2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

An Efficient and Accurate 3D Multiple-Contextual Semantic Segmentation Network for Medical Volumetric Images

He Li[1], Yutaro Iwamoto[1], Xianhua Han[2], Akira Furukawa[3], Shuzo Kanasaki[4], Yen-Wei Chen[1]

Abstract— **Convolutional neural networks have become popular in medical image segmentation, and one of their most notable achievements is their ability to learn discriminative features using large labeled datasets. Two-dimensional (2D) networks are accustomed to extracting multiscale features with deep convolutional neural network extractors, i.e., ResNet-101. However, 2D networks are inefficient in extracting spatial features from volumetric images. Although most of the 2D segmentation networks can be extended to three-dimensional (3D) networks, extended 3D methods are resource and time intensive. In this paper, we propose an efficient and accurate network for fully automatic 3D segmentation. We designed a 3D multiple-contextual extractor (MCE) to simulate multiscale feature extraction and feature fusion to capture rich global contextual dependencies from different feature levels. We also designed a light 3D ResU-Net for efficient volumetric image segmentation. The proposed multiple-contextual extractor and light 3D ResU-Net constituted a complete segmentation network. By feeding the multiple-contextual features to the light 3D ResU-Net, we realized 3D medical image segmentation with high efficiency and accuracy. To validate the 3D segmentation performance of our proposed method, we evaluated the proposed network in the context of semantic segmentation on a private spleen dataset and public liver dataset. The spleen dataset contains 50 patients' CT scans, and the liver dataset contains 131 patients' CT scans.**

I. INTRODUCTION

Segmentation of abdominal organs from computed tomography (CT) images is a crucial and time-consuming task. Recently, deep learning-based methods have achieved impressive success in dealing with segmentation problems of medical images. This success is largely attributed to fully convolutional neural networks (FCNs) [1], [2], [3]. The literature can be classified into two categories. One is based on 2D FCNs [1], [2], and the other category is 3D FCNs [3]. In the field of 2D FCNs, recent works rely on multi-scale contest fusion improve the discriminative ability of feature representations [4], [5], [6], [7]. They all utilize deep convolutional neural networks that aggregate contextual multiscale information, [5] also applies feature pyramid for feature maps fusion. [7] employs ResNet-101 [8] as their multiscale feature extractor, different depth of network layers represent different level image representations. These strategies may help to capture objects at different scales, but 2D convolutions still lack the spatial connection between volumetric images and the spatial information deficiency negatively affects these 2D FCN-based method performance.

Alternatively, in order to solve the lack of spatial information of 2D FCN-based methods, 3D FCN-based methods have been widely studied in recent years. In the field of 3D FCNs, 2D convolutions are replaced by 3D kernels with volumetric data input [9], [10], [11]. Three-dimensional convolutional kernel reflects competitiveness at extracting features from voxel for 3D segmentation. However, 3D FCNs generate numerous training parameters, and standard GPU devices cannot handle large amounts of 3D data for processing with sophisticated backbone structures. Take 3D U-Net [10] for example, its symmetric encoder-decoder structure costs excessive computing resources, and results in low applicability for handling the problems of distribution imbalance between foreground and background, prediction detail losing, and overfitting.

In order to address all these issues, several previous works inspired us to design an efficient and accurate semantic segmentation network for medical volumetric images. First, we employed our multiple-contextual extractor (MCE) as the short-range feature extractor for global feature fusion. Additionally, we set the 3D residual block as a backbone. The block was responsible for extracting local features and increasing network depth. Finally, we optimized a light 3D ResU-Net for generating highly accurate biomedical image segmentation results.

On the basis of the above-mentioned works and insights, in this paper, we propose a unified 3D multiple-contextual semantic segmentation network, which is designed to achieve 3D segmentation from medical volume images in an end-to-end manner. Compared with other segmentation models, our proposed model achieves more advanced accurate results with efficient computation time. Sufficient ablation studies were conducted on two datasets which demonstrated the superiority of our proposed method. Each component from our method validates the following main contributions: (I) we propose a 3D unified segmentation framework for medical volumetric images. The asymmetric framework structure of the proposed method is more efficient and accurate than other symmetric structure segmentation methods; (II) we propose a 3D multiple-contextual extraction (MCE) module to improve global and local feature identification; (III) we employ 3D residual block as the backbone to build a light 3D ResU-Net. The light 3D ResU-Net is able to achieve segmentation efficiently.

[1]Graduate Scool of Information Science and Engineering, Ritsumeikan University, Shiga, Japan
[2]Faculty of Science, Yamaguchi University, Yamaguchi, Japan
[3]Tokyo Metropolitan University, Tokyo, Japan
[4]Koseikai Takeda Hospital, Kyoto, Japan

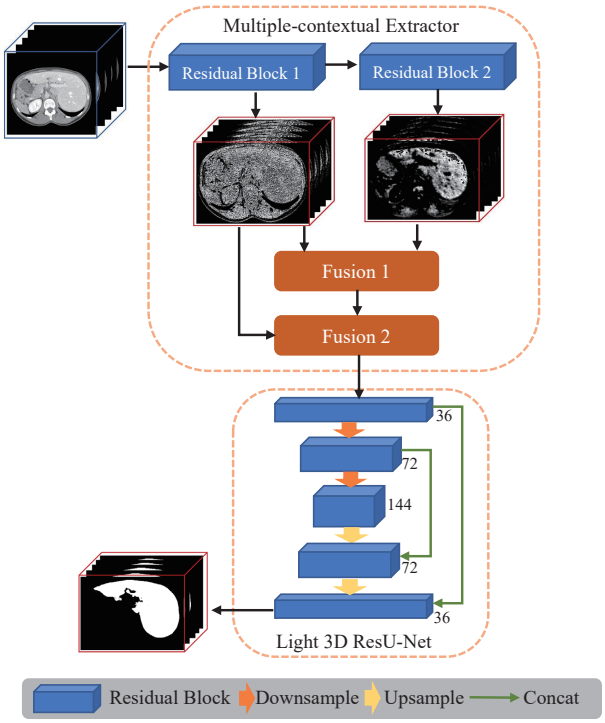

Fig. 1. The illustration of our proposed network. The network consists of a multiple-contextual extractor and a light 3D ResU-Net for efficient and accurate segmentation.

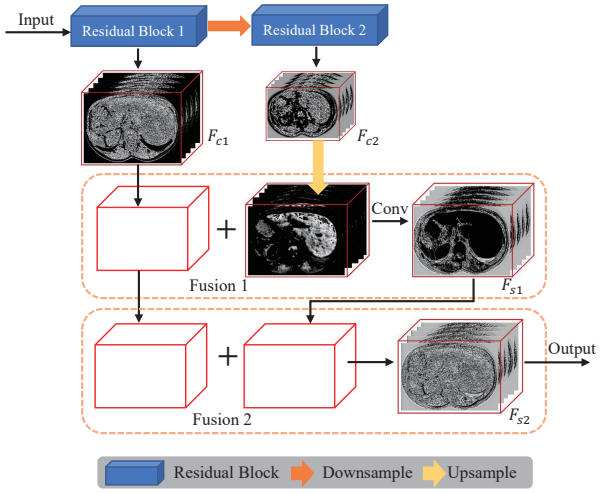

Fig. 2. The structure of our multiple-contextual extractor. The multiple-contextual features are extracted and fused by twice fusion and twice residual block conduction.

II. METHODS

A. Overview

The overview of the proposed framework is depicted in Fig. 1. Traditional 2D FCN-based methods for segmentation generally use long-range feature extractor to generate multiscale feature representations. However, long-range feature extractor may overuse high-level features and generate enormous parameters in 3D FCNs which results in reducing segmentation performance and generating overfitting. To avoid these problems, we propose a multiple-contextual extractor (MCE) as a substitute for long-range feature extractor in 3D FCNs. We first put preprocessed images into MCE for gathering and fusing global contextual features. Meanwhile, the 3D residual block is set as the backbone of the whole network for capturing local features. With the help of our MCE and residual block, global and local features are merged together. Then, the fused features are fed into our light 3D ResU-Net. Our MCE and light 3D ResU-Net are trained to minimize the difference between predictions and ground truths with an end-to-end manner.

B. Multiple-Contextual Extractor

To set up an efficient and accurate 3D segmentation network, we focus on designing a competent network structure for acquiring image features. Inspired by the multiscale feature extraction strategy [7] in 2D FCNs, which helps in collecting global and local features, we propose MCE for our 3D segmentation network. In this setting (Fig. 2), we first obtain different contextual features F_{c1} and F_{c2}, where F_{c1} and F_{c2} indicate low and high-level contextual features, respectively. Because F_{c2} obtained from residual blocks twice and downsampled once, it is a higher feature representation. We employ trilinear interpolation to ensure different levels of information at the same resolution. Then, F_{s1} whcih is obtained from different level contextual features, is added, forming an initial multiple-contextual feature map, $F_{s1} = conv([F_{c1}, F_{c2}])$. To preserve input data informationn, F_{s2} is also used for a more detailed feature map, $F_{s2} = conv([F_{c1}, F_{s1}])$. Thus, F_{s2} encodes low-level detailed contextual features from shallow layers and high-level semantics learned from deeper layers. Then, this new multiple-contextual feature map is combined with different scales information and fed into the light 3D ResU-Net, $Seg = RB_s(conv([F_{c1}, F_{s1}]))$, where RB represents each residual block. Specifically, we set the number of channels invariable in our MCE (e.g. 36.), and this setting is not only for fitting add steps but also for encoding a more efficient feature representation, which means only limited and necessary features can go through our block.

C. Light 3D ResU-Net

As introduced earlier, the efficiency and accuracy of the network are the primary goal which need to be achieved. For utilizing limited computing resources and reducing overfitting, we design a light 3D ResU-Net. Compared with the 3D deep FCNs such as 3D U-Net [10]. To avoid of overfitting, we reduce the steps of high-level feature reusing. Meanwhile, we utilize residual block as the backbone not only for extracting local features but also for increasing the depth of the netwrok. Our light structure gneerates more competitive results, and it is also able to learn hierarchical representations from multiple-contextual informations. Using proposed MCE and 3D residual block, global and local features are extracted and fused. To achieve segmentation, Dice coefficient [12] loss function is employed. The Dice loss function is defined as follows:

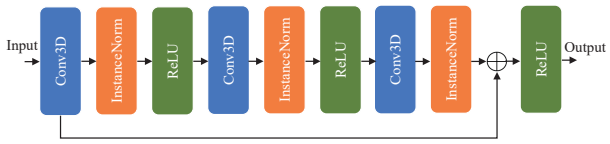

Fig. 3. The illustration of 3D residual block. As the backbone of our proposed network , 3D residual block is responsible for generating local features and offsetting the depth of our network.

$$L_d(P,G) = 1 - 2 \times \frac{\sum_{i=1}^{N} p_i g_i + \sigma}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i + \sigma} \quad (1)$$

where sums are calculated over the N pixels, the predicted volume $p_i \in P$, the ground truth volume $g_i \in G$ and σ is a smoothness term to avoid division by 0. In the optimization stage, the Dice loss is minimized by gradient descend using the following derivate equation:

$$\frac{\partial L_d(P,G)}{\partial p_k} = -2 \times \frac{\sum_{i=1}^{N} p_i g_i - g_k \sum_{i=1}^{N} (p_i + g_i)}{[\sum_{i=1}^{N} (p_i + g_i)]^2} \quad (2)$$

D. 3D Residual Block

The structure of 3D residual block is shown in Fig. 3. We set 3D residual block as the backbone of the network, small 3^3 kernels are used which are faster to convolve with and contain less parameters. When compared to larger kernel size for 3D computation, the small kernels provide high efficiency and accuracy. The 3D residual block can enhance depth of the network, and it also can prevent image detail losing by over using downsample steps to generating high level features. Besides, we adopt instance normalization technique to all hidden layers, which accelerates the convergence of the network and preserve each feature maps instance separately.

III. EXPERIMENTS AND RESULTS

A. Datasets and Implementation

Our proposed network trained and tested on two different datasets: one was a private spleen dataset, and the other was a public liver dataset. The spleen dataset was collected from Shiga University of Medical Science Hospital and had passed ethical approval. It contained 50 CT volume data with a resolution of 1.0 *mm* and slice spacing from 1.0 to 2.0 *mm*, we applied 41 scans for training and 9 scans for testing. The liver segmentation dataset was collected from MICCAI 2017 LiTS Challenge, which contained 131 contrast-enhanced 3D abdominal CT scans. The liver dataset is acquired by different doctors with a big variety resolution of 0.55 *mm* to 1.0 *mm* and slice spacing from 0.45 *mm* to 6.0 *mm*, we applied 103 scans for training and 28 scans for testing. Target spleen and liver areas were labeled by experienced doctors. For image preprocessing, we truncated the image intensity values of all scans to a range of [-200, 250] HU to remove the irrelevant details. For spleen dataset, we resampled images into 2.0×2.0×2.0 mm^3. For liver dataset, we resampled images into 2.0×2.0×5.0 mm^3. The parameters of the network were initialized with random values. Our proposed method was evaluated using three quantitative metrics, including the Dice similarity coefficient (Dice), volumetric overlap error (VOE) and the Jaccard index. We also calculated average prediction times for full 3D data of each model.

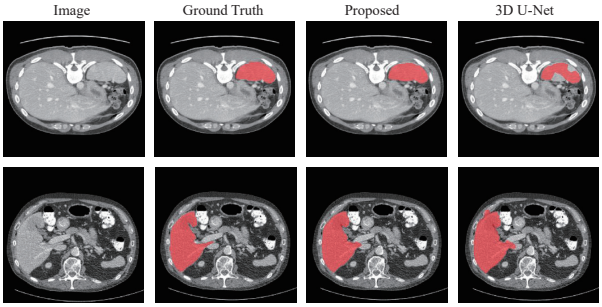

Fig. 4. The prediction compararision between our proposed network and 3D U-Net

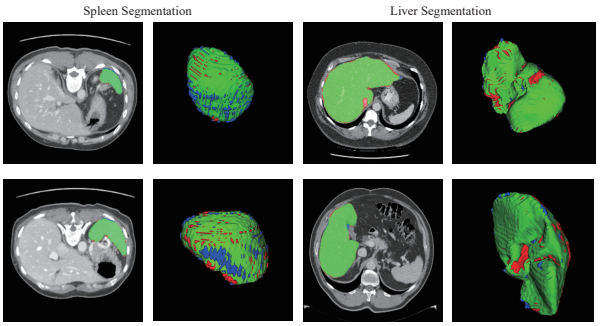

Fig. 5. The 3D prediction examples of our methods on two organs segmentation. Green indicates true positives; red indicates false positives; blue indicates false negatives.

B. Quantitative Comparison

In this section, we compared the proposed network with 3D U-Net [10]. Because 3D U-Net is a standard deep multiscale feature extraction and fusion network, and its encoder-decoder structure is similar to our light 3D ResU-Net. We set the baseline as the light 3D ResU-Net without our multiple-contextual extractor (MCE). We normalized all preprocessed data image intensity into the range of [0, 1] for training and testing.

Table 1 shows that with 3D residual block, our light 3D ResU-Net is better than 3D U-Net. Our light 3D ResU-Net with MCE also achieves a surpassing performance over 3D U-Net. After 5-fold cross-validation, we set the average scores as the final results. Fig. 4 exhibits the segmentation results of proposed network and 3D U-Net on two different datasets. The volume predictions are shown in Fig. 5. It can be observed that our proposed method accurately localizes and predicts organs.

C. Comparison with State-of-the-Art Methods

To demonstrate the predominance of our proposed method, we evaluated the performance of our network by comparing

TABLE I

QUANTITATIVE COMPARISON BETWEEN 3D U-NET AND PROPOSED METHOD. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Methods	Spleen Segmentation				Liver Segmentation			
	Dice	VOE (%)	Jaccard	Time (s)	Dice	VOE (%)	Jaccard	Time (s)
3D U-Net [10]	0.836	15.79	0.743	5.231	0.912	13.61	0.814	5.341
Light 3D ResU-Net	0.877	12.92	0.788	**0.478**	0.927	8.83	0.874	**0.489**
Light 3D ResU-Net + MCE	**0.911**	**9.69**	**0.839**	0.518	**0.947**	**4.53**	**0.915**	0.526

TABLE II

COMPARISON OF PROPOSED METHOD WITH THE STATE OF THE ART METHODS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Methods	Spleen Segmentation				Liver Segmentation			
	Dice	VOE (%)	Jaccard	Time (s)	Dice	VOE (%)	Jaccard	Time (s)
V-Net [9]	0.887	12.34	0.809	5.451	0.936	6.77	0.897	5.495
VoxResNet [13]	0.872	12.76	0.755	2.791	0.921	7.93	0.863	2.483
3D DSD-FCN [14]	0.843	11.67	0.762	4.793	0.929	8.77	0.854	4.677
MultiResUNe [15]	0.883	11.89	0.792	4.895	0.932	6.89	0.892	5.034
Proposed	**0.911**	**9.69**	**0.839**	**0.518**	**0.947**	**4.53**	**0.915**	**0.526**

it with state-of-the-art methods for spleen and liver segmentation: our proposed method, V-Net [9], VoxResNet [13], 3D DSD FCN [14] and MultiResUNet [15]. All results were directly predicted from single-model training and testing without relying on any post processing tools. Moreover, all networks were optimized by the initial loss functions in their own papers, and passed 5-fold cross-validation to get final average scores. We provide these results for reference and emphasize benefits of our proposed method for volumetric medical image segmentation. Table 2 summarizes the comparison results. As it can be seen, attributing to the light design MCE module and ResU-Net, our method outperforms other previous approaches on prediction time cost during the test stage, and it is worth noting that the proposed method also exceeds other results on the liver and spleen segmentation accuracy.

IV. CONCLUSIONS

In this paper, we presented a 3D unified semantic segmentation network for medical volumetric images. Our proposed network consists of a MCE module and a light 3D ResU-Net. The proposed MCE and 3D residual block are capable of gathering global and local features from images, and light 3D ResU-Net is able to complete organ segmentation effectively. Compared with state-of-the-art methods, our proposed method benefiting from its progressive structure, achieved efficient computing resource usage and accurate segmentation. After comprehensive experiments, the comparison results are listed above. We believe that our proposed network can be applied to other medical image segmentation tasks. Some limitations are presented to optimize future work.

ACKNOWLEDGMENT

This work was supported in part by the Grant-in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports (MEXT) under the Grant No.20KK0234, No.20K21821 and No.21H03470.

REFERENCES

[1] O. Ronneberger, et al. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, pp. 234-241, 2015.

[2] H. Huang, et al. "Unet 3+: A full-scale connected unet for medical image segmentation." Proc. of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.1055-1059, 2020.

[3] H. Chen, et al. "3D fully convolutional networks for intervertebral disc localization and segmentation." International Conference on Medical Imaging and Augmented Reality. Springer, Cham, pp.375-382, 2016.

[4] L. 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, 2017.

[5] H. Zhao, et al. "Pyramid scene parsing network." Proceedings of the IEEE conference on computer vision and pattern recognition, pp.2881-2890, 2017.

[6] L. Chen, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." Proceedings of the European conference on computer vision (ECCV), pp.801-818, 2018.

[7] A. Sinha, et al. "Multi-scale self-guided attention for medical image segmentation." IEEE journal of biomedical and health informatics, Vol.25, No.1, pp.121-130, 2020.

[8] K. He, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition, pp.770-778, 2016.

[9] F. Milletari,et al. "V-net: Fully convolutional neural networks for volumetric medical image segmentation." 2016 fourth international conference on 3D vision (3DV), pp. 565-571, 2016.

[10] Ö. Çiçek, et al. "3D U-Net: learning dense volumetric segmentation from sparse annotation." International conference on medical image computing and computer-assisted intervention. Springer, Cham, pp.424-432, 2016.

[11] H. Wang, et. al. Patch-Free 3D Medical Image Segmentation Driven by Super-Resolution Technique and Self-Supervised Guidance, Proc. of MICCAI2021, 2021 (in press)

[12] Lee R. Dice. "Measures of the amount of ecologic association between species." Ecology, Vol.26, No.3, pp.297-302, 1945.

[13] H. Chen, et al. "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images." NeuroImage, Vol.170, pp.446-455, 2018.

[14] B. Wang, et al. "Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation." Medical physics, Vol.46, No.4, pp.1707-1718, 2019.

[15] N. Ibtehaz, et al. "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation." Neural Networks, Vol.121 pp.74-87, 2020.