. Correlation coefficient: 0.94, MI: 0.92.

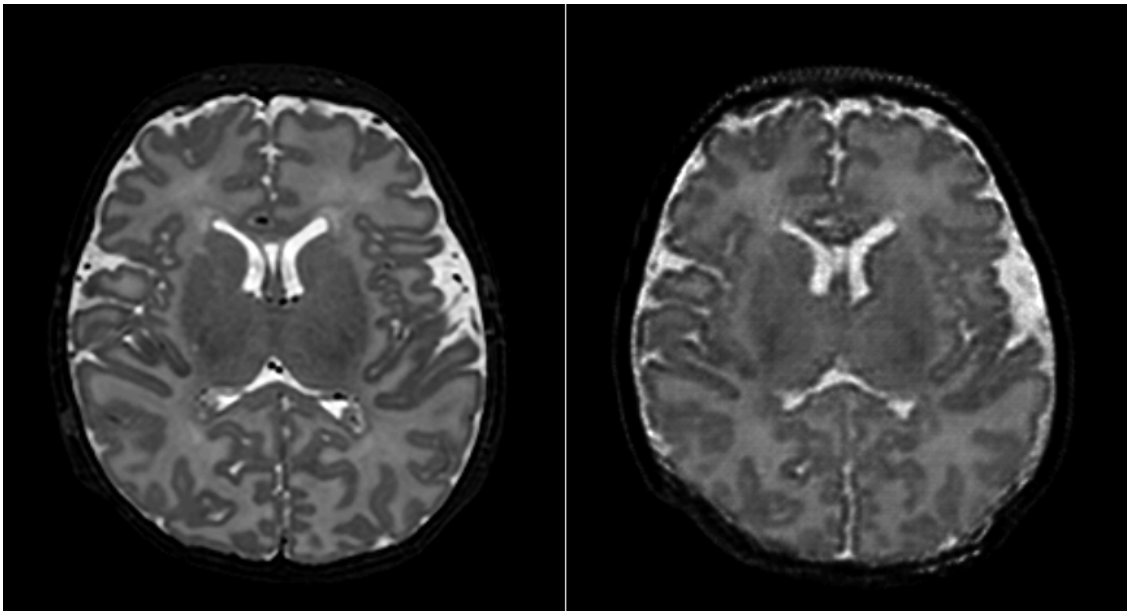


Figure B.2: 115$^{\text{th}}$Axial slice of sub-CC0174, dHCP data. Correlation coefficient: 0.94, MI: 0.90.

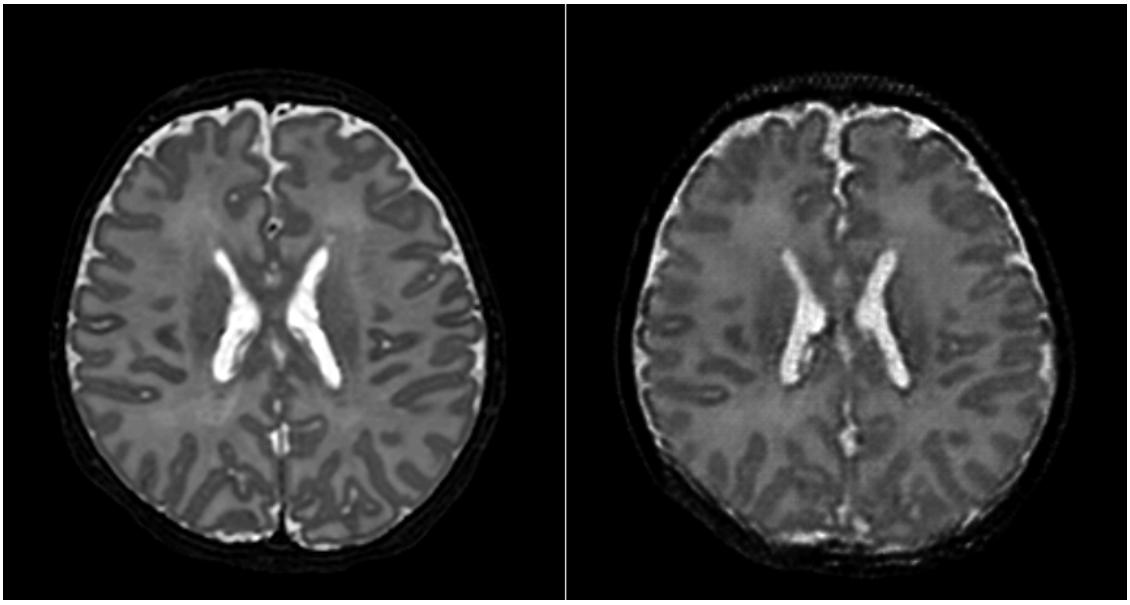


Figure B.3: 133$^{\text{th}}$Axial slice of sub-CC0174, dHCP data. Correlation coefficient: 0.96, MI: 0.93.

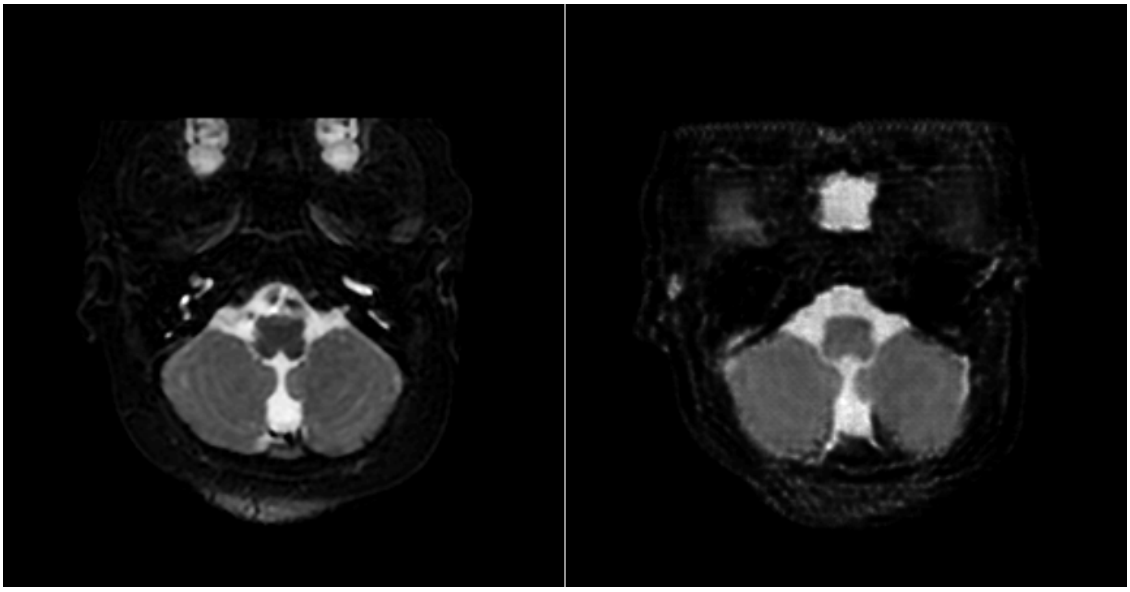


Figure B.4: 25$^{\text{th}}$Axial slice of sub-CC0174, dHCP data.

We move on to some worse examples, assesed visually.

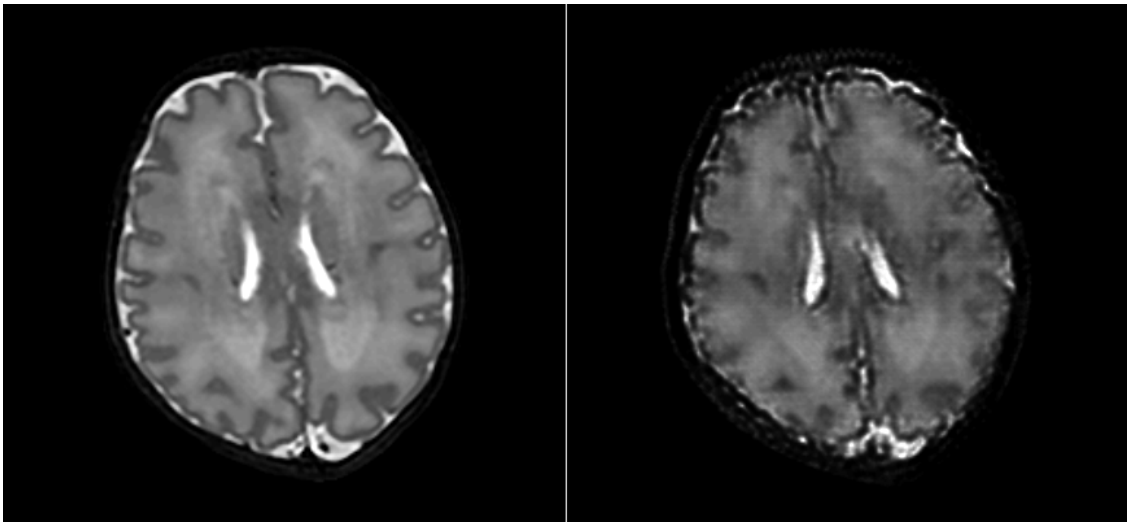


Figure B.5: 127$^{\text{th}}$Axial slice of sub-CC0177, dHCP data. Correlation coefficient: 0.92, MI: 0.70. The CSF of outside of brain is not well generated by the GAN.

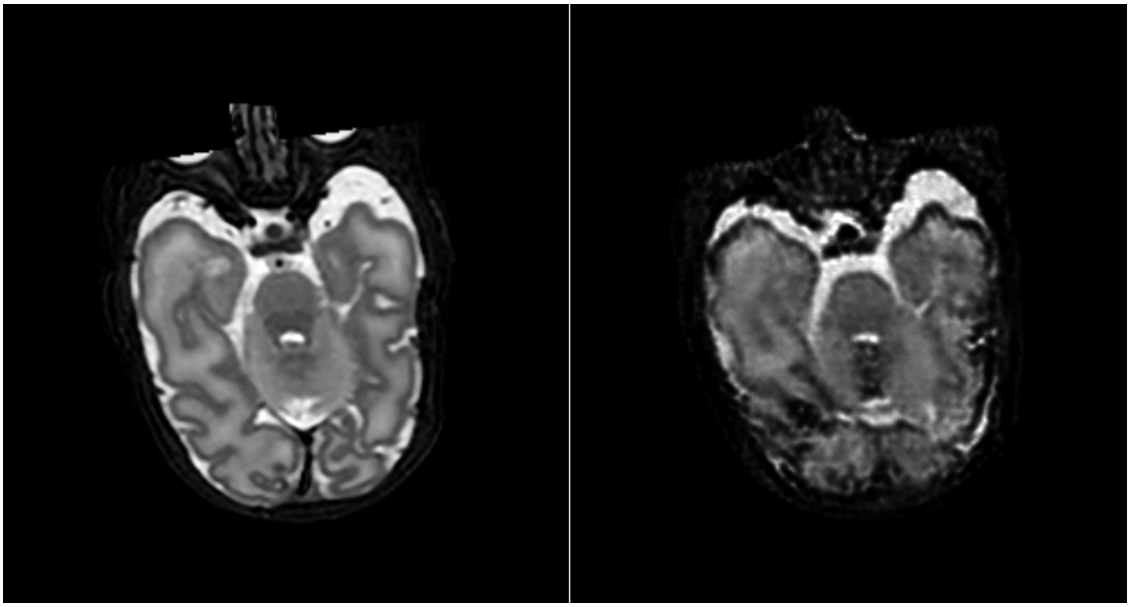


Figure B.6: 64$^{\text{th}}$Axial slice of sub-CC0177, dHCP data. Correlation coefficient: 0.89, MI: 0.53 Here the WM/GM boundary is not clear in the synthetised image.

We now have a look at some T1 generation, results are very impressive in the T1 case and it is hard to find bad examples. On the other hand, one can see the pixelization of the synthetic image better than in the T2 case as they are brighter. Here are some good examples:

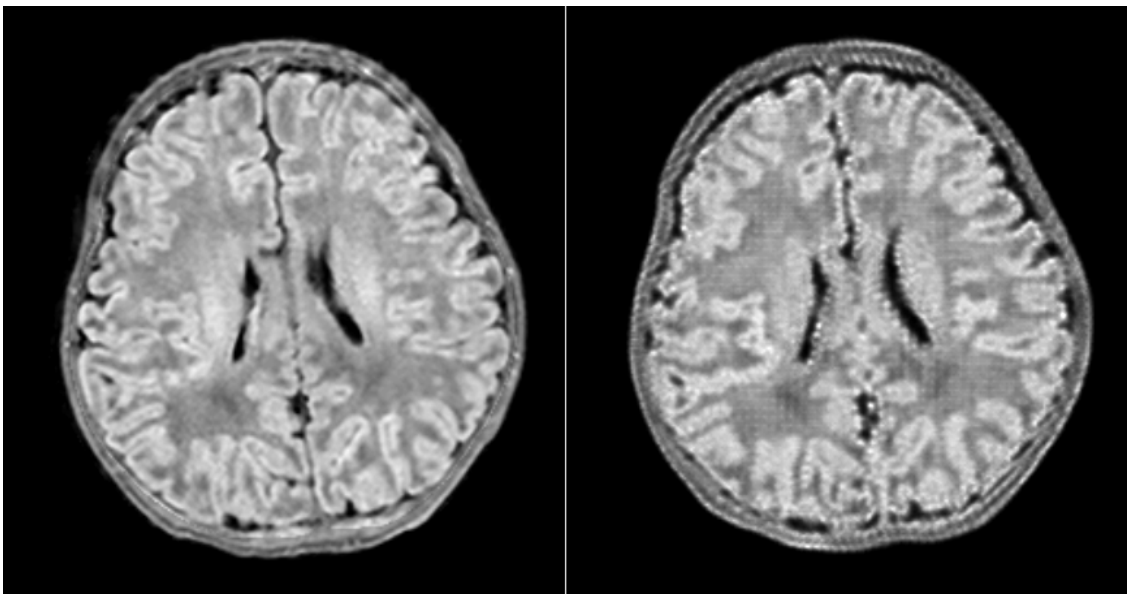


Figure B.7: 116$^{\text{th}}$Axial slice of sub-CC0176, dHCP data. Correlation coefficient= 0.96, MI: 0.89.

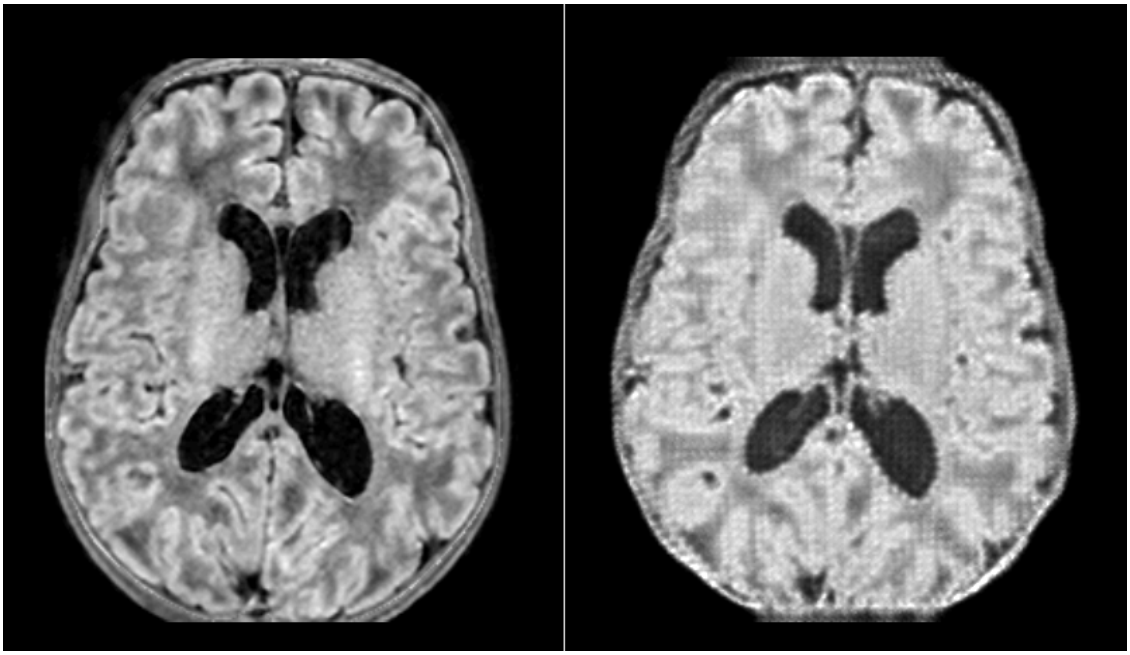


Figure B.8: 101st Axial slice of sub-CC0194, dHCP data. Correlation coefficient: 0.92, MI: 0.76. The GAN translates the image quite well, even though this case has largely inflated ventricles that were not seen in the GAN training set.

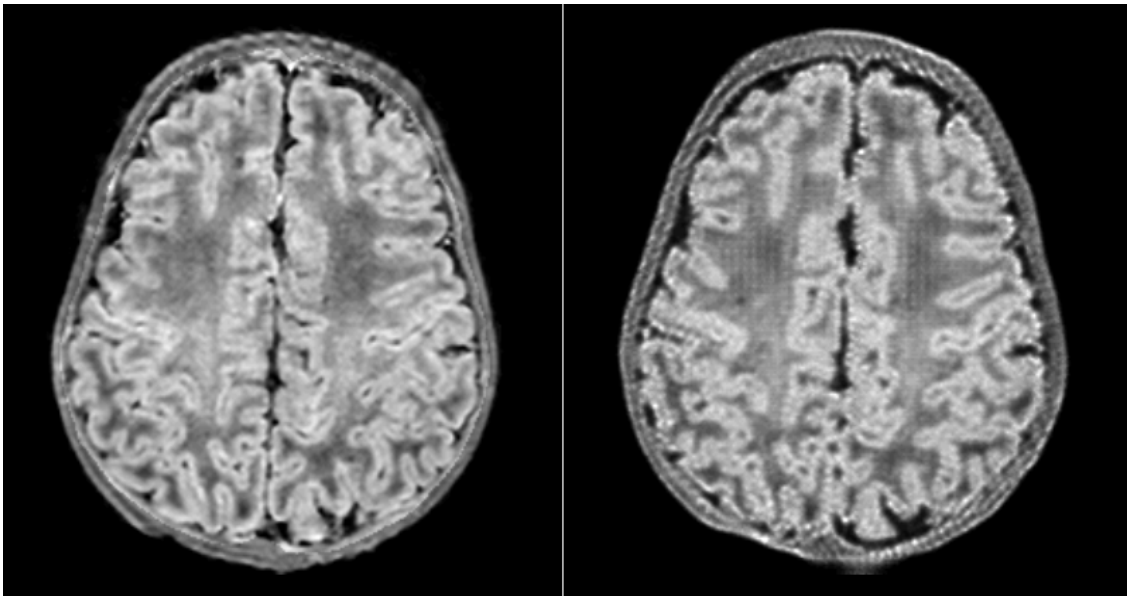


Figure B.9: 151st Axial slice of sub-CC0194, dHCP data. Correlation coefficient: 0.95, MI: 0.9.

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Universität Zürich[UZH]

# Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

---

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

## Title of work

| |
|---|
| Domain adaptation for multi-centric fetal and infant MRI segmentation |

## Authored by

Name(s) First name(s)

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Zürich, October 2022

Charles Moatti