IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 16, 2020, accepted April 24, 2020, date of publication May 8, 2020, date of current version June 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2993504

Hippocampus Segmentation for Preterm and Aging Brains Using 3D Densely Connected Fully Convolutional Networks

DEBIN ZENG, QIONGLING LI, BAOQIANG MA, AND SHUYU LI
School of Biological Science and Medical Engineering, Beihang University, Beijing 100191, China
Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing 100191, China

Corresponding author: Shuyu Li (shuyuli@buaa.edu.cn)

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 81471731 and Grant 81622025.

ABSTRACT Efficient and accurate segmentation of hippocampi from preterm and aging brain MR images is one of the most fundamental steps in understanding hippocampal growth and development or diagnosing and monitoring various clinical conditions. Current hippocampus segmentation methods for preterm and aging brain are limited due to: 1) they can rarely achieve preterm infant hippocampus segmentation; 2) the computation cost is high; 3) current deep learning models cannot well handle the hippocampal feature learning; 4) they are not open obtainable. To deal with these problems, we propose an efficient, open-source algorithm, 3D densely connected fully convolutional network (3D-DCFCN) for the infant and aging hippocampal segmentation. Specifically, we search for a suitable distribution of the hierarchical receptive field size and a joint loss function to balance local and global information and lead to better optimization. In addition, we use image cross-registration for vast augmentation of the infant training data and incorporate multi-modality infant brain information. We compare the performance of our algorithm with those of several state-of-the-art methods. The results show that our method outperforms all comparison methods on infant and aging datasets and achieves much faster speed (less than 0.11s per image). We also provide a notably comprehensive evaluation of the method. Our experiments further demonstrate our model can 1) well generalize to the dataset with different magnetic fields; 2) satisfactorily find hippocampal atrophy in cognitive-decline groups compare with normal controls.

INDEX TERMS Preterm, aging, hippocampus segmentation, hippocampal volume analysis, 3D densely connected fully convolutional network (3D-DCFCN).

I. INTRODUCTION

Hippocampus plays a critical role in high-order cognition functions, including memory, spatial location, and navigation. In preterm children born at less than 37 weeks of gestation, the hippocampus has uniformly been shown to be smaller relative to term-born controls [1], and this decrease in the rate of hippocampal development in children born severely preterm (less than 32 weeks) is associated with impaired cognition and working memory at 2 years of age [1], [2] and school-age [3], [4]. In addition, hippocampus is vulnerable to many psychiatric disorders and neuro degenerative diseases, such as temporal lobe epilepsy, schizophrenia, and Alzheimer's disease, especially in the aging brain [5]–[8]. Therefore, segmentation and quantitative analysis of hippocampus for preterm and aging brains are important to understand and predict hippocampal development in neonates or diagnose and monitor various clinical conditions in aging.

The past ten years have seen increasingly rapid advances in developing automated methods for hippocampus segmentation, and these methods can be categorized into four classes and hybrid versions of these classes: (1) Atlas-based method [9]–[12]. In this method, one or multiple atlases are directly aligned with the target image to obtain the target label. This method enables segmentation in individuals with great anatomical variability, but the disadvantage is that it requires many registration operations, which is computationally expensive, making it impractical for application requiring rapid processing speed. (2) Deformable

The associate editor coordinating the review of this manuscript and approving it for publication was Fan Zhang.

models [13]–[15]. In this technique, a contour is initialized in the image and iteratively deformed to generate a new contour. For example, certain methods used models such as the active contour model [15], active shape model [16] and active appearance model [14] for segmentation. Nevertheless, an obvious problem is that the method is sensitive to contour initialization. (3) Traditional machine-learning method with handcrafted features [17]–[22]. This type of method uses various handcrafted features to classify voxels, such as using spatial and intensity features in a random-forests model [18], using atlas and appearance features in a support vector machine method [20]. However, handcrafted features usually suffer from limited representation capability, and these methods also require careful engineering and specific expertise for accurate recognition. (4) Deep learning-based method [23]–[28]. These approaches allow models to learn the features that optimally represent the data for the problem at hand, such as the stacked autoencoder [27], multiview ensemble 2D convolutional neural network [23], parallelized long short-term memory (LSTM) [28], or 3D convolutional neural network methods [24], [25]. These deep learning-based methods can perform rapidly and obtain accurate segmentation without a manual design of intricate and specific input features. Most of the current methods cannot handle hippocampal segmentation for preterm brain.

In recent years, extraordinary improvements have been achieved by deep learning, especially convolutional neural networks (CNNs) in image segmentation, including hippocampus segmentation. Some of these approaches [26], [27], [29] often replace previous handcrafted image features in atlas-based methods with representation inferred from the deep-learning models, but they are also time-consuming and only focus on 2D images. It should be noted that brain MR images and many other medical image modalities consist of volumetric data. As far as we know, two main categories of end-to-end CNNs are available for volumetric image segmentation. (1) This type of method applies 2D CNNs by taking one slice, assembled adjacent slices, orthogonal planes, or multiple diagonal planes as input to compensate for spatial information [23], [30]. However, these models often contain multistream architectures that require longer training and testing time as well as additional storage. (2) The other methods apply 3D convolution to extract discriminable features from volumetric data and have proved attractive performance [24], [25], [30]. These methods are composed of many down-sampling layers, which denote an overlarge and highly sparse distribution of hierarchical receptive field sizes that are beneficial to the integration of more global context information than fine-grained or local information. However, fine-grained or local information is more important to achieving good voxel-level accuracy of tiny structures such as the hippocampus.

The overwhelming majority of these segmentation methods suffer from at least one of the following restrictions: (i) almost all of these approaches cannot deal with preterm brain with underdevelopment of hippocampus; (ii) the segmentation accuracy is restricted to registration error, contour initialization, or discriminative capability of hand-crafted features; (iii) the computation cost is high due to complex models or registration; (iv) current deep learning models cannot well handle the hippocampal feature learning as these methods cannot extract enough fine-grained information and the imbalance of background and small hippocampus mislead the optimization; (v) the algorithms or the trained models are not openly obtainable.

In this work, we propose an efficient and openly obtainable segmentation algorithm based on 3D densely connected fully convolutional network to deal with the restrictions of segmentation of hippocampi for preterm and aging brain from 3D MR images. First, we incorporate the 3D densely connected block introduced by [31] as well as the bottleneck layer and compression layer [32], [33] in our model. It maximizes information flow such that context information at a certain level can be used directly to inform decisions at other levels, making the model more adaptive to the two segmentation tasks. Additionally, this block and these layers embody fewer parameters and leads to implicit deep supervision, which improves computational efficiency, reduces overfitting, and makes training easier. We also construct a model with a suitable distribution of the hierarchical receptive field sizes to integrate more fine-grained information which is conducive to the segmentation of tiny structures. We also propose a joint loss function to balance the background and small hippocampi and to merge multilevel context information. Furthermore, we balance the number of training images and global information using a random cropping method with a proper cropping size, and we propose the cross-registration to vastly augment the training data and further improve the performance of the infant hippocampal segmentation. Finally, because the infant brain has a much lower tissue contrast, we incorporate multimodality MR images in our model to integrate more comprehensive anatomical information. Our model was trained on 2 datasets consisting of preterm infants and aging subjects. Our method outperforms all of the state-of-the-art methods on the same dataset. The findings from this study make several contributions to the current literature which can be summarized as follows:

i) We propose an efficient 3D-DCFCN-BC model that is more adaptive to the two segmentation tasks, more efficient, easier to train, and reduce overfitting.

ii) We find a suitable distribution of the hierarchical receptive field sizes to integrate more fine-grained and local information to cope with the problem of a relatively small hippocampal size. In addition, we propose a joint loss function to balance the background and small hippocampi and to merge multilevel context information.

iii) We propose cross-registration to vastly augment the infant training data and further improve the performance. In addition, this is the first end-to-end deep learning method for preterm infant hippocampus segmentation.

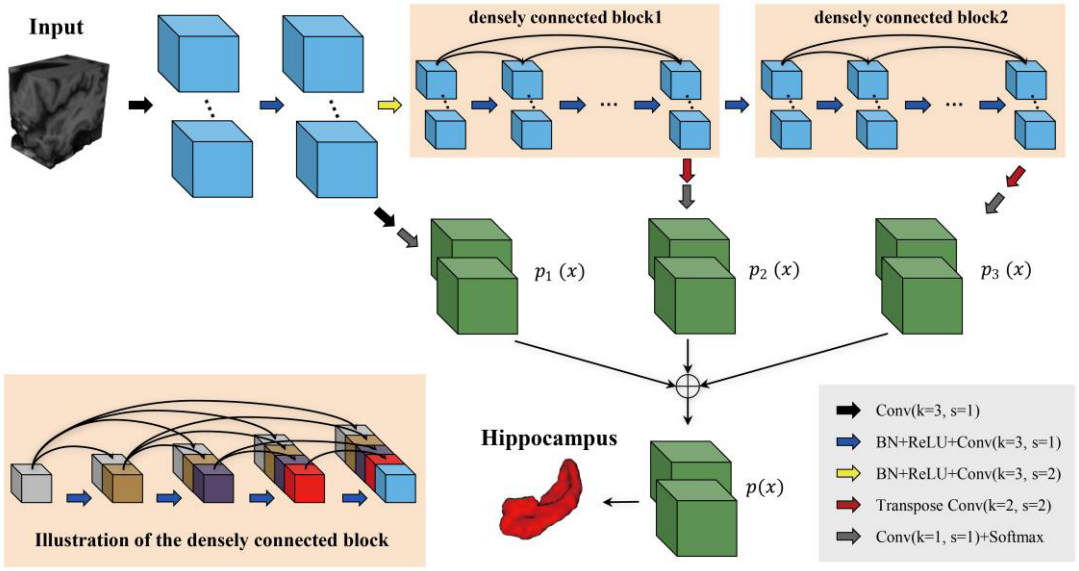


FIGURE 1. The architecture of our proposed 3D densely connected fully convolutional networks (3D-DCFCN).

iv) Our proposed method outperforms the other hippocampal segmentation methods not only in precision but also in speed (near real-time, less than 110ms per image).

II. METHOD

Turning now to the description on the proposed 3D densely connected fully convolutional network (3D-DCFCN), we first present the architecture of 3D-DCFCN and the 3D-DCFCN-BC, i.e., the extended DCFCN model with bottleneck layers and a compression layer added. Second, we introduce our joint loss function for incorporation of deep supervision and balance of the foreground and background. Third, we propose a new method of data augmentation for a small dataset, such as infant data. Finally, we describe the hyperparameter optimization and the details of network training and testing.

A. 3D DENSELY CONNECTED FULLY CONVOLUTIONAL NETWORK (3D-DCFCN)

Our objective is to enable our model to perform preterm infant and aging hippocampi segmentation tasks by proposing a less complicated network architecture with significantly fewer parameters to learn and that can automatically adapt to the two problems. To achieve this objective, we incorporate the densely connected block and fuse multilevel context information in our carefully designed model.

1) DENSELY CONNECTED BLOCK

The lower-left corner of Fig. 1 shows the 3D densely connected block used in our model. Any layer in the block directly connects to all subsequent layers, which is the most distinguishing pattern in this block. Hence, the l_{th} layer has l inputs containing the feature maps of all preceding layers and can be expressed as follows:

$$x_l = F_l([x_0, x_1, \ldots, x_{l-1}]) \tag{1}$$

where x_l represents the feature maps produced in the l_{th} layer, $[x_0, x_1, \ldots, x_{l-1}]$ denotes the concatenation of these feature maps, and F_l is defined as a composite function of three consecutive operations, namely, batch normalization (BN), followed by a rectified linear unit (ReLU) and a $3 \times 3 \times 3$ convolution (Conv). The reason for using these operations is explained later. Each layer produces k feature maps, the l_{th} layer takes $k_0 + k \times (l-1)$ feature maps as input, where k_0 is the number of feature maps in the input layer. Thus, the hyperparameter k is referred to as the growth rate of the block. We incorporate the densely connected block in our model because (i) it maximizes information flow and is conducive to the model adaptability for the two tasks that require context information at different levels; (ii) no need exists to relearn redundant feature maps at the intermediate layers, and this can ease the training burden and reduce the risk of overfitting; (iii) leading to implicit deep supervision [38], and this makes it easier to train a notably deep neural network.

2) 3D-DCFCN

Fig. 1 shows the overall structure of the 3D-DCFCN. $3 \times 3 \times 3$ filters Smaller filters are believed to require fewer parameters and less computational complexity, and they obtain details better than the larger filters [33]. Because $2 \times 2 \times 2$ convolution filters cannot maintain the size of feature-map comparing to the input, we construct our initial model using the feasible and smallest filters of $3 \times 3 \times 3$ convolution. In addition, most of the previous hippocampus segmentation methods use pooling as a downsampling layer [23]–[25], whereas the pooling layer can be viewed

as performing a feature-wise convolution in which the activation function is replaced by the p-norm. We use convolution with stride 2 for replacing the pooling layer because it can add inter-feature dependencies [39]. Because the 3D deep learning model is more difficult to converge compared with 2D models [40], [41], batch normalization (BN) is applied to enable faster and more effective optimization, i.e., robustness to hyperparameter setting and avoidance of gradient explosion or vanishing. The smoothing effect of BN is also believed to facilitate better generalization [42], [43]. In addition, the rectified linear units (ReLU) are applied as the activation function. Finally, we use upsampling layers to upscale the output to the original resolution to enable voxel-wise prediction. Instead of using unpooling or interpolation to perform in-network upsampling without learnable parameters, we use transpose convolution (which is also referred to as deconvolution in certain papers) to achieve this goal [44], [45]. The filter size and stride size of the transpose convolution are each set to 2 to avoid the uneven overlap issue and reduce checkerboard artifacts [46].

Similar to most of the deep-learning segmentation models, our 3D-DCFCN also contains an encoder and a decoder. At the encoder stage, we first stack two convolutional layers with kernel size $3 \times 3 \times 3$, and we use a convolution layer with the same kernel size and stride 2 to downsample the feature maps, we also make each preceding convolution layer outputs 32 feature maps. Later, these feature maps outputted from the third convolution layer are input to another group of layers containing two densely connected blocks. Each block is consisting of 5 layers with a growth rate of 16. A convolution layer with kernel size $3 \times 3 \times 3$ is applied to connect the two blocks, and the connection layer outputs 32 feature maps, which flow into the second block. We zero-pad each side of the inputs by one pixel to hold the feature-map size fixed for $3 \times 3 \times 3$ convolution with stride 1 or downsample the feature map by a factor of 2 for convolution with stride 2. At the decoder stage, we use a three-stream fusion network to integrate multilevel context information. Each stream contains a transpose convolution (the first stream is convolution due to no downsampling) and a softmax layer as an auxiliary classifier to produce a probability map of the hippocampus ($p_1(x)$, $p_2(x)$, $p_2(x)$). All of these maps are fused to obtain the finer probability map ($p(x)$).

3) 3D-DCFCN-BC

Because each layer in the densely connected block produces 16 feature maps, it has many inputs in the relative deep layer of the block and the connection layer between the two blocks. Thus, we introduce the $1 \times 1 \times 1$ convolution before the connection layer and connection layer are all replaced by BN-ReLU-Conv-BN-ReLU-Conv. This architecture, which is referred to as the bottleneck layer and compression layer, can further improve the computation efficiency and reduce the number of parameters. We call the model with the bottleneck layer and compression layer as 3D-DCFCN-BC.

Variations of 3D-DCFCN. At the encoder stage, instead of using only classical out-of-the-box CNN architectures which removed fully connected layers, we design our model for better suitability to our specific tasks. Note that we can create variations of this 3D-DCFCN structure by building different depths of the encoder network and different numbers of downsampling layers. This process represents an inherent trade-off between the integration of fine-grained or local information and global information as well as computation cost. In this work, we define the scale of the feature maps in a specific layer as the local patch in the input image on which the inference of each pixel in the feature maps relies, and it is also referred to as the receptive field. If we use additional downsampling layers or deeper networks, the model acquires feature maps with a larger scale, which means that it incorporates much information from the global context of the image. Global context information can resolve local ambiguities. Nevertheless, this process also leads to using more parameters, computation, and storage and integration of less fine-grained or local information critical to achieving good pixel-level accuracy and vice versa. Experimentally, we find that the structure displayed in Fig. 1 achieves the best performance compared with its other variations (details in Section III-B-1)) because we find a suitable distribution of the scale of the model layers to balance the local information and global information, as well as obtain a small enough computational cost.

Additionally, we suggest that the use of multiscale networks at the encoder stage might aid the model in generalizing to different tasks that require different scale feature maps. Possible architectures [47], [48] are popular models in the domain of natural image segmentation. This method uses multistream networks at the encoder stage that target different scales and subsequently fuse the multiscale feature maps to produce a single output for subsequent prediction. We have organized experiments with this method using two-stream fusion networks designed by us or with the pyramid pooling module proposed by [47] in the early stage of our experiments. We found that the result indicates a performance decline for the task of aging hippocampus segmentation. Considering that the two-stream fusion networks highly increase the number of parameters, it might harm the training process. In addition, the pyramid pooling module can supply feature maps with a larger receptive field, whereas an overly large receptive field is redundant and weakens the local information integration. Consequently, the use of multiscale networks may hamper the model in learning the effective features of the hippocampus.

B. JOINT LOSS FUNCTION

In many image segmentation problems such as that proposed in this work, the numbers of voxels/pixels from different classes usually differ, and the object of interest often occupies a small portion of the image to be analyzed. This scenario might mislead the optimization procedure to overfitting to the classes with many more voxels/pixels, similar to

the background in our problem. The dice objective function proposed by [49] is an efficient way to balance the foreground and background. Consequently, we incorporate the dice objective function into our loss function (see the third component of (2)). In addition, we also construct the auxiliary loss for those previously mentioned auxiliary classifiers, which indicates deep supervision (see the second component of (2)) [38], [40]. Finally, the proposed joint loss function that trains the entire network is written as follows:

$$L = \lambda||W||^2 - \sum_k \sum_{x\in V} \sum_\alpha w_k g^\alpha(x) \log p_k^\alpha(x) - \sum_\alpha \frac{2\sum_{x\in V} p^\alpha(x) \times g^\alpha(x)}{\sum_{x\in V}((p^\alpha(x))^2 + (g^\alpha(x))^2)} \quad (2)$$

where the first component is the L_2 weight decay regularization, W denotes the weights of our networks and λ is the trade-off hyperparameter. The second component is an auxiliary loss, and w_k (where k indicates the index of where the first component is the L_2 weight decay regularization, W denotes the weights of our networks and λ is the trade-off hyperparameter. The second component is an auxiliary loss, and w_k (where k indicates the index of auxiliary classifiers) is the trade-off weights of the auxiliary classifiers, which are set to 0.1 in our experiments. In this work, $p_k^\alpha(x)$ and $p^\alpha(x)$ denote the probability maps of channel α for voxel x in volume V produced by auxiliary classifiers and final fusion, and $g^\alpha(x)$ is the one-hot encoded ground truth for the corresponding voxel.

C. IMAGE CROSS-REGISTRATION FOR DATABASE AUGMENTATION

Database augmentation plays an important role in training tasks with limited data. Because deep-learning models contain a large number of parameters to learn, this method can reduce overfitting and improve generalization. Simple techniques including translating, rotating, flipping and scaling are widely used and are easiest to apply in classification tasks. The other methods [50], [51] generate new and also fake data using neural networks to integrate information from two or more data samples, but they can only be used in classification tasks. In this work, we present a new database augmentation technique using 12 degree-of-freedom (DOF) linear image registration, including rigid body transformation and affine transformation. This strategy can vastly augment the database simply and efficiently: using images after linear registration as the training data. More details of this technique are described in Algorithm 2.1.

Using this algorithm, we can expand n training samples to n^2. Fig. 2 shows that these generated fake MR images combine information from two real brain images using linear registration such that the shape of the hippocampus and the other anatomy are reliable, and as a result, they are feasible for network training. In addition, the generated images are also different from the two corresponding images because the operations involve affine transformations, therefore, this algorithm can add reasonable uncertainty and variation to the training set, which aids the model in improving performance.

Algorithm 2.1 linear image registration for vast augmentation of training data. This algorithm can be applied as a general strategy for varieties of brain MR image segmentation using machine learning.

Let n be the number of subjects for training.
Let X_i^1, X_i^2, Y_i be the original T1 and T2 MR image and label image of the i^{th} subject, respectively.
for $i = 1 \rightarrow n$ **do**
  for $j = 1 \rightarrow n$ **do**
    if $j = i$ **then**
      continue
    end if
    Transform X_i^1 to X_j^1 by 12 DOF linear registration and obtain
    the transformed image X_{ij}^1 and transformation matrix T_{ij}.
    Transform X_i^2 to X_j^2 by T_{ij} and obtain the transformed image X_{ij}^2.
    Transform Y_i to Y_j by T_{ij} and obtain the transformed label image Y_{ij}.
  end for
end for
Collect all X_i^1, X_i^2, Y_i, X_{ij}^1, X_{ij}^2, Y_{ij} to form the new augmented database.

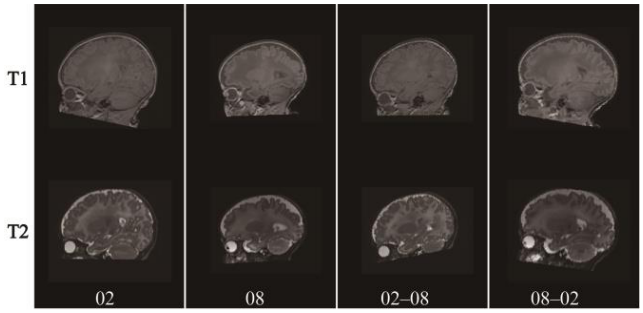


FIGURE 2. First two columns denote sagittal images from T1-weighted and T2-weighted MR of two infants with identity numbers 02 and 08. The two latter columns denote the MR images generated by the 12-DOF linear image registration between subjects 02 and 08. The 02-08 denotes the transformed image obtained by transforming 02 to 08.

The 12 DOF linear registration is implemented by FLIRT, a fully automated robust and accurate tool of FSL (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/) in our experiments. The number of histogram bins is set to 256, and we use the nearest neighbor interpolation and correlation ratio as the cost function.

D. MODEL TRAINING

Previously, random cropping was simply applied as a data augmentation technique in the various training tasks, although in segmentation task it is different from the other data augmentation technique mentioned before. This approach can effectively reduce the contribution of the

background in the model decision and focus more attention on the object, and thus we find that it is quite useful for our segmentation model by improving the precision and the stability. If we use smaller patches instead of the whole image to train the model, the model can use the background regions as negative examples in many cases with little foreground to reduce the contribution of the background. Obviously, when patches with little background are fed into the model, it is conducive to learning the foreground. However, if we use patches that are too small, it can harm the model to integrate information with a necessarily semantic range. Experimentally, we find a compromise between global information and the number of different training samples with random crop patches of size $32 \times 32 \times 32$.

In this study, all of the learnable weights and biases were initialized by a Gaussian (0, 0.01) and a constant (0), respectively. We trained our model by mini-batch gradient descent with momentum, the batch-size was set to 8, and the weight of the previous velocity of updating was set to 0.9. Additionally, we used polynomial decay as the learning rate policy, where the effective learning rate is calculated by base_lr∗(1-iter/max_iter)^(power), and the learning rate approaches zero by the end of the training (max_iter) with this policy, we set power to 0.9 here. Because the training performance can be improved by increasing the max_iter within a certain range due to the learning rate policy, we set max_iter to 51840 (40 epochs) for the infant dataset and 172800 (800 epochs) for the aging dataset experimentally to achieve the best and most stable performance. For the hyperparameters of the base learning rate and weight decay, we used the random search method [52] to select a group of optimized values. We defined a uniform distribution on a log-scale for the 2 hyperparameters, and each of the randomly sampled groups of hyperparameters was used to train the model on a small dataset randomly selected from the original dataset. Specifically, all of our methods and experiments are implemented using MATLAB and Caffe [53] with minor modifications for incorporating the dice objective function. Our trained models and algorithms can be found at https://github.com/DebinZeng/Hippo_seg.

III. EXPERIMENTAL RESULTS

We display the experimental analysis of model design and hippocampal volume analysis and comparisons of the proposed method with state-of-the-art algorithms. We first describe the two datasets and evaluation metrics and subsequently display the experimental analysis for model design. Then we present a comparison of the popular algorithms on the infant and aging, respectively. Additionally, we further tested the model (trained by aging dataset) on the inhouse dataset without any adaptive dataset-specific manipulation. Finally. volumetric analysis of the hippocampus on infant and aging dataset is conducted for volumetric correlation comparison among certain related methods, and cognitive-decline biomarker validation on the aging brain.

A. TWO DATASETS AND PREPROCESSING PROCEDURE

In this paper, we used two datasets composed of preterm infant and aging brain MR images to validate our proposed method.

TABLE 1. Demographic information for the infant dataset (ALBERT, 20 subjects, 2 modalities including T1-weighted and T2-weighted MR images).

	PRETERM (N=15)	Term (N=5)
Gender (F/M)	8/7	2/3
Gestational age at birth (weeks)	29 (26 to 35)	—
Postmenstrual age at scan (weeks)	40 (37 to 43)	41 (39 to 45)
Weight at scan (kg)	3.0 (2.0 to 4.0)	4.0 (3.0 to 5.0)

Abbreviations: F/M, female or male; kg, kilogram.
Note: Values denote median (range).

The infant dataset known as ALBERT (https://brain-development.org/brain-atlases/neonatal-brain-atlases/ neonatal-brain-atlas-gousias/) is composed of 15 preterm and 5 term-born neonates that were all scanned at term [34]. The demographic information for the infant dataset is displayed in TABLE 1. The MR images were obtained using a 3.0T Philips Achieva scanner. The voxel size of the preprocessed T1-weighted image is $0.82\text{mm} \times 0.82\text{mm} \times 0.82$ mm, and the T2 images were coregistered in identical fashion onto the same preprocessed T1 image and resampled to the same voxel size. The images were segmented into 50 regions of interest (ROI) including the hippocampus. A comprehensive and detailed illustration of the segmentation steps can be found in [34]. We extracted the hippocampus mask from the 50 ROIs. Similar to the aging dataset, we also used two bounding boxes with a size of $64 \times 64 \times 48$ around the left and right hippocampus to extract the image patches. The boxes were also located at two-fixed locations in the image. Finally, we split the infant dataset into ten groups to adopt a 10-fold cross-validation strategy in which five term-born infants were randomly assigned to different groups.

The aging dataset consists of harmonized protocol (HarP) data. This dataset contains 135 T1-weighted images obtained from the ADNI database [35] with manual hippocampal labels supplied by the EADC project (http://www.hippocampal-protocol.net/). The MR images were balanced by magnetic field strength, scan manufacturer, diagnosis, qualitative medial temporal atrophy (MTA), and age ranges [36]. The ADNI images were reoriented along the AC-PC line by six-parameter linear registration before manual delineation, and an MNI ICBM152 template with $1\text{mm} \times 1\text{mm} \times 1$ mm voxel dimension was used as the reference space. The hippocampi were segmented by expert raters based on the harmonized hippocampal protocol [37]. The demographic information for the aging dataset is displayed in TABLE 2. Because the hippocampus occupies only a small proportion of the whole brain, we used two bounding boxes with a size of $72 \times 72 \times 48$ around the left and right hippocampus to extract image patches and largely reduce

TABLE 2. Demographic information for the elderly dataset (EADC-ADNI, 135 ADNI subjects, T1-weighted MR images).

	1.5T (N=68)				3.0T (N=67)			
Diagnosis	AD	LMCI	MCI	NC	AD	LMCI	MCI	NC
Number	22	9	15	22	23	8	14	22
Gender (F/M)	11/11	6/3	6/9	10/12	13/10	2/6	5/9	12/10
Age (years)	74.1 (8.04)	72.8 (7.63)	74.2 (8.43)	76.2 (7.52)	74.8 (8.32)	75.1 (9.23)	76.2 (7.99)	76.2 (7.56)
MMSE	23 (2)	27(3)	27(3)	29 (1)	20 (5)	26 (2)	25(3)	29 (1)

Abbreviations: AD, Alzheimer's disease; MCI, mild cognitive impairment; LMCI, late MCI; NC, normal controls; MMSE, mini-mental state examination; F/M, female or male.
Note. Values denote mean (standard deviation), and MMSE was lacking for eight AD, one LMCI, one MCI, and nine NC at 3T.

the computation burden. Note that the boxes are located at two fixed locations according to the image edge. Finally, we split this dataset into five groups with similar distribution of diagnosis and magnetic field strength to implement a 5-fold cross-validation policy.

All training MR images ware augmented 8 times by four rotations (90 degrees) and one flip after each rotation. These images were normalized to zero mean and unit variance to accelerate the learning process.

B. EVALUATION METRICS

1) DICE SIMILARITY COEFFICIENT (DSC)

In our experiments, the dice similarity coefficient (DSC) was used as the performance metric, which measures the overlap ratio between the manual labels and automatic segmentation results, and it is computed as follows:

$$\mathrm{DSC} = 2\times\frac{|M \cap A|}{|M| + |A|} \tag{3}$$

where M and A denote the manual and automatic binary segmentation labels, respectively; $|M \cap A|$ is the number of overlap positive elements between M and A; $|M|$ refers to the number of positive values in M.

2) INTRACLASS CORRELATION COEFFICIENT (ICC)

The regression coefficient and the intraclass correlation coefficient (ICC) [54] are used to evaluate the volumetric similarities between the manual labels and automatically segmented results.

3) COHEN'S D-EFFECT SIZE

Cohen's d-effect size is obtained by computing the distance between two normal distributions:

$$d = |\bar{x}_1 - \bar{x}_2|/s \tag{4}$$

$$s = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}} \tag{5}$$

where $\bar{x}_i$ and s_i are mean and standard deviation, respectively. According to the definition for the magnitudes of d suggested by Cohen, the small, medium, and large effect sizes are defined as $d < 0.5$, $0.5 < d < 0.8$, and $d > 0.8$, respectively [55]." Thus, we can compute Cohen's d-effect size based on the automatically and manually segmented volumes, and these metrics will show the sensitivity of each approach in detecting a hippocampal volume difference between different diagnosis groups.

C. EXPERIMENTAL ANALYSIS OF MODEL DESIGN

1) IMPACT OF THE DISTRIBUTION OF SCALE OF THE MODEL LAYERS

As mentioned in Section II-B, the distribution of the scale (receptive field size) of the model layers can affect the contextual knowledge integration that is crucial to achieving good accuracy. Thus, we conducted experiments to investigate the impact of different scale distribution. We constructed five DCFCN models that contain different numbers of downsampling layers or depth. The largest possible theoretical receptive field of each layer in every model is listed in TABLE 3. We conducted several experiments using these models on both the infant and aging datasets.

The experimental results are displayed in Fig. 3. From the figure, we observe that these models show varying performance on the infant datasets and that the 3D-DCFCN43 achieves the best segmentation accuracy. On the aging dataset, the results of these models are similar, and 3D-DCFCN33 has slight advantages. The results indicate that infant hippocampus segmentation is more sensitive to the distribution of hierarchical receptive field size and needs additional local information because the infant brain has a much lower tissue contrast and a much smaller hippocampal size. In addition, we also supply the computation time per image of each model in the figure, and it can be observed that the segmentation speed of 3D-DCFCN43 is much faster than those of 3D-DCFCN25 and 3D-DCFCN33 and similar to those of 3D-DCFCN59 and 3D-DCFCN131. Therefore, to achieve both better segmentation precision and computational efficiency, we chose the scale distribution of 3D-DCFCN43 to conduct all further experiments.

2) IMPORTANCE OF THE JOINT LOSS FUNCTION

We have described the joint loss function for training our networks in Section II-C. The loss function contains three auxiliary losses of cross-entropy and one dice objective function. To show the effectiveness of the dice objective function and auxiliary loss, we ran the same 3D-DCFCN model on the aging dataset using 5 different loss functions: one cross-entropy loss (without auxiliary loss), one dice loss (without auxiliary loss), four cross-entropy losses, four dice losses, and the proposed joint loss. The experimental results are presented in Fig. 4. The figure shows that the joint loss function outperforms all the other loss functions. Surprisingly, the performance of the four cross-entropy losses is better than one whereas the four dice losses are slightly worse than one. In the former case, the results might occur because the auxiliary loss can help the model to integrate more local information to improve the problem of overfitting to the background. In the latter case, because the dice objective

TABLE 3. Architecture and scale distribution of the five constructed 3D-DCFCN models.

Layers		3D-DCFCN131		3D-DCFCN59		3D-DCFCN43		3D-DCFCN33		3D-DCFCN25	
		K/S	TRF	K/S	TRF	K/S	TRF	K/S	K/S	K/S	TRF
Encoder stage	Conv1and Conv2	3/1	3,5	3/1	3,5	3/1	3,5	3/1	3,5	3/1	3,5
	Conv3	3/2	7	3/2	7	3/2	7	3/1	7	3/1	7
	Dense-block1	(3/1)×4	11,15,19, 23	(3/1)×4	11,15,19,23	(3/1)×4	11,15,19,23	(3/1)×4	9,11,13, 15	(3/1)×4	9,11,13, 15
	Conv4	3/2	27	3/2	27	3/1	27	3/2	17	3/1	17
	Dense-block2	(3/1)×4	35,43,51, 59	(3/1)×4	35,43,51,59	(3/1)×4	31,35,39,43	(3/1)×4	21,25,29,33	(3/1)×4	19,21,23,25
	Conv5	3/2	67	-	-	-	-	-	-	-	-
	Dense-block3	(3/1)×4	83,99,115,131	-	-	-	-	-	-	-	-
Decoder stage: Multistream fusion network	Conv Transpose Conv1	3/1	-	3/1	-	3/1	-	3/1	-	3/1	-
	Transpose Conv2	2/2	-	2/2	-	2/2	-	3/1	-	3/1	-
	Transpose Conv3	4/4	-	4/4	-	2/2	-	2/2	-	3/1	-
		8/8	-	-	-	-	-	-	-	-	-

Abbreviations: K/S, kernel size/stride size; TRF, side length of the biggest possible theoretical receptive field (cubic); Conv, convolution; Dense-block, densely connected block.

Note. Values in model name denote largest size of TRF at encoder stage

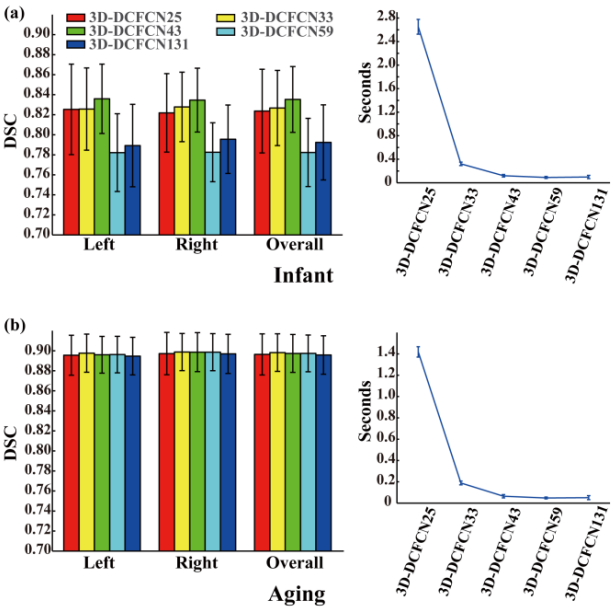


FIGURE 3. Comparison of the five constructed 3D-DCFCN models with different scale distributions on infant and aging datasets. We present both the accuracy (left) and computation time per image (right) of these models.

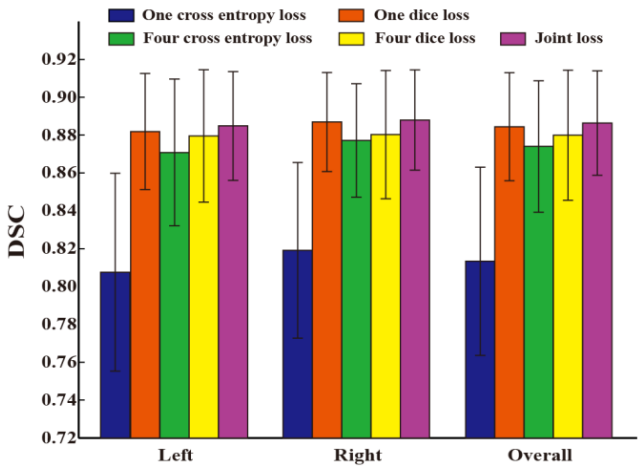


FIGURE 4. Comparison of 3D-DCFCN with five different loss functions on aging dataset.

function can efficiently balance the foreground and background, the auxiliary dice loss might hamper the integration of local information. The proposed loss function contains three auxiliary losses of cross-entropy and one dice loss achieves the best performance.

Besides, we also organized some experiments to investigate the importance of random cropping with proper crop size, the results show that the model gains the best DSC when using 32 × 32 ×32 as crop size. Details can be found in Supplementary Materials and Fig. S1.

D. EXPERIMENTAL RESULTS ON INFANT DATASET

1) IMPACT OF INTEGRATED MULTIMODALITY DATA

The current methods for integrating multimodality information can be categorized into three main classes: (i) the multimodality data are concatenated and fed into one model, and the multimodality information is subsequently fused during forward propagation; (ii) multistream networks are used to process multimodality data separately, and the output feature maps of each stream are subsequently fused to predict the

TABLE 4. Segmentation performance on infant dataset using different methods.

Method	Dataset	DSC		
		Left	Right	Overall
MA with SAE [27]	10 subjects (age at scan: 2 weeks and 3, 6, 9 months)	—	—	0.702 (0.064)
MA [19]	20 subjects (age at scan: 2 weeks and 3, 6, months)	—	—	0.693(0.082)
RF [18]	10 subjects (age at scan: 2 weeks and 3, 6, 9 months)	—	—	0.7206(0.0895)
MAGeT- Brain [56]	197 subjects (168 early in life, 157 term-equivalent)	Early in life: 0.784(0.053) Term equivalent: 0.780(0.814)	Early in life: 0.804(0.036) Term equivalent: 0.814(0.031)	Early in life: 0.794(0.036) Term equivalent: 0.805(0.032)
MA [11]	ALBERT	0.79(0.08)	0.79(0.06)	—
EM [12]	ALBERT	0.783	0.797	—
VoxResNet [57]	ALBERT	0.6439(0.0617)	0.6058(0.0858)	0.6249(0.0763)
3D-DSN [40]	ALBERT	0.7890(0.0639)	0.7820(0.0635)	0.7855(0.0630)
3D-DCFCN	ALBERT	**0.8562(0.0263)**	0.8498(0.0290)	0.8530(0.0275)
3D-DCFCN-BC	ALBERT	0.8536(0.0217)	**0.8579(0.0276)**	**0.8557(0.0246)**

Abbreviations: MA, multi-atlas-based method; SAE, stacked auto-encoders; MAGeT, multiple automatically generated templates; RF, random forest; EM, expectation-maximization algorithm.
Note. Values denote mean (standard deviation)

result; and (iii) multimodality data are processed separately as independent data, and the results of the same hippocampus are averaged to obtain a finer result. To use more training data and construct a lighter-weight algorithm, we choose the third strategy. As shown in Fig. 5, the DSC results of 3D-DCFCN using two modalities show the highest values. Additionally, using T2-weighted images achieves better performance than the case of T1-weighted images. This result indicates that T2-weighted images can supply more hippocampal contour information.

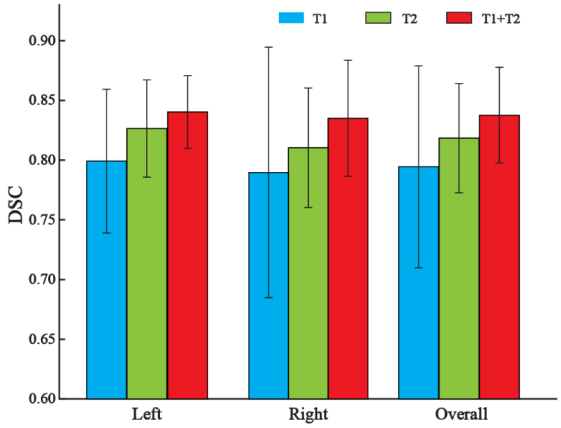


FIGURE 5. DSC results of 3D-DCFCN with respect to different combinations of two imaging modalities.

2) CONSEQUENCE OF THE NEW DATASET AUGMENTATION STRATEGY

As described in Section II-D, we used 12 DOF linear cross-registration to further augment the infant training dataset. The method can vastly augment the number of training samples from n to n^2. We used 3D-DCFCN and 3D-DCFCN-BC to conduct several experiments to reveal the effect of this new dataset augmentation strategy. The experimental results can be found in Fig. 6. As shown in the figure, all of the results with this dataset augmentation strategy outperform those without this technique. We find that this strategy also reduces the performance gap between 3D-DCFCN and 3D-DCFCN-BC, which might indicate that the regularization effect of dataset augmentation is stronger than that of model parameter reduction.

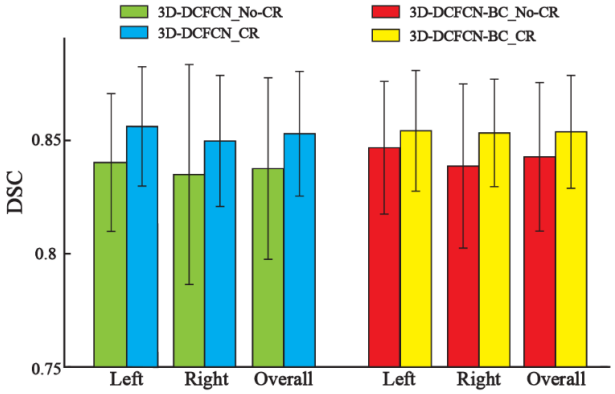


FIGURE 6. Comparison of 3D-DCFCN and 3D-DCFCN-BC with and without the CR (cross-registration) data augmentation strategy.

3) COMPARISON WITH STATE-OF-THE-ART ALGORITHMS

After the optimal scheme was obtained, the proposed 3D-DCFCN and 3D-DCFCN-BC were used to segment the infant hippocampus with the 10-fold cross-validation strategy to present the generalization ability. It takes approximately 4.75 hours to train our designed model on a Titan Xp GPU. We use the dice similarity coefficient (DSC) to evaluate the segmentation performance. The results produced by our methods and different state-of-the-art algorithms can be found in TABLE 4. From TABLE 4 we can see that the 3D-DCFCN-BC achieves the best performance,

TABLE 5. Segmentation performance on elderly dataset using different methods.

Method	Dataset	DSC		
		Left	Right	Overall
LLM [22]	SATA (35)	0.8697(0.0091)	0.8770(0.0176)	0.8734(0.0113)
2D CNN with LSTM [27]	ADNI (110NC)	0.8921(0.0168)	0.8937(0.0146)	0.8929
2D CNN with view ensemble [23]	ADNI (110NC)	0.8948(0.0149)	0.8946(0.0142)	0.8947
Multi task 3D CNN [24]	ADNI (797)	0.898	0.887	0.893(0.013)
MA with RF [21]	EADC-ADNI	0.880(0.020)	0.881(0.020)	—
LML with MA [9]	EADC-ADNI (100)	0.886* (0.028)	0.887* (0.028)	—
ILLM [29]	EADC-ADNI (135)	1.5T: 0.8844(0.0209) 3.0T: 0.8735(0.0335)	1.5T: 0.8853(0.0261) 3.0T: 0.8828(0.0195)	1.5T: 0.8852(0.0203) 3.0T: 0.8783(0.0251)
3D-DSN [40]	EADC-ADNI (135)	0.7400(0.0967)	0.7375(0.0974)	0.7387(0.1009)
VoxResNet [57]	EADC-ADNI (135)	0.8717(0.03238)	0.8667(0.0366)	0.8692(0.0346)
3D-DCFCN	EADC-ADNI (135)	0.8981(0.0189)	0.9003(0.0192)	0.8992(0.0186)
3D-DCFCN-BC	EADC-ADNI (135)	**0.8995(0.0181)**	**0.9008(0.0179)**	**0.9002(0.0180)**

Abbreviations: LLM, local linear mapping; LSTM, long short-term memory; MA, multi-atlas-based method; RF, random forest; ILLM, iterative local linear mapping; LML, local manifold learning.
Note. Values in dataset denote the number of subjects; Values in DSC with or without * denote median (standard deviation) or mean (standard deviation),

yielding a mean DSC of 0.8536 for the left hippocampus, 0.8579 for the right hippocampus, and 0.8557 for the overall hippocampus. We also supply the sagittal views and 3D views of hippocampal segmentation results produced by four automatic segmentation methods for a typical hippocampus in Fig. 7.

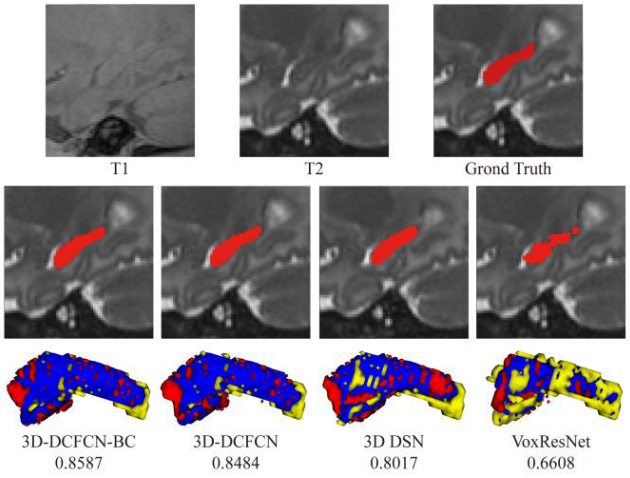


FIGURE 7. Sagittal views and 3D views of hippocampal segmentation results by four different methods for a typical infant subject. In the 3D views, the overlapping areas between ground truth (yellow) and automatic segmentation (red) are shown in blue. The DSC values for the hippocampus are listed below each method.

We additionally executed multiple experiments to analyze the infant hippocampal volumes. The results show that our model has a good volumetric correlation with manually segmented volumes. Detailed information can be seen in Supplementary Materials and Fig. S2, Fig. S3.

E. EXPERIMENTAL RESULTS ON THE AGING DATASET

Similar to the infant dataset, we used the optimal scheme of 3D-DCFCN and 3D-DCFCN-BC to segment the aging hippocampus with the 5-fold cross-validation strategy. This process takes approximately 16.4 hours to train our designed method on a Titan Xp GPU. The performance compared with the state-of-the-art algorithms is reported in TABLE 5. As shown in the table, the proposed 3D-DCFCN and 3D-DCFCN-BC outperform all of the other methods. Specifically, 3D-DCFCN-BC achieves a mean DSC of 0.8995 for the left hippocampus, 0.9008 for the right hippocampus, and 0.9002 for the overall hippocampus.

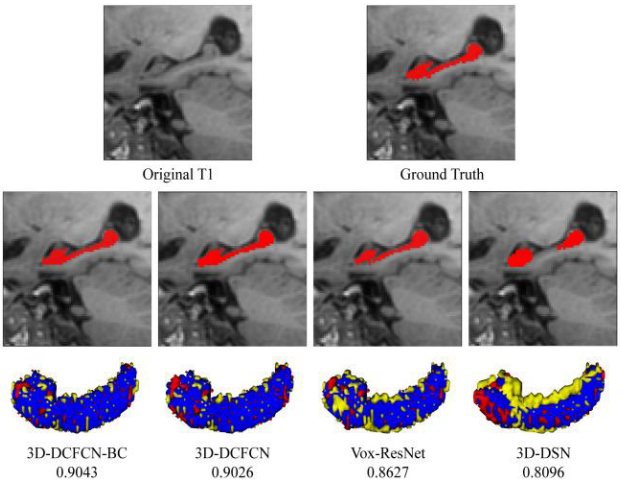


FIGURE 8. Sagittal and 3D views of hippocampal segmentation results on four different methods for a typical subject in aging dataset. In the 3D views, the overlapping areas between ground truth (yellow) and automatic segmentation (red) are shown in blue. The DSC values for the hippocampus are listed below each method.

We also produce a qualitative illustration of the advantage of our method on this dataset. Fig. 8 show the sagittal views and 3D views of the hippocampal segmentation results predicted by four automatic segmentation methods for a typical hippocampus.

We further conduct several experiments for statistical analysis of hippocampal volume and validation the power of that volume as a cognitive-decline biomarker. The results, which can be found in the *Supplementary Materials* and Fig. S4, Fig. S5, Fig. S6, and TABLE S1, show that our methods can satisfactorily reveal hippocampal atrophy in cognitive-decline groups compared with normal controls.

F. GENERALIZATION ACROSS DATASETS WITH DIFFERENT MAGNETIC FIELDS

We further evaluate the generalization capacity of our method. We first split the aging dataset into two groups with different magnetic fields. Second, we train our models on one group and test it on the other one, and vice versa. The results in terms of DSC are displayed in TABLE 6, and it shows that our method could better generalize across the two different magnetic fields.

TABLE 6. Generalization performance across elderly datasets with different magnetic field.

		DSC		
		Left	Right	Overall
1.5T→3.0 T	LML with MA [9]	0.880(0.029)	0.880(0.022)	—
	3D-DCFCN	**0.8882(0.0226)**	**0.8928(0.0189)**	**0.8905(0.0209)**
	3D-DCFCN-BC	**0.8899(0.0209)**	**0.8938(0.0173)**	**0.8918(0.0192)**
3.0T→1.5 T	LML with MA [9]	0.877(0.045)	0.884(0.046)	—
	3D-DCFCN	0.8898(0.0200)	0.8882(0.0226)	0.8890(0.0213)
	3D-DCFCN-BC	**0.8914(0.0199)**	**0.8891(0.0292)**	**0.8902(0.0249)**

Abbreviations: LML, local manifold learning.
Note. Values in DSC denote mean (standard deviation), respectively.

G. TIME, STORAGE COST AND MODEL COMPLEXITY

The computation time of model predicting takes a more important role in evaluation of its performance than offline training time. TABLE 7 shows the parameters, forward time per image, and memory of our models. Note that our method was trained on a workstation with the following settings: memory: 8 × 16 GB DDR4; GPU: Nvidia Titan Xp; CPU: Xeon E5-2667v4; and operating system: Ubuntu 16.04.

Compared with the other methods of TABLE 5 and TABLE 4, where [9] reported a requirement of 40 seconds per image, and [29] required approximately 4 min, our proposed method is at least 40 times faster than these methods.

H. LIMITATION OF OUR METHOD

One of the main limitations of this study is that it shows slightly over-segmentation of the hippocampal head and under-segmentation of the hippocampal tail, perhaps because the different levels of semantic information are not sufficiently fused. The process might be improved by incorporating denser connected networks. Additionally, the coupling effects of various factors in the model design are not considered. Therefore, our method might not be able to exploit a more detailed engineering design of the model on various tasks. Finally, the methods need to be tested in a larger cohort. Further research will be undertaken to investigate these issues.

TABLE 7. Parameters, forward time per image and memory of our models.

Dataset	Method	Parameters	Forward time per image	Memory (GPU)
infant	3D-DCFCN	334.24 KB	0.1111s(0.0114)	**3623 MB**
	3D-DCFCN-BC	**293.74 KB**	**0.1083s**(0.0097)	3637 MB
elderly	3D-DCFCN	same as above	0.0630s(0.0036)	**4446 MB**
	3D-DCFCN-BC	same as above	**0.0621s**(0.0032)	4463 MB

Note: Values in forward time per image denote mean (standard deviation)

IV. CONCLUSION

This study proposes the 3D-DCFCN and 3D-DCFCN-BC techniques for preterm infant and aging hippocampus segmentation of MR brain images. Our methods have a much faster, more accurate, and stable performance than previous methods. We incorporate the densely connected block to make the model more adaptive to the two segmentation tasks. We also propose a suitable distribution of the hierarchical receptive field size to balance the local information and global information as well as obtain a small enough computational cost. Additionally, a joint loss function was presented to improve optimization. Furthermore, we integrate multimodality MR images and present a new dataset augmentation technique to improve the precision and robustness on the infant dataset. By comparing our algorithm with the state-of-the-art approaches, we can see that our method outperforms all comparison approaches both in accuracy and speed on the two main datasets. Moreover, our method can satisfactorily find hippocampal atrophy in cognitive-decline groups compared with normal controls. Overall, this study provides a precise and efficient hippocampus segmentation method and a notably comprehensive evaluation of that.

REFERENCES

[1] D. K. Thompson, S. J. Wood, L. W. Doyle, S. K. Warfield, G. A. Lodygensky, P. J. Anderson, G. F. Egan, and T. E. Inder, "Neonate hippocampal volumes: Prematurity, perinatal predictors, and 2-year outcome," *Ann. Neurol.*, vol. 63, no. 5, pp. 642–651, May 2008.
[2] M. H. Beauchamp, D. K. Thompson, K. Howard, L. W. Doyle, G. F. Egan, T. E. Inder, and P. J. Anderson, "Preterm infant hippocampal volumes correlate with later working memory deficits," *Brain*, vol. 131, no. 11, pp. 2986–2994, Nov. 2008.
[3] C. S. H. Aarnoudse-Moens, H. J. Duivenvoorden, N. Weisglas-Kuperus, J. B. Van Goudoever, and J. Oosterlaan, "The profile of executive function in very preterm children at 4 to 12 years," *Develop. Med. Child Neurol.*, vol. 54, no. 3, pp. 247–253, Nov. 2011.

[4] C. E. Rogers, P. J. Anderson, D. K. Thompson, H. Kidokoro, M. Wallendorf, K. Treyvaud, G. Roberts, L. W. Doyle, J. J. Neil, and T. E. Inder, "Regional cerebral development at term relates to school-age social–emotional development in very preterm children," *J. Amer. Acad. Child Adolescent Psychiatry*, vol. 51, no. 2, pp. 181–191, Feb. 2012.

[5] H. Braak and E. Braak, "Neuropathological stageing of alzheimer-related changes," *Acta Neuropathologica*, vol. 82, no. 4, pp. 239–259, Sep. 1991.

[6] M. Fotuhi, D. Do, and C. Jack, "Modifiable factors that alter the size of the hippocampus with ageing," *Nature Rev. Neurol.*, vol. 8, pp. 189–202, Mar. 2012.

[7] J. Miller, A. J. Watrous, M. Tsitsiklis, S. A. Lee, S. A. Sheth, C. A. Schevon, E. H. Smith, M. R. Sperling, A. Sharan, A. A. Asadi-Pooya, G. A. Worrell, S. Meisenhelter, C. S. Inman, K. A. Davis, B. Lega, P. A. Wanda, S. R. Das, J. M. Stein, R. Gorniak, and J. Jacobs, "Lateralized hippocampal oscillations underlie distinct aspects of human spatial memory and navigation," *Nature Commun.*, vol. 9, no. 1, p. 2423, Jun. 2018.

[8] L. Wang, D. Mamah, M. P. Harms, M. Karnik, J. L. Price, M. H. Gado, P. A. Thompson, D. M. Barch, M. I. Miller, and J. G. Csernansky, "Progressive deformation of deep brain nuclei and hippocampal-amygdala formation in schizophrenia," *Biol. Psychiatry*, vol. 64, no. 12, pp. 1060–1068, Dec. 2008.

[9] X.-W. Li, Q.-L. Li, S.-Y. Li, and D.-Y. Li, "Local manifold learning for multiatlas segmentation: Application to hippocampal segmentation in healthy population and Alzheimer's disease," *CNS Neurosci. Therapeutics*, vol. 21, no. 10, pp. 826–836, Oct. 2015.

[10] V. Dill, P. C. Klein, A. R. Franco, and M. S. Pinho, "Atlas selection for hippocampus segmentation: Relevance evaluation of three meta-information parameters," *Comput. Biol. Med.*, vol. 95, pp. 90–98, Apr. 2018.

[11] I. S. Gousias, A. Hammers, S. J. Counsell, L. Srinivasan, M. A. Rutherford, R. A. Heckemann, J. V. Hajnal, D. Rueckert, and A. D. Edwards, "Magnetic resonance imaging of the newborn brain: Automatic segmentation of brain images into 50 anatomical regions," *PLoS ONE*, vol. 8, no. 4, Apr. 2013, Art. no. e59990.

[12] A. Makropoulos, I. S. Gousias, C. Ledig, P. Aljabar, A. Serag, J. V. Hajnal, A. D. Edwards, S. J. Counsell, and D. Rueckert, "Automatic whole brain MRI segmentation of the developing neonatal brain," *IEEE Trans. Med. Imag.*, vol. 33, no. 9, pp. 1818–1831, Sep. 2014.

[13] M. Hajiesmaeili, J. Dehmeshki, B. B. Nakhjavanlo, and T. Ellis, "Initialisation of 3D level set for hippocampus segmentation from volumetric brain MR images," in *Proc. 6th Int. Conf. Digit. Image Process. (ICDIP)*, 2014, Art. no. 91591D.

[14] S. Hu, P. Coupé, J. C. Pruessner, and D. L. Collins, "Nonlocal regularization for active appearance model: Application to medial temporal lobe segmentation," *Hum. Brain Mapping*, vol. 35, no. 2, pp. 377–395, Feb. 2014.

[15] M. Chupin, A. Hammers, R. S. N. Liu, O. Colliot, J. Burdett, E. Bardinet, J. S. Duncan, L. Garnero, and L. Lemieux, "Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: Method and validation," *NeuroImage*, vol. 46, no. 3, pp. 749–761, Jul. 2009.

[16] Y. Gao, B. Corn, D. Schifter, and A. Tannenbaum, "Multiscale 3D shape representation and segmentation with applications to hippocampal/caudate extraction from brain MRI," *Med. Image Anal.*, vol. 16, no. 2, pp. 374–385, Feb. 2012.

[17] T. Tong, R. Wolz, P. Coupé, J. V. Hajnal, and D. Rueckert, "Segmentation of MR images via discriminative dictionary learning and sparse coding: Application to hippocampus labeling," *NeuroImage*, vol. 76, pp. 11–23, Aug. 2013.

[18] L. Zhang, Q. Wang, Y. Gao, H. Li, G. Wu, and D. Shen, "Concatenated spatially-localized random forests for hippocampus labeling in adult and infant MR brain images," *Neurocomputing*, vol. 229, pp. 3–12, Mar. 2017.

[19] Y. Guo, G. Wu, P. T. Yap, V. Jewells, W. Lin, and D. Shen, "Segmentation of infant hippocampus using common feature representations learned for multimodal longitudinal data," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, vol. 9351, Oct. 2015, pp. 63–71.

[20] A. van Opbroek, H. C. Achterberg, M. W. Vernooij, M. A. Ikram, and M. de Bruijne, "Transfer learning by feature-space transformation: A method for hippocampus segmentation across scanners," *NeuroImage, Clin.*, vol. 20, pp. 466–475, Jan. 2018.

[21] Q. Zheng and Y. Fan, "Integrating semi-supervised label propagation and random forests for multi-atlas based hippocampus segmentation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Oct. 2018, pp. 154–157.

[22] S. Pang, J. Jiang, Z. Lu, X. Li, W. Yang, M. Huang, Y. Zhang, Y. Feng, W. Huang, and Q. Feng, "Hippocampus segmentation based on local linear mapping," *Sci. Rep.*, vol. 7, no. 1, Apr. 2017, Art. no. 45501.

[23] L. Cao, L. Li, J. Zheng, X. Fan, F. Yin, H. Shen, and J. Zhang, "Multi-task neural networks for joint hippocampus segmentation and clinical score regression," *Multimedia Tools Appl.*, vol. 77, no. 22, pp. 29669–29686, Nov. 2018.

[24] B. Thyreau, K. Sato, H. Fukuda, and Y. Taki, "Segmentation of the hippocampus by transferring algorithmic knowledge for large cohort processing," *Med. Image Anal.*, vol. 43, pp. 214–228, Jan. 2018.

[25] M. Kim, G. Wu, and D. Shen, "Unsupervised deep learning for hippocampus segmentation in 7.0 Tesla MR images," in *Proc. Int. Workshop Mach. Learn. Med. Imag. (MLMI)*, vol. 8184, 2013, pp. 1–8.

[26] Y. Guo, G. Wu, L. A. Commander, S. Szary, V. Jewells, W. Lin, and D. Shen, "Segmenting hippocampus from infant brains by sparse patch matching with deep-learned features," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, vol. 8674, 2014, pp. 308–315.

[27] Y. Chen, B. Shi, Z. Wang, T. Sun, C. D. Smith, and J. Liu, "Accurate and consistent hippocampus segmentation through convolutional LSTM and view ensemble," in *Proc. Int. Workshop Mach. Learn. Med. Imag. (MLMI)*, vol. 10541, 2017, pp. 88–96.

[28] Y. Chen, B. Shi, Z. Wang, P. Zhang, C. D. Smith, and J. Liu, "Hippocampus segmentation through multi-view ensemble ConvNets," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 192–196.

[29] S. Pang, Q. Feng, Z. Lu, J. Jiang, L. Zhao, L. Lin, X. Li, T. Lian, M. Huang, and W. Yang, "Hippocampus segmentation based on iterative local linear mapping with representative and local structure-preserved feature embedding," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2271–2280, Oct. 2019.

[30] M. Goubran, E. E. Ntiri, H. Akhavein, M. Holmes, S. Nestor, J. Ramirez, S. Adamo, M. Ozzoude, C. Scott, F. Gao, A. Martel, W. Swardfager, M. Masellis, R. Swartz, B. MacIntosh, and S. E. Black, "Hippocampal segmentation for brains with extensive atrophy using three-dimensional convolutional neural networks," *Hum. Brain Mapping*, vol. 41, no. 2, pp. 291–308, Feb. 2020.

[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

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

[35] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's disease neuroimaging initiative (ADNI)," *Alzheimer's Dementia*, vol. 1, no. 1, pp. 55–66, Jul. 2005.

[36] M. Boccardi, M. Bocchetta, F. C. Morency, D. L. Collins, M. Nishikawa, R. Ganzola, M. J. Grothe, D. Wolf, A. Redolfi, M. Pievani, L. Antelmi, A. Fellgiebel, H. Matsuda, S. Teipel, S. Duchesne, C. R. Jack, and G. B. Frisoni, "Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol," *Alzheimer's Dementia*, vol. 11, no. 2, pp. 175–183, Feb. 2015.

[37] M. Boccardi *et al.*, "Delphi definition of the EADC-ADNI harmonized protocol for hippocampal segmentation on magnetic resonance," *Alzheimer's Dementia*, vol. 11, no. 2, pp. 126–138, Feb. 2015.

[38] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. Artif. Intell. Statist. (AISTATS)*, 2015, pp. 562–570.

[39] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

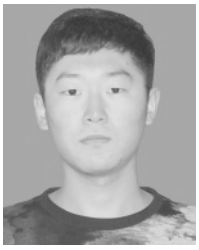
[40] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, and P.-A. Heng, "3D deeply supervised network for automated segmentation of volumetric medical images," *Med. Image Anal.*, vol. 41, pp. 40–54, Oct. 2017.

[41] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Artif. Intell. Statist. (AISTATS)*, 2010, pp. 249–256.

[42] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization? (no, it is not about internal covariate shift)," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2018, pp. 1–15.
[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.
[44] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," 2016, *arXiv:1603.07285*. [Online]. Available: http://arxiv.org/abs/1603.07285
[45] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[46] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, Oct. 2016.
[47] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.
[48] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11211, 2018, pp. 833–851.
[49] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.
[50] J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart augmentation learning an optimal data augmentation strategy," *IEEE Access*, vol. 5, pp. 5858–5869, 2017.
[51] J. Lemley, S. Bazrafkan, and P. Corcoran, "Deep learning for consumer devices and services: Pushing the limits for machine learning, artificial intelligence, and computer vision.," *IEEE Consum. Electron. Mag.*, vol. 6, no. 2, pp. 48–56, Apr. 2017.
[52] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.
[53] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," 2014, *arXiv:1408.5093*. [Online]. Available: http://arxiv.org/abs/1408.5093
[54] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychol. Bull.*, vol. 86, no. 2, p. 420, Mar. 1979.
[55] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Evanston, IL, USA: Routledge, 2013.
[56] T. Guo, J. L. Winterburn, J. Pipitone, E. G. Duerden, M. T. M. Park, V. Chau, K. J. Poskitt, R. E. Grunau, A. Synnes, S. P. Miller, and M. M. Chakravarty, "Automatic segmentation of the hippocampus for preterm neonates from early-in-life to term-equivalent age," *NeuroImage, Clin.*, vol. 9, pp. 176–193, Aug. 2015.
[57] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng, "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images," *NeuroImage*, vol. 170, pp. 446–455, Apr. 2017.

DEBIN ZENG received the B.S. degree in spacecraft design and engineering and the M.S. degree in biomedical engineering from Beihang University, Beijing, China, in 2016 and 2019, respectively, where he is currently pursuing the Ph.D. degree with the School of Biological Science and Medical Engineering. His research interests include neuroimage analysis, machine learning, and network neuroscience.

QIONGLING LI received the B.S. degree in biomedical engineering from Tianjin Medical University, Tianjin, China, in 2013. She is currently pursuing the Ph.D. degree with the School of Biological Science and Medical Engineering, Beihang University, Beijing, China. Her research interests include network neuroscience, brain plasticity, hippocampus, and multimodal MRI.



BAOQIANG MA received the B.S. degree in biomedical engineering from Northeastern University, Shenyang, China, in 2017, and the M.S. degree in biomedical engineering from Beihang University, Beijing, China, in 2020. His research interests include medical image analysis, machine learning, and artificial intelligence.

SHUYU LI received the B.S. degree in biomedical engineering from Capital Medical University, Beijing, China, in 1998, and the Ph.D. degree in biophysics from the Institute of Biophysics, Chinese Academy of Sciences, China, in 2003. From 2003 to 2005, she held a postdoctoral position with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. She is currently a Professor with the School of Biological Science and Medical Engineering, Beihang University, Beijing. Her research interests can be divided into three major areas, including neuroimage analysis, progression of Alzheimer's disease, and adolescent brain structural and functional development.