DAM-AL: Dilated Attention Mechanism with Attention Loss for 3D Infant Brain Image Segmentation

Dinh-Hieu Hoang
University of Science, Ho Chi Minh City, Vietnam
John von Neumann Institute, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
hieu.hoang2020@ict.jvn.edu.vn

Gia-Han Diep
University of Science, Ho Chi Minh City, Vietnam
John von Neumann Institute, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
han.diep@ict.jvn.edu.vn

Minh-Triet Tran
University of Science, Ho Chi Minh City, Vietnam
John von Neumann Institute, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
tmtriet@hcmus.edu.vn

Ngan T.H Le
University of Arkansas, Fayetteville, Arkansas USA
thile@uark.edu

ABSTRACT

While Magnetic Resonance Imaging (MRI) has played an essential role in infant brain analysis, segmenting MRI into a number of tissues such as gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) is crucial and complex due to the extremely low intensity contrast between tissues at around 6-9 months of age as well as amplified noise, myelination, and incomplete volume. In this paper, we tackle those limitations by developing a new deep learning model, named DAM-AL, which contains two main contributions, i.e., dilated attention mechanism and hard-case attention loss. Our DAM-AL network is designed with skip block layers and atrous block convolution. It contains both channel-wise attention at high-level context features and spatial attention at low-level spatial structural features. Our attention loss consists of two terms corresponding to region information and hard samples attention. Our proposed DAM-AL has been evaluated on the infant brain iSeg 2017 dataset and the experiments have been conducted on both validation and testing sets. We have benchmarked DAM-AL on Dice coefficient and ASD metrics and compared it with state-of-the-art methods. Code is available at: https://github.com/DinhHieuHoang/DAM-CA-InfantBrain

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Image segmentation**; *Supervised learning*; *Neural networks*; • **Applied computing** → *Health informatics*;

KEYWORDS

Attention Loss, Channel-wise Attention, Spatial Attention, Infant Brain, Segmentation

1 INTRODUCTION

Accurate infant brain MRI segmentation is a crucial modern technique in recognizing normal and abnormal early brain development [23]. For instance, the report [10] shows that brain overgrowth is associated with an increase in the cortical surface area before two years of age in autistic children. One of the most critical procedures to measure infant brain development and identify biomarkers is accurate segmentation of MRI into different tissue areas, i.e., white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) [13]. Infant brain segmentation is considered to be more challenging than adult brain segmentation due to tissue contrast reduction, amplified noise, myelination, and incomplete volume [37]. Furthermore, the intensity distributions of GM and WM have larger overlapping; thus it is difficult for manual annotation. Hence annotated data is highly limited.

Deep learning, i.e. Convolutional neural networks (CNNs) have obtained great achievement in computer vision including in medical imaging tasks. To address those problems, several approaches leveraged deep learning have been proposed to improve infant brain segmentation accuracy [1, 2, 5, 11, 15–19, 27, 28, 39]. We can generally divide the existing works into two categories: the first category focuses on the network architecture designs, whereas the second category targets proposing loss functions. In the first category, [25, 40] first proposed using 2D CNNs to segment isointense-phase brain images. However, these 2D CNNs are time-consuming since they process each slice independently and fail to capture the spatial contextual information present in the volumetric data. Thus, a 3D network structure has been later developed and outperformed most 2D network architectures. For instance, [5] replaced all 2D operations of the U-Net architecture [31] with 3D counterparts for volumetric biomedical image segmentation without increasing the number of parameters. Later, [11] proposed a tiramisu network as a fully convolutional densenets; [28] proposed 3D FCN for

multimodal isointense infant brain segmentation; [1] proposed skip-connected 3D DenseNet; [7] introduced HyperDenseNet having complex dense connections between paths of different modalities. In the second category, 3D Unet [32] is used as a backbone network, and various proposed loss functions have been evaluated against existing loss functions. Notably, [17] proposed offset curves loss and [15] proposed NB-AC loss for medical image segmentation, and they both mainly target imbalanced class problem. Beyond CNNs, Capnets[33] has been improved to 3D volumetric data to address infant brain segmentation [27]. Some more advanced deep-learning-based methods have also been reported in the iSeg-2017 challenge review [36].

To the best of our knowledge, all existing architectures used in volumetric brain tissue segmentation are based on Unet architecture which contains two paths: encoder path downsamples feature maps to capture the contextual information and decoder path upscales the downsampled feature maps for localization. In those networks, skip-connections is utilized to facilitate information flow from the encoder path to the decoder path. However, those networks treat all data points i.e. voxels equally and there is no mechanism to pay attention on hard data points which are easily misclassified. Infant brain data is captured in low contrast and weak surface, thus segmenting the areas around surface considered as hard-case samples is very challenging. Furthermore, most DL-based segmentation networks have made use of common loss functions, e.g., CE, Dice, Focal. These losses are based on summations over the segmentation regions and are restricted to pixel-wise settings. Not only pixel-wise sensitivity, but these losses are also unfavorable to low contrast which is a big challenge in infant brain segmentation. Furthermore, these losses are working on higher level features of region information and none of them is intentionally designed for lower level features such as edge/surface which play an important role in medical imaging.

In this work, we first present an effective network architecture which is based on UNet [5] and influenced by spatial and channel-wise attention network [4]. Our DAM-AL network contains skip blocks and atrous blocks layers to capture the rich context information at both high-level and low-level features. We then introduce our hard-case attention loss, defined as a surface distance map weighted by hard cases estimation function. Our contribution is summarized as follows:

- Introduce a dilated attention network consists of skip block connection and atrous block layers
- Present spatial attention and channel-wise attention to enrich the high-level context features and low-level spatial structural features.
- Introduce a hard-case attention loss which addresses the low contrast and weak surface problem in infant brain image segmentation.
- The proposed DAM-AL outperforms other state-of-the-art methods on various metrics.

2 PROPOSED METHOD

2.1 Dilated Attention Network

Attention mechanisms have been successfully applied in various computer vision tasks such as object recognition [24], image captioning [4], action detection [35]. In this paper, we leverage SCA-CNN network [4] and propose a dilated attention network that contains channel-wise attention (CA) mechanism to capture context at high level feature and spatial attention (SA) mechanism to capture structure at low level feature. The entire DAM-CA is illustrated in Fig.1.

2.1.1 Multiscale feature extraction. Our feature extraction is based on Unet encoder which contains five layers with skip-block connection. The first two layers (low-level layers) (conv-1, conv-2) conduct a low level feature while the last three layers (high-level layers) (conv-3, conv-4, conv-5) produce a high level feature.

2.1.2 Spatial attention (SA) at the low-level layers: In medical image segmentation, boundaries between objects play an important role; thus, we want to obtain as detailed as possible the boundaries. Instead of considering all spatial positions equally, we adopt SA [3] to focus more on the foreground regions, especially surface areas. Let denote $F_l \in R^{H \times W \times D \times C}$ as feature at a low-level layer. Each voxel in spatial is presented as (x, y, z) in the spatial coordinate. To increase receptive field and extract better global information but not increase parameters, we apply six convolution layers. The kernels of these layers are $1 \times K \times K$, $K \times 1 \times 1$, $K \times 1 \times K$, $1 \times K \times 1$, $K \times K \times 1$, and $1 \times 1 \times K$. Fig.2 illustrates SA applying into the low-level layers to produce low-level features. The final low-level feature is obtained by weighting F_l with SA feature F_{sa}. The process can be presented as follows:

$$F_1 = conv2(conv1(F_l)); F_2 = conv4(conv3(F_l));$$
$$F_3 = conv6(conv5(F_l)); F_{sa} = sigmoid(F_1 + F_2 + F_3); \quad (1)$$
$$F_L = F_{sa}F_l.$$

where $conv1, conv2, conv3, conv4, conv5, conv6$ are corresponding to different kernels.

2.1.3 Channel-wise attention (CA) at the high-level layers: Because different channels of features in CNNs present different semantics, we apply channel-wise attention (CA) [3] to weighted multi-scale at high-level layers. To capture richer structure information with longer range dependence, we adapt atrous convolutions with different dilation rates at the high-level layers. Notably that the CA will assign larger weights to channels that present high response to surface and region-of-interest i.e. WM, GM, CSF. In our case, the dilation rates set to 1, 2, 3, 4. The feature maps from different atrous convolutional layers are then combined by concatenation. Let denote $F_h \in R^{H \times W \times D \times C}$ as feature at a high-level layer, the high-level feature F_H is obtained by weighting F_h with CA feature F_{ca}. The procedure is illustrated as in Fig.3 and formulated as follows:

$$F_{ca} = sigmoid(FC(\delta(FC(Pool(F_h))))); $$
$$F_H = F_{ca}F_h. \quad (2)$$

where FC, δ are fully connected layer and non-linearity function (reLU).
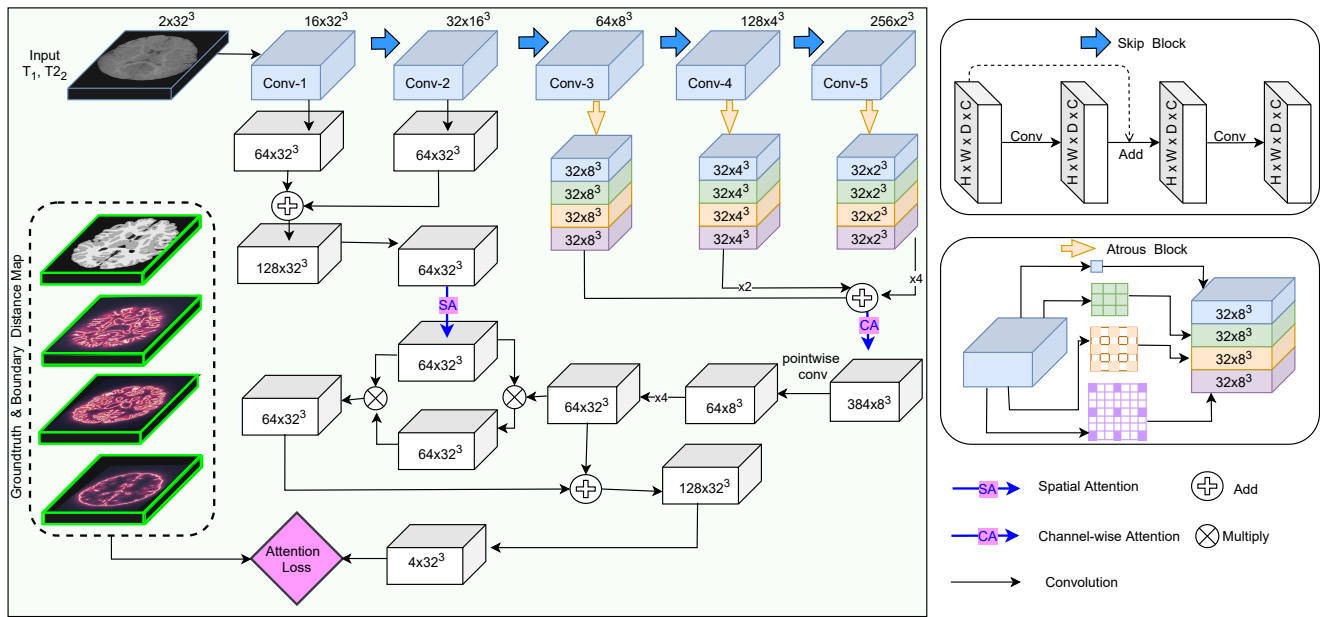
Figure 1: The overall architecture of our proposed DAM-AL. The encoder path is leveraged by Unet contracting path with skip block connections. The low-level layers i.e. Conv-1, Conv-2 are modeled to form low-level feature through spatial attention (SA). The high-level layers i.e. Conv-3, Conv-4, Conv-5 are learnt to generate high-level feature by channel-wise attention (CA). Atrous block is designed to weight high-level layers at multi-scale as well as to extract richer context-aware information.
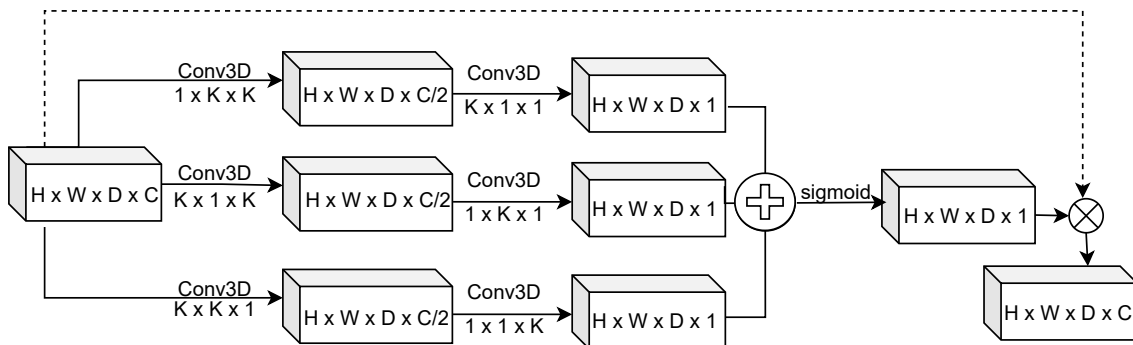

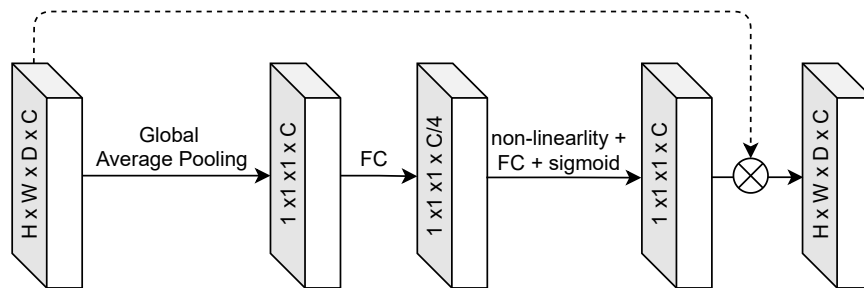Figure 2: The illustration of spatial attention (SA).

Figure 3: The illustration of channel-wise attention (CA).

2.2 Hard-case Attention Loss

To train a Deep Neural Network (DNN), the loss function, known as cost function, plays a significant role. The loss function is to measure the average (expected) divergence between the output of the network (P) and the ground truth (T) being approximated over the entire domain of the input, sized $W \times H \times D$. We denote i as index of each voxel in an volumetric medical image spatial space $N = W \times H \times D$. The label of each class is written as c in C classes.

Herein, we first briefly review and analyze some common loss functions, introduce hard-case samples estimation, and finally present our attention loss function.

2.2.1 Existing loss functions. Cross Entropy (CE), Dice loss, and Focal loss are common loss functions in image segmentation while offset loss, boundary loss are the state-of-the-art loss functions that address the problems of imbalanced data and weak surface in medical image segmentation.

- Cross Entropy (CE) Loss: it was proposed by [26], and is a widely used pixel-wise distance to evaluate the performance of the classification or segmentation model. It is defined as $\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} [T_i \ln(P_i) + (1 - T_i) \ln(1 - P_i)]$. However, for unbalanced data, it typically results in unstable training results and leads to decision boundaries biased towards the majority classes. To deal with the imbalanced-data problem, two variants of the standard CE loss, Weighted CE (WCE) loss and Balanced CE (BCE) loss, are proposed to assign weights to the different classes.

- Dice loss: it was proposed by [22] and defined as $\mathcal{L}_{Dice} = 1 - 2 \frac{\sum_i^N T_i P_i}{\sum_i^N T_i + P_i} = 2 \frac{T \cap P}{T \cup P}$. Despite the Dice loss improvements over the CE loss, Dice loss may undergo difficulties when dealing with very small structures and weak object boundary, as misclassification of a few pixels can lead to a large decrease of the coefficient.

- Focal Loss: it was proposed by [20] to balance between easy and hard samples as $\mathcal{L}_{Focal} = -\frac{\alpha_i}{N} \sum_{i=1}^{N} \left((1 - P_i)^\gamma T_i \ln(P_i) + P_i^\gamma (1 - T_i) \ln(1 - P_i) \right)$. In Focal loss, the loss for confidently correctly classified labels is scaled down, so that the network focuses more on incorrect and low confidence labels than on increasing its confidence in the already correct labels.

- Boundary Loss: Recently, Kervadec, et al. [12] and Le et al. [15, 16] proposed boundary loss, offset curve loss (OsC) loss, and NB-AC loss to address both imbalanced data and weak boundary problems. For instance, OsC loss focuses on narrow band around the boundary. Boundary loss [12] makes use of a distance map in which the weights of voxels far from the boundary with greater than the nearer ones, thus the model learnt to reduce the error on the midst of the regions rather than the boundary. Despite its plausible intuition and competitive performance in practice, it does not help much in dealing with low contrast and weak boundary.

2.2.2 Attention Loss. Our attention loss is leveraged by [12, 15, 17] to pay more attention to hard-case examples or easily misclassified examples i.e. regions which are poorly segmented. In this section, we first show how to estimate hard-case examples and then present our proposed hard-case attention loss.

Hard-case examples estimation:
Infant brain MRI is shown in low contrast, weak boundary/surface; thus hard-case examples are more often presented at the surface regions. We also further observe that the model quickly learns and produces high accuracy on easy-case voxels (inside the object). Still, it is time-consuming and performs poorly on hard-case voxels at the surface. Therefore, it is reasonable to force the model to concentrate on the hard-case voxels while reducing the effect of easy-case voxels in updating the model's parameters. When it comes to Focal loss [20], the authors introduce a new term and integrate it into CE loss so that the model turns its attention from the subjects which are correctly classified with high confidence to the incorrect ones. The motivation of this term is to increase the gradient magnitude of the loss function of hard-case examples and decrease that of the easy-case ones so that the total gradient related to low confident and incorrect labels dominates the total gradient yielded by the correct ones. To make this idea more solid, we compare CE loss and Focal loss. To simply the equation, let consider one voxel at spatial coordinate $i = (x, y, z), x = 1, .., H, y = 1, .., W, z = 1, .., D$ with ground truth label T_i and predicted label P_i. The CE loss and Focal loss at i are as follows:

$$\mathcal{L}_{CE} = -\big[T_i \ln(P_i) + (1 - T_i) \ln(1 - P_i) \big].$$
$$\mathcal{L}_{Focal} = -\big[(1 - P_i)^\gamma T_i \ln(P_i) + P_i^\gamma (1 - T_i) \ln(1 - P_i) \big]. \tag{3}$$

Without any loss of generality, we consider the case where $T_i = 1$ and corresponding losses are simplified into:

$$\mathcal{L}_{CE} = -\ln(P_i).$$
$$\mathcal{L}_{Focal} = -(1 - P_i)^\gamma \ln(P_i). \tag{4}$$

The derivatives of divergence functions w.r.t P_i are as follows:

$$\frac{\partial \mathcal{L}_{CE}}{\partial P_i} = -\frac{1}{P_i}.$$
$$\frac{\partial \mathcal{L}_{Focal}}{\partial P_i} = -\frac{(1 - P_i)^\gamma}{P_i} + \gamma (1 - P_i)^{\gamma-1} \ln(P_i). \tag{5}$$

For $\gamma = 2$ which is the optimal value in [20], the absolute value of the gradient of $\mathcal{L}_{Focal}$ is greater than that of $\mathcal{L}_{CE}$ when P is smaller than approximately 0.298. Therefore, the model focuses on the hard examples whose predicted probability of belonging to true classes is smaller than 29.8%.

Hard-case attention loss: In this section, we first define a surface attention weight map W^c for each class $c \in C$. Let T^c denote as surface of class c. Each element W_i^c at spatial coordinate i in the surface attention weight map W^c is defined as:

$$\mathcal{W}_i^c = \frac{1}{min_{j \in T^c} d(i, j) + 1}. \tag{6}$$

where d is defined as $\mathcal{L}_2$ and $\mathcal{W}_i^c$ measures the distance between surface and voxels in the spatial domain. Unlike existing boundary loss functions, the voxels near the surface receive more weight than those far from the surface in our proposed weight map. Furthermore, to maintain the attention on the surface without introducing noise into the model, we incorporate Dice loss, which helps noise-free regions inside surface.

Figure 4 illustrates the low contrast, weak boundary problem in brain infant MRI segmentation. The most challenging is how to

Table 1: Comparison on iSeg-2017 dataset with Experiment Setting 1: Train on 9 subjects and test on Subject #9. The best is shown in bold and the second best is shown in underline.

Method	Year	DSC ↑			
		WM	GM	CSF	Average
3D-UNet [5]	2016	89.83	90.55	94.39	91.59
DenseVoxNet [11]	2017	85.46	88.51	91.26	89.24
VoxResNet [2]	2018	89.87	90.64	94.28	91.60
CC-3D-FCN [28]	2018	89.19	90.74	92.40	90.79
SegCaps [14]	2018	82.80	84.19	90.19	85.73
3D-SkipDenseSeg [1]	2019	<u>91.30</u>	91.61	94.74	92.55
SemiDenseNet [6]	2020	90.50	**92.05**	**95.80**	<u>92.77</u>
3D-UCaps [27]	2021	90.95	91.34	94.21	92.17
MSCD-UNet [21]	2021	90.47	92.17	<u>95.60</u>	92.74
Our proposed		**91.36**	<u>91.92</u>	95.06	**92.78**

Table 2: Comparison on iSeg-2017 dataset with Experiment Setting 2: Train on 10 subjects and test on 13 Subjects. The best is shown in bold and the second best is shown in underline.

Method	Year	DSC ↑				ASD (mm)↓			
		WM	GM	CSF	Average	WM	GM	CSF	Average
HyberDense [7]	2018	90.1	92.0	95.6	92.57	0.38	0.32	<u>0.12</u>	0.27
FC-DenseNet [9]	2019	90.7	92.6	**96.0**	93.1	0.36	<u>0.31</u>	**0.11**	<u>0.26</u>
D-SkipDenseSeg [1]	2019	90.3	92.2	95.7	92.73	0.38	0.32	<u>0.12</u>	0.27
H-DenseNet [29]	2019	90.0	92.0	**96.0**	92.67	0.36	<u>0.31</u>	**0.11**	<u>0.26</u>
FC-Semi-DenseNet1 [6]	2020	90.0	92.0	**96.0**	92.67	0.38	0.35	0.14	0.29
FC-Semi-DenseNet2 [6]	2020	90.0	92.0	**96.0**	92.67	0.41	0.34	<u>0.12</u>	0.29
V-3D-UNet [30]	2020	91.00	92.00	**96.00**	<u>93.00</u>	0.37	<u>0.31</u>	0.13	0.27
Non-local U-Net [38]	2020	91.03	<u>92.45</u>	95.30	92.29	0.40	0.37	0.14	0.30
HyperFusionNet [8]	2021	<u>90.20</u>	87.80	93.60	90.53	–	–	–	
APRNet[41]	2021	91.10	92.40	95.50	<u>93.00</u>	<u>0.35</u>	0.32	0.12	<u>0.26</u>
Our proposed		**92.60**	**93.49**	<u>95.68</u>	**93.92**	**0.28**	**0.25**	**0.11**	**0.21**

decide the voxels on the surfaces where they belong to the left class or right class. The weight maps (d, e, f) corresponding to CSF, GM, and WM surfaces show the significance of voxels. The voxels far from surfaces have less impact than those which are close to the surface.

Incorporate hard-case estimation into weight map, our attention loss is defined as follows:

$$\mathcal{L}_{attention} = \begin{cases} \sum_{i \in P, c \in C} W_i^c (P_i^c)^2 & \text{if } P_i \neq c \\ \sum_{i \in P, c \in C} W_i^c (1 - P_i^c)^2 & \text{otherwise} \end{cases} \quad (7)$$

The derivative of our $\mathcal{L}_{attention}$ w.r.t P_i corresponding to a particular class c is:

$$\frac{\partial f_{Attention}}{\partial P_i} = -2(1 - P_i). \quad (8)$$

In our attention loss function, the voxels which are wrongly classified (predicted probability of being in the same class as ground truth smaller than 50%) will be strongly attended. Furthermore, our proposed loss will emphasize the true class of each voxel as well as strengthen the confidence of the model prediction without having too much negative effect on rare classes.

3 EXPERIMENTAL RESULTS

3.1 Data and metrics

Dataset: The iSeg17 dataset [36] consists of 10 subjects with ground-truth labels for training and 13 subjects without ground-truth labels for testing. Each subject includes T1 and T2 images with a size of $144 \times 192 \times 256$, and an image resolution of $1 \times 1 \times 1$ mm^3. In iSeg, there are three classes: white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF).

Metrics: For quantitative assessment of the segmentation, the proposed model is evaluated on different metrics, e.g. Dice score (DSC), and average surface distance (ASD).

The DSC measure is defined as:

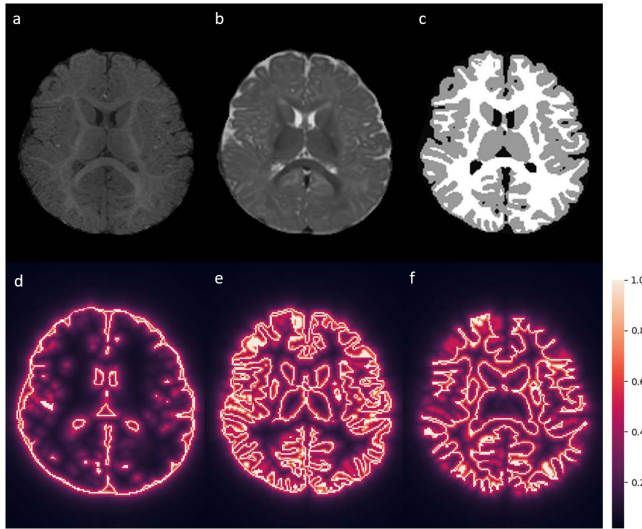$$DSC \frac{2|T \cap P|}{|T| + |P|}. \quad (9)$$

Figure 4: (a) and (b) are T1- and T2-weighted brain MRI scans of the subject 9 in iSeg2017 dataset. (c) is its segmentation ground truth, and (d), (e), and (f) are the corresponding surface attention weight map for cerebrospinal fluid (CSF), gray matter (GM) and white matter (WM).

where T and P are corresponding to groundtruth and predicted segmentation result. A higher value of DSC means better segmentation accuracy.

The ASD is utilized to measure the segmentation boundary distance and defined as:

$$ASD(T,P) = \frac{1}{2}\left(\frac{\sum_{V_i \in S_P} min_{V_j \in S_T} d(V_i, V_j)}{\sum_{V_i \in S_P} 1} + \frac{\sum_{V_j \in S_T} min_{V_i \in S_G} d(V_i, V_j)}{\sum_{V_j \in S_T} 1}\right). \quad (10)$$

where S_T and S_P are the surface of the ground truth and predicted segmentation result. $d(V_i, V_j)$ is the Euclidean distance from a vertex V_i and V_j. A smaller ASD is the better the result.

3.2 Experiment Setting

The input is defined as $N \times C \times H \times W \times D$, where N is the batch size, C is the number of input modalities, and H, W, D are height, width, and depth of the volume patch in the sagittal, coronal, and axial planes. We choose the input as $8 \times 2 \times 32 \times 32 \times 32$. We implemented our network using PyTorch 1.3.0, and our model is trained until convergence by using the SGD optimizer accompanied by the warm restarts technique. More precisely, the training phase consists of 50 200-epoch periods, in which the learning rate is set to 0.01 and reduces by ten times every 40 epochs. As aforementioned, medical images are widely considered as difficult subjects for CNN models to learn meaningful and useful characteristics. In other words, their special domain is far from that of natural images on which common deep learning models and modules are designed. Intuitively, in the training phase, the larger the searched space is, the higher probability the model learns special features, which help to solve the problem efficiently. By repeatedly restarting the learning rate schedule, the model is encouraged to explore a larger space every time the learning rate is set to maximum. After going to a new

area in the search space, the learning rate rapidly drops to find the local optimal parameters in this area. Our DAM-AL makes use of instance normalization [34] and Leaky reLU. The experiments are conducted using an Intel CPU and RTX GPU with two settings.

- Experiment Setting 1: We follow 3D-SkipDenseSeg [1] to have the training set of 9 subjects and testing set of subject #9.
- Experiment Setting 2: We train on 10 annotated subjects and test on 13 unlabeled subjects (Subject #11 - Subject #23).

3.3 Performance and Comparison

The evaluations on iSeg-2017 are given in Table1 and Table2 corresponding to two experiment settings. Regarding Experiment Setting 1, our DAM-AL obtains the best performance on WM and the second-best on GM and CSF compared with the existing state-of-the-art approaches. Overall, our DAM-AL outperforms other methods on DSC in Experiment Setting 1. Fig.5 visualizes the performance of DAM-AL on Subject #9 in the coronal, axial, and sagittal planes. On Subject #9, we randomly crop some patches and compare the performance between the predicted segmentation and ground truth at an enlarged view. In most of cases, the hard-case examples on surface are predicted correctly. Regarding Experiment Setting 2, which was evaluated on testing set, our DAM-AL obtains the best score on both DSC and ASD metrics with considerable gaps compared to the second-best performance, i.e. average DSC gains 0.92% compared with the second-best [30] and average ASD reduces 0.05mm compared with the second-best [9, 29, 41]. Performance on DSC and ASD of individual subject is provided in Fig. 6. Fig.7 visualizes the performance of DAM-AL on Subject #11.

Conclusion

In this work, we introduced DAM-AL, a dilated attention mechanism with attention loss on hard-case examples for medical image segmentation. Our proposed DAL-AL contains spatial attention at high-level feature and channel-wise attention at low-level feature. We tested our framework on the problem of infant brain segmentation and showed that our DAM-AL is effective, robust, and more accurate than existing segmentation methods.

Therefore, exploring spatial and channel-wise attention together with hard-case attention loss is a promising approach to medical image analysis. Future investigations might include other medical image datasets on different modalities, e.g., MRBrainS, Brats, Pancreas, Hippocampus, etc.
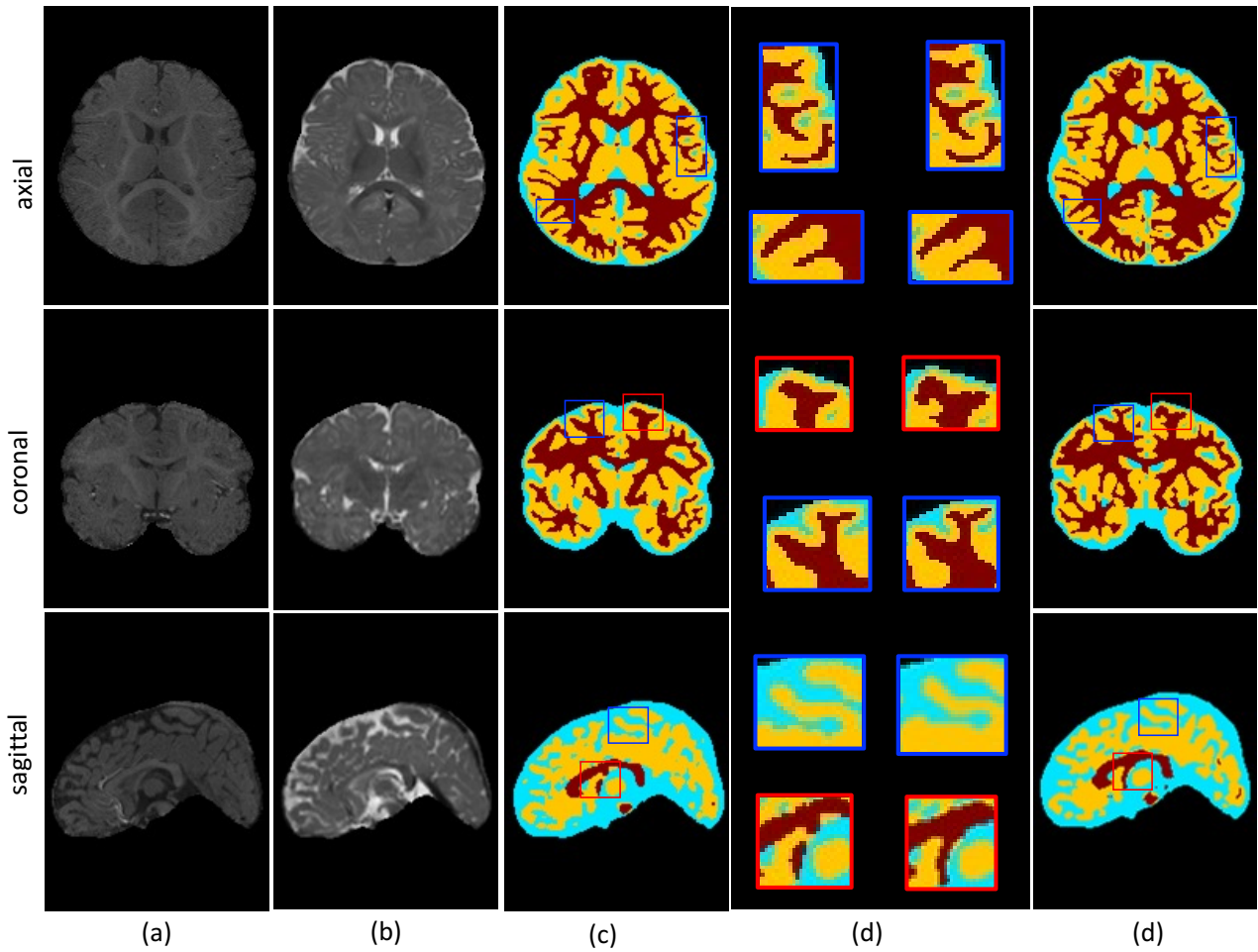
Figure 5: Result of Experiment Setting 1. (a) and (b) are T1- and T2-weighted brain MRI scans of the subject #9 in different views. (c): predicted segmentation results by DAM-AL; (d) Enlarged view of some random regions between (c) and (e). Blue boxes indicate some spots produced correctly by DAM-AL. Some regions where DAL-AL yielded incorrect segmentations are outlined in red ; (e): ground truth. Top-down: visualize in different planes: axial, coronal, and sagittal.
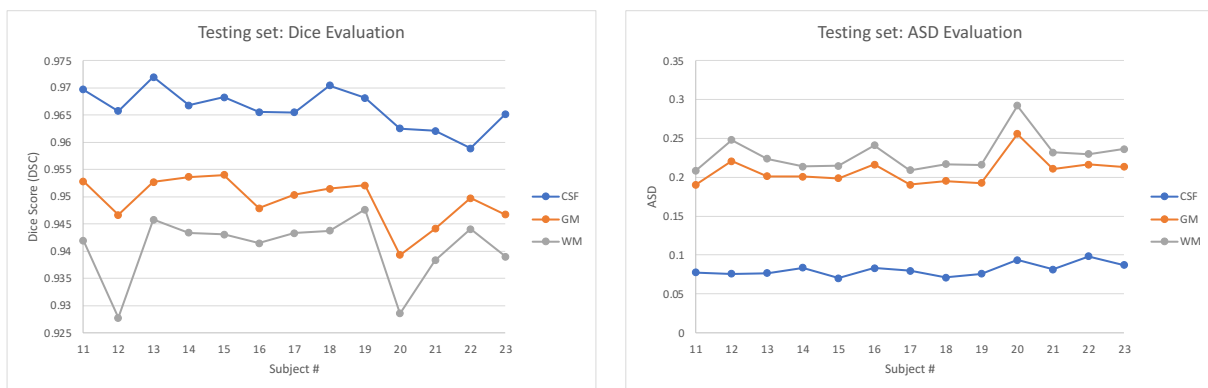

Figure 6: Performance of the proposed DAM-AL on the subjects of iSeg-2017 datasets. left: DSC, right: ASD.
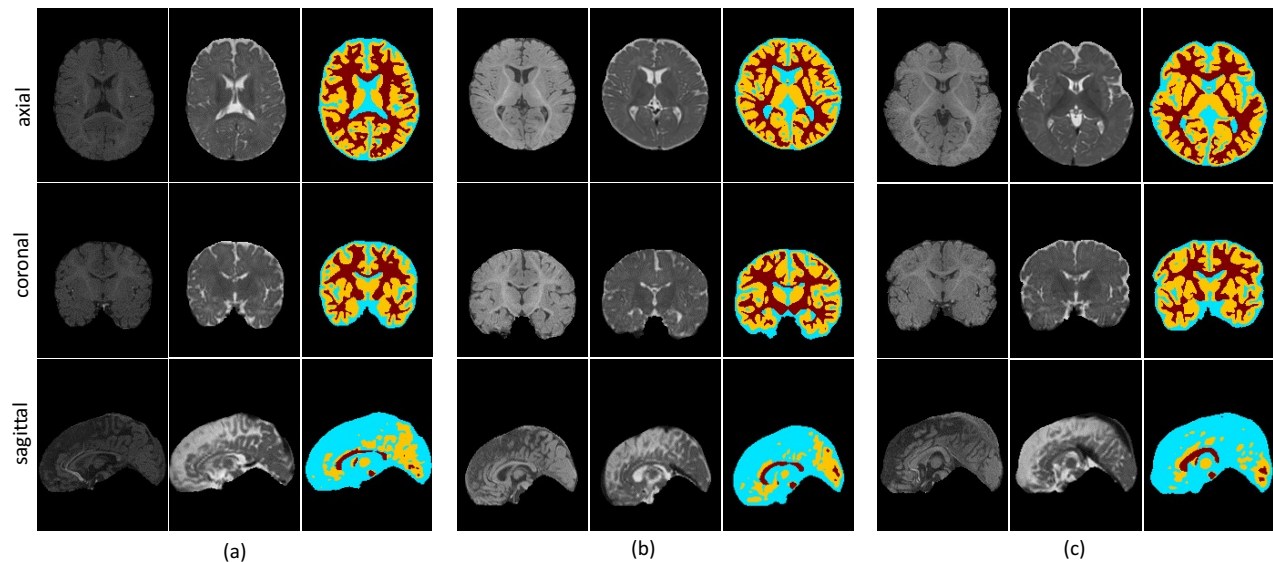
(a) (b) (c)

Figure 7: Result of Experiment Setting 2 with the first three subjects, i.e. #11 (a) , #12 (b) , #13 (c). Top-down: visualize in different planes: axial, coronal, and sagittal. For each subject, from left-right: T1-brain MRI, T2-brain MRI, predicted segmentation results conducted by DAM-AL.

REFERENCES

[1] Toan Duc Bui, Jitae Shin, and Taesup Moon. 2019. Skip-connected 3D DenseNet for volumetric infant brain MRI segmentation. *Biomedical Signal Processing and Control* 54 (2019), 101613.

[2] Hao Chen, Qi Dou, Lequan Yu, Jing Qin, and Pheng-Ann Heng. 2018. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage* 170 (2018), 446–455.

[3] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. Chua. 2017. SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6298–6306.

[4] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5659–5667.

[5] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*. Springer, 424–432.

[6] Jose Dolz, Christian Desrosiers, Li Wang, Jing Yuan, Dinggang Shen, and Ismail Ben Ayed. 2020. Deep CNN ensembles and suggestive annotations for infant brain MRI segmentation. *Computerized Medical Imaging and Graphics* 79 (2020), 101660.

[7] Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. 2018. HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation. *IEEE transactions on medical imaging* 38, 5 (2018), 1116–1126.

[8] Wenting Duan, Lei Zhang, Jordan Colman, Giosue Gulli, and Xujiong Ye. 2021. Multi-modal Brain Segmentation Using Hyper-Fused Convolutional Neural Network. In *International Workshop on Machine Learning in Clinical Neuroimaging*. Springer, 82–91.

[9] Seyed Raein Hashemi, Sanjay P Prabhu, Simon K Warfield, and Ali Gholipour. 2019. Exclusive independent probability estimation using deep 3D fully convolutional DenseNets: Application to IsoIntense infant brain MRI segmentation. In *International Conference on Medical Imaging with Deep Learning*. PMLR, 260–272.

[10] Heather Cody Hazlett, Michele D Poe, Guido Gerig, Martin Styner, Chad Chappell, Rachel Gimpel Smith, Clement Vachet, and Joseph Piven. 2011. Early brain overgrowth in autism associated with an increase in cortical surface area before age 2 years. *Archives of general psychiatry* 68, 5 (2011), 467–476.

[11] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. 2017. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 11–19.

[12] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. 2021. Boundary loss for highly unbalanced segmentation. *Medical Image Analysis* 67 (Jan 2021), 101851. https://doi.org/10.1016/j.media.2020.101851

[13] Rebecca C Knickmeyer, Sylvain Gouttard, Chaeryon Kang, Dianne Evans, Kathy Wilber, J Keith Smith, Robert M Hamer, Weili Lin, Guido Gerig, and John H Gilmore. 2008. A structural MRI study of human brain development from birth to 2 years. *Journal of neuroscience* 28, 47 (2008), 12176–12182.

[14] Rodney LaLonde and Ulas Bagci. 2018. Capsules for object segmentation. *arXiv preprint arXiv:1804.04241* (2018).

[15] Ngan Le, Toan Bui, Viet-Khoa Vo-Ho, Kashu Yamazaki, and Khoa Luu. 2021. Narrow Band Active Contour Attention Model for Medical Segmentation. *Diagnostics* 11, 8 (2021), 1393.

[16] Ngan Le, Trung Le, Kashu Yamazaki, Toan Bui, Khoa Luu, and Marios Savides. 2021. Offset Curves Loss for Imbalanced Problem in Medical Segmentation. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 9189–9195.

[17] Ngan Le, Trung Le, Kashu Yamazaki, Toan Duc Bui, Khoa Luu, and Marios Savides. 2020. Offset Curves Loss for Imbalanced Problem in Medical Segmentation. arXiv:eess.IV/2012.02463

[18] Ngan Le, Kashu Yamazaki, Kha Gia Quach, Dat Truong, and Marios Savvides. 2021. A Multi-task Contextual Atrous Residual Network for Brain Tumor Detection & Segmentation. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 5943–5950.

[19] T Hoang Ngan Le, Raajitha Gummadi, and Marios Savvides. 2018. Deep recurrent level set for segmenting brain tumors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 646–653.

[20] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. 2017. Focal loss for dense object detection. In ICCV 2017. *ICCV*, 2980–2988.

[21] Jiao-Song Long, Guang-Zhi Ma, En-Min Song, and Ren-Chao Jin. 2021. Learning U-Net Based Multi-Scale Features in Encoding-Decoding for MR Image Brain Tissue Segmentation. *Sensors* 21, 9 (2021), 3232.

[22] M.Fausto, N.Nassir, and A.Seyed-Ahmad. 2016. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *the Fourth International Conference on 3D Vision*. 565–571.

[23] Mamta Mittal, Lalit Mohan Goyal, Sumit Kaur, Iqbaldeep Kaur, Amit Verma, and D Jude Hemanth. 2019. Deep learning based enhanced tumor segmentation approach for MR brain images. *Applied Soft Computing* 78 (2019), 346–354.

[24] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. In *Advances in neural information processing systems*. 2204–2212.

[25] Pim Moeskops, Max A Viergever, Adriënne M Mendrik, Linda S De Vries, Manon JNL Benders, and Ivana Išgum. 2016. Automatic segmentation of MR brain images with a convolutional neural network. *IEEE transactions on medical imaging* 35, 5 (2016), 1252–1261.

[26] Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

[27] Tan Nguyen, Binh-Son Hua, and Ngan Le. 2021. 3D-UCaps: 3D Capsules Unet for Volumetric Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 548–558.

[28] Dong Nie, Li Wang, Ehsan Adeli, Cuijin Lao, Weili Lin, and Dinggang Shen. 2018. 3-D fully convolutional networks for multimodal isointense infant brain image segmentation. *IEEE transactions on cybernetics* 49, 3 (2018), 1123–1136.

[29] Saqib Qamar, Hai Jin, Ran Zheng, and Parvez Ahmad. 2019. Multi stream 3D hyper-densely connected network for multi modality isointense infant brain MRI segmentation. *Multimedia Tools and Applications* 78, 18 (2019), 25807–25828.

[30] Saqib Qamar, Hai Jin, Ran Zheng, Parvez Ahmad, and Mohd Usama. 2020. A variant form of 3D-UNet for infant brain segmentation. *Future Generation Computer Systems* 108 (2020), 613–623.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv e-prints* (May 2015), arXiv:1505.04597.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv e-prints* (May 2015), arXiv:1505.04597.

[33] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. *arXiv preprint arXiv:1710.09829* (2017).

[34] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. 2016. Instance Normalization: The Missing Ingredient for Fast Stylization. *CoRR* abs/1607.08022 (2016). http://arxiv.org/abs/1607.08022

[35] Viet-Khoa Vo-Ho, Ngan Le, Kashu Kamazaki, Akihiro Sugimoto, and Minh-Triet Tran. 2021. Agent-Environment Network for Temporal Action Proposal Generation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2160–2164.

[36] Li Wang, Dong Nie, Guannan Li, Élodie Puybareau, Jose Dolz, Qian Zhang, Fan Wang, Jing Xia, Zhengwang Wu, Jiawei Chen, et al. 2019. Benchmark on Automatic 6-month-old Infant Brain Segmentation Algorithms: The iSeg-2017 Challenge. *IEEE TMI* (2019).

[37] Siying Wang, Christian Ledig, Joseph V Hajnal, Serena J Counsell, Julia A Schnabel, and Maria Deprez. 2019. Quantitative assessment of myelination patterns in preterm neonates using T2-weighted MRI. *Scientific reports* 9, 1 (2019), 1–12.

[38] Zhengyang Wang, Na Zou, Dinggang Shen, and Shuiwang Ji. 2020. Non-local U-Nets for biomedical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6315–6322.

[39] Kashu Yamazaki, Vidhiwar Singh Rathour, and T Le. 2021. Invertible Residual Network with Regularization for Effective Medical Image Segmentation. *arXiv preprint arXiv:2103.09042* (2021).

[40] Wenlu Zhang, Rongjian Li, Houtao Deng, Li Wang, Weili Lin, Shuiwang Ji, and Dinggang Shen. 2015. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage* 108 (2015), 214–224.

[41] Yuzhou Zhuang, Hong Liu, Enmin Song, Guangzhi Ma, Xiangyang Xu, and Chih-Cheng Hung. 2021. APRNet: A 3D Anisotropic Pyramidal Reversible Network with Multi-modal Cross-Dimension Attention for Brain Tissue Segmentation in MR Images. *IEEE Journal of Biomedical and Health Informatics* (2021).