Multi-scale Attention Module U-Net liver tumour segmentation method

Fei He1, Gaojian Zhang1, Huamin Yang1* and Zhengang Jiang2*

1 School of Computer Science and Technology, Changchun University of Science and Technology, Changchun, Jilin, 130022, China
2 Corresponding author’s e-mail: yhm@cust.edu.cn
3 Corresponding author’s e-mail: jiangzhengang@cust.edu.cn

Abstract. Liver tumour is a general term for tumour lesions occurring in the liver, which is the third leading cause of tumour death in the world. Most liver tumours are small in size, variable in shape and unfixed in position, and the existing neural network is not fully utilized in the segmentation of liver tumours. In order to make up for the problem of traditional tumour segmentation, improve the precision and effect of tumour treatment. The multi-scale Attention Module U-NET (MAM U-NET) was proposed by analyzing CT image data of liver tumours. The MAM U-NET model consists of a newly designed Combination module and an MAM attention module. Each MAM attention module is composed of two parts: spatial attention model and channel attention model, which enable the model to learn more extensive and rich context-dependent information and spatial information. Finally, a comparison experiment was conducted in the LITS data set to verify the performance of the model. Liver segmentation achieved the optimal result, and the average Dice coefficient reached 96.8282%. The segmentation results of liver tumours were second only to nnU-Net, and the average Dice coefficient was 60.5991%, which proved that the MAM U-NET model had better performance in the segmentation of liver tumours.

1. Introduction
For liver tumours, there are large differences in the size, shape, and location of the tumour in each patient. In addition, most of the tumours have similar contrast to the surrounding normal liver tissue and blurred borders, which determines that three-dimensional liver tumour segmentation remains a challenging task. With the improvement of computer hardware, deep convolutional neural networks (CNNs) have been widely used in computer vision tasks such as segmentation, target detection and classification. Even CNNs have been shown to be highly robust to changing image appearance, and Chlebus et al.[1] proposed a liver tumour segmentation algorithm based on deep learning using a shape-constrained model for post-processing. Sun et al.[2] used CT images of different time series simultaneously as input to the network to increase the information of the tumour. Li et al.[3] proposed a segmentation method of 2.5D (combining two-dimensional and three-dimensional information), which uses 2D information to guide 3D training. Although it solves the problem of calculation time, the problem of excessive calculation space consumption is solved. Many medical images are 3D images, and Çiçek et al.[4] proposed a direct use of all data from the entire 3D volume to train, realizing that few images are needed to train an appropriate generalization network. Roy et al.[5]
proposed a spatial and channel-based attention mechanism for SE modules and experimentally demonstrated that such modules can enhance meaningful features and suppress useless features.

In the process of training, the Multi-scale Attention Module U-Net (MAM U-Net) model can directly learn the three-dimensional data on the one hand, enhancing the ability of the network model to obtain spatial information; on the other hand, the network model can inhibit the useless features when the weight is updated, fully use the useful features, so that the network model pays more attention to the characteristics of the liver region and effectively affects the shape change of the liver. In order to make full use of the features of each layer in the coding and decoding layers, the combination module is designed and added.

2. Method

2.1. Model architecture

Liver tumour segmentation is one of the important steps in liver tumour detection system. The purpose is to extract liver and liver tumours in the segmentation task of CT images and ensure high sensitivity and accuracy. The MAM U-Net model is composed of coding part and decoding part. Each layer of neural network contains a combination module and MAM module. The maximum pooling with step size of 2 is used for down-sampling, and the deconvolution with step size of 2 is used for up-sampling.

As shown in Figure 1, the square block (a) represents the combination module and the square block (b) represents the MAM module. The network contains the coding layer on the left and the decoding layer on the right. The coding part is used to analyze the whole picture and extract and analyze the features, and the corresponding decoding part is to generate a segmented block diagram. The input image size used in this paper is 16*256*256, and the down-sampling uses a maximum pooling operation of 2*2*2 with a step size of 2. In order to collect the high pixel feature information retained in the feature
analysis, so that the image can be better synthesized, each layer output of the coding part is used as the input of each layer of the decoding part, the data from the coding input and the data sampled on the decoding are cascaded, and then the cascaded data is sent to the combination module. Where the upper sampling uses a trilinear interpolation algorithm.

2.2. Combination block
The function of the combination module is to extract the input image features, and the coarse attention mechanism is added to the feature weight update at the channel level. As shown in Figure 2, the first step is to perform Batch Normalization and ReLU on the input data. In the second step, Data input ReLU activation function. The third step performs a three-dimensional convolution operation of the data with the aim of extracting the local features. In the fourth step, the data after the convolution operation is sent to the SegSE (Segmentation of Squeeze-and-Excitation) for processing. In the fifth step, the SegSE processed data and the module input data are multiplied to realize the role of feature fusion. The sixth step performs the convolution operation to continue extracting the relevant features. As the number of network layers increases, the network model is prone to gradient explosion or disappearance, so a skip connection (Skip) is added under the module. Without affecting the input and output dimension, the input is linearly deformed by adding 1*1*1 convolution to the Skip connection, and the nonlinear expression ability of the network is increased by nonlinear processing through ReLU.

2.3. Segmentation of Squeeze-and-Excitation (SegSE) module
The SegSE module acts to add the channel attention mechanism to the weight update on the channel dimension of the combination module. As shown in Figure 3, the SegSE consists of a three-dimensional cavity convolution, a three-dimensional 1*1*1 convolution and ReLU. This paper proposed an improved SegSE module based on the se module, which replaces the global average pooling operation with cavity convolution due to the lack of spatial correlation in the global average pooling, which can effectively expand the receptive field without increasing the computational load. At the same time, the dimension is compressed to obtain spatial information, which has both contextual information and spatial information.

![Combination block](image)

**Figure 2.** Combination block. Grey squares are input data. The green squares are three-digit convolutions with Kernel size of 3*3*3 and Padding of 1. Yellow squares are SegSE modules. The blue squares are three-digit convolutions with a Kernel size of 1*1*1. The letters above the squares indicate the number of channels.

![SegSE module](image)

**Figure 3.** SegSE module. The grey squares represent three-dimensional convolutions with Kernel size of 3*3*3, Padding of 2, and Dilation of 2. The blue squares represent three-dimensional convolutions with Kernel size of 1*1*1. The letter above the square indicates the number of channels.
This module has a good performance in two-dimensional liver tumour segmentation. According to the characteristics of three-dimensional liver tumour segmentation, this paper improves based on SegSE module, switches to three-dimensional cavity convolution, and increases the number of feature channels to 2 times. The feature channels are subsequently compressed to the original number with a three-dimensional 1*1*1 convolution. Finally, it sends the ReLU activation function to output the result.

2.4. Multi-scale Attention Module

The purpose of the MAM module is to strengthen the features of different channels at multiple scales generated by combination to improve the feature extraction and characterization ability of the network model. As shown in Figure 4, the MAM module is a parallel structure of the upper and lower layers, and the upper layer realizes an attention mechanism for three-dimensional data in the spatial dimension. The following layer implements an attention mechanism for three-dimensional data in the channel dimension. The final results fuse spatial and channel mechanisms of attention.

Figure 4. MAM module. Sea blue squares indicate input data. Four-color squares (green, yellow, light blue, dark blue) represent different weight coefficients on the spatial dimension. The purple, grey, and orange squares represent the weight coefficients on the channels, respectively. The grey squares indicate three-dimensional convolutions with Kernel size of 1*1*1.

In the attention mechanism of spatial dimension, the channel is first compressed into 1 using the convolution of three-dimensional 1*1*1, with the aim of compressing the feature map on the spatial dimension. The Sigmoid activation function is then fed into to increase the nonlinear characteristics of the network. The pixel level product is then performed with each channel of the input data of the module to obtain the result with spatial attention weight.

In the attention mechanism of channel dimension, the global pooling operation is first used to compress the data of each channel into 1, in order to compress the feature map on the channel dimension, reduce the computational and improve the accuracy. The channel is then compressed into half of the input using a convolution of the three-dimensional 1*1*1 for the purpose of fusion of features. Then go to the ReLU activation function. Subsequently, use the convolution of 3D 1*1*1 to restore the channel to the original number of channels, feed it into Sigmoid function to obtain the attention weight coefficient of channel domain, finally product the pixel level with the pixels of different channels at the same position of input data of module, so as to obtain the result with the attention weight of channel.

Finally, the results of spatial and channel dimensions are fused.

3. Experiments and results

3.1. Dataset

The experimental data were manually labeled by 4 radiologists using the public liver tumour dataset LITS (Liver Tumour Segmentation, LITS) provided by CodaLab tissue, and there were several hundred CT image sections for each patient. The training dataset contained 131 CT data, all with a resolution size of 512*512, but with different slice thicknesses. The gold standard corresponding to the training CT is also provided in the dataset.
3.2. Indicators
In order to evaluate the performance of liver tumour segmentation, this paper studies the evaluation comparison between the algorithm in this paper and other algorithms with the same research work in the liver tumour segmentation stage under the LITS data set. It is mainly divided into the following two indicators: (1) Accuracy; (2) Sensitivity; the two indicators are used in liver segmentation and liver tumour segmentation, respectively.

\[
\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \tag{1}
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2}
\]

where TP is true class; TN is true negative; FN is false negative; FP is false positive.

3.3. pre-processing
First, the slice thickness was unified to 1 mm. All images in the new dataset were then subjected to ROI extraction according to the gold standard, followed by histogram equalization. Three-dimensional linear interpolation, window transformation, effective distance extraction and sub-image generation are used. Where for sub-image generation, a patch of size 16*256*256 is used in this paper to handle the balance between the available GPU memory and the contextual information retained in the patch.

3.4. Training
In addition to rotation, scaling and grey value enhancement, we also applied histogram equalization on data and ground truth labels. Network output and ground truth labels were compared using soft-max with weighted cross-entropy loss, where we reduced the weight of common background and increased the weight of liver and liver tumours to achieve a balanced effect of liver and liver tumours versus background voxels on loss. We used the stochastic gradient descent solver of the TensorFlow framework for network training. To be able to train large 3D networks, we used a memory-efficient cuDNN3 convolution layer implementation. Data enhancement is done in the run, which produces as many different images as training iterations. We performed 500 training iterations on the NVIDIA Tesla V100 GPU, which took about 2 days.

3.5. Comparison experiments
The segmentation performance of different algorithms for liver tumours segmentation are shown as Table 1 and Table 2.

| Model     | Year | Accuracy | Sensitivity |
|-----------|------|----------|-------------|
| 3D U-Net  | 2016 | 0.923    | ~           |
| RA U-Net  | 2018 | ~        | 0.932       |
| BS-Unet   | 2018 | 0.934    | 0.922       |
| nnU-Net   | 2019 | 0.950    | ~           |
| ours      | 2020 | 0.965    | 0.963       |

| Model     | Year | Accuracy | Sensitivity |
|-----------|------|----------|-------------|
| 3D U-Net  | 2016 | 0.482    | ~           |
| RA U-Net  | 2018 | ~        | 0.530       |
| BS-Unet   | 2018 | 0.518    | 0.512       |
| nnU-Net   | 2019 | 0.740    | ~           |
| ours      | 2020 | 0.596    | 0.617       |

Although the 3D U-Net model[4] has fast processing speed, there are many problems of feature information loss and small receptive field in the process of down-sampling and up-sampling of the network, so the accuracy and sensitivity are lower than those of this algorithm; while RA U-Net[6] processes by using the attention residual mechanism, but only stack the attention modules, resulting in the problem of attention redundancy, while this paper performs the initial screening of attention before
stacking attention, greatly reducing the redundancy of features, so this paper achieves better results; BS-Unet[7] uses bottleneck supervision to achieve good results in shape distortion and reducing false and false negatives, but the network uses densely connected modules, the network is too complex, and the training time is about 10 times higher than that of this algorithm. Compared with nnU-Net[8], although the algorithm in this paper performs inferior in tumour segmentation, it works best in liver segmentation. Through feature utilization by methods such as attention mechanisms, the effective utilization of features is improved, and the liver recognition accuracy is increased by approximately 1.4%. Compared with most algorithms, the accuracy of liver tumour identification is improved by about 1.1%.

The performance of the proposed liver tumour segmentation algorithm in this paper is verified on the LITS data set. The U-type network structure can effectively utilize the deep and shallow information, which makes the network framework more precise than the traditional FCN framework segmentation. Multi-scale attention module not only enhances the edge characteristics of liver and liver tumours, improves the more detailed characteristics of liver and liver tumours, but also helps the feature extraction of each step to be more effective, greatly accelerates the operation speed of the model, and also effectively ensures the performance of the model.

4. Conclusion

From Figure 5, we can see the segmentation results of the segmentation algorithm in this paper. Compared with the algorithm in this paper, 3D U-Net has coarser segmentation results because of the lack of attention to useful weights, resulting in the network lacking the utilization of more useful features. Compared with the improvement of traditional algorithm, the obtained liver segmentation results are more refined and the edge information is more significant. It further shows that the algorithm in this paper can segment better liver tumours with high accuracy and sensitivity, which has certain clinical value.

![Figure 5. Our algorithm segmentation results. The first column is the CT picture of the liver tumour. The second column is the extracted liver ROI. The third column is the 3D U-Net model liver tumour segmentation result map. The fourth column is the resulting plot of liver tumour segmentation for the MAM U-Net model. The fifth list is the gold standard.](image-url)
Although this study resulted in good performance, it also has its limitations. First, the imperfect tags provided by the LITS dataset, which hinders the further improvement of network learning performance, and there is still much future work on how to handle and make full use of such incomplete tags. Secondly, for the segmentation of some liver tumours with small size and variable shape that cannot be identified by the naked eye, there are still errors and omissions, which may be due to the limitation of current hardware level that three-dimensional data cannot be processed in large quantities, as well as the influence of limited three-dimensional training datasets, further research will continue to be done for the above problems in the future.

Acknowledgments
This work is supported by the Science & Technology Development Program of Jilin Province China (Nos. 20180201037SF, 20190201196JC, 20190302112GX, 20200404142YY, 20200403127SF and 20200401078GX).

References
[1] Chlebus, G., Schenk, A., Moltz, G., Ginneken, B., Hahn, H., and Meine, H., J. (2018) Deep learning based automatic liver tumour segmentation in CT with shape-based post-processing. In: Medical Imaging with Deep Learning. Amsterdam. pp.1-9.
[2] SUN C, J., GUO S, ZHANG H, et al. (2017) Automatic segmentation of liver tumours from multiphase contrast-enhanced CT images based on FCNs. Artificial Intelligence in Medicine., 83: 58-66.
[3] LI X, J., CHEN H, QI X, et al. (2018) H-Dense U-Net: hybrid densely connected UNet for liver and tumour segmentation from CT volumes. IEEE Trans Med Imaging., 37(12): 2663-2674.
[4] Çiçek, Ö., Abdulkaadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016) 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M., Unal, G., Wells, W. (eds.). Med Image Comput Comput Assist Interv. Springer. Switzerland. pp. 424-432.
[5] Roy, A. G., Navab, N., & Wachinger. (2018). Concurrent Spatial and Channel ‘Squeeze & Excitation’ in Fully Convolutional Networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Granada. pp.421-429.
[6] Jin, Q., Meng, Z.-P., Sun, C., Wei, L., & Su, R. J. A. (2018). RA-UNet: A hybrid deep attention-aware network to extract liver and tumour in CT scans. http://arxiv.org/abs/1811.01328.
[7] Li, S., & Tso, G. K. F. J. A. (2018). Bottleneck Supervised U-Net for Pixel-wise Liver and Tumour Segmentation. https://arxiv.org/abs/1810.10331.
[8] Fabian Isensee et al. (2020). Automated Design of Deep Learning Methods for Biomedical Image Segmentation. In: arXiv preprint arXiv:1904.08128.