Alma Mater Studiorum · University of Bologna

School of Science
Department of Physics and Astronomy
Master Degree in Physics

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

Supervisor:
Prof./Dr. Name Surname

Co-supervisor:
Dr. Gerard Martí-Juan

Submitted by:
Simone Chiarella

Academic Year 2024/2025

Abstract

Scientific documents often use LaTeX for typesetting. While numerous packages and templates exist, it makes sense to create a new one. Just because.

Contents

List of Figures

List of Tables

Part I

Introduction

Type here the introduction to the part.

1 Magnetic Resonance Imaging

2 The FeTA Challenge

The Fetal Tissue Annotation Challenge (FeTA) [1] was born in 2020, and joined the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) [2] in 2021. Up to now, four editions have been organized (in 2020, 2021, 2022, and 2023), with increasing participation and interest from the medical imaging community. The main contributions of the FeTA challenge are the creation of a benchmark dataset for fetal brain MRI segmentation and biometry, and the promotion of the development of algorithms for the automatic segmentation of fetal brain tissues.

The main task in FeTA is the segmentation of brain tissues in fetal MRI, which is a challenging problem due to the low contrast between tissues, the presence of noise, and the variability in the shape and size of the fetal brain. The dataset used in the challenge is composed of 3D super-resolution (SR) reconstructions of 2D fetal brain MRI images. Participants are asked to segment the fetal brain into seven tissues. The performance is evaluated using different metrics, such as the Dice similarity coefficient between the predicted and ground truth segmentations. [3]

2.1 FeTA 2020

The first edition of the FeTA challenge was organized in 2020, by Payette et al. [4]. The challenge consisted in segmenting fetal brain MRI T2w images. The initial FeTA dataset comprised 40 super-resolution (SR) reconstructions with manual segmentations for training and 10 SR reconstructions without manual segmentation for validation, encompassing both pathological and non-pathological cases. The gestational age (GA) range spanned from 20 to 33 weeks. The scans were acquired in This dataset established a standard in fetal brain tissue parcellation—according to a seven-tissues protocol previously introduced in [5]—that would be used in all the following FeTA editions. The seven tissue types are: external cerebrospinal fluid (CSF), cortical gray matter (GM), white matter (WM), ventricles, cerebellum, deep gray matter, and brainstem.

Four research groups participated, submitting a total of ten algorithms. Nine out of ten used deep learning methods (eight of which were based on 2D and 3D U-Nets), and one used a multi-atlas segmentation method. The assessment of the results was carried out using the Dice similarity coefficient (DSC), the volume similarity (VS), and the Hausdorff 95 distance (HD95). The use of three metrics helps to reduce the reliance on any one metric, which may be misleading in the evaluation of the algorithms. In fact, the authors admit that manual segmentations included in both the training and testing dataset were not perfect, and therefore there are mislabeled voxels, especially in low-resolution scans [6]. All the algortihms had more or less the same issues in segmenting the CSF—especially for the pathological cases, because of not clear tissue boundaries—and the GM, because of its rapidly changing structure. The best performing method was a 3D U-Net made up by the combination of three 2D U-Nets, one per direction (sagittal, coronal, and axial). It is worth noting that the multi-atlas segmentation method performed better than the deep learning methods when the quality of the SR was poor. This is because such method can leverage its prior knowledge even if the structure is not clear in the image.

The dataset used in the first FeTA edition had important limitations:

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

2.2 FeTA 2021

The 2021 edition of the FeTA challenge [6, 7] was the first to join the MICCAI conference. The dataset—hereinafter referred to as Kispi dataset—was expanded to 120 scans from the same institution, with GAs ranging from 20 to 35 weeks. The acquisition was carried out at 1.5 T for a subset of cases, and at 3 T for another subset of cases. 60 scans were reconstructed with the MIALSRTK method [8, 9], while the other 60 cases with the SIMPLE IRTK method [10, 11]. For each reconstruction method, 40 cases were included in the training dataset available to the challenge participants (for a total of 80 cases), and 20 cases were included in testing dataset not available to the participants (for a total of 40 cases). The maternal tissue was excluded from the SR reconstruction, only the fetal brain was reconstructed. There were slightly more pathological than neurotypical cases. In the group with atypical features, a variety of cerebral pathologies of varying severities were included, such as Chiari-II malformation or ventricular dysmorphology seen in ventriculomegaly.

21 algorithms were submitted, of which 19 were U-Nets, with no major differences in the architecture. The main differences across the submissions were in how the training was performed (the use of cross-validation or changes in the learning rate decay), or in the pre-processing (patch size, how the data was normalized) and post-processing (ensemble learning, removal of external label "blobs"). Overall, the most challenging labels to segment were cortical and deep GM—due to limited image resolution and annotation uncertainty—and brainstem—especially in the pathological cases. As in the previous edition, also the segmentation of the CSF was challenging. The results of the image quality and SR reconstruction methods are related to each other, as the majority of the low quality images were done with the MIALSRTK method, and the excellent quality brain volumes included were reconstructed with the SIMPLE IRTK method.

2.3 FeTA 2022

FeTA 2022 [12]—to date, the last edition for which a detailed review is available [13]—introduced a multi-center dataset to address the generalizability of algorithms, which was one of the main limitations of the previous editions. The challenge dataset consisted of fetal brain MRI reconstructions acquired from four different imaging centers. In addition to Kispi, data from Medical University of Vienna was incorporated into both the training and testing datasets. Data from two further centers were included in the testing dataset—University Hospital Lausanne (CHUV), and Benioff Children's Hospital (UC San Francisco, UCSF)—for a total of four centers (Tab 2.1). Data from Vienna differ significantly from all the others because they are not cropped around the brain, meaning that maternal tissues are also visible. The evaluation metrics were the same as in the previous editions.

Inst.	Scanner (field strength in Tesla)	SRR method	TR/TE (ms)	GA range (weeks)
		Training		
Kispi	GE Signa Discovery MR450/MR750 (1.5/3)*	MIALSRTK (40) SIMPLE IRTK (40)	2000-3500 120**	20.0-34.8
Vienna	Philips Ingenia/Intera (1.5) Philips Achieva (3)	NiftyMIC (40)	6000-22000 80-140	19.3-34.4
		Testing		
Kispi	GE Signa Discovery MR450/MR750 (1.5/3)*	MIALSRTK (20) SIMPLE IRTK (20)	2000-3500 120**	21.3-34.6
Vienna	Philips Ingenia/Intera (1.5) Philips Achieva (3)	NiftyMIC (40)	6000-22000 80-140	18.1-35.0
CHUV	Siemens MAGNETOM Aera (1.5)	MIALSRTK (40)	1200 90	21.0-35.0
UCSF	GE Discovery MR750/MR750W (3)	NiftyMIC (40)	2000-3500 100**	20.0-35.1

Table 2.1: Training and testing dataset properties in FeTA 2022. In parenthesis, next to each SRR method, is reported the correspondent number of images. *The field strengths respectively refers to the scanners. **TE values represent the minimum durations.

17 algortihms were submitted. nnU-Net was the most used and effective tool. The most popular loss functions were the DSC loss and cross-entropy loss, or a combination of the two. In regard to the in-domain results, a performance plateau was observed in the DSC scores, similar to FeTA 2021. Compared to the previous edition results, segmentation accuracy improved marginally. The highest DSC in the FeTA 2022 in-domain evaluations was 0.805, while it was 0.786 in 2021. Some submissions demonstrated equivalent performance for both the in-domain and out-of-domain (OOD), while others showed a significant drop in performance. This indicates that the domain shift present in data from different imaging centers can drastically degrade model performance when being

deployed in heterogenous clinical datasets. Overall, the median performance metrics in the OOD setting remain equivalent to the in-domain. The major drops of performance occur in ventricles, and in GM and WM volumes (the drop is observed only in VS). The most challenging labels to segment remained cortical and deep GM, and BS. Notably, some algorithms performed better in the OOD setting than in the in-domain setting. This can be explained with the better quality of the images from the CHUV and UCSF centers, which were included only in the test set. Style and photometric augmentations (contrast, blur, sharpness, etc.) turned out to be effective in improving the generalization of the models. However, "the optimum choice of augmentation techniques remains unclear" [13], standing as a critical factor in achieving domain generalization.

2.4 FeTA 2024

In FeTA 2024 [1], besides a fetal brain biometry task, 20 new scans were added to the test set, in order to have more results on the OOD performance. These scans were acquired at St. Thomas Hospital (King's College London, KCL), with field strength of 0.55 T (Siemens MAGNETOM Free.Max) [3]. This decision follows the recent rise in popularity of low-cost low-field MRI systems [14], which are particularly suitable for fetal imaging due to their lower SAR and acoustic noise. Furthermore, those systems make real-time 3D imaging feasible, allowing the MRI operator to track the fetus as it moves and capture diagnostic quality images during quiescence, which can dramatically shorten exam time [15].

16 algortihms were submitted, of which nine were based on nnU-Net. The top scorer team used an nnU-Net with a residual encoder, generating ensamble predictions from different models—this is fundamental for predictions on different domains, because data augmentation is beneficial to OOD, but at the same time can be detrimental to in-domain. The second top team also used nnU-Net, but training it on real and synthetic data (SynthSeg [16]), and applying post-processing to discard non-brain tissues from the predictions. Complete data are not available yet, but preliminary results show plateaus in all the metrics (Fig. 2.1). DSC median values for the in-domain setting is comparable to OOD. With respect to previous editions, the drop in performance between in-domain and OOD seems to be smaller. The quality of the SR reconstruction is still a critical factor in the segmentation performance, as demonstrated by the many outliers in the in-domain setting, especially in Kispi. Interestingly, even though no 0.55 T scans were included in the training set, the performance on the KCL test set is even better than the other OOD centers. This suggests that the models generalize well across different field strengths, but also in this case SR quality must be taken into account as well, and in general its effect should be more investigated.

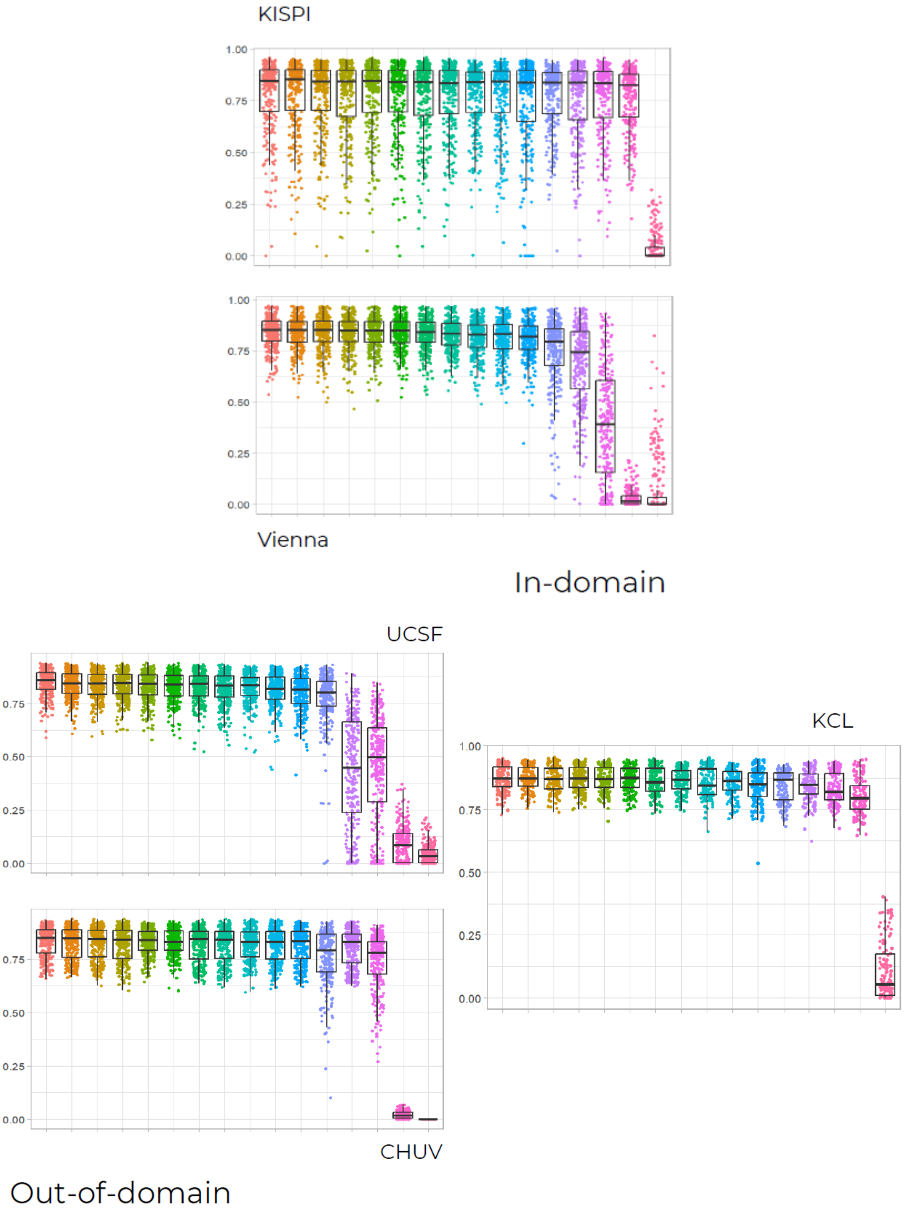

Figure 2.1: Teams' Dice similarity scores in FeTA 2024. From [1].

Part II

Materials and Methods

Type here the introduction to the part.

Part III

Results and Discussion

Type here the introduction to the part.

Conclusions

This is a conclusion.

Acronyms

BS	Brainstem
CHUV	Centre Hospitalier Universitaire Vaudois (Lausanne University Hospital)
CSF	Cerebrospinal fluid
DSC	Dice similarity coefficient
FeTA	Fetal Tissue Annotation Challenge
GA	Gestational age
GM	Gray matter
HD95	Hausdorff 95 distance
KCL	King's College London (St. Thomas Hospital)
Kispi	Universitäts-Kinderspital Zürich (Zurich University Children's Hospital)
OOD	Out-of-domain
SR	Super resolution
UCSF	University of California, San Francisco (Benioff Children's Hospital)
VS	Volume similarity
WM	White matter

Bibliography

1. *FeTA 2024 Challenge.* URL: https://fetachallenge.github.io/ (visited on 01/29/2025) (cit. on pp. 5, 8–9).
2. *MICCAI: Medical Image Computing and Computer-Assisted Intervention.* URL: https://miccai.org/ (visited on 01/29/2025) (cit. on p. 5).
3. M. Bach Cuadra, K. Payette, A. Jakab, et al. *Fetal Tissue Annotation Challenge: Structured description of the challenge design.* 2024. DOI: 10.5281/zenodo.10986046. URL: https://doi.org/10.5281/zenodo.10986046 (cit. on pp. 5, 8).
4. K. Payette, P. de Dumast, H. Kebiri, et al. "An automatic multi-tissue human fetal brain segmentation benchmark using the Fetal Tissue Annotation Dataset". *Scientific Data* 8:167, 1 2021. DOI: 10.1038/s41597-021-00946-3. URL: https://doi.org/10.1038/s41597-021-00946-3 (cit. on p. 5).
5. K. Payette, R. Kottke, and A. Jakab. "Efficient Multi-class Fetal Brain Segmentation in High Resolution MRI Reconstructions with Noisy Labels". In: *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis.* Ed. by Y. Hu, R. Licandro, J. A. Noble, et al. Springer International Publishing, 2020, pp. 295–304. DOI: 10.1007/978-3-030-60334-2_29. URL: https://doi.org/10.1007/978-3-030-60334-2_29 (cit. on p. 5).
6. K. Payette, B. L. Hongwei, P. de Dumast, et al. "Fetal brain tissue annotation and segmentation challenge results". *Medical Image Analysis* 88, 2023. DOI: https://doi.org/10.1016/j.media.2023.102833. URL: https://www.sciencedirect.com/science/article/pii/S1361841523000932 (cit. on p. 6).
7. *FeTA 2021 Challenge.* URL: https://feta.grand-challenge.org/feta-2021/ (visited on 02/21/2025) (cit. on p. 6).
8. S. Tourbier, X. Bressonc, P. Hagmann, et al. "An efficient total variation algorithm for super-resolution in fetal brain MRI with adaptive regularization". *NeuroImage* 118, 2015, pp. 584–597. DOI: 10.1016/j.neuroimage.2015.06.018. URL: https://doi.org/10.1016/j.neuroimage.2015.06.018 (cit. on p. 6).
9. P. Deman, S. Tourbier, R. Meuli, and M. Bach Cuadra. *meribach/mevislabFetalMRI: MEVISLAB MIAL Super-Resolution Reconstruction of Fetal Brain MRI v1.0.* Version v1.0. 2020. DOI: 10.5281/zenodo.3878564. URL: https://doi.org/10.5281/zenodo.3878564 (cit. on p. 6).
10. M. Kuklisova-Murgasova, G. Quaghebeur, M. A. Rutherford, et al. "Reconstruction of fetal brain MRI with intensity matching and complete outlier removal". *Medical Image Analysis* 16, 2012, pp. 1550–1564. DOI: 10.1016/j.media.2012.07.004. URL: https://doi.org/10.1016/j.media.2012.07.004 (cit. on p. 6).

11. M. Kuklisova-Murgasova. *SIMPLE IRTK*. 2019. URL: https://gitlab.com/mariadeprez/irtk-simple (cit. on p. 6).
12. *FeTA 2022 Challenge*. URL: https://feta.grand-challenge.org/feta-2022/ (visited on 02/21/2025) (cit. on p. 7).
13. K. Payette, C. Steger, R. Licandro, et al. "Multi-Center Fetal Brain Tissue Annotation (FeTA) Challenge 2022 Results". *IEEE Transactions on Medical Imaging*, 2024. DOI: 10.1109/tmi.2024.3485554. URL: http://dx.doi.org/10.1109/TMI.2024.3485554 (cit. on pp. 7–8).
14. J. Aviles Verdera, L. Story, M. Hall, et al. "Reliability and Feasibility of Low-Field-Strength Fetal MRI at 0.55 T during Pregnancy". *Radiology* 309:1, 2023. DOI: 10.1148/radiol.223050. URL: https://doi.org/10.1148/radiol.223050 (cit. on p. 8).
15. S. Ponrartana, H. N. Nguyen, S. X. Cui, et al. "Low-field 0.55 T MRI evaluation of the fetus". *Pediatric Radiology* 53, 2023, pp. 1469–1475. DOI: 10.1007/s00247-023-05604-x. URL: https://doi.org/10.1007/s00247-023-05604-x (cit. on p. 8).
16. B. Billot et al. "SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining". *Medical Image Analysis* 86, 2023. DOI: 10.1016/j.media.2023.102789. URL: https://www.sciencedirect.com/science/article/pii/S1361841523000506 (cit. on p. 8).