This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2994114, IEEE Journal of Biomedical and Health Informatics

Discriminative Feature Network Based on a Hierarchical Attention Mechanism for Semantic Hippocampus Segmentation

Jiali Shi, Rong Zhang*, Lijun Guo, Linlin Gao, Huifang Ma, Jianhua Wang

***Abstract*—The morphological analysis of hippocampus is vital to various neurological studies including brain disorders and brain anatomy. To assist doctors in analyzing the shape and volume of the hippocampus, an accurate and automatic hippocampus segmentation method is highly demanded in the clinical practice. Given that fully convolutional networks (FCNs) have made significant contributions in biomedical image segmentation applications, we propose a notably discriminative feature network based on a hierarchical attention mechanism in hippocampal segmentation. First, considering the problem that the hippocampus is a rather small part in MR images, we design a context-aware high-level feature extraction module (CHFEM) to extract high-level features of scale invariance in the encoder stage. Further, we introduce a hierarchical attention mechanism into our segmentation framework. The mechanism is divided into three parts: a low-level feature spatial attention module (LFSAM) is developed to learn the spatial relationship between different pixels on each channel in the low-level stage of the encoder, a high-level feature channel attention module (HFCAM) is to model the semantic information relationship on different channel images in the high-level stage of the encoder, and a cross-connected attention module (CCAM) is designed in the decoder part to further suppress the noisy boundaries of hippocampus and simultaneously utilize the attentional low-level features from the encoder to better guide the high-level hippocampus edge segmentation in the decoder phase. The proposed approach achieves outstanding performance on the ADNI dataset and the Decathlon dataset compared with other semantic segmentation models and existing hippocampal segmentation approaches. Source code is available at https://github.com/LannyShi/Hippocampal-segmentation.**

***Index Terms*—hippocampal segmentation, encoder-decoder network, high-level features, low-level features**

I. Introduction

THE hippocampus is a very important tissue in the human brain and closely related to human cognitive functions such as learning and memory. The morphology analysis of the hippocampus is important for the diagnosis and prediction of various neurological diseases such as Alzheimer's disease (AD) [1], schizophrenia [2], and epilepsy [3].

The early clinical manifestation of these diseases is hippocampal atrophy in the brain [4], [5]. Doctors can use magnetic resonance technology to conduct three-dimensional imaging of the patient's brain in order to diagnose and formulate relevant treatment plans based on the results of image analysis. To determine whether the hippocampus is atrophic, doctors often need to segment it on magnetic resonance (MR) images and analyze its shape and volume [6], [7]. At present, manual segmentation of the hippocampus is still considered the gold standard for the analysis of hippocampal volume and morphology. However, the process is tedious, time-consuming (two to three hours to completion), subjective, and not repeatable [8].

Automatic hippocampal segmentation is a pixel-wise semantic segmentation task. With the recent development of convolutional neural networks (CNNs) [9], [10] pixel-level semantic segmentation tasks have significantly progressed due to their efficient feature extraction capabilities [11]-[13]. However, as the hippocampus is a gray matter structure, it has a low contrast with the surrounding tissues in MR images, and the hippocampus is irregularly shaped, small in size, and without obvious boundary at the edge and has large individual differences. Therefore, automatic segmentation of the hippocampus from MR images is still a challenging task.

Currently, the most effective semantic segmentation methods are based on the fully convolutional network (FCN) [14], which stacks multiple convolution and pooling layers to gradually increase the receptive field and generate high-level semantic

This research work is supported by the Zhejiang Provincial Public Welfare Technology Research Project (No. LGF18F020007), the National Natural Science Foundation of China (No. 61762078) and the Ningbo Municipal Natural Science Foundation of China (No.2018A610 057, 2018A610163). (Corresponding author: Rong Zhang.)

Jiali Shi is studying at Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo Zhejiang, China. (e-mail: 375956258@qq.com)

Rong Zhang is with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo Zhejiang, China. (e-mail: zhangrong@nbu.edu.cn)

Lijun Guo and Linlin Gao are with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo Zhejiang, China. (e-mail: guolijun@nbu.edu.cn, e-mail: gaolinlin@nbu.edu.cn)

Huifang Ma is with the College of Computer Science and Engineering, Northwest Normal University , Lanzhou Gansu, China. (e-mail: mahuifang@nwnu.edu.cn)

Jianhua Wang is with the Department of Radiology, the Affiliated Hospital of Medicine School of Ningbo University, Ningbo Zhejiang, China. (e-mail: woxingw@sina.com)

2168-2194 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Univ of Calif San Francisco. Downloaded on June 02,2020 at 18:05:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2994114, IEEE Journal of Biomedical and Health Informatics

information. However, during encoding high-level features, the pooling layer reduces the size of the feature map and loses boundary information. Fusion of multi-scale features in semantic segmentation has been found to address these problems by effectively aggregating complementary information and complementing the missing boundaries of feature maps [15]-[17]. In the PSPNet [13] or DeepLab [18] system, spatial pyramid pooling is performed at different grid scales or dilating rates. The pyramid pooling module proposed in the PSPNet only fuse different scale features at a single scale level, resulting in insufficient multi-scale information. In addition, although some works [19], [20] capture different scale objects by fusing context, they do not consider effective fusion of local and global information. In medical MR image segmentation, we need to not only learn the relationship between different blocks on a channel but also model the relationship of semantic information from different channel images. Moreover, most multi-scale methods adopt complicated decoder modules that use low-level information to help high-level features recover image details. However, they are indistinguishable to fuse multi-scale features. Due to similarity in the components in the hippocampus and some surrounding tissues, inaccurate boundary information may be obtained from low-level features, leading to poor performance or even incorrect predictions.

In this study, we aim at extracting discriminative features for hippocampal segmentation. First, considering the problem that the hippocampus is a rather small part in MR images, the context features of scale invariance are extracted in the high-level stage of the encoder to further enrich hippocampal semantic information, with maintaining high-resolution representations throughout the process and repeatedly fusing multi-scale subnet features. Furthermore, our segmentation framework is designed specially to capture the spatial and channel dependencies with two attention modules. Specifically, we aggregate and update the features of all locations on the low-level feature maps and all channels on the high-level feature maps by weighted summation, and then model the long-range dependencies between pixels at different locations on the feature map and the long-range dependencies between different channels. Finally, we build an effective decoder module to further suppress the noisy boundaries of hippocampus, and simultaneously take advantage of the attentional low-level features from the encoder to better guide high-level hippocampus features restore the boundaries in the decoder phase.

The main contributions of this study can be summarized as follows:

1. We propose a **context-aware high-level feature extraction module (CHFEM)** on high-level features to extract scale-invariant features. The CHFEM can capture multi-scale discriminative information for a small-sized input sample by repeatedly fusing high-to-low subnet features.
2. We design a **low-level feature spatial attention module (LFSAM)** and a **high-level feature channel attention module (HFCAM)** to capture the spatial and channel dependencies between any two positions of the low-level feature maps and between any two channels of the high-level feature maps, respectively. To the best of our knowledge, this is the first work to apply an attention mechanism to the hippocampal segmentation task. And we achieve state-of-the-art performance on the ADNI dataset.
3. We develop a **cross-connected attention module (CCAM)** between the encoder and decoder to extract the global context of high-level features via global max-pooling and global average-pooling. The global information from high-level features can be used as a guide to weight low-level features, and then select low-level features that are more valuable to segmentation results, thereby helping high-level features restore the boundaries.
4. We achieve new state-of-the-art results on two popular benchmarks, namely ADNI dataset and Decathlon dataset.

The rest of the paper is organized as follows. Section II reviews the latest developments in semantic segmentation tasks. Section III introduces data preprocessing and the details of the four modules. In Section IV, we extensively investigate the performance of the proposed method under different parameters and verify the rationality and effectiveness of each step of the proposed method. Finally, the paper is summarized in Section V.

II. Related Works

Hippocampus Segmentation: Recently, with the development of deep learning, techniques such as convolutional neural networks (CNNs) have been used in hippocampal segmentation. In 2018, Thyreau et al. [21] proposed a deep-learning appearance model by transferring algorithmic knowledge to segment the bilateral hippocampi. Cao et al. [22] proposed a multi-task deep-learning (MDL) method for joint hippocampal segmentation and clinical score regression using MRI scans. In 2019, Liu et al. [23] proposed a new 3D densely connected model based on 3D patches to extract and learn rich hippocampal features. However, the MRI datasets used for AD diagnosis are typically extremely small compared with the datasets used in computer vision and the hippocampus is small in size with no obvious boundary at the edge. Training deeper network models with a large number of parameters for hippocampal segmentation remains a major challenge [24]. Inspired by previous studies, in order to further improve the segmentation accuracy, we aim to extract discriminative features for hippocampal segmentation by using the hierarchical attention mechanism and to enhance the feature representation of the model.

Multi-Scale Representation: Approaches toward the application of encoding the multi-scale context information are widely explored. The typical construction of an image pyramid [25], [26] is frequently used, resulting in various scales of objects in the network. Dilated or atrous convolution [14]-[18] deployed in parallel or in a cascaded structure expands the receptive fields while exhibiting no extra parameters. Further, atrous spatial pyramid pooling (ASPP) modified atrous convolution in parallel within spatial pyramid pooling to efficiently capture features of an arbitrary scale. In particular, in the PSPNet [13] or DeepLab [18] system, spatial pyramid pooling is performed at different grid scales. However, the pyramid pooling module proposed in the PSPNet fuse different scale features only at a single scale level, but not at different

2168-2194 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Univ of Calif San Francisco. Downloaded on June 02,2020 at 18:05:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2994114, IEEE Journal of Biomedical and Health Informatics

scale levels, thus resulting in insufficient multi-scale information. In the ASPP module, dilated convolution is a type of sparse calculation that may cause grid artifacts [27]. In contrast to the above methods, inspired by the HRNet [28], we propose the CHFEM to extract the features of scale invariance. Due to the small size of the hippocampus, it is impossible to construct a deep network in a small input sample model. Therefore, unlike in the HRNet, the proposed CHFEM captures rich discriminative semantic information only on high-level features. Specifically, we connect high-to-low subnets in parallel, which can accurately estimate spatial heat maps by maintaining high-resolution representations throughout the process and repeatedly fusing high-to-low subnet features.

Attention Mechanisms: Attention mechanisms have been successfully applied to various tasks [29], [30]. Ashish et al. [31] first proposed the self-attention mechanism to draw global dependencies of inputs and applied it in machine translation. Meanwhile, attention modules have been increasingly applied in the image vision field. Han et al. [32] introduced the self-attention mechanism for the learning of an improved image generator. Zhao et al. [13] mainly explored the effectiveness of non-local operation in space-time dimension for videos and images with a self-attention module. However, they did not consider how to effectively fuse local and global information. As convolution extracts features through fixed-size local receptive fields, it is difficult to simultaneously consider local and global information. To overcome these problems, an attention model was introduced to the semantic segmentation network. Fu et al. [33] proposed a dual attention network (DANet) to enhance the discriminant ability of high-level feature representations for scene segmentation. However, the DANet ignores the different characteristics of the high-level and low-level features, which may affect the extraction of effective features. Because features of different layers have different semantic values for generating significant feature maps, high-level features usually contain global context-aware information, which is suitable for correct classification, while low-level features contain spatial structure details and are suitable for locating boundaries. Therefore, different from DANet, we introduce the LFSAM and the HFCAM to capture the spatial and channel dependencies between any two positions of low-level feature maps and between any two channels of high-level feature maps, respectively.

Encoder–decoder: Most state-of-the-art segmentation frameworks are based on encoder–decoder networks [12], [34]-[36], which have also been successfully applied to many segmentation tasks. Some types of U-shape networks such as SegNet [37], Refinenet [38], Large Kernel Matters [39], and even U-Net [34], which is widely used in the field of medical image segmentation, involve a complicated decoder module that uses low-level information to help high-level features recover image details. However, most methods of U-shape networks are indistinguishable to fuse multi-scale features. Due to the presence of similar components in the hippocampus and some surrounding tissues, low-level features may provide inaccurate boundary information, which can lead to poor performance or even incorrect predictions. To solve this issue, inspired by the global attention upsample (GAU) [35] module, we build an effective decoder module, i.e., CCAM, during cross connections, which can extract global context of high-level features as guidance to weight low-level feature information. In contrast to GAU, along with average-pooling, we also use max-pooling to obtain global context of high-level features. Our experiments prove that jointly using these two features can greatly improve the representation power of the networks rather than using average-pooling alone.

III. Materials And Methods

Herein, we present the details of the proposed segmentation method, including the data preprocessing and the structure of our network.

A. Data preprocessing

We adopt the public Alzheimer's Disease Neuroimaging Initiative (ADNI) database [40] in this study. Specifically, the baseline ADNI database contains 1.5T T1-weighted structural MRI data of 140 subjects, including 48 normal controls (NC), 45 subjects with mild cognitive impairment (MCI), and 47 AD subjects. We pre-process all studied MRI data using a standard pipeline. Specifically, we first resample all images to have the same size of 192×192×160, followed by intensity inhomogeneity correction via the N3 algorithm [41]. Then, we linearly align all images onto a template image. Note that, in this study, we do not need processes such as skull stripping or cerebellum removal. In addition, no nonlinear registration is required in image pre-processing.

After pre-processing, all MR images are aligned onto a common template space wherein we define a bounding cube for the hippocampus and extract an image patch from this box of size 32×32×32. As a result, an area containing the hippocampus of all test objects is obtained, which is considered the region of interest (ROI) [42]. Because of the relatively small size of the hippocampus in the brain, this process helps us eliminate considerable confounding background information. Otherwise, the number of voxels in the background (i.e., negative samples) will be considerably larger than that of voxels in the hippocampus region (i.e., positive samples), leading to a severe class-imbalance problem. To model the structural similarity between adjacent slices and reduce the computational complexity, we adopt the 2.5D image patch extraction method. Particularly, we believe that there is complementary information between continuous slices of the MR image, therefore, we extract 2D slice containing the hippocampal information from ROI along the depth, and simultaneously extract the two adjacent slices before and after it. Finally, these three 32×32 images are used as three channels of the input image; subsequently, the spatial information of the hippocampus can be learned by 2D convolution. In this study, we use random rotation and translation to increase the number of training images, which not only ensures that enough information is extracted from the MR images but also effectively inhibits the over-fitting problem of the CNN. During the training process, since the output of the network is a single-channel segmentation map, we only perform 2.5D image processing on the training data, and we input the corresponding single-channel ground truth. The test set is processed by the same procedure as that of the training set.

B. Structure of Network

In this section, we present a novel method for hippocampal

2168-2194 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Univ of Calif San Francisco. Downloaded on June 02,2020 at 18:05:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2994114, IEEE Journal of Biomedical and Health Informatics

segmentation, and the overall framework of this study, which involves a process of encoding and decoding. The specific ideas of the algorithm are as follows.

1) Overview

Encoder part: As illustrated in Fig. 1, in the down sampling stage, we adopt different convolutional layers to extract hippocampal features. Inspired by DenseNet [43], we add dense connections to encoders. Specifically, each convolutional block is composed of a repeated cascaded structure of two 3 × 3 convolution layers, all of which are followed by a batch normalization and a ReLU activation function. The shallow information extracted from the first two layers is considered low-level features, whereas the deep information extracted starting from the third layer is considered high-level features. The encoder part contains a **CHFEM** on high-level features to extract the information of scale invariance; it can capture multi-scale discriminative information from small-sized samples. In addition, we adopt an **HFCAM** on high-level features to model the association of semantic information of different channel images, and two **LFSAMs** on low-level features to aggregate similar features at all locations selectively.

Decoder part: As illustrated in Fig. 1, in the up-sampling stage, we restore the edge details of high-level features. Similarly, we add dense connections to decoders. In addition, unlike in U-Net, we set two **CCAMs** during cross connection at this stage. It can select more efficient low-level features to help high-level features restore boundaries.

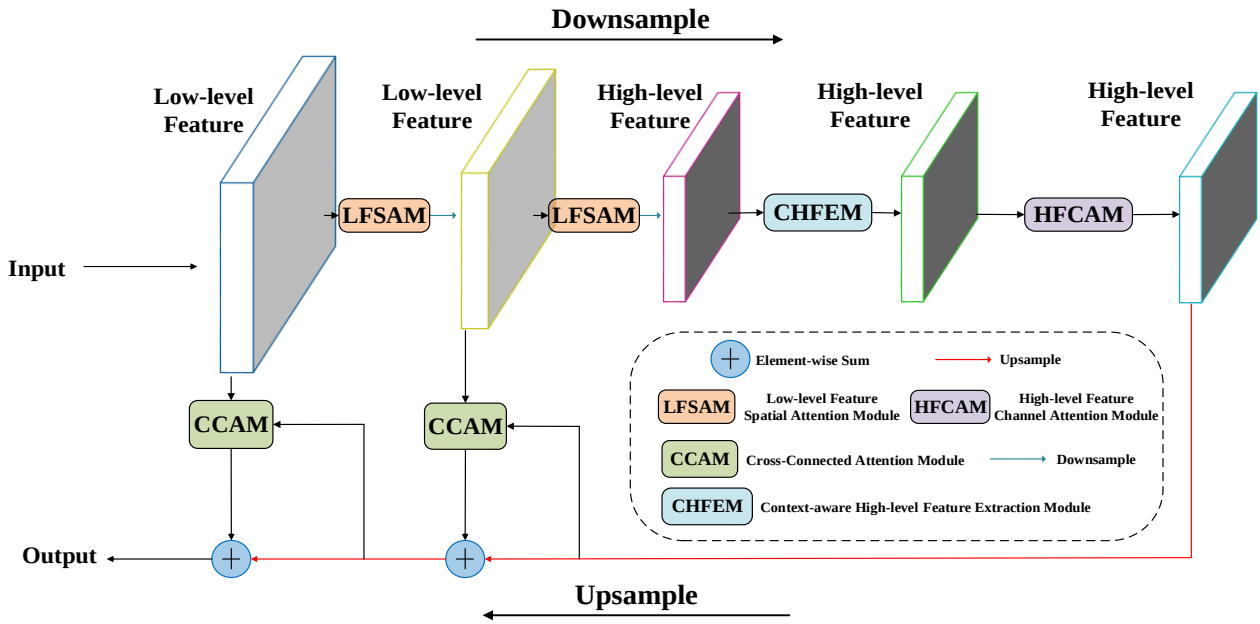

Fig. 1. Overall architecture of the proposed method. (Best viewed in color)

2) Context-aware High-level Feature Extraction Module

Due to small size of the hippocampus, common deep networks cannot achieve accurate segmentation. Therefore, we extract multi-scale context features in the high-level stage of the encoder to further improve the hippocampal semantic information. It has been revealed that features from different layers of the network are complementary [25], [26]. The fusion of multi-scale features in semantic segmentation can effectively aggregate complementary information, and help complement the missing boundaries of feature maps. However, discriminative features cannot be effectively extracted with a deep network based on small-sized samples. Therefore, we propose a CHFEM on a high-level stage to capture the context information of multiple receptive fields, and the final extracted high-level features are scale invariant. Fig. 2 illustrates the structure of the CHFEM.

We consider the third encoder layer information presented in Fig. 1 as the basic high-level features. To extract high-level features of scale invariance, we connect high-to-low subnets in parallel, which can accurately estimate spatial heat maps by maintaining high-resolution representations throughout the process and repeatedly fusing high-to-low subnet features.

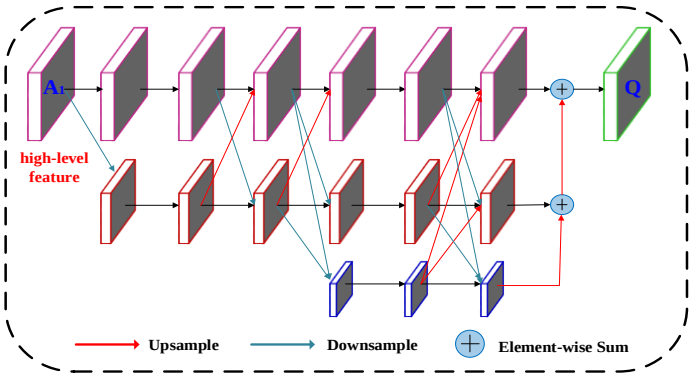

Fig. 2. Details of context-aware high-level feature extraction module. (Best viewed in color)

As illustrated in Fig. 2, we down sample high-level features (A_1 in Fig. 2) into three different resolution features. First, we start from a high-resolution subnetwork as the first stage, and gradually connect high-to-low multi-resolution subnetworks in parallel one by one, forming new stages. The hippocampal feature is extracted by repeated 3 × 3 convolution with dense connections between same resolutions; the resolution is halved by the pooling operation between adjacent resolutions. In

2168-2194 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Univ of Calif San Francisco. Downloaded on June 02,2020 at 18:05:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2994114, IEEE Journal of Biomedical and Health Informatics

addition, we fuse multi-scale features at different scale levels so that each subnetwork repeatedly receives the information from other parallel subnetworks. Subsequently, we obtain three different scale features with context-aware information, of which two smaller ones are then upsampled to the largest one. Finally, we combine them by element-wise summation to obtain the output (Q in Fig. 2) of the CHFEM.

3) *Spatial Attention Module and Channel Attention Module*

In medical MR image segmentation, not only is it important to learn the relationship between different blocks on a channel but also to model the relationship of semantic information on different channel images. However, because convolution extracts feature through fixed-size local receptive fields, it is difficult to simultaneously consider local and global information. To solve this problem, we introduce two attention modules to capture the spatial and channel dependencies between any two positions of the feature maps and between any two channel maps, respectively.

Low-level Feature Spatial Attention Module: Since the convolution of the encoder stage can only model local features, it lacks the ability to model a larger range of spatial relationships. In addition, features of different layers have different semantic values for generating significant feature maps; low-level features contain spatial structural details and are suitable for locating boundaries. By introducing a spatial attention in two low-level layers, we can significantly enhance the spatial structure description ability of the feature of this layer, which plays a guiding role in segmentation. For example, the spatial correspondence between the head and tail of the hippocampus contributes to hippocampus segmentation. Therefore, we only build a spatial attention module for low-level features. Next, we elaborate the process to adaptively aggregate spatial contexts.

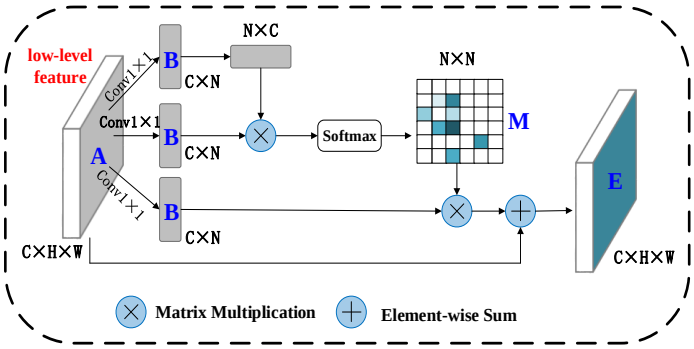

Fig. 3. Details of low-level feature spatial attention module. (Best viewed in color)

For the LFSAM, we introduce the self-attention mechanism to capture the spatial dependencies between any two positions of the low-level feature maps. As illustrated in Fig. 3, the low-level feature map A $\in R^{C\times H\times W}$ first obtains three identical feature maps B through three convolution layers, respectively, where B $\in R^{C\times H\times W}$; these feature maps are then reshaped to $R^{C\times N}$, where N = H × W is the number of pixels. Next, matrix multiplication between the transpose of B and B is performed, and a softmax layer is applied to calculate the spatial attention map M $\in R^{N\times N}$:

$$M_{ji} = \frac{exp\,(B_i \cdot B_j)}{\sum_{i=1}^{N} exp\,(B_i \cdot B_j)} \quad (1)$$

where M_{ji} indicates the i^{th} position's impact on the j^{th} position. In addition, a matrix multiplication between B and the transpose of M is performed and the result is reshaped to $R^{C\times H\times W}$. Finally, this is multiplied by a scale parameter α and an element-wise sum operation with features A is performed to obtain the final output E $\in R^{C\times H\times W}$ as follows:

$$E_j = \alpha \sum_{i=1}^{N} (M_{ji}\, B_i) + A_j \quad (2)$$

where α is initialized to 0 and gradually learns to assign more weight. It can be inferred that the resulting feature E at each position is a weighted sum of the features across all positions and original features. The spatial attention modules enhance their representation capabilities by encoding a wide range of contextual information into local low-level features.

High-level Feature Channel Attention Module: Different channels model different semantic information. In addition, high-level features usually contain global context-aware semantic information, which is suitable for accurate classification. Therefore, by adding a channel attention to high-level different channels, we can enhance the semantic correlation between different channels, which also plays a guiding role in segmentation. Therefore, we build a channel attention module only for high-level features. Our HFCAM is built behind the CHFEM, which aims to further enhance the capability of multi-scale high-level semantic feature representation.

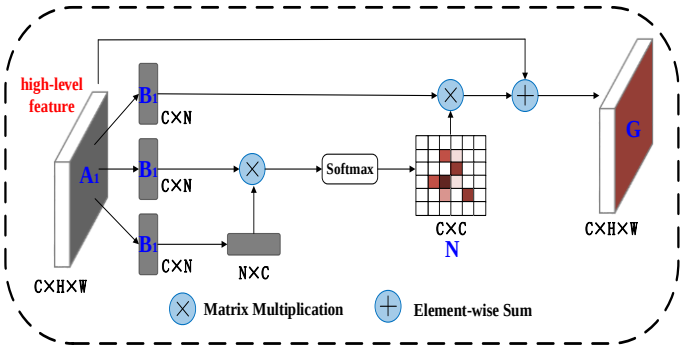

Fig. 4. Details of high-level feature channel attention module. (Best viewed in color)

For the HFCAM, we use the similar self-attention mechanism used for the LFSAM to capture the channel dependencies between any two channel maps, and update each channel map with a weighted sum of all channel maps. The structure of the HFCAM is illustrated in Fig. 4. The high-level feature map $A_1 \in R^{C\times H\times W}$; different from the LFSAM, A_1 is directly reshaped to $R^{C\times N}$, where N = H × W is the number of pixels. Further, this high-level feature map A_1 obtains three identical feature maps B_1, where $B_1 \in R^{C\times N}$. Then, a matrix multiplication is performed between B_1 and the transpose of B_1, and a SoftMax layer is applied to obtain the channel attention map N $\in R^{C\times C}$:

$$N_{ji} = \frac{exp\,(B1_i \cdot B1_j)}{\sum_{i=1}^{C} exp\,(B1_i \cdot B1_j)} \quad (3)$$

where N_{ji} indicates the i^{th} channel's impact on the j^{th} channel. In addition, a matrix multiplication is performed between the transpose of N and B_1 and their result is reshaped to $R^{C\times H\times W}$. Finally, the result is multiplied by a scale parameter β and an

2168-2194 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Univ of Calif San Francisco. Downloaded on June 02,2020 at 18:05:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2994114, IEEE Journal of Biomedical and Health Informatics

element-wise sum operation with the features A_1 is performed to obtain the final output $G \in R^{C\times H\times W}$ as follows:

$$G_j = \beta \sum_{i=1}^{C} (N_{ji} B_{1_i}) + A_{1_j} \quad (4)$$

where β is initialized as 0 and gradually learns to assign more weight. It can be inferred that the resulting feature G at each channel is a weighted sum of the features across all channels and original features. The channel attention module encodes a wider range of semantic dependencies between high-level feature maps, thus boosting feature discriminability.

4) *Cross-Connected Attention Module*

Automatic segmentation of the hippocampus in MR images is a challenging process. The gray levels of the hippocampus in MR images are very similar as other neighboring structures, such as the amygdala, caudate nucleus, and thalamus. In addition, there are no well-defined borders around the hippocampus with these adjacent regions, which increases the difficulty of hippocampus segmentation. Recent research has shown that some types of U-shape networks involve complicated decoder modules that use low-level information to help high-level features recover images detail. However, these methods are indistinguishable to fuse multi-scale features. Due to similar texture between the hippocampus and some surrounding tissues, the boundaries from low-level features may be inaccurate, leading to poor performance or incorrect predictions. To resolve this issue, we build an effective decoder module, the CCAM, during cross connections, which can further suppress the noisy features of the hippocampus, as well as make use of the attentional low-level features from the encoder to better guide the high-level hippocampus features restore the boundaries.

At present, for aggregating spatial information, average-pooling has been commonly adopted. Li et al. [35] used it in their GAU to learn global context, and Hu et al. [44] adopted it in their attention module to compute spatial statistics. In addition to the findings of the previous works, we believe that max-pooling collects other important clues about distinctive object features, which helps achieve more refined channel-wise features. Therefore, we use both average-pooled and max-pooled features simultaneously. We empirically demonstrate that exploiting both features greatly improves the representation capability of the network rather than using average-pooling alone. We describe the details below.

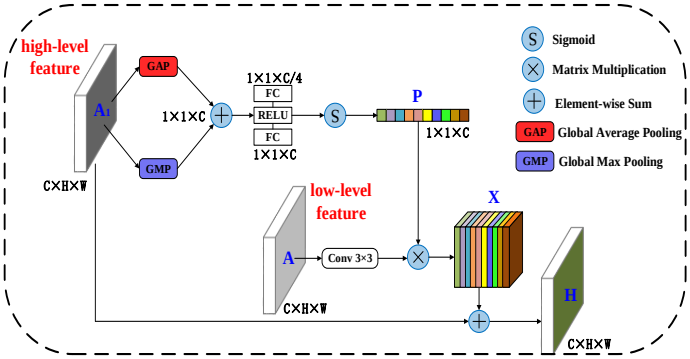

Fig. 5. Details of cross-connected attention module. (Best viewed in color)

As illustrated in Fig. 5, the spatial information of $A_1 \in R^{C\times H\times W}$ is first aggregated by using the average-pooling and max-pooling operations, generating two different spatial context descriptors, F^c_{avg} and F^c_{max}, which represent the average-pooled and max-pooled features, respectively. Then, they are merged using element-wise summation. In addition, to limit the complexity of the model and aid generalization, the gating mechanism is parameterized by forming a bottleneck with two fully connected (FC) layers around the non-linearity. Subsequently, the global context attention map $P \in R^{1\times 1\times C}$ is obtained through sigmoid operation. Next, 3 × 3 convolution on the low-level features (A in Fig. 5) is performed to obtain $O \in R^{C\times H\times W}$. Then, a matrix multiplication is performed between O and the transpose of P and their result is reshaped to obtain $X \in R^{C\times H\times W}$. In short, the global attention is computed as

$$P = \sigma\left(W_1\left(\delta\left(W_0\left(F^c_{avg} + F^c_{max}\right)\right)\right)\right) \quad (5)$$

where W refers to parameters in a channel-wise attention block, $W_0 \in R^{1\times 1\times \frac{C}{4}}$ and $W_1 \in R^{1\times 1\times C}$; σ denotes the sigmoid function; and δrefers to the ReLU function. Finally, we perform an element-wise sum operation between A_1 and X to obtain the final output $H \in R^{C\times H\times W}$.

IV. Experiments And Results

To evaluate the proposed method, we carry out comprehensive experiments on ADNI dataset and dataset from the Medical Segmentation Decathlon challenge. Experimental results demonstrate that proposed method achieves state-of-the-art performance on two datasets. In the following section, we will first introduce the dataset and implementation details, and then performed empirical comparation with a few other competing methods. We also compared the results with different parameter setting.

A. Dataset and Implementation Details

ADNI Dataset: The baseline ADNI database contains 1.5T T1-weighted structural MRI data of 140 subjects, to evaluate the proposed hippocampal segmentation method, we divided 140 subjects into five parts; one (28 subjects) was used for testing, while the other 4 parts (112 subjects) were for training. For the training process, we used cross validation with 20% of the training data used as a validation set. Then, the training parameters were updated iteratively according to each validation result. The testing set was not used for model training and parameter tuning but for general performance evaluation.

Decathlon Dataset: We selected the hippocampus dataset from the Medical Segmentation Decathlon challenge [45], including 265 training data and 130 test data; each training image has a unique label. In the training process, we adopted the same cross-validation method to adjust the network parameters.

Implementation Details: We implemented our method using Keras and TensorFlow. We employed a poly learning rate policy, where the initial learning rate is multiplied by $\left(1 - \frac{iter}{total\_iter}\right)^{0.9}$ after each iteration. The base learning rate was set to 0.01 for the ADNI dataset. Momentum and weight decay coefficients were set to 0.9 and 0.00001, respectively. We trained our model with Synchronized BN. The batch size was

2168-2194 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Univ of Calif San Francisco. Downloaded on June 02,2020 at 18:05:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2994114, IEEE Journal of Biomedical and Health Informatics

set to 32 and the training time was set to 300 epochs for the ADNI dataset. For data augmentation, we applied random rotation and random translation during training for the ADNI dataset. In addition, the dice similarity coefficient (DSC) was measured to evaluate the performance of the proposed method. The segmentation results with a higher DSC represent better segmentation performance.

$$\mathrm{DSC} = \frac{2\left(V_{fcn} \cap V_{manu}\right)}{V_{fcn} \cup V_{manu}} \tag{6}$$

where V_{fcn} represents the results of the segmentation method and V_{manu} represents the information of the corresponding label.

B. Results on Hippocampal Segmentation

1） Results on ADNI Dataset

Comparison of different methods: In this experiment, we compared the results of our method with the latest results of other four hippocampal segmentation methods including multi-alas based method [46], Thyreau's method [21], Cao's method [22] and Liu's method [23]. For these four methods, only the trained model of the Thyreau's method was available online (https://github.com/bthyreau/hippodeep), we downloaded it to test on our dataset for comparison. Experimental results of the other three methods were obtained from related references. Table I shows the comparison of segmentation results in terms of DSC on the ADNI dataset. In addition, we further compared the experimental results with some deep semantic segmentation models on the ADNI dataset. Since U-Net has obtained significant advantages in the field of medical image segmentation, more and more improved models based on U-Net have been developed. Table II shows the different segmentation results of our method and some state-of-the-art improved U-Net models on the ADNI dataset. These models include U-Net [34], U-Net++ [47], Attention U-Net [48] and nnU-Net [49]. Their code was available online, and we downloaded it to test on our ADNI dataset for comparison. Furthermore, Fig. 6 visualizes the segmentation results of different improved U-Net models. The original images, corresponding ground truth and the segmentation results of different deep models are demonstrated in columns from left to right in Fig. 6.

TABLE I
DSC OF DIFFERENT HIPPOCAMPAL SEGMENTATION APPROACHES ON THE ADNI TEST DATASET

Method	DSC%
Multi-alas based on method (2017)	87.11
Thyreau's method (2018)	73.48
Cao's method (2018)	85.63
Liu's method (2019)	87.00
Proposed Method	**91.24**

TABLE II
DSC OF DIFFERENT SEMANTIC SEGMENTATION APPROACHES ON THE ADNI TEST DATASET

Method	DSC%
U-Net (2015)	81.24
U-Net++ (2018)	86.03
Attention U-Net (2018)	88.45
nnU-Net (2019)	87.91
Proposed Method	**91.24**

From Table I and Table II, our method achieves the best performance compared to other methods. In addition, from the original images in Fig. 6, we notice that it is not easy to distinguish the hippocampal regions from the adjacent tissues due to the small difference between their intensity values. Proposed network can capture the overall contour of the hippocampus well after the training is completed, and the segmented hippocampal regions obtained by our method appear to be smoother and more accurate than those by other methods.

Ablation Study of Different Modules: To verify the performance of the CHFEM, HFCAM, LFSAM, and CCAM in hippocampus segmentation, we conducted experiments with different settings presented in Table III.

As shown in Table III, the proposed modules remarkably improved the performance. Compared with the baseline U-net (Residual Connection), employing the CHFEM yielded 80.21% DSC, resulting in 2.18% improvement. Meanwhile, the HFCAM individually outperformed the baseline by 3.36%. In addition, the LFSAM improved the performance from 78.03% to 80.92%. The CCAM yielded 4.42% improvement. When the four modules were integrated, the performance further improved to 85.74%. Furthermore, when we adopted dense connection in each layer, the network with four modules significantly improved the segmentation performance to 91.24% over the baseline model.

Visualization of Attention Module: For spatial attention based on low-level features, we visualized the feature maps of spatial attention in the early stages of training, as shown in Fig. 7c, it can be seen that the attention of network in the training process are mainly activated in the middle area, especially the region close to red, which is consistent with the position of the real hippocampus. It fully illustrates that the spatial attention module could capture clear semantic similarities and long-term relationships, and then locate the approximate position of hippocampus.

For channel attention based on high-level features, it is hard to directly give comprehensible visualization about the attention map. Instead, we show an attended channel to see whether they highlight semantic areas. In Fig. 7d, we displayed the fourth attended channel in the feature map of the encoder's last layer. We found that the area near the red color is the location of the most relevant features of the segmentation task, indicating that the response of specific semantic is noticeable after channel attention module enhances. In short, these visualizations further prove the necessity of capturing long-term dependencies for improving feature representation in hippocampal segmentation.

2168-2194 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Univ of Calif San Francisco. Downloaded on June 02,2020 at 18:05:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2994114, IEEE Journal of Biomedical and Health Informatics

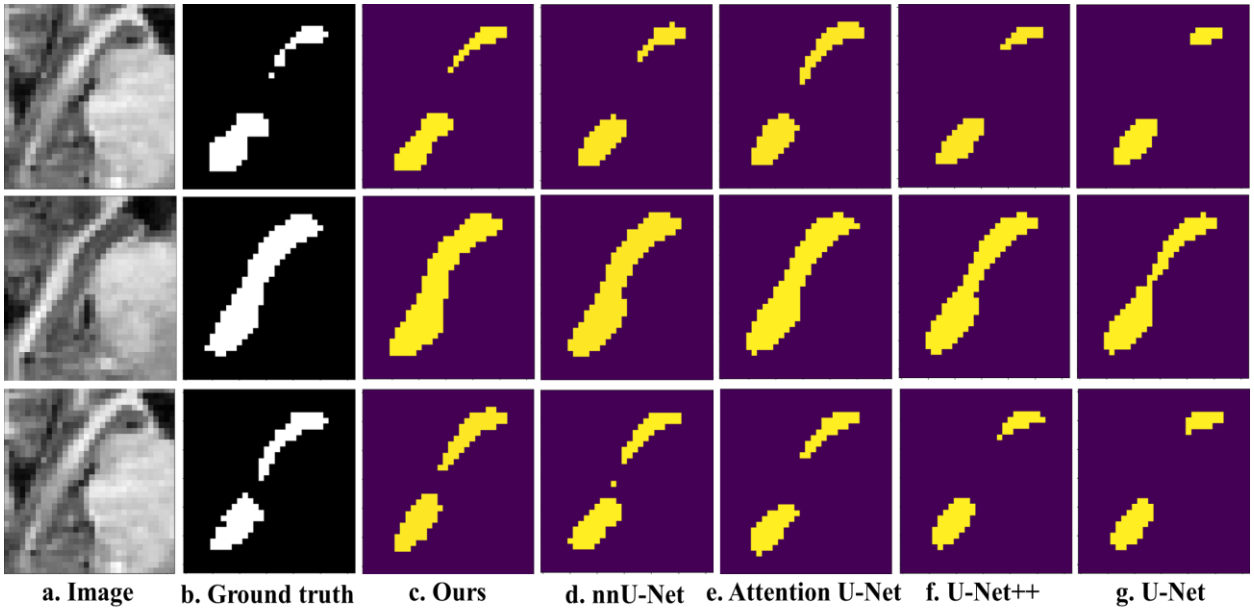

Fig. 6. Visual comparison of the proposed method and the state-of-the-art semantic segmentation algorithms on the ADNI test dataset. For each row, we show the input image and corresponding ground truth. Meanwhile, the segmentation results of the proposed method and nnU-Net, Attention U-Net, U-Net++ and U-net are provided.

TABLE III
ABLATION STUDY OF DIFFERENT MODULES ON THE ADNI DATASET

Method	Connection Type	CHFEM	HFCAM	LFSAM	CCAM	DSC%
U-net	Residual Connection					78.03
Ours	Residual Connection	√				80.21
Ours	Residual Connection		√			81.39
Ours	Residual Connection			√		80.92
Ours	Residual Connection				√	82.45
Ours	Residual Connection	√	√	√	√	85.74
U-net	Dense Connection					83.45
Ours	Dense Connection	√				85.13
Ours	Dense Connection		√			86.56
Ours	Dense Connection			√		85.91
Ours	Dense Connection				√	87.62
Proposed Method	**Dense Connection**	√	√	√	√	**91.24**

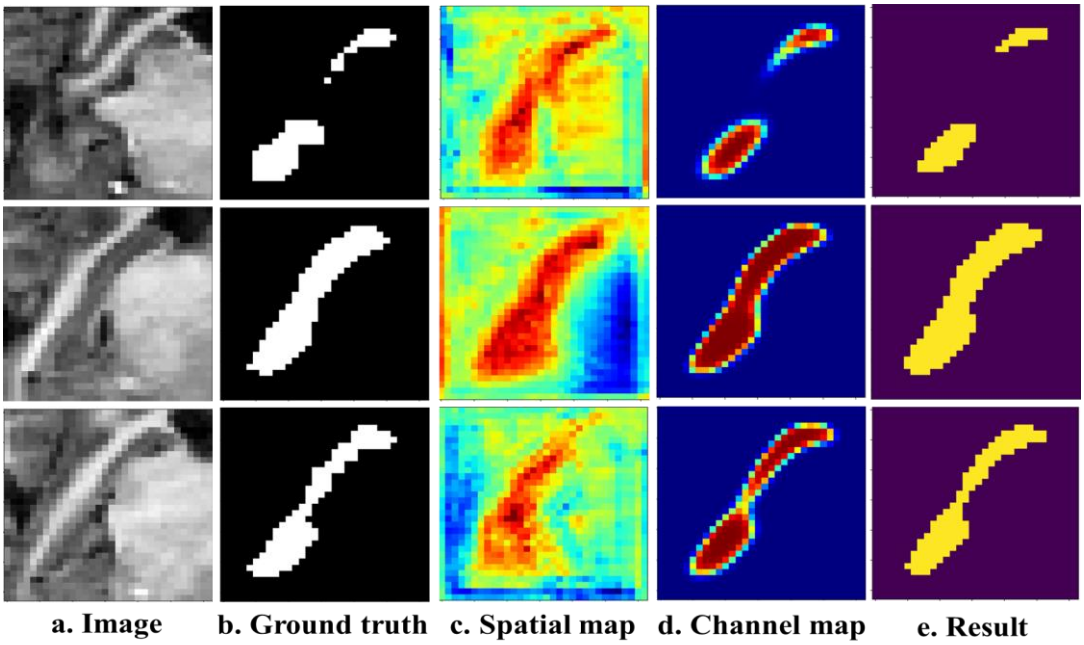

Fig. 7. Visualization results of attention modules on ADNI dataset. For each row, we show an input image and corresponding ground truth. Meanwhile, we give a spatial attention map and a channel attention map. Finally, the segmentation result of proposed method is provided.

2168-2194 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Univ of Calif San Francisco. Downloaded on June 02,2020 at 18:05:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2994114, IEEE Journal of Biomedical and Health Informatics

Comparison of Different Pooling Type in CCAM: We experimentally verified that using both average-pooled and max-pooled features enables finer attention inference in our CCAM. We compared three variants of CCAM: average pooling, max pooling, and combined use of both. Note that the CCAM with an average pooling was the same as the GAU [35] module. We conducted experiments using the different settings presented in Table IV.

TABLE IV
ABLATION STUDY OF DIFFERENT POOLING TYPE IN CCAM ON THE ADNI DATASET

Method	AvgPool	MaxPool	DSC%
U-net (Dense)			83.45
Ours (Dense)	√		89.76
Ours (Dense)		√	88.53
Proposed Method (Dense)	√	√	**91.24**

Experimental results of various pooling methods are presented in Table IV. We can observe that the max-pooled features were as significant as the average-pooled features in terms of the improvement of DSC from the baseline. However, in GAU's work, only the average-pooled features were considered, while ignoring the max-pooled features. We believe that the max-pooled features encoding the most significant part can compensate the average-pooled features that encode global statistics. Therefore, we recommend using both features and then integrating them through element-wise summation. Our empirical results show that our CCAM is an effective way to drive the segmentation performance further from that of GAU.

Analysis of Different Input Scales: This experiment was to test the impact of sample size on the performance of the model. Since the proposed model is built based on 2D image patches, so the size of the image patch has effects on hippocampus segmentation. Here, we gradually increased the image patch size from 24×24, 32×32 to 64×64 to test hippocampal segmentation. The Table V demonstrates that the segmentation performance is improved by decreasing the patch size from 64×64 to 24×24. However, smaller image patches might have limited contextual information, while larger image patches could include unnecessary background pixels increasing both overfitting risk and computational burden. With this observation, we set the image patch size to 32×32 in the following experiments.

TABLE V
STUDY OF DIFFERENT INPUT SCALES ON THE ADNI DATASET

Input Scale	DSC%
24×24	92.83
32×32 (Ours)	**91.24**
64×64	87.62

Analysis of Different Data Processing: To model the spatial structural information and reduce the computational complexity, we constructed two 2.5D image patch extraction methods. One is to learn the spatial information between different dimensions of MR image. Specifically, we extract three slices of the same number from ROI along the x-axis, y-axis, and z-axis, respectively, and they intersect in space. The other is to capture the structural relationship between adjacent slices in the same dimension. We extract the slice containing the hippocampal information from ROI along the depth, and simultaneously extract the two adjacent slices before and after it. Three 32×32 image patches from the previous two methods are used as three channels of the input image; subsequently, the spatial information of the hippocampus can be learned by 2D convolution. In addition, we compared the segmentation results of multi-channel input samples with single-channel input samples. The size of single-channel data is 32×32. The Table VI demonstrates that multi-channel input samples based on adjacent slices achieves the highest segmentation performance in our hippocampal segmentation model.

TABLE VI
STUDY OF DIFFERENT DATA PROCESSING ON THE ADNI DATASET

Method	DSC%
Single-channel Data	86.54
Different Dimensions	89.02
Adjacent Slices (Ours)	**91.24**

2） Results on Decathlon Dataset

Since the same hippocampal label definition was used for training and testing in our experiments, there may be a bias in the comparison of segmentation results because some competitive methods were not trained using this particular label definition. To address this issue, we tested our segmentation method on the hippocampus dataset from the Medical Segmentation Decathlon challenge to further evaluate the effectiveness of our model. Comparisons with state-of-the-art semantic segmentation models are reported in Table VII. Results show that our method achieves the highest segmentation performance, and it can capture the long-range contextual information more effectively and learn better feature representation in hippocampal segmentation.

TABLE VII
DSC OF DIFFERENT SEMANTIC SEGMENTATION APPROACHES ON THE DECATHLON TEST DATASET

Method	DSC%
U-Net (2015)	83.17
U-Net++ (2018)	85.93
Attention U-Net (2018)	87.45
nnU-Net (2019)	89.01
Proposed Method	**90.38**

V. CONCLUSION

This paper has proposed a deep network with a hierarchical attention mechanism to extract notably discriminative features for hippocampal segmentation. Experimental evaluations suggest that this method yields the best segmentation results on the ADNI dataset. The significant improvement stems from the combination of the context-aware multi-scale feature extraction and the hierarchical attention mechanism. Among them, the high-level context-aware features extracted by CHFEM are scale-invariant and solve the problem that deep networks cannot extract features of small-sized samples. In addition, the attention modules, including LFSAM, HFCAM, and CCAM, contribute to the discriminative feature extraction in different stages of the encoder–decoder framework. LFSAM and HFCAM capture the spatial and channel dependence of hippocampus, respectively, and further improve the expression of features. Furthermore, as the hippocampus is a gray matter structure, it has a low contrast with the surrounding tissues in MR images, CCAM not only inhibits the noisy boundaries with similar components in the hippocampus and surrounding tissues but also utilizes the attentional low-level features from the encoder to better guide the high-level hippocampus edge segmentation in the decoder phase. Above all, the proposed

2168-2194 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Univ of Calif San Francisco. Downloaded on June 02,2020 at 18:05:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2994114, IEEE Journal of Biomedical and Health Informatics

approach achieves state-of-the-art performance consistently on two hippocampus segmentation datasets, i.e. ADNI dataset and Decathlon dataset.

VI. References

[1] Braak H and Braak E, "Neuropathological stageing of Alzheimer-related changes," Acta Neuropathologica, vol. 82, no. 4, pp. 239-259, 1991.

[2] Nelson M D et al., "Hippocampal Volume Reduction in Schizophrenia as Assessed by Magnetic Resonance Imaging: A Meta-analytic Study," Archives of General Psychiatry, vol. 55, no. 5, pp. 433-440, 1998.

[3] Wieshmann U C et al., "Development of hippocampal atrophy: A serial magnetic resonance imaging study in a patient who developed epilepsy after generalized status epilepticus," Epilepsia, vol. 38, no. 11, pp. 1238-1241, 1997.

[4] Chupin M et al., "Fully Automatic Hippocampus Segmentation and Classification in Alzheimer's Disease and Mild Cognitive Impairment Applied on Data from ADNI," Hippocampus, vol. 19, no. 6, pp. 579-587, 2009.

[5] Ferreira L K and Busatto G F, "Neuroimaging in Alzheimer's Disease: current role in clinical practice and potential future applications," Clinics, 2011, pp. 19-24.

[6] Coupe P et al., "Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease," NeuroImage, vol. 59, no. 4, pp. 3736-3747, 2012.

[7] Duara R et al., "Medial temporal lobe atrophy on MRI scans and the diagnosis of Alzheimer disease," Neurology, vol.72, no. 24, pp. 1986-1992, 2008.

[8] B. Fischl et al., "Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain," NeuroImage, vol. 33, pp. 341-355, 2002.

[9] Alex Krizhevsky et al., "Imagenet classification with deep convolutional neural networks," In Advances in Neural Information Processing Systems, 2012, pp. 1097-1105.

[10] Kaiming He et al., "Deep residual learning for image recognition," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[11] Ziwei Liu et al., "Semantic image segmentation via deep parsing network," In Computer Vision (ICCV), 2015 IEEE International Conference on, 2015, pp. 1377-1385.

[12] Liang-Chieh Chen et al., "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834-848, 2018.

[13] Hengshuang Zhao et al., "Pyramid scene parsing network," In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2881-2890.

[14] Jonathan Long et al., "Fully convolutional networks for semantic deconvolutionntic segmentation," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431-3440.

[15] Q. Hou et al., "Deeply supervised salient object detection with short connections," In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5300-5309.

[16] P. Zhang et al., "Amulet: Aggregating multi-level convolutional features for salient object detection," In Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 202-211.

[17] Y. Tang et al., "Deeply-supervised recurrent convolutional neural network for saliency detection," In Proceedings of the 2016 ACM on Multimedia Conference, 2016, pp. 397-401.

[18] Liang-Chieh Chen et al., "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.

[19] Chao Peng et al., "Large kernel matters - improve semantic segmentation by global convolutional network," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1743-1751.

[20] Hang Zhang et al., "Context encoding for semantic segmentation," In the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[21] Thyreau B et al., "Segmentation of the Hippocampus by Transferring Algorithmic Knowledge for large cohort processing," Medical Image Analysis, 2018, pp. 214-228.

[22] Cao L et al., "Multi-task neural networks for joint hippocampus segmentation and clinical score regression," Multimedia Tools and Applications, vol. 77, no. 22, pp. 29669-29686, 2018.

[23] Manhua Liu et al., "A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease," NeuroImage, 2019.

[24] Gray K R et al., "Regional analysis of FDG-PET for use in the classification of Alzheimer'S Disease," International Symposium on Biomedical Imaging, 2011, pp. 1082-1085.

[25] G. Lee et al., "Deep saliency with encoded low-level distance map and high-level features," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 660-668.

[26] Y. Tang and X. Wu, "Saliency detection via combining region-level and pixel-level predictions with cnns," In European Conference on Computer Vision, Springer, 2016, pp. 809-825.

[27] Panqu Wang et al., "Understanding convolution for semantic segmentation," arXiv preprint arXiv:1702.08502, 2017.

[28] Sun K et al., "Deep High-Resolution Representation Learning for Human Pose Estimation," arXiv: Computer Vision and Pattern Recognition (CVPR), 2019.

[29] Guosheng Lin et al., "Efficient piecewise training of deep structured models for semantic segmentation," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3194-3203.

[30] Zhouhan Lin et al., "A structured self-attentive sentence embedding," arXiv preprint arXiv:1703.03130, 2017.

[31] Vaswani A et al., "Attention is all you need," Neural Information Processing Systems, 2017, pp. 6000-6010.

[32] Zhang H et al., "Self-Attention Generative Adversarial Networks," International Conference on Machine Learning, 2019, pp. 7354-7363.

[33] Fu J et al., "Dual Attention Network for Scene Segmentation," arXiv: Computer Vision and Pattern Recognition (CVPR), 2018.

[34] Ronneberger O et al., "Liang-Chieh: Convolutional Networks for Biomedical Image Segmentation," Medical Image Computing and Computer Assisted Intervention, 2015, pp. 234-241.

[35] Li H et al., "Pyramid Attention Network for Semantic Segmentation," arXiv: Computer Vision and Pattern Recognition (CVPR), 2018.

[36] Changqian Yu et al., "Learning a discriminative feature network for semantic segmentation," arXiv preprint arXiv:1804.09337, 2018.

[37] Badrinarayanan V et al., "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481-2495, 2017.

[38] Guosheng Lin et al., "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[39] Chao Peng et al., "Large kernel matters–improve semantic segmentation by global convolutional network," arXiv preprint arXiv:1703.02719, 2017.

[40] Jack C R et al., "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods," Journal of Magnetic Resonance Imaging, vol. 27, no. 4, pp. 685-691, 2008.

[41] Sled J G et al., "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," IEEE Transactions on Medical Imaging, vol. 17, no. 1, pp. 87-97, 1998.

[42] Zhenzhou G et al., "Study on RBF neural network based on gray wolf optimization algorithm," Microelectronics & Computer, vol. 34, no. 7, pp. 7-10, 2017.

[43] Huang G et al., "Densely Connected Convolutional Networks," Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261-2269.

[44] Hu J et al., "Squeeze-and-Excitation Networks," arXiv: Computer Vision and Pattern Recognition (CVPR), 2017.

[45] Isensee F et al., "nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation," Computer Vision and Pattern Recognition (CVPR), 2018.

[46] Zhang J et al., "Detecting Anatomical Landmarks from Limited Medical Imaging Data Using Two-Stage Task-Oriented Deep Neural Networks," IEEE Transactions on Image Processing, vol. 26, no. 10, pp. 4753-4764, 2017.

[47] Zongwei Zhou et al., "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," Deep Learning in Medical Image Analysis (DLMIA), 2018.

[48] Oktay O et al., "Attention U-Net: Learning Where to Look for the Pancreas," arXiv: Computer Vision and Pattern Recognition (CVPR), 2018.

[49] Isensee F et al., "nnU-Net: Breaking the Spell on Successful Medical Image Segmentation," arXiv: Computer Vision and Pattern Recognition (CVPR), 2019.

2168-2194 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Univ of Calif San Francisco. Downloaded on June 02,2020 at 18:05:04 UTC from IEEE Xplore. Restrictions apply.