MSAA-Net:A Multi-Scale Attention-Aware U-Net is Used to Segment the Liver

Lijuan Zhang
Changchun University of Technology
Jiajun Liu
Changchun University of Technology
Dongming Li (LDM0214@163.com)
Jilin Agricultural University
Jinyuan Liu
Changchun University of Technology
Xiangkun Liu
Changchun University of Technology

Research Article

Keywords: Deep learning, Liver segmentation, Attention mechanism, U-Net

Posted Date: May 9th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1625950/v1

MSAA-Net:A Multi-Scale Attention-Aware U-Net is Used to Segment the Liver

Lijuan Zhang[1], Jiajun Liu[1], Dongming Li[2*], Jinyuan Liu[1] and Xiangkun Liu[1]

[1]College of Computer Science and Engineering, Changchun University of Technology, Changchun, 130012, Jilin, China.
[2]School of Information Technology, Jilin Agricultural University, Changchun, 130118, Jilin, China.

*Corresponding author(s). E-mail(s): LDM0214@163.com;
Contributing authors: zhanglijuan@ccut.edu.cn;
ljj924726912@163.com; 1354189527@qq.com; 599210633@qq.com;

Abstract

Automatic segmentation of the liver from CT images is a very challenging task because the shape of the liver in the abdominal cavity varies from person to person and it also often fits closely with other organs.In recent years, with the continuous development of deep learning and the proposal of CNN, the neural network-based segmentation models have shown good performance in the field of image segmentation.Among the many network models, U-Net stands out in the task of medical image segmentation.In this paper, we propose a segmentation network MSAA-Net combining multi-scale features and an improved attention-aware U-Net.We extracted features of different scales on a single feature layer and performed attention perception in the channel dimension.We demonstrate that this architecture improves the performance of U-Net while significantly reducing computational costs.To address the problem that U-Net's skip connection is difficult to optimize for merging objects of different sizes, we designed a multi-scale attention gate structure (MAG), which allows the model to automatically learn to focus on targets of different sizes. In addition, MAG can be extended to all structures which contain skip connections, such as U-Net and FCN variants.Our structure was evaluated on the 3Dircadb dataset, and the method obtained a 98.29% Dice coefficient for the liver segmentation task, with a model parametric number only equivalent to 38.4% of Attention U-Net. The experimental results show that MSAA-Net achieves the highest segmentation accuracy while saving arithmetic power.

Keywords: Deep learning, Liver segmentation, Attention mechanism, U-Net

1 Introduction

According to the Global Burden of Cancer Status Report (GLOBOCAN), liver cancer is one of the most common cancers, and its incidence and mortality rates are increasing worldwide[1].Computed tomography (CT) is one of the main tools currently used to detect liver cancer.Segmenting areas of the liver from high-resolution CT images of the abdomen is the first step in treating liver cancer, but this task is too dependent on the experience of the physician and is very energy consuming.With the continuous development of computer technology, the automatic detection and segmentation of liver has become an important research direction in the field of medical image segmentation.Different structures in CT images have the characteristics of uneven gray scale and similar gray scale, which makes the task of accurately segmenting the liver area from the image of multiple organs combined together is challenging.

The methods of liver CT image segmentation can be divided into three categories: manual segmentation, semi-automatic segmentation, and fully automatic self-segmentation[2].Manual segmentation relies entirely on the experience of specialists and physicians.There are dozens to hundreds of CT images of a patient, all of which need to be manually outlined, which is a tedious and very energy-consuming task, and the segmentation results are affected by the subjective experience of the doctor.Semi-automatic segmentation methods rely on less manual effort, but still require manual annotation of some regions, such as region growth and graph cut methods. Fully automatic segmentation methods can segment the liver without manual intervention at all, which is a significant labor saving, and thus fully automatic segmentation methods have become the primary choice for liver image segmentation work.The fully automatic segmentation method of liver based on deep learning is the most hotly researched segmentation method at present.The most commonly used neural network models in the field of medical image segmentation are some variants of U-Net [3] and FCN [4].For example, Cicek et al [5] proposed a 3D U-Net network that could be used for automatic segmentation of the liver, and they extended the architecture of the U-Net to replace all 2D operations by corresponding 3D operations.Also, they labeled only some of the slices in the volume to be segmented to train the model and used it for the segmentation task of the whole volume.Fang et al [6] extracted a U-Net network with multi-scale information combination for liver image segmentation, they used a pyramid structure for feature extraction at different scales, performed equal depth convolution (EDC) on the pyramid feature extraction network to analyze the output, and later combined the analyzed output with multi-scale input information to mitigate the semantic divide of U-Net skip connections.Christ et al [7] proposed a cascaded fully convolutional neural network (CFCN) for automatic segmentation of liver CT and MRI, which uses two FCNs for liver and tumor segmentation, respectively, with tumor training based on liver segmentation,

and this approach was experimentally shown to increase the effectiveness of tumor segmentation.

Skip connections allow the combination of low-level semantic information from the encoding path and high-level semantics from the decoding path, and many models achieve multi-scale feature fusion in this way.Researchers often use multi-level CNNs to cascade the extracted features when the task region appears to be widely different in shape and size.Although these variants including the encoder-decoder structure have the most advanced performance, this approach will lead to an increase in the amount of calculation and redundancy of model parameters.For example, most of the variants with U-Net based on a single feature layer will repeatedly extract semantic information at similar scales. As the network structure deepens, the training difficulty of the model increases. To address these problems, we introduce the bottleneck Res2Net[8] to enhance the multiscale capability of U-Net and reduce the number of parameters of the model, and use the The SE[9] module is used to enhance the network learning capability. In addition, We also design a multi-scale attentional gate structure (MAG) to increase the weight of target regions while suppressing background regions.

The contributions of this study are as follows:

- We propose a MSAA-Net based on U-Net, which combines the attention mechanism and a more granular level multi-scale feature fusion.
- We design an attention gate (MAG) combining multi-scale feature fusion, spatial attention mechanism, and channel attention mechanism to optimize skip connection. MAG can increase the weight of target regions while suppressing irrelevant background regions.MAG can theoretically be used in other structures with skip connections.
- We apply the SE-block to all modules, which improves the learning capability of the network.
- We conducted extensive comparison experiments on the 3Dircadb dataset, and the results show that MSAA-Net achieves the best performance on the segmentation task of the liver with a much reduced number of parameters compared to similar attention networks.

2 Related Work

In this section, we review the results of fully automated segmentation of medical images based on fully convolutional neural networks in recent years.

With the continuous updating and development of hardware technology, deep learning algorithms have flooded into the medical imaging field, and many excellent algorithms based on deep Many excellent algorithms based on deep learning have been proposed and continuously improved.In the past few years, deep learning algorithms based on convolutional neural networks (CNNs) have been widely used in image processing due to their powerful nonlinear feature extraction and data processing capabilities.Alex et al. [10] designed and trained a large deep convolutional neural network, Alexnet, which won the first place in the ILSVRC-2012 competition, scoring much higher than the second place, which created a great impact to the academic

community at that time.Karen et al. [11] explored the relationship between depth and performance of convolutional neural networks and successfully constructed a 16- to 19-layer convolutional network (VGGNet) by stacking small convolutional kernels and maximum pooling operations several times.They found that the effect of two convolutional superpositions of size 3 is equivalent to a convolutional operation of size 5, and the effect of three convolutional superpositions of size 3 is equivalent to a convolutional operation of size 7.VGGNet is a landmark innovation in image processing and is still used by many advanced models to extract image features until now.GoogLeNet [12] was proposed in 2014 by Szegedy et al. They argued that improving the performance of the network requires increasing the depth and width of the network, but this results in large computational effort and overfitting problems. To solve these problems, GoogLeNet transforms convolution and full connectivity into sparse connectivity and improves the computational performance with dense matrices.Long et al. [4] defined a skip structured network (FCN), which combines deep and shallow semantic information to produce accurate and detailed segmentation. This structure implements end-to-end segmentation, where an input image of a certain size will result in a segmentation map of the same size. Ronneberger et al .[3] proposed an FCN-based U-Net architecture for encoder-to-decoder mapping via skip connections between feature maps of the same size. This network structure is the most classical in the field of biomedical images and many excellent models use it as a backbone network.Milletari et al.[13] proposed V-Net by extending the U-Net structure to 3D volumes, and they used random nonlinear variations and histogram matching to extend the data.The experiments demonstrate the good performance of V-Net in medical image processing and greatly reduce the computing time.Wang et al. [14] proposed the No-local U-Net, which reduces the downsampling multiplier of the U-Net, changes the skip connection, and reduces the model parameters. In addition, They embedded a global aggregation module at the bottom of the U-Net to achieve global information fusion of advanced features.With the continuous development of image segmentation models, researchers have found that for the problem of different attention to important and non-important regions of an image, adding attention mechanisms to the model of semantic segmentation can improve the network performance very well.Oktay et al. [15] proposed Attention U-Net by combining the attention mechanism with U-Net for the first time and applied it to the segmentation task of the pancreas.Attention U-Net adds an attention gate (AG) at each skip connection of the U-Net network. The AG is the union of the next layer of the skip connection (a higher level feature layer) and the encoder feature layer of the current layer, and after partial mathematical operation makes these two layers interact to generate a weight matrix, which is multiplied by a single pixel of the current feature layer before the skip connection, and the individual pixel weights of each feature layer are trained after back propagation, thus achieving different weights for different regions of the image.Zhou et al [16] proposed a model UNet++ based on the structure of U-Net integrated according to different depths of U-Net.UNet++ can learn the best depth for the current task through u-net networks with different depths.They improved skip connectivity by interconnecting multiple feature layers at the same scale, which coincides with the dense connectivity structure.

Although attentional mechanisms have been widely used in medical image processing, few models combine different kinds of attention mechanisms.U-Net uses skip connections to fuse multi-scale information, but this structure has the disadvantage of generating semantic conflicts.In addition, researchers have designed models with progressively more complex structures and larger numbers of parameters in order to pursue segmentation accuracy, while work on optimizing and streamlining network models is less common.

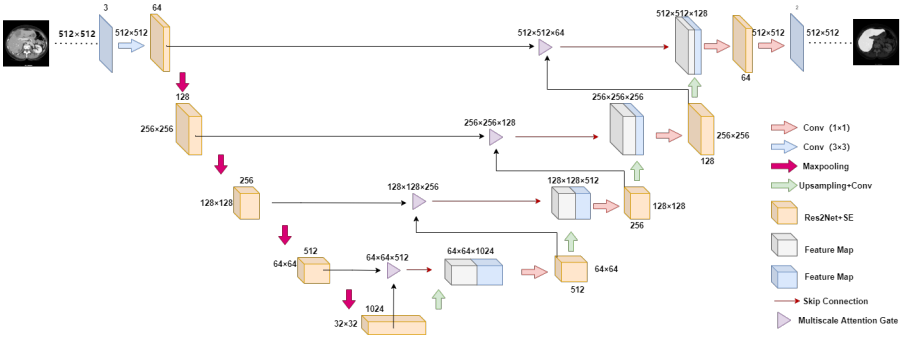
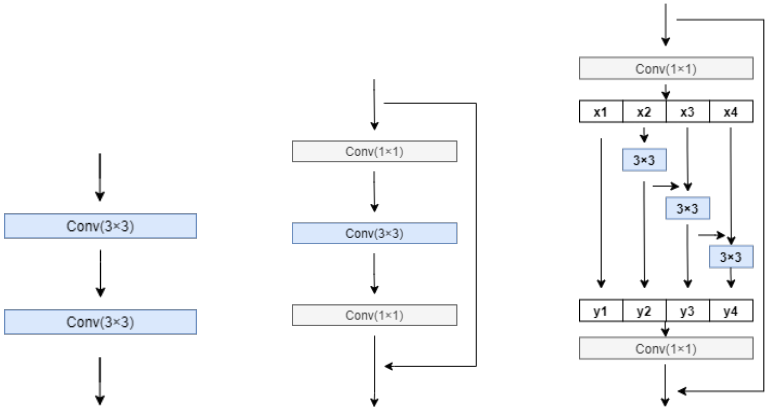


Fig. 1 The proposed MSAA-Net for fine segmentation of the liver. The input is a 512 × 512 three-channel image, and the number of channels is converted to 64 by 3 × 3 convolution, which is output to the Res2Net+SE module, and the final segmentation result map is output by the encoder and decoder.It is important to note that all feature extraction is done by the Res2Net+SE module and the skip connections are optimized using MAG.



(a) A feature layer (b) Bottleneck block (c) Res2Net module

Fig. 2 (a) is a feature extraction layer of the conventional U-Net, consisting of two 3 × 3 convolutions followed by batch normalization (BN) and $RELU$ activation. (b) and(c)can be used to replace (a).

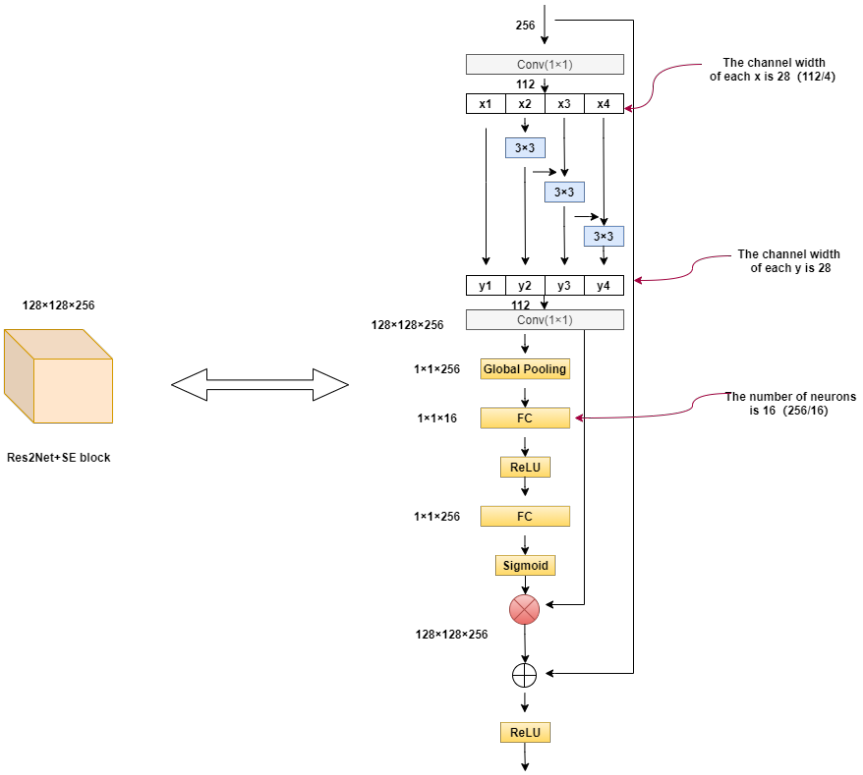


Fig. 3 Details of the MSAA-Net encoder layer 3 are shown. After two feature extractions, the size of the feature map is 128×128 and the number of channels is 256. The Res2Net module obtains multi-scale information and then recalibrates the channels through the SE module to output feature maps with the same specifications.

3 Methodology

This chapter describes in detail the architecture of MSAA-Net.The MSAA-Net is based on the U-Net backbone architecture using the Res2Net bottleneck module and the SE module for multi-scale information fusion and feature recalibration of individual feature layers.This structure improves the learning efficiency of the network and reduces the model parameters.In addition, we improve skip connections using MAG, which integrates the attention mechanism in the channel dimension and the attention mechanism in the spatial dimension.

3.1 MSAA-Net

The U-shaped architecture has been widely used in medical image processing tasks, and its skip-connected structure, which incorporates shallow information and high-level features, has good stability.The structure of MSAA-Net is shown in Figure 1.MSAA-Net can be divided into encoding stage and decoding stage.In the encoder, MSAA-Net uses five feature layers to extract image feature information, but unlike

U-Net where two 3×3 convolutions are repeatedly applied to each layer, MSAA-Net uses the bottleneck structure of Res2Net and the SE module to extract the semantic information of each layer as shown in Figure 3.When one layer of feature extraction is completed, MSAA-Net will use max-pooling with a step size of 2 to compress the feature map and increase the number of channels with a 1 × 1 convolution, and the next layer will continue to extract semantic information in the same way.Similar to the encoder, the decoder uses the same Res2Net+SE feature extraction approach, but the MSAA-Net is optimized using multiscale attention gates in the skip connections.MAG can effectively mitigate semantic conflicts in skip connections, allowing the network to focus on regions of interest and suppress irrelevant background regions.MSAA-Net acquires a 512×512 three-channel image as input and outputs a segmented image of the same size.With the SE module and MAG, MSAA-Net has a more accurate segmentation effect.

Table 1 shows the detailed structure of the proposed MSAA-Net, which includes the layering of the network, the structure of each layer, the activation function, the size of the convolution kernel, the size of the feature map, and the number of channels.

3.2 Res2Net

The bottleneck structure shown in Figure 2(b) is the underlying structure of many advanced network models[17].Res2Net improves on the bottleneck module, replacing a set of 3×3 filters with multiple sets of smaller 3×3 filters and a connection similar to the residual learning framework is made. This allows Res2Net to retain the similar functionality of the bottleneck module while gaining enhanced multi-scale feature fusion capabilities.

Figure 2(c) shows the detailed structure of the Res2Net module. After the 1×1 convolution operation, the Res2Net module divides equally the obtained feature map by number of channels,each partitioned subset is denoted by x_i,$i \in \{1, 2, ..., s\}$,s denotes the number of blocks. Each x_i has the same size and the same number of channels. All x_i except x_1 need to go through different 3×3 convolution operations. Using K_i to denote the i-th set of convolution operations, and y_i denotes the output of the corresponding x_i. $y_i\ (i > 2)$ is calculated from $k_i(x_i + y_{(i-1)})$.y_i and the output out can be expressed as:

$$y_i = \begin{cases} x_i & i = 1 \\ K_i\left(x_i\right) & i = 2 \\ K_i\left(x_i + y_{i-1}\right) & 2 < i \leq s \end{cases} \tag{1}$$

$$out = \sum_{i=1}^{s} y_i \tag{2}$$

Here $\sum$ represents the cumulative stitching operation of all chunks for the channel dimension.

Note that the feature maps processed by $K_{(i+1)}()$ are obtained by stitching $K_i(x_i)$ and $x_{(i+1)}$, so they have a larger sense of receptive field, the output result

Table 1 Details of operations performed, settings of layers in each encoding and decoding stage of the proposed network.

Stages	Encoder Path	Output Shape	Decoder Path	Output Shape
	Input	512×512 $\times 3$	Conv2d (1×1)	512×512 $\times 2$
1	$[Conv2d$ $(3 \times 3)+$ $BN + ReLU]\times 2$ Res2Net[64, scaling=4] $+BN + ReLU$ $SE[$ $avgpool(64)$ $+Linear(64,$ $4) + ReLU$ $+Linear(4, 64)$ $+Sigmoid]$	512×512 $\times 64$	$[Conv2d$ $(1 \times 1)+$ $BN + ReLU]\times 2$ Res2Net[64, scaling=4] $+BN + ReLU$ $SE[$ $avgpool(64)$ $+Linear(64,$ $4) + ReLU$ $+Linear(4, 64)$ $+Sigmoid]$	512×512 $\times 64$
2	$Conv2d$ (1×1) Res2Net[128, scaling=4] $+BN + ReLU$ $SE[$ $avgpool(128)$ $+Linear(128,$ $8) + ReLU$ $+Linear(8, 128)$ $+Sigmoid]$	256×256 $\times 128$	$Conv2d$ (1×1) Res2Net[128, scaling=4] $+BN + ReLU$ $SE[$ $avgpool(128)$ $+Linear(128,$ $8) + ReLU$ $+Linear(8, 128)$ $+Sigmoid]$	256×256 $\times 128$
3	$Conv2d$ (1×1) Res2Net[256, scaling=4] $+BN + ReLU$ $SE[$ $avgpool(256)$ $+Linear(256,$ $16) + ReLU$ $+Linear(16, 256)$ $+Sigmoid]$	128×128 $\times 256$	$Conv2d$ (1×1) Res2Net[256, scaling=4] $+BN + ReLU$ $SE[$ $avgpool(256)$ $+Linear(256,$ $16) + ReLU$ $+Linear(16, 256)$ $+Sigmoid]$	128×128 $\times 256$

out contains scale information of different sizes. This splitting and then multi-scale fusion process is beneficial to extract global and local information.

Stages	Encoder Path	Output Shape	Decoder Path	Output Shape
4	$Conv2d$ (1×1) Res2Net[512, scaling=4] $+BN + ReLU$ $SE[$ $avgpool(512)$ $+Linear(512,$ $32) + ReLU$ $+Linear(32, 512)$ $+Sigmoid]$	64×64 $\times 512$	$Conv2d$ (1×1) Res2Net[512, scaling=4] $+BN + ReLU$ $SE[$ $avgpool(512)$ $+Linear(512,$ $32) + ReLU$ $+Linear(32, 512)$ $+Sigmoid]$	64×64 $\times 512$
5	$Conv2d$ (1×1) Res2Net[1024, scaling=4] $+BN + ReLU$ $SE[$ $avgpool(1024)$ $+Linear(1024,$ $64) + ReLU$ $+Linear(64, 1024)$ $+Sigmoid]$	32×32 $\times 1024$		

3.3 Squeeze-and-Excitation Blocks

Squeeze-and-Excitation (SE) Blocks were proposed by Hu et al [9] and are referred to as channel attention mechanisms in some papers [18][19][20].The SE module recalibrates the channel response by displaying the interdependencies between the modeled channels. In brief, the SE module can update the importance of each channel by learning, strengthening the weights of channels that are of interest to the task and suppressing channels that are less relevant to the task. We introduce the SE Blocks into all modules, which will cost a small amount of computation to improve the segmentation accuracy. In addition, we try to integrate it with the spatial attention mechanism to produce a more efficient gating device for optimizing skip connections. Figure 3 shows the structure of the Res2Net and SE Block, which is inspired by Ms-UNet [21].

3.4 MAG

Attentional mechanisms originated from the study of vision in humans.In medical image segmentation tasks, attention mechanisms can enhance the weights of segmentation locations to suppress task-irrelevant background regions. We designed

the MAG by combining the channel attention mechanism and the spatial attention mechanism. The proposed structure is shown in Figure 4. MAG processes the feature map x_1 used for the skip connections to obtain the optimized feature map $x_{output2}$,$x_{output2}$ with the decoder part of the feature map for channel splicing, and then downscaled into Res2Net+SE module by 1×1 convolution to complete multi-scale information fusion[22][23][24]. Since only the features are scaled, MAG spends a small amount of computation to significantly optimize the skip connections.

A feature map $x_1 \in R^{H\times W\times C}$ with an intermediate layer and a feature map $x_2 \in R^{\frac{H}{2}\times\frac{W}{2}\times 2C}$ containing higher level semantic information are given as inputs. High-level features x_2 resolution is smaller and needs to be up-sampled before subsequent operations can be performed.x_2 is up-sampled to get $x_2^{'} \in R^{H\times W\times 2C}$, $x_2^{'}$ and x_1 have the same size, and the number of channels of x_1 and $x_2^{'}$ are unified by 1×1 convolution, and then summed to $g_1 \in R^{H\times W\times Fint}$ after batch normalization (BN). x_1 obtains g_1 after collecting semantic information on a coarser scale. After the activation of g_1 by the nonlinear activation function ($ReLU$), and then the 1×1 convolutional compression channel, we get $\sigma_1 \in R^{H\times W\times 1}$. σ_1 is activated by $Sigmoid$ to find the 2D spatial attention coefficient α_1. The essence of α_1 is a weight matrix that It has the same size as x_1. Therefore, $x_1 \bigotimes \alpha_1$ achieves spatial attention discovery and outputs $x_{output1}$.

Unlike the structure of Su et al [20], the MAG channel attention was found to be generated by the extrusion and activation of x_1, which is a more concise structure. Since $x_{output1}$ is optimized for x_1 spatial weights, optimizing channels on the basis of $x_{output1}$ only requires squeezing and activation of x_1. As shown in Figure 4, the average global pooling is used for x_1 to obtain $x_1^{'} \in R^{1\times 1\times C}$, $x_1^{'}$ by two fully connected layers to achieve auto-calibrationto obtain $\sigma_2 \in R^{1\times 1\times C}$.$\sigma_2$. σ_2 is activated by $Sigmoid$ to obtain the $1D$ channel attention factor α_2 The essence of α_2 is a weight vector, which is equal to the channel of x_1. $x_{output1} \bigotimes \alpha_2$ implements the channel attention discovery and outputs $x_{output2}$.

Figure 5 shows the mathematical process of MAG, W is the convolution kernel weight,S denotes the Sigmoid function, b represents the bias term, and the whole attention process can be summarized as as:

$$x_2^{'} = F_{up}\left(x_2\right) \tag{3}$$

$$g_1 = F_{sum}\left(W_1(x_1) + b_1, W_2(x_2^{'}) + b_2\right) \tag{4}$$

$$g_1^{'} = Relu\left(g_1\right) \tag{5}$$

$$\sigma_1 = W_3(g_1^{'}) + b_3 \tag{6}$$

$$\alpha_1(x_1, x_2) = S(\sigma_1) \tag{7}$$

The spatial attention factor α_1 can be found by combining (3)(4)(5)(6)(7):

$$\alpha_1(x_1, x_2) = S(W_3(Relu(F_{sum}(W_1(x_1) + b_1, W_2(F_{up}(x_2)) + b_2))) + b_3) \tag{8}$$

$$x_1^{'} = F_{pl}\left(x_1\right) \tag{9}$$

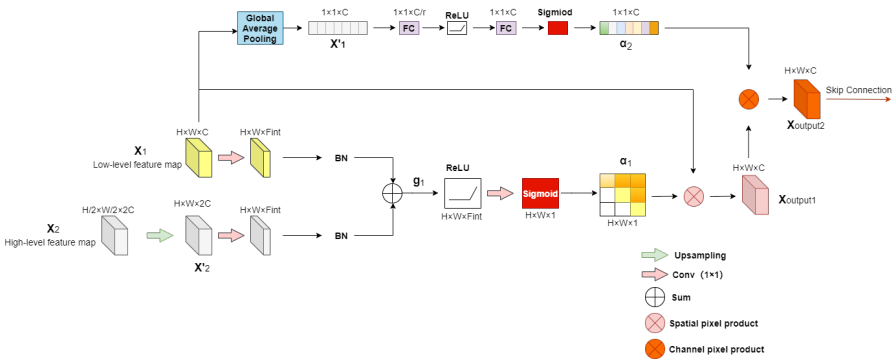


Fig. 4 Schematic diagram of the proposed MAG. the MAG receives input from the encoder part and the decoder part, respectively. x_1 is the low-level feature map used to complete the jump connection, and x_2 has higher-level semantic information. x_1 generates attention coefficients α_1 from the contextual information of x_2, x_2 is the signal collected from the larger sensory field,α_1 performs the first scaling of the spatial dimension of x_1 to obtain $x_{output1}$. x_1 performs a squeeze and excitation operation on its own channel to obtain the attention factor α_2, and α_2 performs a second deflation of the channel dimension on x_1 to obtain $x_{output1}$. $x_{output1}$ is the output of MAG.

$$g_2 = W_{fc1}(x_1^{'}) + b_{fc1} \tag{10}$$

$$g_2^{'} = Relu\,(g_2) \tag{11}$$

$$\sigma_2 = W_{fc2}(g_2^{'}) + b_{fc2} \tag{12}$$

$$\alpha_2(x_1) = S(\sigma_2) \tag{13}$$

The channel attention factor α_2 can be found by combining (9) (10) (11) (12) (13):

$$\alpha_2(x_1) = S(W_{fc2}(Relu(W_{fc1}(F_{pl}(x_1)) + b_{fc1})) + b_{fc2}) \tag{14}$$

The process of the attention factor acting on the feature map can be expressed as:

$$x_{output1} = \alpha_1(x_1, x_2) \bigotimes x_1 \tag{15}$$

$$x_{output2} = \alpha_2(x_1) \bigotimes x_{output1} = \alpha_2(x_1) \bigotimes \left[\alpha_1(x_1, x_2) \bigotimes x_1\right] \tag{16}$$

4 Results

4.1 Experimental Environment

The 9G GeForce RTX 3080 is used for training and prediction.The code was implemented on a windows 10 system with pycharm and jupyter compilers and the model was implemented based on pytorch [25].

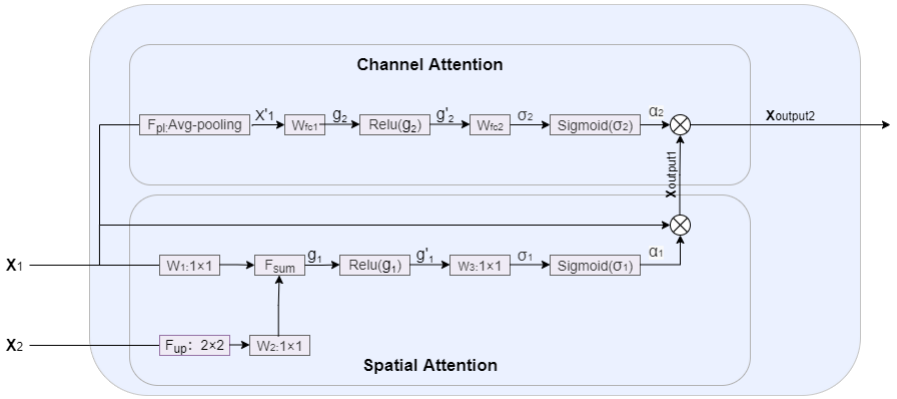


Fig. 5 The mathematical process of MAG, which omits the feature map size and BN, etc. W_i denotes the convolution operation and W_{fci} denotes the fully connected operation.F_{pl},F_{up},F_{sum} represent pooling, upsampling, and pixel summation operations, respectively.

Model	Dice(%)	MIoU(%)	MPA(%)	Params
U-Net[3]	97.58	95.28	97.82	94.97M
Res2+Unet	96.94	94.07	96.59	26.6M
Attention+Res2+Unet	97.17	94.5	97.04	51.2M
Res2+SE+Attention+Unet(ours)	96.72	93.66	97.01	52.08M
AttentionUnet[15]	97.58	95.29	97.67	133.19M
MSAA-Net(ours)	98.29	96.64	98.69	52.25M

Table 2 Evaluation metrics for ablation analysis of our method using test datasets

4.2 Dataset And Data preprocessing

We use the publicly available 3D Image Reconstruction for Comparison of Algorithm Database 3Dircadb[26] for training and testing of the model. The 3Dircadb database consists of CT scans of ten female and ten male patients with liver tumors. The number of CT scan slices for each patient ranged from 74 to 260, respectively, and were stored in DICOM format. The database is divided into 20 folders, each containing labels for locations of interest that have been manually segmented by experts, e.g., liver, liver tumor, portalvein, rightkidney, etc. The images provided by the database exhibit segmentation difficulties such as liver contact with neighboring organs, atypical shape and density, and artifacts, which pose a challenge to the segmentation task.

In CT scanning, Hounsfield Units (HU) are often used to express CT values in a standard and convenient way. The range of HU values is from -1000 to 1000. In the task of liver parenchyma segmentation, it is necessary to set the corresponding HU values to focus the image on the liver region and remove the irrelevant areas. We preprocessed all images of the entire dataset. First, the DICOM formatted image is were converted to 512×512 png images to be used as input for the network. Secondly, we windowed all the images, using (400,50) HU value window to highlight the task area to make the segmented area clean. Finally, we histogram equalized all 2803

original and labeled images each, which solves the problem of overall darkness in the dataset. These preprocessing operations provide good training data and labels for liver parenchyma segmentation tasks, and the preprocessing effect is shown in Figure 6.

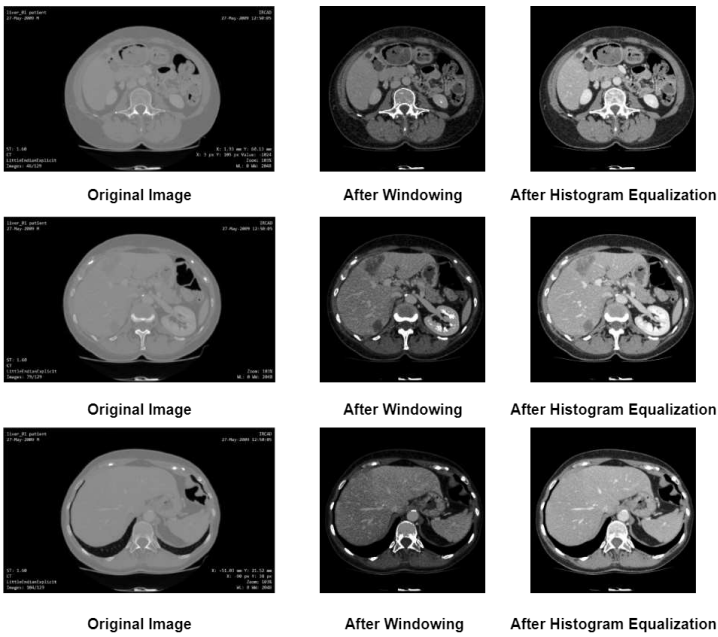


Fig. 6 The pre-processing process of the input image, the first column is the original sample image, the second column indicates the effect after HU Windowing, and the third column is the display after histogram equalization.

4.3 Loss Function

We use a loss function consisting of a weighted cross-entropy loss and a dice loss [27][13] to calculate the gradient. The cross-entropy loss function has been widely used in classification tasks and has its unique advantages. The segmentation task of liver substance is a high-precision binary classification problem; therefore, the cross-entropy loss function should be used as the preferred loss function, which is defined as follows:

$$L_B = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log(p_i) + (1-y_i)\log(1-p_i)\right) \tag{17}$$

where y denotes the ground truth value of the image pixel and p denotes the value of the predicted label.

Semantic segmentation is essentially a pixel-level classification problem, and the cross-entropy loss function can be applied to most scenarios, but it has a drawback

that when the number of foreground is much smaller than the number of background will make the model heavily biased towards the background, affecting the segmentation effect. There are many cases in the liver parenchyma segmentation task where the background is much larger than the segmented region, so we add dice loss and add weights to the cross-entropy loss.The dice loss function is defined as follows:

$$L_{Dice} = 1 - \frac{2\sum_i^N p_i g_i + \varphi}{\sum_i^N p_i^2 + \sum_i^N g_i^2 + \varphi} \tag{18}$$

where p and g denote the predicted binary partition volume and the predicted binary volume of the label, respectively, and φ takes the value $1e-5$. The final loss function is defined as follows:

$$L = w_1 L_B + w_2 L_{Dice} \tag{19}$$

where w_1, w_2 $(0 < w_i \leq 1)$are the weight coefficients that are used to balance the values of the cross-entropy loss function and the dice loss function. In our segmentation task, the best performance is obtained for $w_1 = 0.5, w_2 = 0.5$.

4.4 Evaluation Measures

Mean Pixel Accuracy (MPA), Mean Intersection over Union ($MIoU$) and dice coefficient (DC) are used as evaluation metrics. These metrics are calculated as follows:

$$PA = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

$$MPA = \frac{1}{N} Sum(PA) \tag{21}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{22}$$

$$MIoU = \frac{1}{N} Sum(IoU) \tag{23}$$

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{24}$$

where TP (true positive) indicates the number of foreground pixels correctly classified as foreground (liver), TN (true negative) indicates the number of background pixels correctly classified as background (non-liver region), FP (false positive) indicates the number of background pixels incorrectly identified as foreground, and FN (false negative) indicates the number of foreground pixels incorrectly classified as background. N is the total number of categories, in this article $N = 2$. Sum means the summation operation of the corresponding index scores for all categories.

MPA is the ratio of the average number of correctly classified pixels per class to the total number of pixels and is used to indicate the segmentation accuracy of the task. $MIoU$ reflects the average proportion of common elements at matching positions to the segmentation result, and all inaccurate segmentations reduce the $MIoU$ score. $Dice$ calculates the similarity between the predicted and true results and is used to evaluate the model performance.

4.5 Experimental results and analysis

In this section, we use ablation analysis to determine the effectiveness of MSAA-Net in liver segmentation. We used 128 different CT images as the test set and calculated the average values of $Dice$, IoU and PA respectively. These metrics are commonly used for medical image segmentation. In addition, we also record the number of parameters of the model and the convergence speed of the network, which can help us to evaluate the model more comprehensively.

The backbone architecture of our model is based on the U-Net, which is therefore used as the benchmark for evaluation. We gradually added different modules to U-Net and optimize them, and select the most efficient structure for comparative analysis, and the segmentation results are shown in Table 2. The training environment of all models is the same. As can be seen from Table 2, U-Net and Res2-block greatly reduce the parameter usage at the expense of a small amount of performance. Probably due to the large target region of Attention U-Net segmentation, the simple addition of the Attention Gate structure does not bring about any improvement in model performance, instead, the Attention Gate structure uses more parameters. After adding Attention Gate and SE-block, Res2-block and U-Net obtain significant performance improvement. Combined with the experimental results of Attention U-net, we attribute this performance improvement to SE-block. Finally, we compare the segmentation results of MSAA-Net with other networks, and MSAA-Net obtains the best segmentation accuracy.Obviously, these performance improvements are due to the improved multiscale attention gate structure (MAG). The Res2-block structure in MSAA-Net helps us to greatly reduce the number of parameters, which is only 38.4% of that of Attention U-Net and 55% of that of U-Net.

To evaluate the training speed of MSAA-Net, we record the loss value every 10 epochs for the different models mentioned above. The loss values are shown in Figure 8. Compared with other models, the loss value of MSAA-Net still decreases after 60 epochs. This proves that Res2-block has the effect of improving gradient disappearance, so we can iterate more times to train MSAA-Net. After combining computational overhead and accuracy, we use MSAA-Net after the 100th epochs of training to perform liver segmentation. After combining computational overhead and accuracy, we use the trained MSAA-Net for liver segmentation.

We plotted Box-plot for multiple models, where the data represent the Dice coefficients for each model.As shown in Figure 9, the Box-plot contains the statistical results of multiple experiments.The results show that MSAA-Net possesses stability and does not show any outliers.In addition, MSAA-Net shows a great advantage in both data float and median. This is a good proof of the migrability of MAG.

4.6 Comparison With Other Methods

In this section, we compare MSAA-Net with other classical semantic segmentation architectures, including FCN, UNet, Attention U-Net, and Ms-Unet. The source code of Ms-Net is not available, but the structure is similar to that of MSAA-Net and can be used as a reference. Since our input image is a 512×512 size CT image from 3Dircadb, the number of parameters of the model will be larger and the segmentation

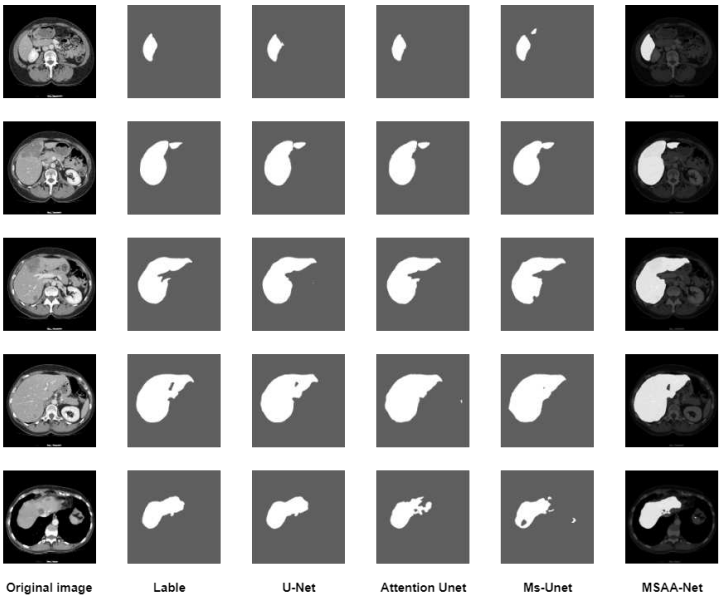


Fig. 7 Comparison of segmentation results of MSAA-Net with other networks. We combine the segmentation results of MSAA-Net with the original image, so that the segmentation effect is more intuitive.

Model	Dice(%)	MIoU(%)	MPA(%)	Params
U-Net[3]	97.58	95.28	97.82	94.97M
Ms-Unet[21]	96.71	95.13	97.64	52.08M
AttentionUnet[15]	97.58	95.29	97.67	133.19M
FCN[4]	91.68	84.64	89.85	512.27M
MSAA-Net(ours)	98.29	96.64	98.69	52.25M

Table 3 Comparison results of other methods

will be better, which does not affect the comparison of the model. As shown in Table 3, MSAA-Net obtained a better segmentation performance, and its number of parameters is much smaller than other models. Figure 7 shows the segmentation results of the liver of MSAA-Net compared with other models. The segmentation results are compared with other models, and it can be seen visually that the segmentation results of MSAA-Net are better than those of the previous models.

5 Discussion

Liver image segmentation is a typical medical image segmentation task, and this work can be used as a pre-processing for treating liver diseases. Adjacent organs, blood vessels, and tumors can affect the segmentation accuracy of the liver. In this paper, we introduce some advanced structures into the U-Net network to improve the segmentation effect and reduce the model parameters. In addition, we also design a

multiscale attention gate structure to optimize optimize skip connections and alleviate the semantic gaps between encoder and decoder. With a new attentional gate (MAG) structure and some optimization strategies, we improve the Dice scores of liver image segmentation. MAG combines the spatial attention mechanism and the channel attention mechanism, and weights the feature maps of the encoder separately in different dimensions, which makes the task-relevant regions and relevant channels have increased weights and are more easily learned and updated. To validate the results of our work, we performed ablation experiments on the 3Dircadb dataset, continuously adding and modifying the structure of the U-Net and analyzing the role of different structures for the task. In order to reduce the cost of the work, our test set was selected as 10% of the total training set, which will result in generally high scores for the evaluation metrics, but will not affect the results of the comparison experiments. Ultimately, we conclude that: Res2-block can replace the traditional convolutional block greatly reducing the model parameters, SE-block can be optimized for all encoder and decoder feature maps, and MAG can be optimized for skip connections (which may be more effective in segmentation tasks with small target regions). The evaluation results are shown in Tables 2 and 3, where MSAA-Net received the highest Dice score of 98.29%, with half the parameters of U-Net. In addition, we record the loss values every 10 epochs for each model and make a line graph as shown in Figure 8. It is obvious that MSAA-Net has a larger initial gradient, and its gradient descent period is longer. After 60 epochs, the loss of MSAA-Net is still decreasing, which indicates that the training can be better with enough computing power.

MAG can be used for other structures containing skip connections, and we speculate that it will perform better in tasks with large volume variations, but due to space limitations, we were unable to add MAG to other good models and evaluate its effectiveness, nor did we evaluate the performance of MSAA-Net in more medical image segmentation tasks. We want MAG to be used for more segmentation tasks.

6 Conclusion

In this paper, we propose an improved U-Net model MSAA-Net for automatic segmentation of the liver. We replace the original feature extraction module with a combination of Res2-block and SE-block, and improve the skip connection with the proposed MAG. MSAA-Net obtains the best evaluation on the 3Dircadb dataset. We also propose the scalability of MAG, the combination of spatial The combination of spatial attention and channel attention is also applicable to other models.

References

[1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **68**(6), 394–424 (2018)

[2] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning

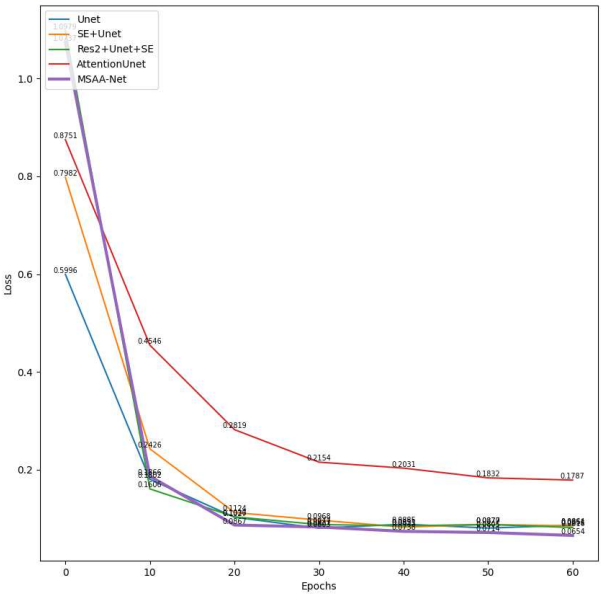


Fig. 8 Loss values per ten epochs for different models

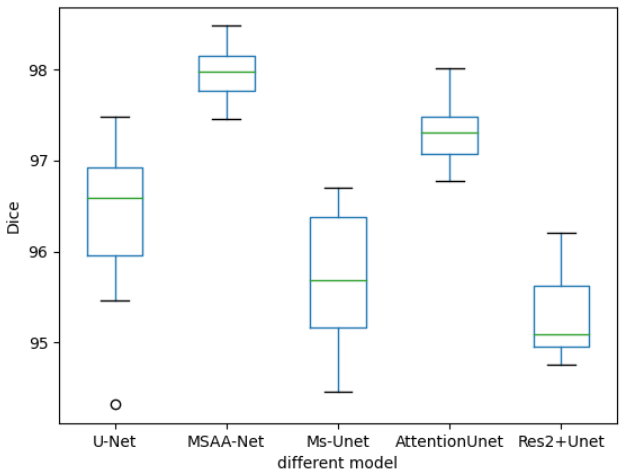


Fig. 9 The Box-plot of the different models

in medical image analysis. Medical image analysis **42**, 60–88 (2017)

[3] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). Springer

[4] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)

[5] Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 424–432 (2016). Springer

[6] Fang, X., Yan, P.: Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. IEEE Transactions on Medical Imaging **39**(11), 3619–3629 (2020)

[7] Christ, P.F., Ettlinger, F., Grün, F., Elshaera, M.E.A., Lipkova, J., Schlecht, S., Ahmaddy, F., Tatavarty, S., Bickel, M., Bilic, P., et al.: Automatic liver and tumor segmentation of ct and mri volumes using cascaded fully convolutional neural networks. arXiv preprint arXiv:1702.05970 (2017)

[8] Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., Torr, P.: Res2net: A new multi-scale backbone architecture. IEEE transactions on pattern analysis and machine intelligence **43**(2), 652–662 (2019)

[9] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

[10] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)

[11] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[12] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. proceedings of the ieee computer society conference on computer vision and pattern recognition (2015)

[13] Milletari, F., Navab, N., Ahmadi, S.-A.: V-net: Fully convolutional neural

networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571 (2016). IEEE

[14] Wang, Z., Zou, N., Shen, D., Ji, S.: Non-local u-nets for biomedical image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 6315–6322 (2020)

[15] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)

[16] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging **39**(6), 1856–1867 (2019)

[17] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[18] Qin, Z., Zhang, P., Wu, F., Li, X.: Fcanet: Frequency channel attention networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 783–792 (2021)

[19] Sun, H., Zeng, X., Xu, T., Peng, G., Ma, Y.: Computer-aided diagnosis in histopathological images of the endometrium using a convolutional neural network and attention mechanisms. IEEE Journal of Biomedical and Health Informatics **24**(6), 1664–1676 (2019)

[20] Su, R., Liu, J., Zhang, D., Cheng, C., Ye, M.: Multimodal glioma image segmentation using dual encoder structure and channel spatial attention block. Frontiers in Neuroscience, 1063 (2020)

[21] Kushnure, D.T., Talbar, S.N.: Ms-unet: A multi-scale unet with feature recalibration approach for automatic liver and tumor segmentation in ct images. Computerized Medical Imaging and Graphics **89**, 101885 (2021)

[22] Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)

[23] Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9413–9422 (2020)

[24] Ni, J., Wu, J., Tong, J., Chen, Z., Zhao, J.: Gc-net: Global context network for medical image segmentation. Computer methods and programs in biomedicine

190, 105121 (2020)

[25] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

[26] Soler, L., Hostettler, A., Agnus, V., Charnoz, A., Fasquel, J., Moreau, J., Osswald, A., Bouhadjar, M., Marescaux, J.: 3d image reconstruction for comparison of algorithm database: A patient specific anatomical and medical image database. IRCAD, Strasbourg, France, Tech. Rep (2010)

[27] Crum, W.R., Camara, O., Hill, D.L.: Generalized overlap measures for evaluation and validation in medical image analysis. IEEE transactions on medical imaging **25**(11), 1451–1461 (2006)