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

3D Deep Attentive U-Net with Transformer for Breast Tumor Segmentation from Automated Breast Volume Scanner

Yiyao Liu, Yi Yang, Wei Jiang, Tianfu Wang, Baiying Lei

Abstract- **Breast cancer has become the primary factor threatening women's health. Automated breast volume scanner (ABVS) is applied for automatic scanning which is meaningful for the rapid and accurate detection of breast tumor. However, accurate segmentation of tumor regions is a huge challenge for clinicians from the ABVS images since it has the large image size and low data quality. Therefore, we propose a novel 3D deep convolutional neural network for automatic breast tumor segmentation from ABVS data. The structure based on 3D U-Net is designed with attention mechanism and transformer layers to optimize the extracted image features. In addition, we integrate the atrous spatial pyramid pooling block and the deep supervision for further performance improvement. The experimental results demonstrate that our model has achieved dice coefficient of 76.36% for 3D segmentation of breast tumor via our self-collected data.**

I. Introduction

The Global Cancer Statistics 2018 showed that the incidence of breast cancer has reached the second place of all kinds of cancer. Among females, breast cancer is the most commonly diagnosed cancer and the leading cause of cancer death[1]. Automated breast volume scanner (ABVS) [2] can deliver a 3D volume image of the whole breast to detect the suspicious breast tumor. However, reading ABVS image will take huge amounts of time of clinicians. Moreover, the various sizes and the low data quality of ABVS image will cause the segmentation error.

To solve these issues and relieve the great work intensity of clinicians for tumor contouring, various automatic segmentation methods have been proposed based on neural network. For example, Roth *et al.* [3] applied the fully convolutional network (FCN) for 2D and 3D multi-organ segmentation and achieved a satisfactory result. Zhu *et al.* [4] utilized pyramid scene parsing network (PSPNet) to deliver the feasible segmentation on coronary angiography image by the ability of catching multi-scale feature. To improve the feature pyramid network (FPN) Wang *et al.* [5] integrated a deep attentive mechanism into FPN and then proposed a deep attentive feature network (DAF) which can obtain better prostate segmentation results. Attention mechanism [6] can redistribute the weights of features by training so that the network can focus more on the region of interest when attention module becomes deeper. Once U-Net is raised, it has been widely used in medical image segmentation, because of characteristics of retaining the sematic information and detail feature simultaneously. Due to the merits of U-Net, Ronneberger *et al.* [7] first made use of U-Net in biomedical area. In addition, Milletari *et al.* [8] designed V-Net specifically for volume segmentation assignments. Furthermore, it is effective to employ an atrous spatial pyramid pooling (ASPP) [9] into network for segmentation since it can acquire more information from multi-scales by expanding receptive field. For instance, Qayyum *et al.* [10] packaged an ASPP in hybrid dense network for computed tomography images segmentation. The deep supervision mechanism was introduced into segmentation network to speed up the convergence and solve the gradient vanishing problem, such as Wang *et al.* [11] applied a supervision model to detect the breast cancer on ABVS images. In 2017, Google group launched the transformer architecture for natural language processing which can acquire global information in parallel by its self-attention mechanism[12]. At present transformer was extended to visual tasks, Dosovitskiy *et al.* [13] proposed a ViT and applied it into image classification.

In this study, to address the challenge for ABVS segmentation, we propose an attentive 3D U-Net with transformer layers and ASPP. Experimental results on our self-collected ABVS data shows that our 3D network achieves quite promising results for breast cancer segmentation, which achieves a dice coefficient of 76.36%.

In summary, this work has three main contributions:

- We integrate an attention module into U-Net to focus on the tumor region.
- We adopt an ASPP module to capture the multi-scales information.
- We apply the transformer layers into the workflow of U-Net to combine the advantages of convolution and transformer.

II. Method

A. 3D U-Net

In this study, we adopt a 3D U-Net as the backbone network as shown in Fig.1. To construct the encoder block, we use two 3D convolutional layers. We utilize the rectified linear unit (ReLU) [14] as the activation function since it can decrease the amount of calculation and reduce dependency between

Y. Liu, T Wang and B. Lei is with School of Biomedical Engineering, Health Science Center, Shenzhen University, National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Shenzhen, China

Y. Yang is with Guangdong Medical University, Zhanjiang, China

W. Jiang is with Department of Ultrasonics, Huazhong University of Science and Technology Union Shenzhen Hospital, China

Correspondence should be addressed to Baiying Lei (leiby@szu.edu.cn).

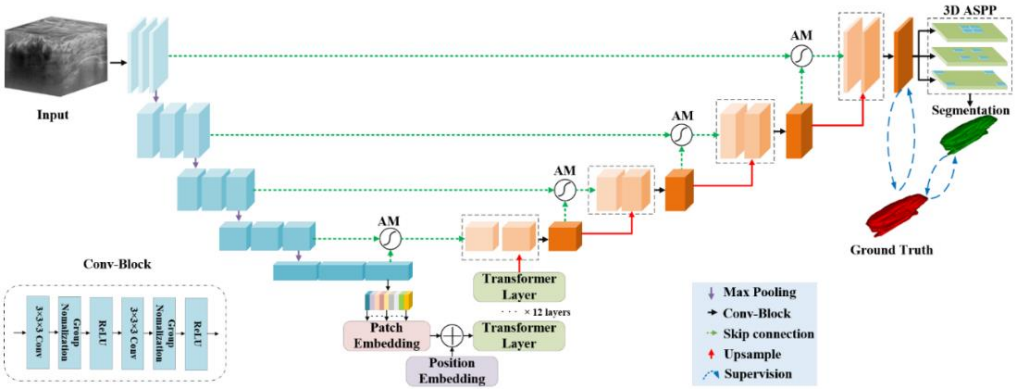

Fig.1. 3D U-Net with attention mechanism, transformer layers and ASPP block AM denotes the attentive module

parameters. Moreover, the downsampling uses a maxpooling layer with a kernel size of 2×2 and stride of 2. As for the decoder block, there is a 3D up-sampling layer with trilinear interpolation.

B. *Feature Transformer*

We apply the transformer layers between the encoder and decoder of U-Net which utilize the self-attention mechanism to obtain the long term and global information. The difference with ViT is that we use the convolutional layers to extract features and put the feature map into transformer instead of the image data. In this way, the advantages of the convolution layer for image data can be retained while the strengths of transformer on global information can be obtained.

The overview of the feature transformer we use is shown as Fig.2. First, we slice the feature map into several patches and do a patch embedding operation. Then to maintain the position information of each patch we need to do position embedding. After that, the feature was transformed into query, key and value vectors through linear layers. Afterwards these vectors will be calculated according to the following formula by a multi-head self-attention layer:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{head\_size}})V, \quad (1)$$

Where Q denotes query vector, K denotes key vector and V denotes value vector. Finally, the output passes through a multi-layer perceptron consisting of multiple linear layers. The normalization layers embedded in it can accelerate the speed of convergence.

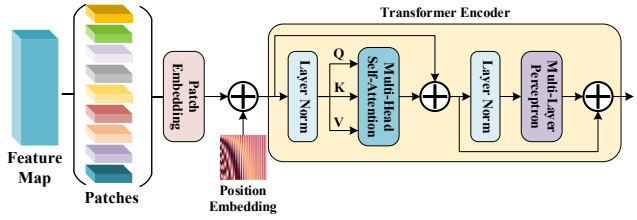

Fig.2.The feature transformer is composed of patch embedding, position embedding, layer normalization, multi-head self-attention and multi-layer perceptron. And Q denotes the query vector, K denotes the key vector and V denotes the value vector.

C. *Attention Mechanism*

To identify the tumor more accurate, an attention mechanism is applied to the backbone. In the decode progress, an attention gate (AG) [15] is used to calculate the attention coefficients, which is shown in Fig.2. The AGs add up the previous decoder layer feature and the same encoder layer feature, and then make the sum-feature to be activated by ReLU and sigmoid function.

$$\mathcal{F}_{att} = \left[\sigma\left(\mathcal{F}_i^E \oplus \mathcal{F}_{i-1}^D\right) \otimes \mathcal{F}_i^E\right] \mathbb{C} \mathcal{F}_{i-1}^D, \quad (2)$$

where $\mathcal{F}_i^E$ denotes the encoder layer feature, $\mathcal{F}_{i-1}^D$ denotes the previous decoder layer feature and $\mathcal{F}_{att}$ is the attention guided feature. The symbol $\oplus$ represents the element-wise summation, $\otimes$ represents the element-wise multiplying and $\mathbb{C}$ represents the concatenate operation. And we can observe that the weights skewed towards the tumor region from the attention map.

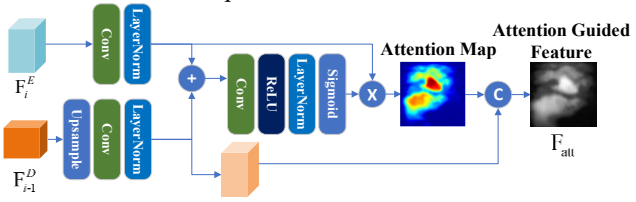

Fig.3.Attention mechanism: $\mathcal{F}_i^E$ denotes the encoder layer feature, $\mathcal{F}_{i-1}^D$ denotes the previous decoder layer feature and $\mathcal{F}_{att}$ is the attention guided feature.

D. *Atrous Spatial Pyramid Pooling*

To acquire more information from different scales, we utilize an ASPP block in our model. The ASPP block is based on the dilated convolutional operation by inserting zero holes into convolution kernels. We use four ASPP blocks with the same kernel size of 3×3×3, but these blocks are in different dilated rates [1,6,12,18]. Firstly, there is a 3D convolution layer in the blocks which is set up as described above. Secondly, after convolution, the feature needs to be normalized by a group normalization layer. Thirdly, the features are concatenated through the four ASPP blocks. Last, we get the output by convolving the result of previous step.

E. Deep Supervision

In our study, we propose a deep supervision mechanism, which can produce a better result efficiently. To achieve it, we need to get a temporary output, which calculates its loss and combines the temporary loss with the last output's loss to supervise the whole parameters. Two types of loss functions, binary cross entropy loss (BCELoss) and dice loss are involved, which are defined as:

$$\mathcal{L}_{bce} = -\sum_i^n (\widehat{y}_i \log y_i + (1-\widehat{y}_i)\log(1-\widehat{y}_i)), \quad (3)$$

$$\mathcal{L}_{dc} = 1 - \frac{2*|\widehat{Y}_i \cap Y_i|}{|\widehat{Y}_i| \cup |Y_i|}, \quad (4)$$

where the $\mathcal{L}_{bce}$ denotes the binary cross entropy loss, $\mathcal{L}_{dc}$ denotes the dice loss, $\widehat{y}_i$ represents the predicted value and y_i represents the ground truth. $\widehat{Y}_i$ and Y_i are sets of $\widehat{y}_i$ and y_i, respectively. Therefore, the final loss is formulated as:

$$\mathcal{L} = \lambda(\theta_t \mathcal{L}_{bce} + \mathcal{L}_{dc)} + (\theta_o \mathcal{L}_{bce} + \mathcal{L}_{dc}), \quad (5)$$

where θ_t and θ_o is coefficient between 0-1 and decided by $\mathcal{L}_{dc}$. And λ is a coefficient between 0-1.

III. Experiments And Results

A. Data and Preprocessing

We collect 103 volumes of ABVS images from patients aged 18-69 years old between Jan. 2019 and Sep. 2020. Each case is delivered by Siemens Acuson OXANA2 system. To mark out the ground truth, our datasets were evaluated independently by three ultrasound doctors. Due to the GPU memory limit, we need to crop the images from 300×500×700 pixels into 64×256×256 pixels by label when training. When testing, we use a sliding window to predict each patch and generate the final result by combining all patches.

B. Experimental Setting-up

In this study, we have 103 valid volumes with labels, and the experiment was carried out by 5-fold cross validation. We use the PyTorch framework to train the models on a single TITAN RTX GPU with 24GB memory. During the training process, the initial learning rate is set to 10^{-4} and utilizes a learning rate decay mechanism with 0.85 decay coefficient. To further verify the effectiveness of our network, we compare it with several current models such as U-Net, V-Net, FCN, PSPNet and Deep Attention Features (DAF) network. To evaluate the result of each network, the metrics such as Dice coefficient (DSC), Hausdorff_95 coefficient (HD_95) [16], Jaccard coefficient (JI), Precision (PRE), Sensitivity (SEN) and Specificity (SPEC) are used.

C. Ablation Study

In order to illustrate the effectiveness of each module clearly, we have done generous ablation experiments. Table 1 shows the effectiveness of different module. We can observe that designed module improves the performance greatly. The attention mechanism, transformer layers and ASPP block improve the segmentation accuracy of original U-Net. And we discovery that the deep supervision module speeds up the convergence.

D. Results

Table 1 illustrates the means and standard deviations (SD) of each state-of-art model used in this paper. It shows that our model has achieved the best score among all networks. From Table 1, we can see that our model has the highest mean dice coefficient of 76.36% with lower SD values. Therefore, we can infer that our model gets a more stable performance on this dataset. In addition, our model obtains the best JI score of 62.14% Furthermore, in other metrics, our model also performs quite well in terms of HD_95, SPEC and PRE.

E. Visualization

In order to have a more intuitive comprehending, we visualize the results of the segmentation in several ways. Fig.4 is the 3D segmentation results with different methods. Each row of Fig.4 shows different volumes image, and images in the first to last columns represent the results of FCN, PSPNet, U-Net, V-Net, DAF and our model, respectively. Fig.4 shows red and blue denotes the further distance and green denote closer distance. We can observe that our model performs well in terms of distance visualization. Fig.5 is the 2D segmentation visualization results with different methods, where each column denotes different volumes image. In Fig.5, the left pictures are cropped original images with the ground truth in red area, the purple, wathet, blue, yellow and pink lines are

Table 1 Ablation analysis of different module in the proposed 3D network and performance comparison of state-of-art methods used in this study (mean± SD). Boldface denotes the best performance. U-Net is the baseline in ablation experiment. AM denotes the attention mechanism, ASPP denotes the atrous spatial pyramid pooling and T denotes the transformer layers.

Method	DSC(%)	JI(%)	HD_95(voxel)	PRE(%)	SEN(%)	SPEC(%)
V-Net	60.83±14.12	45.12±14.49	21.66±10.92	55.26±24.56	83.85±15.82	94.88±3.93
DAF	73.44±11.47	59.20±13.08	15.75±14.33	72.64±16.37	81.32±14.10	97.29±4.46
3D FCN	63.18±10.60	47.03±11.16	25.31±11.96	62.32±19.09	72.55±16.26	97.03±1.92
3D PSPNet	62.79±11.95	46.79±12.18	21.84±11.23	61.56±20.02	72.02±15.58	98.04±1.48
U-Net	61.24±14.61	45.64±14.43	19.99±12.10	51.01±17.56	**88.63±15.85**	97.06±2.06
U-Net+AM	71.30±8.82	56.15±10.87	**13.67±6.58**	67.51±15.68	80.23±12.45	98.22±1.99
U-Net+ASPP	67.31±12.34	52.03±14.22	17.54±10.28	60.16±16.73	83.55±14.32	97.01±2.45
U-Net+AM+ASPP	75.39±6.72	60.94±8.66	15.15±9.06	**87.10±10.57**	68.48±12.21	**99.43±0.73**
U-Net+AM+ASPP+T(Ours)	**76.36±6.11**	**62.14±7.99**	15.47±11.92	78.95±9.91	75.42±9.24	98.85±1.14

labeled by FCN, PSPNet, U-Net, V-Net and DAF networks, respectively. And the green line is labeled by our model.

IV. Conclusion

In this paper, we propose an attention 3D U-Net with feature transformer which aims at breast tumor segmentation. To achieve this model, we integrate attention module, ASPP and feature transformer layers into 3D U-Net. To this end, we address the 3D segmentation challenge of various sizes and low data quality in ABVS data. The experimental results illustrate that our model can deliver the most satisfactory tumor segmentation performance compared with other models.

V. Compliance With Ethical Standards

We wish to confirm that there are no known conflicts of interest associated with this publication. This research study is approved by the ethical review board of the institute.

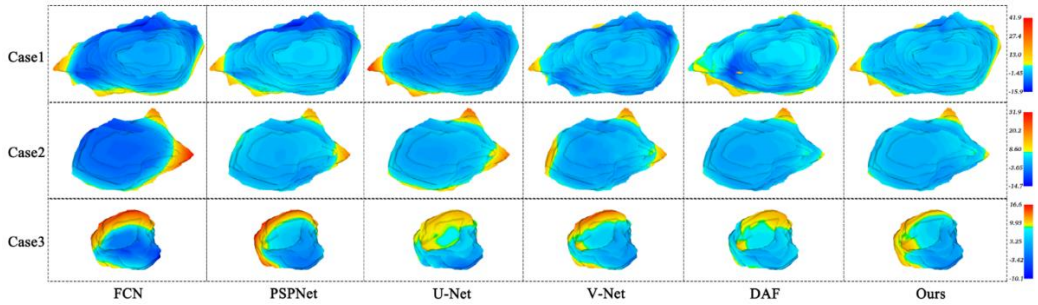

Fig.4. The segmentation results by different methods visualized in 3D view. The red and blue area indicates the greater difference between the result and the ground truth, and green area indicates conversely.

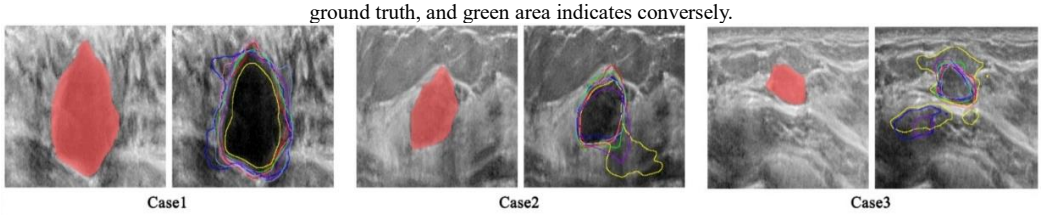

Fig.5. The segmentation results by different methods visualized in 2D view. The left pictures are cropped original images with the ground truth in red area, the purple, wathet, blue, yellow and pink lines are labeled by FCN, PSPNet, U-Net, V-Net and DAF networks, respectively. And the green line is labeled by our model.

References

[1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A.: "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", CA Cancer J Clin, 68, (6), pp. 394-424,2018

[2] Schmachtenberg, C., Fischer, T., Hamm, B., and Bick, U.: "Diagnostic Performance of Automated Breast Volume Scanning (ABVS) Compared to Handheld Ultrasonography With Breast MRI as the Gold Standard", Academic Radiology, 24, (8), pp. 954-961,2017

[3] Roth, H.R., Oda, H., Zhou, X.R., Shimizu, N., Yang, Y., Hayashi, Y., Oda, M., Fujiwara, M., Misawa, K., and Mori, K.: "An application of cascaded 3D fully convolutional networks for medical image segmentation", Comput. Med. Imaging Graph., 66, pp. 90-99,2018

[4] Zhu, X., Cheng, Z., Wang, S., Chen, X., and Lu, G.: "Coronary angiography image segmentation based on PSPNet", Computer Methods and Programs in Biomedicine, 200, pp. 105897,2021

[5] Wang, Y., Dou, H.R., Hu, X.W., Zhu, L., Yang, X., Xu, M., Qin, J., Heng, H.A., Wang, T.F., and Ni, D.: "Deep Attentive Features for Prostate Segmentation in 3D Transrectal Ultrasound", IEEE Trans. Med. Imaging, 38, (12), pp. 2768-2778,2019

[6] Du, J.C., Gui, L., He, Y.L., Xu, R.F., and Wang, X.: "Convolution-Based Neural Attention With Applications to Sentiment Classification", IEEE Access, 7, pp. 27983-27992,2019

[7] Ronneberger, O., Fischer, P., and Brox, T.: "U-Net: Convolutional Networks for Biomedical Image Segmentation", Medical Image Computing and Computer-Assisted Intervention, Pt Iii, 9351, pp. 234-241,2015

[8] Milletari, F., Navab, N., and Ahmadi, S.-A.: "V-net: Fully convolutional neural networks for volumetric medical image segmentation". Proc. 2016 fourth international conference on 3D vision (3DV), 2016 pp. 565-571

[9] Chen, J., Wang, C.Y., and Tong, Y.: "AtICNet: semantic segmentation with atrous spatial pyramid pooling in image cascade network", EURASIP J. Wirel. Commun. Netw., pp. 7,2019

[10] Qayyum, A., Ahmad, I., Mumtaz, W., Alassafi, M.O., Alghamdi, R., and Mazher, M.: "Automatic Segmentation Using a Hybrid Dense Network Integrated With an 3D-Atrous Spatial Pyramid Pooling Module for Computed Tomography (CT) Imaging", IEEE Access, 8, pp. 169794-169803,2020

[11] Wang, Y., Wang, N., Xu, M., Yu, J., Qin, C., Luo, X., Yang, X., Wang, T., Li, A., and Ni, D.: "Deeply-Supervised Networks With Threshold Loss for Cancer Detection in Automated Breast Ultrasound", IEEE Trans Med Imaging, 39, (4), pp. 866-876,2020

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I.: "Attention is all you need", arXiv preprint arXiv:1706.037622017

[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., and Gelly, S.: "An image is worth 16x16 words: Transformers for image recognition at scale", arXiv preprint arXiv:2010.119292020

[14] Jiang, X.H., Pang, Y.W., Li, X.L., Pan, J., and Xie, Y.H.: "Deep neural networks with Elastic Rectified Linear Units for object recognition", Neurocomputing, 275, pp. 1132-1139,2018

[15] Zhang, J.X., Jiang, Z.K., Dong, J., Hou, Y.Q., and Liu, B.: "Attention Gate ResU-Net for Automatic MRI Brain Tumor Segmentation", IEEE Access, 8, pp. 58533-58545,2020

[16] Marosevic, T.: "The Hausdorff distance between some sets of points", Math. Commun., 23, (2), pp. 247-257,2018