Improved 3D fully convolutional network based on squeeze-excitation method for liver segmentation

Xiaoling Xia¹ and Nana Hao¹,*

¹ School of Computer Science and Technology, Donghua University, Shanghai, Shanghai, 201620, China

*Corresponding author’s e-mail: 2191931@mail.dhu.edu.cn
*Corresponding author’s ORCID: https://orcid.org/0000-0001-8715-1669

Abstract. Due to the variable shape and size of the liver in abdominal CT and the complex surrounding tissues, accurate liver segmentation in CT remains a challenge. In the field of two-dimensional image analysis, the squeeze-excitation (SE) method effectively adjusts the input feature information. Our work is based on SE method to obtain feature information that is more relevant to the liver region through improved feature recalibration methods. This work adds the MSCR block, which can do space and channel feature recalibration, to 3D full convolution network. The 3D space part of the MSCR block re-calibrates the feature of liver voxel space. And in the channel recalibration part, it changes the original spatial compression operation according to the none-local method, so the spatial sampling results contain more global context information. Multiple metrics are used to evaluate the proposed model on the LiTS dataset, and it achieves better segmentation performance than other comparison models such as Attention U-Net. The method is further tested on the 3DIRCADb dataset, which proves its effectiveness and stability. Thus, the model proposed in the paper is effective for improving the performance of liver segmentation.

1. Introduction

According to the latest statistics of the World Health Organization in 2020, the incidence of hepatocellular carcinoma ranks fifth among all cancers in the world, with 830,000 deaths due to hepatocellular carcinoma, second only to lung and intestinal cancer [1]. Medical imaging technology such as computed tomography (CT) is one of the frequently used methods to evaluate the condition of liver tumors and estimate the survival rate of patients in clinical practice. Segmentation of organs or lesion areas from medical images in the process of computer-assisted therapy can help physicians effectively understand the condition and thus adopt more appropriate treatment regimens to improve the survival probability of patients further. And achieving precise segmentation of the liver region can prepare for further effective tumor segmentation.

In order to effectively improve the efficiency of segmentation in this field, some researchers applied traditional image processing methods such as threshold segmentation [2]-[3], region growth [4]-[5] to the field of medical image segmentation in the early days. However, these methods based on hand-designed features cannot effectively extract high-level semantic feature information from images. Moreover, especially for some liver regions with complex and variable features in the image, these methods are difficult to obtain high-quality automatic segmentation results.

With the continuous development of deep learning, convolutional neural networks have played a major role in image classification with strong feature abstraction and learning capabilities. Long [6] and others
applied convolution neural networks to the field of semantic segmentation and proposed full convolutional networks (FCN). The depth neural network was used to extract the depth features of the input image of any size, and then the deconvolution was used to up sample the feature image. Finally, the mask with the same size as the input image is obtained through effective reasoning of pixel level classification of feature image, which provides a new idea for end-to-end semantic segmentation. The two-stage cascade method is used to separate the liver and the lesion region independently, and used the FCNs with dense connections to complete the localization of the liver region[7]. Delmoral[8] and others proposed the introduction of dilated convolution in the FCNs, with the goal of introducing richer context information into the feature map, thereby improving the segmentation performance of liver tissue. Ronneberger[9] and others improved the structure of FCNs, and proposed U-Net with a symmetrical encoding-decoding structure. This method uses skip connections to combine the high-resolution feature maps output by each stage of the encoding module with the feature maps obtained by the up-sampling of the corresponding decoding module, and then go through the convolutional layer to obtain a more refined segmentation result. This method is very applicable in the field of medical image segmentation where training data is lacking. 3D U-Net[10] is a three-dimensional extension of U-Net, using volumetric data directly as the input of the model. Compared with the two-dimensional input, it can learn the information between slices and ensure the continuity of the segmentation results. Oktay[11] and others proposed the Attention U-Net. It added attention gating (AG) module to the skip connection to suppress non-target region features and give more weight to the target region. It achieved good performance in abdominal CT multi-organ segmentation. ResU-Net proposed by Xiao[12] and others and Dense U-Net proposed by Wang[13] and others. They both established multi-layer information connections by adding residual modules or dense connection modules to U-Net. The feature information between multiple layers was further utilized, which improved the segmentation performance of the model to a certain extent.

Although the model methods proposed above had achieved good performance, due to the influence of noise caused by imaging equipment or injection of contrast agents in liver CT, or the changeable shape of the liver itself and the complex characteristics of adjacent tissues around it, there is room for further improvement in the performance of liver segmentation. The feature information captured by convolution is limited, and the global context information cannot be effectively used.

In our work, in order to further improve the accuracy of liver segmentation in response to the complex situation of liver in CT, we proposed a model called 3D Modified Space and Channel Recalibration(MSCR) U-Net, which mainly used the block that performs the squeeze -excitation methods[14-15]on the three-dimensional space and the channel respectively, and obtained the weight according to the operation to re-calibrate the feature information on the space and the channel. Regarding the channel recalibration block, we referred to the non-local method used in[16], and changed the spatial context modelling module to replace the original global average pooling operation of the SE module to obtain more effective spatial downsampling results. The block proposed in the paper aims to capture effective and comprehensive contextual information through the above methods and improve the limitations of extracting information from the convolutional layer, and enhance the feature information related to the segmentation target, thereby improving the effect of liver segmentation.

![Figure 1. The proposed model](image-url)
2. Method

The model mainly includes the encoders for feature extraction and the decoders for progressive up sampling for liver segmentation. The focus of this paper is how to comprehensively learn the features that are effective for the liver region in the process of feature learning and inference; the model structure is shown in Figure 1. First, before inputting the data into the model, we do some pre-processing operations such as setting the window width and level, resampling and so on. The processed data are input into the model in the form of three-dimensional patches. After a series of convolution operations and nonlinear transformations, the spatial and channel level information of the feature map is continuously learned and integrated. After up-sampling and inference, the prediction result consistent with the size of the input is finally obtained. We added the MSCR block, which consists of the 3D space SE block and the modified channel recalibration (MCR) block, after each encoder or decoder to accept their output, as shown in Figure 2. The purpose is to use this block to calculate the importance of each location or each channel, and to re-calibrate the spatial feature map or channel feature map according to the weight to obtain feature information more relevant to the target region. The output of the block after encoder is passed to the next encoder and the corresponding decoder, and the output of the module after decoder is input to the next decoder. The implementation details of MSCR block will be further described below.

**Figure 2.** MSCR block

2.1. 3D space SE block

The weight of each position learned through the process of compressing the feature map on the channel and then exciting on the spatial position helps to achieve the fine segmentation of the target region[14].

We consider the original input $I \in \mathbb{R}^{C \times D \times H \times W}$ as the collection of pixels at all positions in the space, $\{i^{x,y,z} \in \mathbb{R}^{C \times 1 \times 1} \}$ is the feature information of a position $(x, y, z)$ in the three-dimensional space, $x \in \{1, 2, ..., D\}, y \in \{1, 2, ..., H\}, z \in \{1, 2, ..., W\}$. Then let $I$ through a $1 \times 1 \times 1$ convolution to reduce the number of channels $C$ to 1, the process can be formulated as follows,

$$m = W_p \ast I$$

where $W_p \in \mathbb{R}^{C \times 1 \times 1 \times 1}$ is weight of convolution, and $m \in \mathbb{R}^{1 \times D \times H \times W}$ is the output after convolution. $m_{x,y,z}$ is the value position $(x, y, z)$, which is got by linearly integrating C channels at this position in $m$. Use the activation function to convert the values in m into $[0, 1]$, indicating the importance of each position in the context, and then multiply it with the original input $I$, the process can be formulated as follows,

$$O_p = [\sigma(m_{1,1,1})i^{1,1,1}, ..., \sigma(m_{x,y,z})i^{x,y,z}]$$

where $\sigma(\cdot)$ is sigmoid activation of $\cdot$, the value of $\sigma(m_{x,y,z})i^{x,y,z}$ represents the importance of the spatial position $(x, y, z)$ at all positions in the input, and $O_p$ is the output of the module. This process can be simply described as re-calibrating the original feature map by using the learned weight information of each location to pay more attention to the spatial location related to the task, thereby suppressing the influence of those irrelevant location features. The structure of this block is as shown in Figure 3.

2.2. 3D MCR block

Re-calibrate the importance of each channel after spatial squeeze and channel excitation of the feature map, so as to pay more attention to the effective channel in the learning process[15]. Previously, the space squeeze achieved by global average pooling operation, but the non-local module can effectively
fuse context information, and can effectively model long-range dependence [16]. We use the simplified non-local method instead of global average pooling to do space squeeze. The input \( I = \{i^1, ..., i^x\}, x \in \{1, 2, ..., N\} \), \( N = D \times H \times W \) is the number of all positions in the feature map. \( I \) is calculated by a \( 1 \times 1 \times 1 \) convolution and softmax to obtain a weight map, which contains global context information, and the operation can be formulated as follows,

\[
n = \sum_{p=1}^{N} \frac{\exp(W_s i^p)}{\sum_{q=1}^{N} \exp(W_s i^q)}
\]

where \( W_s \) is the weight of the \( 1 \times 1 \times 1 \) convolution which is a linear transformation matrix, and \( n \) is the output, which is the result of spatial squeeze based on the weight of the global attention pooling. Then learn the inter-channel dependence information of \( n \), the process can be formulated as follows,

\[
O_c = i^p \cdot \sigma(W_2 \cdot \text{ReLU}(W_1 n))
\]

where \( (W_2 \cdot \text{ReLU}(W_1 \cdot \cdot)) \) changes the number of channels of feature map \( f \) from \( C \) to \( C' \), and then to \( C \), which is like a bottleneck. \( W_s \) and \( W_2 \) respectively represent the weight of the two convolutions. We can get the weight dependency between these channels of the feature map. Then, the output containing the channel weight information obtained by this process is multiplied with the original input \( I \) to perform feature fusion, so that the channels in the feature map are given different attention. The structure of the block is as shown in Figure 4.

Finally, the two blocks are combined to the whole MSCR block. Specifically, Max-Out is used to fuse the output feature maps of two blocks, so as to re-calibrate the input \( I \) in space and channel through the weights learned by the two blocks. And then, the model can learn the most effective feature information for the target region through both of them.

3. Experiment

We build the model based on Pytorch 1.8.1 in the computer with windows10 64-bit operating system and CPU is Intel(R) Gold 5218R, and we train models on the RTX 2080 Ti GPU with 16 GB memory. The task is the segmentation of liver organs in abdominal CT images. The datasets, training parameters, evaluation metrics and results of this work are described as follows.

3.1. Dataset

In order to effectively evaluate the model, we use some public datasets in our experiment. (1) The dataset mainly used in the paper is from LiTS-ISBI2017 [17], which is currently a larger public data set related to liver and tumor segmentation. The data set contains 131 3D abdominal CT scan images with annotations. The data is collected from 6 medical institutions in different countries and regions. The
liver region in each slice of each case is labelled by a professional doctor. The data and labels are real and valid. In the experiment, 98 of the 131 data and labels will be randomly selected as the training dataset, and the remaining 33 cases will be used as the test dataset to evaluate these trained models. (2) This paper uses the 3DIRCADb dataset [18] to further evaluate the generalization of models trained. The dataset contains 20 3D abdominal CT scans from various European hospitals, and the CT of each case has effectively annotated by professional doctors as the ground truth (GT).

3.2. Parameter setting and training details
When training the model, we use the dice loss between the input and the prediction as the optimization target, the optimizer is Adam. For both the datasets, the batch size is 2, and the learning rate change can be formulated as follows,

\[ lr_{\text{new}} = lr_{\text{initial}} \times \gamma^{\text{epoch}} \]

where \( lr_{\text{initial}} \) is the initial learning rate, which is set to 0.0001. \( lr_{\text{new}} \) is the updated learning rate, \( \gamma \) is the multiplication factor, which is set to 0.95, \( \text{epoch} \) is the number of times that all training data in the training set is input to the model for training.

During the experiment, we first pre-process 98 CT used for training in LiTS dataset, including cutting the slices containing only the liver region at transverse section, and performing data enhancement processing such as window width and window level value setting on the slices. Finally, the original 3D abdominal CT image is divided into a series of 3D image patches with the size of \( 16 \times 256 \times 256 \) and input them into the model for training. We also performed similar data preprocessing on the data in the test dataset. The CT in the test dataset cover representative situations such as different liver shapes and sizes, and complex adjacent organs. Then we applied the model trained on the first dataset to the second dataset to further verify the segmentation performance of the model and observe the generalization of the model.

3.3. Evaluation Metrics
We use some metrics to evaluate the liver segmentation performance of these models. The main metric is Dice score, which can be formulated as follows,

\[ DICE(A, B) = \frac{2|A \cap B|}{|A| + |B|} \]

where, A and B respectively represent GT and prediction mask. This formula can get the degree of overlap between the two. The value is between [0,1], when a value closer to 1 represents the perfect division.

In addition, we have added some other indicators, including the volumetric overlap error (VOE), average symmetric surface distance (ASSD), false negative rate (FNR) and false positive rate (FPR) to more comprehensively evaluate models.

3.4. Results and Analysis
We analyze the results on the segmentation performance of the model on the test dataset, which are randomly divided from the LiTS dataset, and the 3DIRCADb dataset.

Table 1 shows the segmentation results of the model proposed in the paper and some other models on the test dataset divided in the LiTS dataset. The preprocessing of the input data of all models and other settings such as training parameters are kept consistent to ensure the effectiveness of the comparison experiment. Compared with the other two models, our proposed model achieves a higher Dice score, the average volume overlap error is also the lowest, and the average surface distance is also much lower than other models, which means the contours of the results of our model segmentation are closer to the GT. Although the false positive rate of segmentation is slightly higher than other models, the false negative rate is significantly lower than other models, which shows that our model has less under-segmentation. Our model for liver segmentation outperforms the other two models in most evaluation metrics.
Table 1. Results on the LiTS dataset

| Model            | Dice | VOE  | ASSD | FNR  | FPR  |
|------------------|------|------|------|------|------|
| Res U-Net[12]    | 0.936| 0.123| 2.778| 0.092| 0.031|
| Attention U-Net[11]| 0.934| 0.132| 2.510| 0.093| 0.038|
| MSCR U-Net       | **0.948** | **0.108** | **1.978** | **0.068** | 0.040 |

Table 2. Results on the 3DIRCADb dataset

| Model            | Dice | VOE  | ASSD | FNR  | FPR  |
|------------------|------|------|------|------|------|
| Res U-Net[12]    | 0.913| 0.143| 3.694| 0.112| 0.031|
| Attention U-Net[11]| 0.907| 0.152| 2.680| 0.126| **0.025** |
| MSCR U-Net       | **0.942** | **0.105** | **1.778** | **0.063** | 0.041 |

In order to further verify the effectiveness of our model, we use another liver dataset 3DIRCADb dataset to test these models, and the results are shown in Table 2. According to the results, we get that the proposed model is still superior to other models in most metrics. And the model still maintains a relatively stable segmentation performance in additional datasets, which shows that the model is effective and robust for liver segmentation.

![Comparison of qualitative results between different models](image_url)

Figure 5. Comparison of qualitative results between different models

We randomly choose some slices from results obtained by each model on the test data set to evaluate their qualitative results, as shown in Figure 5. We show the slices of GT and liver masks obtained from the models, Attention U-Net, Res U-Net and MSCR U-Net, from left to right. First of all, in the region marked by the white rectangle in the first row of the figure, it can be seen that the comparison models have different degrees of under-segmentation, but our model has no such situation, which is closer to GT. In the region marked by the white rectangle in the second row of the figure, the liver features are more complicated, and there are some unsegmented liver regions in the comparison model. Our model can segment the liver in this region better. In the third row of the figure, there is a large region of mis-segmentation in the segmentation result of Attention U-Net. The mis-segmentation part in the segmentation result of Res U-Net is significantly reduced, but there are still some mis-segmented parts around the liver. But in this case, the segmentation results obtained by our model hardly have mis-segmented regions.
4. Conclusion
In the paper, we propose the 3D MSCR U-Net of liver segmentation, which includes improved space and channel squeeze-excitation blocks. First, in space, we use the learned weight of each position to recalibrate the original feature map spatially, which assigns more weight to the position related to the liver and reduces the impact of other regional characteristics. And secondly, similar operations are used on the channel. In this part, we change the original space squeeze operation and use a simplified non-local method to fuse context information, which can effectively model long-range dependencies. Finally, the whole block can integrate the feature maps after calibration of them simultaneously to effectively learn the related feature information for target segmentation.

In the experiment, we use two public liver datasets, LiTS dataset and 3DIRCADb dataset, to evaluate models. On the first dataset, our model is compared with other liver segmentation methods based on deep learning. It has been proved by experiments that the method in this paper is superior to other methods in Dice, VOE, ASSD and other metrics. The model is further tested on the second dataset, and the results prove that our model in the dataset is stable, and has higher segmentation performance than other models.

References
[1] International Agency for Research on Cancer, World Health Organization. (2020) Cancer today. https://gco.iarc.fr/today/home.
[2] Moltz, J. H., Bornemann, L., Dicken, V., & Peitgen, H. (2008) Segmentation of liver metastases in CT scans by adaptive thresholding and morphological processing. In: MICCAI workshop. New York. pp. 195.
[3] Zhu, H., Zhuang, Z., Zhou, J., Zhang, F., Wang, X., & Wu, Y. (2017) Segmentation of liver cyst in ultrasound image based on adaptive threshold algorithm and particle swarm optimization. Multimedia Tools and Applications, 76(6): 8951-8968.
[4] Wong, D., Liu, J., Fengshou, Y., Tian, Q., Xiong, W., Zhou, J., ... & Wang, S. C. (2008) A semi-automated method for liver tumor segmentation based on 2D region growing with knowledge-based constraints. In: MICCAI workshop. New York. pp. 159.
[5] Qi, Y., Xiong, W., Leow, W. K., Tian, Q., Zhou, J., Liu, J., ... & Wang, S. C. (2008) Semi-automatic segmentation of liver tumors from CT scans using Bayesian rule-based 3D region growing. In: MICCAI workshop. New York. pp. 201.
[6] Long, J., Shelhamer, E., & Darrell, T. (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Boston. pp. 3431-3440.
[7] Kaluva, K. C., Khened, M., Kori, A., & Krishnamurthi, G. (2018) 2D-densely connected convolution neural networks for automatic liver and tumor segmentation. arXiv preprint arXiv:1802.02182.
[8] Delmoral, J. C., Costa, D. C., Borges, D., & Tavares, J. M. R. (2019) Segmentation of pathological liver tissue with dilated fully convolutional networks: A preliminary study. In: 2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG). Lisbon. pp. 1-4.
[9] Ronneberger, O., Fischer, P., & Brox, T. (2015) U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Munich. pp. 234-241.
[10] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016) 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. Athens. pp. 424-432.
[11] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... & Rueckert, D. (2018) Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.
[12] Xiao, X., Lian, S., Luo, Z., & Li, S. (2018) Weighted res-Unet for high-quality retina vessel segmentation. In: 2018 9th international conference on information technology in medicine and education (ITME). Hangzhou. pp. 327-331.
[13] Wang, Z. H., Liu, Z., Song, Y. Q., & Zhu, Y. (2019) Densely connected deep u-net for abdominal multi-organ segmentation. In: 2019 IEEE International Conference on Image Processing (ICIP). Taipei. pp. 1415-1419.

[14] Roy, A. G., Navab, N., & Wachinger, C. (2018) Re-calibrating fully convolutional networks with spatial and channel “squeeze and excitation” blocks. IEEE transactions on medical imaging, 38(2): 540-549.

[15] Hu, J., Shen, L., & Sun, G. (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, Utah. pp. 7132-7141.

[16] Cao, Y., Xu, J., Lin, S., Wei, F., & Hu, H. (2019) GCnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Seoul. pp. 0-0.

[17] Bilic, P., Christ, P. F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., ... & Menze, B. H. (2019) The liver tumor segmentation benchmark (lits). arXiv preprint arXiv:1901.04056.

[18] Soler, L., Hostettler, A., Agnus, V., Charnoz, A., Fasquel, J., Moreau, J., ... & Marescaux, J. (2010) 3D image reconstruction for comparison of algorithm database: a patient-specific anatomical and medical image database. IRCAD, Strasbourg. France, Tech. Rep.