IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON DEEP LEARNING FOR COMPUTER-AIDED MEDICAL DIAGNOSIS

Received March 8, 2019, accepted March 26, 2019, date of publication April 1, 2019, date of current version April 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2908386

Nested Dilation Network (NDN) for Multi-Task Medical Image Segmentation

LIANSHENG WANG[1,2], (Member, IEEE), RONGZHEN CHEN[2], SHUXIN WANG[2], NIANYIN ZENG[3], XIAOYANG HUANG[2], AND CHANGHUA LIU[4]

[1]Fujian Key Laboratory of Sensing and Computing for Smart City, School of Information Science and Engineering, Xiamen University, Xiamen 361005, China
[2]Department of Computer Science, School of Information Science and Engineering, Xiamen University, Xiamen 361005, China
[3]Department of Instrumental and Electrical Engineering, Xiamen University, Xiamen 361005, China
[4]Department of Medical Imaging, Chenggong Hospital Affiliated to Xiamen University, Xiamen 361005, China

Corresponding authors: Xiaoyang Huang (xyhuang@xmu.edu.cn) and Changhua Liu (liuxingc@126.com)

This work was supported by the National Natural Science Foundation of China under Grant 61671399.

ABSTRACT The deep convolutional network has shown excellent performance in medical image analysis. However, almost all network variants are presented for one specific task, e.g., segment pancreas on computerized tomography (CT). In this paper, we propose a nested dilation network (NDN) which is applied to multiple segmentation tasks even for different modalities, including CT, magnetic resonance imaging (MRI), and endoscopic images. We design residual blocks nested with dilations (RnD Blocks) that catch larger receptive field in the first few layers to boost shallow semantic information. Besides, we apply the modified focal loss to help the network to provide more accurate segmentation results. We evaluate our method on five subtasks from medical segmentation decathlon challenge and GIANA2018, and the results show that our method achieves a better performance than the latest methods in each task. A lot of research works have been done recently to strengthen the learning power of the convolutional neural network (CNN) to get better performance.

INDEX TERMS Deep learning, medical segmentation, residual blocks nested with dilations, multi-task, focal loss.

I. INTRODUCTION

In recent years, with the historical opportunities brought by the growth of computing power and a large number of annotation data, deep learning has gained considerable attention [1]. Deep learning, especially deep convolutional neural networks (DCNN) has brought about great progress to computer vision in many aspects including image classification [2], detection [3], and segmentation [4], which is credited to its ability to learn abstractions in all levels from raw data automatically. In 1998, LeCun *et al.* proposed the LeNet [5] consisting of convolution layers, pooling layers and fully-connected layers which define the prototype of the CNN. After many years, AlexNet [6] presented by Krizhevsky *et al.* won the 2012 ImageNet competition by a large margin, which promotes ConvNets to become the mainstream of visual research. In subsequent years, with the emergence of deeper network architectures and various network variants such as VGGNet [7], Inception networks [8]–[11] and Residual networks [12], [13], deep convolutional networks have rapidly become a prominent methodology for computer vision.

Inspired by remarkable progress acquired by CNN, many medical image analysis groups enter this field and try to apply DCNN to the computer-aided diagnosis (CAD) systems [14]–[20]. Automated segmentation for organs and lesions is an important task in medical image analysis, which provides visualization of various anatomical structures in different modalities and the region information obtained by segmentation can be utilized for such purposes as computer-aided diagnosis or computer-assisted surgery [21]. Since Olaf *et al.* proposed UNet [22] for biomedical image segmentation, researchers have already published a lot of excellent work in this area. Li *et al.* proposed H-DenseUnet [23] to segment liver and tumor from CT volumes and achieved very competitive performance for liver segmentation even with a single model. Kamnitsas *et al.* employed DeepMedic [24], a 3D CNN architecture for brain lesion segmentation in MRI that helps improve disease diagnosis and treatment planning. CNNs also have been utilized to analyze the endoscopic

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

images. Wickstrom *et al.* presented a method on semantic segmentation of polyps by colonoscopy images [25].

However, as can be seen from the above, almost all methods are developed to perform one specific task, which typically focuses on a single organ in a single modality (e.g., CT or MRI). Inspired by wide applications of transfer learning in natural images, we consider that maybe there are similar features between different medical images and whether we can develop a framework that works out-of-the-box on many tasks. Moeskops *et al.* have investigated a similar topic before, using a single CNN architecture for different medical image segmentation tasks in different imaging modalities [26]. However, they only experimented with a small number of tasks and did not quantitatively analyze the gap between their results and the latest results for each task. From a methodological point of view, there are some main constraints as follows. First, the size of the different targets varies greatly, which causes difficulty in selecting the input size and designing the receptive field of network. For instance, the volume of the liver is much larger than the volume of the pancreas. Second, according to the doctor's diagnostic need, targets are imaged in different modalities where each pixel represents different information, and these distinctions require the network to re-understand to extract the corresponding features. Third, due to the characteristics of different lesions, for example, lung nodules are usually diagnosed on CT Volumes [27] while colonic polyps are detected on endoscopic slices, different tasks need to analyze various images range from two dimensions to three dimensions. Besides, small data, unbalanced labels, and multi-class labels also increase the challenge.

In this paper, we propose an end-to-end Nested Dilation Network (NDN) that can be used to segment organs or lesions in several highly different tasks. We design the Residual Blocks Nested with Dilations (RnD Blocks) in the first few layers to help network adapt to targets of any size. What's more, we insert squeeze-and-excitation blocks [28] into multiple nodes in the network to boost essential features for each task. Besides, we also modify the focal loss [29] to overcome small data, unbalanced labels, and multi-class labels. We evaluate our method on five subtasks from Medical Segmentation Decathlon challenge and GIANA2018 challenge presented in Fig. 5. Our method achieves better performance than the latest method in each task.

The remainder of this paper is organized as follows. In Section II, we introduce background and related technologies. In Section III, we set forth our method in detail. In Section IV, we present the experimental results and make a comparison between our method and other latest methods for each task. Finally, we present a conclusion in Section VI.

II. RELATED WORK

A. SEMANTIC SEGMENTATION

Semantic segmentation is intended to assign a categorical label to every pixel in an image. Most approaches are training an end-to-end fully convolutional network, which replaces the last fully-connected few layers with convolutional layers and recovers the downsampled feature maps to original maps through a decoder path consisting of deconvolutional layers or other upsampling layers. On this basis, many variants [30]–[32] have made great progress by merging richer features at various levels. What's more, some researchers [33] apply Conditional Random Fields (CRFs) to refine the prediction maps further.

B. DILATED CONVOLUTION

Dilated convolution is referred to as ''convolution with a dilated filter,'' which is first introduced in [34] for wavelet decomposition. In general, the network enlarges the receptive field through a series of downsampling operations. However, the loss of resolution or coverage brought by downsampling operations has a negative impact on the segmentation result. Dilated convolution handles this conflict by increasing the kernel size without increasing its calculation. Dilated convolution is first applied to semantic segmentation by Yu *et al.* in [35]. In [33], Chen *et al.* highlight the significance of dilated convolution in dense prediction and propose atrous spatial pyramid pooling (ASPP) to aggregate multi-scale receptive fields. Their subsequent work [36] further explore the application of dilated convolution in previous ASPP. After that, to alleviate the ''gridding issue'' caused by the standard dilated convolution operation, Wang *et al.* [37] develop the hybrid dilated convolution (HDC) module that consists of well-designed groups of dilation rates. Due to the mentioned advantages, dilated convolution has been extensively used to other computer vision tasks.

C. MEDICAL IMAGE SEGMENTATION WITH CNN

Medical image segmentation requires accurate dense prediction for organs or lesions to assist the computer-aided diagnosis or computer-assisted surgery. A broad range of work proves that CNNs outperforms previous methods which utilize the hand-craft features in medical image segmentation. In 2015, U-Net architecture demonstrated the dominance of dense end-to-end prediction for medical images. From then on, various variants of U-Net have been developed to segment lesions, organs or tissues. Due to the difference in anatomical structures of various parts, they need to be displayed in different imaging modalities such as CT, MRI, and Endoscopic where each pixel represents different information. Therefore, there is still a challenging task for CNNs to perform segmentation tasks on various data modalities.

III. METHOD

In this section, we will introduce our method in detail. Network function modules are described in III-A, III-B and III-C. The overall network structure is shown in III-D. At last, III-E illustrates our modified focal loss.

A. RESIDUAL BLOCKS NESTED WITH DILATIONS

In semantic segmentation, dilated convolution is seen as inserting ''holes''(zeros) between each pixel in the

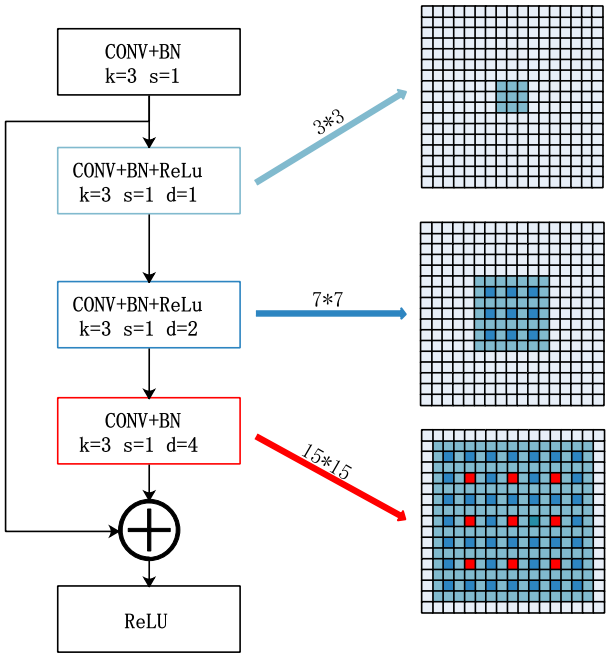


FIGURE 1. The schema of the RnD blocks (left) and corresponding receptive field (right) for each dilated convolution.

convolutional kernel [37]. For a $k * k$ dilated convolution kernel with dilate rate r, the receptive field of resulting dilated convolution filter is $k_r * k_r$, where $k_r = k + (k-1) * (r-1)$. Dilated convolution is employed in segmentation to maintain more extensive local information while keeping the same resolution. Fig. 1 depicts the schema of an RnD block, the RnD module superimposes three convolution layers with dilation rates 1, 2, 4 and obtains a 15×15 receptive field which is twice that of superimposes three standard convolution layers. To avoid the loss of detail result from sparse dilated convolution, we add a short-distance residual connection inside the RnD block, which helps recovery details. The RnD module operates like the equation defined below:

$$X_{i+1} = F(r_3, F(r_2, F(r_1, X_i))) + X_i \tag{1}$$

where F represents the convolution function and r_i denotes the dilation rate, *e.g.* r_1, r_2, r_3 correspond to 1, 2, 4.

B. RESIDUAL BLOCKS

Convolution neural network's performance is closely related to its depth, but simply stacking convolution layers may result in vanishing/exploding gradients when back propagation [12]. Residual connections alleviate this issue by directly detouring the input information to the output and protect the integrity of the gradient flow. The equation 2 illustrates the working principle, where x is the input, and F is the residual mapping function. Instead of learning a complex nonlinear mapping directly, residual modules asymptotically approximate a residual function which has the similar effect as the desired mapping. Residual function simplifies the optimization objective without extra parameters but improves the performance, which inspires the applications of deeper

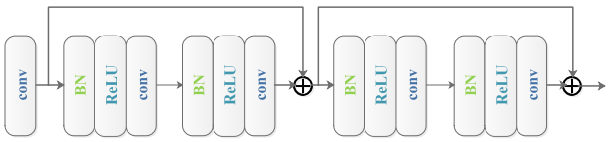


FIGURE 2. The schema of the residual blocks.

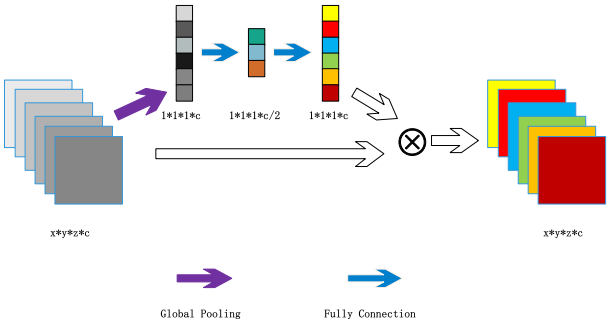


FIGURE 3. The schematic diagram of the proposed network architecture. Arrows represent different operations explained below.

residual-based networks. As shown in Fig. 2, we stack five convolutional layers in each residual block to leverage advantages of the residual learning.

$$y = F(x, W_i) + x \tag{2}$$

C. SQUEEZE-AND-EXCITATION BLOCKS

As the core of the convolutional neural network, the convolution kernel is usually regarded as the aggregation of spatial information and channel-wise information on the local receptive field. Hu *et al.* [38] consider the relationships between feature channels and put forward a novel module, namely Squeeze-and-Excitation Blocks (SE Blocks). SE Blocks learn channel weights through global spatial information that emphasizes the effective feature maps and suppresses the low-effect feature maps. Given a 4D input X of size $W * H * D * C$ where W, H, D correspond to different spatial dimensions and C represents the number of channels. As shown in Fig. 3, squeeze-and-excitation blocks recalibrate the previously obtained features through three operations. Firstly, the squeeze operation compresses features in spatial dimensions, responds to global distribution in each feature channel, and has the global receptive field to some extent. Then the excitation operation generates the weight for each channel, which works like the mechanism of gates in recurrent neural networks. Two fully-connection layers achieve the excitation vector and are activated by a sigmoid function. Finally, reweight input feature maps with excitation vector channel by channel.

D. NETWORK ARCHITECTURE

Inspired by [22], our network adopts the classic U-Net structure. Fig. 4 is a schematic diagram of network architecture. Firstly, Szegedy *et al.* [8] have demonstrated that the deeper the network, the stronger the learning ability. So we stack a

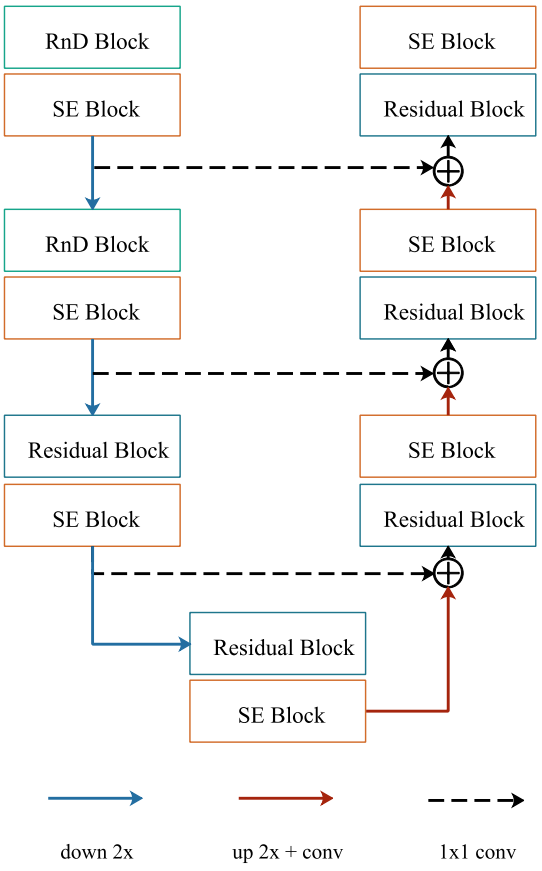


FIGURE 4. The schematic diagram of the proposed network architecture. Arrows represent different operations explained below. The details are illustrated in Fig. 1, Fig. 2, and Fig. 3.

series of residual blocks as the backbone, and the network contains more than 50 convolution layers. Secondly, skip connection between encode path and decode path transfer low-level semantic information to the decoder, which is greatly important to such an U-Net type architecture. However, there exist some large targets such as liver tumors and colon polyps that beyond the receptive field of shallow layers. Filters in shallow layers couldn't capture complete information, which weakens the first two skip connection in Fig. 4. Therefore, we apply the RnD blocks in shallow layers to obtain enough receptive field after a few convolution layers. Thirdly, we embed SE blocks after every block on the encoder-decoder path to adaptively increase sensitivity to informative feature and suppress less useful features in different tasks.

E. MODIFIED FOCAL LOSS

In segmentation tasks, cross entropy loss and dice loss are commonly used to promote the training. They are defined as

$$ce\_loss = -\sum(y) * \log(p) + (1-y) * \log(1-p) \quad (3)$$

$$dice\_loss = 1 - \frac{2 * \sum py + \varepsilon}{\sum y + \sum p + \varepsilon} \quad (4)$$

where p is the prediction score map and y is the corresponding ground truth. However, as defined in equation (3), cross entropy loss has an obvious disadvantage. When the foreground ($y = 1$) takes up a very small part, the component of $y = 0$ in the loss function will dominate, making the model severely biased towards the background. There is also a problem with dice loss that the penalty factor for positive samples is greater than the negative sample, which results in the possibility of false positive.

Focal loss [29] is first proposed for dense object detection to reshape the standard cross entropy loss such that it down-weights the loss assigned to well-classified examples. In this paper, we modify the standard focal loss to apply in segmentation tasks to overcome above two problems. It is defined as follows:

$$focal\_loss^{(i)} = \frac{-\sum_{X,Y,Z}(1-p^{(i)}_{x,y,z})^2 * \log(p^{(i)}_{x,y,z})}{max(\sqrt{\sum_{X,Y,Z} 1\left\{y_{x,y,z} \neq p_{x,y,z}\right\}}, 1)} \quad (5)$$

where $focal\_loss^{(i)}$ denotes the focal loss for an input image $X^{(i)}$ and the function $\sum_{X,Y,Z} 1\left\{y_{x,y,z} \neq p_{x,y,z}\right\}$ is to count the number of error predicted pixels in $X^{(i)}$. The modified focal loss has a very high penalty for error pixels regardless of foreground or background which breaks limitations of above two common losses.

IV. EXPERIMENTS

In this section, we demonstrate the effectiveness of our method on five different segmentation tasks. There are the segmentation of brain tumors in multimodal multisite MRI (FLAIR, T1w, T1gd, T2w), the segmentation of hippocampus in mono-modal MRI, the segmentation of liver tumors in portal venous phase CT, the segmentation of pancreas in portal venous phase CT and colon polyps in endoscopic images.

A. EXPERIMENT DESIGN

1) DATASET

Medical Segmentation Decathlon Challenge provides the above first four datasets and the last dataset for colon polyps is obtained from the GIANA2018 challenge. There are 484, 260, 118, 281, 190 image data for brain tumor, hippocampus, liver tumor, pancreas, and colon polyp tasks respectively. All data have been pixel-wise labeled to mimic the accuracy required for clinical use. In this experiment, each dataset is divided into training, validation, and test in a ratio of 7:2:1.

2) DATA PREPROCESSING AND AUGMENTATION

Data used in experiments are from five different datasets and in different modalities. Under different modalities, every pixel represents unique situational information, therefore we do preprocessing separately for each dataset to attempt to eliminate this difference. For CT data, we adjust window width and window level. For instance, CT value of liver tumor data is limited to $[-110, 190]$ and that of pancreas data are in the range $[-100, 200]$. As for hippocampus imaged in MRI, all pixels in an image are sorted by pixel value firstly, only pixels whose value is in the range $[5\%, 95\%]$ are finally retained. And brain tumor data are processed by

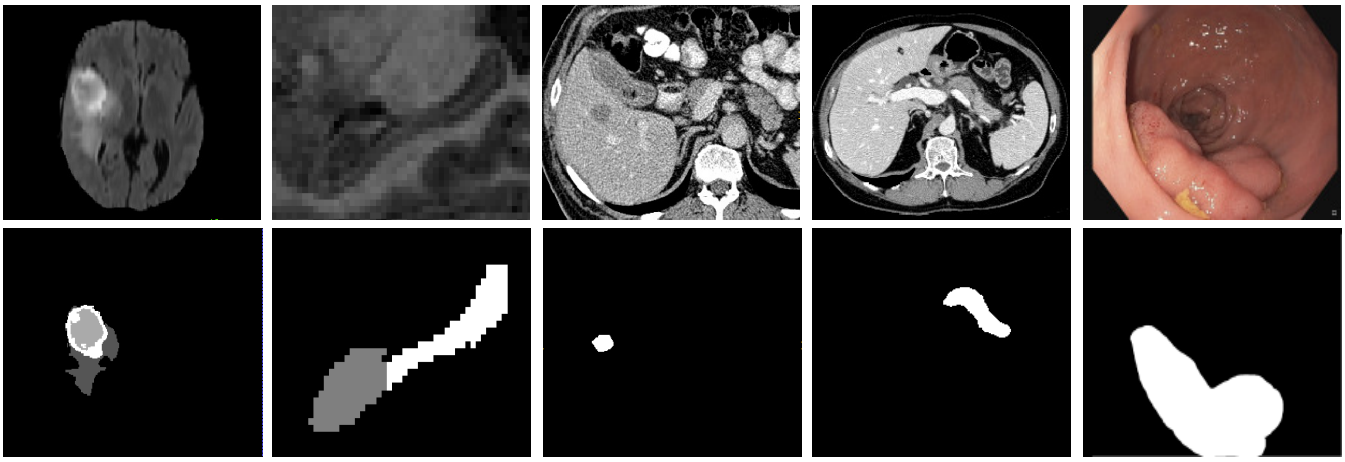


FIGURE 5. Five samples for each segmentation task. The first row shows the original images and the second row shows the ground truth. (a) Brain Tumor. (b) Hippocampus. (c) Liver Tumor. (d) Pancreas. (e) Colon Polyp.

whitening. Besides, all five datasets are finally normalized to [0, 1]. Data augmentation only includes random rotation and random flipping.

3) IMPLEMENTATION DETAILS

All tasks use the identical network structure except for the last convolution layer and they don't share the model weights each other. All five tasks are training from scratch. For the polyp task, considering the format of endoscopic images, we reduce all convolution kernels from 3D to 2D without any further changes. Due to GPU memory limitation, we make a prediction using a smaller sliding prediction window over the whole image for brain tumor, liver tumor, and pancreas segmentation task. We utilize the Adam optimizer and the learning rate is initialized to 1e-4. Training time for each task is less than 10 hours and we select the model that performs best on validation dataset as the final model.

4) EVALUATION METRICS

Three metrics are utilized to assess the performance of each model. Our main metric is dice score, which is evaluated as:

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{6}$$

where TP, FP, FN represent true positive, false positive and false negative respectively. Additionally, we use the Jaccard similarity coefficient and standard deviation of dice as two optional measurement criteria. The Jaccard is defined as equation 7 and the standard deviation is as equation 8, where N is the total number of test samples, and $\hat{D}$ is the mean value of dice. The standard deviation can reflect the degree of dispersion for a dataset, of which the value is expected to close to 0.

$$Jaccard = \frac{TP}{TP + FP + FN} \tag{7}$$

$$StdDev = \sqrt{\frac{\sum_i (D_i - \hat{D})^2}{N}} \tag{8}$$

B. BRAIN TUMOR

To evaluate the effectiveness of NDN, we first perform experiments on MRI images for brain tumor diagnosis. Glioma is the most common primary brain tumor, and glioma segmentation is a challenging task because of the complex and heterogeneously-located targets like samples in Fig.5(a). The purpose of this task is to segment glioma into three subregions, namely edema, non-enhancing tumor and enhancing tumor, according to the activity of tumor cells.

In our task, the uninformative black border is removed, and only brain area is retained for model training. We employ a cascade strategy to segment substructures of gliomas hierarchically and sequentially. A subvolume with a size of $96 \times 96 \times 48 \times 4$ (where 4 means different modalities) random samples from the MRI and is used as an input to segment the whole tumor. Subsequent training takes the entire tumor volume as input and differentiates non-enhancing tumor and enhancing tumor respectively. Three classes are segmented using the same architecture with different learned weights. Besides, the following two methods are considered as comparative trials to illustrate the segmentation effect. Wang *et al.* [39] propose a triple cascaded framework with anisotropic convolution to deal with 3D brain images. WNet, TNet, and ENet are designed to segment substructures of the brain tumor, and each of these networks settles a binary segmentation problem. One-Pass Multi-task Network (OM-Net) [40] exploits the correlation between classes in training and simplifies the cascade inference processed by one-pass computation. A comparison between NDN and these two current published algorithms is listed in TABLE 1 and the corresponding example images are shown in Fig. 6. The dice scores obtained by NDN surpass the comparative methods in all three classes.

C. HIPPOCAMPUS

The hippocampus, which locates between the thalamus and the medial temporal lobe of the brain, is primarily responsible

TABLE 1. Comparison of results on brain tumor segmentation.

Method	Edema	Non-enhancing	Enhancing
Ours	**0.7122**	**0.6018**	**0.7213**
Triple cascaded	0.6919	0.5504	0.6793
OM-Net	0.6894	0.5376	0.6861

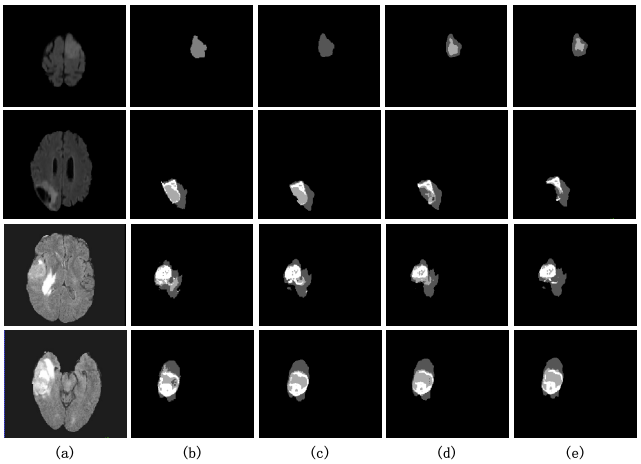


FIGURE 6. Results for each of the experiments in TABLE 1. (a) Original images. (b) Ground truth. (c) Results for NDN. (d) Results for triple cascaded. (e) Results for OM-Net.

for long-term memory storage, conversion, and orientation. The study of the hippocampus has implications for the treatment of Alzheimer's Disease (AD) because it bears the brunt of the damage. As shown in Fig. 5(b), the challenge of hippocampus segmentation is how to segment two neighboring small structures with high precision.

An entire MRI image for hippocampus serves as input to the network without special processing in this work. The anterior and posterior of the hippocampus are segmented simultaneously in a single NDN model. There exists two baseline approaches for hippocampus segmentation. Cao *et al.* [41] employ a 3D-Unet to joint hippocampus segmentation and clinical score regression. Due to the limitation of the label, we cut the regression branch in the experiment. In [42], the authors utilized 2D U-Seg-Net (a modified U-Net) to predict planar slices along different views and the predicted slices are combined to form a 3D result in an Ensemble-Net. As shown in TABLE 2, either anterior or posterior of the hippocampus, our method achieves the best result compared to the other two lastest methods. Fig. 7 displays the predictions for all methods that [41] do not perform well in the posterior and [42] shows poor performance in anterior. In contrast, our method perfectly segments these two neighboring small structures with high precision.

D. LIVER TUMOR

We next conduct experiments on CT scans for the task of liver tumor segmentation. The liver has an abundant supply of blood flow, which is closely related to the important blood vessels in the human body. The liver is one of the areas where

TABLE 2. Comparison of results on hippocampus segmentation.

	Dice		Jaccard	
Method	Anterior	Posterior	Anterior	Posterior
Ours	**0.8793**	**0.8869**	**0.7871**	**0.7951**
3D U-Net	0.8748	0.8492	0.7841	0.7922
U-Seg-Ensemble-Net	0.8575	0.8829	0.7529	0.7409

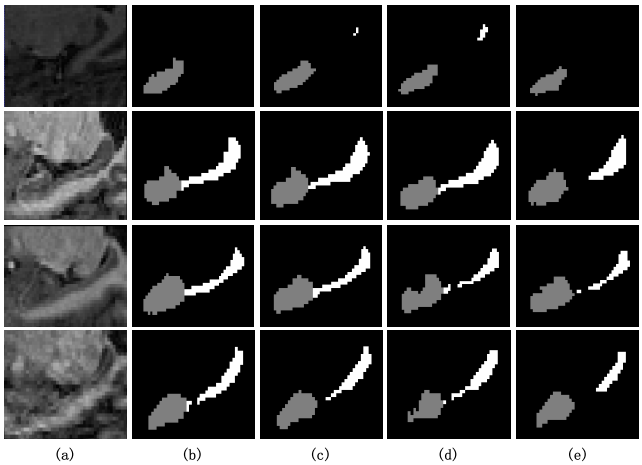


FIGURE 7. Results for each of the experiments in TABLE 2. (a) Original images. (b) Ground truth. (c) Results for NDN. (d) Results for 3D U-Net. (e) Results for U-Seg-Ensemble-Net.

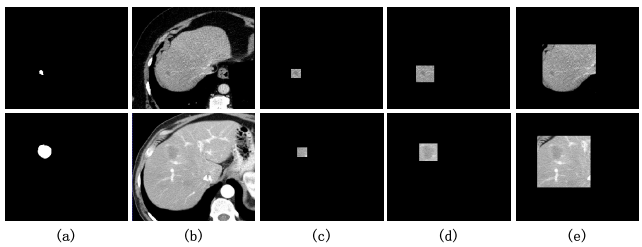


FIGURE 8. Observing the liver tumor under different size fields of view. (a) Ground truth. (b) Original images. (c) Small window of view. (d) Middle window of view. (e) Large window of view.

tumors often occur, among which metastatic tumors are more common in the malignant tumors. Moreover, malignant liver tumors are insidious and grow rapidly, so early detection of the tumor is quite urgent in clinical practice. Liver tumor segmentation usually faces with all kinds of problems such as unbalanced label with a large liver and small tumor targets (Fig. 5(c)).

To mitigate the adverse effects of a complex background, we utilize annotations of liver from dataset to train an NDN liver segmentation network in advance. We first predict a segmentation map of the whole liver and then inference the liver tumor based on the location of the liver. This working workflow is similar to [43], which uses a cascaded fully convolutional neural networks (CFCNs) to first segment liver as ROI input for a second FCN. So does the particular DCNN model presented in [44] by Han *et al.*, which combines the long-distance skip connection of U-Net and the short-distance skip

TABLE 3. Comparison of results on liver tumor segmentation.

Method	Dice	Jaccard	StdDev
Ours	**0.6795**	**0.5433**	**0.2418**
CFCNs	0.4342	0.3054	0.3019
DCNN	0.4955	0.3728	0.2779

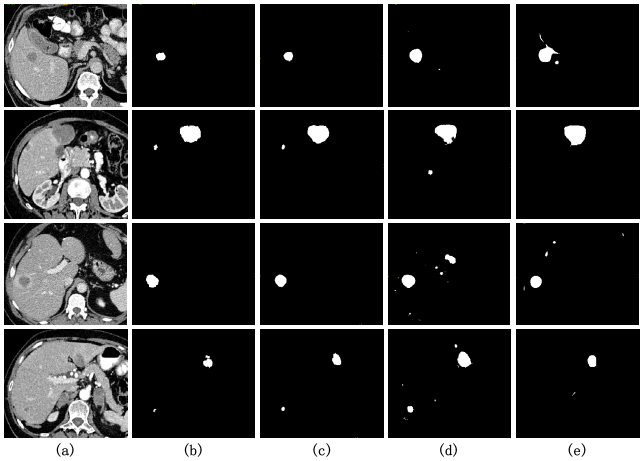


FIGURE 9. Results for each of the experiments in TABLE 3. (a) Original images. (b) Ground truth. (c) Results for NDN. (d) Results for CFCNs. (e) Results for DCNN.

connection of ResNet. However, instead of using 3D images as input, Han *et al.* use 2.5D to produce the segmentation map corresponding to the center slices. However, due to the low contrast of tumor in the liver, it is usually necessary to combine a large amount of surrounding information to discover the liver tumor region. As shown in Fig. 8, the more surrounding information, the easier it is to segment liver cancer. Therefore, our method accumulates a large receptive field by RnD blocks and Residual blocks and takes advantage of this large region information to highlight the tumor region inside the liver.

The result shown in TABLE 3 demonstrates that our method achieves better performance in dice and Jaccard by a large margin compared to the above two methods. Due to the massive difference in the characterization of liver tumors between various patients, the algorithm's performance generally fluctuates greatly between different data. In order to verify the robustness of the method, we add the standard deviation of the dice score to evaluation metrics and our method also achieves the smallest standard deviation. Fig. 9 shows that our approach understands liver tumors more precisely and provides a cleaner segmentation result without too many false positives.

E. PANCREAS

The Pancreas is a small but extraordinary organ. Its physiological function and pathological changes are closely bound up with life. Up to now, pancreas segmentation has faced opportunities and challenges due to the following: 1) there exist great anatomical difference in shape and size between patients, 2) the pancreas is a deformable organ and 3) its boundary is not clear. The pancreas is involved with a complex abdominal environment, which leads to further difficulties for its segmentation.

TABLE 4. Comparison of results on pancreas segmentation.

Method	Dice	Jaccard	StdDev
Ours	**0.8476**	**0.7392**	**0.0549**
ResDSN C2F	0.8040	0.6812	0.0926
3D U-Net	0.8315	0.7173	0.0700

TABLE 5. Comparison of results on polyp segmentation.

Method	Dice	Jaccard	StdDev
Ours	**0.8934**	**0.8178**	**0.0917**
EFCN-8	0.8633	0.7752	0.1190
Deeplabv3 based	0.8755	0.7942	0.1147

The network is hard to receive a complete pancreas CT image as input due to GPU memory. We use a cascade inference strategy in this task. Served as a preprocessing step, the network uses a $128 \times 128 \times 48$ sliding window to scan on the whole CT image to generate a rough segmentation map. Based on the maximum connected domain in the rough map, we crop a subvolume that contains the whole pancreas target and send it to the network for a refined prediction. In addition, the ResDSN C2F network developed in [45] introduces the twofold coarse-to-fine strategy, namely ResDSN Coarse and ResDSN Fine, to segment the pancreas leveraging the rich spatial information. What's more, Roth *et al.* describe a custom-built 3D U-Net in [46] and utilize a random forest (RF) to remove the background.

In our consideration, the pancreas is almost in a fixed position, so the surrounding environment is roughly similar. It is important to utilize surrounding anatomical structural information to locate the pancreas which is still needed as large a receptive field as possible. Therefore, NDN's strategy to accumulate a large area. Then focal loss is applied to outline clear boundaries. We reproduce the previous two algorithms and make a comparison with our method. The result is shown in TABLE 4 and Fig. 10.The proposed NDN is obviously superior in all evaluation metrics.

F. COLON POLYP

Endoscopic images for colon polyp diagnosis are chosen for one of the experiments because of the variations in the size and shape of the polyp. A colon polyp is a type of benign tumor, but some of them are prone to malignant ones. Detecting colon polyp lesions in their early stage has become a serious medical issue.

In this experiment, the NDN is further investigated in 2D medical image segmentation. The network structure is consistent except for the change of replacing 3D convolution kernels with 2D convolution kernels to fit 2D data. The whole 2D images are fed into the network without any further processing but normalization. For comparison, strong results

TABLE 6. Comparison of dice score on each task using different losses.

Method	Brain tumor			Hippocampus		Liver tumor	Pancreas	Polyps
	Edema	Non-enhancing	Enhancing	Anterior	Posterior			
Focal loss	**0.7122**	**0.6018**	**0.7213**	0.8793	**0.8869**	**0.6795**	**0.8476**	**0.8934**
Dice loss	0.7043	0.5889	0.7206	0.8604	0.8644	0.6623	0.8303	0.8648
CE loss	0.7001	0.5203	0.7069	**0.8820**	0.8751	0.5984	0.8349	0.8460

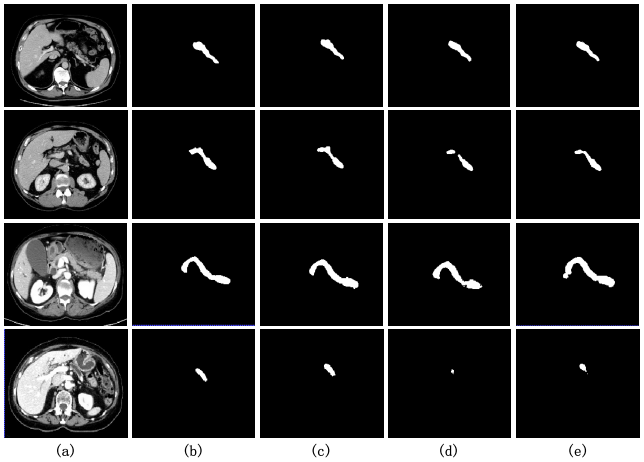


FIGURE 10. Results for each of the experiments in TABLE 4. (a) Original images. (b) Ground truth. (c) Results for NDN. (d) Results for ResDSN C2F. (e) Results for 3D U-Net.

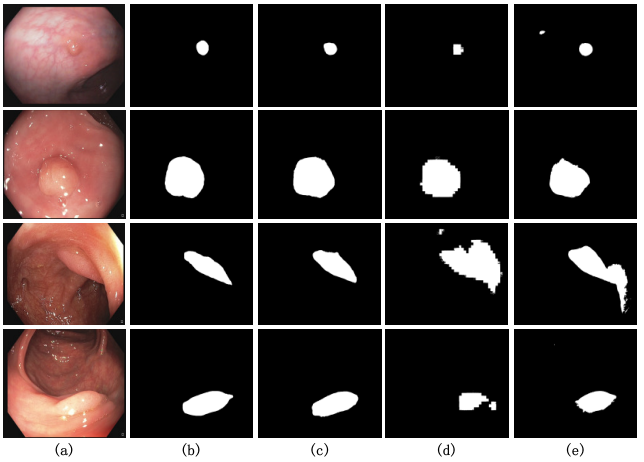


FIGURE 11. Results for each of the experiments in TABLE 5. (a) Original images. (b) Ground truth. (c) Results for EFCN-8. (d) Results for ResDSN C2F. (e) Results for Deeplabv3 based model.

are included that we are currently aware of the published literature. Wickstrøm *et al.* [25] combine FCN-8 and SegNet with recent development in the deep learning like Batch Normalization and Dropout respectively. The enhanced FCN-8 and SegNet are corresponding named to EFCN-8 and ESegNet. According to the result provided by Zhu *et al.*, the EFCN-8 perform better than ESegNet, so EFCN-8 is reproduced as a comparative approach. Nguyen and Lee [47] develop a deep encoder-decoder network which uses Deeplabv3 as an encoder to segment polyps. They train the model three times with different resolutions and ensemble the results as the final segmentation map.

We evaluate the results of NDN, EFCN-8, Deeplabv3 based model and provide a comparison between these models in TABLE 5. It is observed that NDN achieves a dice of 0.8934, a Jaccard of 0.8178 and a standard deviation of 0.0917, which presents the most robust reported results. As shown in Fig. 11, folds in the colon are easily mistaken for polyps because of their similar appearance. Our model can well address this issue.

G. THE COMPARISON OF DIFFERENT LOSS

To highlight the importance of our modified focal loss, we adopt the variable-controlling method that only changes the supervision loss. In the above experiments, we fine-tune the NDN for each task with cross-entropy(CE) loss and dice loss that are generally used in medical image segmentation. As shown in Table 6, it is obvious that the focal loss performs better against dice loss and CE loss in all. The focal loss is used to help network focus on error regions including missing area and false positive area. According to the row 3, 4 of Fig. 12, liver tumors, and pancreas are small and fuzzy, which results in frequent omissions. However, our method supervised by focal loss is able to detect these hard regions

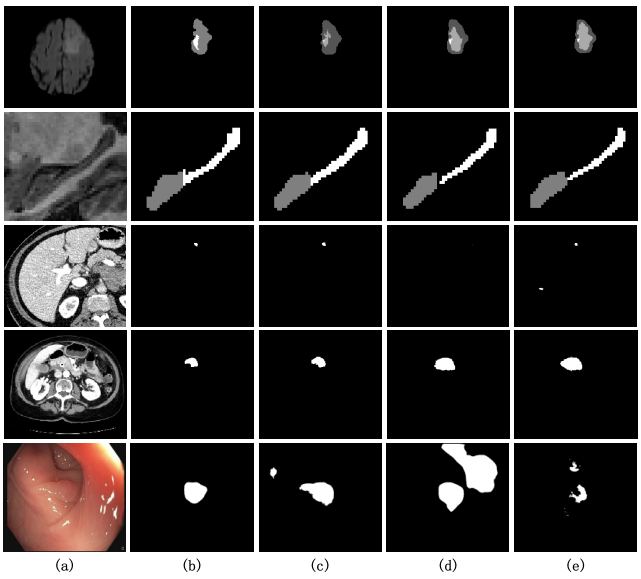


FIGURE 12. Results for each of the experiments in TABLE 6. Rows represent different tasks. (a) Original images. (b) Ground truth. (c) Results for focal loss. (d) Results for dice loss. (e) Results for cross-entropy loss.

compared to the same framework supervised by the dice loss and CE loss. Besides, the focal loss also plays an important role in suppressing false positive. In colon polyp segmentation task, affected by light and bubbles, the result usually contains some false detection areas. Focal loss shows its role and suppresses the false positive as shown in row 5 of Fig. 12.

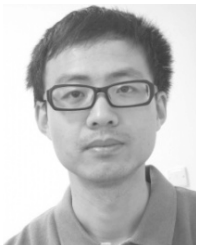
V. CONCLUSION

In this paper, we propose a new network named Nested Dilation Networks (NDN) to cope with multi-task medical image segmentation imaging in different modalities. The innovation of the article is mainly reflected in two aspects. Firstly, the Residual Blocks Nested with Dilations (RnD Blocks) is designed to enlarge the receptive field, which can handle targets of different sizes simultaneously. Secondly, the focal loss is carefully modified to adopt the segmentation tasks. The modified focal loss can overcome the challenges of small targets and unbalanced labels, which perform better than previous dice loss. Besides, five experiments are conducted to prove the generalization of NDN, and the results demonstrate that our algorithm outperforms other currently published methods which focus only one specific task. The proposed approach is helpful for more medical segmentation tasks.

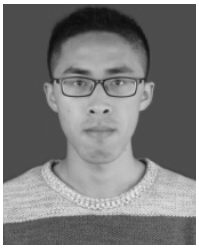
REFERENCES

[1] X.-W. Chen and X. Lin "Big data deep learning: Challenges and perspectives," *IEEE Access*, vol. 2, pp. 514–525, 2014.
[2] Y.-D. Zhang, K. Muhammad, and C. Tang, "Twelve-layer deep convolutional neural network with stochastic pooling for tea category classification on GPU platform," *Multimedia Tools Appl.*, vol. 77, no. 17, pp. 22821–22839, 2018.
[3] Z. Xie. (2018). "Towards single-phase single-stage detection of pulmonary nodules in chest CT imaging." [Online]. Available: https://arxiv.org/abs/1807.05972
[4] S.-H. Wang, J. Sun, P. Phillips, G. Zhao, and Y.-D. Zhang, "Polarimetric synthetic aperture radar image segmentation by convolutional neural network using graphical processing units," *J. Real-Time Image Process.*, vol. 15, no. 3, pp. 631–642, Oct. 2018.
[5] 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.
[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.
[8] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1–9.
[9] S. Ioffe and C. Szegedy. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167
[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.
[11] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. (2016). "Inception-v4, inception-ResNet and the impact of residual connections on learning." [Online]. Available: https://arxiv.org/abs/1602.07261
[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[13] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 636–644.
[14] H. Greenspan, B. V. Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1153–1159, Mar. 2016.
[15] G. Wang. (2016). "A perspective on deep imaging." [Online]. Available: https://arxiv.org/abs/1609.04375
[16] S.-H. Wang, Y.-D. Lv, Y. Sui, S. Liu, S.-J. Wang, and Y.-D. Zhang, "Alcoholism detection by data augmentation and convolutional neural network with stochastic pooling," *J. Med. Syst.*, vol. 42, no. 1, p. 2, 2018.
[17] Y.-D. Zhang, C. Pan, X. Chen, and F. Wang, "Abnormal breast identification by nine-layer convolutional neural network with parametric rectified linear unit and rank-based stochastic pooling," *J. Comput. Sci.*, vol. 27, pp. 57–68, Jul. 2018.
[18] Y.-D. Zhang, C. Pan, J. Sun, and C. Tang, "Multiple sclerosis identification by convolutional neural network with dropout and parametric ReLU," *J. Comput. Sci.*, vol. 28, pp. 1–10, Sep. 2018.
[19] S.-H. Wang *et al.*, "Multiple sclerosis identification by 14-layer convolutional neural network with batch normalization, dropout, and stochastic pooling," *Frontiers Neurosci.*, vol. 12, p. 818, Nov. 2018.
[20] S.-H. Wang, K. Muhammad, J. Hong, A. K. Sangaiah, and Y.-D. Zhang, "Alcoholism identification via convolutional neural network based on parametric ReLU, dropout, and batch normalization," in *Neural Computing and Applications*. London, U.K.: Springer-Verlag, 2019, pp. 1–16.
[21] C. Chu *et al.*, "Multi-organ segmentation based on spatially-divided probabilistic atlas from 3D abdominal CT images," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 16. Nagoya, Japan: Springer, no. 2, 2013, pp. 165–172.
[22] 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.*, 2015, pp. 234–241.
[23] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes", *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.
[24] K. Kamnitsas *et al.*, "DeepMedic for brain tumor segmentation," in *Proc. Int. Workshop Brainlesion, Glioma, Multiple Sclerosis, Stroke Traumatic Brain Injuries*, 2016, pp. 138–149.
[25] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen, "Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation," in *Proc. IEEE 28th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2018, pp. 1–6.
[26] P. Moeskops *et al.*, "Deep learning for multi-task medical image segmentation in multiple modalities," in *Medical Image Computing and Computer-Assisted Intervention*. Athens, Greece: Springer, 2016, pp. 478–486.
[27] A. El-Baz, A. Elnakib, M. A. El-Ghar, G. L. Gimel'farb, R. Falk, and A. Farag, "Automatic detection of 2D and 3D lung nodules in chest spiral CT scans," *Int. J. Biomed. Imag.*, vol. 2013, p. 517632, Dec. 2013.
[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2999–3007.
[30] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *IEEE Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6230–6239.
[31] V. Badrinarayanan, A. Handa, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," in *Proc. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–10.
[32] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1520–1528.
[33] 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.
[34] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in *Wavelets*. Berlin, Germany: Springer, 1990, pp. 286–297.
[35] F. Yu and V. Koltun. (2015). "Multi-scale context aggregation by dilated convolutions." [Online]. Available: https://arxiv.org/abs/1511.07122
[36] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. (2017). "Rethinking atrous convolution for semantic image segmentation." [Online]. Available: https://arxiv.org/abs/1706.05587
[37] P. Wang *et al.*, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2018, pp. 1451–1460.

[38] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. (2017). "Squeeze-and-excitation networks." [Online]. Available: https://arxiv.org/abs/1709.01507
[39] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, "Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks," in *Proc. Int. MICCAI Brainlesion Workshop*. Cham, Switzerland: Springer, 2017, pp. 178–190.
[40] C. Zhou, C. Ding, Z. Lu, X. Wang, and D. Tao, "One-pass multi-task convolutional neural networks for efficient brain tumor segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 637–645.
[41] L. Cao *et al.*, "Multi-task neural networks for joint hippocampus segmentation and clinical score regression," *Multimedia Tools Appl.*, vol. 77, no. 22, pp. 29669–29686, 2018.
[42] 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.*, Apr. 2017, pp. 192–196.
[43] P. F. Christ *et al.* (2017). "Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1702.05970
[44] X. Han. (2017). "Automatic liver lesion segmentation using a deep convolutional neural network method." [Online]. Available: https://arxiv.org/abs/1704.07239
[45] Z. Zhu, Y. Xia, W. Shen, E. K. Fishman, and A. L. Yuille. (2017). "A 3D coarse-to-fine framework for volumetric medical image segmentation." [Online]. Available: https://arxiv.org/abs/1712.00201
[46] H. Roth *et al.*, "Towards dense volumetric pancreas segmentation in CT using 3D fully convolutional networks," *Proc. SPIE, Med. Imag., Image Process.*, vol. 10574, p. 105740B, Mar. 2018.
[47] Q. Nguyen and S.-W. Lee, "Colorectal segmentation using multiple encoder-decoder network in colonoscopy images," in *Proc. IEEE 1st Int. Conf. Artif. Intell. Knowl. Eng. (AIKE)*, Sep. 2018, pp. 208–211.



LIANSHENG WANG received the Ph.D. degree in computer science from The Chinese University of Hong Kong, in 2012. He is currently an Associate Professor with the Department of Computer Science, Xiamen University, Xiamen, China. His research interest includes medical image processing and analysis.



RONGZHEN CHEN received the B.S. degree from Xiamen University, Xiamen, China, in 2017, where he is currently pursuing the master's degree with the Department of Computer Science. His research interests include medical image processing and machine learning.

SHUXIN WANG received the B.S. degree from Shandong Normal University, in 2017. She is currently pursuing the master's degree with the Department of Computer Science, Xiamen University, Xiamen, China. Her research interests include medical image processing and machine learning.

NIANYIN ZENG received the B.Eng. degree in electrical engineering and automation and the Ph.D. degree in electrical engineering from Fuzhou University, in 2008 and 2013, respectively. From 2012 to 2013, he was an RA with the Department of Electrical and Electronic Engineering, The University of Hong Kong. From 2017 to 2018, he was an ISEF Fellow founded by the Korea Foundation for Advance Studies and also a Visiting Professor with the Korea Advanced Institute of Science and Technology. He is currently an Associate Professor with the Department of Instrumental and Electrical Engineering, Xiamen University. His current research interests include intelligent data analysis, computational intelligent, and time-series modeling and applications.

XIAOYANG HUANG received the Ph.D. degree in computer science from Xiamen University, Xiamen, China, where he is currently an Assistant Professor with the Department of Computer Science. His research interests include medical image processing and analysis, and computerized cardiology.

CHANGHUA LIU is currently an Associate Professor and the Head of the Department of Medical Imaging, Chenggong Hospital Affiliated to Xiamen University. His research interests include medical imaging and medical image processing.