Benchmark on Automatic Six-Month-Old Infant Brain Segmentation Algorithms: The iSeg-2017 Challenge

Li Wang, *Senior Member, IEEE*, Dong Nie, Guannan Li, Élodie Puybareau, Jose Dolz, Qian Zhang, Fan Wang, Jing Xia, Zhengwang Wu, Jia-Wei Chen, *Member, IEEE*, Kim-Han Thung, Toan Duc Bui, Jitae Shin, Guodong Zeng, Guoyan Zheng, *Member, IEEE*, Vladimir S. Fonov, Andrew Doyle, Yongchao Xu, Pim Moeskops, Josien P. W. Pluim, *Fellow, IEEE*, Christian Desrosiers, Ismail Ben Ayed, Gerard Sanroma, Oualid M. Benkarim, Adrià Casamitjana, Verónica Vilaplana, Weili Lin, Gang Li, *Senior Member, IEEE*, and Dinggang Shen, *Fellow, IEEE*

***Abstract*—Accurate segmentation of infant brain magnetic resonance (MR) images into white matter (WM), gray matter (GM), and cerebrospinal fluid is an indispensable foundation for early studying of brain growth patterns and morphological changes in neurodevelopmental disorders. Nevertheless, in the isointense phase (approximately 6–9 months of age), due to inherent myelination and maturation process, WM and GM exhibit similar levels of intensity in both T1-weighted and T2-weighted MR images, making tissue segmentation very challenging. Although many efforts were devoted to brain segmentation, only a few studies have focused on the segmentation of six-month infant brain images. With the idea of boosting methodological development in the community, iSeg-2017 challenge (http://iseg2017.web.unc.edu) provides a set of six-month infant subjects with manual labels for training and testing the participating methods. Among the 21 automatic segmentation methods participating in iSeg-2017, we review the eight top-ranked teams, in terms of Dice ratio, modified Hausdorff distance, and average surface distance, and introduce their pipelines, implementations, as well as source codes. We further discuss the limitations and possible future directions. We hope the dataset in iSeg-2017, and this paper could provide insights into methodological development for the community.**

***Index Terms*—Infant, brain, segmentation, isointense phase, challenge.**

I. Introduction

THE first year of life is the most dynamic phase of the postnatal human brain development, along with rapid tissue growth and development of a wide range of cognitive and motor functions [1], [2]. The increasing availability of non-invasive infant brain multimodal magnetic resonance images (MRI), e.g., T1-weighted (T1w) and T2-weighted (T2w) images, provides unprecedented opportunities for accurate and reliable charting of dynamic early brain developmental trajectories in understanding normative and aberrant growth. For example, the Baby Connectome Project[1] (BCP) [3] is acquiring and releasing both cross-sectional

[1] http://babyconnectomeproject.org

Manuscript received December 20, 2018; revised February 17, 2019; accepted February 17, 2019. Date of publication February 27, 2019; date of current version August 30, 2019. This work was supported in part by the National Institutes of Health under Grant MH109773, Grant MH117943, Grant MH100217, Grant MH070890, Grant EB006733, Grant EB008374, Grant EB009634, Grant AG041721, Grant AG042599, Grant MH088520, Grant MH108914, Grant MH116225, and Grant MH107815. *(Li Wang, Dong Nie, Guannan Li, Élodie Puybareau, Jose Dolz, Qian Zhang, Fan Wang, Jing Xia, Zhengwang Wu, Jia-Wei Chen, and Kim-Han Thung are co-first authors.) (Corresponding authors: Li Wang; Dinggang Shen.)*

L. Wang, D. Nie, G. Li, Q. Zhang, F. Wang, J. Xia, Z. Wu, J.-W. Chen, K.-H. Thung, W. Lin, and G. Li are with the Department of Radiology and the Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA (e-mail: li_wang@med.unc.edu).

É. Puybareau and Y. Xu are with the EPITA Research and Development Laboratory, 94270 Le Kremlin-Bicêtre, France.

J. Dolz, C. Desrosiers, and I. Ben Ayed are with the Laboratory for Imagery, Vision and Artificial Intelligence, École de Technologie Supérieure, Montreal, QC H3C 1K3, Canada.

T. D. Bui and J. Shin are with the Media System Laboratory, School of Electronic and Electrical Engineering, Sungkyunkwan University, Seoul 16419, South Korea.

G. Zeng and G. Zheng are with the Information Processing in Medical Intervention Laboratory, University of Bern, 3012 Bern, Switzerland.

V. S. Fonov is with the NeuroImaging and Surgical Technologies Laboratory, Montreal Neurological Institute, McGill University, Montreal, QC H3A 0G4, Canada.

A. Doyle is with the McGill Centre for Integrative Neuroscience, Montreal Neurological Institute, McGill University, Montreal, QC H3A 0G4, Canada.

P. Moeskops and J. P. W. Pluim are with the Medical Image Analysis Group, Department of Biomedical Engineering, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands.

G. Sanroma is with the Population Sciences Department, German Center of Neurodegenerative Diseases (DZNE), 53127 Bonn, Germany.

O. M. Benkarim is with the Simulation, Imaging and Modelling for Biomedical Systems, Universitat Pompeu Fabra, 08002 Barcelona, Spain.

A. Casamitjana and V. Vilaplana are with the Image and Video Processing Unit, Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain.

D. Shen is with the Department of Radiology and the Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA, and also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, South Korea (e-mail: dgshen@med.unc.edu).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2019.2901712

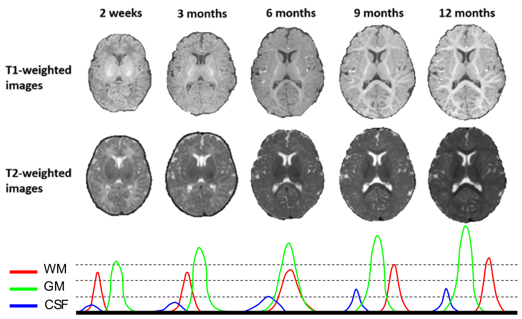

Fig. 1. The T1- and T2-weighted MR images of an infant, longitudinally scanned at 2 weeks, 3, 6, 9 and 12 months of age. At around 6 months of age (i.e., the isointense phase), the MR images show the lowest tissue contrast, implying the most challenging for tissue segmentation. The corresponding tissue intensity distributions from T1w MR images are shown at the bottom row, where the WM and GM intensities are highly overlapped in the isointense phase.

and longitudinal high-resolution multimodal MRI data from 500 typically-developing children from birth to 5 years of age. The Developing Human Connectome Project[2] (dHCP) in the UK is releasing MRI data from 1500 subjects acquired from 20 to 44 weeks post-conceptional age. These large-scale datasets will undoubtedly greatly increase our limited knowledge on normal early brain development, and provide important insights into the origins and abnormal developmental trajectories of neurodevelopmental disorders, such as autism [4], schizophrenia, bipolar disorder, and attention-deficit/hyperactivity disorder.

One fundamentally important step in studying the normal and abnormal early brain development is accurate segmentation of infant brain MR images into different regions of interest (ROIs) [5], [6], e.g., white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF), which is also very important for registration [7] and atlas building [8], [9]. There are three distinct phases in the first-year brain MRI, as shown in Fig. 1. During the infantile phase ($<=$5months), GM shows higher signal intensity than WM in T1w images. The isointense phase (6-9 months) corresponds to the myelination and maturation process of the brain, yielding an increase of the intensity of WM in T1w images and thus a low signal differentiation between GM and WM (which is also the case for T2w images). The last phase is the early adult-like phase ($>$9 months), where GM intensity is much lower than that of WM in T1w images, similar to the pattern of tissue contrast in the adult MR images. The corresponding tissue intensity distributions of three phases are shown in the third row of Fig. 1, from which we can observe the relative good contrast for the infantile and early adult-like phases. However, in the isointense phase, the intensity distributions of voxels in GM and WM are largely overlapping (especially in the cortical regions), thus leading to the lowest tissue contrast and creating the main challenge for tissue segmentation, in comparison to images at other phases of brain development. Also, the appearance of exact isointense contrast varies across different brain regions due to nonlinear brain development [10]. These patterns, along with various factors, such as motion artifacts or severe partial volume effect due to smaller brain size and ongoing white matter myelination, make automatic segmentation of isointense infant brain MRI a highly challenging task, thus causing that existing computational tools typically developed for processing and analyzing adult brain MRI [11], e.g., SPM, FSL, BrainSuite, CIVET, FreeSurfer and HCP pipeline, often perform poorly on infant brain MRI [12].

We have witnessed the spread and rise in popularity of Grand Challenges in the medical imaging community during the past years (e.g., NeoBrainS12,[3] [13], MRBrainS,[4] [14], ISLES,[5] [15], and BRATS [6] [16]). These challenges have allowed development of public benchmarks that serve as fair and up-to-date comparisons for the methods proposed by colleagues around the world. For example, the MICCAI challenge on neonatal MRI segmentation (NeoBrainS12[3]) and the MICCAI challenge on adult MRI segmentation (MRBrainS[4]) mainly focused on the infantile and adult-like phases, respectively, rather than the challenging isointense phase. To date, only a few studies focused on the segmentation of 6-month infant brain image [13], [17]–[19]. In iSeg-2017challenge (http://iseg2017.web.unc.edu), researchers were invited to participate with their automatic algorithms to segment WM, GM and CSF on isointense (6-month) infant brain MR scans, which remains scarce in the field. At the time of writing this paper, 21 teams had submitted their results on the iSeg-2017 website. In this paper, we focus only on those methods that were ranked among the 8 top-ranked teams in terms of Dice Coefficient (DICE), modified Hausdorff distance (HD95) and Average Surface Distance (ASD). In the next section, we introduce the cohort employed for this challenge. Then, in Section III, the metrics used to evaluate the performance of the proposed methods are detailed. Section IV provides a complete description of the top-ranked methods selected for this review. Section V discusses their performance, limitations and possible future directions.

II. Data

Selected MR scans for training and testing were randomly chosen from the pilot study of Baby Connectome Project (BCP, http://babyconnectomeproject.org). All infants were term born (40±1 weeks ofgestational age) without any pathology. At the time of scanning, the average age is 6.0±0.5 months. MR scans were acquired on a Siemens head-only 3T scanners with a circular polarized head coil. During the scan, infants were asleep, unsedated, fitted with ear protection, and their heads were secured in a vacuum-fixation device.

1) T1-weighted MR images were acquired with 144 sagittal slices using parameters: TR/TE = 1900/4.38 ms, flip angle = 7°, resolution = $1\times1\times1$ mm^3;
2) T2-weighted MR images were obtained with 64 axial slices: TR/TE = 7380/119 ms, flip angle = 150°, resolution $=1.25\times1.25\times1.95$ mm^3.

[2] http://www.developingconnectome.org

[3] http://neobrains12.isi.uu.nl

[4] http://mrbrains13.isi.uu.nl

[5] http://www.isles-challenge.org

[6] https://www.med.upenn.edu/sbia/brats2017/data.html

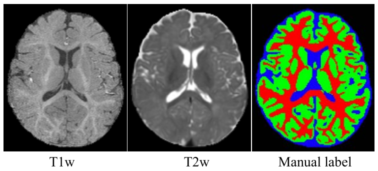

Fig. 2. T1w and T2w MR images of an infant subject scanned at 6 months of age (isointense phase), provided by iSeg-2017. From left to right: T1w MR image, T2w MR image, and manual label image.

For image preprocessing, T2w images were firstly resampled into an isotropic 1×1×1 mm^3 resolution and rigidly aligned onto their corresponding T1w images. Next, standard image preprocessing steps were performed before manual segmentation, including skull stripping [20], intensity inhomogeneity correction [21], and manual removal of the cerebellum and brain stem by experts.

To generate reliable manual segmentations, we first took advantage of the follow-up 24-month scans of the same subjects, with high tissue contrast, to generate an initial automatic segmentation for 6-month scans [22], by using a publicly available software iBEAT(www.nitrc.org/projects/ibeat/) [23]. This is based on the fact that, at term birth, the major sulci and gyri are already present in the neonates [24]. The pattern of the major sulci and gyri are generally preserved but are fine-tuned during postnatal brain development [25]. Specifically, the cortical convolutions emerge in the late gestation before birth [26], with extensive folding occurring during the third trimester [27], [28]. At term birth, although the brain is only one-third of the adult brain volume [29], the major sulci and gyri present in the adult are already established [24]. Second, based on the obtained initial automatic segmentation, manual editing was performed, under the guidance of an experienced neuroradiologist (Dr. Valerie Jewells, UNC-Chapel Hill), to correct segmentation errors (based on both T1w and T2w MR images) and geometric defects using ITK-SNAP, with the help of surface rendering. For example, if there is a hole/handle in the surface, we will first localize the related slices, and then check the segmentation maps of both T1w and T2w images to determine whether to fill the hole or cut the handle. Generally, it took almost one week for correcting one subject. Fig. 2 shows an example of a 6-month infant subject with T1w and T2w MR images, and manual labels of WM, GM and CSF, where WM includes both unmyelinated and myelinated white matter; GM includes cortical and subcortical gray matter; and CSF includes the ventricles and cerebrospinal fluid in the extracerebral space. Finally, 10 infant subjects (5 females/5 males) with manual labels were provided for training and 13 infant subjects (7 females/6 males) with manual labels were provided for testing. Note that the manual labels of testing subjects are not provided to the participants for fair comparison. All testing subjects were segmented off-site and uploaded for evaluation.

III. Evaluation

To evaluate the performance of different methods, we use Dice coefficient (DICE), 95th-percentile Hausdorff distance (HD95), and average surface distance (ASD), as metrics to evaluate the performance.

A. DICE

$$\mathrm{DICE} = \frac{2|A \cap B|}{|A| + |B|}$$

where A and B denote the binary segmentation labels generated manually and computationally, respectively, $|A|$ denotes the number of positive elements in the binary segmentation A, and $|A \cap B|$ is the number of shared positive elements by A and B.

B. HD95

$$\mathrm{HD}\left(C, D\right) = \max\left(h\left(C, D\right), h\left(D, C\right)\right)$$

where C and D are the two sets of vertices identified manually and computationally, respectively, for one tissue class of a subject. $h\left(C, D\right)$ is given by:

$$h\left(C, D\right) = \max_{c \in C} \max_{d \in D} \|c - d\|$$

The modified Hausdorff distance is defined as the 95th-percentile Hausdorff distance (HD95).

C. ASD

$$ASD = \frac{1}{2}\left(\frac{\sum_{V_i \in S_A} min_{V_j \in S_B} d\left(V_i, V_j\right)}{\sum_{V_i \in S_A} 1} + \frac{\sum_{V_j \in S_B} min_{V_i \in S_A} d\left(V_j, V_i\right)}{\sum_{V_j \in S_B} 1}\right)$$

where S_A is the surface of the ground-truth label map, S_B is the surface of the automatically segmented label map, and $d\left(V_j, V_i\right)$ indicates the Euclidean distance from vertex V_j to the vertex V_i.

IV. Methods and Implementations

First, we give an overview of all the participants of the iSeg-2017 Challenge, along with a very short description of each participating approach. A total of 21 teams successfully submitted their results to iSeg-2017 before the official deadline. Please refer to Appendix Table I,[7] in which we describe all the participating teams with affiliations and features used in their methods. In Appendix Table II, we summarize the performance of all these teams in terms of DICE, HD95 and ASD. An interesting finding is that 20 out of 21 teams employed convolutional neural networks for segmentation, while only 1 team utilized a classic atlas-based segmentation method. Among those 20 teams using convolutional neural networks, 8 teams adopted the U-Net architecture [30]. As explained earlier, we will review only the 8 top-ranked methods according to these metrics.

[7] Supplementary materials are available in the supplementary files/multimedia tab.

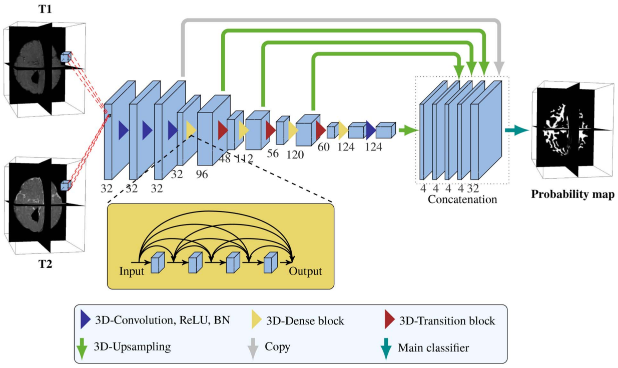

Fig. 3. 3D densely convolutional network architecture for infant brain segmentation [31].

A. MSL_SKKU: Media System Laboratory at Sungkyunkwan University (SKKU), Korea [31]

Bui *et al.* [31] extended the densely connected convolutional network [32] to deal with segmentation of 6-month infant brain MRI. By concatenating information from shallow to deep dense blocks, the proposed network allows capturing multiple contextual information and yields accurate segmentation results. Their proposed network architecture for infant brain segmentation is shown in Fig. 3.

The network consists of two paths: 1) the down-sampling path and 2) the up-sampling path. The down-sampling path includes four dense blocks. Each dense block comprises of four 3×3×3 convolutional kernels, each of which is preceded by a batch normalization (BN) layer [33] and a rectified linear unit (ReLU) nonlinearity [34]. A bottleneck layer is introduced before each 3×3×3 convolution to improve computational efficiency. They use a dropout layer [35] with the dropout rate of 0.2 after each 3×3×3 convolution layer to avoid overfitting. Between two contiguous dense blocks, a transition block that has 1×1×1 convolution with the compression rate of half and a convolution layer of stride 2 is used to reduce the feature map resolutions while preserving the spatial information. In the up-sampling path, the 3D-upsampling operators are used to recover the input resolution. In particular, the shallower layers provide fine output maps, while the deeper layers contain the coarse output maps [36]. To combine multiple levels of contextual information, up-sampling is performed after each dense block and then those up-sampled feature maps are concatenated. A classifier consisting of a 1×1×1 convolution is used to classify the concatenated feature maps into target classes. In total, this network has 47 layers with 1.55 million learnable parameters.

In the implementation, T1w and T2w images were normalized to zero mean and unit variance before inputting them into the network. Due to the limited GPU memory, sub-volume samples of size 64×64×64 were used as input of the network. The network was trained with Adam [37] with a mini-batch size of 4. The weights were initialized as in [38]. The learning rate was initially set to 0.0002 and was decreased by a factor of γ =0.1 every 50,000 iterations. Weight decay of 0.0005 and a *momentum* of 0.97 were set up for the network. The final segmentation results were obtained using the majority voting strategy from the predictions of the overlapped sub-volumes with stride of 8×8×8. It took about 2 days for training and 5 minutes for segmenting each subject on a TitanX Pascal GPU and Caffe framework [30], [39].

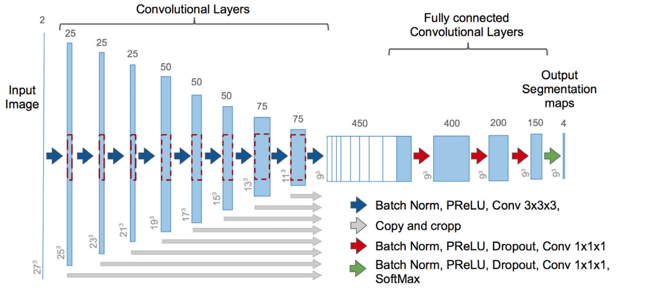

Fig. 4. Architecture of the proposed SemiDenseNet [40], which takes the input sub-patches of size 27 × 27 × 27 from T1w and T2w images and provides segmentation maps of size 9 × 9 × 9.

B. LIVIA: Laboratory for Image, Vision and Artificial Intelligence (LIVIA), at the École de Technologie Supérieure (ETS) in Montreal [40]

Inspired by the recent success of dense networks in image segmentation problems, Dolz *et al.* [40] proposed an ensemble of semi-dense deep architectures to segment 6-month infant brain MRI. In this novel architecture called SemiDenseNet, the outputs of all convolutional layers are connected directly to the last block of the network. This semi-dense connectivity brings some advantages: 1) efficient propagation of gradients during training, and 2) reducing the number of trainable parameters.

Their proposed method (Fig. 4) extends the recent deep architecture proposed in [41], which is composed of many convolutional layers, each containing several 3D convolution filters. To avoid losing resolution when down-sampling the data, the proposed architecture is a fully convolutional network (FCN) without any pooling operations. In addition, multi-scale context is modeled by embedding the outputs from all layers into a dense feature map that is provided to the first fully connected layer, which gives to the architecture the appearance of a semi-dense CNN. A notable difference of the proposed approach with respect to most existing works is the adopted sampling strategy. Instead of employing a whole 3D MR scan as the input, they sub-sample the whole image into smaller sub-volumes, which are then fed into the network. This allows: 1) avoiding memory issue if pooling is not employed and 2) avoiding data augmentation for training, since a high number of samples can be extracted from each image. Further, to achieve a more robust segmentation, an ensemble of several architectures is employed to combine their outputs via a majority voting strategy.

The proposed *SemiDenseNet* is composed of 13 layers in total: 9 convolutional layers in each path, 3 fully-connected layers, and a classification layer. The number of kernels (with the size of 3×3×3) in each convolutional layer, from shallow to deeper, is 25, 25, 25, 50, 50, 50, 75, 75 and 75, respectively. The fully-connected layers are composed of 400, 200 and 150 hidden units, respectively, followed by the final classification layer. To preserve spatial resolution, a unit

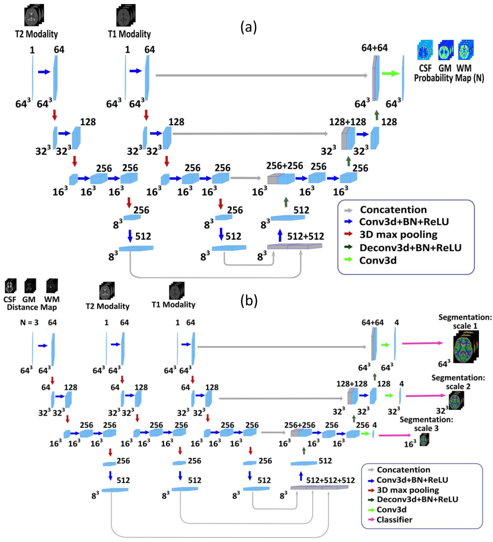

Fig. 5. A schematic illustration of the proposed two-stage method [42], consisting of (a) 3DFCN-1 at stage one and (b) 3DFCN-2 at stage two. For each block, the number above represents the number of feature stacks, and the number on the left side indicates the data size.

stride is used for all convolutional layers. Each convolutional block is composed by a batch normalization step followed by a Parametric Rectified Linear Unit (PReLU) and several convolutional filters in the convolutional layers. Further, in the fully convolutional connected layers, dropout is employed right after PReLU activations. The optimization of network parameters is performed via RMSprop optimizer. *Momentum* was set to 0.6 and the initial learning rate to 0.001, reduced by a factor of 2 after every 5 epochs (starting from epoch 10). Weights in layer l were initialized based on a zero-mean Gaussian distribution of standard deviation $2/nl$, where nl denotes the number of connections to units in that layer. The proposed 3D FCN was trained for 30 epochs, each composed of 20 subepochs. At each subepoch, a total of 1000 samples were randomly selected from the training images, and processed in batches of size 20. An ensemble composed by 10 identical CNNs was employed, each trained with a different combination of subjects. No data augmentation was employed to increase the size of the dataset. Experiments were performed in a computational server equipped with a NVIDIA Tesla P100 GPU with 16 GB of RAM memory. Training the proposed network took around 25 min per epoch, and around 13 hours to have a single CNN. Segmentation of a whole 3D MR scan was performed in 10 seconds per CNN model in average.

C. Bern_IPMI: Information Processing in Medical Intervention Lab., University of Bern, Switzerland [42]

Zeng and Zheng [42] proposed a two-stage, 3D fully convolutional networks (3DFCN)-based method for segmentation of 6-month infant brain MRI. In order to alleviate the potential gradient vanishing problem during training, they designed multi-scale deep supervision. Moreover, context information was used to further improve the performance.

Fig. 5 illustrates their proposed two-stage method. Both 3DFCN-1 and 3DFCN-2 adopt an encoder (contracting path)-decoder (expansive path) structure [43]. More specifically, 3DFCN-1 is used in the first stage to learn the probability map of each brain tissue from multimodal MR images (T1w and T2w images). An initial segmentation of different brain tissues is then obtained from the probability map, which further allows us to compute a distance map for each tissue [44]. The computed distance maps can be used to model the spatial context information. At the second stage, 3DFCN-2 is employed to get the final segmentation by using both the spatial context information and the multimodal MR images. To effectively integrate multimodal information, separate encoder paths are constructed for different modalities and then their outputs of the encoder paths are concatenated at the beginning of the expansive path such that the decoder can fuse complementary information from different sources. At both stages, long and short skip connections are employed to recover spatial context lose in the contracting encoder. See Fig. 5 for details. For 3DFCN-2, two down-scaled branch classifiers are further injected into the networks in addition to the classifier of the main network. By doing this, segmentation is performed at multiple output layers. As a result, classifiers in different scales can take advantage of multi-scale context.

Their proposed method was implemented in Python using TensorFlow framework and trained on a desktop with a 3.6 GHz Intel®i7 CPU and a GTX 1080 Ti graphics card with 11 GB GPU memory. In order to enlarge the training samples, data augmentation was utilized. Specifically, each training data was rotated for (90, 180, 270) degrees around the y-axis of the image and flipped horizontally (by taking the z axis as the vertical direction). The network was trained for 10,000 iterations. All weights were updated by the stochastic gradient descent algorithm (*momentum*=0.9, *weight_decay*=0.005). Learning rate was initialized as 1×10^{-3} and reduced by a factor of 2 every 3,000 times. After training, the proposed method took about 8 seconds in average to segment one subject.

D. TU/e IMAG/e: Medical Image Analysis Group (IMAG/e) of Eindhoven University of Technology (TU/e) [45]

A convolutional neuronal network was used for the segmentation of 6-month infant brain MRI into WM, GM and CSF [45]. Unlike previous work [46], the network does not include pooling layers, but uses dilated convolutions to achieve a large receptive field using a limited number of trainable weights.

The method combines 2D triplanar and 3D input using four network branches (Fig. 6). All network branches use the T1w and T2w images as 2-channel input. The triplanar input is included in three branches with dilated 2D convolutions. Each of these branches consists of 7 layers of 3×3 convolutions with increasing dilation factors, resulting in a receptive field of 67×67 [47], as previously also used for cardiac segmentation [48] and adult brain MRI segmentation [45]. The 3D input is included in the fourth branch that consists of 12 layers of 3×3×3 convolutions, resulting in a receptive field of 25×25×25. The output features from the four branches

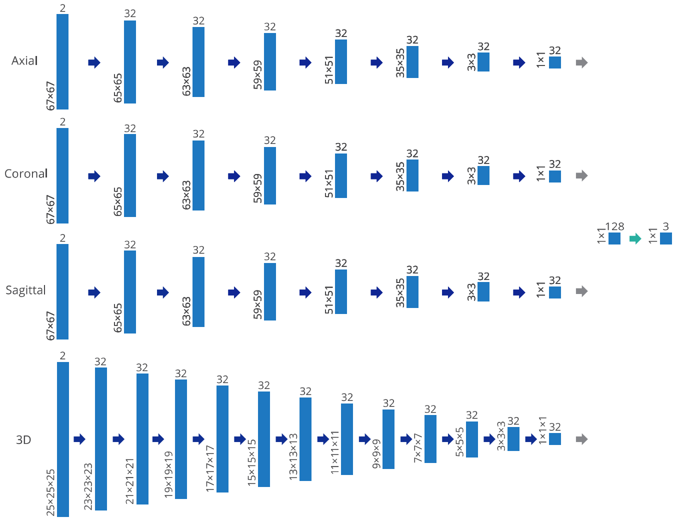

Fig. 6. Network architecture [45]. The colors of the arrows indicate, from left to right: 3 × 3 or 3 × 3 × 3 convolutions, concatenation, and 1 × 1 convolutions. Dilation factors are shown above the arrows. During the training, single voxels are used as output. During the testing, arbitrarily sized outputs can be used, because of the fully convolutional nature of the network.

are concatenated and combined in the output layer with 1×1 convolutions.

Batch normalization and ReLUs were used throughout. Dropout was used before the output layer. The network was trained with Adam based on the cross-entropy loss, using mini-batches of 200 or 300 samples in 10 epochs of 50,000 random samples per class per training image. The network was trained with a patch-based approach, randomly sampling from all images in the training set. During the testing, arbitrarily sized inputs can be used, because of the fully convolutional nature of all four branches. The method took about 1 minute to segment a 3D MRI on a NVIDIA Titan X Pascal GPU. The segmentation results were obtained without any data augmentation. Data augmentation could possibly improve the results in scenarios not well represented in the training set.

E. UPF_Simbiosys: Simbiosys Research Lab at Universitat Pompeu Fabra (UPF), Barcelona [49]

There exist many segmentation approaches, such as multi-atlas label fusion [50], [51] and learning-based methods [52], [53]. Each method has its own strength, and different segmentation approaches may potentially complement each other. The motivation of the proposed method is to combine the strengths of complementary methods in a cascaded fashion.

The pipeline of the method is shown in Fig. 7. The 0-level of the cascade segments the multi-modal (T1w and T2w) input images independently with joint label fusion (JLF) [50]. The estimated probability maps in level-0, along with the original images, are inputted to the level-1 of the cascade. In level-1, first, multi-scale features are extracted from both input images and probability maps of level-0. Image features consist of 1) Gaussian, 2) Laplacian-of-Gaussian, and 3) gradient magnitude images convolved with Gaussians at multiple scales for each modality. Probability features are obtained by convolving the level-0 probability maps with Gaussians at multiple scales.

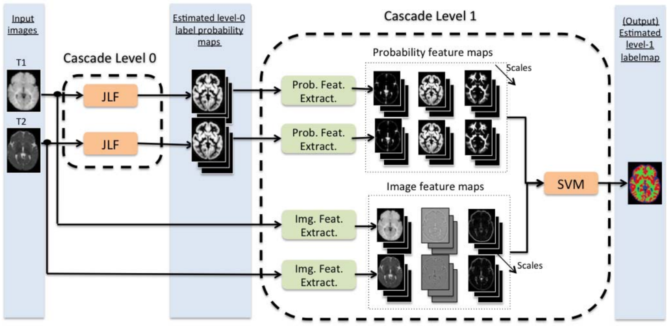

Fig. 7. Dashed blocks correspond to the different levels of the cascade [49]. Blue columns denote input, intermediate output, and final results. Rounded rectangles denote segmentation methods (orange) and feature extraction processes (green), respectively.

The multi-scale image and probability-map features are fed into a SVM classifier for outputting the final estimated label map. Each sample of the SVM classifier is composed of the features extracted from each voxel. The SVM classifier is trained during the training phase using the features extracted from the training set.

Pre-processing steps include 1) histogram matching of all the images to the UNC 1-year-old infant template [20], and 2) non-rigid registration to the same template using ANTs [54]. Pair-wise registrations for multi-atlas JLF are computed by concatenating registrations through the template. No post-processing steps are applied. The parameters for the segmentation methods in each level (i.e., JLF and SVM) are chosen by cross-validation in the training set. Specifically, for JLF, the patch radius is set to be 2 for both modalities and the search window is set to be 7 and 5 for T1w and T2w images, respectively. For SVM, we set the regularization constant to C=5, use an RBF kernel, and normalize the features to zero-mean and unit standard deviation. The computational time for segmenting each subject is ∼30 minutes.

The performance of the SVM classifier in level-1 is highly influenced by the features derived from JLF in level-0. This suggests the advantage of combining multiple complementary methods in the proposed cascaded scheme. A slight drop in performance is experienced by adding an extra layer in the cascade by the level-1 outputs using as the input, so the two-levels scheme is kept as the final model. Among different combination strategies, the proposed cascaded scheme performed better than an alternative ensembling strategy [55].

F. NeuroMTL: Montreal Neurological Institute, McGill University, Montreal QC Canada[8] [56]

First, an extended training dataset was created by applying existing tissue classification to scans from the longitudinal dataset of infants at-risk of autism and control subject in the Infant Brain Imaging Study (IBIS) [57] where scans of 24-month old infants for whom 6 and 12 month scans were available and had T1w and T2w scans acquired at all time points (n=216).

[8] Fonov et al. acknowledged imaging data was collected as part of the Infant Brain Imaging Study (IBIS). Fonov et al. also thank IBIS children and families for their ongoing participation in this longitudinal study.

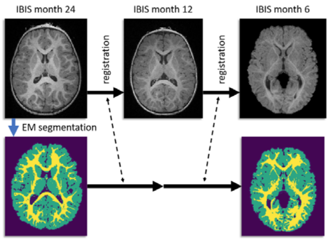

Fig. 8. Automatic segmentation of 6-month old infant MRI data.

TABLE I
PARAMETERS OF 3D U-NET

Layer	Input Channels	Output Channels	Convolution kernel 2	Convolution kernel 2	Upsampling kernel
1	4	64	5x5x5	5x5x5	5x5x5
2	16	64	5x5x5	5x5x5	3x3x3
3	16	64	3x3x3	3x3x3	3x3x3
4	16	64	3x3x3	3x3x3	3x3x3
5	32	64	1x1x1	3x3x3	-

Tissue classification method is shown in Fig. 8: i) An unbiased population average of T1w scans for each age group (6 months, 12 months and 24 months) was created [58]. ii) The group average for the 24-month old scans was manually segmented into areas of high probability of WM, GM and CSF. iii) All 24-month-old T1w scans were non-linearly registered to the template, and tissue priors from the template were transformed to the space of each subject's scan. iv) An expectation-maximization algorithm was run to obtain tissue classification. v) Longitudinal non-linear registrations between scans at 6 and 12 months and then between 12 and 24 months were performed using ANTs with mutual information [54], using both T1w and T2w scans. Using these registration transformations, tissue classification maps from 24 months were transformed to the 6-month scans. Segmentations from the 24-month scans were propagated back to the 6-month scans via non-linear registration. Then, a 3D U-Net [30] was trained in two stages with the extra dataset to automatically segment healthy tissues. U-Net with 5 downsampling and upsampling blocks with skip connections was trained on 80×80×80 image patches for tissue classification, with the parameters listed in Table I. Each block contained two convolutional layers with ReLU activations, with 5×5×5 convolution layers in the first two blocks, 3×3×3 convolution layers in the next two blocks, and a combination of 3×3×3 and 1×1×1 convolution layers in the fifth block, with max pooling at each block. Additionally, a 3×3×3 convolution layer with 64 input and output channels was added, followed by a1×1×1 convolutional layer with 64 input and 32 output channels and then another 1×1×1 layer with 32 input and 4 output channels with dropout, optimizing categorical cross-entropy with Adam. The output patch was cropped to 64×64×64 to remove edge effects. Training was done in two stages, first on the IBIS dataset, and then fine-tuned on the iSeg-2017 challenge data (n=10).

All experiments were performed on a computer with Xeon CPU E5-2620 v4 @ 2.10 GHz with 64GB of ram and NVIDIA Titan-X GPU, with deep-net implemented in Torch7. Training on ACE-IBIS dataset took approximately 32 hours (10000 mini-batches), and final training on iSeg-2017 data took 11 hours (4000 mini-batches). Application on a single subject, using GPU, took 8 seconds.

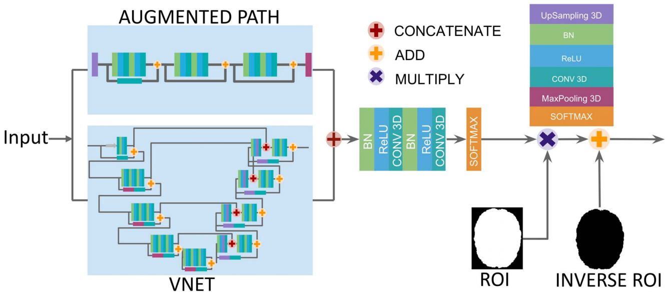

Fig. 9. Augmented V-Net. It builds upon the concatenation of the V-Net core network [59] with an augmented path with higher resolution. Augmented V-Net uses a ROI-mask to train only in brain tissue voxels. Layer types are color-coded as shown in the top-right corner.

G. UPC_DLMI: Signal Theory and Communications Department, Universitat Politècnica de Catalunya, Barcelona

Milletari and Ahmadi [59] proposed a convolutional neural network, named Augmented V-Net (Fig. 9), which is an extension of the V-Net architecture. The main changes with respect to the original V-Net model can be summarized as follows:

1) *Augmented path:* An upsampled version of the input is used to exploit high resolution features. This is done by upsampling by repetition the input (factor of 2) and stacking several convolutional layers after the upsampling. The resultant features are concatenated in the last layers.
2) *Modified residual connections:* The residual connections are reformulated such that the propagation of the input signal through the network is minimally modified.
3) *Mask:* A mask is used before the final prediction in order to constrain the network to train on relevant voxels.
4) *Input concatenation:* The raw input image is used as feature map in the last stages of the network.

The key part of the network is the augmented path, which has been shown to boost the performance of the standard V-Net for the infant brain segmentation task. It provides high-resolution features by keeping small filter sizes and adding redundancy in the input, helping to detect finer regions such as boundaries. Later in the network, the authors use the input image as raw features, since voxel's intensities already contain valuable information. Finally, the mask is used to train/predict only on voxels of brain tissue.

T1w and T2w MRIs are used as input images. Both are normalized to zero mean and unit variance. From the normalized T1w image, a mask is created to mask out background voxels. When training such a big and deep network, there are two main problems: GPU memory constraints and the scarcity of data. Patch-wise training arises as a possible solution for the first issue. The memory required to train Augmented V-Net does not allow using dense-training, which is also discouraged when data is scarce. Larger patch sizes are preferred because they can encode localization features (brain structures) across the network, while smaller patches allow increasing the batch size

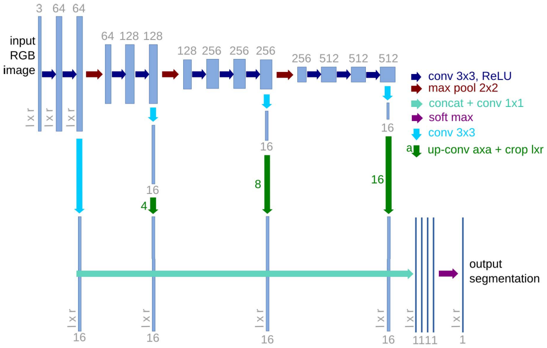

Fig. 10. Visualization of the proposed segmentation network [60].

in the optimization process. The authors finally choose patches of size 64×64×64 and sample uniformly across the brain, forcing the central voxel to belong to brain tissue (WM, GM and CSF). This size provides the best trade-off between local and contextual information providing faster convergence and lower generalization error.

The authors used data augmentation to increase the size of the training set, by making sagittal reflections of each subject. Other reflections have been shown to produce worse results, and no other datasets were used to train the network. In the optimization process, they used Adam optimizer with initial learning rate of lr=0.0005. The loss function used was the weighted cross-entropy, where loss weights were computed as the normalized inverse of the class frequency. At inference time, the whole subject can be used as input for the trained model, performing dense inference and using the mask to indicate brain tissue voxels. The method is fully automatic, taking from 5 to 7 seconds to process one subject.

H. LRDE: Epita Research & Development Laboratory [60]

Xu *et al.'s* method is an extension from single modality to multi-modality of the authors' previous work on neonatal infant brain MRI segmentation [60]. This automatic method uses fully convolutional network (FCN) and transfer learning (see details in Fig. 10), and is very fast: the segmentation of a whole volume only takes a few seconds. The core part of the 16 layers VGG network [55] is used, which was pre-trained on millions of 2D color natural images in ImageNet (for image classification purpose), and fine-tuned with the MRI training dataset. The key contribution is to show how to build 2D color images from a 3D MRI volume, so that VGG effectively gives state-of-the-art segmentation results.

The combination of the T1w and T2w slices to obtain a set of 2D color (RGB) images is very simple. For each slice (indexed by n), the fake color image is constructed in such a way that the "green" channel is the T2w slice n, and the red and the blue are T1w slices respectively at indices *n-1* and *n+1*. Each 2D color image thus forms a *3D-like* representation of a part (3 consecutive slices) of the MR volume. This representation enables incorporating some 3D information, while avoiding the expensive computational and memory requirements of fully 3D CNN. For this specific application, the fully connected layers at the end of VGG network are discarded; only the 4 stages of convolutional parts called "base network" are retained. This base network is composed of convolutional layers, ReLU layers and max pooling layers between two successive stages. The three max pooling layers divide the base network into four stages of fine to coarse feature maps. A stack of specialized layers is obtained, 1 from each stage, and a softmax function yields the segmentation result.

TABLE II
SOURCE CODES FROM TOP-RANKED TEAMS IN ISEG-2017

TEAM	LINK
MSL_SKKU	https://github.com/tbuikr/3D_DenseSeg
LIVIA	https://github.com/josedolz/SemiDenseNet
Bern_IPMI	https://github.com/zengguodong/iSeg_Bern_IPMI
TU/e IMAG/e	https://github.com/pimmoeskops/iSeg_dilatedCNN
NeuroMTL	https://github.com/vfonov/NeuroMTL_iSEG
UPC_DLMI	https://github.com/imatge-upc/segmentation_DLMI/
LRDE	https://www.lrde.epita.fr/wiki/NeoBrainSeg

Before creating the set of 2D color images, a pre-processing of the T1w and T2w sequences was performed, which consists of: 1) shifting the voxel values of the MRI volumes to center their histograms on their maximal histogram value, and 2) requantizing the voxel values on 8 bit (values lower than 0 and greater than 255 are saturated). For the training, the classical data augmentation strategy by scaling and rotating images were adopted. 2D images were then computed for each volume of the augmented training base using the pre-processed T1w and T2w slices as described before. The network was fine-tuned for the first 50K iterations with a learning rate of $lr = 10^{-8}$, and the last 100K with a smaller learning rate ($lr = 10^{-10}$). Stochastic gradient descent was employed to minimize the loss function with $momentum = 0.99$ for the first 50K iterations and 0.999 for the next 100k, and $weight\_decay = 0.0005$. The loss function was averaged over 20 images. During test, the runtime on a 3D volume was 1.8 seconds on average; note that this included the pre-processing step, the computation of the set of 2D color input images, and after inference, the reconstruction of a 3D volume (the expected segmentation output) by stacking the set of 2D output images.

I. Source Codes

A proactive goal of this paper is to encourage authors to make their codes publicly available for reproducible research. By far, most of teams have shared their codes, as summarized in Table II. For readers who seek to come up to speed with deep learning, these codes can be also served as good starting points to understand how deep learning algorithms can be implemented for image segmentation.

V. DISCUSSION

Based on Section IV, 7 out of 8 top-ranked teams adopted deep learning based algorithms. Moreover, most of the deep learning related algorithms are based on 3D U-Net (or U-Net-like structures). Thanks to the use of GPUs, most of these

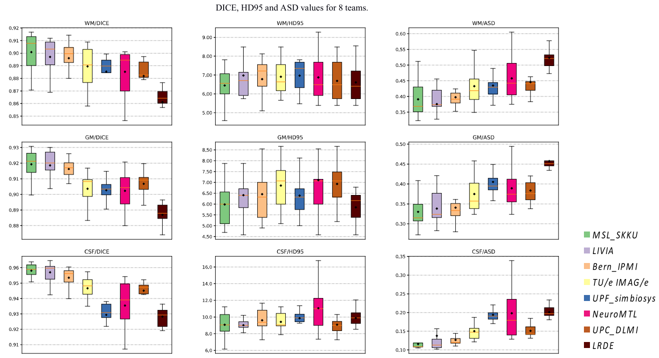

Fig. 11. Performances of the eight top-ranked teams, in terms of DICE, HD95 and ASD, using box-plots. Besides medians, the means are also indicated by the dark dots.

algorithms have inference times between 5-10 seconds for a whole MR scan. The only non-deep learning based method is developed by Sanroma et al. (*UPF_simbiosys*), which employs a multi-atlas based method followed by an SVM to design a cascade learning segmentation algorithm.

A. Evaluation in Terms of the Whole Brain

We first evaluate the performance in terms of the whole brain. Fig. 11 reports performances of the 8 top-ranked teams in terms of DICE, HD95 and ASD by employing box-plots. Besides medians, the means are also indicated by the black diamonds. To know whether any method performs significantly better than the others, we calculated Wilcoxon signed-rank test, as shown in Appendix Table III with all-against-all diagram in terms of three metrics (DICE, HD95 and ASD). Interestingly, we did not find any method achieving strong statistically significant better performances compared to all other methods, for segmentation of WM, GM and CSF across any metric (DICE, HD95 or ASD). For example, we found that the results from *MSL_SKKU* present the highest median in terms of DICE for WM. Nevertheless, their differences with the results obtained by *LIVIA* and *Bern_IPMI* are not strongly statistically significant. In terms of HD95, the results obtained with the networks proposed by *LRDE* and *MSL_SKKU* have the lowest median for WM and GM, respectively, but still there is no strong, statistically significant difference with any other methods.

B. Evaluation in Terms of ROIs

Besides evaluation in terms of the whole brain, we further evaluate the performances based on 80 ROIs. Specifically, a total of 33 two-year-old subjects were employed as individual atlases (www.brain-development.org) [61]. Each atlas consists of a T1w MR image and a label image of 80 ROIs (excluding cerebellum and brainstem). We first employ FreeSurfer [62] to segment each T1w MR image into WM, GM, and CSF. Then, we warp all atlases to each testing subject's space based on their tissue segmentation maps using ANTs [63]. Finally, we employ a majority voting to parcellate each testing subject into 80 ROIs. For each ROI, we employed DICE to measure the performance between automatic segmentations and manual segmentation. ROI-based DICE values for 8 top-ranked teams are shown in Appendix Table IV. Due to large number of ROIs, p-values for each ROI are not reported in this paper. However, to better interpret these ROI-based evaluations, we have generated error maps for each method, as shown in Fig. 12. They were estimated by aligning all the error maps from 13 testing subjects to a 6-month template [64]. The higher value of error map, the higher probability for miss-classification. From all these error maps, we can see all methods consistently produce small errors in the subcortical regions, while with larger errors in the cortical regions, which is actually consistent with the fact that tissue contrast is much lower in the cortical regions than subcortical regions. Average error maps for all 8 top-ranked methods were further generated, as shown in the right bottom of Fig. 12. The most error-prone ROI regions are the straight gyrus, lingual gyrus, and medial orbital gyrus. These regions are also consistently confirmed with results given in Appendix Table IV, where DICE values of these ROIs are relatively low, i.e., around 0.84. By contrast, the DICE values of subcortical regions, such as putamen and thalamus, are higher, i.e., around 0.94.

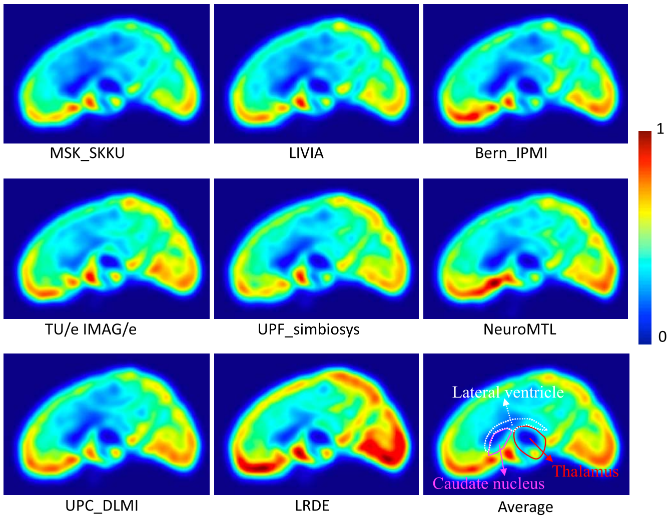

Fig. 12. Error maps: all 8 top-ranked methods produce small errors in the subcortical regions, but large errors in the cortical regions. The most error-prone regions are the straight gyrus, lingual gyrus, and medial orbital gyrus. Average error map for all 8 top-ranked methods is shown in the right bottom, with the subcortical mask (caudate nucleus and thalamus). Color bar is from 0 to 1, with the high values indicating large errors.

C. Evaluation in Terms of Gyral Landmark Curves

To better reflect the accuracy of the 8 top-ranked methods, we further measured the distance of gyral landmark curves on the cortical surfaces. Large curve distance indicates poor performance on the gyral crest. We selected two major gyri, i.e., the superior temporal gyral curve and the postcentral gyral curve, as the landmarks to measure the accuracy. We manually labeled two sets of gyral curves on the inner cortical surfaces from different tissue segmentation results [65]. One typical example is shown in Fig. 13, in which curves were delineated by the experts on the superior temporal gyrus and postcentral gyrus: the white curves indicate the ground truth, and the colored curves indicate results by different methods. We employed HD95 to calculate the curve distance,

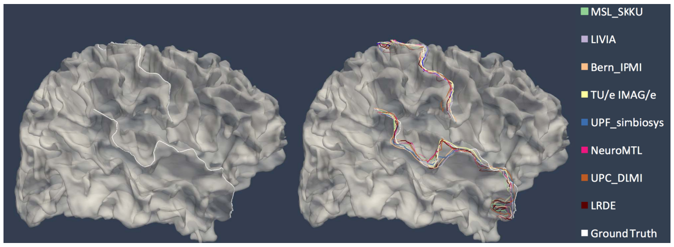

Fig. 13. Evaluations on gyri for 8 top-ranked methods. The left one shows the manually labeled postcentral and superior temporal gyral landmark curves, used as ground truth; and the right one shows the gyral curves from the segmentation results of 8 top-ranked methods, compared with the ground truth.

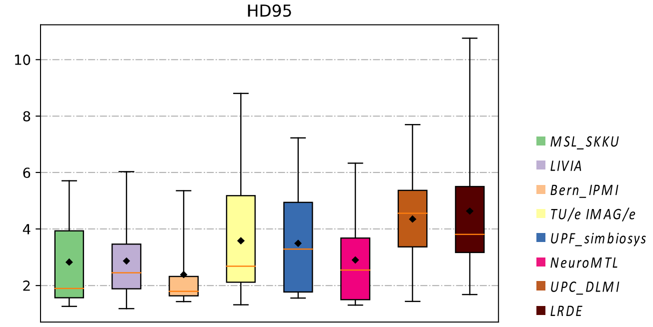

Fig. 14. The boxplot shows HD95 evaluations of 8 top-ranked methods on the superior temporal gyrus and the postcentral gyrus of the 13 testing subjects. Besides medians, the means are also indicated by the dark diamonds.

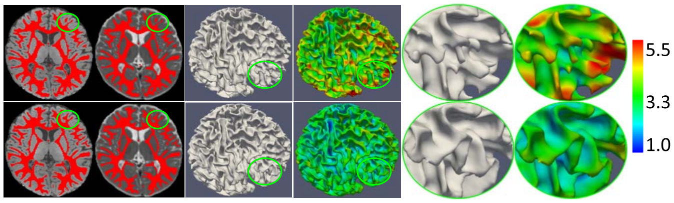

Fig. 15. Comparison with *MSL_SKKU* [31] in 2017 MICCAI Grand Segmentation Challenge (iSeg-2017). The results by *MSL_SKKU* and manual segmentation are shown in the 1st and 2nd rows, respectively. From left to right: segmentation overlaid on T1w and T2w images, inner cortical surface, cortical thickness map, and zoomed views of inner cortical surface and cortical thickness map (in mm).

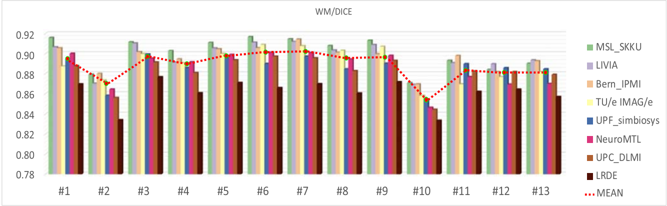

Fig. 16. WM DICE values for each subject by different methods, with the average DICE represented by the red dashed line.

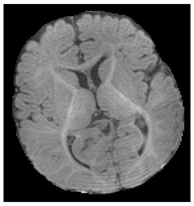

Fig. 17. The 10th testing subject with motion and unusual scan pose.

with the median HD95 over 13 testing subjects reported in Fig. 14. The *p*-values were calculated based on Wilcoxon signed-rank test, as shown in Appendix Table V. We find that *Bern_IPMI* achieves the lowest median HD95, but no statistically significant difference with *MSL_SKKU*, *LIVIA*, *UPF_simbiosys*, and *NeuroMTL*.

Based on the above evaluations, in terms of the whole brain, small ROIs, and gyral curves, we can observe that none of these 8 top-ranked methods has achieved a strong, statistically significant better performance than all other methods. Especially, from the error maps in Fig. 12, these methods consistently have a poor performance along the cortical regions. Therefore, there is still opportunity for improvement.

First, all methods directly apply well-established models (e.g., U-Nets) on the challenge, without considering any *prior* knowledge of infant brain images [66], [67], e.g., cortical thickness is within a certain range. Especially, due to low contrast between WM and GM in the 6-month infant brain images, WM voxels may be under/over segmented. Given a voxel with a resolution of $1\times1\times1$ mm^3, although one voxel error will have a negligible impact on DICE or HD95, it will result in ±1 mm estimation error of cortical thickness. Fig. 15 shows a segmentation result on a testing subject obtained by *MSL_SKKU* [31]. Without anatomical guidance, there are many missing gyrus in the reconstructed inner surface by *MSL_SKKU* [31]. Consequently, the estimated cortical thickness is abnormally thicker. It is worth noting that this type of error should be seriously considered, especially for possible biomarker identification, since this will lead to difficulty of accurately characterizing brain developmental attributes, i.e., cortical thickness, gyrification, and convexity. For example, the cortical thickness of the zoomed regions (the last column of Fig. 15) is abnormally larger than the ground truth.

Second, all methods ignore a fact that tissue contrast between CSF and GM is much higher than that between GM and WM. Therefore, it might be reasonable to identify CSF first from infant brain images to reconstruct the outer cortical surface and use it as a guidance to estimate the inner cortical surface, since cortical thickness is within a certain range. Preliminary work on 6-month infant subjects with risk of autism demonstrates the effectiveness of this kind of strategy [66], [67].

Third, we have inspected the performances of different methods for each subject. Fig. 16 shows DICE values for each subject, with the average DICE represented by the red dashed line. Among all 13 testing subjects, we find that all methods consistently performed badly on the 2^{nd} and 10^{th} testing subjects, which were acquired with motion artifacts. Especially, the 10^{th} testing subject presents severe motion artifacts, with one representative slice shown in Fig. 17. Another possible reason could be the different scan pose of this subject, compared to other testing subjects. Therefore, the models with robustness to the motion or the scan pose are highly desired, since the motion is inevitable and these types of scan variation are normal during image acquisition. A possible solution to address these issues is to augment the training images with different rotation degrees, flipping, and simulated motion artifacts.

Fourth, to better compare these 8 top-ranked methods, Appendix Table VI further lists their key highlights, as well as various detailed implementations. For example, all these 8 top-ranked methods randomly selected samples (2D/3D patches) from the training images using moving windows, without evaluating the importance of each sample. For example, in the conventional machine learning algorithms, adaptive boosting is an effective strategy to learn features from those error-prone regions to improve the performance [68]. Similarly, the average error map shown in Fig. 12 could potentially provide guidance for selecting effective samples for training. For example, by selecting more training samples from those error-prone regions, the performance of these segmentation algorithms could be further improved. In addition, from Appendix Table VI, we can see the patch size used in these 8 top-ranked methods varies dramatically from 24×24×24 to 80×80×80, which could be further optimized for achieving better results.

Finally, we would like to indicate limitations for iSeg-2017. First, we only reviewed the 8 top-ranked teams. Some works from the remaining teams are also interesting but not included in this paper, due to space limit. For example, Bernal *et al.* [69] extended a multi-resolution fully convolutional neural to deal with segmentation of 6-month infant brain MRI. Hashemi *et al.* [70] proposed an exclusive multi-label multi-class training strategy to deal with classes that have highly overlapping features. Second, the number of training subjects and the number of testing subjects are small. Third, image resolution is low, especially for T2w images with 1.25×1.25×1.95 mm^3 of voxel size. Actually, in the current BCP imaging protocol [3], T1w and T2w images are acquired with 0.8×0.8×0.8 mm^3. These limitations will be alleviated, such as by including subjects acquired in BCP, for our planned 2019 iSeg Grand Challenge (https://iseg2019.web.unc.edu).

VI. Conclusion

In this paper, we have reviewed and summarized 21 automatic segmentation methods participating in iSeg-2017. Especially, we have elaborated the details of 8 top-ranked methods: including the pipeline, implementation, and source code. We further pointed out limitations and possible future directions. The iSeg-2017 website is always open and we hope our manual labels in iSeg-2017, this review article and source codes could greatly advance methodological development in the community.

References

[1] G. Li *et al.*, "Mapping region-specific longitudinal cortical surface expansion from birth to 2 years of age," *Cerebral Cortex*, vol. 23, pp. 2724–2733, Nov. 2013.

[2] G. Li *et al.*, "Mapping longitudinal hemispheric structural asymmetries of the human cerebral cortex from birth to 2 years of age," *Cerebral Cortex*, vol. 24, no. 5, pp. 1289–1300, May 2014.

[3] B. R. Howell *et al.*, "The UNC/UMN Baby Connectome Project (BCP): An overview of the study design and protocol development," *NeuroImage*, vol. 185, pp. 891–905, Jan. 2019.

[4] G. Li *et al.*, "Early diagnosis of autism disease by multi-channel CNNs," in *Machine Learning in Medical Imaging*. Cham, Switzerland: Springer, 2018, pp. 303–309.

[5] L. Wang *et al.*, "Segmentation of neonatal brain MR images using patch-driven level sets," *NeuroImage*, vol. 84, pp. 141–158, Jan. 2014.

[6] L. Wang *et al.*, "LINKS: Learning-based multi-source IntegratioN frameworK for Segmentation of infant brain images," *NeuroImage*, vol. 108, pp. 72–160, Mar. 2015.

[7] S. Hu *et al.*, "Learning-based deformable image registration for infant MR images in the first year of life," *Med. Phys.*, vol. 44, no. 1, pp. 158–170, Jan. 2017.

[8] F. Shi *et al.*, "Neonatal atlas construction using sparse representation," *Hum. Brain Mapping*, vol. 35, no. 9, pp. 4663–4677, Sep. 2014.

[9] F. Shi *et al.*, "Construction of multi-region-multi-reference atlases for neonatal brain MRI segmentation," *NeuroImage*, vol. 51, no. 2, pp. 684–693, Jun. 2010.

[10] T. Paus, D. L. Collins, A. C. Evans, G. Leonard, B. Pike, and A. Zijdenbos, "Maturation of white matter in the human brain: A review of magnetic resonance studies," *Brain Res. Bull.*, vol. 54, pp. 255–266, Feb. 2001.

[11] Z. Xue, D. Shen, and C. Davatzikos, "CLASSIC: Consistent longitudinal alignment and segmentation for serial image computing," *NeuroImage*, vol. 30, no. 2, pp. 388–399, Apr. 2006.

[12] G. Li *et al.*, "Computational neuroanatomy of baby brains: A review," *NeuroImage*, vol. 185, pp. 906–925, Jan. 2018.

[13] I. Išgum *et al.*, "Evaluation of automatic neonatal brain segmentation algorithms: The NeoBrainS12 challenge," *Med. Image Anal.*, vol. 20, no. 1, pp. 135–151, Feb. 2015.

[14] A. M. Mendrik *et al.*, "MRBrainS challenge: Online evaluation framework for brain image segmentation in 3T MRI scans," *Comput. Intell. Neurosci.*, vol. 2015, Jan. 2015, pp. 1–16.

[15] S. Winzeck *et al.*, "Isles 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI," *Frontiers Neurol.*, vol. 9, p. 679, Sep. 2018.

[16] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[17] W. Zhang *et al.*, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, Mar. 2015.

[18] L. Wang *et al.*, "Integration of sparse multi-modality representation and anatomical constraint for isointense infant brain MR image segmentation," *NeuroImage*, vol. 89, pp. 64–152, Apr. 2014.

[19] D. Nie, L. Wang, E. Adeli, C. Lao, W. Lin, D. Shen, "3-D fully convolutional networks for multimodal isointense infant brain image segmentation," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1123–1136, Mar. 2019.

[20] F. Shi, L. Wang, Y. Dai, J. H. Gilmore, W. Lin, and D. Shen, "LABEL: Pediatric brain extraction using learning-based meta-algorithm," *NeuroImage*, vol. 62, no. 3, pp. 1975–1986, 2012.

[21] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *IEEE Trans. Med. Imag.*, vol. 17, no. 1, pp. 87–97, Feb. 1998.

[22] L. Wang, F. Shi, P.-T. Yap, W. Lin, J. H. Gilmore, and D. Shen, "Longitudinally guided level sets for consistent tissue segmentation of neonates," *Hum. Brain Mapping*, vol. 34, no. 4, pp. 956–972, Apr. 2013.

[23] Y. Dai, F. Shi, L. Wang, G. Wu, and D. Shen, "iBEAT: A toolbox for infant brain magnetic resonance image processing," *Neuroinformatics*, vol. 11, no. 2, pp. 211–225 , Apr. 2013.

[24] J. G. Chi, E. C. Dooling, and F. H. Gilles, "Gyral development of the human brain," *Ann. Neurol.*, vol. 1, no. 1, pp. 86–93, Jan. 1977.

[25] E. Armstrong, A. Schleicher, H. Omran, M. Curtis, and K. Zilles, "The ontogeny of human gyrification," *Cerebral Cortex*, vol. 5, pp. 56–63, Jan./Feb. 1995.

[26] J. Hill *et al.*, "A surface-based analysis of hemispheric asymmetries and folding of cerebral cortex in term-born human infants," *J. Neurosci.*, vol. 30, pp. 2268–2276, Feb. 2010.

[27] J. Dubois *et al.*, "Mapping the early cortical folding process in the preterm newborn brain," *Cerebral Cortex*, vol. 18, pp. 1444–1454, Jun. 2008.

[28] S. Abe, K. Takagi, T. Yamamoto, Y. Okuhata, and T. Kato, "Assessment of cortical gyrus and sulcus formation using MR images in normal fetuses," *Prenatal Diagnosis*, vol. 23, no. 3, pp. 225–231, Mar. 2003.

[29] C. Lebel, L. Walker, A. Leemans, L. Phillips, and C. Beaulieu, "Microstructural maturation of the human brain from childhood to adulthood," *NeuroImage*, vol. 40, no. 3, pp. 1044–1055, Apr. 2008.

[30] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. MICCAI*, 2016, pp. 424–432.

[31] T. D. Bui, J. Shin, and T. Moon. (2017). "3D densely convolutional networks for volumetric segmentation." [Online]. Available: https://arxiv.org/abs/1709.03199

[32] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, Jun. 2017, pp. 2261–2269.

[33] S. Ioffe and C. Szegedy. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167

[34] X. Glorot, A. Bordes, and Y. Bengio, "Deep Sparse Rectifier Neural Networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

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

[37] D. P. Kingma and J. Ba. (2017). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. ICCV*, 2015, pp. 1026–1034.

[39] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[40] J. Dolz, C. Desrosiers, L. Wang, J. Yuan, D. Shen, and I. Ben Ayed. (2017). "Deep CNN ensembles and suggestive annotations for infant brain MRI segmentation." [Online]. Available: https://arxiv.org/abs/1712.05319

[41] J. Dolz, C. Desrosiers, and I. B. Ayed, "3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study," *NeuroImage*, vol. 170, pp. 456–470, Apr. 2018.

[42] G. Zeng and G. Zheng, "Multi-stream 3D FCN with multi-scale deep supervision for multi-modality isointense infant brain MR image segmentation," in *Proc. ISBI*, Apr. 2018, pp. 136–140.

[43] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[44] R. Kimmel, N. Kiryati, and A. M. Bruckstein, "Sub-pixel distance maps and weighted distance transforms," *J. Math. Imag. Vis.*, vol. 6, nos. 2–3, pp. 223–233, Jun. 1996.

[45] P. Moeskops and J. P. W. Pluim. (2017). "Isointense infant brain MRI segmentation with a dilated convolutional neural network." [Online]. Available: https://arxiv.org/abs/1708.02757

[46] 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.

[47] F. Yu and V. Koltun. (2015). "Multi-scale context aggregation by dilated convolutions." [Online]. Available: https://arxiv.org/abs/1511.07122

[48] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Dilated convolutional neural networks for cardiovascular MR segmentation in congenital heart disease," in *Reconstruction, Segmentation, and Analysis of Medical Images*. Cham, Switzerland: Springer, 2016, pp. 95–102.

[49] G. Sanroma, O. M. Benkarim, G. Piella, and M. A. G. Ballester, "Building an ensemble of complementary segmentation methods by exploiting probabilistic estimates," in *Machine Learning in Medical Imaging*. Cham, Switzerland: Springer, 2016, pp. 27–35.

[50] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, and P. A. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 611–623, Mar. 2013.

[51] G. Wu, Q. Wang, D. Zhang, F. Nie, H. Huang, and D. Shen, "A generative probability model of joint label fusion for multi-atlas based brain segmentation," *Med. Image Anal.*, vol. 18, no. 6, pp. 881–890, 2014.

[52] Y. Hao *et al.*, "Local label learning (LLL) for subcortical structure segmentation: Application to hippocampus segmentation," *Hum. Brain Mapping*, vol. 35, no. 6, pp. 2674–2697, Jun. 2014.

[53] P. Moeskops *et al.*, "Automatic segmentation of MR brain images of preterm infants using supervised classification," *NeuroImage*, vol. 118, pp. 628–641, Sep. 2015.

[54] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ANTs similarity metric performance in brain image registration," *NeuroImage*, vol. 54, no. 3, pp. 2033–2044, Feb. 2011.

[55] K. Simonyan and A. Zisserman. (2015). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[56] V. Fonov, A. Doyle, A. C. Evans, and L. Collins. (2018). *NeuroMTL iSEG Challenge Methods*. [Online]. Available: https://www.biorxiv.org/content/10.1101/278465v1

[57] H. C. Hazlett, H. Gu, and The IBIS Network, "Early brain development in infants at high risk for autism spectrum disorder," *Nature*, vol. 542, pp. 348–351, Feb. 2017.

[58] V. Fonov, A. C. Evans, K. Botteron, C. R. Almli, R. C. McKinstry, and D. L. Collins, "Unbiased average age-appropriate atlases for pediatric studies," *NeuroImage*, vol. 54, no. 1, pp. 313–327, Jan. 2011.

[59] F. Milletari, N. Navab, and S.-A. Ahmadi. (2016). "V-Net: Fully convolutional neural networks for volumetric medical image segmentation." [Online]. Available: https://arxiv.org/abs/1606.04797

[60] Y. Xu, T. Géraud, and I. Bloch, "From neonatal to adult brain MR image segmentation in a few seconds using 3D-like fully convolutional network and transfer learning," in *Proc. ICIP*, Sep. 2017, pp. 4417–4421.

[61] I. S. Gousias *et al.*, "Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest," *NeuroImage*, vol. 40, pp. 672–684, Apr. 2008.

[62] B. Fischl, "FreeSurfer," *NeuroImage*, vol. 62, pp. 774–781, Aug. 2012.

[63] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Med. Image Anal.*, vol. 12, no. 1, pp. 26–41, Feb. 2008.

[64] Y. Zhang, F. Shi, G. Wu, L. Wang, P.-T. Yap, and D. Shen, "Consistent spatial-temporal longitudinal atlas construction for developing infant brains," *IEEE Trans. Med. Imag.*, vol. 35, no. 12, pp. 2568–2577, Dec. 2016.

[65] G. Li, L. Guo, J. Nie, and T. Liu, "An automated pipeline for cortical sulcal fundi extraction," *Med. Image Anal.*, vol. 14, pp. 343–359, Jun. 2010.

[66] L. Wang *et al.*, "Volume-based analysis of 6-month-old infant brain MRI for autism biomarker identification and early diagnosis," in *Proc. MICCAI*, 2018, pp. 411–419.

[67] L. Wang *et al.*, "Anatomy-guided joint tissue segmentation and topological correction for 6-month infant brain MRI with risk of autism," *Hum. Brain Mapping*, vol. 39, no. 6, pp. 2609–2623, Jun. 2018.

[68] S. Shalev-Shwartz. (2017). "SelfieBoost: A boosting algorithm for deep learning." [Online]. Available: https://arxiv.org/abs/1411.3436

[69] J. Bernal, K. Kushibar, M. Cabezas, S. Valverde, A. Oliver, and X. Lladó. (2018). "Quantitative analysis of patch-based fully convolutional neural networks for tissue segmentation on brain magnetic resonance imaging." [Online]. Available: https://arxiv.org/abs/1801.06457

[70] S. R. Hashemi, S. P. Prabhu, S. K. Warfield, and A. Gholipour, "Exclusive independent probability estimation using deep 3D fully convolutional DenseNets: Application to IsoIntense infant brain MRI segmentation," in *Proc. 2nd Int. Conf. Med. Imag. Deep Learn.*, 2019, pp. 1–14.