Two Independent Teachers are Better Role Model

Afifa Kahled
School of Computer Science and Technology
Huazhong University of Science and Technology
Wuhan 430074, China
`afifakhaied@tju.edu.cn`

Ahmed A. Mubarak
College of Applied Sciences and Educational
IBB University
IBB, Yemen
`ahmedmubarak@ibbuniv.edu.ye`

Kun He*
School of Computer Science and Technology
Huazhong University of Science and Technology
Wuhan 430074, China
`brooklet60@hust.edu.cn`

June 9, 2023

Abstract

Recent deep learning models have attracted substantial attention in infant brain analysis. These models have performed state-of-the-art performance, such as semi-supervised techniques (e.g., Temporal Ensembling, mean teacher). However, these models depend on an encoder-decoder structure with stacked local operators to gather long-range information, and the local operators limit the efficiency and effectiveness. Besides, the MRI data contain different tissue properties (TPs) such as $T1$ and $T2$. One major limitation of these models is that they use both data as inputs to the segment process, i.e., the models are trained on the dataset once, and it requires much computational and memory requirements during inference. In this work, we address the above limitations by designing a new deep-learning model, called 3D-DenseUNet, which works as adaptable global aggregation blocks in down-sampling to solve the issue of spatial information loss. The self-attention module connects the down-sampling blocks to up-sampling blocks, and integrates the feature maps in three dimensions of spatial and channel, effectively improving the representation potential and discriminating ability of the model. Additionally, we propose a new method called Two Independent Teachers ($2IT$), that summarizes the model weights instead of label predictions. Each teacher model is trained on different types of brain data, $T1$ and $T2$, respectively. Then, a fuse model is added to improve test accuracy and enable training with fewer parameters and labels compared to the Temporal Ensembling method without modifying the network architecture. Empirical results demonstrate the effectiveness of the proposed method.

1 Introduction

Deep learning is on the horizon with the promise to enhance the segmentation of medical images due to the tedious and expensive annotating images [1, 2]. Numerous studies have been conducted on various deep learning models that have accomplished a wide variety of tasks [3, 4], including brain segmentation, tumor segmentation, etc. Many of these tasks

---

*Afifa Kahled and Kun He are with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China (E-mail: afifakhaied@tju.edu.cn; brooklet60@hust.edu.cn). Corresponding author: Kun He. Ahmed A. Mubarak is with the College of Applied Sciences and Educational, IBB University, IBB, Yemen (E-mail: ahmedmubarak@ibbuniv.edu.ye).

have stringent quality of result requirements. Consequently, advanced deep learning models have wide applications in brain segmentation.

In line with that, semi-supervised techniques are undergoing a paradigm shift towards adding auxiliary tasks to improve the segmentation result and to alleviate problems in brain segmentation, particularly, formed from the U-net structure, as an effective alternative to other deep learning models, as U-net models can learn both global information and local information. Though effective, we observe some limitations on recent U-net-based methods.

First, most of them perform segmentation on both tissue properties, called $T1$ and $T2$, based on the fact that merging $T1$ and $T2$ images together could capture more tissue information and also make segmentation an easy process. Since the differences in image contrast in MRI types, the $T1$-weighted MRI enhances the fatty tissue's signal and suppresses the water's signal. In contrast, $T2$-weighted MRI improves the signal of the water. In addition, the volumes of $3D$ medical images are normally too large to fit on graphics processing unit (GPU) memory at their native resolution, and the number of model parameters will be large, requiring enormous computation during training. Accordingly, analysis of the information provided by these modalities induces a loss of local information conducting to extreme and redundant low-level features.

Second, by adopting convolutions and down-sampling operators, we could get local operators, scan inputs and extract local information by applying small kernels. However, the cascade of stacking local information results in large effective kernels. The encoder generally stacks size-preserving convolutional layers, interlaced with downsampling operators, to progressively decrease the spatial sizes of feature maps. In addition, most of the models have an encoder with many stacked local operators but it suffers from a large amount of training parameters. This is because the number of feature maps generally grow doubly or even exponentially after every down-sampling operation. A very crucial issue in medical image segmentation is the loss of spatial information during encoding and this happens with more down-sampling, particularly when dealing with both $T1$ and $T2$ as one input.

Besides, several major challenges emerge while segmenting brain data into one or more tissues. Brain datasets are divided into $T1$ and $T2$ [5] that focus on different textures. Most existing models employ a one-stage method to segment medical data and predict the class probability and position information on both data, i.e., $T1$ and $T2$. Moreover, many existing popular technologies are expected to play a pivotal role in the semi-supervised learning methodology on medical images. Temporal ensembling [6] aggregates the predictions of multiple network evaluations into an ensemble prediction, and the mean teacher method [7] takes a supervised architecture and makes a copy of the model. In the medical image community, the mean teacher model is good for getting better predictions and has the ability to optimally exploit unlabeled data during training [7, 8].

This paper addresses the above limitations and challenges by proposing a 3D-DenseUNet model based on the U-net model, which focuses on loosening information loss and can leverage overall relationships between structures using multi-head attention. Moreover, two independent teachers are employed for deep supervision on each type of brain data. The first teacher model focuses on $T1$ and another on $T2$. Then, a fused model is adopted that takes advantage of the weights of two teacher models, leading to a decreasing number of learning parameters in the depth of network and better performance. The performance of the proposed model is compared with the state-of-the-art models evaluated by the Dice Coefficient metric [9], and empirical results demonstrate that our proposed model achieves competitive performance with fewer parameters.

Our main contributions in terms of designing the model architecture are as follows:

- We propose a 3D-DenseUNet model based on the U-net structure to form the down-sampling and upsampling modules, working with multi-head attention as a self-attention, which not only can effectively encode the wide context information into local features but also can mine the interdependency between channel maps and improve the feature representation of specific semantics. In addition, the skip-connections between modules to sum the multiple contextual information instead of concatenation are employed in the U-net model to further boost the performance.

- We construct Two Independent Teacher models (2IT), with each model operating on a different data type to learn the deep feature information of each modality to reduce the information uncertainty, boundary and extract weights.

- A Fuse model is built with weights updated from the two teacher modules, which could help to deeply analyze both brain data ($T1$ and $T2$), meanwhile avoid computational and memory requirements during inference and time training.

- To avoid oscillation throughout the training stage of the fusion model as well as the overfitting and noising issues resulting from the summation of the teacher models' weights, we propose a function to calculate the fusing coefficient α value based on metric values of the model and some parameters' current epochs.

2

The rest of this paper is organized as follows. Section 2 presents the related work. Section 3 presents the detailed design of our proposed model. Section 4 presents our results and discussion. Finally, Section 5 concludes the paper and discusses directions for future work.

2   Related Work

Deep learning has become an integral part for brain segmentation. In the realm of medical imaging, segmented images consume enormous time. It is estimated to take 9-11 hours to segment 5-20 brain images [10]. Many proposed models based on U-net [11] have achieved state-of-the-art performances in brain segmentation, which adopt a serial encoding–decoding structure and use hybrid dilated convolution (HDC) modules. Their results are in line with a recent study in brain segmentation, which investigates concatenation between each module of two serial networks [11].

In addition, considerable efforts have been devoted to utilize unlabeled data in medical applications using semi-supervised techniques. For example, Zhang et al. [12] propose a deep adversarial network (DAN) model that aims to attain consistently good segmentation results. Besides, another semi-supervised method [13] explores the learning equivalence to elastic deformations, and finds that segmentation performance improves when all training data is labeled and unlabeled data is added for training. Wang et al. [14] propose a multi-task semi-supervised segmentation model that includes a segmentation task and a regression task. A new idea is developed in [15], where only a few number of training data is annotated. Their proposed method depends on forming a consensus prediction.

Regarding the teacher model, there are many proposed models, which try to take advantage of historical information to improve the segmentation accuracy. For example, Wang et al. [16] propose two auxiliary tasks, and use multi-task learning to improve performance in medical segmentation. Moreover, they propose to use uncertainty constraint to improve the feature representation. In addition, Cui et al. [17] propose a semi-supervised method by an adapted mean teacher model, and empirical results demonstrate the effectiveness of their proposed teacher and student model.

In contrast to previous works, we aim to focus on solving the problem of missing spatial information. To this end, we propose a 3D-DenseUNet model that acts as two teacher modules to extract the weights and a fusing module that integrates the weights of two independent teacher modules instead of label predictors. Each teacher module is trained on different types of brain data ($T1$ and $T2$ respectively). Then, a fuse module takes advantage of the two modules, which improves the test accuracy and enables training with fewer labels than Temporal Ensembling [16] without modifying the network structure.

3   Methodology

The U-net architecture developed in 2015 [18] comprises a down-sampling as an encoder and an up-sampling as a decoder, with skip connections between them, used to segment biomedical images. The U-net model is built to integrate local and global contextual information through the encoding-decoding stages. Other recently proposed U-net models are based on an encoder-decoder structure with stacked local operators to gather long-range information. However, there are still two limitations in recent U-net models when dealing with brain segmentation tasks. First, spatial information loss is caused by down-sampling in the encoding stage. Second, the network cannot accurately locate the category of each voxel due to the loss of semantically related output and input positions caused by the network decoding phase.

In line with that, we propose a 3D-DenseUNet model for segmenting 3D medical images. The model structure comprises a down-sampling as an encoder, which contains three blocks to build a fully residual network in each block; and an up-sampling as a decoder, which also contains three blocks to build a fully residual network in each block, with skip connections between them as long-term residual connections. In addition, a global attention block is proposed to fully use the multi-scale context features. Moreover, our framework comprises two independent teachers' models and a fused model, each using the 3D-DenseUNet architecture. The details of each model are presented in the following subsections. Figure. 1 depicts the framework of the 3D-DenseUNet model with the global attention block employed in the proposed model.

3.1   The Framework of 3D-DenseUNet Model

3.1.1   Down-Sampling

In the proposed 3D-DenseUNet, the down-sampling module contains three blocks to build a fully residual network in each block. The residual network has three different convolutions with batch normalization and ReLU6 activation function employed before each convolutional layer. Specifically, the first block is employed as the input layer with $3 \times 3 \times 3$ convolution and a stride of 1, and the output block is formed all the same followed by a $1 \times 1 \times 1$ convolution
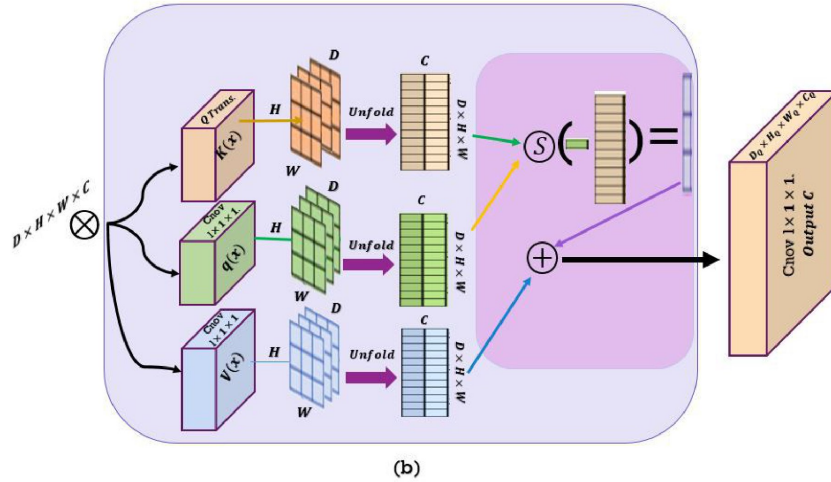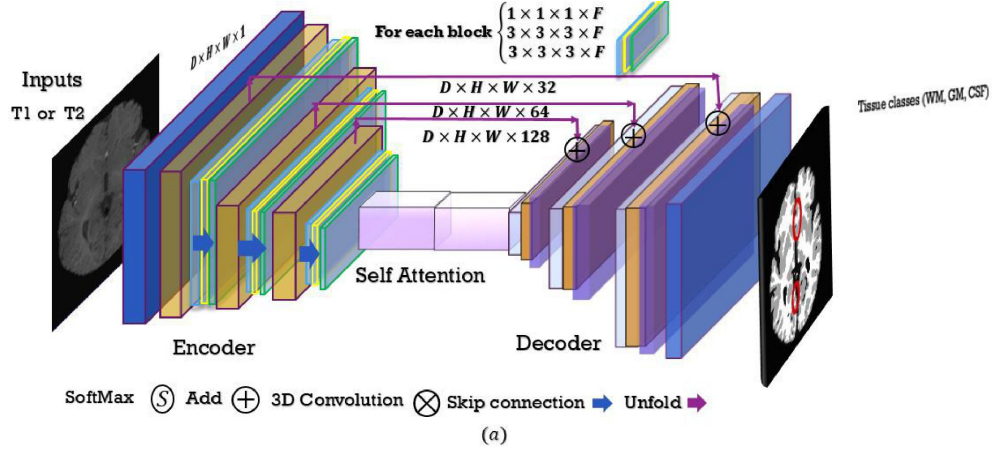
Figure 1: Framework of the proposed model with (a) 3D-DenseUNet architecture and (b) a global attention block with 3D layers based on multi-head attention as the self-attention.

with a stride of 1. To solve the first issue mentioned above, the features are extracted from multiple scales using the first residual block after the sum of skip connections to avoid loss of spatial information.

Here we use the summation instead of concatenation used in the U-net standard model for the following reasons. First, the summation works on the limit from increased feature maps, which contributes to reducing trainable parameters in the next layer. Second, the summation with skip connections could be regarded as long-term residual connections, which reduces the time of model training. The three different convolutions are used to build three-level features, which are combined to form the final feature for each block in down-sampling. The three different features allow down-sampling to be performed which preserves different features of the next level. An end block then collects global information, resulting in the encoder output.

3.1.2 Multi-head Self-attention Mechanism

To fully use the multi-scale context features, we collect the global information of the low-level layers with the high-level feature of deep layers for final feature representations in the end block encoder. Nevertheless, a one-step sequence correlation may bypass some valuable points in the large-scale upsampling process, in which a lot of information is lost. Therefore, we need to joint feature maps of output of each apiece by attending to every position of the input semantically. We propose the attention mechanism based on the multi-head self-attention function from Transformer [19] as shown in Figure. 1 (b). The attention mechanism is adopted after the encoder stage to concatenate global information from feature maps of any size and helps to adaptively integrate information on local features and edges of the image. This block can handle the end block of down-sampling and up-sampling blocks as a block.

In general, we have an input of 3D medical image $X^{h \times w \times d \times c}$ where h, w, and d are the image shapes and c the number of image channels. The end block of the down-sampling stage takes the latent cube representation X as input and then computes by multi-head attention with the $1 \times 1 \times 1$ convolution and a stride of 1. First, the attention mechanism operates to produce the query (Q), key (K) and value (V) matrices by Eq. 1:

$$
\begin{aligned}
Q &= Q\_Transform_{C_K}(X), \\
K &= Conv_{C_K}(X), \\
V &= Conv_{C_V}(X),
\end{aligned}
\tag{1}
$$

where operation $Q\_Transform_{C_K}(\cdot)$ produces C_K feature maps, and C_K, C_V are hyper-parameters representing the dimensions of keys and values.

These matrices are decomposed in multiple heads in the second stage by parallel and independent computations. This allows a simpler stacking of multiple transformer blocks as well as identity skip connections. Therefore, the $Unfold(\cdot)$ is employed to unfold the $D \times H \times W \times C$ tensor into $(D \times H \times W) \times C$ matrix, then the V and K dimensions are $(H \times W \times D) \times C_V$ and $(H \times W \times D) \times C_K$, respectively.

Next, a scaled dot-product operation with SoftMax normalization between Q and the transposed version of K is conducted to generate the matrix of contextual attention map A with dimension $(D_Q \times H_Q \times W_Q) \times (D \times H \times W)$, which defines the similarities of given features from Q concerning global elements of K. To calculate the aggregation of values weighted by attention weights, A could be multiplied by V, which produces the output matrix O as the following:

$$
O = Softmax\left(\frac{QK^T}{\sqrt{c_K}}\right)V,
\tag{2}
$$

where $\sqrt{c_K}$ is the dimension of query Q and K the key-value sequence. Finally, dropout is used to avoid over-fitting and Y reshapes the optimized feature maps to obtain the final output:

$$
Y = Conv_{CO}(Fold(O)),
\tag{3}
$$

where $Fold(\cdot)$ is the reverse operation of $Unfold(\cdot)$ and CO represents the output dimension. So, the size of output Y is $D_Q \times H_Q \times W_Q \times C_O$.

At the end of an up-sampling residual block, a $3 \times 3 \times 3$ deconvolution with a stride of 2 is employed instead of the identity residual connection, which works on taking global information into consideration, and the up-sampling block will recover more accurate details.

3.1.3 Up-Sampling

The up-sampling module of 3D-DenseUNet contains three blocks to build a fully residual network in each block. The residual network has three different convolutions with batch normalization and ReLU6 activation function employed before each convolutional layer. The global attention block is employed as the input layer in the up-sampling module, a $1 \times 1 \times 1$ convolution with a stride of 2. Then, the output block is formed within $3 \times 3 \times 3$ convolution with a stride of 1. To solve the first issue mentioned above, the features are extracted from multiple scales using the first residual block after the sum of skip connections to avoid the loss of spatial information. In a similar manner to the down-sampling phase, we use the summation instead of the concatenation for the limit from increased feature maps to reduce trainable parameters in the next layer. Then, we conduct a summation of skip connections as long-term residual connections, which reduces the time of training of models. In our model for up-sampling, the $Q\_Transform_C(\cdot)$ in the global attention block is formed by a $3 \times 3 \times 3$ deconvolution with a stride of 2 to avoid the information loss problem and keep more accurate details. Finally, the output of the decoder block is the segmentation probability map for producing probabilities for each segmentation class using a single $\arg\max$ operation.

3.2 Two-Independent-Teacher Model

In the field of computer vision, large datasets require further analysis and extraction of features by employing deep learning models to achieve desired performance. Therefore, the models may need more training time and hyperparameters with increasing network depth, which may simply lead to degradation and is prone to over-fitting. Also, the resolution and image size of input affects the prediction accuracy, particularly for images in magnetic resonance MRI types that share both $T1$ and $T2$ data. The $T1$-weighted images are produced by using short Time to Echo (TE) and Repetition Time (TR). The contrast and brightness of the images are predominately determined by $T1$ properties of tissue. Conversely, the $T2$-weighted images are produced using longer TE and TR times. In these images, the contrast and brightness are predominately determined by the $T2$ properties of tissue. Moreover, $T1$- and $T2$-weighted
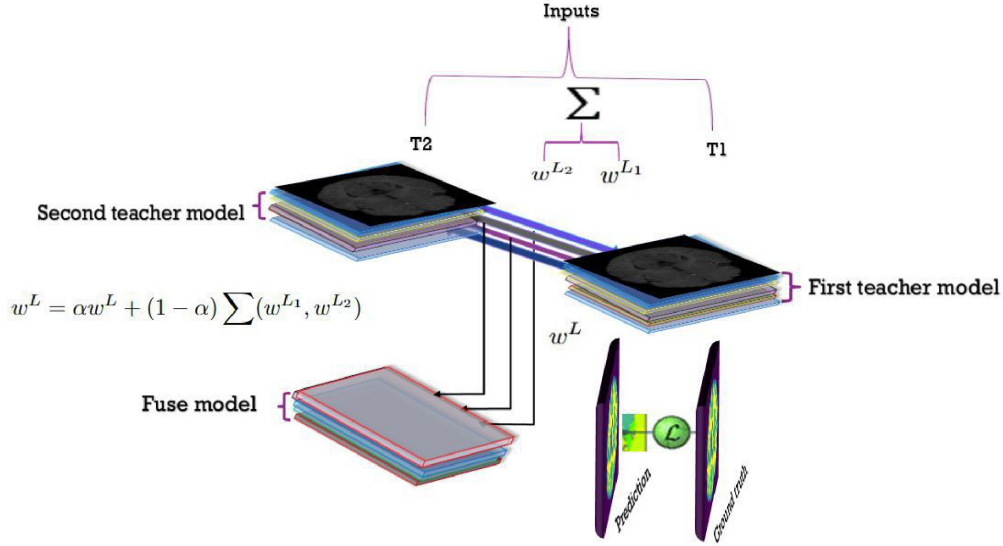
Figure 2: Illustration of the fused weights from teacher models to the fuse model.

images can be easily differentiated by looking the cerebrospinal fluid (CSF). CSF is dark on $T1$-weighted imaging and bright on $T2$-weighted imaging. Besides, $T1$-weighted MRI enhances the signal of the fatty tissue and suppresses the signal of the water, while $T2$-weighted MRI enhances the signal of the water.

In order to more precisely distinguish cerebral structures in neuroimaging studies and enhance tissue contrast, it may be the case that training the model with individual $T1$ or $T2$ can be an optimal way than combining $T1$ and $T2$ modalities. However, most semi-supervised learning models consider all information these modalities provide for MRI image analysis and diagnosis. Since the segmentation cost is undefined for unlabeled examples, the noise regularization by itself does not assist in semi-supervised learning [20]. Some recent studies use noise, and then use a consistency cost between the two predictions. These studies assume a dual role as a student (it learns as before), and a teacher (it generates targets).

There exist at least two methods to enhance the target results. One is to select the perturbation of the representations carefully instead of barely applying additive or multiplicative noise [21]. The other is to select the teacher model carefully instead of ensembling models [7]. To the best of our understanding, these two approaches are consistent, and their combination may yield better results.

In line with that, We take the second approach and show that it provides significant benefits. Our objective is to build two models, denoted as $TM1$ and $TM2$, as independent teachers. To adopt every model to work on different data ($TM1$ on $T1$-Weight and $TM2$ on $T2$-Weight), several changes must be made. At the first training step for each model, for the initial weights of the model, we use the strategy proposed in [22] which results in a rapid convergence of the model structure based on specific data characteristics.

3.3 The Fuse Model

Based on the structure of the teacher model proposed by [7], which used averages consecutive student models instead of predictions without additional training, and also to overwhelm the limitations of Temporal Ensembling [15]. The authors used an exponential moving average (EMA) to calculate the averages of weights of the student model rather than sharing the weights. Their method is resulting better test accuracy and the ability to handle massive datasets.

In keeping with that idea, we propose a Fuse model that benefits from the output of teacher models that are weights of model layers over the training steps. We can use this during training to construct a more accurate model instead of using the final weights directly according to Eq. 4. Therefore the fuse model updates weights by fusing the weight from the two independent teacher models per iteration as shown in Figure. 2. Besides, through the practice of experiment, we notice that there are two advantages when using the group of independent teachers. First, summation does increase the feature maps of $T1$ and $T2$. Second, the summation of the current weight can be considered as more information, which is known to be capable of increasing the performance of the proposed model [23]. Since the labeled data is usually very

difficult to acquire, this merging weights from more than one model can open the door for a new method to improve the performance of segmentation without increasing the labeled data. Moreover, the main idea of our proposed model is to update each parameter's weight of the fuse model by computing the summation of weights of the same parameter in independent teacher models plus the parameter's weight multiply α to reduce noise and to avoid over-fitting, as shown in Eq. 4.

$$W^L = \alpha W^L + (1 - \alpha) \sum (W^{L_1}, W^{L_2}), \tag{4}$$

where W^{L_1} is the weight from first teacher model and W^{L_2} is weight from second teacher model. α is defined in Eq. 5. Furthermore, at each iteration, the 3D-DenseUNet model tunes the model weights by utilizing α, which can be formalized as:

$$\alpha = \frac{(P + 1)}{((1/L) + N)}, \tag{5}$$

where P is the accuracy, L is the loss and N is the batch size for each iteration. As accuracy and loss values depend on weights of the parameters, the α value resulted dynamically from these values in addition to the batch size of each iteration. A careful selection of α can affect the performance. In the early phase of training the proposed model, the accuracy and loss of the model obtained have not fit well enough. Thus, to avoid losing global information, α takes small values and the model takes full advantage of the weights of teacher models.

On the other hand, if α is much large, the model relies more on the current weights, leads to slow convergence and tends to overfit with low accuracy. Therefore, α should be close to 1. In line with that, we run the model several times and notice the model gains better performance when starting with 0.01 and gradually grows along with the training until approaching 1.

The cross-entropy used as a cost function to measure the average (expected) divergence between the predicted output P and the ground truth T being approximated over the entire domain of the input sized $W \times H \times D \times C$. Let θ denote the network parameters (i.e., convolution weights, biases and P_i from the parametric rectifier units), and Y_s^v the label of voxel V in the S-th image segment, we optimize the cross-entropy equation as following:

$$J(\theta) = -\frac{1}{S \cdot V} \sum_{s=1}^{S} \sum_{v=1}^{V} \sum_{c=1}^{C} \delta(Y_s^v = C) \cdot \log P_c^v(X_S), \tag{6}$$

where $P_c^v(X_s)$ is the SoftMax output of the network for voxel V and class C, when the input segment is X_s. The overall fuse process is presented in Algorithm 1.

Algorithm 1: The fuse model algorithm.

---

1 **Input:** Initial model parameters P_0, dataset X, number of epochs N, data batch $D^i = (x_i, y_i)$, initial weights w.
2 **Output:** Prediction for segmenting MRI into a number of tissues GM, WM, and CSF.
3 **for** $n_i \in N$ **do**
4      $x_i \leftarrow X$
5   **if** $trainmodel\ is\ TM1$ **then**
6   **for** $j \in W^i$ **do**
7              $w^{L_1} \leftarrow w^{ij}$
8   **end**
9   **end**
10   **if** $trainmodel\ is\ TM2$ **then**
11   **for** $j \in W^i$ **do**
12              $w^{L_2} \leftarrow w^{ij}$
13   **end**
14   **end**
15   **if** $trainmodel\ is\ Fuse$ **then**
16   **for** $j \in W^i$ **do**
17              $Calculate\ \alpha$ by Eq. 5
18              $Update\ W^{ij}$ by Eq. 4
19   **end**
20   **end**
21      $loss = \mathcal{L}_{\mathcal{CE}}(x_i, y_i)$
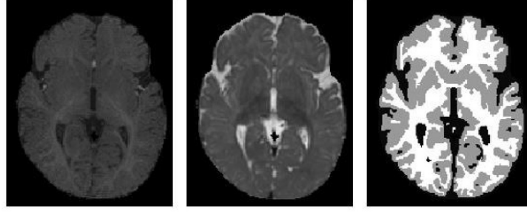22 **end**

Figure 3: An example of the $MICCAI\ iSEG$ dataset ($T1$,$T2$, manual reference contour).

4 Experiments and Discussion

This section studies the outstanding results of the proposed model by using metrics for evaluation, and compare our method with state-of-the-art baselines. First, the evaluation metrics and datasets are described. Then, the results of the proposed form are discussed.

4.1 Evaluation Metric

4.1.1 Dice Coefficient (DC)

In the research on brain segmentation, many comparison metrics have been broadly employed to assess the accuracy of segmentation methods in terms of effectiveness and efficiency. Accordingly, Dice Coefficient (DC) [9] is a powerful and important metric to assess the accuracy between predicting the model and the ground truth of data. Many recent studies used it as a benchmark and have achieved remarkable results as evaluated by the Dice Coefficient, especially, in an exploration of gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). It evaluates the overlap between the reference segmentation and automated segmentation. This metric is defined as follows:

$$DC(V_{ref}, V_{auto}) = \frac{2|V_{ref} \bigcap V_{auto}|}{|V_{ref}| + |V_{auto}|}, \tag{7}$$

where V_{ref} denotes for the reference segmentation and V_{auto} denotes for the automated segmentation. The DC values are given in the range of [0, 1], where 1 denotes perfect overlap and 0 denotes total mismatch.

4.1.2 Datasets

To evaluate the performance of our proposed model, we conduct experiments on $MICCAI\ iSEG$ dataset. During analysis of the $MICCAI\ iSEG$ dataset, there is a big difference in image data characteristics. The $MICCAI\ iSEG$ dataset was described in [9], that has a total of 10 images, including $T1$-1 through $T1$-10, $T2$-1 through $T2$-10, and a ground truth for the training set. And there are 13 images, T-11 through T-23, including T-11 through T-23 for test set. Figure 3 illustrates an example of the $MICCAI\ iSEG$ dataset. The parameters used to create $T1$ and $T2$ are listed in Table 1.

Table 1: Parameters used to generate T1 and T2.

Data	TR/TE	Flip angle	Resolution
T1	1,900/4.38 ms	7	1×1×1
T2	7,380/119 ms	150	1.25×1.25×1.25

The $MRBrainS$ is described by "https://mrbrains13.isi.uu.nl", which is an adult dataset and contains only 20 subjects. In this paper, $T1$- and $T2$-fluid-attenuated inversion recovery ($FLAIR$) images are used for segmentation. In $MRBrainS$, the number of training images are five (i.e., 2 male and 3 female). And the number of testing images are fifteen. The dataset follows the eight tissues of segmentation of (a) peripheral cerebrospinal fluid, (b) basal ganglia, (c) cerebellum, (d) white matter lesions, (e) brainstem, (f) lateral ventricles, (g) white matter, and (h) cortical gray matter.

4.2 Experimental Results

The proposed model is extensively assessed on the $3D$ multi-modal isointense infant brain tissues in $T1$- and $T2$-weighted brain MRI scans task. The task is to conduct automatic segmentation of MRI images into the cerebrospinal
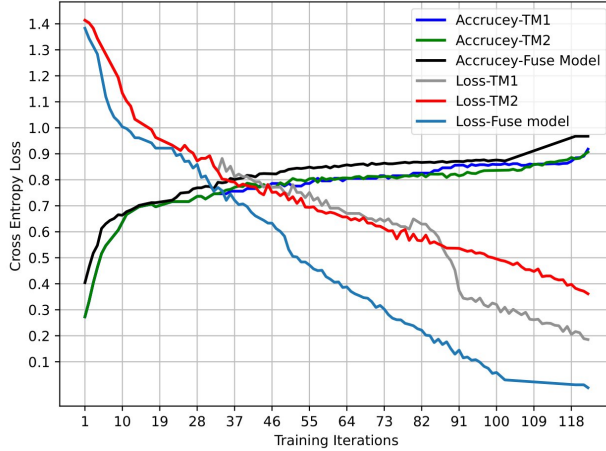
Figure 4: The accuracy and loss validation for propose model at $MICCAI\ iSEG$ dataset.

fluid (CSF), gray matter (GM) and white matter (WM) regions. In our experiment, the $MICCAI\ iSEG$ dataset is used. For each tissue type, we first replicate the model as independent teacher models. Each model was trained on a determined data type (the first teacher model trained on $T1$ and the second on $T2$) with one sample as a validation set and the remaining nine as training sets separately. At each iteration, the 2000 patches of $32 \times 32 \times 32$ were randomly selected as the training and validating datasets and processed in a batch size of 8. In the Fuse model process, we take patches the same size as that employed in both teacher models through the original image with a fixed overlapping step size. The size of the overlapping step should not be greater than the patch size. Thus, the third model as the fuse model benefited teacher models' weight for accurate tissue segmentation and refined tissue probability maps for every voxel in the original image for GM, WM and CSF. The proposed model was trained for 5000 epochs on the training and validation datasets.

We can see that, Figure. 4 plots the proposed model's accuracy and loss validation stage at different iterations. The figure demonstrates that the teacher models obtained bad accuracy in the early training phase, and the loss percentage was high. On the contrary, the fuse model results were better in the early training phase. With the continuation of the training and validation phase, it was noted that any improvement in the teacher models led to an improvement in the fuse model.

The Dice Coefficient (DC) is used as the evaluation metric to evaluate the accuracy of the fuse model for the validation subject. DC investigates the overlap level among the segmentation space and ground label. As we notice in Table 2, the results of the proposed model were compared to advanced state-of-the-art deep learning models in terms of metrics for the segmentation of CSF, WM, and GM brain tissues. Higher DC values denote more extensive overlap between manual and automatic segmentation boundary. As a result, our model has equivalent values to some models in some cases and outperforms others. Particularly, our model yields the best DC values for CSF and GM and the second ranked result for WM. Similarly, the 13 unlabeled images were used as the test set during the testing phase. The proposed model outperforms the state-of-the-art models in segmenting for GM, WM and CSF, as shown in Table 3. The results reveal excellent performance compared to other methods. As we can see, the proposed model produces the best DC values in GM, equal values in CSF and is slightly smaller in WM brain tissues.

Table 2: Segmentation performance in Dice Coefficient (DC) obtained on the $MICCAI\ iSEG$ dataset for the image used as a validation set. The best performance for each tissue class is highlighted in bold.

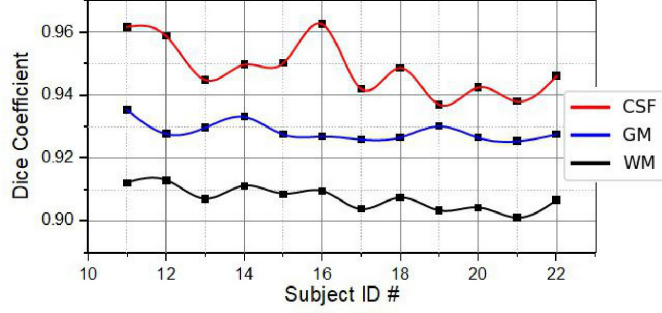| Model	Dice Coefficient (DC) Accuracy		
	CSF	**GM**	**WM**
Wang et al. [24]	0.95	**0.92**	**0.91**
Hoang et al. [25]	0.95	0.91	**0.91**
Dolz et al. [10]	**0.96**	0.92	0.90
Qamar et al. [26]	**0.96**	0.92	**0.91**
First Teacher Model	0.80	0.79	0.90
Second Teacher Model	0.89	0.82	0.81
Fuse Model	**0.96**	**0.92**	0.90

9

Figure 5: Performance of the proposed model on 13 different subjects of $MICCAI\ iSEG$ dataset as test set.

Table 3: Segmentation performance in Dice Coefficient (DC) obtained on the $MICCAI\ iSEG$ dataset for the 13 unlabeled images used for the test set. The best performance for each tissue class is highlighted in bold.

Model	Dice Coefficient (DC) Accuracy		
	CSF	**GM**	**WM**
Wang et al. [24]	0.95	0.92	**0.91**
Hoang et al. [25]	0.95	0.91	**0.91**
Dolz et al. [10]	**0.96**	0.92	0.90
Qamar et al. [26]	**0.96**	0.92	**0.91**
First Teacher Model	0.82	0.80	0.89
Second Teacher Model	0.88	0.90	0.81
Fuse Model	**0.96**	**0.93**	0.90

Furthermore, Figure. 5 illustrates the accuracy metric values of the proposed model for the 13 subjects in the test set. As the figure shows, the DC values have some variance in different subjects. This can be justified because providing additional training information by fused weights from two teacher models can enhance the segmentation results to significantly distinguish the edges between the white and gray matter.

In addition, we conduct an experiment on $MRBrainS$ dataset to further verify the proposed model's efficiency. Table 4 shows the performance of our proposed model in the testing set compared with the state-of-the-art models in terms of DC metric for the segmentation of CSF, WM, and GM brain tissues. Our model produces good DC values compared to other models, where it obtains the top results for CSF and GM and slightly fewer results for WM.

Table 4: Segmentation performance (Dice Coefficient, DC) on the $MRBrainS$ dataset. The best performance for each tissue class is highlighted in bold.

Model	Dice Coefficient (DC) Accuracy		
	CSF	**GM**	**WM**
STH [27]	0.83	0.85	0.88
HyperDenseNet [28]	0.83	0.86	**0.89**
First Teacher Model	0.39	0.40	0.50
Second Teacher Model	0.43	0.60	0.44
Fuse Model	**0.86**	**0.87**	0.87

To be more accurate in efficiency and effectiveness, the training time was taken as a metric as shown in Table 5 with the average training time (in hours) and standard deviation (SD). To study that, we compare our model with related works that reported the training time. As can be seen from the table, the average execution time of the proposed framework, which is the summation of the two teacher models and the fuse model, is lower than that of the HyperDenseNet [28]. On the contrary, it is larger than Qamar et al. [26], despite our framework being composed of three independent models. This exemplifies that our proposed framework's architecture has fewer learned parameters.

Accordingly, we investigate the number of parameters used in our model. As noticed in Table 6, our model has a number of parameters that are 20% less than the state-of-the-art models. The architecture of our proposed model is much deeper than all state-of-the-art models because it is composed of three independent models, which consist of 82 layers with 2.12 million learned parameters. Therefore, the proposed architecture is deeper compared to other existing approaches, and it achieves the 96% accuracy. In terms of parameters, the model of Wang et al. [24] is close to ours.

Table 5: Average execution time (in hours) and standard deviation (SD) on the $MICCAI\ iSEG$ dataset.

Model	Time (SD)
HyperDenseNet [28]	105.67 (14.7)
Qamar et al. [26]	38.00 (0.00)
First Teacher Model	36.45 (0.12)
Second Teacher Model	36.45 (0.12)
Fuse Model	**22.61 (0.21)**

Table 6: Comparison on the number of parameters

Model	Number of Parameters
Wang et al. [24]	2,534,276
HyperDenseNet [28]	10,349,450
First Teacher Model	2,331,160
Second Teacher Model	2,331,160
Fuse Model	**2,123,211**

Moreover, Figure. 6 depicts the visualization result of the proposed model on the image used as the validation set. We can see the similarity between the results achieved by the proposed model and the ground truth fairly close. Specifically, the segmentation accuracy is high at the boundary.

Overall, our results have matched the results of some models and are much better than others regarding the image segmentation accuracy. Despite our model architecture being much deeper than the state-of-the-art approaches, it contains fewer parameters and takes less execution time which leads to improvement in performance comparable to the state-of-the-art approaches.

4.3 Ablation Study

To evaluate the effectiveness of the component blocks used in our proposed model with the fusion method, we conduct ablation studies on the $MICCAI\ iSEG$ training and testing set. We apply different α values to examine what suits for our model. The two models with α from Eq. 5 or with constant values for α are constructed and to ensure a suitable procedure. These models are trained for 5000 epochs with a batch size of 32 and supervised by the proposed combination loss function. Moreover, the same number of down$-$sampling and up$-$sampling blocks are used. The lowest number of alpha starts with α =0.1 using multiple runs by changing the α 0.2, 0.3, up to 0.9.

We can see from the results that the model is improving in some values, while it is fluctuating in the values of the loss and the accuracy at some of them, especially at the minimum and maximum values which is why we use Eq. 5, in that Eq. 5 leads to better performance. Table 7 and Figure. 7 report comparisons of the ablation studies on the validation set in accuracy and loss metrics. In the early training steps, the accuracy and loss of the model obtained have not fit well enough and α has very small values. In contrast, when α is much large, it leads to slow convergence and tends to overfit with low accuracy. Thus, the α values are taken according to Eq. 5 to alleviate the fluctuations in the early stage and the overfitting issue throughout the training stage.

Also, from the point of view of the Dice metric in Table 7, it can be observed that adopting the fusion method with α from Eq. 5 is an effective method to improve the performance of brain MRI segmentation. These experimental
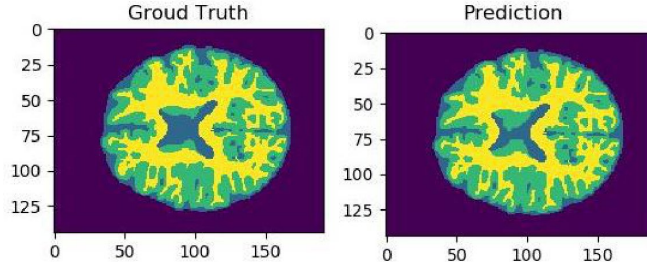
Figure 6: Comparison on the 150-th slice results in $MICCAI\ iSEG$ dataset.
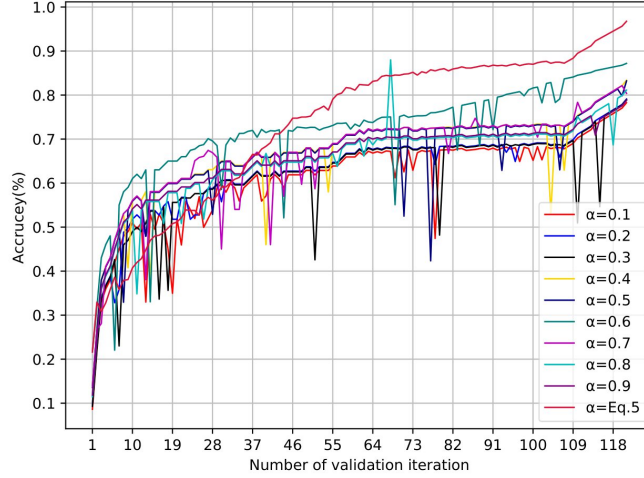
Figure 7: The accuracy of validation dataset on $MICCAI\ iSEG$ with various setting of α.

results show that α hyper-parameter enhances the performance of the proposed model components for brain MRI segmentation.

Table 7: Segmentation performance in Dice Coefficient (DC) for the $MICCAI\ iSEG$. The best performance for each tissue class is highlighted in bold.

Fuse model with different α	Dice Coefficient (DC) Accuracy		
	CSF	GM	WM
0.1	0.78	0.74	0.70
0.2	0.79	0.75	0.71
0.3	0.79	0.75	0.71
0.4	0.83	0.79	0.76
0.5	0.83	0.79	0.76
0.6	0.87	0.83	0.80
0.7	0.80	0.75	0.71
0.8	0.81	0.77	0.70
0.9	0.79	0.74	0.70
Defined in Eq. 5	**0.96**	**0.92**	**0.90**

5 Conclusion

Despite the great success of recent U-net models in medical image segmentation, it still suffers from some limitations. This work proposed a new deep-learning model, termed 3D-DenseUNet, which works as adaptable global aggregation blocks in down-sampling to solve spatial information loss. A self-attention module was connected with down/up-sampling blocks to integrate the feature maps in three spatial and channel dimensions, effectively improving its representation ability and discrimination. Moreover, a new method called Two Independent Teachers ($2IT$) is proposed, which summarizes the model weights instead of label predictions. Each teacher model is trained on different types of brain data, $T1$ and $T2$, respectively. Then, a fused model was added to improve test accuracy and enable training with fewer parameters and labels without modifying the network architecture compared to an advanced method called Temporal Ensembling. Empirical results showed that the proposed model outperformed other mainstream models. Overall, the results of the proposed model match the results of other models in the segmentation of brain images. This may be considered a weakness in the model, but the proposed model outperformed in terms of reducing the number of parameters and labels in addition to the implementation time.

References

[1] Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3D semantic segmentation for medical images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 552–561, 2020.

[2] Afifa "Khaled, Chung-Ming Own, Wenyuan Tao, and Taher Ahmed" Ghaleb. "improved brain segmentation using pixel separation and additional segmentation features". In Xin "Wang, Rui Zhang, Young-Koo Lee, Le Sun, and Yang-Sae" Moon, editors, *"Web and Big Data"*, pages "85–100", "2020".

[3] Afifa Khaled, Jian-Jun Han, and Taher A. Ghaleb. Multi-model medical image segmentation using multi-stage generative adversarial networks. *IEEE Access*, 10:28590–28599, 2022.

[4] Taher A. Ghaleb Afifa Khaled, Jian-Jun Han. Learning to detect boundary information for brain image segmentation. *BMC Bioinformatics*, 23(1):1–15, 2022.

[5] Hao Zheng, Faming Fang, and Guixu Zhang. Cascaded dilated dense network with two-step data consistency for MRI reconstruction. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] Jing Li, Jiaqi Pu, Xiaogen Zhou, Haonan Zheng, Qinquan Gao, Ensheng Xue, and Tong Tong. Multi-angle dual-task consistency for semi-supervised medical image segmentation. In *The 9th International Conference on Information Technology: IoT and Smart City*, pages 61–66, 2021.

[7] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, 2018.

[8] Taher A Ghaleb Afifa Khaled, Jian-Jun Han and Radman Mohamed. Learning to detect boundary information for brain image segmentation. *Arabian Journal for Science and Engineering*, 48:2133–2146, 2023.

[9] Li Wang, Dong Nie, Guannan Li, Élodie Puybareau, Jose Dolz, Qian Zhang, Fan Wang, Jing Xia, Zhengwang Wu, Jia-Wei Chen, et al. Benchmark on automatic six-month-old infant brain segmentation algorithms: the iseg-2017 challenge. *IEEE Transactions on Medical Imaging*, 38(9):2219–2230, 2019.

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

[11] Ping Zheng, Xunfei Zhu, and Wenbo Guo. Brain tumour segmentation based on an improved u-net. *BMC Medical Imaging*, 22(1):1–9, 2022.

[12] Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P Hughes, and Danny Z Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 408–416, 2017.

[13] Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen De Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 810–818, 2019.

[14] Kaiping Wang, Bo Zhan, Yanmei Luo, Jiliu Zhou, Xi Wu, and Yan Wang. Multi-task curriculum learning for semi-supervised medical image segmentation. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 925–928, 2021.

[15] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[16] Kaiping Wang, Bo Zhan, Chen Zu, Xi Wu, Jiliu Zhou, Luping Zhou, and Yan Wang. Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. *Medical Image Analysis*, 79:102447, 2022.

[17] Wenhui Cui, Yanlin Liu, Yuxing Li, Menghao Guo, Yiming Li, Xiuli Li, Tianle Wang, Xiangzhu Zeng, and Chuyang Ye. Semi-supervised brain lesion segmentation with an adapted mean teacher model. In *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, pages 554–565, 2019.

[18] Olaf Ronneberger, Philipp Fischer, Brox, Nassir Thomas", editor="Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, 2015.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[20] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[21] Takeru Miyato, Shin ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning, 2018.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification, 2015.

[23] Yuyin Zhou, Yan Wang, Peng Tang, Song Bai, Wei Shen, Elliot Fishman, and Alan Yuille. Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 121–140, 2019.

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

[25] Dinh-Hieu Hoang, Gia-Han Diep, Minh-Triet Tran, and Ngan T. H Le. DAM-AL. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, April 2022.

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

[27] Örjan Smedby Amirreza Mahbod, Manish Chowdhury and Chunliang Wang. Automatic brain segmentation using artificial neural networks with shape context. *Pattern Recognition Letters*, 101:74–79, 2018.

[28] José Dolz, Karthik Gopinath, Jing Yuan, H. Lombaert, Christian Desrosiers, and Ismail ben Ayed. HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation. *IEEE Transactions on Medical Imaging*, 38:1116–1126, 2018.