IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 8, 2020, accepted March 21, 2020, date of publication March 31, 2020, date of current version April 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2984661

Fusion of Multi-Size Candidate Regions Enhances Two-Stage Hippocampus Segmentation

PING CAO[1,2], QIUYANG SHENG[2,5], SIQI FANG[3], XINYI LI[2], GANGMIN NING[4,6], (Senior Member, IEEE), AND QING PAN[2], (Member, IEEE)

[1]Zhijiang College, Zhejiang University of Technology, Shaoxing 312030, China
[2]College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China
[3]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, CB3 OWA, U.K.
[4]Department of Biomedical Engineering, Zhejiang University, Hangzhou 310027, China
[5]Deepwise AI Laboratory, Hangzhou 310000, China
[6]Zhejiang Laboratory, Hangzhou 310000, China

Corresponding author: Qing Pan (pqpq@zjut.edu.cn)

This work was supported in part by the Zhejiang Province Key Research and Development Program under Grant 2020C03073, and in part by the National Natural and Science Foundation of China under Grant 31870938 and Grant 81871454.

ABSTRACT The hippocampus plays an important role in the memory and cognition abilities of humans. Precise three-dimensional (3D) segmentation of the hippocampus from magnetic resonance imaging scans is of great importance in the diagnosis of neurological diseases. Conventional automatic segmentation methods poorly achieve satisfactory performance because of the irregular shape and small volume of the hippocampus. We propose a novel two-stage segmentation method, which includes a localization stage and a segmentation stage, to handle the task of the 3D segmentation of the hippocampus. In the localization stage, a novel strategy for localizing multi-size candidate regions was developed to improve the sample balance for the 3D segmentation task. In the segmentation stage, a method which fuses the multi-size candidate regions was proposed to improve the accuracy in predicting the hippocampal boundary, after which we aggregated the segmentation results from three orthogonal views to further improve the performance. Quantitative evaluation was performed on the Alzheimer's Disease Neuroimaging Initiative dataset. The experimental results achieved Dice similarity coefficients of 92.48 ± 0.61% and 92.90 ± 0.51% for the left and right hippocampus, respectively, outperforming state-of-the-art studies in hippocampus segmentation tasks.

INDEX TERMS Hippocampus segmentation, hippocampus localization, fully convolutional network, two-stage segmentation.

I. INTRODUCTION

The hippocampus, located between the thalamus and medial temporal lobes, has an important influence on the memory and cognition abilities of human. Many studies report that patients with Alzheimer's disease [1], schizophrenia [2], or major depression [3] develop symptoms of changes in the morphology of the hippocampus. Morphological analysis of the hippocampus, which relies on the precise segmentation of the hippocampus in magnetic resonance imaging (MRI) scans, facilitates the diagnosis of related neurological diseases noninvasively [4]. Manual segmentation is generally regarded as the gold standard for morphological analysis and evaluation [5], but its clinical application is limited by the high requirements for professionals [6]. Automatic segmentation shall bring substantial gains to this field and is thereby highly necessary.

The hippocampus occupies a small volume on the MRI scans. The performance of segmentation methods based on fully convolutional networks (FCNs) [7] may be compromised by the imbalance between the organ voxels and the background voxels. Some two-stage methods address this problem by locating the candidate regions in the first stage and then performing segmentation within these candidate regions in the second stage. In the candidate regions, the number of positive and negative voxel samples are relatively balanced, which helps improve the segmentation performance. Existing two-stage methods [8]–[10] simply use fixed-size

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

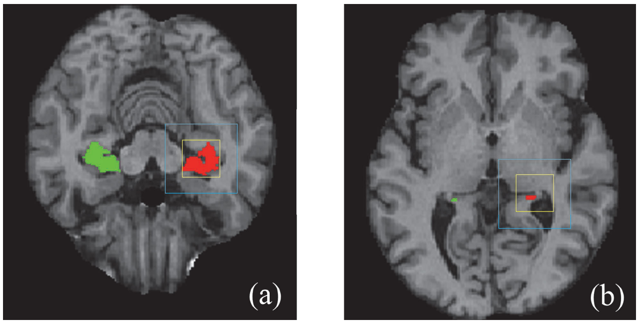

FIGURE 1. Schematic diagram of organ morphological differences in different slices. (a) An intermediate slice with a large hippocampus region. (b) A near-end slice with a small hippocampus region. The red region is the left hippocampus and the green region is the right hippocampus. In large candidate regions (blue box), sample imbalance problems are better resolved for large hippocampus regions than for small hippocampus regions. The opposite holds in small candidate regions (yellow box).

candidate regions. However, as the morphology of the hippocampus varies along consecutive slices, it is difficult to keep the positive and negative samples balanced in each slice using a fixed-size candidate region strategy. As shown in Fig. 1, a small candidate region does not include sufficient global features of large objects but is good at representing features of small objects. In contrast, a large candidate region contains more global features of an object that matches its region size but weakly represents the local features of smaller objects.

To overcome this challenge, we propose a novel two-stage hippocampus segmentation framework. This framework combines the essences of multi-view FCNs and localization-based coarse segmentation. The localization stage aims to reduce non-hippocampus areas and generate candidate regions covering the hippocampus. The segmentation stage aims to produce a segmentation with high precision within the candidate regions.

To keep positive and negative voxel samples balanced in a majority of the consecutive slices and obtain features matching the morphology of the hippocampus, our framework generates multiple candidate regions of different sizes for every hippocampus slice. Multi-size candidate regions provide a variety of conditions for sample balance. Larger candidate regions are beneficial for segmenting larger hippocampus areas, while smaller candidate regions are beneficial for segmenting smaller hippocampus areas. Segmentation of larger areas requires more features surrounding the organ as classification evidence, while fewer are employed for segmentation of smaller areas in order to reduce the risk of misjudgment. We perform segmentation operations within multi-size candidate regions and fuse the results to make full use of their complementary advantages. Specifically, we determine the approximate centroid of hippocampus regions in every MRI slice in the localization stage by coarse segmentation. With the centroid as the center, bounding boxes of multiple sizes covering the hippocampus are determined. Next, the candidate regions defined by the bounding boxes are fed into the pretrained FCNs. Then, the corresponding segmentation outputs are fused to achieve greater segmentation precision. This process behaves like an ensemble of multiple deep classifiers, where each deep classifier in the ensemble makes predictions under different surroundings, and their fusion enhances segmentation performance. The above operations are performed on axial view, coronal view and sagittal view one by one. Finally, the outputs from these orthogonal views are fused once more to enhance the segmentation performance further by majority voting (MV).

The rest of this paper is organized as follows. An overview of the related work is described in II. RELATED WORKS. We describe the technical motivation and details of the proposed approach in III. METHOD. Analysis of the experimental results is provided in IV. EXPERIMENT & RESULTS. Finally, extended discussions and conclusions are given in V. DISCUSSION and VI. CONCLUSIONS, respectively.

II. RELATED WORKS

The long-established hippocampus segmentation methods are usually based on three types of techniques: (1) conventional image processing; (2) atlas registration; and (3) machine learning. For a more detailed discussion of the related works, readers can refer to comprehensive reviews of hippocampus segmentation [11], [12].

Methods based on conventional image processing usually require the user to participate in the initialization of the segmentation process. For example, thresholding [13] and region growing [14] methods require manual seed selection. Deformable model techniques [15], [16] require contour placement by the user. However, these methods are limited in terms of practical application on account of the high dependence on human-computer interaction, which are subjective and nonreproducible.

The representative atlas registration method is the multi-atlas segmentation (MAS) approach [17]–[19]. The performance of the MAS relies on both registration accuracy and the label fusion (LF) strategy. Consequently, in addition to optimizing the registration, many researchers focus on exploring more effective LF strategies. Recently, the patch-based LF strategy [20] has received much attention, and numerous improvement methods, e.g., the nonlinear [21], set partition strategy [22] and template local ranking strategy [23], have arisen. However, the performance of the MAS is sensitive to the atlas selection [24], [25], which reduces its applicability in practical applications.

Methods based on machine learning usually combine some local features (e.g., image intensities, gradients [26], and Gaussian features [27]) and use a classifier (e.g., k-nearest neighbor [27], random forest [28], Bayesian classifier [29], and support vector machine (SVM) [26]) to segment the hippocampus. Combined with MAS, some machine learning technologies, e.g., SVM [30] and dictionary learning [31], [32], have been used to determine the optimal weights of the LF or compile patches for more efficient similarity matching. Some object-oriented segmentation methods have hierarchical characteristics. First, object description is realized based on local features, and then object segmentation is realized by constructing inter-object relations. In the field of brain segmentation, the relationships among brain tissues can be constructed by Bayesian law [33] or tree-like structures [34]. These methods often use a number of handcrafted features as the evidence of classification or for weight calculation. However, these handcrafted features may not reflect the general morphology of the hippocampus. In addition, their performance is largely dependent on heuristic tuning parameters and empirical pretreatment techniques, thus limiting their generality.

The emerging convolutional neural network (CNN), one of a number of deep learning techniques, has contributed to the sphere of medical image processing greatly in recent years [35]–[37]. It benefits from the fact that the end-to-end learning of salient feature representations maybe more effective than that of handcrafted features with heuristic tuning parameters [38]. FCN [7] (i.e., a CNN in which all layers are convolution layers), in which both learning and inference are performed whole-image-at-a-time, has played an increasingly important role in the field of semantic segmentation. FCN and its variants, including U-Net [39], DeepLab [40], etc., have demonstrated promising performances in many semantic segmentation tasks [41], [42]. However, limited by their architectures (i.e., they can only be applied to 2D image segmentation), these networks are unable to make full use of the 3D context, hampering their direct application in hippocampus segmentation to some extent.

Several recent studies have used FCNs directly with 3D convolution kernels to segment the prostate [43], liver [44], lung tumors [45], and brain [46]. Three-dimensional FCN has attracted much attention in organ segmentation on MRI because it can take full advantage of the 3D context. However, compared with 2D FCNs, 3D FCNs require a large number of parameters and an enormous computational cost. In addition, its training phase also boasts a high requirement for training data. Multi-view FCNs, which combine decisions from multiple 2D views (i.e., axial view, coronal view, and sagittal view), are effective alternatives for making full use of the 3D context. In the processing of tissues with complex morphologies, e.g., pulmonary nodule detection [47], [48], knee cartilage segmentation [49], and pancreas segmentation [8], 2D CNN models may be more generalizable than their 3D counterparts.

Hippocampus segmentation methods based on multi-view FCN has been unfolding [50], [51] recently. To counteract the inherent voxel imbalance, some approaches [50], [51] focus on segmentation in a smaller hand-selected space to further improve precision, yet their practical application may be limited by human-computer interaction. Two-stage methods provide various ways to automate the hand-selected operation and generate candidate regions in the localization stage. In a pancreas segmentation task, localization was implemented through coarse segmentation based on the super-pixel technique [9] or an FCN [8], [10].

III. METHOD

A. OVERVIEW OF THE FRAMEWORK

The proposed framework includes two stages: localization and segmentation. The outline of the proposed framework is shown in Fig. 2. The localization stage consists of two steps: (1) slice of interest (SOI) fetch; and (2) candidate region generation. Slices containing the hippocampus are defined as SOIs. In the first step, a Bi-LeNet model is applied to fetch t SOIs from hippocampus MRI scans in one view. Based on the t SOIs, k groups of candidate regions are then generated in the second step. Every group consists of t candidate regions of a specific size. The groups of candidate regions are denoted as $\{\{x_1, x_2, \cdots, x_t\}_1, \cdots, \{x_1, x_2, \cdots, x_t\}_k\}^v$, where $v \in \{A, C, S\}$, A represents an axial view, C represents a coronal view and S represents a sagittal view. The localization operation is performed in the three views.

The segmentation stage consists of two steps: (1) planar fusion; and (2) multi-view decision. In the planar fusion stage, we use modified U-Nets (hereafter referred to as U-Net) to take k groups of candidate regions as input to produce probability maps $\{\{h_1, h_2, \cdots, h_t\}_1, \cdots, \{h_1, h_2, \cdots, h_t\}_k\}^v$ with k groups. Then, the k groups of probability maps are fused to produce a segmentation mask sequence $Y^v = \{y_1, y_2, \cdots, y_t\}^v$ per view. The above operations are carried out in the axial view, coronal view and sagittal view. All the slices are processed one by one in their original order, and three 2D mask sequences are generated for the different views. In the second step, we aggregate the mask sequences of the three views to make the final decision for 3D hippocampus segmentation Υ.

In the localization stage, three Bi-LeNets and three U-Nets are employed, while $3k$ U-Nets are used in the segmentation stage. All the nets are trained individually.

Details regarding Bi-LeNet are described in Sec.B1. And the architecture of the U-Nets is described in detail in Sec.B2.

B. NETWORKS

1) BI-LENET

The proposed Bi-LeNet was modified from LeNet-5 [52], which was designed for handwritten digit recognition. The Bi-LeNet model contains two convolution-pooling layers, as shown in Fig. 3. The first convolution layer has 6 channels with a 5 × 5 kernel. The second convolution layer has 16 channels with a 5 × 5 kernel. The activation functions are both rectified linear units (ReLU). In addition, the two max-pooling layers are 2 × 2 windows. There are two fully connection layers, one with 120 and the other with 84 neurons. The final output layer implements the softmax activation function and 2 output channels. In the training phase of Bi-LeNet, binary cross entropy is employed as the loss function.

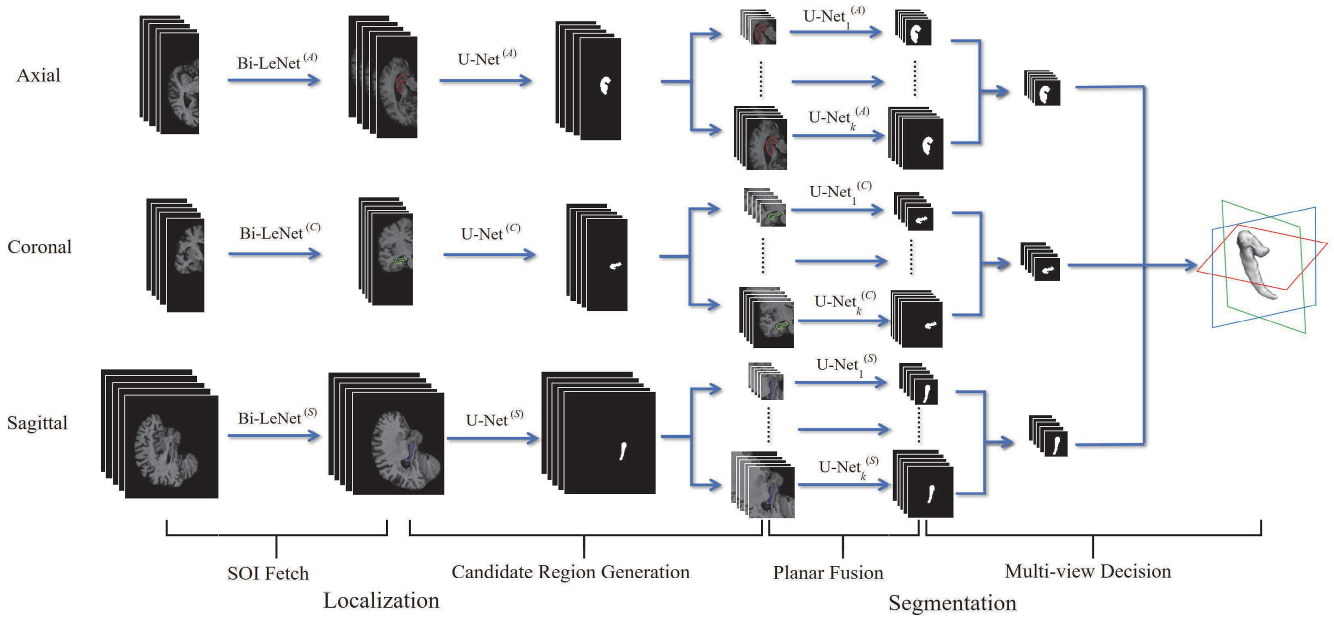

FIGURE 2. **Overview of the proposed framework. In the SOI Fetch step, MRI slices containing the hippocampal region are determined using Bi-LeNets. Multiple groups of candidate regions of the SOIs are determined during the Candidate Region Generation step with U-Nets. The superscripts next to the networks mark the corresponding orthogonal plane. In the Planar Fusion step, segmentation and fusion on multiple groups of candidate regions are performed using their respective U-Nets (indicated by the subscript). All the above steps are repeated in axial, coronal and sagittal views. Finally, the output segmentation results from the three views are aggregated to generate a final 3D segmentation in the Multi-view Decision step.**

2) MODIFIED U-NET

The modified U-Net is based on the classical encoder and decoder architecture of U-Net [39], as shown in Fig. 4. The encoder, following the typical architecture of a convolutional network, encodes the image features. The convolution layer has a 3×3 kernel and a stride of 1. Down-sampling, realized by a max-pooling layer with a 2×2 window and no stride, reduces the resolution of the feature map and enlarges the receptive field of the network.

To keep the dimensions of the input and output of the U-Net consistent, all feature maps are processed by zero-padding. To perform internal covariate shift alleviation, training acceleration and overfitting lessening, batch normalization (BN) is included into both the encoder and decoder. BN occurs between the convolution layer and activation layer, i.e., ReLU, to accelerate the network training.

A skip architecture, which concatenates the feature maps from different depths in the encoder with the corresponding reconstructed resolutions from the decoder, is included as part of the decoder to achieve feature fusion. Generally, the features extracted by the convolution kernels at different depths of the encoder network have distinct characteristics. The low-depth convolution kernel extracts low-level features with high-resolution but low semantic information, while the high-depth convolution kernel extracts high-level features with low-resolution and high semantic information. Combining the features from different depths helps to refine the prediction [39]. In addition, all network parameters are initialized according to the rectifier method [53].

As a frequently used evaluation metric in medical image segmentation, the Dice similarity coefficient (DSC) [54] measures the overlap between the predicted segmentation and the ground truth:

$$DSC(A_P, A_G) = \frac{2 \times |A_P \cap A_G|}{|A_P| + |A_G|} \times 100\% \tag{1}$$

where A_P is the predicted segmentation mask, and A_G is the ground truth.

The Dice loss, used as the loss function in our U-Net training, is defined as follows,

$$C = 1 - \frac{2 \times \sum_{i=1}^{n} p_i g_i}{\sum_{i=1}^{n} p_i + \sum_{i=1}^{n} g_i} \tag{2}$$

where $p_i, g_i \in \{0, 1\}$ and p_i is value of pixel i in the predicted segmentation mask P, while g_i is the true value of pixel i in the ground truth mask G.

The gradient of the Dice loss can be computed as:

$$\frac{\partial C}{\partial p_j} = -2 \times \frac{g_j(\sum_{i=1}^{n} p_i + \sum_{i=1}^{n} g_i) - \sum_{i=1}^{n} p_i g_i}{(\sum_{i=1}^{n} p_i + \sum_{i=1}^{n} g_i)^2} \tag{3}$$

For network training, stochastic gradient descent (SGD) is applied, and the learning rate is initialized to 0.1 and decreases as the training proceeds. Spatial dropout is applied in training phase to avoid overfitting.

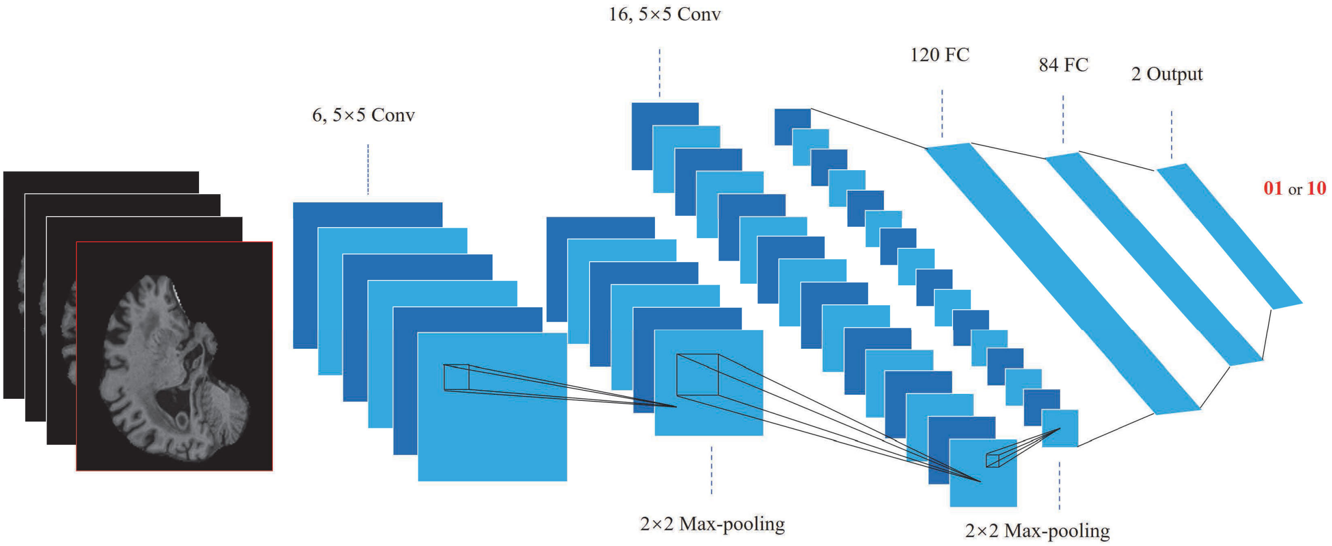

FIGURE 3. Schematic diagram of the architecture and parameters of the proposed Bi-LeNet. It mainly consists of two convolutional layers, two max-pooling layers, two fully connection layers and one output layer. Bi-LeNet takes an MRI slice and predicts whether it contains a hippocampus region (01 or 10).

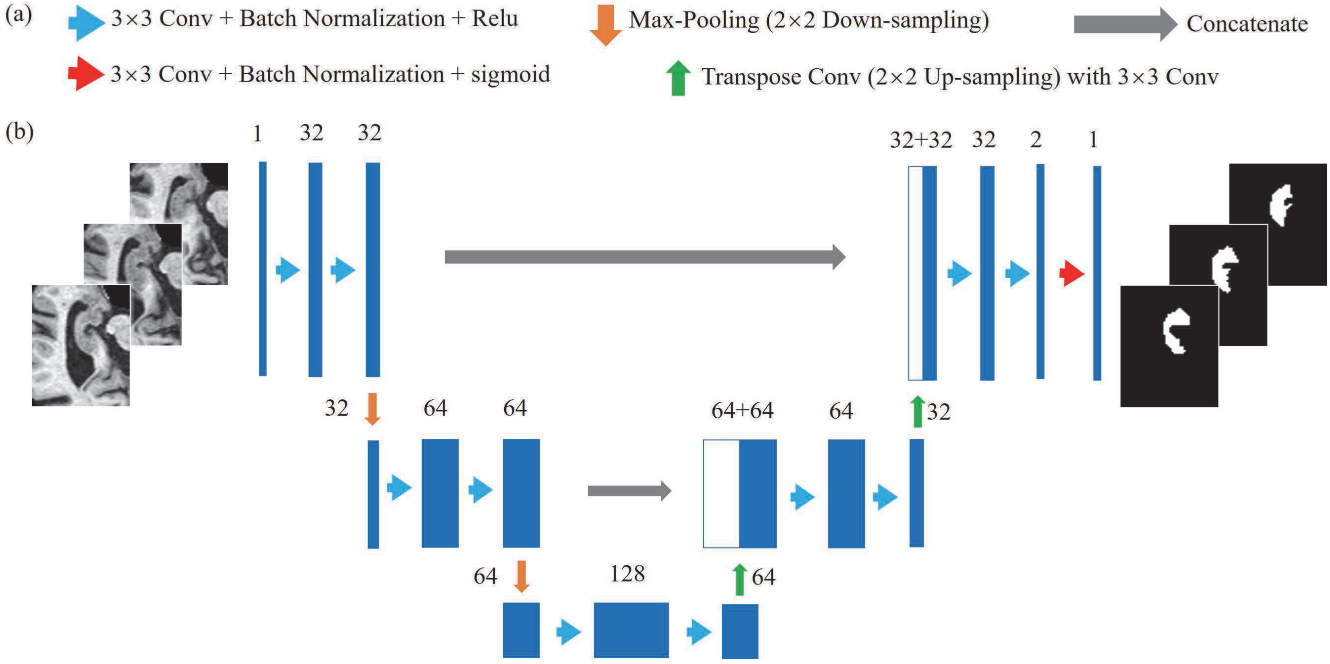
(a)
3×3 Conv + Batch Normalization + Relu
Max-Pooling (2×2 Down-sampling)
Concatenate
3×3 Conv + Batch Normalization + sigmoid
Transpose Conv (2×2 Up-sampling) with 3×3 Conv
(b)
1
32
32
32+32
32
2
1
32
64
64
64+64
64
32
64
128
64

FIGURE 4. Architecture diagram of the proposed U-Net. (a) The arrows represent involved operations: convolution, max-pooling, transpose convolution and concatenation. (b) The vertical bars represent input or output data, and the numbers above the bars denote the channel dimensions.

C. LOCALIZATION

1) SOI FETCH

According to the statistics of the dataset we used, the left and right parts of hippocampus occupy 0.0169% ± 0.0037 and 0.0163% ± 0.0035 of the whole MRI volume, respectively. Due to the wide gap in the number of foreground and background voxels, it is almost impossible to train an FCN for segmentation. In the meantime, we notice that the hippocampus is not present in every slice of the MRI scan. Therefore, we first fetch the SOIs that contain the hippocampus by Bi-LeNet classification. Every slice is fed to the Bi-LeNet to determine whether it is an SOI. The non-SOIs are excluded from further analysis.

2) CANDIDATE REGION GENERATION

The SOIs fetched by Bi-LeNet are fed into the U-Nets to produce probability maps. Using 0.5 as the threshold, we binarize the probability maps to obtain segmentation masks and then determined the centroids of the masks. The segmentation is coarse because the U-Net is trained on SOIs with

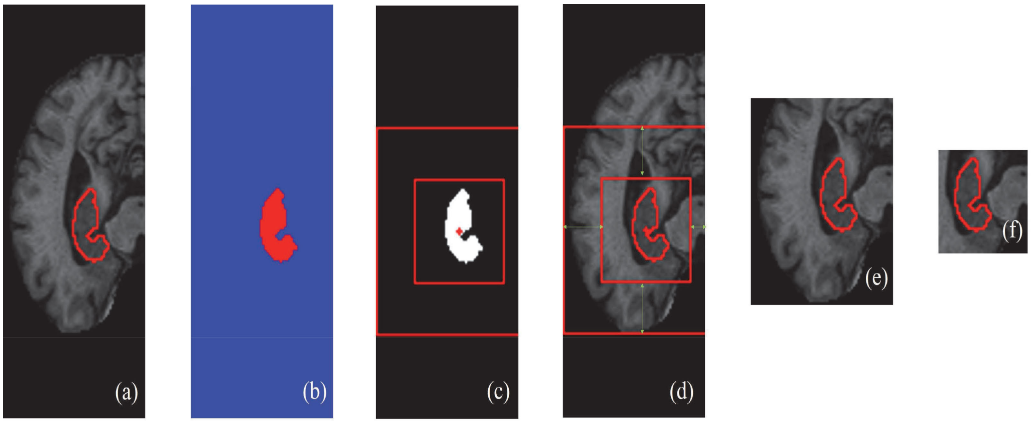

FIGURE 5. Pipeline of candidate region generation. The hippocampus and bounding boxes are outlined in red. (a) The selected SOI. (b) The probability map output by U-Net. (c) The centroid and two bounding boxes calculated based on the mask. (d) The centroid and two bounding boxes superimposed on the SOI. The increases of the borders are indicated with green arrows. (e) and (f) are candidate M- and S-regions, respectively, cropped from the SOI according to the two bounding boxes.

imbalanced data. With the centroid as the center, multiple bounding boxes of different sizes are established for each SOI, which are then used to generate candidate regions of different sizes for each SOI. To find the position of the centroid of the mask, the low-order geometric moments m_{00}, m_{01}, m_{10} must be calculated. For an $M \times N$ grayscale image $I(i, j)$, its pq order geometric moment can be calculated as follows:

$$m_{pq} = \sum_{i=1}^{M}\sum_{j=1}^{N} i^p j^q I(i, j) \tag{4}$$

The centroid is positioned at:

$$(x, y) = (\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}) \tag{5}$$

Taking this centroid (x, y) as the center, we can generate multiple candidate regions of multiple sizes. In this work, we generated two candidate regions for each SOI, one small sized and the other medium sized. Physically, the sizes of the hippocampal regions on all MRI scans lie within a fixed range. The upper limit of the statistical range is used as the size of the small-size bounding box, which covers the hippocampal regions across all the scans and contains as little background as possible. Based on the small-sized bounding boxes, SOIs are cropped to generate small-size candidate regions (S-regions). Then, bounding boxes twice the size of the S-region is used to generate medium-sized candidate regions (M-regions). Finally, the whole SOIs are defined as L-regions. Fig. 5 shows the process of candidate region generation.

D. SEGMENTATION

1) PLANAR FUSION

In this phase, we perform segmentation once more and named it fine segmentation to distinguish it from the coarse segmentation performed in the candidate region generation phase. The fine segmentation delineates the hippocampus morphology better than the coarse segmentation. The performance improvement of U-Net depends on the improved sample balance of candidate regions compared to the original SOIs.

U-Net models $\{H_M, H_S\}^v$ assume the role of fine segmentation operation, which takes the candidate regions $\{\{x_1, x_2, \cdots, x_t\}_M, \{x_1, x_2, \cdots, x_t\}_S\}^v$ as input and generates the probability maps $\{\{h_1, h_2, \cdots, h_t\}_M, \{h_1, h_2, \cdots, h_t\}_S\}^v$ of different sizes, where $v \in \{A, C, S\}$ indicate axial, coronal and sagittal directions, and the subscripts M and S indicate the M-region and S-region, respectively. Planar fusion refers to the fusion of the prediction results given by the fine segmentation performed on the S-regions and M-regions. It is infeasible to fuse the $\{\{h_1, h_2, \cdots, h_t\}_M\}^v$ and $\{\{h_1, h_2, \cdots, h_t\}_S\}^v$ directly because of the size difference of the probability maps. The probability maps of the M-regions $\{\{h_1, h_2, \cdots, h_t\}_M\}^v$ are cropped to the same size as that of the S-regions so that the two probability maps can be summed to generate the final segmentation mask sequence $Y^v = \{y_1, y_2, \cdots, y_t\}^v$, where the decision threshold is equal to 0.5, and $y_i \in \{0, 1\}$.

2) MULTI-VIEW DECISION

The three segmentation mask sequences $\{Y^A, Y^C, Y^S\}$ obtained in the axial, coronal and sagittal views are finally aggregated to generate an overall hippocampus segmentation Υ by MV. We assume that p represents a voxel in the 3D MRI image space. p^a, p^c and p^s are the projections of p in the axial view space Y^A, the coronal view space Y^C and the sagittal view space Y^S, respectively. Let $m(p^a)$, $m(p^c)$ and $m(p^s)$ denote the corresponding values of p_a, p_c and p_s in spaces Y^A, Y^C and Y^S, where $m(p^a) \in \{0, 1\}$, $m(p^c) \in \{0, 1\}$ and $m(p^s) \in \{0, 1\}$. Therefore, for a voxel p in Υ space, its value is determined by

$$m(p) = \begin{cases} 1, & m(p^a) + m(p^c) + m(p^s) \geq 2 \\ 0, & otherwise \end{cases} \tag{6}$$

The MV contributes to the cancellation of the random errors and enhancement of the prediction. Aggregation of complementary information from distinct views usually

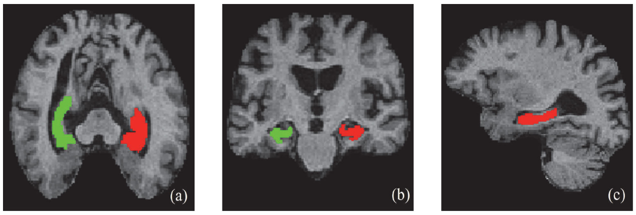

FIGURE 6. The morphology of the hippocampus in different views. The red region is the left hippocampus and the green region is the right hippocampus. (a) The hippocampus in an axial view, (b) coronal view and (c) sagittal view.

leads to more precise segmentation of the organ boundary. Fig. 6 shows the morphology of the hippocampus in the brain in different views.

IV. EXPERIMENT & RESULTS

A. DATA

The dataset used was obtained from the publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership, aiming to test whether serial MRI, positron emission tomography, other biological markers, clinical and neuropsychological assessment can be combined to evaluate the progression of mild cognitive impairment and early Alzheimer's disease.

The dataset consists of T1-weighted MRI scans of 74 patients with segmentation labels. The size of each scan is 192 × 192 × 160. Each MRI scan was first centrally cropped into a sub-volume of 160 × 144 × 128 to remove the black margins, which do not include any useful information. Because the original images were processed by affine transformation, the brain was located in the center of the MRI stack. We longitudinally split every 3D MRI scan into left and right parts. Each part contained a hippocampus. The sizes of the left and right parts were 160 × 144 × 64. For each part, the size of sagittal slices was 160×144, the size of axial slices was 144 × 64 and the size of coronal slices was 160 × 64. In addition, the axial slices and coronal slices of the right part were flipped horizontally to ensure that the appearance of the left and right halves of the brain was consistent in the training and testing phases. In addition, all the images were preprocessed through normalization and mean value removal before being used.

The training dataset of Bi-LeNet was generated based on the manual segmentation labels. An MRI slice was considered an SOI if its corresponding label mask indicated the presence of hippocampus blobs. According to the statistical results of the whole dataset, the hippocampus-containing slices accounted for 10.6% (± 2.4%) of the whole slice sequence in the axial view, 17.5% (± 1.5%) in the coronal view, and 10.1% (± 1.1%) in the sagittal view.

We evaluated the performance of the proposed method with 10-fold cross-validation throughout the experiment, i.e., 66 MRI scans were used as training sets and 8 scans as test sets. No validation set was used.

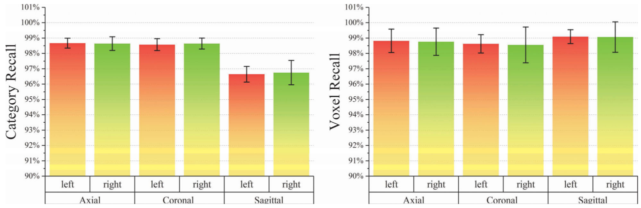

FIGURE 7. Mean values and standard deviations of Category Recall and Voxel Recall.

In the following descriptions, L and R are used to denote left and right. In addition, A, S and C stand for axial view, coronal view and sagittal view, respectively. Two-letter combinations are used to indicate the performance of the method in each step. For example, LA stands for the performance of the method in segmenting the left hippocampus in axial view.

B. EVALUATION OF LOCALIZATION

1) EVALUATION OF SOI FETCH

We measured the performance of Bi-LeNet with two metrics, i.e., category recall and voxel recall. Category recall refers to the quantity ratio of the predicted positive samples to the total positive samples. Voxel recall, on the other hand, refers to the volume ratio of the hippocampus area in the predicted positive samples to that in the ground truth samples. Fully trained Bi-LeNets show high performance on both category recall and voxel recall, as illustrated in Fig. 7.

In the experiments, it can be observed that Bi-LeNet had almost the same performance on the left and right hippocampi. The lowest category recalls are 96.64±0.51% on LS and 96.75±0.79% on RS, whereas the voxel recalls on sagittal view were greatest, 99.09±0.45% on LS and 99.06±0.99% on RS. Fetching SOIs with Bi-LeNet recalls, on average, more than 98.82% of the target organ volume on all three views. It shows ideal stability; even the worst metric reaches 98.56%±1.16 on voxel recall. A volume loss of less than 2% has little effect on subsequent procedures.

2) EVALUATION OF CANDIDATE REGION GENERATION

We used the DSC and Euclidean distance to measure the performances of coarse segmentation and centroid localization in this step. DSC was defined in (1), and the Euclidean distance (Ed) is defined as:

$$Ed = \sqrt[2]{(x_G - x_P)^2 + (y_G - y_P)^2} \tag{7}$$

where (x_P, y_P) are the coordinates of the centroid predicted in coarse segmentation, and (x_G, y_G) are the coordinates of the ground truth centroid.

Table 1 demonstrates the mean value μ and the standard deviation σ of the DSC of coarse segmentation and the corresponding Ed between the predicted centroid and the ground truth centroid. Basically, DSC and Ed are negatively correlated, i.e., the higher the DSC, the lower the Ed is, and the more precise the localization is. In the experiment, it can be seen that the predicted centroid is close to the ground truth centroid on all views. The localization bounding box

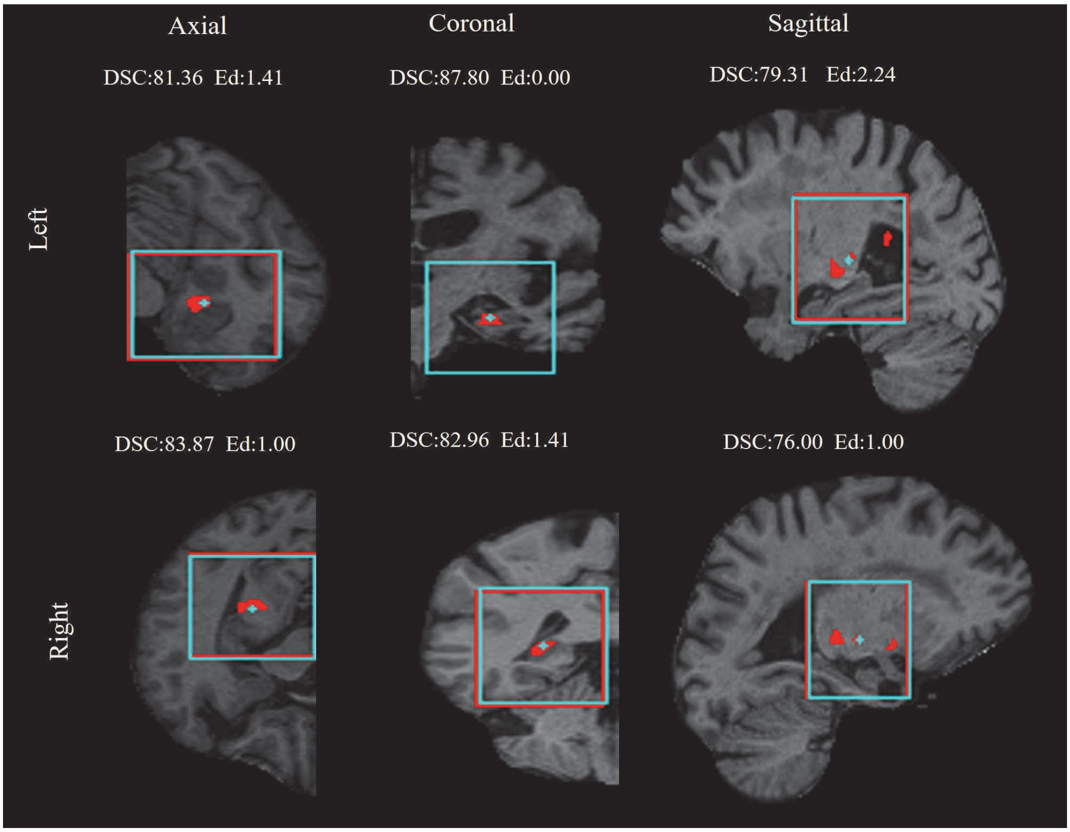

FIGURE 8. Examples of centroid localization. Prediction results are marked in bluish green, and the ground truth is marked in red. It can be seen that the higher the DSC score, the better that the predicted centroid and the bounding box coincide with the ground truth.

TABLE 1. Coarse segmentation and centroid localization results.

	μ_{DSC}	σ_{DSC}	μ_{Ed}	σ_{Ed}
LA	86.59	0.83	0.81	0.12
RA	86.50	0.66	0.91	0.17
LC	89.28	0.21	0.51	0.03
RC	89.60	0.25	0.46	0.04
LS	86.53	1.32	1.71	0.31
RS	87.01	0.83	1.50	0.28

covers the whole hippocampus region. The hippocampus is located roughly in the center of the box, which is beneficial for maintaining the integrity of the surrounding features of the target. Some example localization results are shown in Fig. 8.

C. EVALUATION OF SEGMENTATION

1) EVALUATION OF FINE SEGMENTATION

We applied three evaluation metrics, i.e., the DSC, the Jaccard similarity coefficient (JSC) [55] and the Hausdorff distance [56], to measure the performance of the fine segmentation. The DSC reflects the overall overlap between the prediction and ground truth in 3D space. The JSC focuses on describing the overlap in 2D cases. Finally, the Hausdorff distance describes the morphological similarity of 2D segmentations by measuring the closeness between the predicted segmentation and the ground truth, with smaller results indicating better segmentations. Among them, the DSC is the most important metric and is defined in (1). The JSC is defined as:

$$JSC(A_P, A_G) = \frac{|A_P \cap A_G|}{|A_P \cup A_G|} \times 100\% \tag{8}$$

The Hausdorff distance is defined as:

$$\begin{aligned} & d_H(A_P, A_G) \\ & = \max\{\sup_{p \in A_P} \inf_{g \in A_G} d(p, g), \sup_{g \in A_G} \inf_{p \in A_P} d(p, g)\} \end{aligned} \tag{9}$$

where sup represents the supremum, inf represents the infimum, and A_P and A_G represents the segmentation prediction and ground truth, respectively.

Our experimental results are described in the form of boxplots, as shown in Fig. 9. L stands for the L-regions, M stands for the M-regions and S stands for the S-regions. F(X,Y) represents the planar fusion of candidate regions X and Y. From the comparisons among L, M and S, it can be seen that fine segmentation completely outperforms coarse segmentation because of the reduction in the search space and the increase in the ratio of positive samples. Almost all the indicators of S are better than those of M. This means that

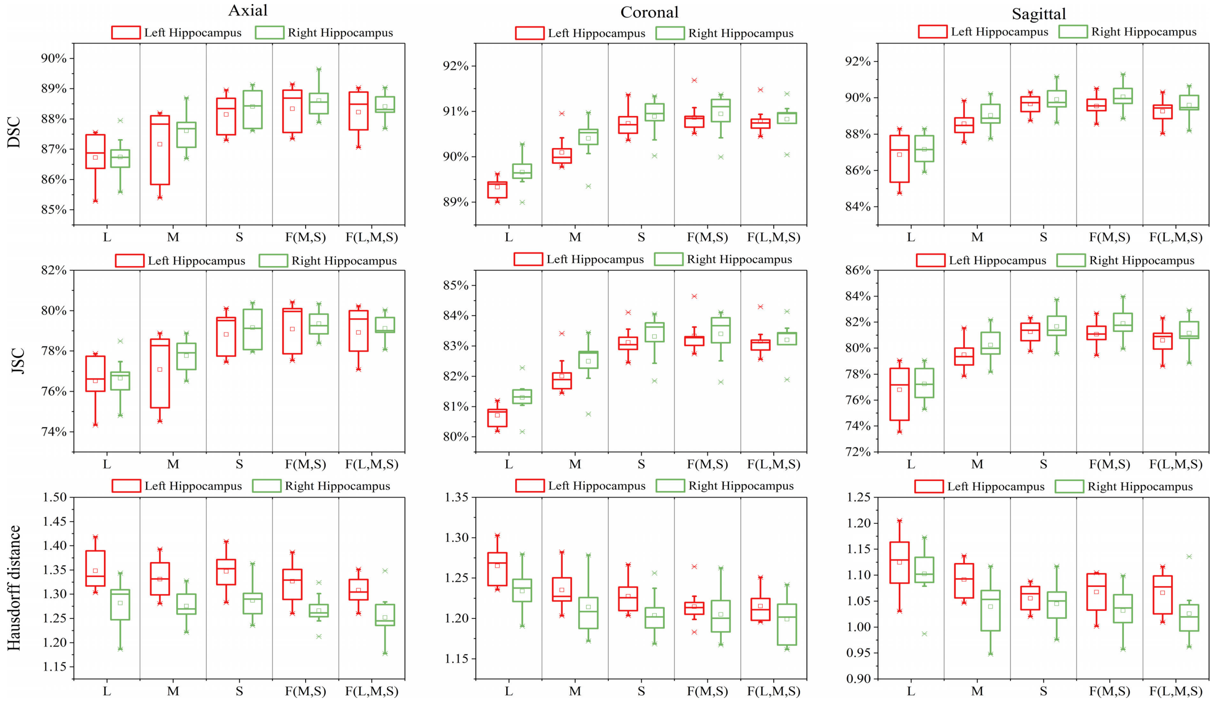

FIGURE 9. Boxplots of the DSCs, JSCs and Hausdorff distances of multi-size left and right hippocampus in three views. In each box, the central mark is the median and the edges of the box denote the 25th and 75th percentiles. Whiskers extend from each end of the box to adjacent values in the dataset and the extreme values within 1 interquartile range from the ends of the box. Outliers are data with values beyond the ends of the whiskers and are represented with stars.

TABLE 2. Comparison between coarse segmentation and fine segmentation (DSC).

	μ_{coarse}	μ_{fine-M}	μ_{fine-S}
LA	86.59	87.16	88.15
RA	86.50	87.61	88.41
LC	89.28	90.11	90.74
RC	89.60	90.40	90.89
LS	86.53	88.57	89.66
RS	87.01	89.03	89.90

S-regions demonstrate better sample balance compared to M-regions and L-regions. Comparisons of the DSCs and JSCs in all views show that segmentation performance in the coronal view is usually better than that in other two views, which may be due to the distinct morphological features of hippocampus in the different views.

In terms of the mean values of the DSC and JSC, F(M,S) exhibits enhanced performance on all three views. In addition, F(L,M,S) performs worse than F(M,S) in terms of the DSC and JSC because of the poor segmentation performance in the L-regions. In terms of the Hausdorff distance, F(L,M,S) scores similarly to F(M,S), and on axial view F(L,M,S) scores slightly higher.

Table 2 shows the results of coarse segmentation (direct segmentation) of SOIs and fine segmentation using different candidate regions. It shows that fine segmentation outperforms direct segmentation in all the views.

2) ANALYSIS OF PLANAR FUSION

We analyzed the segmentation performance in hippocampi with different morphological appearances. The difference is mainly reflected in the area occupied by the hippocampus voxels. Candidate regions were sorted in decreasing order based on their hippocampus areas. The top 5% and bottom 5% candidate regions were labeled T-5% and B-5%, respectively. The receiver operating characteristic (ROC) curves and areas under the curve (AUCs) were applied to measure the segmentation performance for different appearances of the hippocampus and the effect of planar fusion and are shown in Fig. 10.

Conventional two-stage methods use fixed-size candidate regions, whose performance is equivalent to that using S-regions or M-regions shown in Fig. 10. In axial view and coronal view, when segmenting T-5% regions, the effect of using M-regions is better than that using S-regions, which indicates that the size of the candidate region needs to be larger when segmenting an image where the hippocampus occupies a larger area in order to extract more global features. For B-5% regions, the fact that using S-regions is substantially better than using M-regions means that smaller candidate regions are beneficial for segmenting smaller targets due to the higher proportion of positive samples. In the sagittal view, M-regions have no obvious advantage for T-5% regions, while using

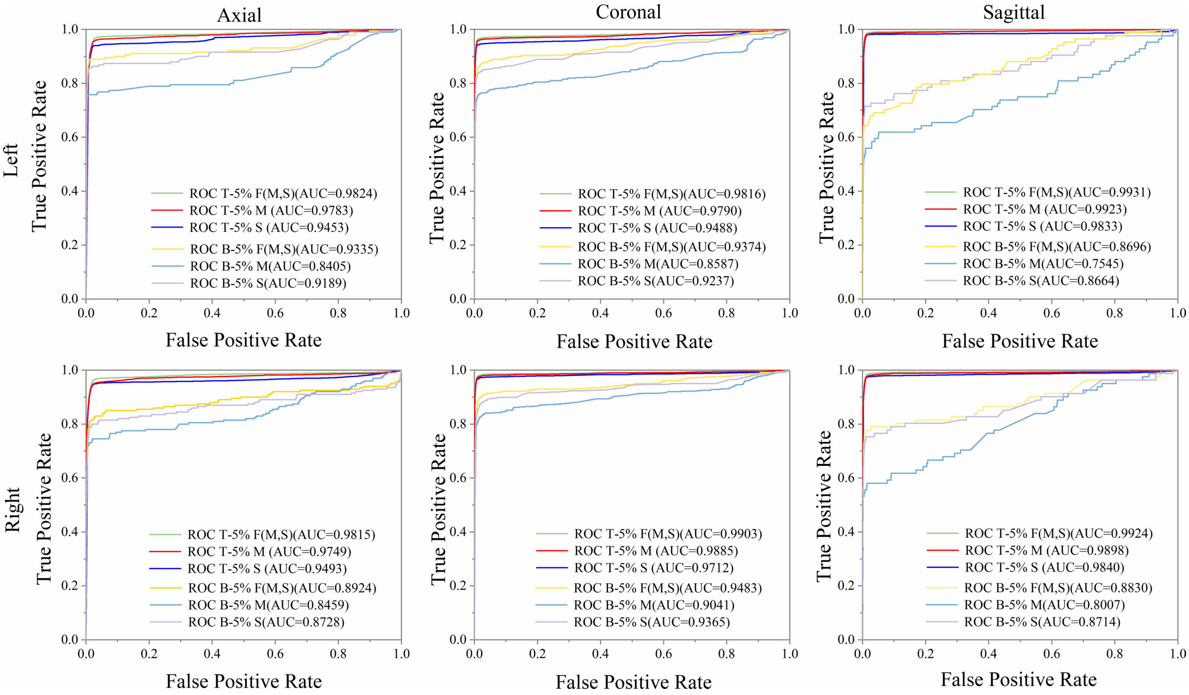

FIGURE 10. Receiver operating characteristic (ROC) curves of the segmentation in different morphologies of the hippocampus and the effects of planar fusion.

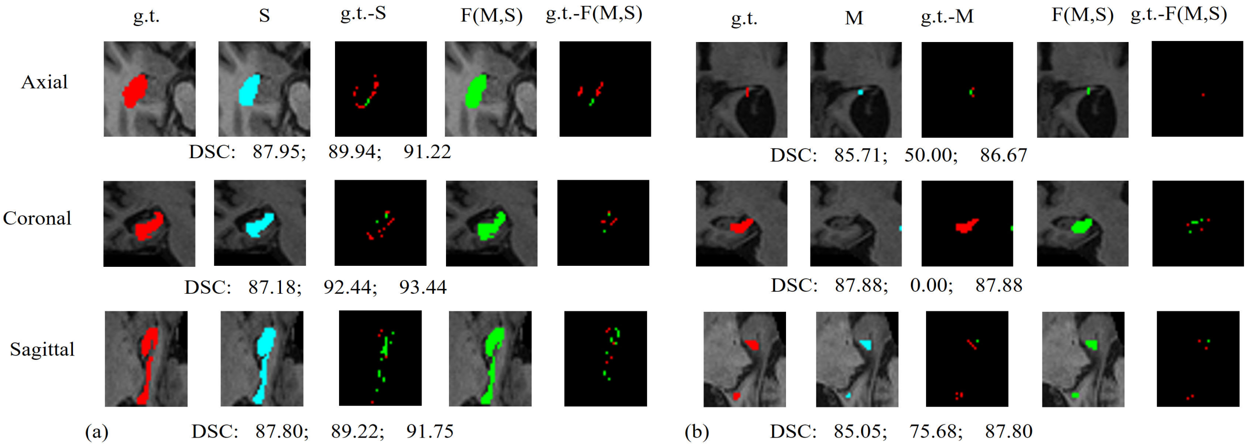

FIGURE 11. (a) The enhancement effect of F(M, S) with S-regions. (b) The enhancement effect of F(M,S) with M-regions. In the 3rd and 5th columns, false positive samples of the segmentation are marked in green, while false negative samples are marked in red. DSC scores corresponding to the S-regions, M-regions, and F(M,S) are given in order under each set of images.

S-regions results in better performance for the B-5% regions. All these findings indicate the necessity of planar fusion. Planar fusion reduces both false positive and false negative predictions, as Fig. 11 shows. The AUC based on F(M,S) is larger than that based on both M-regions and S-regions.

3) EVALUATION OF MULTI-VIEW DECISION

After performing planar fusion, enhanced segmentation masks are produced in the three. Next, in the multi-view decision step, the masks from the three views are combined to generate the final segmentation result. The results of the multi-view decision step are summarized in Table 3. Some surface rendering examples are shown in Fig. 12.

We compared our experimental results with those of related works on the same dataset. Notably, the subsets of the dataset used by different studies are not exactly the same. Therefore, to ensure a fair comparison, the amount of data we employed was the median of the amounts of data used in the compared studies. Table 4 shows that our method achieved the best overall segmentation performance.

V. DISCUSSION

In this paper, compared with that of previous approaches, the proposed method achieved better performance on the ADNI dataset. We attribute these improvements to the fetching of SOIs and the segmentation fusion on multi-size candidate

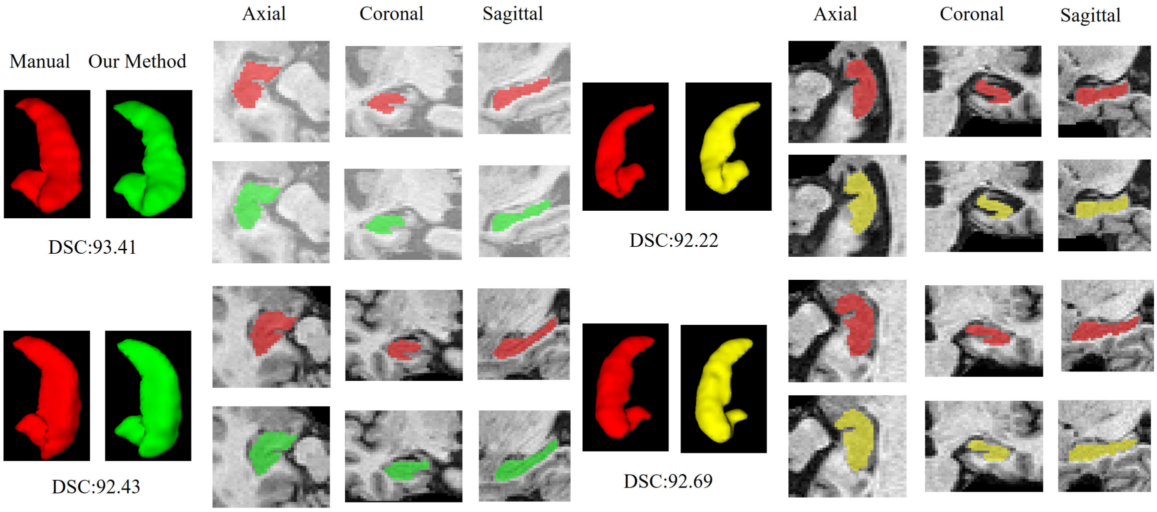

FIGURE 12. Surface rendering results of some samples of the ground truth and those of the corresponding prediction segmentation. Red: manual hippocampus segmentation. Green: prediction segmentation of the left hippocampus (top view). Yellow: prediction segmentation of the right hippocampus (top view).

TABLE 3. Multi-view decision results.

	MIN_{DSC}	MAX_{DSC}	μ_{DSC}	σ_{DSC}
LEFT	91.44	93.43	92.48	0.61
RIGHT	92.01	93.56	92.90	0.51

regions. For the SOI fetch step in the localization stage, our main concern is to select candidate slices for segmentation to reduce the negative effects caused by excessive background voxels. In some patch-based methods, including F-DDLS [32], Progressive SPBL [31], manifold learning MAS [57], etc., searching for similar patches in a complete MRI patch space is a cumbersome and unstable task. Due to the large search space and difficulty in matching similar patches, these patch-based methods have lower performance compared to our method. We believe that the idea of SOI fetch can also be applied to improve patch-based methods.

Removing redundant information is an effective strategy to reduce the difficulty of similarity matching. Although more advanced classification networks [58], [59] can be applied to fetch SOI, in our proposal, the simple LeNet-based network performs sufficiently well, with a volume loss of less than 2%. The small voxel volume loss is caused by SOIs containing the end regions of the hippocampus. These end regions usually consist of only a few voxels, making it difficult to capture target features even through more sophisticated networks. The proposed method inherits and extends the localization stage of conventional two-stage segmentation methods, which is lacking in some current deep learning-based hippocampus segmentation methods [50], [51]. The lack of a localization stage limits the application of these methods in practical situations. In addition, with regard to the U-Net, we attempted to train a single U-Net model that used the same weights in all three views. However, convergence was difficult. We suggest that this was likely due to the highly heterogeneous features of the hippocampus in the three orthogonal views.

TABLE 4. Comparison between our method and those of other related works (DSC).

METHODS	LEFT	RIGHT	AVERAGE
Ada-SVM [26]	81.40	82.20	82.80
F-DDLS [32]	87.20	87.02	87.20
Progressive SPBL [31]	88.02	88.50	88.35
U-Seg-Net+ CLSTM [51]	89.21±1.68	89.37±1.44	89.29
Ensemble -Net [50]	89.48±1.49	89.46±1.42	89.47
STEPS [23]	90.30	90.30	90.30
Manifold Learning MAS [57]	91.14±0.02	91.14±0.02	91.14
Ours	92.48±0.61	92.90±0.51	92.69

Planar fusion on candidate regions of multiple sizes can more accurately segment tissues of different sizes on different slices. Fig. 10 shows that the AUC of F(M,S) is better than that of the M- and S-regions, which means that the planar fusion segmentation on multi-size candidate regions has higher accuracy than the segmentation on fixed-size candidate regions. We hypothesize that it is difficult to use fixed-size candidate regions to represent organs with complex morphologies. Specifically, both large and small candidate regions have their own advantages and disadvantages. For example, the AUC of T-5% on M-regions is larger than that on S-regions, while the opposite is true for B-5%. This is exactly what the conventional two-stage approaches ignore. The segmentation performance based on either M-regions or

S-regions is roughly similar to that of the conventional two-stage approaches. M-regions usually contain more global features, which is advantageous for segmenting larger regions of the hippocampus but disadvantageous for segmenting smaller regions due to sample imbalance. Conversely, few negative samples in the S-region, thus resulting in a small impact of the negative samples, will help in the segmentation of smaller regions of the hippocampus. Smaller regions provide more local features, which is not conducive to the accurate segmentation of large regions. The fusion segmentation predictions from two candidate regions of different sizes can further improve the segmentation accuracy, enabling our method to perform better than deep learning segmentation methods using fixed-size regions [50], [51]. In addition, it is worth mentioning that, as summarized in Table 4, all deep learning methods outperform the manual feature-based method [26], which proves the effectiveness of automatic features extraction.

Fig. 11 highlights the importance of our multi-size fusion strategy to reduce the errors of segmentation. Fusion is very effective in reducing false positives and false negatives and helps to more accurately describe the hippocampal boundary. For segmenting larger hippocampal regions using S-regions (Fig. 11 (a)), voxels in the contour region are easily misjudged. However, fusion with M-region segmentation can greatly reduce the number of misclassified samples. The high-confidence predictions obtained from using M-regions in the segmentation of a large area can compensate for some ''wobbly'' judgments (false positive and false negative samples with prediction probabilities close to the threshold value) of predictions that use S-regions alone. Using the M-regions to segment small-area targets is usually ineffective, and often it is not even possible to segment such targets. For example, the second line of Fig. 11(b) shows that the DSC of the M-regions is 0.00; therefore, the contribution of F(M,S) comes only from the segmentation using the S-regions.

Fusion of additional candidate regions is a promising way to further improve the prediction performance. In our work, the use of candidate regions of two sizes (S-regions and M-regions) results in an obvious performance improvement over the use of a single candidate region (Fig. 9). However, the size selection is critical. If more candidate regions of other sizes are to be used, determining the appropriate sizes is a problem that requires specific research for extracting more complementary features from such additional candidate region.

The multi-view decision mechanism guarantees the refined segmentation of tissues with complex morphology. In particular, this step helps to correctly classify voxels at the edges of organs. From Fig. 6, we can see that the hippocampus is an organ with a complex appearance that differs in the three views. Some edges of the organ are blurry in one view but clear in another. A multi-view observation provides obvious global information on the overall structure to improve the segmentation. Our approach achieved a higher DSC score than STEPS [23], which may be because the latter uses only information from a cross-sectional view or a longitudinal view, not information from a multi-view decision.

The multi-view decision used in our method may be further optimized. As shown in Fig. 9, the segmentation performance differs substantially among the three views, so view selection may be a factor that needs to be considered to ensure ideal segmentation results. The boxplots in Fig. 9 show that the DSC score obtained from the coronal view is always higher than those obtained from the other views, suggesting that the use of weighted MV may be more helpful than the unweighted MV used in the proposed method.

VI. CONCLUSION

The main contribution of this paper is the proposal of a novel two-stage hippocampus segmentation method using the fusion of information from multi-size candidate regions. This method contains a hippocampus localization stage and a segmentation stage that fuses multi-size information and makes a multi-view decision. Segmentation fusion based on multi-size candidate regions results in an effective representation of the complex morphological features of the hippocampus. The method achieves high precision segmentation of the hippocampus that is superior to that of recently reported algorithms run on the ADNI dataset. The proposed method is expected to automatically segment and measure the volume of the hippocampus in clinical settings to help doctors diagnose diseases such as Alzheimer's disease, schizophrenia and major depression. In addition, this method can be extended to segmentation tasks of other small organs such as the pancreas, gallbladder, and so on.

REFERENCES

[1] M. Bobinski, ''Neurofibrillary pathology—Correlation with hippocampal formation atrophy in Alzheimer disease,'' *Neurobiol. Aging*, vol. 17, no. 6, pp. 909–919, Nov. 1996.

[2] M. Styner, J. A. Lieberman, D. Pantazis, and G. Gerig, ''Boundary and medial shape analysis of the hippocampus in schizophrenia,'' *Med. Image Anal.*, vol. 8, no. 3, pp. 197–203, Sep. 2004.

[3] J. D. Bremner, M. Narayan, E. R. Anderson, L. H. Staib, H. L. Miller, and D. S. Charney, ''Hippocampal volume reduction in major depression,'' *Amer. J. Psychiatry*, vol. 157, no. 1, pp. 115–118, 2000.

[4] G. B. Frisoni, ''Structural imaging in the clinical diagnosis of Alzheimer's disease: Problems and tools,'' *J. Neurol., Neurosurgery Psychiatry*, vol. 70, no. 6, pp. 711–718, Jun. 2001.

[5] R. E. Hogan, K. E. Mark, L. Wang, S. Joshi, M. I. Miller, and R. D. Bucholz, ''Mesial temporal sclerosis and temporal lobe epilepsy: MR imaging deformation-based segmentation of the hippocampus in five patients,'' *Radiology*, vol. 216, no. 1, pp. 291–297, Jul. 2000.

[6] O. T. Carmichael, H. A. Aizenstein, S. W. Davis, J. T. Becker, P. M. Thompson, C. C. Meltzer, and Y. Liu, ''Atlas-based hippocampus segmentation in Alzheimer's disease and mild cognitive impairment,'' *NeuroImage*, vol. 27, no. 4, pp. 979–990, Oct. 2005.

[7] J. Long, E. Shelhamer, and T. Darrell, ''Fully convolutional networks for semantic segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Madison, WI, USA, Jun. 2015, pp. 3431–3440.

[8] H. R. Roth, L. Lu, N. Lay, A. P. Harrison, A. Farag, A. Sohn, and R. M. Summers, ''Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation,'' *Med. Image Anal.*, vol. 45, pp. 94–107, Apr. 2018.

[9] H. R. Roth, L. Lu, A. Farag, A. Sohn, and R. M. Summers, ''Spatial aggregation of holistically-nested networks for automated pancreas segmentation,'' in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Athens, Greece: Springer, 2016, pp. 451–459.

[10] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille, "A fixed-point model for pancreas segmentation in abdominal CT scans," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Quebec City, QC, Canada: Springer, 2017, pp. 693–701.

[11] V. Dill, A. R. Franco, and M. S. Pinho, "Automated methods for hippocampus segmentation: The evolution and a review of the state of the art," *Neuroinformatics*, vol. 13, no. 2, pp. 133–150, Apr. 2015.

[12] B. S. He, X. P. Zhang, and Y. G. Shi, "Hippocampus segmentation techniques: A survey," *Adv. Mater. Res.*, vol. 760, pp. 2086–2090, Sep. 2013.

[13] P. A. Freeborough, N. C. Fox, and R. I. Kitney, "Interactive algorithms for the segmentation and quantitation of 3-D MRI brain scans," *Comput. Methods Programs Biomed.*, vol. 53, no. 1, pp. 15–25, May 1997.

[14] M. Chupin, A. R. Mukuna-Bantumbakulu, D. Hasboun, E. Bardinet, S. Baillet, S. Kinkingnéhun, L. Lemieux, B. Dubois, and L. Garnero, "Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: Method and validation on controls and patients with Alzheimer's disease," *NeuroImage*, vol. 34, no. 3, pp. 996–1019, Feb. 2007.

[15] E. A. Ashton, K. J. Parker, M. J. Berg, and C. W. Chen, "A novel volumetric feature extraction technique with applications to MR images," *IEEE Trans. Med. Imag.*, vol. 16, no. 4, pp. 365–371, Aug. 1997.

[16] A. Ghanei, H. Soltanian-Zadeh, and J. P. Windham, "A 3D deformable surface model for segmentation of objects from volumetric data in medical images," *Comput. Biol. Med.*, vol. 28, no. 3, pp. 239–253, May 1998.

[17] J. Barnes, R. G. Boyes, E. B. Lewis, J. M. Schott, C. Frost, R. I. Scahill, and N. C. Fox, "Automatic calculation of hippocampal atrophy rates using a hippocampal template and the boundary shift integral," *Neurobiol. Aging*, vol. 28, no. 11, pp. 1657–1663, Nov. 2007.

[18] A. Akhondi-Asl, K. Jafari-Khouzani, K. Elisevich, and H. Soltanian-Zadeh, "Hippocampal volumetry for lateralization of temporal lobe epilepsy: Automated versus manual methods," *NeuroImage*, vol. 54, pp. 218–226, Jan. 2011.

[19] R. E. Hogan, K. E. Mark, I. Choudhuri, L. Wang, S. Joshi, M. I. Miller, and R. D. Bucholz, "Magnetic resonance imaging deformation-based segmentation of the hippocampus in patients with mesial temporal sclerosis and temporal lobe epilepsy," *J. Digit. Imag.*, vol. 13, no. S1, pp. 217–218, May 2000.

[20] P. Coupé, J. V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins, "Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation," *NeuroImage*, vol. 54, no. 2, pp. 940–954, Jan. 2011.

[21] A. Zandifar, V. Fonov, P. Coupé, J. Pruessner, and D. L. Collins, "A comparison of accurate automatic hippocampal segmentation methods," *NeuroImage*, vol. 155, pp. 383–393, Jul. 2017.

[22] G. Wu, M. Kim, G. Sanroma, Q. Wang, B. C. Munsell, and D. Shen, "Hierarchical multi-atlas label fusion with multi-scale feature representation and label-specific patch partition," *NeuroImage*, vol. 106, pp. 34–46, Feb. 2015.

[23] M. Jorge Cardoso, K. Leung, M. Modat, S. Keihaninejad, D. Cash, J. Barnes, N. C. Fox, and S. Ourselin, "STEPS: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation," *Med. Image Anal.*, vol. 17, no. 6, pp. 671–684, Aug. 2013.

[24] 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.

[25] P. Aljabar, R. A. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert, "Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy," *NeuroImage*, vol. 46, no. 3, pp. 726–738, Jul. 2009.

[26] J. H. Morra, Z. Tu, L. G. Apostolova, A. E. Green, A. W. Toga, and P. M. Thompson, "Comparison of AdaBoost and support vector machines for detecting Alzheimer's disease through automated hippocampal segmentation," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 30–43, Jan. 2010.

[27] Y. Wang, G. Ma, X. Wu, and J. Zhou, "Patch-based label fusion with structured discriminant embedding for hippocampus segmentation," *Neuroinformatics*, vol. 16, nos. 3–4, pp. 411–423, Oct. 2018.

[28] Y. Guo, Z. Wu, and D. Shen, "Learning longitudinal classification-regression model for infant hippocampus segmentation," *Neurocomputing*, Apr. 2019.

[29] F. van der Lijn, T. Den Heijer, M. M. B. Breteler, and W. J. Niessen, "Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts," *NeuroImage*, vol. 43, no. 4, pp. 708–720, Dec. 2008.

[30] Y. Hao, T. Wang, X. Zhang, Y. Duan, C. Yu, T. Jiang, and Y. Fan, "Local label learning (LLL) for subcortical structure segmentation: Application to hippocampus segmentation," *Hum. Brain Mapping*, vol. 35, no. 6, pp. 2674–2697, Jun. 2014.

[31] Y. Song, G. Wu, K. Bahrami, Q. Sun, and D. Shen, "Progressive multi-atlas label fusion by dictionary evolution," *Med. Image Anal.*, vol. 36, pp. 162–171, Feb. 2017.

[32] 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.

[33] K. M. Pohl, S. Bouix, M. Nakamura, T. Rohlfing, R. W. McCarley, R. Kikinis, W. E. L. Grimson, M. E. Shenton, and W. M. Wells, "A hierarchical algorithm for MR brain image parcellation," *IEEE Trans. Med. Imag.*, vol. 26, no. 9, pp. 1201–1212, Sep. 2007.

[34] A. Akselrod-Ballin, M. Galun, J. M. Gomori, A. Brandt, and R. Basri, "Prior knowledge driven multiscale segmentation of brain MRI," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Brisbane, QLD, Australia: Springer, 2007, pp. 118–126.

[35] F. Jiang, "Medical image semantic segmentation based on deep learning," *Neural Comput. Appl.*, vol. 29, no. 5, pp. 1257–1265, 2018.

[36] H. Jiang, H. Ma, W. Qian, M. Gao, and Y. Li, "An automatic detection system of lung nodule based on multigroup patch-based deep learning network," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 4, pp. 1227–1237, Jul. 2018.

[37] A. Kumar, J. Kim, D. Lyndon, M. Fulham, and D. Feng, "An ensemble of fine-tuned convolutional neural networks for medical image classification," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 31–40, Jan. 2017.

[38] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Las Condes, Chile, Dec. 2015, pp. 1529–1537.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Munich, Germany: Springer, 2015, pp. 234–241.

[40] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[41] Y. Wang, Z. Sun, C. Liu, W. Peng, and J. Zhang, "MRI image segmentation by fully convolutional networks," in *Proc. IEEE Int. Conf. Mechatronics Autom.*, Harbin, China, Aug. 2016, pp. 1697–1702.

[42] L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, and D. Feng, "Dermoscopic image segmentation via multistage fully convolutional networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2065–2074, Sep. 2017.

[43] 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)*, Stanford, CA, USA, Oct. 2016, pp. 565–571.

[44] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "3D deeply supervised network for automatic liver segmentation from CT volumes," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Athens, Greece: Springer, 2016, pp. 149–157.

[45] Z. Zhong, Y. Kim, L. Zhou, K. Plichta, B. Allen, J. Buatti, and X. Wu, "3D fully convolutional networks for co-segmentation of tumors on PET-CT images," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Washington, DC., USA, Apr. 2018, pp. 228–231.

[46] 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. 2018.

[47] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sanchez, and B. van Ginneken, "Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1160–1169, May 2016.

[48] B. van Ginneken, A. A. A. Setio, C. Jacobs, and F. Ciompi, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," in *Proc. IEEE 12th Int. Symp. Biomed. Imag. (ISBI)*, New York, NY, USA, Apr. 2015, pp. 286–289.

[49] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Nagoya, Japan: Springer, 2013, pp. 246–253.

[50] 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)*, Melbourne, VIC, Australia, Apr. 2017, pp. 192–196.

[51] 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.* Quebec City, QC, Canada: Springer, 2017, pp. 88–96.

[52] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[54] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, Jul. 1945.

[55] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[56] D. P. Huttenlocher, W. J. Rucklidge, and G. A. Klanderman, "Comparing images using the Hausdorff distance," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Champaign, IL, USA, Sep. 1992, pp. 654–656.

[57] A. K. H. Duc, M. Modat, K. K. Leung, T. Kadir, and S. Ourselin, "Manifold learning for atlas selection in multi-atlas-based segmentation of hippocampus," in *Proc. SPIE Medical Imaging*, vol. 8314. San Diego, CA, USA: International Society for Optics and Photonics, 2012, Art. no. 83140Z.

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

[59] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

PING CAO was born in China, in 1968. She received the B.S. degree in biomedical engineering and the M.S. degree in computer science and technology from Zhejiang University, China, in 1991 and 2004, respectively.

Since 2008, she has been an Associate Professor with the Information Engineering Department, Zhijiang College, Zhejiang University of Technology. She is currently the PIs of more than ten research programs and the author of more than 20 articles and inventions. Her research interests include biomedical image processing, deep learning, cloud computing, and mobile computing.

QIUYANG SHENG was born in China, in 1994. He received the B.S. degrees from the School of Biomedical Engineering, Wenzhou Medical University, in 2013. He is currently pursuing the M.S. degree with the School of Information Engineering, Zhejiang University of Technology. His main research interests include medical image segmentation, signal processing, deep learning, and computer vision.

SIQI FANG received the B.A. degree in mathematics from the University of Cambridge, U.K., in 2019, where she is currently pursuing the M.S. degree in systems biology.

Her research interests include deep learning applications in medical image segmentation, gene expression analysis, and prediction of regulatory network structure.

XINYI LI received the B.S. degree in communication engineering from Lanzhou Jiaotong University. She is currently pursuing the master's degree in information and communication engineering with the Zhejiang University of Technology.

Her research interests include deep learning, biomedical signal processing, and semantic segmentation.

GANGMIN NING (Senior Member, IEEE) received the Ph.D. degree in biomedical engineering from Ilmenau Technical University, Germany, in 2001.

He is currently a Professor with the Department of Biomedical Engineering, Zhejiang University, where is currently a Ph.D. Supervisor. His research interests include multiscale cardiovascular system function modeling, mobile health technology, health promotion strategies, and medical intelligence. He is the Vice Chairman of the Vascular and Inflammatory Biology Committee of the Zhejiang Biomedical Society of American Physiological Society (APS), and a Core Member of the Zhejiang Critical Medical Technology Innovation Team. He is a Reviewer of various international academic journals. In the past five years, he has published more than 30 articles and received a series of medical device registration certificates (in cooperation with enterprises), national invention patent authorization, computer software copyright, and other independent intellectual property rights. He received the First Prize of Science and Technology Progress of Zhejiang Province, in 2005, the First Prize of Medicine and Health Science and Technology of Zhejiang Province, in 2012, and the Second Prize of Science and Technology Progress of Zhejiang Province, in 2014.

QING PAN (Member, IEEE) was born in China, in 1985. He received the B.S. and Ph.D. degrees in biomedical engineering from Zhejiang University, China, in 2008 and 2013, respectively.

From 2013 to 2016, he was a Lecturer with the College of Information Engineering, Zhejiang University of Technology, China. From 2016 to 2017, he was a Postdoctoral Researcher with the Institute of Physiology, Charité Universitätsmedizin Berlin, Germany. Since 2018, he has been an Associate Professor with the College of Information Engineering, Zhejiang University of Technology. His research interests include biomedical signal and image processing, physiological system modeling, deep learning, and medical device development.