Automatic Segmentation of Posterior Fossa Structures in Pediatric Brain MRIs

Hugo Oliveira*, Larissa Penteado*, José Luiz Maciel*, Suely Fazio Ferraciolli†
Marcelo Straus Takahashi†, Isabelle Bloch‡, Roberto Cesar Junior*
*Instituto de Matemática e Estatística
Universidade de São Paulo, R. do Matão, 1010, São Paulo, Brazil
Email: {oliveirahugo, lariop, joselmp}@ime.usp.br, rmcesar@usp.br
†Hospital das Clínicas, Faculdade de Medicina
Universidade de São Paulo, Av. Dr. Enéas Carvalho de Aguiar, São Paulo, Brazil
Email: {suely.ferraciolli, marcelo.straus}@hc.fm.usp.br
‡Sorbonne Université, CNRS, LIP6
F-75005 Paris, France
Email: isabelle.bloch@sorbonne-universite.fr

***Abstract*—Pediatric brain MRI is a useful tool in assessing the healthy cerebral development of children. Since many pathologies may manifest in the brainstem and cerebellum, the objective of this study was to have an automated segmentation of pediatric posterior fossa structures. These pathologies include a myriad of etiologies from congenital malformations to tumors, which are very prevalent in this age group. We propose a pediatric brain MRI segmentation pipeline composed of preprocessing, semantic segmentation and post-processing steps. Segmentation modules are composed of two ensembles of networks: generalists and specialists. The generalist networks are responsible for locating and roughly segmenting the brain areas, yielding regions of interest for each target organ. Specialist networks can then improve the segmentation performance for underrepresented organs by learning only from the regions of interest from the generalist networks. At last, post-processing consists in merging the specialist and generalist networks predictions, and performing late fusion across the distinct architectures to generate a final prediction. We conduct a thorough ablation analysis on this pipeline and assess the superiority of the methodology in segmenting the brain stem, 4^{th} ventricle and cerebellum. The proposed methodology achieved a macro-averaged Dice index of 0.855 with respect to manual segmentation, with only 32 labeled volumes used during training. Additionally, average distances between automatically and manually segmented surfaces remained around $1mm$ for the three structures, while volumetry results revealed high agreement between manually labeled and predicted regions.**

I. Introduction

The brain develops rapidly during the third trimester of pregnancy and can be analyzed by magnetic resonance imaging (MRI) in the prenatal and postnatal periods [1]–[3]. It continues to evolve, especially in the first two years when most of the myelination process happens. Different myelination stages in pediatric brain Magnetic Resonance Imaging (MRI) result in distinct water, lipid, and cholesterol content in white matter, yielding distinct visual patterns. Pediatric brain segmentation is hence challenging not only due to its smaller size – which may result in partial volume effect – but also because of its particular contrast between white and gray matter [4]–[6]. The lack of myelination in children younger than two years old has also major implications because of low image contrast between normal and abnormal appearing white matter. The difficulties related to the myelination process in the pediatric brain impact both computer-aided segmentation and manual methods, leading to potentially significant disagreement among experts [7]. Thus, despite the huge amount of work on segmentation [8], the methods proposed in the adult context cannot be directly used for segmenting pediatric brain images. This also applies to white matter hyperintensities, where the methods developed for adult brain images (e.g. [9]–[11]) are not appropriate, since this T2/FLAIR hyperintensity may be physiological in children under two years of age. This highlights the need for a robust knowledge of normal versus pathological patterns in terms of volume, morphology, and signal intensities.

Most existing semi-automatic or automatic methods for classical pediatric brain segmentation are based on atlases [12], [13] and/or shallow learning [6], [14], [15]. However, these methods are often unable to cope with inter-individual variability and are generally not adapted to pathological cases. Additionally, most of these methods focus on high-quality images, acquired with 3T or stronger field MRI scanners. Yet, until recently in clinical routine most MRI scanners operate at field strengths around 1.5T and result in images of inferior quality. Tasks related to segmentation concern volumetry and regional analysis of specific structures, such as the corpus callosum which can be involved in various pathologies [6], [16], and of neuro-pathologies such as tumors. Thus, developing new methods for neuro-oncology would be of particular interest to answer important clinical needs [9], [17], such as treatment response [9] and the effects of low-intensity radiation [17] treatments.

Convolutional Neural Networks [18] (CNNs) have become ubiquitous in traditional visual learning tasks such as object classification, segmentation [19], [20], and detection [21]. Related fields also benefited from the large representation

This work was partly done while I. Bloch was with LTCI, Télécom Paris, Institut Polytechnique de Paris, France.

capabilities of CNNs, including a plethora of medical image tasks [22]. Radiology is arguably the largest beneficiary of deep learning in medical imaging, mainly because it is a highly data-driven field [23], and due to the inherent difficulties in reading radiological exams. This is especially true in volumetric data (e.g. MRI, Computer Tomography – CT, and Positron Emission Tomography – PET), as monitors only allow for 2D slices or projections of these 3D data to be viewed at any one time, making the visual assessment a laborious task.

In an effort to alleviate the burden of physicians, the main contribution of this work is the pipeline composed of a generalist and a pair of specialist networks for fine-grained automatic segmentation and volumetry of posterior fossa structures. Secondary contributions include: 1) a standard methodology for acquiring and annotating posterior fossa structures in pediatric brain MRIs; 2) a segmentation benchmark for posterior fossa structure segmentation; 3) a first effort in standardizing the volumetry measurements of the 4^{th} ventricle, brain stem and cerebellum for patients with ages ranging from 0 to 18 years; and 4) experiments reported in this manuscript have been performed in a real dataset of 32 pediatric images obtained in a large public hospital.

II. Related Work

Considering the posterior fossa as a region of interest is a common choice due to its clinical interest. Some of the research in this region has focused on biometric analysis, like [24] which uses a semi-automatic segmentation based on a region growing technique to distinguish posterior fossa, vermis, and brainstem, and then to measure its structures due to the growth disorders that affect this region. Other works focus on the analysis of this region of interest to find pediatric brain tumors [25], [26] which develop more often in that region, reaching about 55% to 70% of cases [25].

Over the last decade, deep feature learning has become the state-of-the-art for most computer vision applications. Since the resurgence of CNNs [18] from their first proposal [27], this family of network architectures has become ubiquitous in traditional visual learning tasks, such as object and scene classification. Around the middle of the 2010's, variations of traditional CNNs were adapted to a broader set of problems, such as object detection [21], semantic segmentation [19], and video understanding [28]. Related fields also benefited from the large representation capabilities of CNNs, including medical image analysis [22]. Radiology is arguably the largest beneficiary of deep learning in medical imaging, as mentioned in the introduction. Fully Convolutional Networks (FCNs) and variations of Encoder-Decoder architectures were quickly adapted to perform volumetric image segmentation [20], [29]–[32], effectively aiding physicians in diagnosis and surgery planning.

III. Dataset Description

There is a lack of publicly available neonatal and pediatric brain MRI datasets, mainly due to patient privacy and ethical issues. Also, existing public datasets are not suitable for our proposed task, since data are annotated for other purposes. For instance, BraTS [33] contains data for tumor segmentation in adults MRIs, and iSeg [34] consists of data for brain tissue segmentation of neonates with ages ranging from 2 weeks to 12 months. Therefore, we found it necessary to build our own labeled dataset.

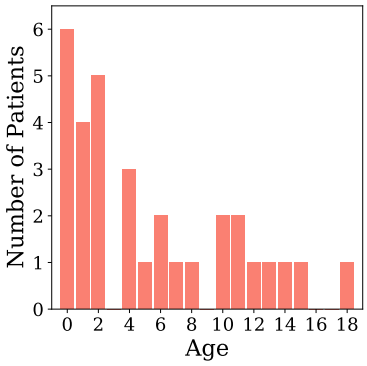

Fig. 1. Histogram of patient ages in the dataset.

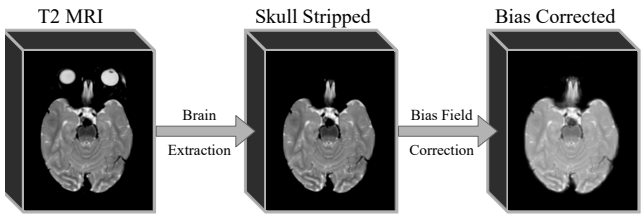

Fig. 2. Preprocessing steps for our pediatric MRI dataset.

The dataset used in this project consists of 32 T2-weighted MRI volumes acquired using a 1.5T Philips Ingenia scanner, 11 of them being obtained in the axial plane and the other 21 in the sagittal plane. We standardized the dataset to the axial plane view by transposing the 21 sagittal samples. Voxel sizes varied from $0.5 \times 0.5 \times 0.9$ to $0.9 \times 0.9 \times 2\ mm^3$. Patient's ages in the dataset range from a few months (0 years) to 18 years. As Figure 1 shows, there is a predominance of smaller children (0 to 4 years). Such a diversity of ages contrasts with other datasets, such as the previously mentioned iSeg [34].

Figure 2 shows the preprocessing pipeline applied to the dataset. First, the Brain Extraction Tool (BET) [35] is used to skull-strip the volume, in order to remove the non-brain voxels of the image. Then, Bias-Field Correction (BFC) [36] is applied to compensate for inhomogeneities in the magnetic field. The result is the input for the subsequent segmentation methods.

MRI volumes were annotated by radiology residents using medical image processing tools, such as Insight Toolkit (ITK)[1] and 3D Slicer[2], to provide reference segmentations of the structures of interest: 4^{th} ventricle, brain stem, and cerebellum. The residents were supervised by two radiology experts, who reviewed and corrected the residents' annotations.

[1]https://itk.org/

[2]https://www.slicer.org/

IV. Proposed Pipeline for Posterior Fossa Segmentation

We assess the performance of 5 architectures for the fossa segmentation task. One of the most common architectures in medical image segmentation is the U-Net [20], an Encoder-Decoder architecture that uses a multitude of skip connections via concatenation linking symmetric shallow and deep layers to yield high resolution segmentations. Classical U-Nets were designed for 2D images, but this architecture has been adapted for 3D volumes all throughout the literature. Concatenations, however, add a large amount of trainable parameters to the network, which needs to learn trainable convolutional filters for both shallow and deep activations on the decoders. V-Nets [29] were proposed to mitigate this problem by replacing concatenation by addition as their identity mapping, leveraging the proven advantages of residual blocks [37]. HighRes3DNet (HR3N) [30] exploits the varying receptive fields of dilated convolutions coupled with residual connections to yield a highly versatile and more compact architecture without losing performance when compared with other networks in the literature. SkipDenseNet (SDN) [31] takes advantage of densely connected blocks [38] to extract multiscale features from the data, while, at the same time, mitigating vanishing gradients by providing shortcuts in the backpropagation. At last, Med3D [32] showed highly adaptable representation learning capable of performing volumetric image segmentation in a multitude of domains.

As different architectures use distinct data flow strategies (e.g. filter sizes, dilation, upsampling strategies, skip connections, etc.), the features learned by each of them tend to highlight varied characteristics in the data. By exploiting this variety among voxel-wise classifiers, one can build more robust segmentation models via late fusion. The most common late fusion technique is a majority voting among the predictions of an ensemble of classifiers, which is the strategy employed in our segmentation framework. More details regarding the late fusion procedure can found in Section IV-B.

Our MRI dataset contains samples with a large variation in the resolution along the x-, y- and z-axes, thus, in order to standardize the sizes of the inputs, we resized all volumes and segmentation masks during training to $128 \times 128 \times 64$ voxels, regardless of voxel size. This volume size was chosen because exploratory experiments showed that it offered an acceptable compromise between memory budget and volume size. We performed online data augmentation by randomly removing up to 10% at the beginning and end of each axis before resizing, slightly rotating the images and masks around the craniocaudal axis by angles $\theta \sim N(0, 1)$, as well as randomly flipping the image/mask pairs across the sagittal plane, leveraging the quasi-symmetric nature of the brain. We tested a variety of patch-wise approaches, however, as the borders of the posterior fossa structures are often delineated in relation to surrounding tissues, patches hampered the convergence of the networks instead of helping. It was observed that the volume size $128 \times 128 \times 64$ after simply resizing the volumes was too low to achieve smooth segmentation boundaries. To overcome this limitation we devised a scheme composed of specialist networks able to perform fine-grained structure segmentation on smaller Regions of Interest (ROIs), as further explained in Section IV-A.

A. Specialist Networks

Early experiments showed that patching smaller volumes of the MRIs severely hampered the performance of the models, as the segmentation tasks at hand proved to be highly dependent on the spatial context of the voxels. In other words, the properties of the non-immediate surrounding tissue are very important to the predicted segmentation class. In order to avoid patching and leverage all annotated pixels in our small labeled set, we propose a 2-step architecture: 1) rough prediction by a generalist network ($\mathcal{G}$); 2) fine-grained predictions from a couple of specialist networks ($\mathcal{S}_{(1)}$ and $\mathcal{S}_{(2)}$). $\mathcal{S}_{(1)}$ specializes in the region in and around the brain stem and 4^{th} ventricle, learning to segment only a small ROI predicted by $\mathcal{G}$ and resized to $64 \times 64 \times 128$ voxels. $\mathcal{S}_{(2)}$ learns to perform fine-grained segmentation on the cerebellum, using the bounding box predicted by $\mathcal{G}$ and resized to a volume of size $64 \times 128 \times 64$. Specialist patch sizes were chosen empirically and based on the fact that the brain stem and ventricle extend more on the craniocaudal axis (i.e. z-axis) than on the posterior-anterior and latero-lateral axes, while the cerebellum main axis stretches across the latero-lateral axis (i.e. y-axis), not extending as far in the craniocaudal and posterior-anterior axes. ROIs fed to both specialists are padded with 5 voxels on each side before resizing in order to include the full structures even in the case where $\mathcal{G}$ misses large regions in the boundaries of the desired structures. A depiction of the generalist/specialist pipeline can be seen in Figure 3.

In addition to the preservation of the spatial context provided by our strategy, resizing coupled with specialist networks is also simpler and less expensive to implement to be applied on samples with varied resolutions and voxel sizes as the ones in this study. It is also cheaper to merge predictions for ROI segmentations from specialist networks into a single global prediction than joining a large set of smaller overlapping patches, a process that often requires a post-processing heuristic (e.g. mode filtering, image morphology, or even additional Bayesian models) to be applied in the predicted volume. Our fusion function $\Phi(\mathcal{S}_{(1)}, \mathcal{S}_{(2)})$ simply overwrites the predicted cerebellum in $\hat{y}_{\mathcal{G}}$ with $\hat{y}_{(2)}$, and the ventricle and brain stem by $\hat{y}_{(1)}$. At last, yet another advantage of specialist training is the inherent mitigation of label imbalance when using smaller volumes surrounding the ROI in the specialists because the structures occupy larger relative sizes in comparison with the volume.

Our pipeline is closely related to Med3D [39] in the use of cropped ROIs, with three crucial distinctions. First, Med3D relies heavily on transfer learning from related tasks, while the proposed methodology did not see advantages in pretraining from adult MRIs from BraTS2018 [33]. This could

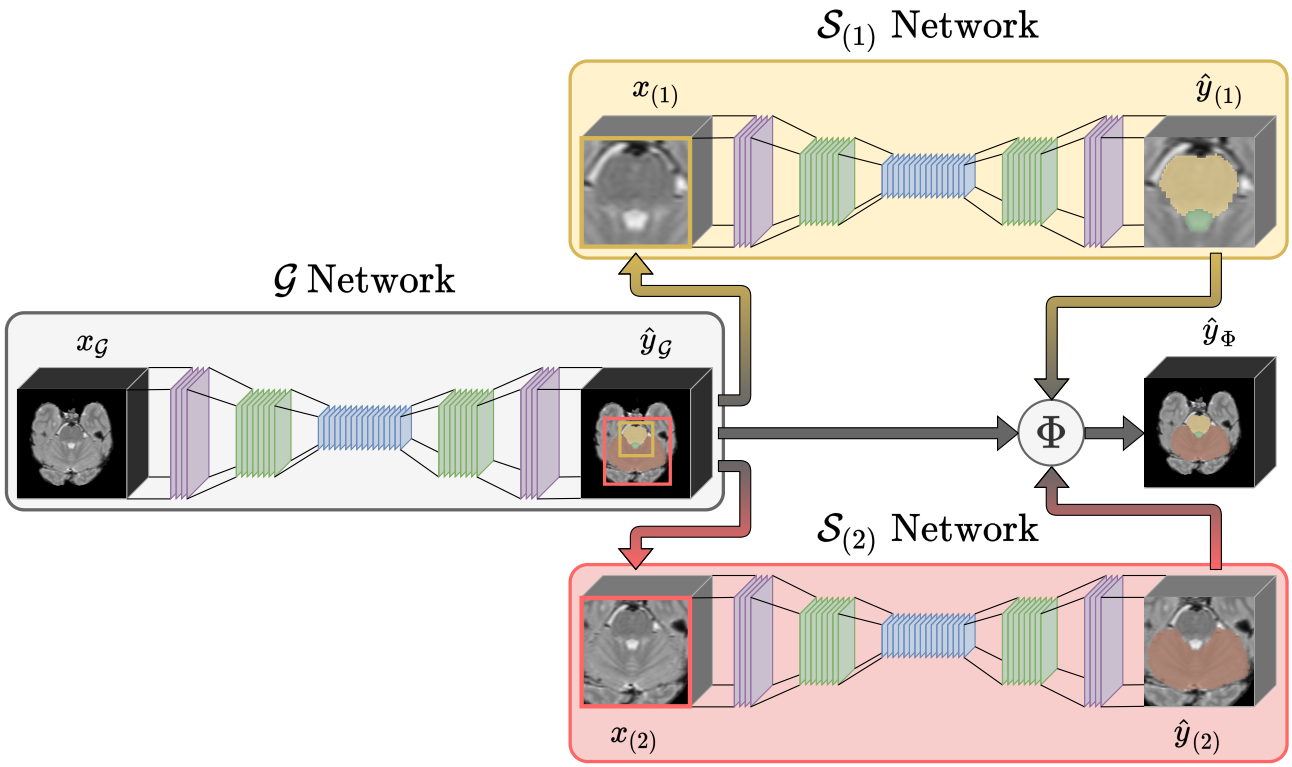

Fig. 3. Proposed segmentation pipeline for posterior fossa structures. The whole volume $X_{\mathcal{G}}$ is first fed to the generalist ($\mathcal{G}$), yielding a coarse prediction $\hat{y}_{\mathcal{G}}$. From $\hat{y}_{\mathcal{G}}$, two volumes are extracted: $X_{(1)}$, which feeds $\mathcal{S}_{(1)}$ to learn predictions $\hat{y}_{(1)}$ for the 4^{th} ventricle and brain stem; and $X_{(2)}$, fed to $\mathcal{S}_{(2)}$ to yield cerebellum segmentation $\hat{y}_{(2)}$. Predictions are combined via a function Φ, resulting in $\hat{y}_{\Phi}$.

be attributed to the large differences in visual patterns due to distinct myelination development stages. The second main difference between our method and Med3D is that we do not repurpose the generalist network for the specialized tasks. While this has the advantage of leveraging the whole MRI data, even outside the ROIs for the specialists, it also forces the same network to learn multiscale information; potentially hampering its performance. At last, we do not fix the architecture of the network, instead leveraging the advantages of ensembles by performing late fusion on 5 off-the-shelf architectures, as presented at the beginning of Section IV.

In an effort to improve reproducibility and foster future applications in pediatric brain MRI for automatic assessment of child development, our implementation of semantic segmentation with specialist networks can be found on the project webpage[5]. We also publicize pretrained weights for generalist and specialist networks, so other researchers can leverage the advantages of transfer learning to related tasks and/or private datasets.

B. Late Fusion

The late fusion strategy adopted in our pipeline can be divided into two distinct stages: 1) fusion of intra-architecture generalist and specialist predictions, and 2) cross-architecture voting. While the first stage aims to improve the delineation of organ boundaries, the second stage raises the objective segmentation metrics by considering an ensemble of models instead of a single network. Algorithm 1 details both stages of the late fusion procedure for both training and test data.

Stage 1: The first stage receives training data and labels $(\mathbf{x}^{tr}, \mathbf{y}^{tr})$ and test samples $(\mathbf{x}^{ts})$. The ensemble of generalist models $\mathbf{f}_{\mathcal{G}}$ is trained on $\{\mathbf{x}^{tr}, \mathbf{y}^{tr}\}$ to yield a coarse segmentation prediction $\hat{\mathbf{y}}^{tr}_{\mathcal{G}}$, which can be used to crop the full volumes into specialist Regions of Interest ($\mathbf{x}^{tr}_{(1)}$ and $\mathbf{x}^{tr}_{(2)}$) together with the cropped masks ($\mathbf{y}^{tr}_{(1)}$ and $\mathbf{y}^{tr}_{(2)}$). $\{\mathbf{x}^{tr}_{(1)}, \mathbf{y}^{tr}_{(1)}\}$ is then used to train $\mathbf{f}_{\mathcal{S}_{(1)}}$, while $\{\mathbf{x}^{tr}_{(2)}, \mathbf{y}^{tr}_{(2)}\}$ feeds the training of $\mathbf{f}_{\mathcal{S}_{(2)}}$. A fusion function Φ is then used to merge predictions $\hat{\mathbf{y}}^{tr}_{\mathcal{S}_{(1)}}$ and $\hat{\mathbf{y}}^{tr}_{\mathcal{S}_{(2)}}$ to generate predictions $\hat{\mathbf{y}}^{tr}_{\Phi}$. In our implementation Φ is simply a function that first copies $\hat{\mathbf{y}}^{tr}_{\mathcal{S}_{(2)}}$ and then overlaps it with $\hat{\mathbf{y}}^{tr}_{\mathcal{S}_{(1)}}$. This order was defined because it was observed that $\hat{\mathbf{y}}^{tr}_{\mathcal{S}_{(1)}}$ better delineates the boundary between the cerebellum and other structures.

Stage 2: With ensembles $\mathbf{f}_{\mathcal{S}_{(1)}}$ and $\mathbf{f}_{\mathcal{S}_{(2)}}$ properly fit, one can then generate predictions $\hat{\mathbf{y}}^{ts}_{(1)}$ and $\hat{\mathbf{y}}^{ts}_{(2)}$ for cropped samples $\mathbf{x}^{ts}_{(1)}$ and $\mathbf{x}^{ts}_{(2)}$. As in **Stage 1**, $\hat{\mathbf{y}}^{ts}_{(1)}$ and $\hat{\mathbf{y}}^{ts}_{(2)}$ are joined into $\hat{\mathbf{y}}^{ts}_{\Phi}$. Pixel-wise predictions from individual network architectures ($\hat{\mathbf{y}}^{ts}_{\Phi}$) are then merged via majority voting to yield the final prediction $\hat{\mathbf{y}}^{ts}_{maj}$.

[5]https://github.com/hugo-oliveira/STAP-3DSegmentation

Algorithm 1 Late fusion algorithm for specialist and generalist networks for an ensemble of segmentation architectures on one single fold division between training ($\{\mathbf{x}^{tr}, \mathbf{y}^{tr}\}$) and test ($\{\mathbf{x}^{ts}\}$) sets.

Require: $\mathbf{x}^{tr}, \mathbf{y}^{tr}$: training volumes and labels
Require: $\mathbf{x}^{ts}, \mathbf{y}^{ts}$: test volumes and labels
Require: $\mathbf{f}_{\mathcal{G}}$, $\mathbf{f}_{\mathcal{S}_{(2)}}$, $\mathbf{f}_{\mathcal{S}_{(2)}}$: generalist and specialist ensembles
  Randomly initialize $\mathbf{f}_{\mathcal{G}}$, $\mathbf{f}_{\mathcal{S}_{(2)}}$ and $\mathbf{f}_{\mathcal{S}_{(2)}}$
  // Training Procedure.
  for all $f^i_{\mathcal{G}}$, $f^i_{\mathcal{S}_{(1)}}$, $f^i_{\mathcal{S}_{(2)}}$ in $\mathbf{f}_{\mathcal{G}}$, $\mathbf{f}_{\mathcal{S}_{(1)}}$, $\mathbf{f}_{\mathcal{S}_{(2)}}$ **do**
    Train $f^i_{\mathcal{G}}$ on $\{\mathbf{x}^{tr}, \mathbf{y}^{tr}\}$
    Compute prediction $\hat{\mathbf{y}}^{tr}_{\mathcal{G}} \leftarrow f^i_{\mathcal{G}}(\mathbf{x}^{\mathbf{tr}}_{\mathcal{G}})$
    Obtain $\mathbf{x}^{tr}_{(1)}$ and $\mathbf{x}^{tr}_{(2)}$ by cropping $\mathbf{x}^{tr}$ according to $\hat{\mathbf{y}}^{tr}_{\mathcal{G}}$
    Obtain $\mathbf{y}^{tr}_{(1)}$ and $\mathbf{y}^{tr}_{(2)}$ by cropping $\hat{\mathbf{y}}^{tr}_{\mathcal{G}}$
    Train $f^i_{\mathcal{S}_{(1)}}$ on $\{\mathbf{x}^{tr}_{(1)}, \mathbf{y}^{tr}_{(1)}\}$
    Train $f^i_{\mathcal{S}_{(2)}}$ on $\{\mathbf{x}^{tr}_{(2)}, \mathbf{y}^{tr}_{(2)}\}$
  end for
  $\hat{\mathbf{Y}} \leftarrow \{\}$ *// Prediction list.*
  // Evaluation Procedure.
  for all $f^i_{\mathcal{G}}$, $f^i_{\mathcal{S}_{(1)}}$, $f^i_{\mathcal{S}_{(2)}}$ in $\mathbf{f}_{\mathcal{G}}$, $\mathbf{f}_{\mathcal{S}_{(1)}}$, $\mathbf{f}_{\mathcal{S}_{(2)}}$ **do**
    Compute prediction $\hat{\mathbf{y}}^{ts}_{\mathcal{G}} \leftarrow f^i_{\mathcal{G}}(\mathbf{x}^{ts})$
    Obtain $\mathbf{x}^{ts}_{(1)}$ and $\mathbf{x}^{ts}_{(2)}$ by cropping $\mathbf{x}^{ts}$ according to $\hat{\mathbf{y}}^{ts}_{\mathcal{G}}$
    Compute $\hat{\mathbf{y}}^{ts}_{\mathcal{S}_{(1)}} \leftarrow f^i_{\mathcal{S}_{(1)}}(\mathbf{x}^{ts}_{(1)})$ and $\hat{\mathbf{y}}^{ts}_{\mathcal{S}_{(2)}} \leftarrow f^i_{\mathcal{S}_{(2)}}(\mathbf{x}^{ts}_{(2)})$
    $\hat{\mathbf{y}}^{ts}_{\Phi} \leftarrow \Phi(\hat{\mathbf{y}}^{ts}_{(1)}, \hat{\mathbf{y}}^{ts}_{(2)})$ *// Stage 1 fusion.*
    $\hat{\mathbf{Y}} \leftarrow \hat{\mathbf{Y}} + \hat{\mathbf{y}}^{ts}_{\Phi}$
  end for
  // Majority voting across architectures (Stage 2 fusion).
  $\hat{\mathbf{y}}^{ts}_{maj} \leftarrow majority\_voting(\hat{\mathbf{Y}})$
  return fused prediction $\hat{\mathbf{y}}^{ts}_{maj}$

V. Experimental Procedure and Metrics

Individual architectures were modified (e.g. number of filters in each layer, number of layers, etc.) in order to fit a mini-batch containing 2 samples into one single GPU with 8GB of memory or in a couple of GPUs with a mini-batch size of 4. This setup was implemented in order to prevent individual architectures to perform better than others simply because of parameter capacity. Hyperparameters were also standardized across architectures, with training lasting 400 epochs using the Adam optimizer [40], initial learning rate of 1×10^{-2} halved each 80 epochs, L2 weight decay of 5×10^{-5} and a momentum of 0.5. The total loss function $\mathcal{L}_T$ used for training of generalist and specialist networks is a composite of the Cross Entropy ($\mathcal{L}_{CE}$) and Dice ($\mathcal{L}_{Dice}$) losses:

$$\mathcal{L}_T = \mathcal{L}_{CE} + \mathcal{L}_{Dice}. \quad (1)$$

All neural networks were implemented using Pytorch[3] and post-processing used the scikit-image[4] library.

In order to quantify the segmentation errors, we employ three evaluation measures: 1) voxel-wise error in the form of the macro-averaged Dice score (DSC) – also known as the F1-score; 2) distances between surfaces, as the average distance (μSD) and the 95^{th} percentile of the robust Hausdorff distance ($HD95$); and 3) R^2 correlation between predicted and reference volumetry. While the voxel-wise measures report overall scores in the performance of the algorithm, the surface distances can estimate the upper bounds of the distances between the manually segmented objects and network predictions, thus complementing each other. Dice is the less expensive measure to compute and yields an objective per-voxel performance evaluation, thus we drive the baseline architecture comparison via the DSC and report surface distances and structure volumes only for the results of majority voting.

As our dataset only contains a small set of labeled data, our experimental setup aimed to use the largest possible amount of samples in training, while also leveraging the whole labeled set to conduct evaluation. Therefore, we employed a 5-fold cross-validation scheme on the 32 samples of our dataset. We report average DSC computed for the validation folds as a whole, but surface distances and structure volumes are computed separately for each sample, as these are inherently instance-level annotations.

VI. Results and Discussion

Tables I and II present the results from experiments according to the voxel-wise and surface distance measures, respectively. Table I depicts the improvements brought by using the specialist networks and majority voting in comparison with generalist single architectures. The best DSC results from single generalist architectures reached 0.810 (HR3N), while majority voting across $\mathcal{G}$ architectures achieved an DSC of 0.823. Employing specialist networks on a single architecture showed considerable gains in performance – up to 0.07 of DSC in the case of Med3D – however combining majority voting with specialist networks yielded both the best overall score and the lowest standard deviation (0.855 ± 0.012), indicating consistent gains in segmentation performance.

Surface distances shown in Table II indicate a consistent improvement in the use of the specialist networks, as their focus on precomputed ROIs from $\mathcal{G}$ allows them to precisely delineate fine-grained segmentation masks of posterior fossa structures. A more precise border delineation between classes allows for better μSD and $HD95$, as they measure the average and 95^{th} percentile of the distance of $\hat{Y}$ and Y. Generalist networks are able to achieve average distances of $1.5mm$, $1.2mm$, and $1.0mm$ for 4^{th} ventricle, brain stem, and cerebellum, respectively. $\Phi(\mathcal{S}_{(1)}, \mathcal{S}_{(2)})$ shrinks these values to $1.1mm$, $0.8mm$, and $0.8mm$, respectively. For comparison, the lower voxel size in all our 1.5T MRIs had around $0.5mm$, while around half of our samples were acquired with a voxel size along their finer resolution axis around $0.92mm$ or $0.96mm$. In other words, the distance μSD for all organs is close to the scale of 1 or 2 pixels in the worst case.

Larger surface distances for the 4^{th} ventricle can be attributed to a more difficult standardization in the labeling of the lower point in this structure, as there is no clear physiological landmark unequivocally pointing to where the ventricle ends.

[3]https://pytorch.org/
[4]https://scikit-image.org/

TABLE I

DSC SCORES FOR THE PREDICTIONS OF OUR PIPELINE IN PEDIATRIC BRAIN MRIS. VALUES ARE PRESENTED IN THE FORMAT $\mu \pm \sigma$ ACROSS THE 5-FOLD DIVISION DESCRIBED IN SECTION V. THE SOLE BOLD VALUE INDICATES THE BEST OVERALL DSC.

Strategy	V-Net	U-Net	HR3N	SDN	Med3D	Majority
$\mathcal{G}$	.808 ± .035	.796 ± .025	.810 ± .014	.809 ± .022	.767 ± .022	.823 ± .017
$\Phi(\mathcal{G}, \mathcal{S}_1)$	.840 ± .021	.832 ± .029	.817 ± .039	.829 ± .023	.825 ± .024	.850 ± .018
$\Phi(\mathcal{S}_1, \mathcal{S}_2)$	.842 ± .014	.840 ± .019	.819 ± .037	.842 ± .018	.839 ± .018	**.855 ± .012**

This inherent inconsistency in the inter-specialist annotation procedure may prove to be a downside of using human observers instead of our automated methodology, as the models will tend to average the inputs of the radiologists. In this scenario, the model would effectively create its own standard for delineating the lower point of the 4^{th} ventricle and other structures that present lower contrast in comparison with surrounding tissue.

TABLE II

SURFACE DISTANCE IN mm FOR INDIVIDUAL POSTERIOR FOSSA STRUCTURES PREDICTED FROM THE MAJORITY VOTING OF GENERALIST AND SPECIALIST STRATEGIES.

Metric	Strategy	4^{th} ventricle	Stem	Cerebellum
μSD	$\mathcal{G}$	1.5 ± 1.7	1.2 ± 0.6	1.0 ± 0.5
	$\mathbf{\Phi}(\mathcal{G}, \mathcal{S}_{(\mathbf{1})})$	**1.0 ± 0.8**	**0.8 ± 0.5**	1.0 ± 0.5
	$\mathbf{\Phi}(\mathcal{S}_{(\mathbf{1})}, \mathcal{S}_{(\mathbf{2})})$	1.1 ± 0.8	**0.8 ± 0.5**	**0.8 ± 0.7**
MHD	$\mathcal{G}$	7.4 ± 5.3	5.3 ± 3.1	3.4 ± 1.8
	$\mathbf{\Phi}(\mathcal{G}, \mathcal{S}_{(\mathbf{1})})$	**5.6 ± 3.6**	**4.4 ± 3.1**	3.4 ± 1.8
	$\mathbf{\Phi}(\mathcal{S}_{(\mathbf{1})}, \mathcal{S}_{(\mathbf{2})})$	6.2 ± 3.3	4.5 ± 3.2	**3.2 ± 2.3**

The last set of objective measures in this work is composed of volumetry comparisons. Figure 4 shows these volumes for the three structures annotated in our dataset, both in correlation to the manually labeled images and by age. Correlations between $\mathcal{G}$ and reference segmentation masks are much smaller than the R^2 values from the specialist networks, again evidencing that our 2-step scheme indeed quantitatively improves predictions and may lead to automatic extraction of valuable clinical information in child development. The influence of age in the volume of the brain stem and cerebellum is clearly seen in the rightmost column of the figure, while, due to its smaller size and larger discrepancies in the annotation procedure, the growth of the 4^{th} ventricle is not as evident.

Figure 5 shows qualitative results from 3 samples in our dataset, their respective ROIs surrounding the brain stem and 4^{th} ventricle, and reference segmentations and predictions by distinct models. $\mathcal{G}$ and $\Phi(\mathcal{G}, \mathcal{S}_{(2)})$ produce coarse class borders, in contrast to $\Phi(\mathcal{S}_{(1)}\mathcal{S}_{(2)})$, which yields more accurate delineations between classes with fine-grained resolution, also taking into account the physiological properties of the structures. For instance, there is no gap between the 4^{th} ventricle and brain stem/cerebellum and $\Phi(\mathcal{G}, \mathcal{S}_{(1)})$ incorrectly predicts background voxels in this region due to the low resolution of $\hat{Y}_{\mathcal{G}}$.

VII. CONCLUSION

Our pipeline composed of a generalist network and a couple of specialist networks proved to be the best alternative both quantitatively and qualitatively to segment posterior fossa structures. Specialist networks were able to achieve better voxel-wise, surface distance, and structure volume agreement scores than the baseline approaches. Hence, automatic volumetry of posterior fossa structures seems to be a viable approach for aiding physicians in determining healthy biometry standards in neurological child development.

Future works include performing the same evaluations on a larger unlabeled sample of pediatric MRIs and extracting standards for structure volumes and other measurements that can be computed according to the automatic segmentation masks. Additionally, leveraging the models developed in this work, we aim to grow our dataset by integrating pathological samples and comparing them with our healthy pediatric MRIs.

ACKNOWLEDGEMENT

The authors would like to acknowledge the funding provided by *Fundação de Amparo à Pesquisa do Estado de São Paulo* (FAPESP – grants 2015/22308-2, 2017/50236-1, 2018/07386-5, 2019/16112-9 and 2020/06744-5) and *Agence Nationale de la Recherche* (ANR, grant ANR-17-CE23-0021). We also would like to thank *Hospital das Clínicas* from the University of São Paulo for providing the pediatric brain MRI samples used in our experiments.

REFERENCES

[1] C. Parazzini, A. Righini, M. Rustico, D. Consonni, and F. Triulzi, "Prenatal magnetic resonance imaging: brain normal linear biometric values below 24 gestational weeks," *Neuroradiology*, vol. 50, no. 10, pp. 877–883, 2008.

[2] B. Tilea, C. Alberti, C. Adamsbaum, P. Armoogum, J. Oury, D. Cabrol, G. Sebag, G. Kalifa, and C. Garel, "Cerebral biometry in fetal magnetic resonance imaging: new reference data," *Ultrasound in Obstetrics and Gynecology*, vol. 33, no. 2, pp. 173–181, 2009.

[3] A. Viola, S. Confort-Gouny, J. Schneider, Y. Le Fur, P. Viout, F. Chapon, S. Pineau, P. Cozzone, and N. Girard, "Is brain maturation comparable in fetuses and premature neonates at term equivalent age?" *American Journal of Neuroradiology*, vol. 32, no. 8, pp. 1451–1458, 2011.

[4] L. Gui, R. Lisowski, T. Faundez, P. S. Hüppi, F. Lazeyras, and M. Kocher, "Morphology-driven automatic segmentation of MR images of the neonatal brain," *Medical Image Analysis*, vol. 16, no. 8, pp. 1565–1579, 2012.

[5] R. Tadeusiewicz, M. Ogiela, and P. Szczepaniak, "Notes on a linguistic description as the basis for automatic image understanding," *International Journal of Applied Mathematics and Computer Science*, vol. 19, no. 1, pp. 143–150, 2009.

[6] D. K. Thompson, T. E. Inder, N. Faggian, L. Johnston, S. K. Warfield, P. J. Anderson, L. W. Doyle, and G. F. Egan, "Characterization of the corpus callosum in very preterm and full-term infants utilizing MRI," *Neuroimage*, vol. 55, no. 2, pp. 479–490, 2011.

[7] B. Morel, G. Antoni, J. Teglas, I. Bloch, and C. Adamsbaum, "Neonatal brain MRI: how reliable is the radiologist's eye?" *Neuroradiology*, vol. 58, no. 2, pp. 189–193, 2016.

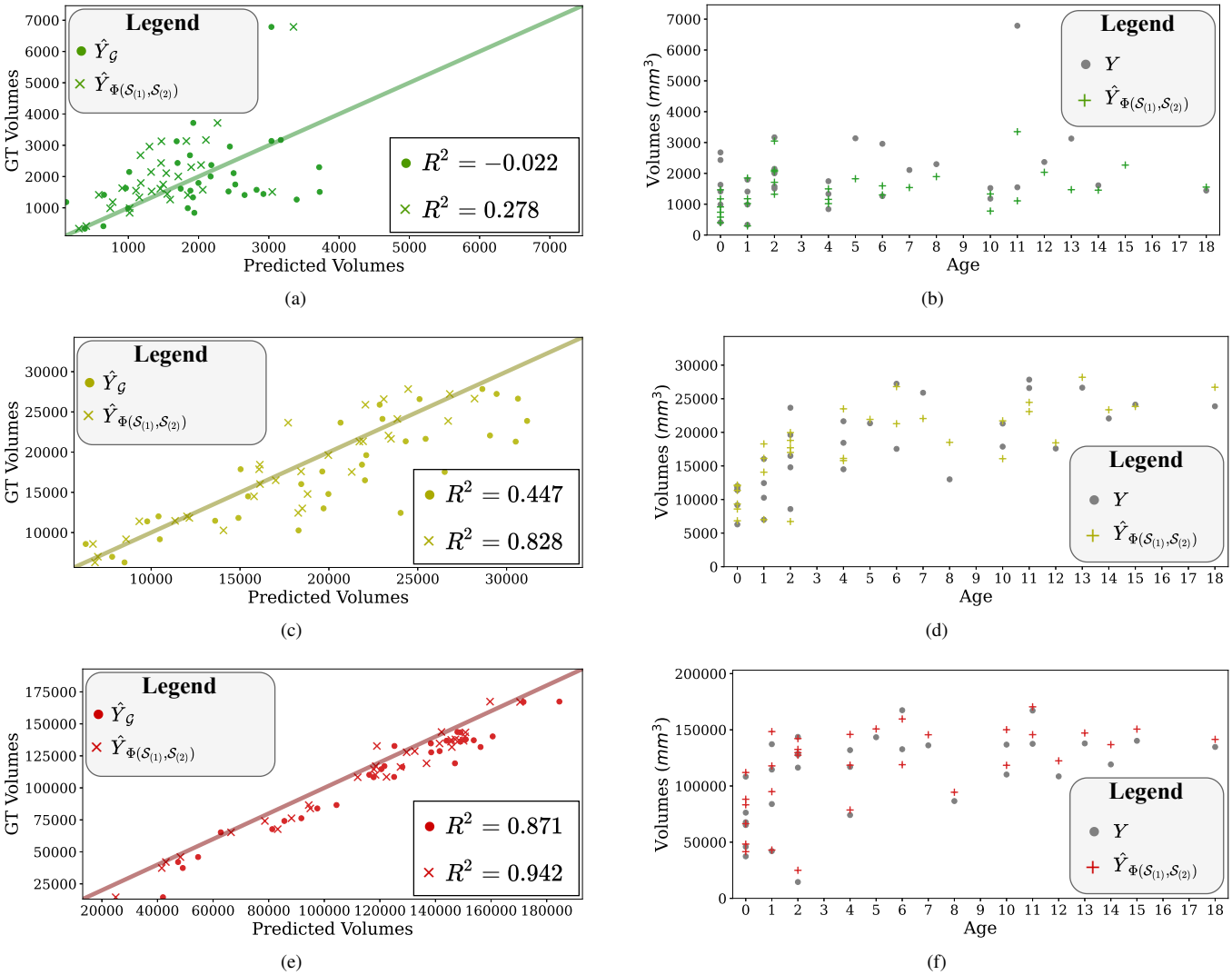

Fig. 4. 4^{th} ventricle (1^{st} row), brain stem (2^{nd} row) and cerebellum (3^{rd} row) volumes in mm^3. First and second columns indicate respectively: correlations between volumes of reference segmentations and predictions from $\mathcal{G}$ and $\Phi(\mathcal{S}_{(1)}, \mathcal{S}_{(2)})$; and volumes by age.

[8] I. Despotović, B. Goossens, and W. Philips, "MRI segmentation of the human brain: challenges, methods, and applications," *Computational and Mathematical Methods in Medicine*, vol. 2015, 2015.

[9] D. Rodriguez Gutierrez, M. Manita, T. Jaspan, R. A. Dineen, R. G. Grundy, and D. P. Auer, "Serial MR diffusion to predict treatment response in high-grade pediatric brain tumors: a comparison of regional and voxel-based diffusion change metrics," *Neuro-oncology*, vol. 15, no. 8, pp. 981–989, 2013.

[10] M. E. Caligiuri, P. Perrotta, A. Augimeri, F. Rocca, A. Quattrone, and A. Cherubini, "Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: a review," *Neuroinformatics*, vol. 13, no. 3, pp. 261–276, 2015.

[11] L. Griffanti, G. Zamboni, A. Khan, L. Li, G. Bonifacio, V. Sundaresan, U. G. Schulz, W. Kuker, M. Battaglini, P. M. Rothwell *et al.*, "Bianca (brain intensity abnormality classification algorithm): A new tool for automated segmentation of white matter hyperintensities," *Neuroimage*, vol. 141, pp. 191–205, 2016.

[12] I. S. Gousias, A. D. Edwards, M. A. Rutherford, S. J. Counsell, J. V. Hajnal, D. Rueckert, and A. Hammers, "Magnetic resonance imaging of the newborn brain: manual segmentation of labelled atlases in term-born and preterm infants," *Neuroimage*, vol. 62, no. 3, pp. 1499–1509, 2012.

[13] M. J. Cardoso, A. Melbourne, G. S. Kendall, M. Modat, N. J. Robertson, N. Marlow, and S. Ourselin, "AdaPT: an adaptive preterm segmentation algorithm for neonatal brain MRI," *NeuroImage*, vol. 65, pp. 97–108, 2013.

[14] P. Moeskops, M. J. Benders, S. M. Chiţă, K. J. Kersbergen, F. Groenendaal, L. S. de Vries, M. A. Viergever, and I. Išgum, "Automatic segmentation of MR brain images of preterm infants using supervised classification," *NeuroImage*, vol. 118, pp. 628–641, 2015.

[15] L. Wang, Y. Gao, F. Shi, G. Li, J. H. Gilmore, W. Lin, and D. Shen, "Links: Learning-based multi-source integration framework for segmentation of infant brain images," *NeuroImage*, vol. 108, pp. 160–172, 2015.

[16] M. Tardieu and K. Deiva, "Rare inflammatory diseases of the white matter and mimics of multiple sclerosis and related disorders," *Neuropediatrics*, vol. 44, no. 06, pp. 302–308, 2013.

[17] I. Moxon-Emre, E. Bouffet, M. D. Taylor, N. Laperriere, M. B. Sharpe, S. Laughlin, U. Bartels, N. Scantlebury, N. Law, D. Malkin *et al.*, "Vulnerability of white matter to insult during childhood: evidence from patients treated for medulloblastoma," *Journal of Neurosurgery: Pediatrics*, vol. 18, no. 1, pp. 29–40, 2016.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *NIPS*, vol. 25, pp. 1097–1105, 2012.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks

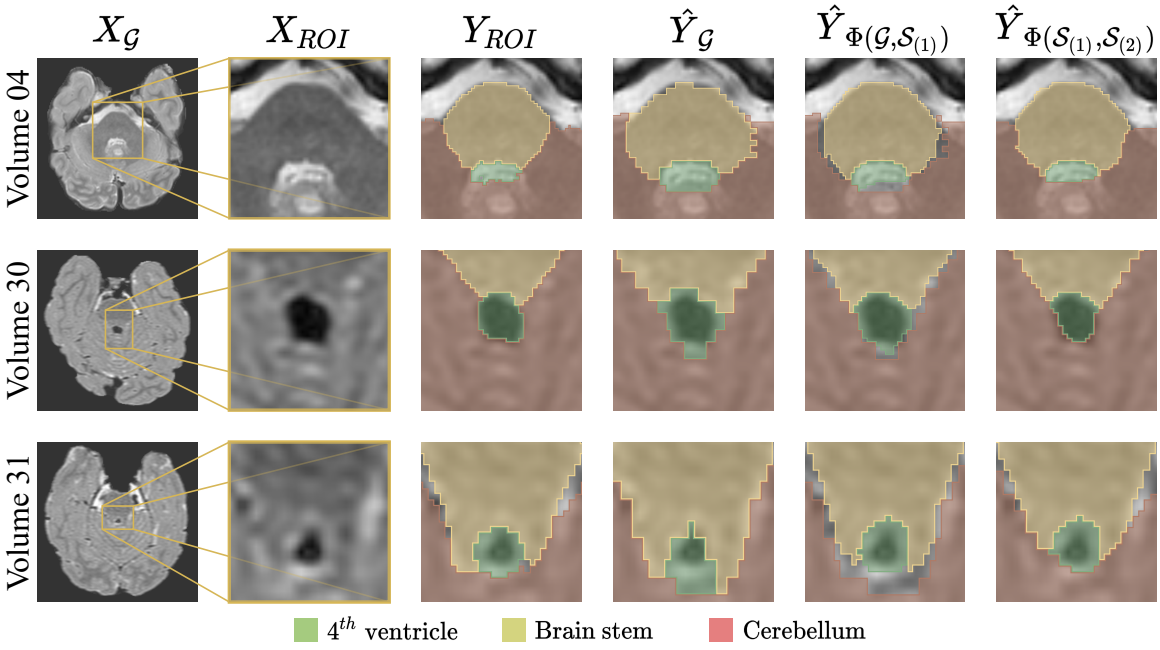

Fig. 5. Posterior fossa segmentations. Columns indicate in order: axial slices, $\mathcal{S}_{(1)}$ ROIs, reference segmentations and predictions from $\mathcal{G}$, $\Phi(\mathcal{G}, \mathcal{S}_{(1)})$ and $\Phi(\mathcal{S}_{(1)}, \mathcal{S}_{(2)})$.

for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.

[22] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[23] M. P. McBee, O. A. Awan, A. T. Colucci, C. W. Ghobadi, N. Kadom, A. P. Kansagra, S. Tridandapani, and W. F. Auffermann, "Deep learning in radiology," *Academic Radiology*, vol. 25, no. 11, pp. 1472–1480, 2018.

[24] I. Claude, J.-L. Daire, and G. Sebag, "Fetal brain MRI: segmentation and biometric analysis of the posterior fossa," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 4, pp. 617–626, 2004.

[25] S. Ahmed, K. M. Iftekharuddin, and A. Vossough, "Efficacy of texture, shape, and intensity feature fusion for posterior-fossa tumor segmentation in MRI," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 2, pp. 206–213, 2011.

[26] A. Islam, K. M. Iftekharuddin, R. J. Ogg, F. H. Laningham, and B. Sivakumar, "Multifractal modeling, segmentation, prediction, and statistical validation of posterior fossa tumors," in *Medical Imaging 2008: Computer-Aided Diagnosis*, vol. 6915. International Society for Optics and Photonics, 2008, p. 69153C.

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

[28] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014, pp. 1725–1732.

[29] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *International Conference on 3D Vision*. IEEE, 2016, pp. 565–571.

[30] W. Li, G. Wang, L. Fidon, S. Ourselin, M. J. Cardoso, and T. Vercauteren, "On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task," in *MICCAI*. Springer, 2017, pp. 348–360.

[31] T. D. Bui, J. Shin, and T. Moon, "3D densely convolutional networks for volumetric segmentation," *arXiv preprint arXiv:1709.03199*, 2017.

[32] S. Chen, K. Ma, and Y. Zheng, "Med3d: Transfer learning for 3d medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019.

[33] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[34] Y. Sun, K. Gao, Z. Wu, Z. Lei, Y. Wei, J. Ma, X. Yang, X. Feng, L. Zhao, T. L. Phan *et al.*, "Multi-site infant brain segmentation algorithms: The iSeg-2019 challenge," *arXiv preprint arXiv:2007.02096*, 2020.

[35] M. Jenkinson, M. Pechaud, S. Smith *et al.*, "BET2: MR-based estimation of brain, skull and scalp surfaces," in *Eleventh Annual Meeting of the Organization for Human Brain Mapping*, vol. 17. Toronto., 2005, p. 167.

[36] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4ITK: improved N3 bias correction," *IEEE Transaction on Medical Imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017, pp. 4700–4708.

[39] J. L. Cheong, D. K. Thompson, A. J. Spittle, C. R. Potter, J. M. Walsh, A. C. Burnett, K. J. Lee, J. Chen, R. Beare, L. G. Matthews *et al.*, "Brain volumes at term-equivalent age are associated with 2-year neurodevelopment in moderate and late preterm children," *The Journal of Pediatrics*, vol. 174, pp. 91–97, 2016.

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