Fetal-BET: Brain Extraction Tool for Fetal MRI

Razieh Faghihpirayesh ⓘ, Davood Karimi ⓘ, Deniz Erdoğmuş ⓘ, *Senior Member, IEEE,* and Ali Gholipour ⓘ, *Senior Member, IEEE*

Abstract—Goal: **In this study, we address the critical challenge of fetal brain extraction from MRI sequences. Fetal MRI has played a crucial role in prenatal neurodevelopmental studies and in advancing our knowledge of fetal brain development *in-utero*. Fetal brain extraction is a necessary first step in most computational fetal brain MRI pipelines. However, it poses significant challenges due to 1) non-standard fetal head positioning, 2) fetal movements during examination, and 3) vastly heterogeneous appearance of the developing fetal brain and the neighboring fetal and maternal anatomy across gestation, and with various sequences and scanning conditions. Development of a machine learning method to effectively address this task requires a large and rich labeled dataset that has not been previously available. Currently, there is no method for accurate fetal brain extraction on various fetal MRI sequences.** *Methods:* **In this work, we first built a large annotated dataset of approximately 72,000 2D fetal brain MRI images. Our dataset covers the three common MRI sequences including T2-weighted, diffusion-weighted, and functional MRI acquired with different scanners. These data include images of normal and pathological brains. Using this dataset, we developed and validated deep learning methods, by exploiting the power of the U-Net style architectures, the attention mechanism, feature learning across multiple MRI modalities, and data augmentation for fast, accurate, and generalizable automatic fetal brain extraction.** *Results:* **Evaluations on independent test data, including data available from other centers, show that our method achieves accurate brain extraction on heterogeneous test data acquired with different scanners, on pathological brains, and at various gestational stages.** *Conclusions:* **By leveraging rich information from diverse multi-modality fetal MRI data, our proposed deep learning solution enables precise delineation of the fetal brain on various fetal MRI sequences. The robustness of our deep learning model underscores its potential utility for fetal brain imaging.**

Index Terms—Deep learning, brain extraction, fetal MRI.

Impact Statement— This study introduces a robust deep learning model for accurate fetal brain extraction from MRI, enhancing fetal brain imaging across various conditions and stages.

I. INTRODUCTION

FETAL Magnetic Resonance Imaging (MRI) is a critical tool in studying prenatal neurodevelopment due to its superior soft tissue contrast compared to ultrasound [1]. However, MRI is very susceptible to motion and fetuses can move significantly during MRI scans. To mitigate this problem, fast MRI acquisition techniques are used to obtain stacks of 2D slices [2], [3], [4]. Brain extraction in these MRI slices is a fundamental step in various applications, including fetal head motion tracking [5], slice-level motion correction [6], [7], and slice-to-volume reconstruction [8], [9], [10], [11], [12], [13], [14], [15], [16]. However, automated fetal brain extraction remains challenging due to the variability in brain size, shape, and structure across gestational age, unpredictable fetal and maternal motion, image distortions, intensity non-uniformity, and the image contrast that changes in various fetal MRI sequences such as diffusion-weighted MRI. As can be seen in the examples presented in Fig. 1, motion artifacts can significantly degrade the quality of fetal MRI acquisitions. Fetal MRI scans also typically exhibit anisotropic resolutions across different axes, featuring high in-plane resolution but lower inter-slice resolution. Consequently, achieving accurate fetal brain extraction from such MR images poses a significant challenge.

Over the years, various approaches have been proposed to tackle the challenging task of fetal brain extraction, ranging from classical image processing techniques to modern machine learning-based techniques. Classical techniques often rely on thresholding, region growing, and morphological operations for fetal brain segmentation [17]. Although these methods may yield acceptable outcomes in specific scenarios, they often face difficulties related to intensity variations, image artifacts, and the intricate anatomical structures inherent to fetal brain imaging. Classical machine learning techniques have also been explored for fetal brain extraction in MRI [18]. However, these methods exhibit limitations in terms of efficiency, accuracy, and their

Razieh Faghihpirayesh is with the Electrical and Computer Engineering Department, Northeastern University, Boston, MA 02115 USA, and also with the Radiology Department, Boston Children's Hospital, and Harvard Medical School, Boston, MA 02115 USA (e-mail: razieh.faghih@ece.neu.edu).

Davood Karimi and Ali Gholipour are with the Radiology Department, Boston Children's Hospital, and Harvard Medical School, Boston, MA 02115 USA.

Deniz Erdoğmuş is with the Electrical and Computer Engineering Department, Northeastern University, Boston, MA 02115 USA.

For easy accessibility, we have dockerized our method and made it available at: https://github.com/IntelligentImaging/fetal-brain-extraction; https://github.com/Faghihpirayesh/fetal-brain-extraction

Digital Object Identifier 10.1109/OJEMB.2024.3426969

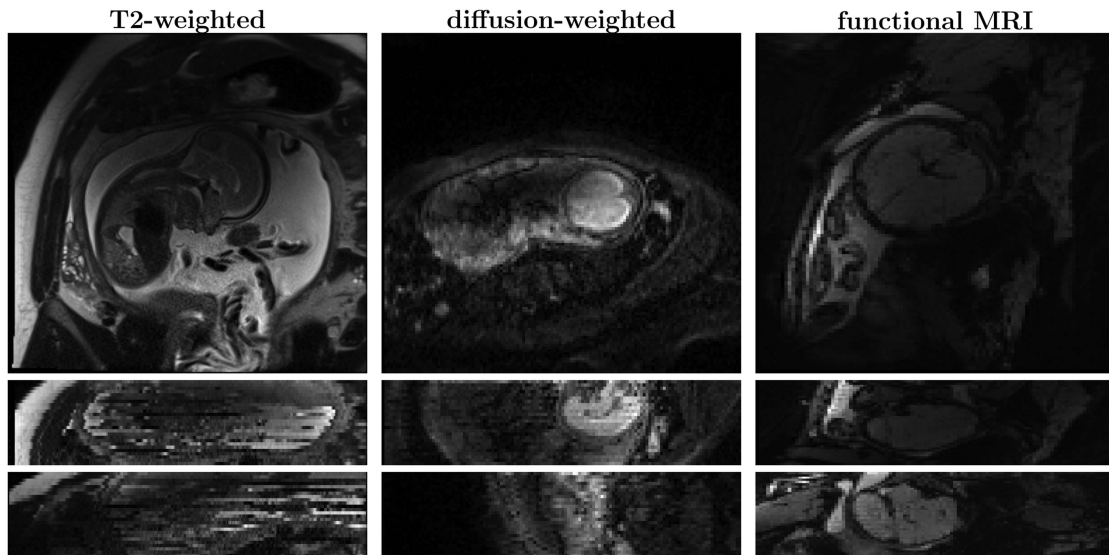| T2-weighted	diffusion-weighted	functional MRI

Fig. 1. Examples of multi-modal in-utero MRI images including T2-weighted, diffusion-weighted, and functional MRI. The first row shows in-plane views and the second and the third rows show out-of-plane views. These examples highlight some of the factors that make fetal brain extraction especially challenging such as motion artifacts, anisotropic resolution, heterogeneous contrast, and the highly variable shape and features of the anatomy based on the gestational age and the position of the fetus.

ability to perform well in diverse settings. Typically, they follow a two-stage process: first, a technique is applied to detect the fetal brain or a reference object within the MRI, and then the extraction phase isolates the fetal brain from the identified region. One drawback of these approaches is their reliance on manually crafted features to interpret MR images, which may not adequately capture the intricate visual patterns present in the fetal brain and its surrounding tissues [19]. The emergence of deep learning has brought significant improvements in the field of medical imaging [20], including fetal brain extraction [21], [22]. These techniques have demonstrated remarkable performance in handling the complexity and variability inherent in fetal brain MRI scans.

While significant strides have been made in addressing the complexities of fetal brain extraction through various methodologies, the quest for more robust, efficient, and accurate techniques persists. The focus of this study is to leverage the power of deep learning to further advance the field of fetal brain MRI extraction, particularly across diverse MRI sequences including T2-weighted (T2W), diffusion-weighted (DWI), and functional MRI (fMRI).

In this work, we propose a deep learning framework to tackle the challenges posed by fetal brain MRI scans. By capitalizing on the vast information encoded in diverse fetal MRI data across multiple modalities, our method aims to surpass the limitations of conventional techniques and classical machine learning approaches. This work has two primary objectives. First, to achieve superior accuracy in fetal brain extraction. Our goal is to develop methods that work well in the presence of motion artifacts, intensity variations, and complex anatomical structures observed in T2w, DWI, and fMRI sequences. Second, to provide a standardized, efficient, and adaptable solution for seamless integration into the workflow of clinical practitioners and researchers working with diverse MRI modalities. To the best of our knowledge, there is currently no model that can effectively extract fetal brain from MRI images of varying contrasts from different sequences.

In the following sections, we will review related research on fetal brain MRI extraction, explain our framework's design and implementation, present the results of our model's performance, and discuss the broader implications and future research directions.

II. RELATED WORK

Brain extraction refers to computational methods for removing the skull (skull stripping) and other non-cerebral tissues from head scans [23]. It is an important step that can significantly impact all downstream processing and analyses. Accurate skull stripping is crucial for various downstream tasks in medical image analysis, such as brain tissue segmentation, volumetric measurements, and the study of brain development and abnormalities [1]. Precise delineation of the fetal brain is essential for assessing brain growth, detecting developmental delays, and characterizing structural anomalies, which are critical for prenatal diagnosis and monitoring. Skull stripping serves as a critical preprocessing step that enables reliable and accurate analysis of brain structures and pathologies. In the context of fetal MRI brain extraction, 2D segmentation techniques are often preferred over 3D approaches as acquisitions are based on 2D sequences to minimize through-slice motion artifacts. These acquisitions are often anisotropic in resolution (with relatively thick slices) and almost always exhibit significant inter-slice motion artifacts. Such artifacts as well as maternal respiratory dynamics can result in errors when 3D segmentation methods are used [19]. Brain extraction in these MRI slices is a fundamental step in various fetal MRI applications, including fetal head motion tracking [5], slice-level motion correction [6], [7], and slice-to-volume reconstruction [8], [9], [10], [11], [12], [13], [14], [15], [16].

In the case of 2D fetal MRI scans, the fetal brain must be recognized and separated in each slice from a widely variable set of structures such as the uterus, placenta, amniotic fluid, maternal tissue and organs, or fetal body and extremities. In 2D views many of these structures may resemble a sectional view of a developing fetal brain. Therefore, fetal brain extraction on the original MRI slices is a challenging task. Although several studies have addressed fetal brain extraction on 3D reconstructed fetal brain MRI images, e.g. [24], only a few have tackled the more difficult task of extracting the fetal brain in each slice of the original acquisitions [25]. Additionally, most existing models for fetal brain extraction have focused on a single sequence, often T2W MRI, which limits their applicability to other MRI sequences.

To address the fetal brain extraction task, several techniques employ a preliminary step of localizing the fetal brain to direct brain extraction. Subsequently, two strategies are presented: either extracting the fetal brain from the entire field of view or after a brain localization step. Keraudren et al. [18] pinpointed the fetal brain region using a Bag-of-Words model combined with SIFT (Scale-Invariant Feature Transform) features [26]. Subsequently, a combination of a sparse patch-based method and a conditional random field was used for brain segmentation from 2D MRI slices. However, these approaches relied on handcrafted features, which often introduced inaccuracies in fetal brain extraction due to the inherent heterogeneity between these features and the subsequent extraction algorithms [19].

Taimouri et al. [27] proposed a block matching technique to simultaneously detect a bounding box around the brain and match the orientation of the brain to a template. Due to a search in the space of possible orientations, this technique was also computationally expensive. Tourbier et al. [28] proposed an atlas-based fetal brain segmentation approach that required a predefined bounding box around the brain. This method was also time consuming as it relied on deformable registration to multiple atlases.

Recent studies have focused mainly on deep learning (DL) and, in particular, convolutional neural networks (CNNs). In comparison to conventional methods, deep CNNs can be utilized to learn the features of fetal brain MR images that are pertinent to the task of fetal brain extraction. The U-Net [29] style architecture is the baseline model that is commonly used for medical image segmentation [25], [30].

Lou et al. [31] presented a multistage 2D U-Net, DS U-Net, for fetal brain extraction in T2W MRI. The approach uses a three-step process, all employing the U-Net architecture [29]. First, a coarse segmentation is performed to outline a 3D bounding box around the fetal brain. Next, a more detailed segmentation is carried out for precise brain extraction. Finally, a refined segmentation is conducted using a local patch strategy. Li et al. [32] designed a two-step framework using two 2D FCNs (Fully Convolutional Networks) [33] for fetal brain extraction from MRI slices. One FCN locates and extracts the region of interest (ROI) containing the brain, while a deeper FCN further refines the segmentation. Dudovitch et al. [34] introduced a DL method with two CNN types: a custom 3D U-Net [35] for bounding box

definition and brain extraction, and a 2D U-Net [29] that refines segmentation considering adjacent slice results.

Liao et al. [36] showcased a multistage DL model for both image quality assessment and fetal brain extraction, with modules that detect and extract the brain using the U-Net and deformable convolutional layers. Zhang et al. [37] proposed a confidence-aware cascaded framework with two U-Net [29] modules, one for localization and another for fine-tuning. The framework evaluates slice-specific confidence for extraction, using higher-confidence slices to guide the extraction of lower-confidence ones. Salehi et al. [21] utilized a 2D U-Net [29] to efficiently extract the fetal brain from T2W MR images. Khalili et al. [38] presented a multiscale CNN, influenced by the architecture from [39], using three parallel 2D convolutional pathways that analyze 2D patches of different sizes. In a subsequent study, Khalili et al. [40] utilized a 2D CNN based on a scaled-down U-Net architecture [29] for both fetal and neonatal T2W MR scans, and a post-processing algorithm from their previous work [38].

Faghihpirayesh et al. [41] proposed RFBSNet, a U-Net style [29] architecture designed for real-time fetal brain segmentation, emphasizing speed and accuracy. Rutherford et al. [22] introduced a CNN model adapted from the U-Net architecture specifically for fetal brain extraction in fMRI.

While all of the above-referenced studies focused on only one type of MRI modality, either T2W or fMRI; in this study we aimed to train a model on a pool of heterogeneous data from diverse MRI modalities including T2W, fMRI, and DWI. We anticipated that this approach should improve the model's performance by 1) enabling the model to recognize features that are specific to each MRI sequence, and 2) allowing the model to learn patterns that are common to all modalities. We hypothesized that this holistic training approach would help improve model robustness in handling varying imaging scenarios and enhance generalizability in real-world settings.

To build a holistic fetal brain extraction tool for fetal MRI (Fetal-BET), we built deep learning models based on some of the best-performing CNN architectures. Specifically, the U-Net, the nnU-Net (here the dynamic U-Net), and the Attention U-Net. We critically evaluated the performance of these models on our heterogeneous test sets. Since none of the previous methods [21], [22] were designed to segment the brain across multiple MRI modalities, a direct comparison was not appropriate. Nonetheless, we critically evaluated the performance of Fetal-BET on any specific image type, where we specifically evaluated the performance gain achieved by feature learning across multiple MRI modalities and data augmentation. We built and trained models based on variations of the U-Net that was used in most of the previous studies.

III. MATERIALS AND METHODS

A. Data

The data utilized in this study were sourced from fetal MRIs conducted at Boston Children's Hospital over a span of approximately 20 years. These MRI acquisitions encompassed a range of MRI scanners, including 1.5 T GE, Philips, and

Siemens, as well as 3 T Siemens scanners, specifically Skyra, Prisma, and Vida models. The study was approved by the Institutional Review Board. For all prospective fetal MRIs, written informed consent was obtained from the participants.

The MRI acquisition protocols typically involved the acquisition of multiple types of images, including T2-weighted (T2W) 2D half-Fourier single-shot turbo spin echo sequences with in-plane resolutions ranging from 1 to 1.25 mm and slice thicknesses between 2 to 4 mm, capturing detailed structural information. Additionally, diffusion-weighted imaging (DWI) was acquired with an in-plane resolution of 2 mm and slice thickness ranging from 2 to 4 mm, enabling the assessment of water diffusion in fetal brain tissues. Functional MRI (fMRI) images were also included, featuring an isotropic resolution of 2-3 mm, allowing the study of brain activity and neural circuit activity development in the developing fetal brain.

Our dataset included a total of 38,038 2D MRI slices (from 100 subjects) for T2W imaging, 22,902 2D MRI slices (from 65 subjects) for DWI, and 4756 2D MRI slices (from 36 subjects) for fMRI. The fetal scans in this dataset span a wide gestational age range, ranging from 22 to 38 weeks, resulting in considerable variations in brain size and shape–approximately a five-fold increase in brain volume over this period. Moreover, the dataset exhibits diversity by encompassing a spectrum of conditions, including both typical and abnormal brains, various artifacts, and twin pregnancies. This diversity mirrors the complexities encountered in real-world fetal MRI imaging, facilitating robust evaluation and analysis of the developed algorithms. For ground truth annotations, a skilled annotator meticulously segmented the fetal brain on each MRI slice.

To ensure proper evaluation, we partitioned the data into three subsets: training, validation, and testing. This partitioning was performed on a subject-wise basis such that there was no overlap between subjects in different subsets. This was necessary to ensure independent evaluation of our model's performance. Additionally, we augmented our test dataset with a separate collection of fetal MRI scans sourced from different scanners and distinct imaging sites. This supplementary dataset was entirely excluded from the training phase, ensuring its independence from our model development process. This was critical to test the generalization performance of our proposed methods to unseen data from different sites and scanners. For a comprehensive overview of the data distribution across these subsets, please refer to Table I.

B. Model Architecture

In our study, we investigated three state of the art neural network architectures for medical image segmentation: U-Net [29], Dynamic U-Net [42], [43], which is an adaptation of the nnU-Net framework [30], and Attention U-Net [44].

U-Net [29] is a standard benchmark in the field of biomedical image segmentation. It features a symmetrical encoder-decoder structure, where the encoder extracts hierarchical features from the input data, and the decoder progressively upsamples and refines these features to produce the final segmentation map. U-Net is widely recognized for its adaptability through skip connections that concatenate feature maps from the encoder to

TABLE I
SUMMARY OF THE FETAL MRI DATA USED IN THIS WORK

Modality/Type	Subjects (Stacks, Slices)	Resolution (mm)	Scanner
Training			
T2W	44 (483, 19083)	$\sim 1 \times 1 \times 2$	Siemens3T
DWI	41 (492, 14597)	$\sim 2 \times 2 \times 2 - 4$	Siemens3T
fMRI	22 (85, 2790)	$\sim 2 - 3$ isotropic	Siemens3T
Validation			
T2W	9 (82, 3236)	$\sim 1 \times 1 \times 2$	Siemens3T
DWI	9 (108, 3260)	$\sim 2 \times 2 \times 2 - 4$	Siemens3T
fMRI	5 (25, 750)	$\sim 2 - 3$ isotropic	Siemens3T
Testing			
T2W/Typical	18 (143, 5362)	$\sim 1 \times 1 \times 2$	Siemens3T
T2W/Abnormality	4 (30, 1124)	$\sim 1 \times 1 \times 2$	Siemens3T
T2W/Artifacts	6 (34, 1238)	$\sim 1 \times 1 \times 2$	Siemens3T
T2W/Twins	3 (32, 922)	$\sim 1 \times 1 \times 2$	Siemens3T
T2W	3 (14, 391)	$\sim 1 \times 1 \times 2$	GE1.5T
T2W	5 (84, 2732)	$\sim 1 \times 1 \times 2$	Phillips1.5T
T2W	3 (28, 947)	$\sim 1 \times 1 \times 2$	Siemens1.5T
T2W/SiteW	5 (91, 3003)	$\sim 1 \times 1 \times 2$	Siemens3T
DWI/B0	14 (46, 1312)	$\sim 2 \times 2 \times 2 - 4$	Siemens3T
DWI/B1	14 (122, 3733)	$\sim 2 \times 2 \times 2 - 4$	Siemens3T
fMRI	9 (38, 1216)	$\sim 2 - 3$ isotropic	Siemens3T
fMRI/External	77 (477, 17649)	~ 3 isotropic	Siemens3T

We used three different MRI sequences including T2-weighted (T2W), diffusion-weighted MRI (DWI), and functional MRI (fMRI). T2W/Typical refers to most common seen data types in fetal T2W MRIs including normal and challenging cases. T2W/SiteW involved independent T2W scans acquired at a remote site. DWI/B0 represents the non-diffusion sensitized baseline images in DWI, while DWI/B1 refers to the collection of diffusion-sensitized images. fMRI/External refers to independent fMRI data that we accessed through Open-Neuro [22].

the decoder, allowing the model to combine multi-scale contextual information with fine-grained details. The adaptability of U-Net lies in its ability to capture both local and global context effectively. However, it may struggle with handling intricate anatomical structures and fine details in certain scenarios due to its fixed architecture and limited contextual awareness. In Fig. 2, we illustrate an instance of a U-Net architecture without attention gates, showcasing its standard structure for segmentation tasks.

Dynamic U-Net [42], [43] is an adaptation of the nnU-Net framework [30]. It showcases its capacity to autonomously customize its architecture according to the input data. Notably, Dynamic U-Net demonstrates intelligent adaptability in selecting optimal parameters such as kernel sizes, strides, and channel dimensions, all tailored to the specific input image size. This flexibility is expected to enhance network efficiency and its ability to seamlessly adapt to diverse datasets.

Attention U-Net [44] enhances the base U-Net by introducing an attention gate (depicted in Fig. 2) within the decoder section. The attention gate processes the encoder's feature map before concatenation in the decoder block, determining the significance of regions in the encoder feature map relative to the context of the preceding decoder block. This determination is facilitated by the multiplication of the encoder feature map with attention gate-computed weights, which range between 0 and 1, representing the neural network's focus on specific pixels. The attention mechanism essentially acts as a gatekeeper, facilitating
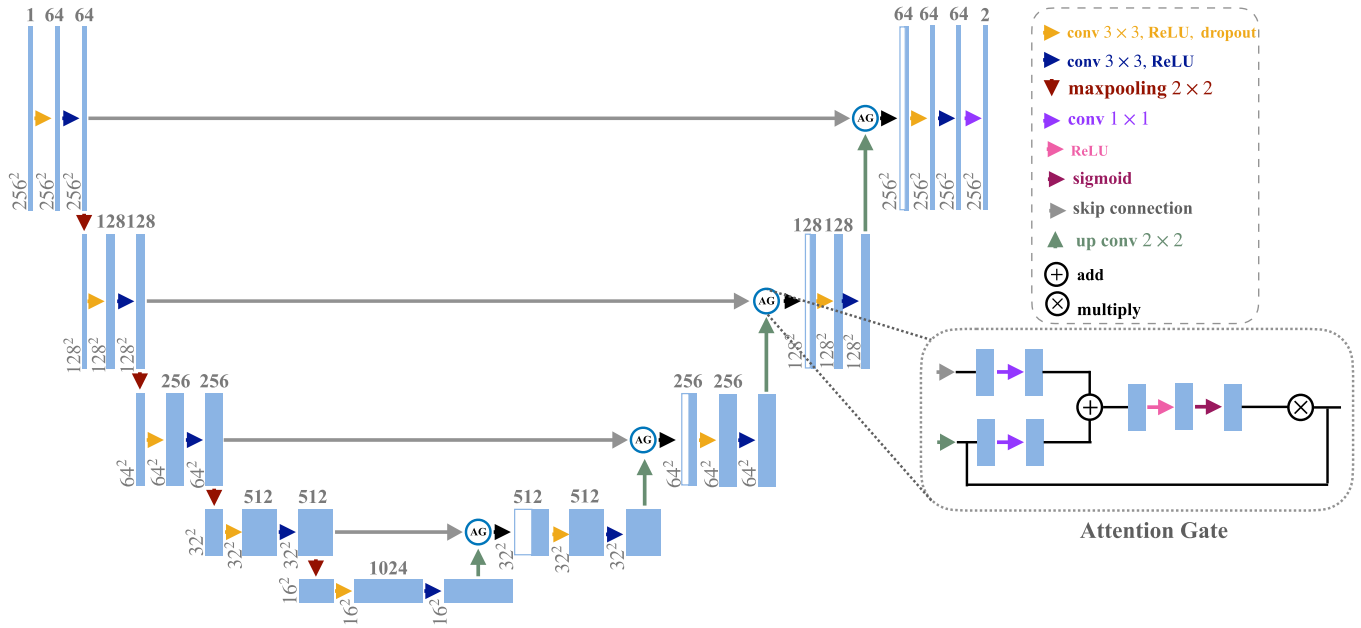
Fig. 2. Architecture of the U-Net with Attention Gates (AG) known as Attention U-Net. The backbone U-Net architecture can be achieved if the AG units are ignored. In the Attention U-Net, AGs filter the features that are propagated through the skip connections by using the contextual information of features extracted in coarser scales. This is achieved by adding the decoder output of a coarser scale to the output of every skip connection from the encoder after 1×1 convolutions. The output then passes through ReLU and sigmoid activation functions and is multiplied to the coarser level decoder input.

the model's ability to attend to the most salient features, ultimately improving segmentation accuracy by adaptively selecting and combining information from various spatial locations. Fig. 2 presents a detailed visual representation of the Attention U-Net structure, highlighting the integrated attention modules and demonstrating how they interact with the traditional U-Net (bypassing Attention Gate) layers to achieve more discerning and accurate segmentations.

C. Model Training and Inference

1) Data Processing and Augmentation: First, to establish uniformity across all three modalities (T2W, DWI, and fMRI) for training (only), we resampled each slice in two dimensions to a consistent pixel size of 1 mm and resized them to a uniform image size of 256×256 pixels. This ensured that all preprocessed slice images adhered to the same dimensions, denoted as (H, W, C), with H=W=256 representing width and height, and C=1 signifying the number of image channels. Furthermore, the data is normalized on a per-slice basis by its variance, ensuring uniform data scaling and facilitating improved model convergence.

Second, to bolster the robustness of our method against challenges that are inherent in fetal MRI, we used a comprehensive set of data augmentation techniques. We designed these techniques specifically to tackle the challenges that are raised by the widely variable position, orientation, size, and shape of the fetal brain and its surrounding structures, as well as the pronounced effects of fetal movements and maternal breathing motion. Depending on the sequence type, intermittent fetal and maternal motion often result in pronounced artifacts that may

include partial or complete signal loss, in-plane blur, ringing, slice cross-talk, and spin-history artifacts.

In our image augmentation pipeline, we applied a range of transformations to enhance the diversity of our training dataset. Specifically, we generated augmented scans using the following techniques:

1. Spatial Augmentation: This category consisted of five spatial transformations, including random flips, rotations, zooming, and affine transformations. For each augmentation technique, we applied these transformations with a probability of 0.5 or 0.6, resulting in a significant increase in the variability of the dataset.
2. Intensity Augmentation: In this category, we introduced variations in image intensity through three distinct techniques. Gaussian noise was added with a standard deviation of 0.4 and a probability of 0.5. Multiplicative bias fields were incorporated with a degree of 4 and coefficients ranging between 0.05 and 0.1, applied with a probability of 0.6. Gaussian smoothing was performed with sigma values ranging from 0.5 to 1.0 and a probability of 0.4.

These augmentations collectively contributed to a versatile training dataset, allowing our deep learning models to better adapt to diverse fetal brain MRI images during training.

2) Training: In the training stage, we initiated the process by using an independent validation dataset to determine the best hyperparameters. Subsequently, we trained the models with the training dataset. All models were trained with a batch size of 8 and input image size fixed at 256×256. The learning rate for each of the compared networks was fine-tuned separately. For the Attention U-Net, we set a learning rate of 1×10^{-4}. We

conducted training for 300 epochs using Adam optimization [45] of stochastic gradient descent.

To guide the training process, we utilized a weighted sum of binary cross-entropy and Dice loss. Specifically, we employed the batched variant of the Dice loss, where the loss was calculated over all samples in the batch, as opposed to averaging the Dice loss for each 2D sample individually. This approach considers samples in the batch as a pseudo-volume and computes the loss function over all voxels in the batch [30]. This was done for improved computations only, as models only worked on 2D slices. Since the original fetal MRI acquisitions were 2D (slice-based), to minimize the effects of inter-slice motion, no 3D context was used in training, inference, or evaluation. We used foreground Dice, which measures the overlap between the predicted and ground truth segmentations. Our loss function was

$$L_{total} = 0.4 \times L_{CE} + 0.6 \times L_{dice} \qquad (1)$$

$$L_{CE} = -\sum_{i=1}^{n} v_i \log(u_i) + (1 - v_i) \log(1 - u_i) \quad (2)$$

$$L_{dice}(u, v) = -\frac{2}{|K|} \sum_{k \in K} \frac{\sum_{i \in I} u_i^k v_i^k}{\sum_{i \in I} u_i^k + \sum_{i \in I} v_i^k} \qquad (3)$$

where u is the softmax output of the network and v is a one hot encoding of the ground truth segmentation map. Both u and v have shape $I \times K$ with $i \in I$ being the number of pixels in the training batch and $k \in K$ being the classes.

3) Inference: During the inference stage, we apply the same resampling and normalization procedures as in the training data (see Section: III-C1), ensuring consistency in data preparation. However, unlike the training dataset, we do not perform resizing on these datasets. We utilize a sliding window approach with a window size matching the training patch size (256×256), allowing flexibility in input image sizes. Each consecutive window overlaps the previous one by half its size, and predictions within these overlapping regions are averaged to produce the final prediction. This method ensures consistent and accurate results across varying input sizes.

4) Implementation: All our experiments and model training were conducted with NVIDIA RTX A5000 GPUs on a workstation with 128 GB of system memory. Our implementations leveraged the PyTorch framework [46] and harnessed the capabilities of the MONAI toolkit [47].

5) Evaluation: We evaluated the performance of all methods using four key metrics: the Dice Similarity Coefficient (DSC), Intersection-over-Union (IoU), Average Surface Distance (ASD), and 95th percentile Hausdorff Distance (HD95). DSC and IoU quantify the overall overlap between the predicted and ground truth segmentations, while ASD and HD95 assess the spatial accuracy and worst-case scenario, respectively. DSC, IoU, ASD, and HD95 are defined as [48], [49]:

$$DSC(P, R) = \frac{2|P \cap R|}{|P| + |R|} = \frac{2TP}{2TP + FP + FN} \qquad (4)$$

$$IoU(P, R) = \frac{|P \cap R|}{|P \cup R|} = \frac{TP}{TP + FP + FN} \qquad (5)$$

$$ASD(P, R) = \frac{1}{|S(P)| + |S(R)|} \left(\sum_{p \in S(P)} \min_{r \in S(R)} d(p, r) \right.$$

$$\left. + \sum_{r \in S(R)} \min_{p \in S(P)} d(r, p) \right) \qquad (6)$$

$$HD95(P, R) = \max(h(P, R), h(R, P)) \qquad (7)$$

where P is the predicted mask, R is the Reference (ground truth) mask, TP, FP, and FN are the true positive, false positive, and false negative rates, respectively, $S(P)$ and $S(R)$ are the surface points of the predicted and ground truth segmentations, $d(p, r)$ is the Euclidean distance between points p and r, and $h(P, R)$ is the directed Hausdorff distance from P to R, defined as the 95th percentile of the distances from each point in $S(P)$ to its nearest point in $S(R)$.

IV. RESULTS

In this section, we first present the results of training the three models, i.e., the U-Net, the dynamic U-Net (DynU-Net), and the attention U-Net (AttU-Net) on all sequences. We compared the performance of the three models on the different test sets that we built. Next, we conducted ablation studies to evaluate the relative impact of our proposed feature learning across multiple MRI modalities and data augmentation strategies. To this end, we compared the best model performance with models trained on each individual sequence, as well as models trained without data augmentation. Specifically, we assessed how our best model trained on all sequences performed compared to models trained on each sequence individually. This comparative analysis provides insights into the impact of training across multiple MRI modalities and data augmentation, demonstrating how they increase the robustness and generalizability of deep learning models.

Table II provides a detailed breakdown of Dice and IoU scores for each of the considered architectures and sequences when evaluated on an independent test dataset. In this experiment, all models were trained on a diverse dataset encompassing all three sequences (T2W, DWI, and fMRI). Overall, the results show that the standard U-Net and Attention U-Net models are very accurate in extracting the fetal brain in all MRI sequences. Notably, U-Net and Attention U-Net achieved a DSC of 95.72% and 95.70% and an IoU of 92.06% and 92.03% for the T2W sequence respectively, demonstrating their effectiveness in fetal brain extraction on this sequence. For the DWI sequence, Attention U-Net achieved a DSC of 93.27% and an IoU of 87.64%, whereas U-Net achieved a DSC of 92.58% and an IoU of 86.66%, showcasing Attention U-Net's robustness across diverse MRI sequences. In the case of fMRI, the Attention U-Net model yielded a DSC of 95.25% and an IoU of 90.99%, underlining its capability to handle complex functional MRI data. While the dynamic U-Net also delivered commendable results, with slightly lower DSC and IoU scores, this could be attributed to its inherent sensitivity to dynamic changes in the input data. The dynamic U-Net adapts its architecture to the

TABLE II
AVERAGE DICE (DICE), AVERAGE IoU, AND INFERENCE TIME ON TEST DATA FOR DIFFERENT MODELS

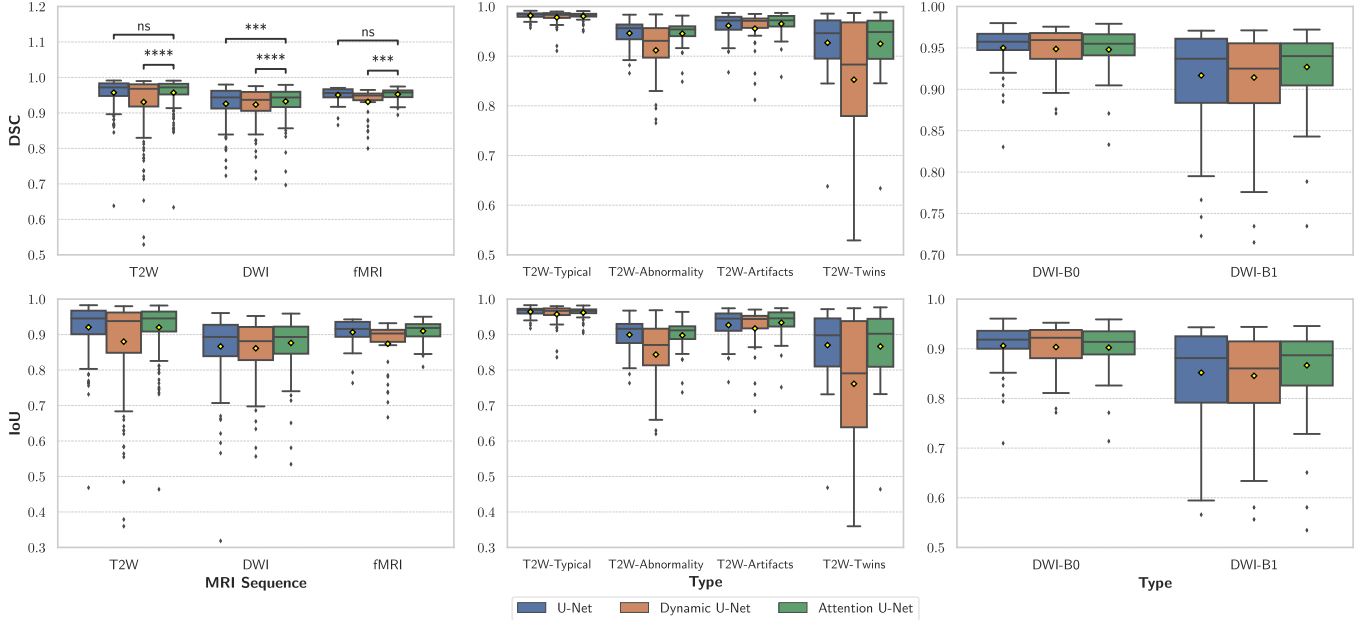| Data	T2W		DWI		fMRI		Inference Time
	Dice (%)	IoU (%)	Dice (%)	IoU (%)	Dice (%)	IoU (%)	(ms)
U-Net [29]	95.72±4.24	92.06±6.93	92.58±5.83	86.66±8.78	95.05±2.27	90.66±3.96	8.18±0.74
DynUNet [30]	93.07±8.31	88.01±12.56	92.36±4.63	86.14±7.54	93.12±4.18	87.40±6.89	10.49±0.72
AttU-Net [50]	**95.70±4.28**	**92.03±6.97**	**93.27±4.11**	**87.64±6.69**	**95.25±1.77**	**90.99±3.16**	11.48±1.11

Fig. 3. Boxplots of Dice Similarity Coefficient (DSC) and Intersection-over-Union (IoU) for different sequences (T2-weighted (T2W), diffusion-weighted (DWI), and functional MRI (fMRI)), T2W data characteristics (Typical, Abnormalities, Artifacts, and Twins pregnancies), and model architectures (U-Net, Dynamic U-Net, and Attention U-Net). Higher DSC and IoU values indicate greater segmentation accuracy. The U-Net and Attention U-Net models achieved higher median Dice scores overall compared to the Dynamic U-Net model. The asterisks displayed on the top left plot serve as visual indicators of the statistical significance associated with the differences observed between the groups using a paired t-test. The asterisks displayed on the first plot serve as visual indicators of the statistical significance associated with the differences observed between the groups. Significance levels: (ns) $p > 0.05$, (*) $0.01 < p \leq 0.05$, (**) $0.001 < p \leq 0.01$, (***) $0.0001 < p \leq 0.001$, (****) $p \leq 0.0001$.

input size dynamically, which might lead to minor fluctuations in performance [42].

Fig. 3 shows the boxplots that provide an illustration of the distribution and variability of the DSC and IoU scores across the different test datasets for each of the trained models. The first column of boxplots in Fig. 3 represents results obtained on all T2W, DWI, and fMRI test data. The second column offers a detailed analysis of the performance of the three models on the T2W images including typical fetal brains, abnormal cases, images with artifacts, and twin pregnancies. The boxplots provide insights into the adaptability of the trained models to varying T2W data scenarios. In the third column, we focused on the evaluation of the DWI sequences, specifically considering B0 and B1 data.

We used paired t-tests to assess the statistical significance of the difference between the performance of different models. For comparing U-Net with Attention U-Net, the p-value for T2W, DWI, and fMRI were, respectively, $8.585e-01$, $2.643e-04$, and $2.250e-01$. For comparing Dynamic U-Net and Attention U-Net, the p-values were $2.957e-08$, $5.301e-06$, and $8.494e-04$, respectively. The asterisks displayed on the top left plot serve as visual indicators of the statistical significance

associated with the differences observed between the groups. (ns) indicate no significant difference (p-value > 0.05). (∗) suggests moderate significance ($0.01 <$ p-value ≤ 0.05). (∗∗) signify high significance ($0.001 <$ p-value ≤ 0.01). (∗ ∗ ∗) represent very high significance ($0.0001 <$ p-value ≤ 0.001). (∗ ∗ ∗∗) indicate extremely high significance (p-value ≤ 0.0001).

Table III presents the average ASD and HD95 values along with their standard deviations for the U-Net, Dynamic U-Net, and Attention U-Net models on the T2W, DWI, and fMRI test datasets. The Attention U-Net model achieved the lowest ASD and HD95 values across all three MRI sequences, demonstrating its superior segmentation accuracy compared to the other models. The U-Net model also performed well, with slightly higher ASD and HD95 values than the Attention U-Net. The Dynamic U-Net model generally exhibited higher ASD and HD95 values, indicating lower segmentation accuracy.

Fig. 4 shows the boxplots illustrating the distribution of ASD and HD95 values for different MRI sequences (T2W, DWI, and fMRI), T2W data characteristics (typical, abnormalities, artifacts, and twins pregnancies), and model architectures (U-Net, Dynamic U-Net, and Attention U-Net). The U-Net and Attention U-Net models generally achieved lower ASD and HD95 values

TABLE III
AVERAGE SURFACE DISTANCE (ASD) AND 95TH PERCENTILE HAUSDORFF DISTANCE (HD95) FOR DIFFERENT MODELS

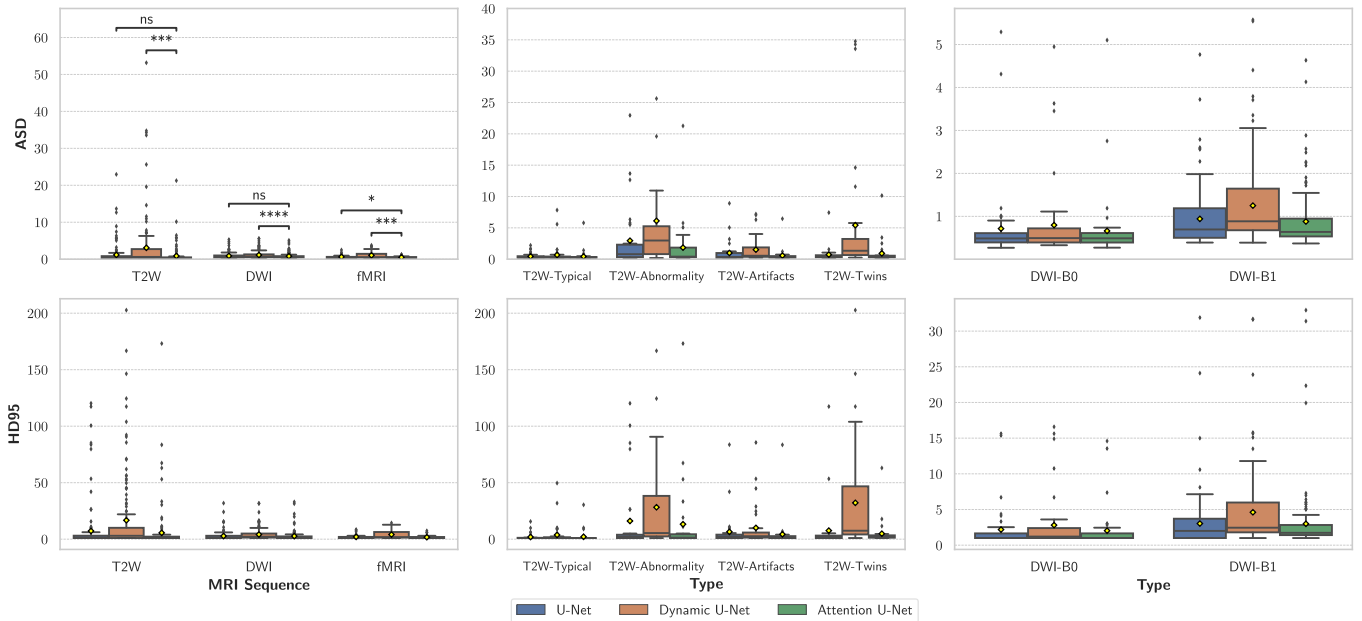| Data	T2W		DWI		fMRI	
	ASD	HD95	ASD	HD95	ASD	HD95
U-Net [29]	1.16 ± 2.72	7.25 ± 20.35	0.88 ± 0.76	2.80 ± 3.71	0.62 ± 0.49	2.03 ± 1.87
DynUNet [30]	3.08 ± 7.18	16.73 ± 33.51	1.13 ± 0.97	4.12 ± 4.87	1.05 ± 0.93	4.10 ± 3.84
AttU-Net [50]	$\mathbf{0.87 \pm 2.16}$	$\mathbf{5.59 \pm 18.28}$	$\mathbf{0.82 \pm 0.71}$	$\mathbf{2.73 \pm 4.28}$	$\mathbf{0.51 \pm 0.28}$	$\mathbf{1.76 \pm 1.48}$

Fig. 4. Boxplots of Average Surface Distance (ASD) and 95th percentile Hausdorff Distance (HD95) for different MRI sequences (T2-weighted (T2W), diffusion-weighted (DWI), and functional MRI (fMRI)), T2W data characteristics (Typical, Abnormalities, Artifacts, and Twins pregnancies), and model architectures (U-Net, Dynamic U-Net, and Attention U-Net). Lower ASD and HD95 values indicate better segmentation accuracy. The U-Net and Attention U-Net models generally achieved lower ASD and HD95 values compared to the Dynamic U-Net model. However, the HD95 values were relatively high for some cases, particularly in the T2W images, likely due to the presence of missed or misaligned slices caused by motion artifacts. The asterisks displayed on the top left plot serve as visual indicators of the statistical significance associated with the differences observed between the groups using a paired t-test. Significance levels: (ns) $p > 0.05$, (*) $0.01 < p \leq 0.05$, (**) $0.001 < p \leq 0.01$, (***) $0.0001 < p \leq 0.001$, (****) $p \leq 0.0001$.

compared to the Dynamic U-Net model, indicating better segmentation accuracy. However, the HD95 values were relatively high for some cases, particularly in the T2W images. This can be attributed to the presence of missed or misaligned slices caused by motion artifacts, which are more common in T2W acquisitions compared to DWI and fMRI. These artifacts can introduce discrepancies between the ground truth and predicted segmentations, leading to higher HD95 values. Additionally, the HD95 metric is sensitive to outliers and can be greatly affected by even a small number of misaligned or missing slices, as it represents the 95th percentile of the Hausdorff distances between the two segmentations. In contrast, the ASD metric provides an average measure of surface distance, making it more robust to such outliers.

In the ablation studies we focused on gauging the effectiveness of 1) our models trained on multiple sequences when compared to models trained on each sequence separately, and 2) our image augmentation strategies. The DSC and IoU plots on the left side of Fig. 5 illustrate the results of our ablation studies on the test sets of every sequence (T2W, DWI, fMRI). Each box plot represents the performance distribution of models trained

on individual sequences (Attention U-Net, Single Aug), a model trained on all sequences but without data augmentation (Attention U-Net, All, No Aug), and our final model (Attention U-Net, All, Aug), which used all the sequences along with our data augmentation for training. The results show that our final model, Attention U-Net (All, Aug), performed significantly better than Attention U-Net without data augmentation (All, No Aug) for all of the sequences, and performed significantly better than Attention U-Net (Single, Aug) for the DWI and fMRI sequences.

The DSC and IoU plots on the right side of Fig. 5 show the results on set-aside test sets from different scanners. Since no images from those scanners or sites were included in the training dataset, this served as a test of the generalization performance of the models for data from new domains. Overall, the results show that our final model, which exploited feature learning across multiple MRI modalities and data augmentation, performed significantly better than its counterparts that did not use these strategies. The results in Fig. 5 vividly display how the aggregated model outperformed its counterparts across various test datasets, underscoring the advantages of feature learning across
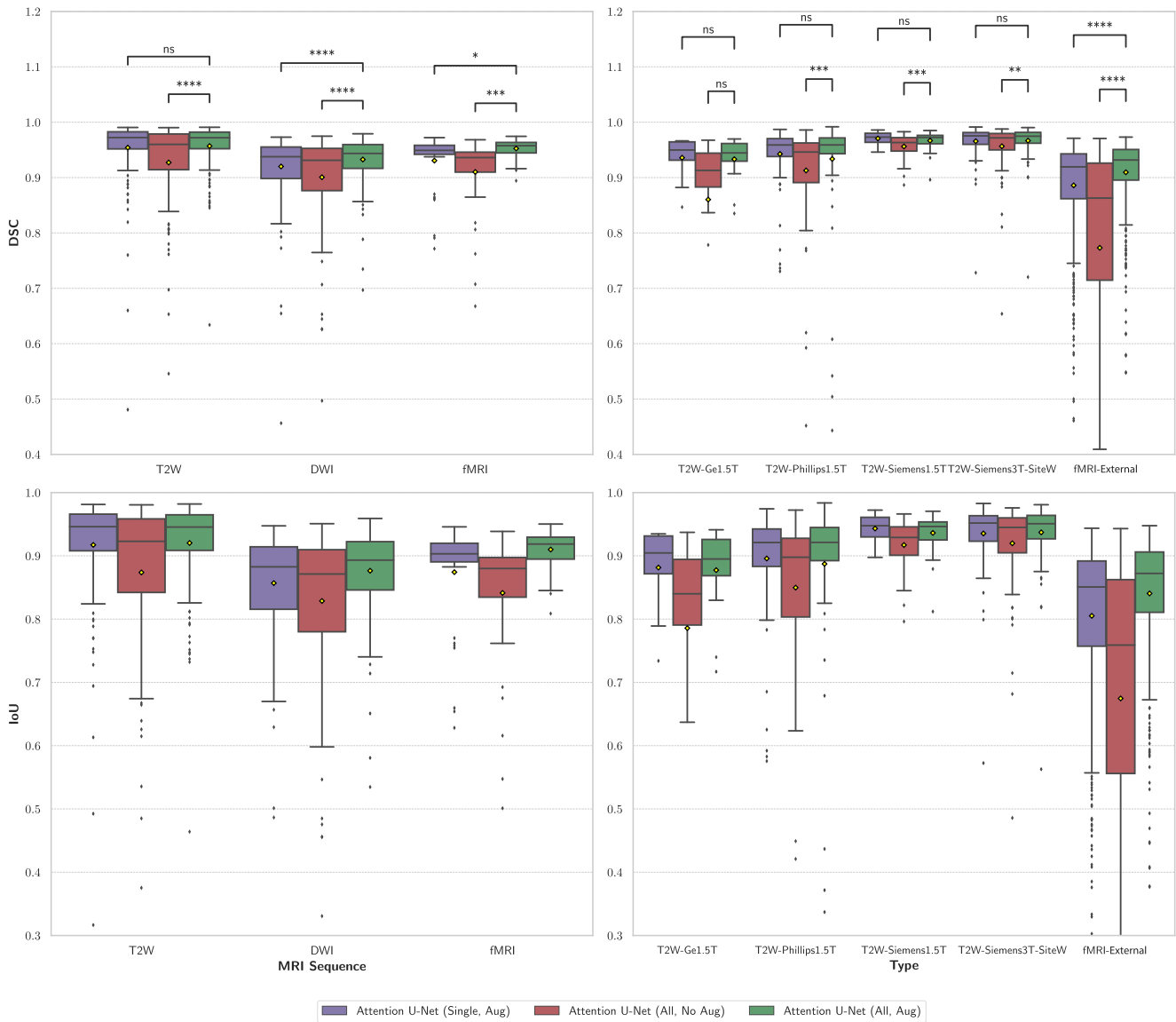
Fig. 5. Boxplots of Dice similarity coefficient (DSC) and Intersection-over-Union (IoU) for different MRI sequences (T2-weighted (T2W), diffusion-weighted (DWI), and functional MRI (fMRI)), comparing the extraction performance of different Attention U-Net model architectures ((Single, Aug): trained on a single sequence (corresponding) with augmentation, (All, No Aug): trained on all sequences with no augmentation, and (All, Aug): trained on all sequences with augmentation). Left plots show the results on our test data and right plots show the results on our out-of-distribution test data. The best result is achieved by Attention U-Net trained on all sequences with augmentation. This indicates that leveraging multiple complementary sequences and expanded training data through augmentation can enhance deep learning model performance.

multiple MRI modalities and data augmentation in improving generalization.

Figs. 6 and 7 provide sample representative outcomes of our trained model (Fetal-BET) on a variety of test data including images of twins and brains with abnormalities on T2w images as well as DWI and fMRI scans.

V. DISCUSSION

Our findings underscore the effectiveness of our approach in fetal brain extraction. In particular, our experiments with three powerful deep convolutional neural network architectures trained with multiple sequences and comprehensive data augmentation strategies demonstrate that we can achieve accurate automatic fetal brain extraction. Both U-Net and the Attention U-Net models exhibited high DSC and IoU scores, particularly on T2W images, with Attention U-Net showcasing robust generalizability across various MRI contrasts and outperforming the other two models on DWI and fMRI. Importantly, our ablation study results, as depicted in Fig. 5, highlight the advantage of training across multiple MRI modalities and data augmentation. The model trained on the amalgamation of all three sequences consistently outperformed models trained individually on each sequence when tested on that specific sequence alone. In several instances, this difference was statistically significant, particularly in achieving higher DSC scores for the more challenging sequences, i.e., DWI and fMRI, when compared to models trained on a single modality.
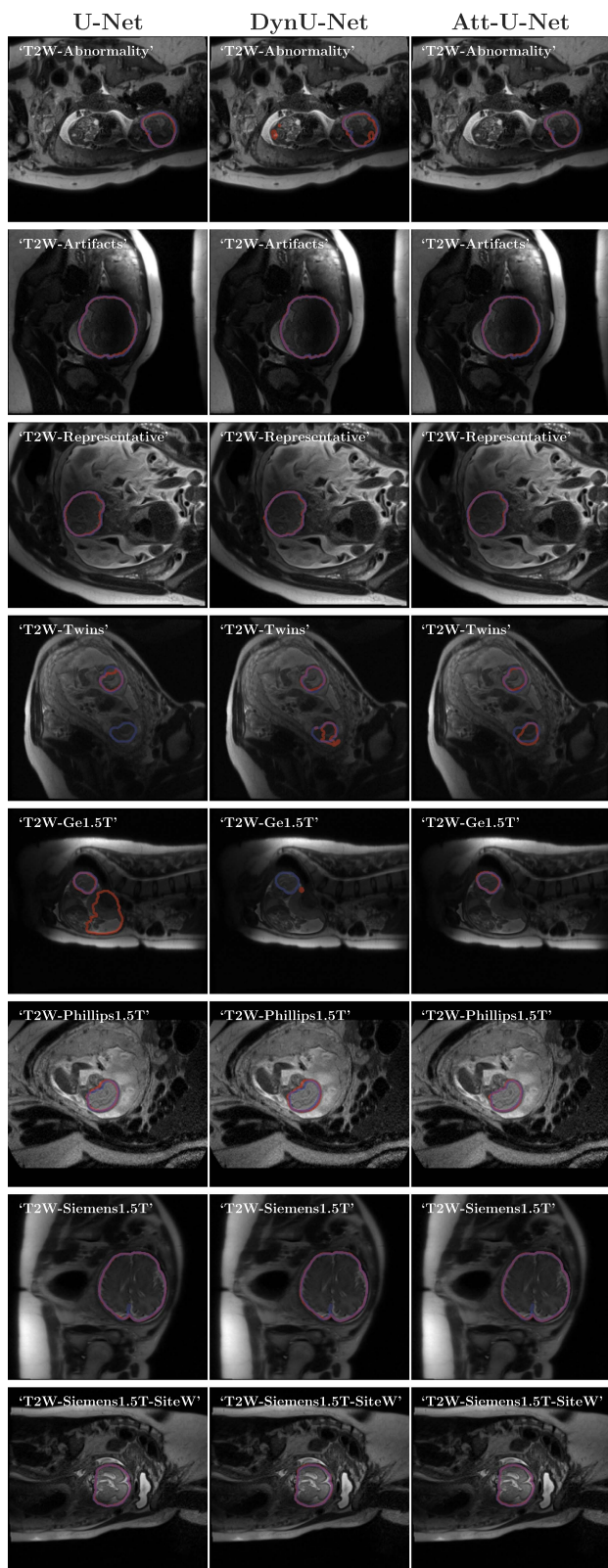
Fig. 6. Representative examples of segmentations produced by different models (U-Net, DynU-Net, AttU-Net) on T2w slices. On each image, the blue curve shows the outline of the reference brain mask drawn manually by an experienced annotator, while the red curve shows the contour of the segmentation mask predicted by the deep learning method.
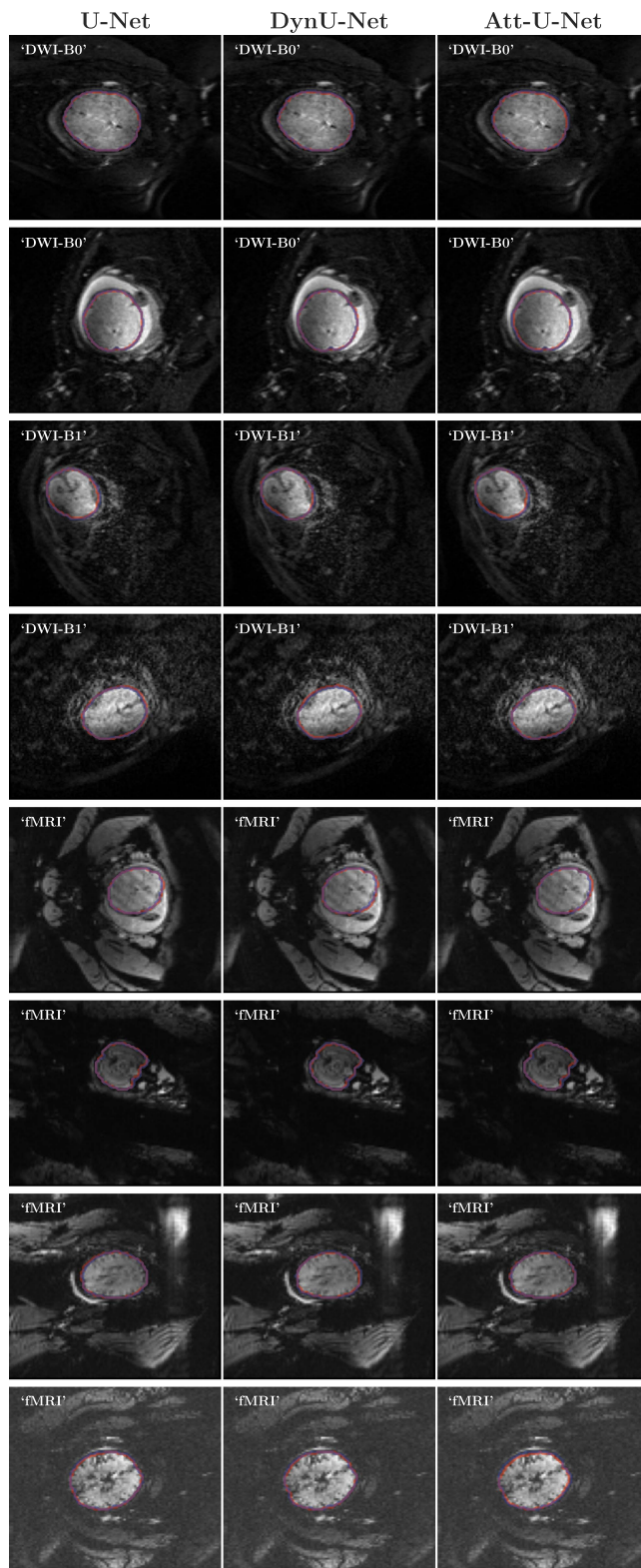

Fig. 7. Representative examples of segmentations produced by different models (U-Net, DynU-Net, AttU-Net) on DWI and fMRI test slices. The blue curves show the outline of the reference (ground truth) brain mask, while the red curve shows the contour of the mask predicted by the deep learning method.

While our deep learning models have demonstrated promising results in fetal brain extraction, there are certain limitations to consider. One important limitation is the potential challenge in handling rare pathological cases that may not be well-represented in the training data. Although we have included a diverse range of pathologies in our dataset, there may be some rare or complex cases that the models have not encountered during training, which could affect their performance in such scenarios. Another limitation is the need for further validation on larger, multi-center datasets to assess the generalizability of our models across different populations, scanner types, and imaging protocols. While we have evaluated our models on data from multiple scanners and sites, a more extensive validation would provide additional insights into their robustness and potential for clinical adoption. Moreover, the training and inference of deep learning models often require significant computational resources, which may be a consideration for widespread implementation in clinical settings with limited infrastructure. Future work could explore techniques for model compression and optimization to improve computational efficiency without compromising performance.

In summary, our study addresses the complex task of fetal brain extraction in fetal MRI analysis. We have developed an innovative solution using deep learning techniques, attention mechanisms, feature learning across multiple MRI modalities, and data augmentation. Crucially, we have created a substantial and diverse dataset that includes various MRI sequences and pathological cases, significantly contributing to the advancement of our approach. Through rigorous evaluation, we demonstrated the reliability and robustness of our method, achieving accurate fetal brain extraction across different scanning conditions, stages of pregnancy, and brain conditions. Despite the limitations discussed above, the adaptability and precision of our deep learning model hold significant promise for the field of fetal brain imaging and analysis. By overcoming long-standing challenges in fetal brain extraction, our work has the potential to improve automatic workflows for quantitative fetal MRI analysis, including, for example, image reconstruction and segmentation. Therefore, it has the potential to profoundly impact clinical practices and advance our understanding of prenatal brain development and developmental disorders, ultimately improving the quality of prenatal care and diagnostics.

VI. Conclusion

In conclusion, our study presents a comprehensive evaluation of fetal brain extraction techniques using a range of MRI sequences. We have demonstrated the efficacy of both U-Net and Attention U-Net architectures, with Attention U-Net excelling, particularly in challenging sequences like DWI and fMRI. The advantages of learning across multiple MRI modalities and augmentation training were clearly evident, with the aggregated model consistently outperforming individual sequence models. As we move forward, further research may explore fine-tuning these models for specific clinical applications and expanding the dataset to encompass even more variations. The promising results obtained in this study provide a strong foundation for

future advancements in the field of fetal brain extraction, ultimately benefiting both healthcare professionals and expectant parents in ensuring the well-being of unborn children.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

R.F. developed the methodology, conducted experiments, performed data preprocessing, contributed to writing the original draft and revisions, and implemented coding. D.K. helped with idea development, provided feedback, and contributed to writing the original draft and revisions. D.E. provided supervision and feedback. A.G. contributed to idea development, supervised the project, provided feedback, contributed to revisions, and assisted with writing. All authors contributed to the final version of the manuscript.

ACKNOWLEDGMENT

REFERENCES

[1] M. Rutherford et al., "MR imaging methods for assessing fetal brain development," *Devlop. Neurobiol.*, vol. 68, no. 6, pp. 700–711, 2008.

[2] A. Gholipour et al., "Fetal MRI: A technical update with educational aspirations," *Concepts Magn. Reson. Part A*, vol. 43, no. 6, pp. 237–266, 2014.

[3] F. Machado-Rivas et al., "Fetal MRI at 3 T: Principles to optimize success," *Radiographics*, vol. 43, no. 4, 2023, Art. no. e220141.

[4] C. Calixto et al., "Advances in fetal brain imaging," *Magn. Reson. Imag. Clin.*, 2024.

[5] B. Marami et al., "Temporal slice registration and robust diffusion-tensor reconstruction for improved fetal brain structural connectivity analysis," *NeuroImage*, vol. 156, pp. 475–488, 2017.

[6] S. Tourbier et al., "Automated template-based brain localization and extraction for fetal brain MRI reconstruction," *NeuroImage*, vol. 155, pp. 460–472, 2017.

[7] J. Xu, D. Moyer, P. E. Grant, P. Golland, J. E. Iglesias, and E. Adalsteinsson, "SVoRT: Iterative transformer for slice-to-volume registration in fetal brain MRI," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2022, pp. 3–13.

[8] A. Gholipour, J. A. Estroff, and S. K. Warfield, "Robust super-resolution volume reconstruction from slice acquisitions: Application to fetal brain MRI," *IEEE Trans. Med. Imag.*, vol. 29, no. 10, pp. 1739–1758, Oct. 2010.

[9] M. Kuklisova-Murgasova, G. Quaghebeur, M. A. Rutherford, J. V. Hajnal, and J. A. Schnabel, "Reconstruction of fetal brain MRI with intensity matching and complete outlier removal," *Med. Image Anal.*, vol. 16, no. 8, pp. 1550–1564, 2012.

[10] B. Kainz et al., "Fast volume reconstruction from motion corrupted stacks of 2D slices," *IEEE Trans. Med. Imag.*, vol. 34, no. 9, pp. 1901–1913, Sep. 2015.

[11] M. Ebner et al., "An automated localization, segmentation and reconstruction framework for fetal brain MRI," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2018, pp. 313–320.

[12] A. Uus et al., "Deformable slice-to-volume registration for motion correction of fetal body and placenta MRI," *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2750–2759, Sep. 2020.

[13] F. Rousseau et al., "Registration-based approach for reconstruction of high-resolution in utero fetal MR brain images," *Academic Radiol.*, vol. 13, no. 9, pp. 1072–1081, 2006.

[14] R. Faghihpirayesh, D. Karimi, D. Erdogmus, and A. Gholipour, "Automatic brain pose estimation in fetal MRI," *Proc. SPIE*, vol. 12464, pp. 149–154, 2023.

[15] J. Xu et al., "NeSVoR: Implicit neural representation for slice-to-volume reconstruction in MRI," *IEEE Trans. Med. Imag.*, vol. 42, no. 6, pp. 1707–1719, Jun. 2023.

[16] A. U. Uus, A. E. Collado, T. A. Roberts, J. V. Hajnal, M. A. Rutherford, and M. Deprez, "Retrospective motion correction in foetal MRI for clinical applications: Existing methods, applications and integration into clinical practice," *Brit. J. Radiol.*, vol. 96, no. 1147, 2023, Art. no. 20220071.

[17] M. Ison, R. Donner, E. Dittrich, G. Kasprian, D. Prayer, and G. Langs, "Fully automated brain extraction and orientation in raw fetal MRI," in *Proc. Workshop Paediatric Perinatal Imag.*, 2012, pp. 17–24.

[18] K. Keraudren et al., "Automated fetal brain segmentation from 2D MRI slices for motion correction," *NeuroImage*, vol. 101, pp. 633–643, 2014.

[19] L. Sun et al., "Multi-scale multi-hierarchy attention convolutional neural network for fetal brain extraction," *Pattern Recognit.*, vol. 133, 2023, Art. no. 109029.

[20] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, 2017.

[21] S. S. M. Salehi et al., "Real-time automatic fetal brain extraction in fetal MRI by deep learning," in *Proc. IEEE 15th Int. Symp. Biomed. Imag.*, 2018, pp. 720–724.

[22] S. Rutherford et al., "Automated brain masking of fetal functional MRI with open data," *Neuroinformatics*, vol. 20, no. 1, pp. 173–185, 2022.

[23] A. Hoopes, J. S. Mora, A. V. Dalca, B. Fischl, and M. Hoffmann, "Synthstrip: Skull-stripping for any brain image," *NeuroImage*, vol. 260, 2022, Art. no. 119474.

[24] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Auto-context convolutional neural network (Auto-Net) for brain extraction in magnetic resonance imaging," *IEEE Trans. Med. Imag.*, vol. 36, no. 11, pp. 2319–2330, Nov. 2017.

[25] T. Ciceri et al., "Review on deep learning fetal brain segmentation from magnetic resonance images," *Artif. Intell. Med.*, vol. 143, 2023, Art. no. 102608.

[26] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.

[27] V. Taimouri, A. Gholipour, C. Velasco-Annis, J. A. Estroff, and S. K. Warfield, "A template-to-slice block matching approach for automatic localization of brain in fetal MRI," in *Proc. IEEE 12th Int. Symp. Biomed. Imag.*, 2015, pp. 144–147.

[28] S. Tourbier et al., "Automatic brain extraction in fetal MRI using multi-atlas-based segmentation," *Proc. SPIE*, vol. 9413, pp. 248–254, 2015.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. image Comput. Comput.- Assist. Interv.*, 2015, pp. 234–241.

[30] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[31] J. Lou et al., "Automatic fetal brain extraction using multi-stage U-Net with deep supervision," in *Proc. Mach. Learn. Med. Imaging: 10th Int. Workshop, MLMI 2019, Held Conjunction MICCAI*, Shenzhen, China, 2019, pp. 592–600.

[32] J. Li et al., "Automatic fetal brain extraction from 2D in utero fetal MRI slices using deep neural network," *Neurocomputing*, vol. 378, pp. 335–349, 2020.

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[34] G. Dudovitch et al., "Deep learning automatic fetal structures segmentation in MRI scans with few annotated datasets," in *Proc. 2020 Med. Image Comput. Comput. Assist. Interv.–MICCAI, 23rd Int. Conf.*, Lima, Peru, 2020, pp. 365–374.

[35] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Med. Image Comput. Comput.-Assist. Interv.–MICCAI, 19th Int. Conf.*, Athens, Greece, 2016, pp. 424–432.

[36] L. Liao et al., "Joint image quality assessment and brain extraction of fetal MRI using deep learning," in *Proc. Med. Image Comput. Comput. Assist. Interv.–MICCAI, 23rd Int. Conf.*, Lima, Peru, 2020, pp. 415–424.

[37] X. Zhang et al., "Confidence-aware cascaded network for fetal brain segmentation on MR images," in *Proc. Med. Image Comput. Comput. Assist. Interv.–MICCAI, 24th Int. Conf.*, Strasbourg, France, 2021, pp. 584–593.

[38] N. Khalili et al., "Automatic segmentation of the intracranial volume in fetal MR images," in *Proc. Fetal, Infant Ophthalmic Med. Image Anal.*, 2017, pp. 42–51.

[39] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. N. L. Benders, and I. Išgum, "Automatic segmentation of MR brain images with a convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1252–1261, May 2016.

[40] N. Khalili et al., "Automatic extraction of the intracranial volume in fetal and neonatal MR scans using convolutional neural networks," *NeuroImage: Clin.*, vol. 24, 2019, Art. no. 102061.

[41] R. Faghihpirayesh, D. Karimi, D. Erdoğmuş, and A. Gholipour, "Deep learning framework for real-time fetal brain segmentation in MRI," in *Proc. Perinatal, Preterm Paediatric Image Analysis: 7th Int. Workshop, PIPPI 2022, Held Conjunction MICCAI 2022*, Singapore, 2022, pp. 60–70.

[42] M. Ranzini, L. Fidon, S. Ourselin, M. Modat, and T. Vercauteren, "Monaifbs: Monai-based fetal brain MRI deep learning segmentation," 2021, *arXiv:2103.13314*.

[43] M. Futrega, A. Milesi, M. Marcinkiewicz, and P. Ribalta, "Optimized U-Net for brain tumor segmentation," in *Proc. Int. MICCAI Brainlesion Workshop*, 2021, pp. 15–29.

[44] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[46] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 8024–8035.

[47] M. J. Cardoso et al., "Monai: An open-source framework for deep learning in healthcare," 2022, *arXiv:2211.02701*.

[48] V. Yeghiazaryan and I. Voiculescu, "Family of boundary overlap metrics for the evaluation of medical image segmentation," *J. Med. Imag.*, vol. 5, no. 1, 2018, Art. no. 015006.

[49] T. Heimann et al., "Comparison and evaluation of methods for liver segmentation from CT datasets," *IEEE Trans. Med. Imag.*, vol. 28, no. 8, pp. 1251–1265, Aug. 2009.

[50] R. P. K. Poudel, S. Liwicki, and R. Cipolla, "Fast-SCNN: Fast semantic segmentation network," 2019, *arXiv:1902.04502*.