MELBA Journal Machine Learning for Biomedical Imaging

Evaluating Synthetic Data Generation for Domain Generalization in Fetal Brain MRI Segmentation

Vladyslav **Zalevskyi** [1,2,*], Thomas **Sanchez** [1,2,*], Margaux **Roulet** [1,2], Busra **Bulut** [1,2], Hélène **Lajous** [1,2], Jordina **Aviles Verdera** [3,4], Sara **Neves Silva** [4], Georg **Langs** [6,7,8], Gregor **Kasprian** [7,9], Roxane **Licandro** [6,7,8], Jana **Hutter** [3,4], Hamza **Kebiri** [1,2], Meritxell **Bach Cuadra** [1,2]

1 Department of Radiology, Lausanne University Hospital and University of Lausanne (UNIL), Lausanne, Switzerland
2 CIBM Center for Biomedical Imaging, Lausanne, Switzerland
3 Institute for Information Processing, Leibniz University Hannover, Hannover, Germany
4 Department of Biomedical Engineering, School of Biomedical Engineering & Imaging Sciences, King's College London, London, United Kingdom
5 Department of Biomedical Imaging and Image-Guided Therapy, Division of Neuroradiology and Musculoskeletal Radiology, Medical University of Vienna, Vienna, Austria
6 Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research Lab (CIR), Medical University of Vienna, Vienna, Austria
7 Christian Doppler Laboratory for Mathematical Modelling and Simulation of Next-Generation Medical Ultrasound Devices, Medical University of Vienna, Vienna, Austria
8 Comprehensive Center for Artificial Intelligence in Medicine, Medical University of Vienna, Vienna, Austria
9 Division of Neuroradiology and Musculoskeletal Radiology, Department of Biomedical Imaging and Image–guided Therapy, Medical University of Vienna, Vienna, Austria
***** Equal contribution

Abstract

Fetal brain tissue segmentation from magnetic resonance imaging (MRI) is crucial for studying neurodevelopment, but remains challenging due to data heterogeneity and limited annotations. Domain randomization (DR) has recently emerged as a promising strategy for single-source domain generalization by synthesizing training images with randomized artifacts, contrast, and resolution. In this work, we investigate how to maximize the out-of-domain (OOD) generalization of DR-based methods. We evaluate several synthetic data generation strategies for DR, with a particular focus on our recently proposed framework, `FetalSynthSeg`. We show that simple Gaussian mixture-based intensity modeling outperforms more complex physics-based simulations, and that *intensity clustering* (subdividing tissue classes based on intensity) improves OOD robustness. Evaluated on 348 fetal subjects from four sites spanning 0.55–3T and both T1w and T2w contrasts, `FetalSynthSeg` reaches state-of-the-art performance on several FeTA 2024 testing datasets (80–85 Dice score) and, for the first time, offers robust segmentation on modalities other than T2w for fetal brain segmentation (80 Dice on dHCP-T1w dataset). Compared with state-of-the-art methods such as BOUNTI, nnU-Net ensemble, and the FeTA 2024 winner, `FetalSynthSeg` delivers comparable or superior accuracy while maintaining strong robustness across domain shifts. Our code, model weights, and Docker image ready for easy inference are available at `https://hub.docker.com/r/vzalevskyi/fetalsynthseg`.

Keywords

Segmentation, Fetal Brain, MRI, Domain shifts, Synthetic Data, Domain Randomization

Article informations

`https://doi.org/10.59275/j.melba.2026-6e4e`
Volume 3, Received: 2025-12-02, Published 2026-07-01
Corresponding author: vladyslav.zalevskyi@unil.ch

1. Introduction

Fetal development is a critical period shaping lifelong neurological and physiological outcomes (Halfon et al., 2014). Monitoring brain maturation and detecting atypical development during this period enables early diagnosis and intervention (Saleem, 2014). While ultrasonography (US) remains the default fetal imaging modality because of its broad availability and its low cost, fetal magnetic resonance imaging (MRI) complements US: it provides superior soft-tissue contrast and is less operator-dependent (Glenn and Barkovich, 2006). MRI has been used to study fetal brain development (Jakab et al., 2021; Aviles Verdera et al., 2023; Machado-Rivas et al., 2022; Shen et al., 2022) and for the detection of abnormalities such as ventriculomegaly, spina bifida, corpus callosum agenesis, edema, and hemorrhage (Garel and Garel, 2004; Pfeifer et al., 2019). Although, automated methods for fetal brain MRI analysis are promising to enable rapid and reliable scan processing (Uus et al., 2023), they struggle with limited data availability and domain shifts, which are particularly pronounced in the fetal population (Dockès et al., 2021; Varoquaux and Cheplygina, 2022; Guan and Liu, 2022).

Domain shifts, or distribution shifts — a mismatch between the distribution of training and testing data — are quite severe in MRI and stem from differences in scanner hardware, imaging protocols, and reconstruction methods (Yan et al., 2020). Fetal MRI exacerbates these challenges as the fetal brain morphology heavily changes during gestation and can be drastically altered by pathological processes (Stiles and Jernigan, 2010; Dubois et al., 2020). Furthermore, fetal MR images are typically acquired as T2-weighted orthogonal stacks of two-dimensional (2D) slices using fast spin echo sequences to mitigate motion artifacts. Post-processing using super-resolution (SR) reconstruction algorithms is then needed to create a high-resolution three-dimensional (3D) volume (Rousseau et al., 2010; Gholipour et al., 2010; Kuklisova-Murgasova et al., 2012; Tourbier et al., 2015; Ebner et al., 2020), which induces an additional layer of heterogeneity. As a result, training models that generalize across these dimensions (Figure 1) remains challenging (Payette et al., 2024; Zalevskyi et al., 2025).

Many strategies have been proposed to mitigate domain shifts, including diffusion-based models (Kaandorp et al., 2025a; Niemeijer et al., 2024) and data augmentation-based approaches (Shorten and Khoshgoftaar, 2019). Recently, domain randomization (Tobin et al., 2017) techniques have seen great success in the field of MR image analysis (Billot et al., 2021, 2023b; Liu et al., 2024a). These methods begin with label maps rather than real images, generating synthetic images with randomized contrasts using Gaussian mixture models, alongside common artifact corruptions, and demonstrated excellent OOD generalization performance, especially in MRI (Billot et al., 2023b).

Contributions This work extends our MICCAI 2024 publication (Zalevskyi et al., 2024) by expanding the evaluation of the proposed `FetalSynthSeg` model to additional datasets and presenting new ablation studies and comparisons with state-of-the-art (SoTA) methods. Specifically, in this work, we aim to investigate two main research hypotheses:

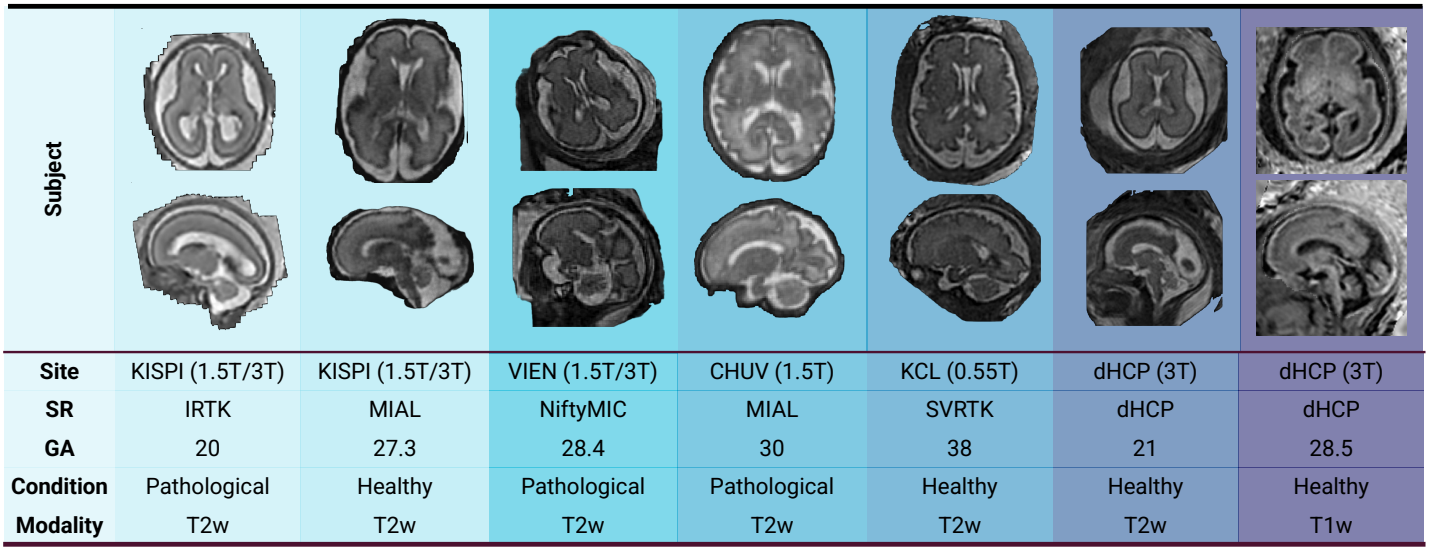
Hypothesis 1 *Do domain randomization methods, by simulating a wide range of contrasts, provide more effective data augmentation than physics-based approaches?*

Hypothesis 2 *Do domain randomization methods enable more robust OOD generalization than models trained using only real data, while maintaining competitive in-domain (ID) performance?*

Our results explore and confirm both hypotheses, and also disentangle which components of the data generation pipeline contribute most to out-of-domain robustness, what types of synthetic contrast modeling are most effective, and under which conditions such approaches may fail when applied to diverse clinical and research cohorts. In this work, we focus on the data synthesis and augmentation strategies rather than the design of new neural network architectures, while the main methodological contributions lie in the non-learning-based components that generate and randomize the training data. By isolating and evaluating these components, our study aims to provide insights that remain applicable across future generations of segmentation networks. We summarize our key contributions are as follows:
(i) We compare **Gaussian mixture-based** contrast simulation with **physics-based** approaches to assess the benefits and drawbacks of physically grounded modeling (Figure 2).
(ii) We benchmark `FetalSynthSeg` against leading SoTA methods: BOUNTI (Uus et al., 2022), the FeTA 2024 winner (Zalevskyi et al., 2025), and an nnU-Net ensemble (Isensee et al., 2021), on 348 subjects spanning multiple scanners, SR algorithms, and modalities.
(iii) We show that `FetalSynthSeg` enables new imaging applications by robustly segmenting T2-weighted images across varying echo times and T1-weighted acquisitions, making new downstream tasks such as T2 mapping in fetal brain imaging possible.

The code, trained models, and a ready-to-use Docker image are available at: `https://github.com/Medical-Image-Analysis-Laboratory/FetalSynthSeg`.

2. Related Works



Subject							
Site	KISPI (1.5T/3T)	KISPI (1.5T/3T)	VIEN (1.5T/3T)	CHUV (1.5T)	KCL (0.55T)	dHCP (3T)	dHCP (3T)
SR	IRTK	MIAL	NiftyMIC	MIAL	SVRTK	dHCP	dHCP
GA	20	27.3	28.4	30	38	21	28.5
Condition	Pathological	Healthy	Pathological	Pathological	Healthy	Healthy	Healthy
Modality	T2w	T2w	T2w	T2w	T2w	T2w	T1w

Figure 1: **Domain shifts are ubiquitous in fetal brain super-resolution MRI.** Variations in acquisition protocols, super-resolution methods, gestational age and pathology distributions, field strengths, and specific MRI modalities contribute to distributional differences that affect model generalization.

2.1 Tackling domain shifts

Various strategies have been proposed to address domain shifts in medical imaging. Data augmentation and fine-tuning are widely used solutions (Guan and Liu, 2022), but many task-specific approaches have been designed based on a wide variety of techniques like meta-learning (Li et al., 2018), harmonization (Hu et al., 2023) or style transfer (Zhou et al., 2021). Although effective, these methods often rely on target domain data and annotations, which are costly and difficult to obtain in fetal MRI (Payette et al., 2023). Manual segmentation of a single case can take several hours (e.g., eight hours in Kyriakopoulou et al. (2017)). Moreover, approaches requiring multiple source domains are limited when only one domain is available or when existing domains fail to represent new ones (Zhou et al., 2023). This limitation is particularly pronounced in fetal imaging, where new scanner field strengths and SR methods introduce previously unseen data distributions. Poor generalization to these domains hinders the adoption of low-field ($<$1T) and ultra-low-field ($<$0.1T) MR systems, which could democratize access to advanced neuroimaging techniques to a broader population (Aviles Verdera et al., 2023). Furthermore, the scarcity and the small size of current fetal brain MRI datasets exacerbate this problem (Payette et al., 2023; Lajous et al., 2022).

2.2 Single-source domain generalization (SSDG)

To mitigate these issues, SSDG techniques aim to train models on a *single source domain* while ensuring generalization to unseen target domains. In medical imaging, common strategies include texture and intensity augmentations and synthetic view blending. For example, Ouyang et al. (2022) proposed a global intensity non-linear augmentation (GIN) module using shallow convolutional networks to enhance intensity diversity, while Li et al. (2023) simulated inter-domain frequency discrepancies through frequency mixing and self-supervised learning. Other notable efforts include the adversarial domain synthesizer (ADS) by Xu et al. (2022), which generated synthetic domains using mutual information regularization to preserve semantic consistency, and the test-time augmentation approach by Liu et al. (2022), which leveraged dictionary learning to extract semantic priors for segmentation refinement.

Despite advances, SSDG methods remain limited by the lack of semantic diversity in training data and their focus on 2D images. The FeTA Challenge 2021 clearly showed that 3D models outperform 2D ones for fetal brain segmentation, emphasizing the need for SSDG methods that extend to 3D data (Payette et al., 2023).

2.3 Synthetic data and domain randomization

Synthetic data generation and domain randomization have emerged as effective ways to overcome SSDG limitations (Al Khalil et al., 2023; Paproki et al., 2024; Gopinath et al., 2024). Deep generative models such as generative adversarial networks (GANs) and diffusion models can produce realistic data but require extensive datasets and are challenging to train (Kazerouni et al., 2023; Kazeminia et al., 2020; Kaandorp et al., 2025b). Alternatively, numerical phantoms such as `FaBiAN` generate realistic MR images through physical simulations based on anatomical priors

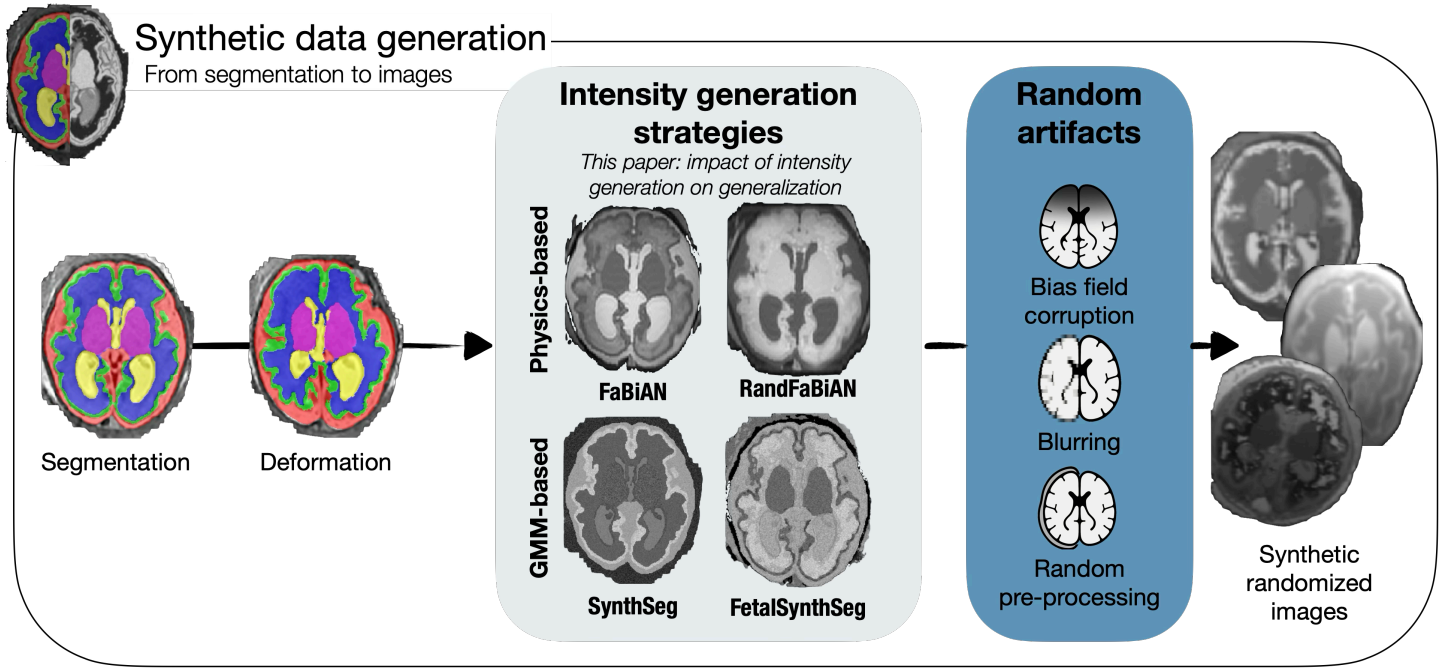


Figure 2: **Summary of the contribution of this paper:** we systematically investigate how strategies for intensity generation impact the OOD generalization and show that our proposed method, `FetalSynthSeg`, achieves strong generalization ability across sites, reconstruction methods and imaging modalities.

without pre-training (Lajous et al., 2022), but at the cost of longer generation times. `FaBiAN` can simulate multiple acquisitions for the same subject with different physical parameters and model fetal motion, allowing to complement scarce clinical datasets (Bhattacharya et al., 2024; Lajous et al., 2024b,a).

In contrast, `SynthSeg` (Billot et al., 2021) adopts a domain randomization paradigm: instead of relying on a costly physical modeling, it generates anatomically plausible but visually unrealistic synthetic MRIs with fully randomized contrasts and artifacts (Tobin et al., 2017) using Gaussian mixture models. The resulting diversity allows to train models that become contrast-agnostic and robust to real-world variability. `SynthSeg` has been widely used across various MRI tasks, including brain segmentation, skull stripping, and anatomical representation extraction of healthy and pathological subjects (Omidi et al., 2024; Kelley et al., 2024; Liu et al., 2024b,a; Shang et al., 2022; Valabregue et al., 2024; Gopinath et al., 2024). Importantly, because `SynthSeg` relies only on anatomical annotations rather than intensity images, the generation process is largely independent of the source imaging domain, aside from anatomical or population differences.

Most synthetic-data-based methods, however, target adult or infant brains. Fetal MRI introduces unique challenges such as a fast-changing morphology across gestation, limited tissue contrast, fewer label classes than in adult imaging, strong motion and artifacts, making realistic simulation and robust segmentation of these data particularly challenging.

2.4 Fetal brain tissue segmentation

The FeTA Challenge, organized at MICCAI, has been instrumental in benchmarking fetal brain segmentation algorithms (Payette et al., 2024; Zalevskyi et al., 2025). The 2024 edition emphasized the importance of data augmentation and topology-enhancing post-processing. Top-performing methods used ensembles of 3D architectures such as nnU-Net, and incorporated synthetic or augmented data like `SynthSeg` (Billot et al., 2023a) and GIN (Ouyang et al., 2022). The winning team in 2024, `cesne-digair`, achieved a mean Dice of 0.816 and 95th percentile Hausdorff Distance (HD95) of 2.317, using a 3D U-Net trained on skull-stripped, affine-registered T2w images with handcrafted augmentations and synthetic registration-based sample generation. A denoising autoencoder was then used to correct segmentation artifacts.

Using FeTA's public dataset, Huang et al. (2023) proposed a hybrid convolution-transformer achieving 0.837 ± 0.03 Dice on a held-out *in-domain* validation set. Beyond the FeTA dataset, BOUNTI (Uus et al., 2023) combined U-Net and Attention U-Net trained on 380 fetal MRIs (public and private), reaching 0.89 ± 0.02 Dice, thanks to extensive manual refinement of ground truth segmentations that contributed to the model's performance. However, its reliance on private datasets and custom annotation protocols limits its comparability with FeTA-trained models.

These studies show that high performance can be achieved

Table 1: Dataset properties. N_n – number of neurotypical subjects, N_p – number of pathological subjects. SR algorithms are MIALSRTK (MIAL (Tourbier et al., 2020)), IRTK (Kuklisova-Murgasova et al., 2012), NiftyMIC (Ebner et al., 2020), and SVRTK (Uus et al., 2022)

Site	**Scanner**	**Acquisition Parameters**	**SR algorithm**	**Resolution (mm^3)**	**GA (weeks)**	N_n/N_p
KISPI (public)	GE Signa Discovery MR450/MR750 (1.5T/3T)	SS-FSE TR/TE: 2500–3500/120 ms $0.5 \times 0.5 \times 3.5$ mm^3	MIAL IRTK	$0.5 \times 0.5 \times 0.5$ $0.5 \times 0.5 \times 0.5$	20–34 20–34	25/15 24/16
VIEN (private)	Philips Ingenia/Intera (1.5T) Philips Achieva (3T)*	SS-FSR, TR/TE: 6000–22000/80–140 ms	NiftyMIC	$1.0 \times 1.0 \times 1.0$	19–35	33/7
CHUV (private)	Siemens MAGNETOM Aera (1.5T)	HASTE, TR/TE: 1200/90 ms $1.13 \times 1.13 \times 3\ mm^3$	MIAL	$1.1 \times 1.1 \times 1.1$	21–35	25/15
KCL (private)	Siemens MAGNETOM FREE.MAX (0.55T)	HASTE, TR/TE: 2600/106 ms $1.45 \times 1.45 \times 4.5\ mm^3$	SVRTK	$0.8 \times 0.8 \times 0.8$	21–35	15/5
dHCP (T2w) (public)	Philips Achieva (3T)	MB-TSE, TR/TE: 2265/250 ms $1.1 \times 1.1 \times 2.2\ mm^3$	IRTK	$0.8 \times 0.8 \times 0.8$	21–38	248/0
dHCP (T1w) (public)	Philips Achieva (3T)	bSSFP, TR/TE: 3.6/7.2 (8479*) ms $1.5 \times 1.5 \times 4\ mm^3$	IRTK	$0.8 \times 0.8 \times 0.8$	21–38	208/0

*Slice package duration (effective TR for slab selective IR pulse)

through extensive data augmentation, the inclusion of diverse multi-site data, and expert-curated labels. Yet, such resources are not always available. In contrast, our work investigates a more common and resource-constrained scenario: **how to train fetal brain segmentation models that generalize well when using only public data from a single-domain with standard annotation quality**. We aim to characterize the limitations of this setting and propose practical strategies that enhance generalization in data-scarce environments.

3. Methods

3.1 Data

We use open-access data from the FeTA challenge (Payette et al., 2023) and the developing Human Connectome Project (dHCP) (Edwards et al., 2022), complemented with private clinical datasets from three institutions (Table 1). All acquisitions were approved by local ethics committees and performed without maternal or fetal sedation. For all datasets, several 2D single-shot fast spin-echo stacks were acquired in at least three orthogonal orientations (axial, coronal, sagittal) to mitigate motion artifacts (Table 1, *Acquisitions parameters*). A super-resolution reconstruction algorithm then combined these stacks into a high-resolution 3D volume suitable for volumetric segmentation.

FeTA Challenge Datasets. We use subsets of both training and testing datasets employed in the FeTA 2024 Challenge. Acquisition protocols and site-specific characteristics for KISPI, VIEN, CHUV, and KCL are detailed in Table 1 and in Zalevskyi et al. (2024); Payette et al. (2023). Among these, only the KISPI dataset is publicly available on Synapse (Payette and Jakab, 2021). In contrast, CHUV and KCL were used as held-out test datasets in the FeTA challenge and were not shared with participants, serving exclusively for final evaluation. Although VIEN was part of the challenge training set, its use is restricted to the scope of the challenge, and it is therefore not publicly available. Consequently, we refer to CHUV, KCL, and VIEN as private datasets.

Fetal dHCP Dataset. The fetal dHCP dataset, acquired at St Thomas' Hospital, London, includes T2w and T1w structural images obtained on a Philips Achieva 3T scanner with a 32-channel cardiac coil (Price et al., 2019; Karolis et al., 2025). T1w data were acquired using eight stacks across six unique orientations with an inversion-recovery sequence and interleaved wide-slab preparation pulses, followed by a bSSFP readout (in-plane resolution 1.5×1.5 mm^2, slice thickness 4 mm, slice gap -1.2 mm, flip angle 35°, total scan ∼8 min). T2w data were acquired using a zoomed multiband single-shot TSE sequence (in-plane resolution 1.1×1.1 mm^2, slice thickness 2.2 mm, slice gap -1.1 mm, flip angle 30° or 130°, total scan ∼12 min). All images were reconstructed to 0.5 mm isotropic using the fetal branch of the dHCP structural pipeline (Makropoulos et al., 2017; Cordero-Grande et al., 2022).

Prospective multi-echo KCL dataset Data were prospectively acquired for this study from five fetuses (GA = 24, 32, 32, 33, 37 weeks) using SST2w sequences at St. Thomas' Hospital, London, on a Siemens FreeMax 0.55T scanner, with similar parameters as the KCL dataset in Table 1, except for echo times (TEs), with data acquired at 300 ms, 397 ms, 600 ms, in order to do T2 mapping. At each echo time, three stacks (axial, coronal, sagittal) were acquired. The study was conducted under the ethically approved MEERKAT [REC: 21/LO/0742], MiBirth [REC: 23/LO/0685], and NANO [REC: 22/YH/0210] projects. Data sharing was approved by the Ethics Committee London Bromley (Ethics code 21/LO/0742). The data were then reconstructed using the method of Bulut et al. (2025), which yielded a volume at each TE.

Pre-processing and labeling. All structural images used for training and inference were resampled to 0.5 mm isotropic resolution and standardized to a size of 256^3 voxels via cropping or zero-padding. For FeTA datasets, we used the original FeTA labels comprising seven classes: cerebrospinal fluid (CSF), white matter (WM), gray matter (GM), subcortical gray matter (SGM), ventricles (LV), brainstem (BSM), and cerebellum (CBM) (Payette et al., 2021). For dHCP, we remapped Draw-EM annotations (Makropoulos et al., 2018, 2017) to the FeTA scheme using: 1→CSF, 2→GM, 3→WM, 4→background, 5→LV, 6→CBM, 7→SGM, 8→BSM, 9→WM. Thus, hippocampi and amygdala (9) are merged with WM, and the skull (4) with the background. Draw-EM and FeTA differ slightly in anatomical definitions (e.g., ventricles coverage), but all final datasets share a unified seven-tissue labeling consistent with FeTA.

3.2 Synthetic Data Generation Frameworks

We start by testing Hypothesis 1 by comparing various synthetic data generation frameworks, to assess how physics-based and domain randomization-based frameworks behave as data augmentation strategies. Our experimental framework builds on three components: `SynthSeg` (Billot et al., 2021), the `FaBiAN` numerical phantom (Lajous et al., 2022), and `FetalSynthSeg`, our fetal-tailored extension of `SynthSeg` (Zalevskyi et al., 2024). These components isolate how different intensity generation strategies impact single-source domain generalization. Note that these components can generate images that appear unrealistic, as illustrated in Figure 2. Exaggerated anatomies or non-physical contrasts, artifacts, and resolutions are an intentional feature of domain randomization, and the value of these generated images will be assessed through the downstream segmentation performance of models using them throughout training, not through their visual appearance. Overall, for each synthetic data generation approach, we follow a two-stage process: (1) generate synthetic images using different intensity simulation strategies, and (2) train a standard 3D U-Net segmentation network on these images.

3.2.1 SynthSeg

`SynthSeg` generates synthetic MR images directly from label maps using domain randomization (Billot et al., 2021). Given segmentations $\{S_n\}$, a segmentation is randomly sampled, spatially transformed by ϕ (affine and non-linear diffeomorphic transforms), and intensities at voxel (x, y, z) are drawn from a Gaussian mixture model (GMM) with label-specific parameters:

$$G(x, y, z) \sim \mathcal{N}(\mu_{L(x,y,z)}, \sigma^2_{L(x,y,z)}),$$

where $\mu_L \sim \mathcal{U}(a_\mu, b_\mu)$ and $\sigma_L \sim \mathcal{U}(a_\sigma, b_\sigma)$. The image is then corrupted by adding a bias field, intensity transformations and by simulating various image resolutions. By sampling a broad range of appearances beyond what is realistic, `SynthSeg` promotes robustness to the severe domain shifts faced in practice.

3.2.2 FetalSynthSeg

FeTA annotations contain only seven tissue classes, which is coarse compared to adult segmentations with 30 classes used in `SynthSeg`. This limits the ability to reproduce heterogeneous fetal tissue appearances, particularly within developing WM (Lajous et al., 2024a). `FetalSynthSeg` addresses this by augmenting the label space via unsupervised intensity-based clustering. Following Zalevskyi et al. (2024), we first group labels into four meta-labels (or meta-classes): **WM** (WM, CBM, BSM), **GM** (cortical and deep GM), **CSF** (ventricles and external CSF), **non-brain** (skull, uterus, fetal body, maternal tissue). This grouping fuses labels for structures that have similar macro-structural composition and hence have similar intensity profiles on the structural images during the late gestation period. All non-zero voxels are assigned to one of these meta-labels, allowing realistic modeling of both brain and surrounding anatomy as encountered in fetal pipelines (Figure 3).

Each meta-label is then **randomly and independently partitioned** into a number of subclasses sampled from $\mathcal{U}(a_{\text{subcl}}, b_{\text{subcl}})$. Sub-clustering is performed with an Expectation–Maximization algorithm conditioned on the input intensities (Dempster et al., 1977) (Fig. 3). Each subclass is assigned its own GMM parameters, introducing controlled intra-class variability. Importantly, intensity-based clustering itself is not new and has previously been used in domain-randomisation approaches such as SynthSeg (Billot et al., 2021). In contrast to prior work that performs clustering within existing anatomical labels, our approach performs clustering within meta-labels that group anatomically related tissues and background structures. This allows subclasses to span boundaries between structures with sim-

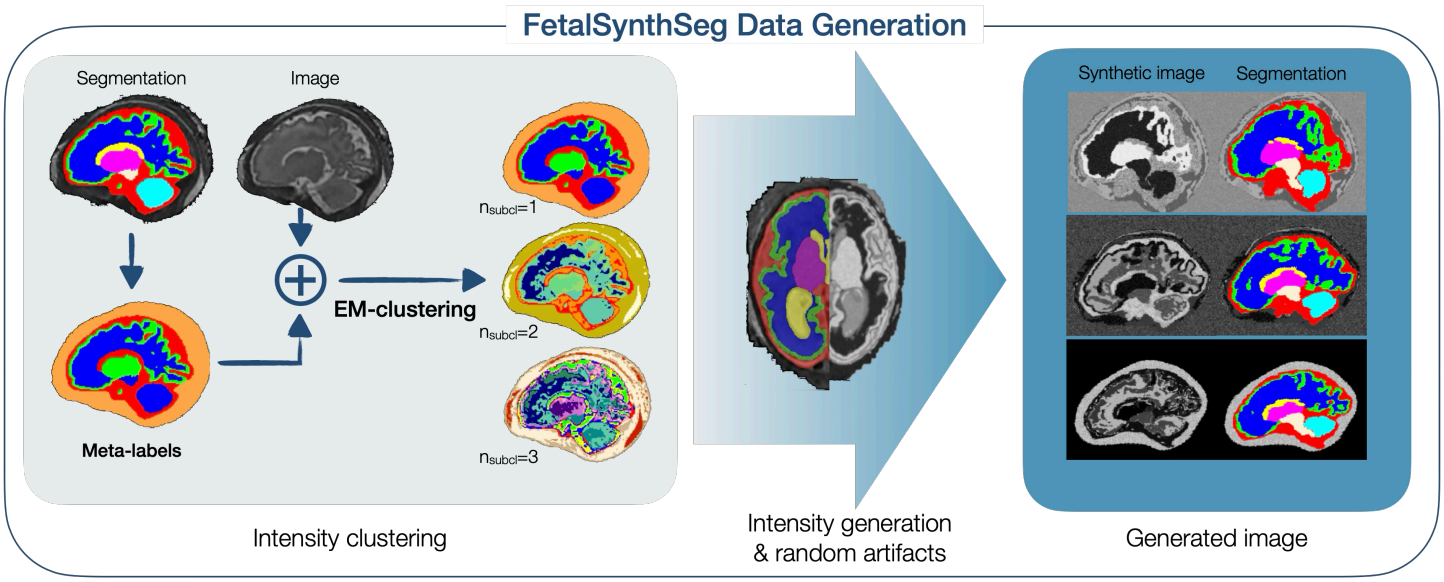


Figure 3: **FetalSynthSeg's pipeline for meta-label splitting into sub-classes used for training image generation.** Original segmentation labels are merged into four meta-labels (WM, GM, CSF, non-brain tissue). Then, each meta-label is randomly split into sub-classes based on EM-clustering using the original image. Finally, an independent GMM is sampled for each of the sub-classes and used to sample intensities for voxels inside of it. Synthesized training images are then fed into the U-Net. During testing only real images are used for inference.

ilar intensity profiles (e.g., WM and CBM) while preserving realistic contrast transitions. In the fetal setting, where annotations are coarse and tissue appearance is highly heterogeneous, this strategy enables more flexible modelling of intensity variability without introducing artificial boundaries.

Implementation. SynthSeg and FetalSynthSeg are implemented using the PyTorch SynthSeg generator from Brain-ID (Liu et al., 2024a), with hyperparameters kept identical to the original implementation. Our generator and modifications are available at: `https://github.com/Medical-Image-Analysis-Laboratory/fetalsyngen`.

3.2.3 FaBiAN and randFaBiAN

FaBiAN is an open-source numerical phantom that simulates clinical T2w FSE fetal brain images using tissue segmentations and the extended phase graph (EPG) formalism (Lajous et al., 2022). It models signal evolution based on literature T1/T2 values and sequence parameters, generating WM, GM, and CSF contrasts based on the physical tissue decay properties.

We adapt FaBiAN to include the non-brain meta-class consistent with FetalSynthSeg, and draw T1/T2 values from reference distributions before physical simulation, introducing controlled variability. We further define randFaBiAN, where all T1/T2 values are sampled from a broad, unconstrained range, making it more strongly randomized and conceptually closer to FetalSynthSeg.

Implementation. We use the Matlab implementation of FaBiAN_v2 (Lajous et al., 2024a,b) via a Python wrapper: `https://github.com/Medical-Image-Analysis-Laboratory/fabian_utils`. Examples for all generators are shown in Fig. 2, and details of the parameters used are available in the Supplementary Table S3.

3.3 Baselines: Real Image Training and Hybrid Training

After evaluating the data generation process, we move onto Hypothesis 2, where we aim to benchmark the ID and OOD performance of domain-randomization-based models. To quantify the value of randomized intensity simulation over conventional model training, we train two baseline models:

1. FetalRealSeg is trained exclusively on real images using the same augmentation pipeline as synthetic-data models. FetalRealSeg uses precisely the real scans whose segmentations serve as anatomical priors for the synthetic generators. Apart from the absence of randomized intensities, all transformations and training settings are identical, enabling a controlled comparison between purely real-data training and synthetic-data-driven approaches.

2. RealSynthHybrid is trained using a mixture of 50% synthetic and 50% real data, and is essentially a hybrid between FetalSynthSeg and FetalRealSeg, randomly generating either a random synthetic sample or a data-augmented intensity image, each with 50% probability.

3.4 Experimental Design

We conduct four main experiments:

I. Selecting an intensity generation strategy (4.1) We benchmark `FetalRealSeg`, `FaBiAN`, `randFaBiAN`, `SynthSeg`, and `FetalSynthSeg` in a single-source domain generalization setting. Due to the computational cost of `FaBiAN`, this experiment uses three domains: KISPI-IRTK (n=40), KISPI-MIAL (n=40), and CHUV-MIAL (n=40). We train on one domain and test on the remaining two, covering mild (site or SR-only) and severe (site+SR) domain shifts. The goal is to identify which generator yields the most robust out-of-domain performance.

II. Impact of intensity clustering (4.2). We perform an ablation on the impact of *meta-label fusion and EM-clustering* by varying the number of EM clusters in `SynthSeg` and `FetalSynthSeg`. Using the splits from Experiment I, we assess how sub-clustering impacts performance and whether the `FetalSynthSeg` design consistently improves robustness.

III. Comparison with state-of-the-art models with failure mode analysis (4.3). We retrain `FetalSynthSeg`, `FetalRealSeg` and `RealSynthHybrid` on the full FeTA 2024 training set (n=120) and compare them against:

1. the FeTA 2024 winning method `cesne-digair` (*FeTA24*) (Zalevskyi et al., 2025)
2. BOUNTI (Uus et al., 2023),
3. an nnU-Net ensemble (Isensee et al., 2021).

All models are evaluated on 348 subjects from the dHCP dataset and a subset of FeTA 2024 test data, spanning T1w/T2w, 0.55T/1.5T/3T, and both pathological and neurotypical cases. We use the default implementations and publicly available weights for the FeTA 2024 challenge winner[1] and the nnU-Net model [2], both trained on the full FeTA 2024 training dataset (n = 120). For BOUNTI, we use the publicly available Docker image [3] with weights trained on a larger collection of private and public datasets unrelated to the FeTA 2024 Challenge (n = 360).

IV. Prospective validation (4.4). Finally, we showcase the abilities of `FetalSynthSeg` to tackle challenging data by validating its ability to consistently segment prospectively acquired data at different echo times (TEs) (Bulut et al., 2025).

3.5 Model Architecture and Training

All models use the same 3D U-Net backbone, implemented in PyTorch with MONAI (Paszke et al., 2019; Cardoso et al., 2022). The network has five levels with 32 feature maps at the first level (doubling per level), 3^3 convolutions with LeakyReLU activations, skip connections, and a softmax output layer. We deliberately use a simple, consistent architecture to focus the analysis on data generation strategies rather than architectural optimizations.

We train with Adam (learning rate 10^{-3}) and a combined Dice + cross-entropy loss (Valabregue et al., 2024). A `ReduceLROnPlateau` scheduler (factor 0.1, patience 10 epochs) and `EarlyStopping` (patience 100 iterations) are applied. Batch size is 1. Training is implemented with PyTorch Lightning and run on NVIDIA RTX 3090/6000 GPUs.

For `SynthSeg` and `FetalSynthSeg`, synthetic images are generated on-the-fly for up to 80,000 iterations. For `FaBiAN` and `randFaBiAN`, we pre-generate 50 synthetic images per real scan (n=120; 6000 images total) and train for up to 80,000 iterations. Online generation with `FaBiAN` is infeasible (∼287 s per volume) compared to ∼1 s for `SynthSeg`/`FetalSynthSeg`. For `FetalRealSeg` and `RealSynthHybrid`, we train on for up to 500 epochs to limit overfitting.

Data augmentation. All models use the `SynthSeg` augmentation pipeline: random non-linear deformations, gamma-based contrast changes, Gaussian noise, bias field perturbations, and random isotropic resampling to emulate varying resolutions.

3.6 Evaluation Protocol

Data splitting. For Experiment 1, we perform cross-domain evaluation by training on one of KISPI-IRTK, KISPI-MIAL, or CHUV-MIAL (40 scans each) and testing on the other two domains. For each training domain, 35 scans are used for training and 5 for validation; all 40 scans of each held-out domain are used for testing. We report per-domain and pooled performance. Experiment 2 reports results broken down per split, while Experiment 3 relies on standard train/validation/test split.

Metrics. We evaluate segmentations using:

- **Dice Similarity Coefficient (DSC):**

$$DSC = \frac{2|A \cap B|}{|A| + |B|},$$

with A and B the predicted and ground-truth voxel sets.

- **95th-percentile Hausdorff Distance (HD95):**

$$HD95 = \max\left(\max_{x\in A}\min_{y\in B}\|x-y\|, \max_{y\in B}\min_{x\in A}\|x-y\|\right),$$

with A and B boundary point sets.

1. https://hub.docker.com/u/fetachallenge2024
2. https://github.com/mic-dkfz/nnunet
3. https://hub.docker.com/r/fetalsvrtk/segmentation

Table 2: Mean Dice scores (multiplied by 100 for readability) for models trained on different data sources, evaluated across multiple testing splits. Each cell reports the mean and standard deviation across test subjects. **Bold** indicates the best-performing method per column, and underlined denotes the second-best. An asterisk (*) marks cases where the best method is statistically significantly better than the second best according to a Wilcoxon rank-sum test ($p < 0.05$).

Testing split	**CHUV-MIAL**		**KISPI-IRTK**		**KISPI-MIAL**		Global
Training split	KISPI-IRTK	KISPI-MIAL	CHUV-MIAL	KISPI-MIAL	CHUV-MIAL	KISPI-IRTK	
`FaBiAN`	74.2±4.2	73.1±5.7	53.6±13.3	56.6±12.9	60.6±17.1	61.5±20.1	63.3±15.5
`randFaBiAN`	79.1±2.3	78.2±2.9	55.1±12.7	68.9±7.8	60.7±7.3	68.1±14.7	68.3±13.8
`SynthSeg`	75.9±3.9	73.7±3.9	70.9±9.2	74.8±7.8	60.5±15.7	63.4±16.8	72.2±13.0
`FetalSynthSeg`	**80.7±2.0***	76.9±3.3	**79.2±9.0***	**76.8±6.9***	67.5±16.0	**68.5±15.6**	**74.9±11.5***
`FetalRealSeg`	77.2±4.0	**78.5±3.3***	70.6±13.9	71.9±11.5	**68.0±19.3**	64.3±19.1	73.0±12.6

When not stated otherwise, the metrics that we report are first averaged across the 7 labels for each subject, and then averaged across all subjects in experiments where we compare different models.

Statistical analysis. Normality is assessed with the Shapiro–Wilk test. Depending on distribution, we use either a paired t-test or Wilcoxon rank-sum test with Bonferroni correction. Results with $p < 0.05$ are considered statistically significant.

4. Results

4.1 Selecting an intensity generation strategy

A global comparison of all intensity generation methods is reported in Table 2. Across all cross-validation splits, `FetalSynthSeg` achieves the highest global average Dice score of 74.9, consistently outperforming all other simulation-based approaches. Across individual splits, the only significant exception is the KISPI-mial→CHUV-mial split, where `FetalRealSeg` outperforms `FetalSynthSeg`, suggesting that training on real images can provide advantages under certain domain shift conditions (same SRR method in this experiment). Overall, `FetalRealSeg` remains the second-best performing model, with an average Dice of 73.0, indicating that synthetic data generation alone does not necessarily guarantee improved generalization, and models trained on real images may remain competitive when the domain gap between training and testing data is relatively limited.

Focusing on synthetic generation methods, the progression across generation strategies highlights the impact of increasing intensity variability. `FaBiAN`, which uses tissue-specific relaxometry values with realistic physics-based modeling, yields the lowest mean Dice (63.3). Randomizing T1/T2 within `randFaBiAN` improves performance by +5.0 (68.3). Replacing physics-based simulation with GMM-based sampling in `SynthSeg` leads to a further increase of +3.9 (72.2). Incorporating meta-classes and intensity sub-clustering in `FetalSynthSeg` brings an additional +2.7 gain, reaching a mean Dice of 74.9.

These differences are not solely explained by training set size. Supplementary Experiment S2 shows that, even when matched to `FaBiAN` with 6,000 synthetic samples, `FetalSynthSeg` still improves by 3.9 Dice points. Moreover, `FaBiAN` 's physics-based simulation is computationally demanding (∼280s per volume), whereas the GMM-based sampling of `SynthSeg` and `FetalSynthSeg` requires ∼1s, enabling efficient on-the-fly generation during training.

Qualitative results in Fig. 4 corroborate the quantitative findings: methods with stronger contrast randomization (`randFaBiAN`, `FetalSynthSeg`) are noticeably more robust to changes in reconstruction and skull-stripping quality than approaches relying solely on realistic or real intensities.

Per-label Dice and HD95 scores are provided in Supplementary Tables S4 and S5. While HD95 differences between `SynthSeg` and `FetalSynthSeg` are small (mean HD95 3.15 ± 0.52 vs. 3.16 ± 0.52), all domain-randomized models outperform purely real-data or strictly physics-based approaches, confirming that randomized synthetic intensities are key for robust, cross-domain fetal brain segmentation.

4.2 Impact of intensity clustering

Figure 5 summarizes the effect of intensity sub-clustering on `SynthSeg` and `FetalSynthSeg`. Intensity-clustering is done on the seven tissue classes for `SynthSeg` and on the four meta-labels for `FetalSynthSeg`. Starting from the baseline configuration (a single class per label), gradually increasing the number of subclasses consistently improves Dice scores for both methods. Performance gains saturate beyond approximately six subclasses, suggesting an optimal range of sub-cluster granularity.

Across most configurations, `FetalSynthSeg` maintains a systematic advantage over `SynthSeg`, demonstrating that meta-label fusion and sub-clustering act synergistically. Over the four tested sub-cluster settings, `FetalSynthSeg`

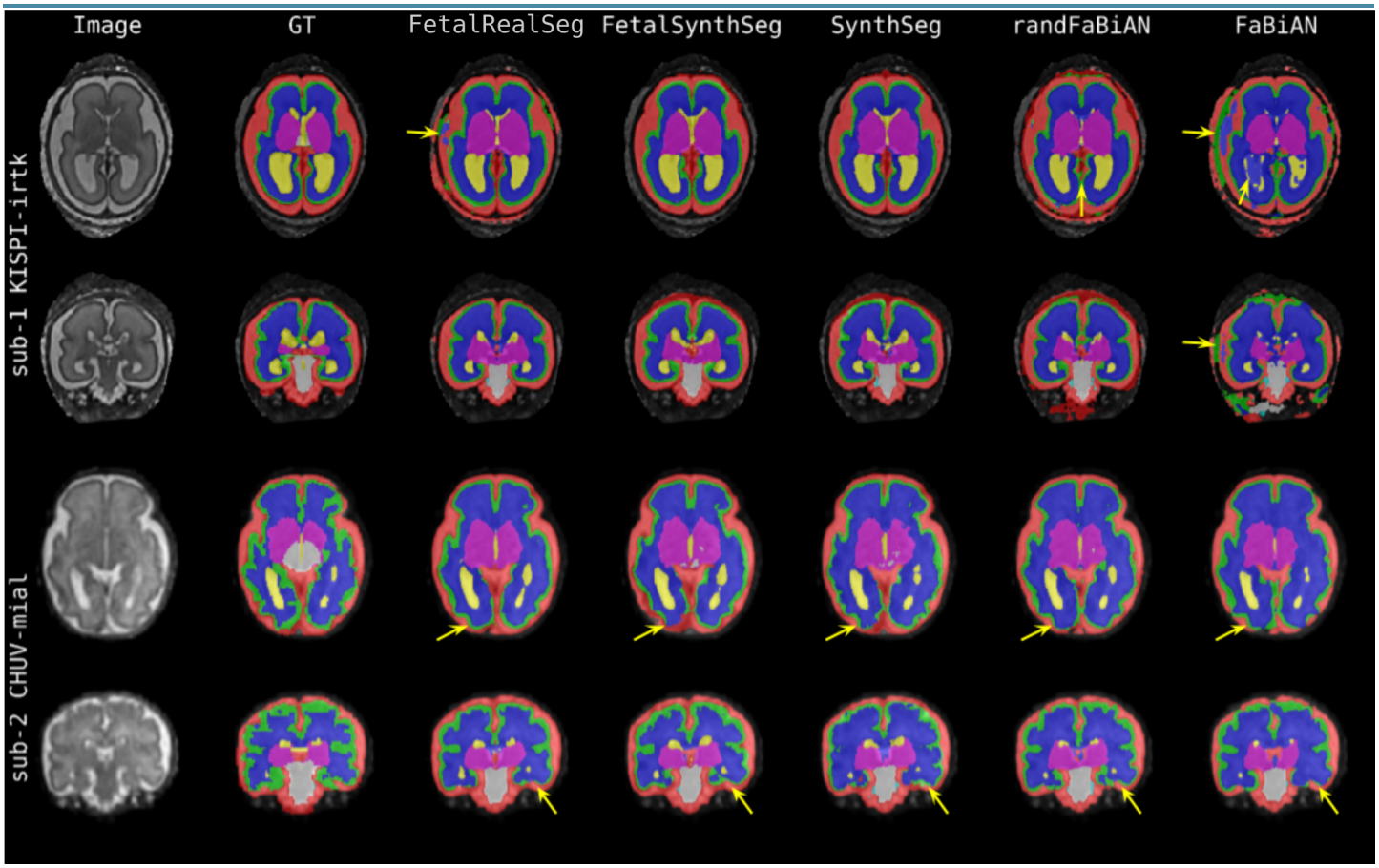


Figure 4: Segmentation results for models trained on the KISPI-MIAL data split, evaluated on a subject from the KISPI-IRTK split and on another from the CHUV-MIAL split. Arrows point to regions of substantial discrepancies across models.

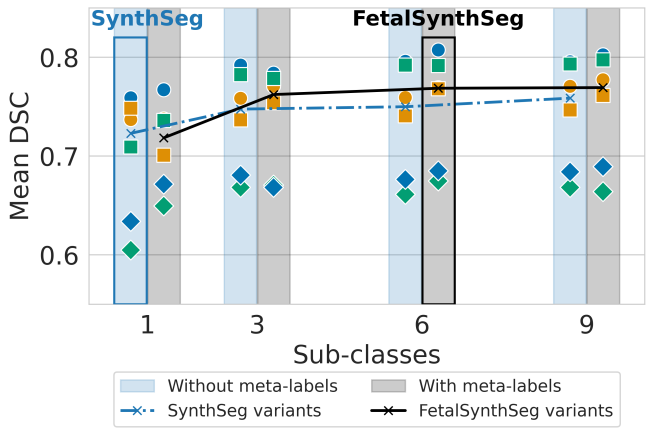


Figure 5: **Ablation study on the number of EM sub-clusters.** For `SynthSeg`-variants, the original 7 labels used as generation classes are randomly split into 1-9 subclasses, while in `FetalSynthSeg`-variants the meta-labels are split. The blue and black rectangles show respectively what we refer to as `SynthSeg` (one sub-class with no meta-labels) and `FetalSynthSeg` (six sub-classes with meta-labels) in the paper. We report mean Dice across all tissues for each combination of training-testing splits. The different labels denote these splits:
● CHUV-MIAL→KISPI-IRTK ● CHUV-MIAL→KISPI-MIAL
■ KISPI-MIAL→KISPI-IRTK ■ KISPI-MIAL→CHUV-MIAL
◆ KISPI-IRTK→KISPI-MIAL ◆ KISPI-IRTK→CHUV-MIAL

outperforms `SynthSeg` in 7/12 train–test splits (p-value < 0.05), `SynthSeg` is superior in 3/12, and the remaining 2/12 show no significant difference. This indicates that enriching the label space with structured yet randomized subclasses is a key ingredient for strong SSDG performance.

Furthermore, based on our experiments, performance improvements from additional subclass splitting plateau at six subclasses, so we use this setting for all following experiments involving `FetalSynthSeg`.

4.3 Comparison with state-of-the-art models

Table 3 compares `FetalSynthSeg`, `FetalRealSeg` and `RealSynthHybrid` (trained on the full FeTA 2024 training set) against BOUNTI, the FeTA24 winning method (`cesne-digair`), and an nnU-Net ensemble. On T2w datasets that are close to the training distribution (e.g., CHUV_{T2w}; VIEN_{T2w} sharing both scanner/SRR; KCL with SVRTK being similar to IRTK), `FetalRealSeg` and `RealSynthHybrid` remain slightly ahead of `FetalSynthSeg`, with differences typically within 0.1–2.5 Dice points. `FetalSynthSeg`, `RealSynthHybrid` and `FetalRealSeg` are competitive with or superior to existing SoTA models in these settings: for instance, on CHUV, they surpass all reference methods.

Table 3: Dice performance (in %) and surface distance (in mm) across datasets for each model. Values are mean± standard deviation. Bold indicates best and underlined second-best performance per column. An asterisk (*) marks cases where the best method is statistically significantly better than the second-best method according to a Wilcoxon rank-sum test ($p < 0.05$).

Model	DSC					HD95				
	CHUV$_{T2w}$	KCL$_{T2w}$	VIEN$_{T2w}$	dHCP$_{T2w}$	dHCP$_{\mathbf{T1w}}$	CHUV$_{T2w}$	KCL$_{T2w}$	VIEN$_{T2w}$	dHCP$_{T2w}$	dHCP$_{\mathbf{T1w}}$
BOUNTI	79.9±2.2	83.4±1.1	70.1±12.3	**85.1±1.3***	16.4±2.1	3.15±0.46	2.62±0.27	4.51±2.58	**3.56±0.42***	15.86±3.36
FeTA24	83.0±1.8	85.6±0.9	81.4±4.1	82.4±1.5	23.3±3.8	2.49±0.42	1.82±0.27	2.47±0.86	4.07±0.58	10.35±2.03
nnU-Net	81.8±1.8	**86.7±1.2**	**83.6±3.4***	84.3±1.2	13.5±2.5	2.55±0.35	1.64±0.15	**2.07±0.54**	4.08±0.47	25.16±8.74
`FetalRealSeg`	**84.6±1.8**	86.5±1.1	82.0±3.3	84.1±1.5	43.7±7.0	**1.79±0.23**	**1.55±0.16***	2.11±0.50	4.11±0.65	9.10±4.08
`RealSynthHybrid`	84.2±1.5	85.9±1.2	80.1±4.3	83.1±1.7	78.1±2.5	1.84±0.22	1.76±0.26	2.46±0.90	4.32±0.55	4.42±0.85
`FetalSynthSeg`	83.2±1.7	85.3±1.2	79.5±4.1	84.0±2.1	**80.1±2.1***	1.89±0.22	1.76±0.25	2.17±0.40	4.08±0.50	**4.27±0.53***

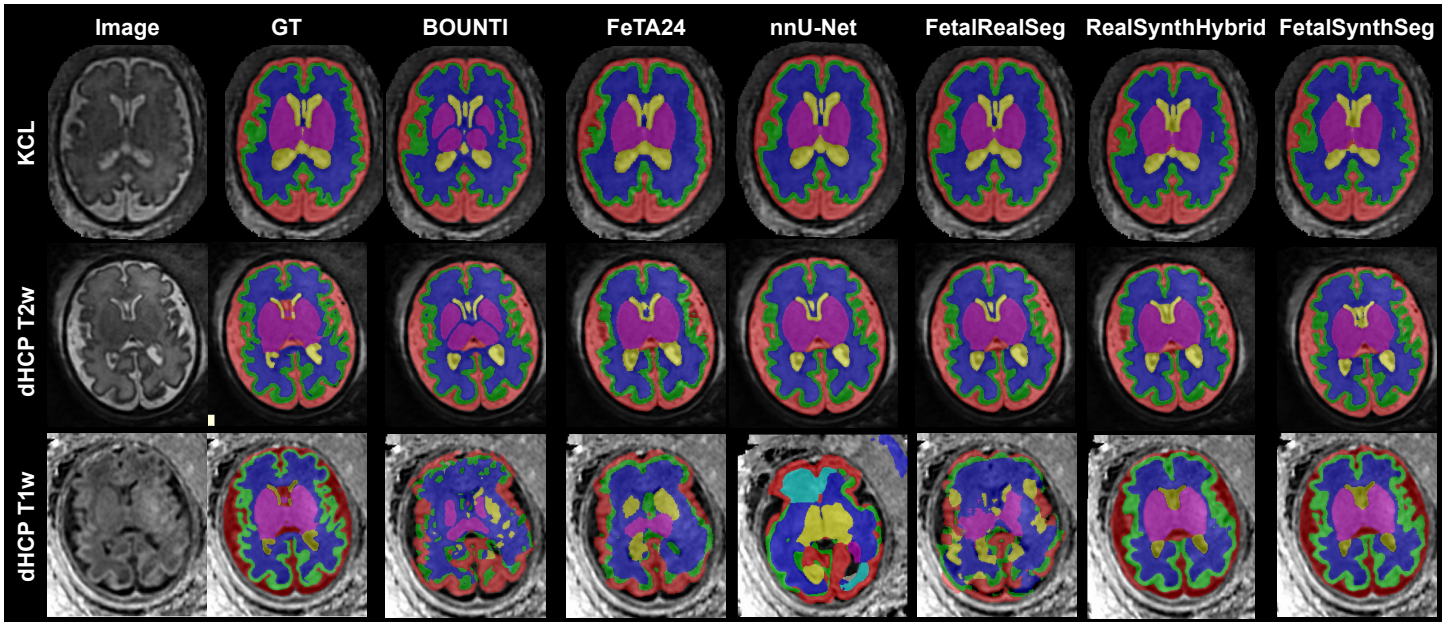


Figure 6: Qualitative comparison of the proposed models and `FeTA24` on a test case from the dHCP dataset, showing T2w images (top two rows) and T1w images (bottom two rows) from the same subject. White arrows highlight discrepancies between the dHCP and FeTA annotation schemes—particularly in the definitions of SGM and LV —as all models were trained using FeTA annotations, while the ground truth reflects dHCP labels converted to FeTA format. Yellow arrows indicate prominent segmentation errors made by the model fine-tuned on real T2w images when applied to the T1w contrast.

On dHCP T2w, BOUNTI attains the highest performance: this could be explained by a closer alignment between the dHCP and BOUNTI labels (the latter were refined based on the dHCP annotations) as well as by a data leakage problem, as BOUNTI's training data included some of the dHCP T2w data (though the exact data splits of dHCP used for BOUNTI training were not published). In contrast, `FetalRealSeg` and `FetalSynthSeg` reach performance comparable to nnU-Net without such overlap, highlighting the strength of the proposed training schemes. Additional HD95 and per-class results are provided in Table S6 in the Supplementary materials, which show that all methods exhibit similar relative performance patterns across labels, despite differing in their absolute metric values.

The most striking effect appears in the OOD evaluation on dHCP T1w. Here, `FetalSynthSeg` achieves a Dice of 80.1%, while all other methods, including `FetalRealSeg`, nnU-Net, BOUNTI, and FeTA24, drop below 25%. `FetalSynthSeg` is thus the only evaluated model capable of reliably segmenting this challenging modality using a single, contrast-agnostic network trained exclusively on synthetic data. Figure 6 illustrates these findings qualitatively: models trained on real T2w data struggle on T1w images, whereas `FetalSynthSeg` preserves anatomical plausibility and label consistency. Besides a quantitative evaluation, visual inspection also shows that `FetalSynthSeg` produces highly consistent segmentations across T1w and T2w images. On T1w images, a visual assessment shows that these

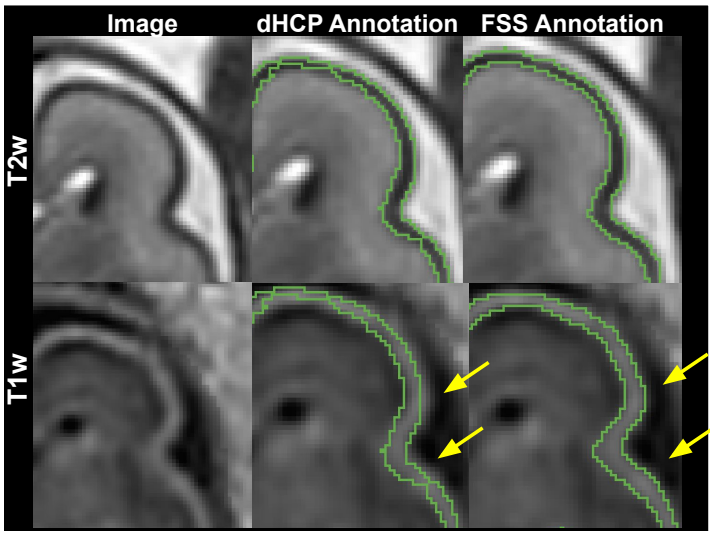


Figure 7: Cortical GM segmentations on T1w and T2w dHCP images. dHCP Annotation: reference annotations obtained via registration of T2w-based labels to T1w. Bottom: `FetalSynthSeg` predictions obtained independently for each modality using the same model, showing improved alignment with modality-specific tissue boundaries.

segmentations improve the quality of cortical gray matter segmentations compared to reference annotations obtained from propagating T2w labels on the T1w images, which are often imprecise, as illustrated on Figure 7.

Failure mode analysis

When analyzing the failure cases of `FetalSynthSeg`, three main categories emerge, as illustrated in Fig. 8.

First, **severely pathological cases** remain challenging for all methods. As shown in Fig. 9, segmentation performance degrades and variance increases for pathological subjects, particularly in cases with strong anatomical alterations (e.g., ventriculomegaly; Fig. 8, top row). Notably, both `FetalSynthSeg` and the FeTA 2024 model outperform BOUNTI in this regime. This can be attributed to the limited presence of extreme anatomical abnormalities in BOUNTI's training data (Uus et al., 2023), in contrast to the FeTA dataset, which contains a substantial proportion of pathological cases, including severe abnormalities such as spina bifida and ventriculomegaly.

Second, reduced performance can often be attributed to **imperfect ground-truth annotations**. Strong partial volume effect and small structure sizes make manual annotation particularly challenging, and automated methods may produce anatomically more consistent segmentations than the provided ground truth. This is reflected in our results approaching or even exceeding the inter-rater variability reported for the FeTA dataset (Zalevskyi et al., 2025).

Finally, `FetalSynthSeg` shows reduced performance in **older fetuses**, particularly in structures undergoing significant changes during late gestation, such as cortical gray matter. In these cases, segmentations tend to be overly smooth and deviate from the ground truth (Fig. 8, bottom row). This effect is especially pronounced in the dHCP dataset, which contains a higher proportion of older subjects that lie outside the gestational age range represented in the FeTA training data. Retraining the model with a

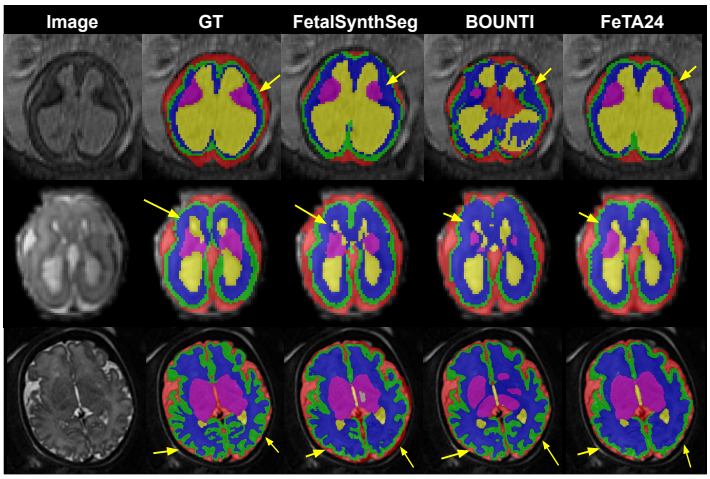


Figure 8: Representative failure cases of `FetalSynthSeg` compared to two state-of-the-art methods (BOUNTI and FeTA24). **Top row:** severely pathological case (ventriculomegaly), with errors highlighted in subcortical gray matter (yellow arrows). **Middle row:** very young subject with annotation errors due to challenging manual segmentation (yellow arrow indicates ventricle over-segmentation in the ground truth). **Bottom row:** older gestational age subject, where cortical gray matter segmentation is overly smooth (yellow arrows).

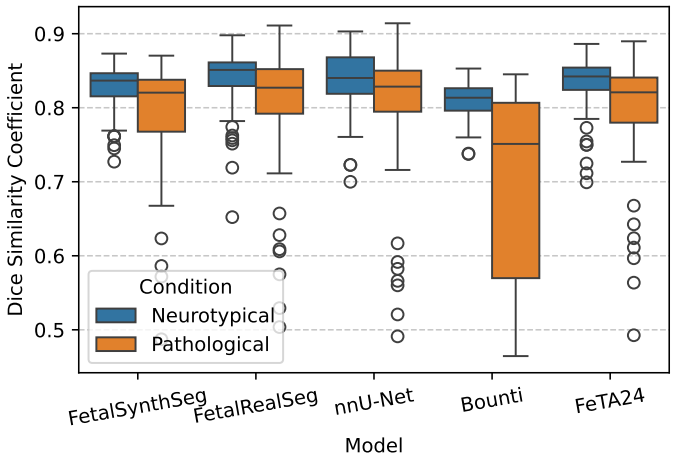


Figure 9: Dice score distributions of selected state-of-the-art methods on CHUV and KCL datasets, stratified by healthy and pathological subjects. Pathological cases show lower median performance and higher variability across all methods. All methods except for BOUNTI trained on the FeTA 2024 training data.

larger proportion of older subjects should however address this issue.

4.4 Prospective validation.

Finally, we carried out a prospective validation on multi-TE data acquired at 0.55 T at KCL, and reconstructed using the method of Bulut et al. (2025). The main challenge in segmenting these data lies in the contrast change across echo times, and algorithms often fail at higher echo times, where the tissue across contrast is low.

As shown in Figure 10, `FetalSynthSeg` delivers more consistent segmentations across echo times, especially at higher TEs, where BOUNTI and FeTA24 fail to segment the occipital lobe (subject 1), as well as the brainstem (especially the pons) and CSF around it (subject 2). This is confirmed by quantitative results: since no ground-truth annotations are available for these low-field acquisitions, we compute cross-TE metrics. As the Shapiro-Wilk test for normality does not reject the null hypothesis, we compare the results using a paired t-test. The results show that `FetalSynthSeg` achieves a DSC of 0.89±0.02 compared to 0.84±0.04 (BOUNTI; p=0.018) and 0.82±0.02 (FeTA24; p=0.038), while also obtaining a lower HD95 (9.18±5.20 vs. 12.5±0.30 and 13.20±2.57, respectively), although in this case the results are not statistically significant.

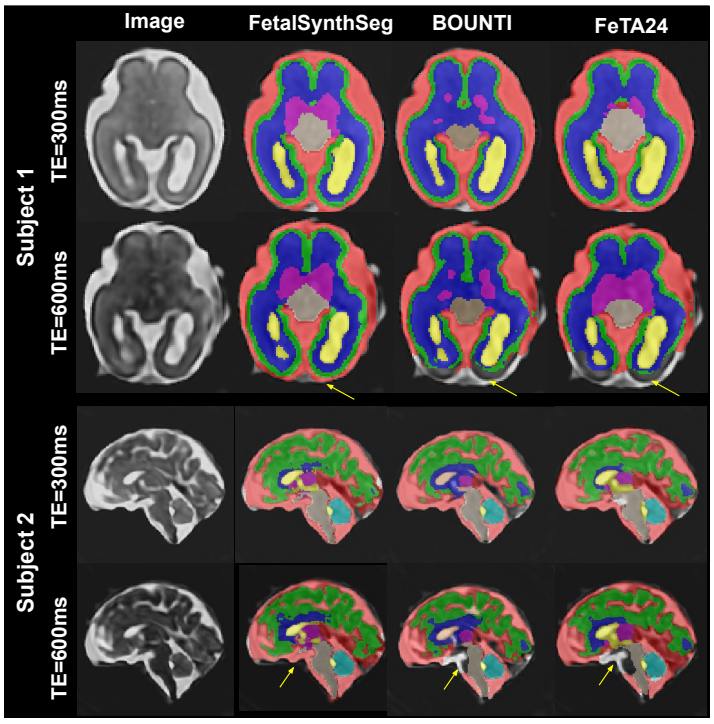


Figure 10: **Example of TE-invariant segmentation for two subjects.** Two rows show images of the same subject acquired at different TEs. Columns show segmentations from different methods. `FetalSynthSeg` provides more consistent segmentations across TEs, supporting reliable T2 mapping.

5. Discussion

In this work, we systematically analyzed single-source domain generalization for fetal brain tissue segmentation, building on `SynthSeg`-style domain randomization (Billot et al., 2021, 2023a). Using more than 300 subjects across multiple sites, field strengths, reconstruction methods, and modalities, we disentangled the roles of intensity simulation, clustering, meta-classes, and real-image training. We also confirmed both of our hypotheses: domain randomization is indeed more efficient that physics-based simulation for SSDG, and it enables strong OOD generalization while retaining a competitive ID performance. Below, we summarize the main findings.

(i) Intensity clustering is essential for strong domain randomization. Introducing intensity-based clustering within labels yielded an average improvement of ∼5 Dice points and was critical for enabling `SynthSeg`-based models to outperform, on average, models trained solely on real data. Clustering increases intra-class variability and better captures subtle fetal tissue heterogeneity, particularly when only few anatomical labels are available. Consistent with observations in cardiac MRI (Billot et al., 2023b) and recent extensions (Valabregue et al., 2024), our results confirm that structured intensity diversification is a key ingredient for robust SSDG.

(ii) Meta-classes improve the trade-off between realism and randomization. Aggregating labels into meta-classes (WM, GM, CSF, non-brain) and then sub-clustering them, as done in `FetalSynthSeg`, consistently provided a small but statistically significant improvement over `SynthSeg` (∼2 Dice points). This design balances flexibility and prior knowledge: tissues expected to share similar intensities (e.g., CSF in ventricles and extra-axial spaces) are modeled jointly rather than forced to diverge, avoiding unnecessary use of capacity and yielding more coherent yet still highly varied synthetic contrasts.

(iii) Physics-based simulators underperform randomized GMM-based generation for SSDG. Using `FaBiAN` (Lajous et al., 2022) with realistic T1/T2 values did not translate into strong cross-domain performance. Randomizing its parameters (`randFaBiAN`) improved results (to 68.4 Dice on average), but remained below `SynthSeg` (72.2 Dice) and `FetalSynthSeg` (74.9 Dice). Direct GMM-based sampling more easily spans the large diversity of real-world contrasts than perturbations of physically grounded parameters alone. These findings suggest that, for SSDG under strong domain shifts, broad and sometimes non-physical intensity randomization is more beneficial than faithful but constrained physical simulation.

(iv) Training with real data is sufficient for some ap-

plications. While FetalSynthSeg's performance remains strong in-domain, its generalization ability comes at a small price: its performance in-domain is often slightly below models trained using purely real data, as was noticed in Billot et al. (2023a). When evaluated on T2w datasets, FetalSynthSeg's DSC and HD95 is often slightly below FetalRealSeg, RealSynthHybrid and nnU-Net (the latter using an ensemble of models). The hybrid training of RealSynthHybrid provides another layer of trade-offs: it is weaker in-domain compared to FetalRealSeg, worse than FetalSynthSeg out-of-domain, but much more robust than FetalRealSeg out-of-domain. Nevertheless, for all cases, an essential component lies in the data augmentation strategies used to train FetalSynthSeg and FetalRealSeg (S1): they allow a stronger generalization than MONAI-based augmentations in every setting, which warrants further works to study these augmentations strategies more rigorously.

(v) FetalSynthSeg enables new applications in fetal brain MRI. FetalSynthSeg shows robust generalization across sites, SR methods, field strengths, and, importantly, contrasts. Trained exclusively on synthetic data, it:

1. Matches or closely approaches strong real-data baselines on in-domain T2w datasets;
2. Is, to our knowledge, the only evaluated model providing reliable segmentation on fetal dHCP T1w dataset (80% Dice, versus $< 25\%$ for all other methods);
3. Maintains consistency across varying echo times (TEs) in T2-mapping acquisitions. These properties allow FetalSynthSeg to act as a contrast-invariant backbone for: inter-modality registration support and QC, detection of label misregistrations (Fig. 7), and stable tissue delineation in quantitative mapping (Fig. 10).

We hypothesize that the performance gains achieved by FetalSynthSeg primarily arise from the proposed data generation and augmentation strategies, as the network architecture, loss function, and training procedure were kept unchanged throughout our experiments. In principle, these improvements should therefore be largely model-agnostic and transferable to other segmentation frameworks. Consequently, FetalSynthSeg data generation strategy could potentially be combined with alternative architectures, including more recent transformer-based models, ensembles, or different optimization and training strategies, which may further improve performance. However, validating this generalizability across architectures and training paradigms warrants further investigation.

Limitations and future work. Our experiments focus on fetal brain MRI only, a domain with pronounced shifts and limited labels. While this makes it an informative stress-test for SSDG, extrapolation to other anatomies or modalities should be done cautiously. Evaluations on additional public datasets and non-fetal applications are a natural next step.

Our work deliberately fixed a simple 3D U-Net backbone to isolate the effects of data generation. Architectural choices, ensembling, or topology-aware losses could interact with synthetic data strategies in non-trivial ways, and may further boost performance.

Another limitation relies on hand-crafted generators, GMM-based intensity sampling with user-defined ranges. More expressive generative models (e.g., conditional or diffusion-based approaches) could in principle learn richer priors over anatomy, artifacts, and contrasts. However, their computational cost and complexity currently hinder large-scale, on-the-fly generation (Wang et al., 2023), which is central to our efficiency-focused SSDG setting. Exploring hybrid schemes that combine learned generative models with lightweight randomization is an important direction for future work.

Although our study includes a diverse set of datasets, most are available only after site-specific SRR, due to privacy constraints and the lack of access to raw low-resolution data. As a result, each dataset is typically associated with a single SRR pipeline, introducing a potential confounding factor. This makes it difficult to disentangle the respective contributions of acquisition domain and reconstruction method to the observed segmentation performance. A natural direction for future work is to systematically reconstruct datasets using multiple SRR methods and explicitly evaluate their impact on segmentation methods.

6. Conclusions

We investigated how to best leverage synthetic data for fetal brain tissue segmentation under single-source domain generalization. Comparing physics-based, purely randomized, and hybrid generation strategies, we found that:

- Structured intensity clustering and meta-class design are crucial for effective domain randomization;
- Broad GMM-based randomization outperforms physically constrained simulations for handling strong domain shifts;

Guided by these insights, we introduced FetalSynthSeg, a SynthSeg-based framework tailored to fetal MRI. FetalSynthSeg attains competitive performance on in-domain T2w data and, notably, strong out-of-domain generalization across sites, reconstruction methods, and contrasts—including accurate segmentation of T1w fetal brain MRI using only synthetic training data.

Overall, our results demonstrate that carefully designed domain randomization, grounded in intensity clustering and meta-class modeling, provides a practical and efficient path toward contrast- and protocol-robust fetal brain segmentation in data-scarce settings.

Acknowledgments

This research was funded by the Swiss National Science Foundation (182602 and 215641), ERA-NET Neuron MULTIFACT project (SNSF 31NE30 203977), UKRI FLF (MR/T018119/1) and DFG Heisenberg funding (502024488); we acknowledge the Leenaards and Jeantet Foundations as well as CIBM Center for Biomedical Imaging, a Swiss research center of excellence founded and supported by CHUV, UNIL, EPFL, UNIGE and HUG. This research was also supported by grants from NVIDIA and utilized NVIDIA RTX6000 ADA GPUs.

The Developing Human Connectome Project (dHCP) was funded by the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007–2013), Grant Agreement No. 319456. The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust [Grant No. 203139/Z/16/Z]. The views expressed in this work are those of the authors and do not necessarily reflect those of the NHS, the NIHR, King's College London, or the Department of Health and Social Care. The authors gratefully acknowledge the families and participants involved in the dHCP, as well as the neonatal staff at the Evelina Newborn Imaging Centre, St Thomas' Hospital, Guy's & St Thomas' NHS Foundation Trust, London, UK, for their work in acquiring and processing the data.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

The author discloses receiving an NVIDIA Academic Hardware Grant, through which NVIDIA supplied GPU hardware that supported the experiments reported in this manuscript.

Data availability

The KISPI and dHCP are publicly available dataset (Payette and Jakab, 2021; Edwards et al., 2022). KCL and CHUV datasets used for evaluation in this study are part of the private testing set from the FeTA challenge and are not publicly accessible. VIEN data is not publicly available.

References

Yasmina Al Khalil, Sina Amirrajab, Cristian Lorenz, Jürgen Weese, Josien Pluim, and Marcel Breeuwer. On the usability of synthetic data for improving the robustness of deep learning-based segmentation of cardiac magnetic resonance images. *Medical Image Analysis*, 84:102688, 2023. ISSN 1361-8415. . URL https://www.sciencedirect.com/science/article/pii/S1361841522003164.

Jordina Aviles Verdera, Lisa Story, Megan Hall, Tom Finck, Alexia Egloff, Paul T. Seed, Shaihan J. Malik, Mary A. Rutherford, Joseph V. Hajnal, Raphaël Tomi-Tricot, and Jana Hutter. Reliability and feasibility of low-field-strength fetal mri at 0.55 t during pregnancy. *Radiology*, 309(1), October 2023. ISSN 1527-1315. . URL http://dx.doi.org/10.1148/radiol.223050.

Suryava Bhattacharya, Anthony N Price, Alena Uus, Helena S Sousa, Massimo Marenzana, Kathleen Colford, Peter Murkin, Maggie Lee, Lucilio Cordero-Grande, Rui Pedro AG Teixeira, et al. In vivo t2 measurements of the fetal brain using single-shot fast spin echo sequences. *Magnetic resonance in medicine*, 92(2):715–729, 2024.

Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, and Juan Eugenio Iglesias. Synthseg: domain randomisation for segmentation of brain scans of any contrast and resolution. *arXiv preprint arXiv:2107.09559*, 2021.

Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, Juan Eugenio Iglesias, et al. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. *Medical image analysis*, 86:102789, 2023a.

Benjamin Billot, Colin Magdamo, You Cheng, Steven E Arnold, Sudeshna Das, and Juan Eugenio Iglesias. Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain mri datasets. *Proceedings of the National Academy of Sciences*, 120(9):e2216399120, 2023b.

Busra Bulut, Maik Dannecker, Thomas Sanchez, Sara Neves Silva, Vladyslav Zalevskyi, Steven Jia, Jean-Baptiste Ledoux, Guillaume Auzias, François Rousseau, Jana Hutter, et al. Physics-informed joint multi-te super-resolution with implicit neural representation for robust fetal t 2 mapping. In *International Workshop on Preterm, Perinatal and Paediatric Image Analysis*, pages 61–72. Springer, 2025.

M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

Lucilio Cordero-Grande, Juan Enrique Ortuño-Fisac, Alejandra Aguado Del Hoyo, Alena Uus, Maria Deprez, Andres Santos, Joseph V Hajnal, and María J Ledesma-Carbayo. Fetal mri by robust deep generative prior reconstruction and diffeomorphic registration. *IEEE Transactions on Medical Imaging*, 42(3):810–822, 2022.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. . URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x.

Jérôme Dockès, Gaël Varoquaux, and Jean-Baptiste Poline. Preventing dataset shift from breaking machine-learning biomarkers. *GigaScience*, 10(9):giab055, 2021.

Jessica Dubois, Marianne Alison, Serena J. Counsell, Lucie Hertz-Pannier, Petra S. Hüppi, and Manon J.N.L. Benders. MRI of the neonatal brain: A review of methodological challenges and neuroscientific advances. *Journal of Magnetic Resonance Imaging*, 53(5):1318–1343, May 2020. ISSN 1522-2586. . URL http://dx.doi.org/10.1002/jmri.27192.

Michael Ebner, Guotai Wang, Wenqi Li, Michael Aertsen, Premal A Patel, Rosalind Aughwane, Andrew Melbourne, Tom Doel, Steven Dymarkowski, Paolo De Coppi, et al. An automated framework for localization, segmentation and super-resolution reconstruction of fetal brain mri. *NeuroImage*, 206:116324, 2020.

A. David Edwards, Daniel Rueckert, Stephen M. Smith, Samy Abo Seada, Amir Alansary, Jennifer Almalbis, Joanna Allsop, Jesper Andersson, Tomoki Arichi, Sophie Arulkumaran, Matteo Bastiani, Dafnis Batalle, Luke Baxter, Jelena Bozek, Eleanor Braithwaite, Jacqueline Brandon, Olivia Carney, Andrew Chew, Daan Christiaens, Raymond Chung, Kathleen Colford, Lucilio Cordero-Grande, Serena J. Counsell, Harriet Cullen, John Cupitt, Charles Curtis, Alice Davidson, Maria Deprez, Louise Dillon, Konstantina Dimitrakopoulou, Ralica Dimitrova, Eugene Duff, Shona Falconer, Seyedeh-Rezvan Farahibozorg, Sean P. Fitzgibbon, Jianliang Gao, Andreia Gaspar, Nicholas Harper, Sam J. Harrison, Emer J. Hughes, Jana Hutter, Mark Jenkinson, Saad Jbabdi, Emily Jones, Vyacheslav Karolis, Vanessa Kyriakopoulou, Gregor Lenz, Antonios Makropoulos, Shaihan Malik, Luke Mason, Filippo Mortari, Chiara Nosarti, Rita G. Nunes, Camilla O'Keeffe, Jonathan O'Muircheartaigh, Hamel Patel, Jonathan Passerat-Palmbach, Maximillian Pietsch, Anthony N. Price, Emma C. Robinson, Mary A. Rutherford, Andreas Schuh, Stamatios Sotiropoulos, Johannes Steinweg, Rui Pedro Azeredo Gomes Teixeira, Tencho Tenev, Jacques-Donald Tournier, Nora Tusor, Alena Uus, Katy Vecchiato, Logan Z. J. Williams, Robert Wright, Julia Wurie, and Joseph V. Hajnal. The developing human connectome project neonatal data release. *Frontiers in Neuroscience*, 16, May 2022. ISSN 1662-453X. . URL http://dx.doi.org/10.3389/fnins.2022.886772.

Catherine Garel and Catherine Garel. *MRI of the Fetal Brain*. Springer, 2004.

Ali Gholipour, Judy A Estroff, and Simon K Warfield. Robust super-resolution volume reconstruction from slice acquisitions: application to fetal brain mri. *IEEE transactions on medical imaging*, 29(10):1739–1758, 2010.

O.A. Glenn and A.J. Barkovich. Magnetic resonance imaging of the fetal brain and spine: An increasingly important tool in prenatal diagnosis, part 1. *American Journal of Neuroradiology*, 27(8):1604–1611, 2006. ISSN 0195-6108. URL https://www.ajnr.org/content/27/8/1604.

Karthik Gopinath, Andrew Hoopes, Daniel C Alexander, Steven E Arnold, Yael Balbastre, Adrià Casamitjana, You Cheng, Russ Yue Zhi Chua, Brian L Edlow, Bruce Fischl, et al. Synthetic data in generalizable, learning-based neuroimaging. *Imaging Neuroscience*, 2024.

Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, March 2022. ISSN 1558-2531. . URL http://dx.doi.org/10.1109/TBME.2021.3117407.

Neal Halfon, Kandyce Larson, Michael Lu, Ericka Tullis, and Shirley Russ. Lifecourse health development: past, present and future. *Matern. Child Health J.*, 18(2):344–365, February 2014.

Fengling Hu, Andrew A. Chen, Hannah Horng, Vishnu Bashyam, Christos Davatzikos, Aaron Alexander-Bloch, Mingyao Li, Haochang Shou, Theodore D. Satterthwaite, Meichen Yu, and Russell T. Shinohara. Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. *NeuroImage*, 274:120125, 2023. ISSN 1053-8119. . URL https://www.sciencedirect.com/science/article/pii/S1053811923002719.

Xiaona Huang, Yang Liu, Yuhan Li, Keying Qi, Ang Gao, Bowen Zheng, Dong Liang, and Xiaojing Long. Deep

learning-based multiclass brain tissue segmentation in fetal mris. *Sensors*, 23(2), 2023. ISSN 1424-8220. . URL https://www.mdpi.com/1424-8220/23/2/655.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Andras Jakab, Kelly Payette, Luca Mazzone, Sonja Schauer, Cécile Olivia Muller, Raimund Kottke, Nicole Ochsenbein-Kölble, Ruth Tuura, Ueli Moehrlen, and Martin Meuli. Emerging magnetic resonance imaging techniques in open spina bifida in utero. *European Radiology Experimental*, 5(1), June 2021. ISSN 2509-9280. . URL http://dx.doi.org/10.1186/s41747-021-00219-z.

Misha P. T Kaandorp, Damola Agbelese, Hosna Asma-ull, Hyun-Gi Kim, Kelly Payette, Patrice Grehten, Gennari Antonio Giulio, Levente István Lánczi, and Andras Jakab. Pathological mri segmentation by synthetic pathological data generation in fetuses and neonates, 2025a. URL https://arxiv.org/abs/2501.19338.

Misha PT Kaandorp, Damola Agbelese, Hosna Asma-ull, Hyun-Gi Kim, Kelly Payette, Patrice Grehten, Gennari Antonio Giulio, Levente István Lánczi, and Andras Jakab. Pathological mri segmentation by synthetic pathological data generation in fetuses and neonates. *arXiv preprint arXiv:2501.19338*, 2025b.

Vyacheslav R Karolis, Lucilio Cordero-Grande, Anthony N Price, Emer Hughes, Sean P Fitzgibbon, Vanessa Kyriakopoulou, Alena Uus, Nicholas Harper, Denis Prokopenko, Devi Bridglal, et al. The developing human connectome project fetal functional mri release: Methods and data structures. *Imaging Neuroscience*, 3: imag_a_00512, 2025.

Salome Kazeminia, Christoph Baur, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay. Gans for medical image analysis. *Artificial Intelligence in Medicine*, 109:101938, 2020. ISSN 0933-3657. . URL https://www.sciencedirect.com/science/article/pii/S0933365719311510.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023. ISSN 1361-8415. . URL https://www.sciencedirect.com/science/article/pii/S1361841523001068.

William Kelley, Nathan Ngo, Adrian V. Dalca, Bruce Fischl, Lilla Zöllei, and Malte Hoffmann. Boosting skull-stripping performance for pediatric brain images, 2024. URL https://arxiv.org/abs/2402.16634.

Maria Kuklisova-Murgasova, Gerardine Quaghebeur, Mary A Rutherford, Joseph V Hajnal, and Julia A Schnabel. Reconstruction of fetal brain MRI with intensity matching and complete outlier removal. *Medical image analysis*, 16(8):1550–1564, 2012.

Vanessa Kyriakopoulou, Deniz Vatansever, Alice Davidson, Prachi Patkee, Samia Elkommos, Andrew Chew, Miriam Martinez-Biarge, Bibbi Hagberg, Mellisa Damodaram, Joanna Allsop, et al. Normative biometry of the fetal brain using magnetic resonance imaging. *Brain Structure and Function*, 222:2295–2307, 2017.

Hélène Lajous, Andrés le Boeuf Fló, Pedro M. Gordaliza, Oscar Esteban, Ferran Marqués, Vincent Dunet, Mériam Koob, and Meritxell Bach Cuadra. Maturation-informed synthetic magnetic resonance images of the developing human fetal brain. *bioRxiv*, 2024a. . URL https://www.biorxiv.org/content/early/2024/04/09/2024.04.08.588566.

Hélène Lajous, Christopher W. Roy, Tom Hilbert, Priscille de Dumast, Sébastien Tourbier, Yasser Alemán-Gómez, Jérôme Yerly, Thomas Yu, Hamza Kebiri, Kelly Payette, Jean-Baptiste Ledoux, Reto Meuli, Patric Hagmann, Andras Jakab, Vincent Dunet, Mériam Koob, Tobias Kober, Matthias Stuber, and Meritxell Bach Cuadra. A fetal brain magnetic resonance acquisition numerical phantom (fabian). *Scientific Reports*, 12(1), May 2022. ISSN 2045-2322. . URL http://dx.doi.org/10.1038/s41598-022-10335-4.

Hélène Lajous, Andrés le Boeuf Fló, Oscar Esteban, and Meritxell Bach Cuadra. Dataset Maturation-informed Synthetic Magnetic Resonance Images of the Developing Human Fetal Brain, April 2024b. URL https://doi.org/10.5281/zenodo.10940427.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: meta-learning for domain generalization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Heng Li, Haojin Li, Wei Zhao, Huazhu Fu, Xiuyun Su, Yan Hu, and Jiang Liu. Frequency-mixed single-source domain

generalization for medical image segmentation. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 127–136, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43987-2.

Peirong Liu, Oula Puonti, Xiaoling Hu, Daniel C. Alexander, and Juan E. Iglesias. Brain-id: Learning contrast-agnostic anatomical representations for brain imaging. In *European Conference on Computer Vision (ECCV)*, 2024a.

Peirong Liu, Oula Puonti, Annabel Sorby-Adams, William T. Kimberly, and Juan E. Iglesias. Pepsi: Pathology-enhanced pulse-sequence-invariant representations for brain mri, 2024b. URL https://arxiv.org/abs/2403.06227.

Quande Liu, Cheng Chen, Qi Dou, and Pheng-Ann Heng. Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. *Proc. Conf. AAAI Artif. Intell.*, 36(2):1756–1764, June 2022.

Fedel Machado-Rivas, Jasmine Gandhi, Jungwhan John Choi, Clemente Velasco-Annis, Onur Afacan, Simon K. Warfield, Ali Gholipour, and Camilo Jaimes. Normal growth, sexual dimorphism, and lateral asymmetries at fetal brain mri. *Radiology*, 303(1):162–170, 2022. . URL https://doi.org/10.1148/radiol.211222. PMID: 34931857.

Antonios Makropoulos, Emma C. Robinson, Andreas Schuh, Robert Wright, Sean Fitzgibbon, Jelena Bozek, Serena J. Counsell, Johannes Steinweg, Jonathan Passerat-Palmbach, Gregor Lenz, Filippo Mortari, Tencho Tenev, Eugene P. Duff, Matteo Bastiani, Lucilio Cordero-Grande, Emer Hughes, Nora Tusor, Jacques-Donald Tournier, Jana Hutter, Anthony N. Price, Maria Murgasova, Christopher Kelly, Mary A. Rutherford, Stephen M. Smith, A. David Edwards, Joseph V. Hajnal, Mark Jenkinson, and Daniel Rueckert. The developing human connectome project: a minimal processing pipeline for neonatal cortical surface reconstruction. *bioRxiv*, 2017. . URL https://www.biorxiv.org/content/early/2017/04/07/125526.

Antonios Makropoulos, Emma C. Robinson, Andreas Schuh, Robert Wright, Sean Fitzgibbon, Jelena Bozek, Serena J. Counsell, Johannes Steinweg, Katy Vecchiato, Jonathan Passerat-Palmbach, Gregor Lenz, Filippo Mortari, Tencho Tenev, Eugene P. Duff, Matteo Bastiani, Lucilio Cordero-Grande, Emer Hughes, Nora Tusor, Jacques-Donald Tournier, Jana Hutter, Anthony N. Price, Rui Pedro A.G. Teixeira, Maria Murgasova, Suresh Victor, Christopher Kelly, Mary A. Rutherford, Stephen M. Smith, A. David Edwards, Joseph V. Hajnal, Mark Jenkinson, and Daniel Rueckert. The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. *NeuroImage*, 173:88–112, 2018. ISSN 1053-8119. . URL https://www.sciencedirect.com/science/article/pii/S1053811918300545.

Joshua Niemeijer, Manuel Schwonberg, Jan-Aike Termöhlen, Nico M Schmidt, and Tim Fingscheidt. Generalization by adaptation: Diffusion-based domain extension for domain-generalized semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2830–2840, 2024.

Abbas Omidi, Amirmohammad Shamaei, Anouk Verschuu, Regan King, Lara Leijser, and Roberto Souza. Unsupervised domain adaptation of brain MRI skull stripping trained on adult data to newborns: Combining synthetic data with domain invariant features. In *Medical Imaging with Deep Learning*, 2024. URL https://openreview.net/forum?id=vu4LsiSpf7.

Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106, 2022.

Anthony Paproki, Olivier Salvado, and Clinton Fookes. Synthetic data for deep learning in computer vision & medical imaging: A means to reduce data bias. *ACM Comput. Surv.*, may 2024. ISSN 0360-0300. . URL https://doi.org/10.1145/3663759. Just Accepted.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

K. Payette, P. de Dumast, H. Kebiri, I. Ezhov, J.C. Paetzold, S. Shit, A. Iqbal, R. Khan, R. Kottke, P. Grehten, H. Ji, L. Lanczi, M. Nagy, M. Beresova, T.D. Nguyen, G. Natalucci, T. Karayannis, B. Menze, M. Bach Cuadra, and A. Jakab. An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. *Scientific Data*, 8(1), 2021. ISSN 2052-4463. .

Kelly Payette and Andras Jakab. Fetal tissue annotation dataset, 2021. URL https://repo-prod.prod.sagebase.org/repo/v1/doi/locate?id=syn23747212&type=ENTITY.

Kelly Payette, Hongwei Bran Li, Priscille de Dumast, Roxane Licandro, Hui Ji, Md Mahfuzur Rahman Siddiquee, Daguang Xu, Andriy Myronenko, Hao Liu, Yuchen Pei, et al. Fetal brain tissue annotation and segmentation challenge results. *Medical Image Analysis*, 88:102833, 2023.

Kelly Payette, Céline Steger, Roxane Licandro, Priscille de Dumast, Hongwei Bran Li, Matthew Barkovich, Liu Li, Maik Dannecker, Chen Chen, Cheng Ouyang, Niccolò McConnell, Alina Miron, Yongmin Li, Alena Uus, Irina Grigorescu, Paula Ramirez Gilliland, Md Mahfuzur Rahman Siddiquee, Daguang Xu, Andriy Myronenko, Haoyu Wang, Ziyan Huang, Jin Ye, Mireia Alenyà, Valentin Comte, Oscar Camara, Jean-Baptiste Masson, Astrid Nilsson, Charlotte Godard, Moona Mazher, Abdul Qayyum, Yibo Gao, Hangqi Zhou, Shangqi Gao, Jia Fu, Guiming Dong, Guotai Wang, ZunHyan Rieu, HyeonSik Yang, Minwoo Lee, Szymon Płotka, Michal K. Grzeszczyk, Arkadiusz Sitek, Luisa Vargas Daza, Santiago Usma, Pablo Arbelaez, Wenying Lu, Wenhao Zhang, Jing Liang, Romain Valabregue, Anand A. Joshi, Krishna N. Nayak, Richard M. Leahy, Luca Wilhelmi, Aline Dändliker, Hui Ji, Antonio G. Gennari, Anton Jakovčić, Melita Klaić, Ana Adžić, Pavel Marković, Gracia Grabarić, Gregor Kasprian, Gregor Dovjak, Milan Rados, Lana Vasung, Meritxell Bach Cuadra, and Andras Jakab. Multi-center fetal brain tissue annotation (feta) challenge 2022 results, 2024.

Cory M Pfeifer, Scott D Willard, and Patricia Cornejo. Mri depiction of fetal brain abnormalities. *Acta Radiologica Open*, 8(12):205846011989498, December 2019. ISSN 2058-4601. . URL http://dx.doi.org/10.1177/2058460119894987.

Anthony N Price, Lucilio Cordero-Grande, Emer Hughes, Suzanne Hiscocks, Elaine Green, Laura McCabe, Jana Hutter, Giulio Ferrazzi, Maria Deprez, Thomas Roberts, et al. The developing human connectome project (dhcp): fetal acquisition protocol. In *Proceedings of the annual meeting of the International Society of Magnetic Resonance in Medicine (ISMRM)*, volume 244. International Society for Magnetic Resonance in Medicine (ISMRM), 2019.

F. Rousseau, K. Kim, C. Studholme, M. Koob, and J. L. Dietemann. *On Super-Resolution for Fetal Brain MRI*, page 355–362. Springer Berlin Heidelberg, 2010. ISBN 9783642157455. . URL http://dx.doi.org/10.1007/978-3-642-15745-5_44.

Sahar N. Saleem. Fetal mri: An approach to practice: A review. *Journal of Advanced Research*, 5(5):507–523, 2014. ISSN 2090-1232. . URL https://www.sciencedirect.com/science/article/pii/S2090123213000805.

Ziyao Shang, Md Asadullah Turja, Eric Feczko, Audrey Houghton, Amanda Rueter, Lucille A Moore, Kathy Snider, Timothy Hendrickson, Paul Reiners, Sally Stoyell, et al. Learning strategies for contrast-agnostic segmentation via synthseg for infant mri data. In *International Conference on Medical Imaging with Deep Learning*, pages 1075–1084. PMLR, 2022.

Liyue Shen, Jimmy Zheng, Edward H. Lee, Katie Shpanskaya, Emily S. McKenna, Mahesh G. Atluri, Dinko Plasto, Courtney Mitchell, Lillian M. Lai, Carolina V. Guimaraes, Hisham Dahmoush, Jane Chueh, Safwan S. Halabi, John M. Pauly, Lei Xing, Quin Lu, Ozgur Oztekin, Beth M. Kline-Fath, and Kristen W. Yeom. Attention-guided deep learning for gestational age prediction using fetal brain mri. *Scientific Reports*, 12 (1), January 2022. ISSN 2045-2322. . URL http://dx.doi.org/10.1038/s41598-022-05468-5.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Joan Stiles and Terry L. Jernigan. The basics of brain development. *Neuropsychology Review*, 20(4):327–348, November 2010. ISSN 1573-6660. . URL http://dx.doi.org/10.1007/s11065-010-9148-4.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

Sébastien Tourbier, Xavier Bresson, Patric Hagmann, Jean-Philippe Thiran, Reto Meuli, and Meritxell Bach Cuadra. An efficient total variation algorithm for super-resolution in fetal brain mri with adaptive regularization. *NeuroImage*, 118:584–597, 2015.

Sebastien Tourbier, Priscille De Dumast, Hamza Kebiri, Patric Hagmann, and Meritxell Bach Cuadra. Medical-image-analysis-laboratory/mialsuperresolutiontoolkit: MIAL super-resolution toolkit v2.0.1. *Zenodo*, 2020. URL https://zenodo.org/record/4392788.

Alena U Uus, Irina Grigorescu, Milou PM van Poppel, Johannes K Steinweg, Thomas A Roberts, Mary A Rutherford, Joseph V Hajnal, David FA Lloyd, Kuberan Pushparajah, and Maria Deprez. Automated 3d reconstruction of the fetal thorax in the standard atlas space from

motion-corrupted mri stacks for 21–36 weeks ga range. *Medical image analysis*, 80:102484, 2022.

Alena U. Uus, Vanessa Kyriakopoulou, Antonios Makropoulos, Abi Fukami-Gartner, Daniel Cromb, Alice Davidson, Lucilio Cordero-Grande, Anthony N. Price, Irina Grigorescu, Logan Z. J. Williams, Emma C. Robinson, David Lloyd, Kuberan Pushparajah, Lisa Story, Jana Hutter, Serena J. Counsell, A. David Edwards, Mary A. Rutherford, Joseph V. Hajnal, and Maria Deprez. Bounti: Brain volumetry and automated parcellation for 3d fetal mri. *bioRxiv*, 2023. . URL https://www.biorxiv.org/content/early/2023/04/27/2023.04.18.537347.

R Valabregue, F Girka, A Pron, F Rousseau, and G Auzias. Comprehensive analysis of synthetic learning applied to neonatal brain mri segmentation. *Human Brain Mapping*, 45(6):e26674, 2024.

Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1): 48, 2022.

Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang Wang, Weizhu Chen, Mingyuan Zhou, et al. Patch diffusion: Faster and more data-efficient training of diffusion models. *Advances in neural information processing systems*, 36:72137–72154, 2023.

Yanwu Xu, Shaoan Xie, Maxwell Reynolds, Matthew Ragoza, Mingming Gong, and Kayhan Batmanghelich. *Adversarial Consistency for Single Domain Generalization in Medical Image Segmentation*, page 671–681. Springer Nature Switzerland, 2022. ISBN 9783031164491. . URL http://dx.doi.org/10.1007/978-3-031-16449-1_64.

Wenjun Yan, Lu Huang, Liming Xia, Shengjia Gu, Fuhua Yan, Yuanyuan Wang, and Qian Tao. Mri manufacturer shift and adaptation: Increasing the generalizability of deep learning segmentation for mr images acquired with different scanners. *Radiology: Artificial Intelligence*, 2 (4):e190195, July 2020. ISSN 2638-6100. . URL http://dx.doi.org/10.1148/ryai.2020190195.

Vladyslav Zalevskyi, Thomas Sanchez, Margaux Roulet, Jordina Aviles Verdera, Jana Hutter, Hamza Kebiri, and Meritxell Bach Cuadra. Improving cross-domain brain tissue segmentation in fetal mri with synthetic data. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 437–447, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72378-0.

Vladyslav Zalevskyi, Thomas Sanchez, Misha Kaandorp, Margaux Roulet, Diego Fajardo-Rojas, Liu Li, Jana Hutter, Hongwei Bran Li, Matthew Barkovich, Hui Ji, Luca Wilhelmi, Aline Dändliker, Céline Steger, Mériam Koob, Yvan Gomez, Anton Jakovčić, Melita Klaić, Ana Adžić, Pavel Marković, Gracia Grabarić, Milan Rados, Jordina Aviles Verdera, Gregor Kasprian, Gregor Dovjak, Raphael Gaubert-Rachmühl, Maurice Aschwanden, Qi Zeng, Davood Karimi, Denis Peruzzo, Tommaso Ciceri, Giorgio Longari, Rachika E. Hamadache, Amina Bouzid, Xavier Lladó, Simone Chiarella, Gerard Martí-Juan, Miguel Ángel González Ballester, Marco Castellaro, Marco Pinamonti, Valentina Visani, Robin Cremese, Keïn Sam, Fleur Gaudfernau, Param Ahir, Mehul Parikh, Maximilian Zenk, Michael Baumgartner, Klaus Maier-Hein, Li Tianhong, Yang Hong, Zhao Longfei, Domen Preloznik, Žiga Špiclin, Jae Won Choi, Muyang Li, Jia Fu, Guotai Wang, Jingwen Jiang, Lyuyang Tong, Bo Du, Andrea Gondova, Sungmin You, Kiho Im, Abdul Qayyum, Moona Mazher, Steven A Niederer, Andras Jakab, Roxane Licandro, Kelly Payette, and Meritxell Bach Cuadra. Advances in automated fetal brain mri segmentation and biometry: Insights from the feta 2024 challenge, 2025. URL https://arxiv.org/abs/2505.02784.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6xHJ37MVxxp.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2023. .

Appendix

Table of contents:

Appendix S1. Ablation on data augmentations

In this additional experiment, we compared the SynthSeg-based augmentations to a common set of augmentations from MONAI (referred to as *simple aug.*). We re-trained the models using the following augmentations, each applied with a random probability of 0.5: random affine deformation (rotation=0.2, scale=0.1, translate=30, sheer=0.1), random contrast change (Gamma transformation) ($\gamma \in [0.5, 1.5]$), random Gaussian noise ($\mu = 0, \sigma = 0.1$), random blurring ($\sigma \in [0.5, 1.5]$), and scaling to the 0-1 range. Compared to these *simple aug.*, SynthSeg-based augmentation has three additional components: a random non-linear deformation, random bias field, and a random resampling simulating an acquisition at a different resolution.

The results in Table S1 show a different picture than the results in the main paper. When comparing models trained using simple augmentations, the model trained using real data outperforms FetalSynthSeg (69.9 vs 66.9 Dice). Moving from simple augmentation increases the performance of all methods considered: FaBiAN (+2.9), randFaBiAN (+3.9), FetalSynthSeg (+8.0), Real Data (+1.9). This illustrates clearly how these data augmentation strategies are instrumental to the performance of FetalSynthSeg. They are also strong contenders for a model trained on real data and could be used in the future. A detailed ablation study of augmentations would be an interesting future step.

Table S1: Mean Dice for models trained on different sources of data. The variance is computed over all testing subjects within a split. The best performing method is shown in **boldface** in each column, and the second best is underlined. An asterisk (*) indicates that the best method is statistically significantly better than the second best (Wilcoxon test with correction).

Augmentation type	**Testing split**	**CHUV-MIAL**		**KISPI-IRTK**		**KISPI-MIAL**		**Global**
	Training split	KISPI-IRTK	KISPI-MIAL	CHUV-MIAL	KISPI-MIAL	CHUV-MIAL	KISPI-IRTK	
Simple	FaBiAN	66.9 ± 5.1	70.3 ± 6.9	51.6 ± 14.1	56.7 ± 14.2	59.8 ± 15.2	57.3 ± 17.9	60.4 ± 14.3
	randFaBiAN	74.1 ± 3.1	69.0 ± 6.4	63.8 ± 10.3	54.0 ± 9.3	62.0 ± 13.8	62.8 ± 13.0	64.3 ± 12.1
	FetalSynthSeg	72.9 ± 4.0	70.3 ± 6.2	71.4 ± 9.3	68.3 ± 8.2	58.7 ± 15.5	60.0 ± 16.1	66.9 ± 11.8
SynthSeg	FaBiAN	74.2 ± 4.2	73.1 ± 5.7	53.6 ± 13.3	56.6 ± 12.9	60.6 ± 17.1	61.5 ± 20.1	63.3 ± 15.5
	randFaBiAN	$\underline{79.1 \pm 2.3}$	$\underline{78.2 \pm 2.9}$	55.1 ± 12.7	68.9 ± 7.8	60.7 ± 7.3	$\underline{68.1 \pm 14.7}$	68.3 ± 13.8
	SynthSeg	75.9 ± 3.9	73.7 ± 3.9	$\underline{70.9 \pm 9.2}$	$\underline{74.8 \pm 7.8}$	60.5 ± 15.7	63.4 ± 16.8	72.2 ± 13.0
	FetalSynthSeg	$\mathbf{80.7 \pm 2.0}^*$	76.9 ± 3.3	$\mathbf{79.2 \pm 9.0}^*$	$\mathbf{76.8 \pm 6.9}^*$	$\underline{67.5 \pm 16.0}$	$\mathbf{68.5 \pm 15.6}$	$\mathbf{74.9 \pm 11.5}^*$
Simple	Real Data	76.5 ± 3.2	75.2 ± 3.5	69.6 ± 13.6	67.7 ± 12.9	67.2 ± 17.6	63.4 ± 17.0	69.9 ± 12.6
SynthSeg	Real Data	77.2 ± 4.0	$\mathbf{78.5 \pm 3.3}^*$	70.6 ± 13.9	71.9 ± 11.5	$\mathbf{68.0 \pm 19.3}$	64.3 ± 19.1	$\underline{71.8 \pm 13.9}$

Appendix S2. Comparing FaBiAN vs FetalSynthSeg on the same number of images

To ensure a fair comparison between FaBiAN and FetalSynthSeg we equalize the computational budget used to generate images for these two approaches. We generate offline 6000 synthetic images per split, as in our FaBiAN experiments, and use them to train a FetalSynthSeg-6k model. Table S2 shows that although this model achieves slightly lower results than FetalSynthSeg, it follows similar trends in performance across splits and still outperforms even the randFaBiAN approach overall (average Dice of FetalSynthSeg-6k 72.2±8.8 is vs average Dice of randFaBiAN of 68.3±8.5 across all splits).

Table S2: Dice scores comparing FetalSynthSeg model trained on the same amount of synthetic images as FaBiAN models.

Testing Split	**CHUV-MIAL**		**KISPI-IRTK**		**KISPI-MIAL**		**Global**
Training Split	KISPI-IRTK	KISPI-MIAL	CHUV-MIAL	KISPI-MIAL	CHUV-MIAL	KISPI-IRTK	
FaBiAN	74.2±4.2	73.1±5.7	53.6±13.3	56.6±12.9	60.6±17.1	61.5±20.1	63.3±15.5
randFaBiAN	79.1±2.3	**78.2±2.9***	55.1±12.7	68.9±7.8	60.7±7.3	68.1±14.7	68.3±13.8
FetalSynthSeg-6k	79.0±2.6	76.0±3.6	78.3±9.2	75.1±6.6	66.5±15.9	67.7±15.4	72.2±13.0
FetalSynthSeg	**80.7±2.0***	76.9±3.3	**79.2±9.0***	**76.8±6.9***	**67.5±16.0***	**68.5±15.6***	**74.9±11.5***

Appendix S3. Generation parameters

In the Table S3 we report parameters used to create synthetic images with the FaBiAN generator.

Contrast	
Effective echo time (ms)	[90,300]
Echo spacing (ms)	6.12
Echo train length	150
Excitation flip angle (°)	90
Refocusing pulse flip angle (°)	[150,180]
Geometry	
Slice thickness (mm)	0.8
Slice gap (mm)	0
Number of slices	112
Phase oversampling (%)	80
Shift of field-of-view (mm)	0
Resolution	
Field-of-view (mm^2)	120x120
Base resolution (voxels)	150
Phase resolution (%)	70
Reconstruction matrix	150
Zero-interpolation filling	1
Acceleration technique	
Reference lines	42
Acceleration factor	2
Noise	
Mean	0
Standard deviation	<0.01

Table S3: FaBiAN Simulation Parameters for HASTE Imaging

Appendix S4. Additional quantitative results

We present additional quantitative results related to the conducted experiments, reporting the Dice scores per tissue for all of the tested models in Tables S4 as well as the average 95th-percentile Hausdorff distance score in Table S5.

Table S4: Mean and standard deviation of Dice score of the explored models for different testing and training splits. LV - lateral ventricles, CBM - cerebellum, SGM - sub-cortical gray matter, BS - brainstem, Mean DSC - value averaged across all segmentation labels.

Testing Splits	Training Split	Experiment	Mean DSC	CSF	GM	WM	LV	CBM	SGM	BS
CHUV-MIAL	KISPI-IRTK	FetalSynthSeg	**80.7±2.0**	**79.3±2.7**	**70.7±3.6**	**86.4±2.3**	78.7 ± 5.4	88.5 ± 2.3	**79.6±5.4**	**81.9±2.7**
		Real data	77.2±4.0	72.3±7.1	68.4±4.7	86.0±2.8	75.8±7.9	84.3±9.2	76.4±5.9	77.5±4.7
		FaBiAN	74.2±4.2	75.3±4.2	64.2±4.8	81.3±4.4	71.9±9.9	85.3±3.6	62.8±14.2	78.7±3.3
		randFaBiAN	79.1±2.3	77.0±3.3	69.7±3.6	85.9±2.3	75.7±5.3	87.4±3.1	78.5±8.4	79.5±3.9
		SynthSeg	75.9±3.9	73.7±4.5	64.1±6.3	83.1±4.4	71.9±8.5	85.8±2.9	74.0±10.6	78.8±3.9
	KISPI-MIAL	FetalSynthSeg	76.9±3.3	73.1±8.1	60.5±4.8	84.9±3.0	**77.3±6.3**	86.2±4.3	83.7±4.6	72.8±3.9
		Real data	**78.5±3.3**	**79.0±5.5**	**65.6±4.4**	**85.8±3.2**	77.1±7.3	86.5±6.1	**83.9±3.8**	71.4±5.1
		FaBiAN	73.1±5.7	73.3±4.9	61.9±5.9	81.1±4.8	66.3±10.6	81.0±14.9	76.3±7.3	71.5±7.1
		randFaBiAN	78.2±2.9	77.8±3.9	64.1±5.2	84.4±3.2	75.5±5.4	**87.5±4.6**	83.0±5.0	**75.3±5.0**
		SynthSeg	73.7±4.4	67.9±12.0	55.1±8.7	81.3±4.2	72.1±7.2	84.0±3.0	80.9±5.2	74.6±3.9
KISPI-IRTK	CHUV-MIAL	FetalSynthSeg	**79.2±9.0**	**77.6±10.9**	**69.4±12.6**	**85.0±12.9**	**79.7±5.7**	**87.5±9.1**	**74.8±14.4**	**80.3±6.6**
		Real data	70.6±13.9	58.6±18.4	60.9±15.9	84.8±9.8	72.8±11.7	74.3±26.3	74.5±11.0	68.1±19.9
		FaBiAN	53.6±13.3	51.4±18.0	40.9±11.9	74.9±10.7	36.0±13.8	45.9±27.0	66.9±12.1	59.0±19.3
		randFaBiAN	55.1±12.7	54.2±17.2	41.5±11.5	76.9±10.5	37.2±13.7	51.8±28.0	66.5±11.7	57.8±18.5
		SynthSeg	70.9±9.2	69.9±11.6	58.6±9.0	79.2±11.8	66.4±7.0	80.2±11.3	68.8±13.7	73.2±9.2
	KISPI-MIAL	FetalSynthSeg	**76.8±6.9**	**78.7±9.2**	**68.5±9.1**	**86.5±10.9**	**80.8±4.7**	85.1±9.1	68.2±12.3	69.9±6.8
		Real data	71.9±11.5	72.3±19.5	65.3±13.4	85.5±7.6	71.6±9.0	80.3±20.2	64.8±13.3	63.2±14.5
		FaBiAN	56.6±12.9	58.0±17.6	39.3±13.0	74.3±9.5	50.6±15.0	64.7±27.7	58.7±13.9	50.6±18.7
		randFaBiAN	68.9±7.8	60.8±16.0	52.8±11.1	82.2±6.3	68.1±9.7	80.1±10.9	68.0±12.3	70.2±7.9
		SynthSeg	74.8±7.8	74.4±12.7	65.7±10.8	85.4±11.0	71.6±7.3	**86.3±7.5**	**68.5±13.1**	**72.0±8.0**
KISPI-MIAL	CHUV-MIAL	FetalSynthSeg	67.5±16.0	59.5±28.3	48.7±19.4	74.3±17.3	**78.6±12.3**	**67.2±29.7**	**79.3±9.8**	**64.9±16.0**
		Real data	**68.0±19.3**	**62.1±32.4**	**58.3±15.7**	**81.8±12.3**	77.0±14.5	64.6±37.3	78.8±9.9	53.4±29.8
		FaBiAN	60.6±17.1	52.9±29.1	48.7±15.2	77.6±12.3	63.2±17.3	53.7±33.1	73.4±14.5	54.8±25.8
		randFaBiAN	60.7±16.3	52.3±28.6	49.6±14.5	78.2±12.4	63.7±18.0	52.8±32.5	73.8±14.3	54.6±22.4
		SynthSeg	60.5±15.7	53.7±26.0	43.8±14.1	70.0±15.4	65.9±12.1	59.9±28.5	71.1±15.5	59.0±17.0
	KISPI-IRTK	FetalSynthSeg	**68.5±15.6**	**63.2±25.9**	**57.2±16.8**	82.1±13.5	79.4±12.2	**73.2±24.9**	**63.6±18.4**	**60.8±17.0**
		Real data	64.3±19.1	59.8±29.0	55.0±17.9	79.1±15.5	72.6±12.9	64.4±33.8	62.5±16.3	57.1±21.6
		FaBiAN	61.5±20.1	57.5±31.5	53.1±19.0	80.6±11.8	72.3±16.1	59.1±35.1	52.8±20.9	54.9±23.4
		randFaBiAN	68.1±14.7	62.5±27.3	59.4±12.4	**83.8±12.2**	**80.4±11.0**	70.6±25.0	62.3±21.2	57.4±15.1
		SynthSeg	63.4±16.8	57.3±26.8	50.2±17.6	77.4±16.0	77.7±13.4	70.4±22.0	54.7±23.9	56.2±18.2

Table S5: Mean and standard deviation of 95-th percentile Hasudorff distance for the explored models on different testing and training splits. LV - lateral ventricles, CBM - cerebellum, SGM - sub-cortical gray matter, BS - brainstem, mHD95 - average value across all labels.

Testing Splits	Training Split	Experiment	mHD95	CSF	GM	WM	Ventricles	Cerebellum	Deep_GM	Brainstem
CHUV - MIAL	KISPI-IRTK	FetalSynthSeg	**2.0±0.5**	**1.7±0.3**	**1.6±0.3**	2.5±3.3	**2.0±0.9**	**1.5±0.2**	**2.8±0.6**	**2.1±0.6**
		Real data	2.9±1.2	2.6±0.8	1.9±0.5	**2.0±0.3**	4.2±4.4	2.9±4.7	4.0±2.8	2.8±1.1
		FaBiAN	3.6±1.3	2.6±0.9	2.2±0.4	3.6±0.8	5.3±4.1	3.5±6.0	5.0±2.3	2.9±1.0
		randFaBiAN	2.1±0.3	1.8±0.3	**1.6±0.3**	2.1±0.3	2.1±0.7	1.6±0.4	3.0±0.8	2.6±1.0
		SynthSeg	2.7±0.6	2.0±0.4	2.1±1.9	2.5±0.3	3.7±1.8	2.3±0.8	3.3±0.8	2.9±1.4
	KISPI-MIAL	FetalSynthSeg	3.1±0.7	2.7±1.7	2.5±0.5	**2.2±0.4**	2.6±1.7	2.9±2.9	**2.6±0.5**	6.3±0.9
		Real data	3.0±0.7	2.2±1.6	2.3±0.4	**2.2±0.5**	3.1±3.0	**1.9±0.8**	2.9±0.6	6.6±1.1
		FaBiAN	4.1±1.4	3.1±1.9	2.8±0.6	4.0±1.2	4.6±3.4	2.8±1.7	5.7±5.2	5.4±1.3
		randFaBiAN	**2.7±0.6**	**1.9±0.6**	**2.2±0.5**	2.4±0.5	**2.1±0.6**	2.5±2.2	2.8±0.7	5.0±1.2
		SynthSeg	3.4±0.6	3.5±2.9	2.4±0.4	2.9±0.3	4.6±1.9	3.1±1.2	3.2±0.6	**3.9±0.9**
KISPI - IRTK	CHUV-MIAL	FetalSynthSeg	2.5±1.6	2.4±2.2	1.6±0.8	2.2±1.4	1.9±1.0	1.5±0.9	4.3±4.0	3.4±3.6
		Real data	9.8±5.1	9.0±2.3	7.6±4.3	5.7±4.2	6.3±6.0	16.1±16.1	9.5±9.6	14.1±9.7
		FaBiAN	14.8±5.3	9.2±2.1	10.3±2.2	9.6±2.7	14.2±4.3	28.7±16.6	15.1±8.7	16.3±11.1
		randFaBiAN	14.8±5.1	9.0±2.0	10.7±2.3	10.0±3.2	13.2±5.5	24.4±16.5	17.1±10.1	18.9±11.3
		SynthSeg	**1.5±0.5**	**1.3±1.0**	**1.2±0.3**	**1.4±0.4**	**1.4±0.6**	**1.3±0.8**	**2.0±0.8**	**1.7±0.7**
	KISPI-MIAL	FetalSynthSeg	**3.4±1.7**	**2.4±1.5**	**1.7±0.8**	**2.1±0.7**	**3.1±3.4**	**3.4±6.5**	5.5±2.9	5.9±1.4
		Real data	5.9±4.0	6.3±4.0	4.1±4.8	5.4±6.8	4.3±4.8	6.3±10.8	6.8±4.2	7.8±4.7
		FaBiAN	14.0±5.0	9.0±2.1	10.1±2.7	11.1±3.4	8.9±6.3	24.9±18.4	18.6±11.0	15.4±9.1
		randFaBiAN	5.5±2.6	6.8±2.6	3.9±1.9	3.8±2.3	5.3±5.6	6.7±10.0	6.3±5.6	5.8±3.2
		SynthSeg	3.5±1.4	2.9±2.0	**1.7±0.8**	2.3±0.7	3.8±3.2	3.7±6.3	**5.3±2.0**	**4.4±1.3**
KISPI - MIAL	CHUV-MIAL	FetalSynthSeg	4.1±2.4	4.8±4.5	3.2±1.9	3.4±1.8	2.4±2.3	4.5±3.9	3.8±3.3	6.4±3.3
		Real data	5.1±3.4	5.0±4.9	3.1±1.7	2.8±1.3	4.1±4.1	7.7±9.8	4.1±3.0	8.6±6.3
		FaBiAN	8.2±4.3	7.3±4.9	6.5±2.4	4.1±1.4	8.9±4.4	11.2±12.4	7.3±7.7	12.3±8.7
		randFaBiAN	7.6±4.2	7.0±4.8	6.0±1.8	3.9±1.4	8.0±4.4	9.9±10.6	7.8±8.1	10.5±8.1
		SynthSeg	**2.6±1.6**	**3.1±3.2**	**2.2±2.2**	**2.6±3.0**	**1.7±1.0**	**2.5±1.7**	**2.7±3.4**	**3.6±4.3**
	KISPI-IRTK	FetalSynthSeg	**3.9±2.0**	**3.9±3.6**	**2.4±1.5**	**2.6±1.3**	2.3±2.0	3.6±3.1	**5.6±2.6**	6.7±2.7
		Real data	4.7±2.7	5.2±4.6	2.8±1.7	2.7±1.4	4.7±3.9	5.9±6.5	5.8±2.6	**5.9±1.8**
		FaBiAN	6.5±3.9	6.5±4.9	4.7±3.3	4.4±2.4	6.4±5.4	8.3±7.9	7.0±5.5	8.4±6.0
		randFaBiAN	4.1±2.0	4.4±4.3	2.8±1.3	3.1±2.8	**2.2±1.7**	**3.5±2.3**	5.7±2.8	7.0±2.6
		SynthSeg	5.2±2.8	4.9±4.2	3.6±2.0	3.7±3.0	4.4±6.5	5.1±8.1	6.7±3.9	7.9±2.9

Appendix S5. Out-of-domain evaluation and comparison to state of the art

Table S6 reports the label-wise DSC and HD95 metrics for all evaluated methods, separated by modality (T1w and T2w). The results are averaged across the corresponding out-of-domain datasets to provide a robust comparison against state-of-the-art (SoTA) fetal brain MRI segmentation models.

On the **T1w datasets**, we observe a substantial performance gap between `FetalSynthSeg` and all competing methods. `FetalSynthSeg` is the only model that consistently performs well across *all* labels, whereas other SoTA approaches show highly variable performance and generally fail to generalize to T1-weighted data. Notably, WM is the only label where some methods achieve moderate performance (with DSC values around 60%), yet this remains far below the 89% achieved by `FetalSynthSeg`. All other labels exhibit even larger drops. These results highlight that domain shift affects different anatomical structures unequally, with some tissues (e.g., WM) being significantly more robust to cross-domain variability than others.

On the **T2w datasets**, the performance across models is more homogeneous. For each label, the differences between SoTA methods typically remain within 1–3% DSC, indicating that all methods generalize reasonably well to unseen T2-weighted data. Nevertheless, consistent label-specific trends are still present: WM is generally the easiest structure to segment, whereas VM and GM tend to be more challenging across all models. This reinforces that label-dependent difficulty persists even in settings with smaller domain gaps.

T2w datasets			
Model	**Label**	**DSC [%]**	**HD95**
Bounti	BS	88.0±6.0	2.7±2.6
	CBM	89.0±2.0	2.3±2.3
	CSF	88.0±8.0	1.6±1.4
	GM	81.0±9.0	1.2±0.8
	SGM	69.0±8.0	5.9±1.6
	VM	77.0±9.0	7.9±3.5
	WM	88.0±10.0	3.4±0.8
FRS	BS	83.0±5.0	4.2±1.4
	CBM	89.0±5.0	2.6±1.0
	CSF	85.0±6.0	1.9±2.1
	GM	79.0±6.0	1.6±1.8
	SGM	81.0±6.0	4.8±1.5
	VM	78.0±7.0	8.6±4.9
	WM	91.0±3.0	2.0±2.1
FSS	BS	81.0±5.0	4.2±1.6
	CBM	90.0±6.0	2.2±1.0
	CSF	84.0±9.0	1.8±1.1
	GM	78.0±9.0	1.4±0.5
	SGM	80.0±7.0	4.8±1.7
	VM	77.0±9.0	6.9±4.5
	WM	90.0±4.0	1.8±0.4
FeTA24	BS	81.0±6.0	4.0±1.5
	CBM	87.0±6.0	2.5±1.0
	CSF	83.0±8.0	2.3±1.9
	GM	77.0±7.0	1.6±0.8
	SGM	80.0±8.0	4.5±1.7
	VM	77.0±9.0	6.6±4.8
	WM	90.0±4.0	2.0±0.7
nnU-Net	BS	81.0±8.0	4.3±1.6
	CBM	89.0±8.0	2.5±1.1
	CSF	84.0±9.0	2.3±3.9
	GM	80.0±7.0	1.4±0.6
	SGM	81.0±8.0	4.7±1.7
	VM	80.0±8.0	7.0±4.9
	WM	91.0±3.0	1.8±1.9

T1w datasets			
Model	**Label**	**DSC [%]**	**HD95**
Bounti	BS	0.0±1.0	32.3±9.5
	CBM	2.0±4.0	36.3±11.2
	CSF	21.0±3.0	5.3±1.1
	GM	8.0±3.0	6.5±1.4
	SGM	19.0±11.0	13.4±3.1
	VM	5.0±3.0	10.8±3.2
	WM	60.0±6.0	7.3±1.3
FRS	BS	51.0±12.0	11.1±6.8
	CBM	63.0±24.0	7.5±7.9
	CSF	31.0±6.0	6.3±4.1
	GM	22.0±5.0	5.4±4.2
	SGM	58.0±10.0	12.2±7.3
	VM	15.0±8.0	13.4±7.8
	WM	67.0±4.0	7.8±6.9
FSS	BS	78.0±3.0	5.5±0.6
	CBM	89.0±2.0	2.7±0.8
	CSF	79.0±6.0	2.0±0.5
	GM	75.0±5.0	1.6±0.6
	SGM	80.0±4.0	5.8±0.5
	VM	69.0±7.0	10.3±2.0
	WM	89.0±3.0	1.9±0.5
FeTA24	BS	5.0±8.0	16.5±5.8
	CBM	33.0±18.0	10.7±6.0
	CSF	28.0±4.0	5.4±1.7
	GM	15.0±4.0	5.2±1.6
	SGM	11.0±11.0	17.9±5.6
	VM	7.0±3.0	12.8±3.4
	WM	65.0±6.0	5.7±2.1
nnU-Net	BS	0.0±0.0	41.3±8.2
	CBM	2.0±7.0	41.1±10.5
	CSF	29.0±4.0	10.8±8.7
	GM	9.0±4.0	14.2±9.1
	SGM	1.0±2.0	25.4±10.4
	VM	3.0±2.0	27.4±16.7
	WM	51.0±10.0	21.0±12.2

Table S6: Segmentation performance on T2w (left) and T1w (right) datasets. DSC in percent, HD95 in original units (mean±std).