

EffTransUNet: One Method for Medical Image

Tasks Based on TransUNet

Zhuohang Chen

*Sussex Artificial Intelligence Institute, Zhejiang Gongshang University, Hangzhou, China
Email: 15246194000@163.com*

Abstract: With the advancement of deep learning technologies, these techniques have been widely applied across various fields. Medical image segmentation is one of the crucial tasks in the medical domain. In this paper, the task is to modify the TransUNet model for Synapse multi-organ segmentation dataset and LeftAtrium dataset to achieve higher segmentation accuracy. The model is also applied to the Brain Tumor dataset to perform the image classification task. After experimentation, EffTransUNet the modified model achieved an accuracy of 91.33% on LeftAtrium dataset. The model performed well on the dataset, achieving a classification accuracy of 99.69%. The paper successfully use the modified model to improve segmentation accuracy in LeftAtrium dataset and successfully classified brain tumor Magnetic Resonance Imaging (MRI). It indicates that this method has generalization.

Keywords: Medical Images, Image Segmentation, Efficientnet, TransUNet.

1. Introduction

With the development of deep learning, convolutional neural networks (CNNs) are widely used in various fields such as medicine. Now CNNs are widely used in medical image segmentation task. In this paper, the parameters and structure of TransUNet will be modified to obtain higher accuracy and applied to other medical image tasks. Studies [1, 2] propose U-net and Efficientnet for medical image segmentation. Studies [3, 4] propose networks with encoder-decoder structures. However, studies cannot capture global features of medical images and these methods of segmentation accuracy still needs to be improved. The contributions in this paper are described as follow:

(a) *Higher Segmentation Accuracy:* (i) *Parameter variation of TransUNet:* The accuracy is slightly improved by modifying the parameters. (ii) *Modify the structure of network:* We propose EffTransUNet. Its accuracy can be increased by at least 2.8% on multi-organ datasets.

(b) *Complete New Dataset Tasks:* The medical segmentation dataset LeftAtrium is applied to EffTransUNet and the accuracy is higher than 91%. Brain Tumor Data is classified by EffTransUNet to solve the classification problem of brain tumors, the accuracy reached 99.69%.

2. Related Work

The network structures can be divided into CNNs [1-3, 5, 6], Transformer integrated structures which can capture global features [4, 7-9] and attention mechanism [10-12].

(a) CNNs: Study [1, 3, 6] respectively proposed U-net, RescueNet and DeepMRSeg which has encoder-decoder architecture. U-net is not enough to complex tasks. RescueNet uses an training method to label large-scale data but cannot solve multi-task problems. DeepMRSeg perform well in complex tasks but has limitation in small sample scenarios. Study [2] proposed Efficientnet which is suitable for image tasks, but it cannot flexibly capture global context.

(b) CNNs with Transformer: Studies [4, 7-9] proposed transformer which has been improved. Studies [7, 8] proposed models combining Transformer and CNN, which are limited to processing local information and have poor interpretability respectively. Studies [4, 9] respectively proposed TransUNet and Swin-Unet which can focus on both global and local information, but they are not suitable for small targets. TransUNet is treated as baseline in this paper.

(c) Attention Mechanism: Studies [10-12] respectively proposed Convolutional Block Attention Module (CBAM), Coordinate Attention (CA) and Squeeze-and-Excitation Networks (SE). SE focuses on channels but ignores spatial dimension. CA which obtains spatial information is hard to complete complex tasks. The computational cost of CBAM which combines channel and spatial attention is high.

3. Methodology

Given medical images as input, we complete the segmentation tasks by EffTransUNet. It consists of an encoder and a decoder which are shown in Figure 1. Figure 2 shows the details of network.

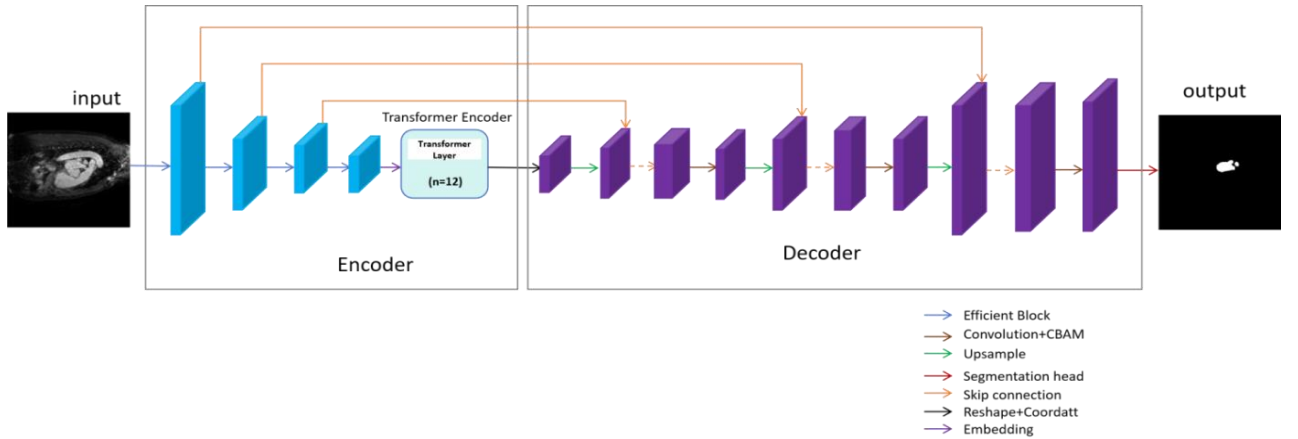


Figure 1: The overall flow of EffTransUNet. In this paper, the input is a 3-channel image.

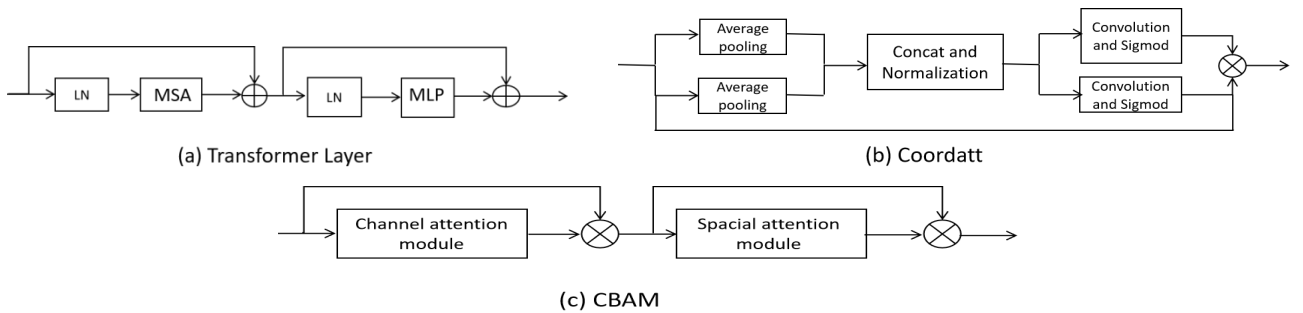


Figure 2: Details of EffTransUNet. (a) shows the structure of Transformer layer. (b) and (c) show the details of attention mechanisms which are added to the decoder.

3.1. Overview of EffTransUNet

In experiments, the medical image x ($x \in \mathbb{R}^{H \times W \times 3}$) is input into the network, and input is a 3-channel RGB image. The objective of the method is to generate the predicted label maps with the relevant medical images whose sizes are $H \times W$. The way to complete this task is to combine Transformer and CNNs. The network is divided into two parts: encoder and decoder. The encoder consists of Efficientnet and Transformer. CBAM and CA are introduced to decoder. The last part of the network is the segmentation head, which is used to obtain the segmentation result consistent with the resolution of input.

3.2. EffTransUNet Encoder

The TransUNet encoder combines Efficientnet and Transformer, which is a hybrid encoder structure of CNN and Transformer. The first step of image processing is that the shape of the image changes from the input $3 \times 224 \times 224$ to $768 \times 7 \times 7$ by Efficientnet. The next step is to do image sequentialization and input the sequence into the Transformer encoder. Input is changed into a vector, and the form of the vector is 49×768 .

3.2.1. Efficientnet

Efficientnet is used for feature extraction. Efficientnet adopts a compound scaling strategy, which simultaneously optimizes the depth, width and resolution of the network. The core module in it is Mobile Inverted Bottleneck, which is mainly composed of Depthwise convolution and Squeeze-and-Excitation Network [2].

3.2.2. Transformer Layers

There are a total of 12 Transformer layers in the hybrid encoder structure. Transformer encoder consists of multiple layers of Multihead Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks shown in Figure 2 (a). The structure of the Transformer layer can be represented by the following equations (1) (2). In the equations, the MSA structure is regarded as a function. After being given the input, the corresponding output is obtained by the function. The same applies to LN and MLP.

$$x' = input + MSA(LN(input)) \quad (1) \quad output = x' + MLP(LN(x')) \quad (2)$$

Input obtained from patch embedding is first normalized using Layer Normalization (LN) before entering MSA. The output of MSA is then processed by a residual connection. Then the sequence is further processed by MLP, which also includes a residual connection, resulting in the final output of the current Transformer layer [4]. The MLP consists of two fully connected layers with a non-linear activation function between them.

3.3. EffTransUNet Decoder

The decoder is mainly composed of the upsampling layer and the attention mechanism. The output of the needs to be restored to the size of the image resolution. Then CA is used to obtain more position information for feature. The upsampling layers and the segmentation head are simply connected to form the main body of decoder. Each upsampling layer consists of one upsampling operation, one skip connection, two convolution operations and CBAM. Skip connection is the fusion of symmetrical encoder features and decoder features. The feature is processed through CBAM to focus on the effective information of channels and spaces. Finally, the final result is output by the segmentation head.

From Figures 2 (b) and (c), CA uses the average pooling layer to process information in both horizontal and vertical directions and captures the long-distance dependencies in the two directions. The features after convolution processing are split, and the original feature map is multiplied by the weights in the two directions to obtain the final result [11]. CBAM consists of channel attention module and spacial attention module. Channel attention contains global average pooling and global maximum pooling on the feature maps. Spacial attention performs maximum pooling and average pooling on all channels at each position on the feature map, and then successively obtains the weights of spatial attention through convolution [10].

3.4. Difference between EffTransUNet and TransUNet

EffTransUNet is a network which change the part for extracting features based on TransUNet. In this paper, Efficientnet is used to replace Resnet in TransUNet. EffTransUNet-B3, EffTransUNet-B4 and EffTransUNet-B5 respectively use Efficientnet-B3, Efficientnet-B4 and Efficientnet-B5 as the feature extraction network to replace Resnet. In addition, CA is added to the model after the encoder, the input of CA is the output of the encoder. CBAM is introduced to each upsampling layer. The advantage of the new model structure is the compound scaling strategy of Efficientnet and attention mechanism, which enables the model to extract more effective features compared to TransUNet and pay more attention to the effective information in space and channels.

4. Experiments

4.1. Experiment Setup

The GPU model used in the experiment is an NVIDIA GPU with 16GB of GPU memory. The experiment was implemented using the python language.

4.1.1.Dataset

The experiments in this paper are based on three datasets, and the datasets used are as follows.

(a) *Synapse multi-organ segmentation dataset* is a nine-class 3D medical image segmentation dataset. This is the dataset used by the original model [4].

(b) *LeftAtrium* is a binary classification 3D medical image segmentation dataset [13].

(c) *Brain Tumor Data* is a dataset related to brain tumors for classification tasks, and it is a four-class task dataset [14].

4.1.2.Algorithm

The performance of TransUNet and proposed EffTransUNet in this paper on different segmented datasets was compared through experiments. In TransUNet, it consists of Resnet, encoder of Transformer and decoder of U-net. TransUNet loads the pre-trained weights of Imagenet and trains for 150 epoch [4].

EffTransUNet consists of Efficientnet, encoder of Transformer and decoder. CBAM and CA are added to the decoder. The algorithm uses the Adamw optimizer, freezes Efficientnet parameters for pre-training for 300 epoch, then unfreezes Efficientnet parameters and continues to train EffTransUNet for 150 epoch.

4.1.3.Evaluation Metric

In this experiment, the performance of the model is evaluated by two metrics, which are as follows.

(a) *Dice Coefficient* is a set similarity measure function. The higher the Dice Coefficient value, the better the experiment effect [15].

(b) *Hausdorff Distance* is to measure the similarity between two point sets. The higher the value of Hausdorff Distance, the better the experimental effect [16].

4.1.4. Variation of Parameters in TransUNet

We modified three parameters in order to obtain higher segmentation accuracy of TransUNet. Synapse multi-organ segmentation dataset is used in the experiments. The three parameters to be modified are as follows.

- (a) *Epoch* is changed from 101 to 200 and test them to find the epoch with the highest accuracy.
- (b) *Learning rate* is used to control the step size of the model when updating parameters.
- (c) *Batch size* defines the number of samples to be used for each parameter update.

4.2. Experiment Results

4.2.1. Results of Parameter Variation

The performance results of the model after modifying three parameters are shown as follows.

(a) Epoch: From Figure 3 (a), the dice which is 0.7848 is highest when epoch is 105.

(b) Learning rate: Learning rate are { 0.2, 0.1, 0.05, **0.01**, 0.001 }, and dice metrics were { 0.7572, 0.7733, 0.7785, **0.7788**, 0.7328 }. As the learning rate increases, the model reaches its peak when learning rate is 0.01. When learning rate is higher than 0.01, dice gradually decreases. It might be that during the training process, if the learning rate is too low, the model cannot converge rapidly, and if it is too high, the model cannot reach the optimum.

(c) Batch size: The results corresponding to batch size { 8, 16, **24**, 32, 48 } are {0.7740, 0.7771, **0.7788**, 0.7686, 0.7599}. When batch size is higher than 24, accuracy of the model decreases as the batch size increases. The possible reason is that the batch size is too large and the convergence speed is too slow, resulting in underfitting or the model entering a suboptimal solution.

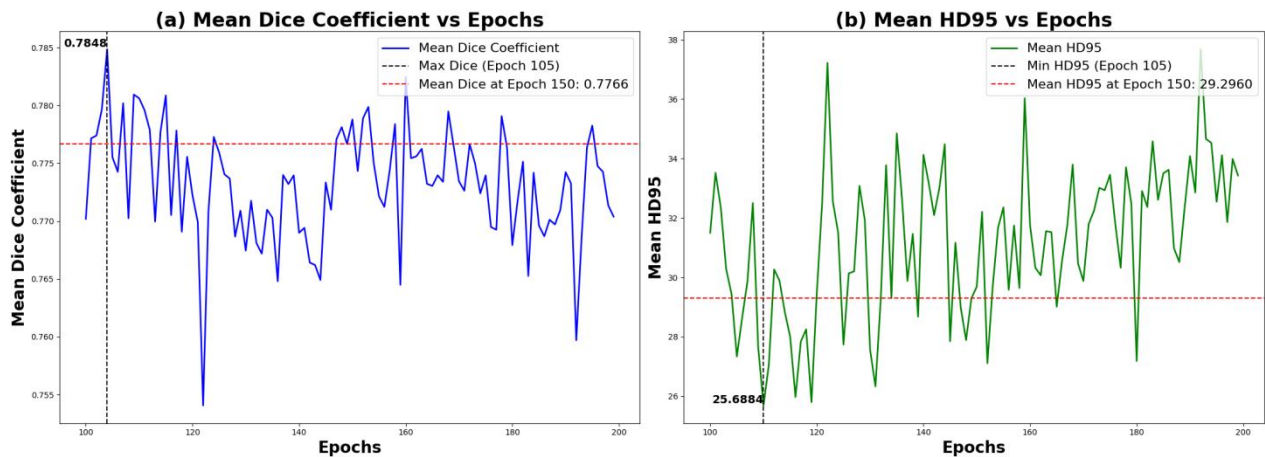


Figure 3: The diagram shows the change of metrics by changing epoch.

4.2.2. Results of Different Models

After experiments, the results of the EffTransUNet and TransUNet models are as follows.

(a) Dice metrics of *TransUNet* in the multi-organ dataset were 0.7788 and 0.7328 when the learning rates were 0.01 and 0.001. The metrics of LeftAtrium was 0.9009 and 0.8777 in the same condition.

(b) Due to space limitations, visualization results of *EffTransUNet* are placed in the technical report [17]. The test results of *EffTransUNet* based on the three datasets are as follows.

- *Synapse multi-organ segmentation dataset*

The experimental results of EffTransUNet and TransUNet are showed in Table 1. When learning rate is 0.01, dice of EffTransUNet-B4 is 0.8006 (2.8% higher than TransUNet). EffTransUNet-B3 reaches 0.8134 which is better than TransUNet (0.7328). The possible reason for the higher accuracy of EffTransUNet is the efficient compound scaling strategy of Efficientnet. It can obtain rich semantic and spatial information by uniformly and coordinately scaling the depth, width and input resolution of the network, while Resnet only improves performance from a single dimension. The reason why the segmentation accuracy of TransUNet is not high at a learning rate of 0.001 might be that during the training of the model, due to the small learning rate, the model has not converged or is in a local optimum.

Table 1: Performance comparison of EffTransUNet and TransUNet on multi-organ dataset (left).

Table 2: Performance comparison of EffTransUNet and TransUNet on LeftAtrium (right).

Model	Lr = 0.01		Lr = 0.001	
	Dice	HD95(mm)	Dice	HD95(mm)
EffTransUNet-B3	0.7904	27.5911	0.8134	23.6755
EffTransUNet-B4	0.8006	21.4563	0.8086	20.1242
EffTransUNet-B5	0.7515	36.8063	0.8109	21.6749
TransUNet	0.7788	29.6834	0.7328	36.4315

Model	Lr = 0.01		Lr = 0.001	
	Dice	HD95(mm)	Dice	HD95(mm)
EffTransUNet-B3	0.9088	3.1507	0.9113	2.9120
EffTransUNet-B4	0.9133	2.9093	0.9079	2.9251
EffTransUNet-B5	0.8904	3.9922	0.9111	2.7719
TransUNet	0.9009	3.6201	0.8777	4.2967

- *LeftAtrium*

The difference between this experiment and the experiment on multi-organ datasets is that 200 epoch are trained after unfreezing the parameters of Efficientnet. The results are showed in Table 2. When learning rate is 0.01, dice of EffTransUNet-B4 reaches 0.9133. When learning rate is 0.001, dice of EffTransUNet-B3 reaches 0.9113, which is 3.8% higher than that of TransUNet. EffTransUNet performs better than TransUNet in binary classification problems. This is because the model can extract effectively detailed information and global context information.

- *Brain Tumor Data*

Brain tumor image classification task is completed by EffTransUNet. When completing the task, skip connections are removed. Accuracy of EffTransUNet reached 99.69%. This indicates that EffTransUNet has the potential to become a basic model for completing multi-task learning and transfer learning. This also indicates that the algorithm is robust, and the network can extract features with rich semantic information and strong generalization ability.

5. Conclusion

This paper proposes the EffTransUNet model for medical image segmentation. After experimental verification, the segmentation accuracy of this method is higher than that of TransUNet based on the two segmented datasets. The model was applied to a classification dataset and it can achieve a high classification accuracy. The medical image segmentation tasks and medical image classification tasks have been successfully completed by using EffTransUNet. The experimental results show that EffTransUNet is robust and has the potential to solve multi-task learning problems and complete transfer learning tasks.

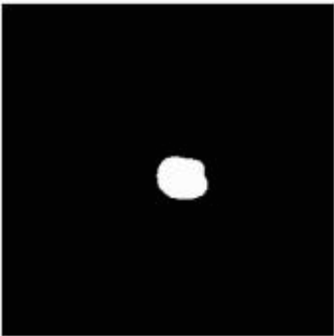
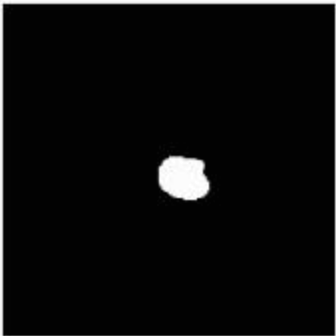

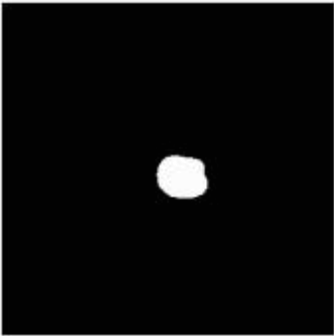




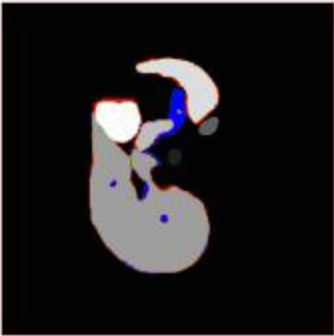
References:

- [1] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18* (pp. 234-241). Springer international publishing.
- [2] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PMLR.
- [3] Nema, S., Dudhane, A., Murala, S., & Naidu, S. (2020). RescueNet: An unpaired GAN for brain tumor segmentation. *Biomedical Signal Processing and Control*, 55, 101641.
- [4] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.
- [5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
- [6] Neelima, G., Chigurukota, D. R., Maram, B., & Girirajan, B. (2022). Optimal DeepMRSeg based tumor segmentation with GAN for brain tumor classification. *Biomedical Signal Processing and Control*, 74, 103537.
- [7] 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*.
- [8] Gao, Y., Zhou, M., & Metaxas, D. N. (2021). UTNet: a hybrid transformer architecture for medical image segmentation. In *Medical image computing and computer assisted intervention—MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, Part III 24* (pp. 61-71). Springer International Publishing.
- [9] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2022, October). Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision* (pp. 205-218). Cham: Springer Nature Switzerland.
- [10] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3-19).
- [11] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13713-13722).
- [12] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141).
- [13] Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., ... & Cardoso, M. J. (2022). The medical segmentation decathlon. *Nature communications*, 13(1), 4128.
- [14] Ghaffar, A. (2024). Brain Tumor Data. Mendeley Data, V1.
- [15] Andrews, S., & Hamarneh, G. (2015). Multi-region probabilistic dice similarity coefficient using the Aitchison distance and bipartite graph matching. *arXiv preprint arXiv:1509.07244*.
- [16] Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9), 850-863.

Appendix

A. Visualization results

Here are the visual segmentation results obtained in the two datasets when using Efficientnet-B3 and Efficientnet-B4 respectively as the feature extraction networks and the learning rates are 0.01 and 0.001.

EffTransUNet		Label	Prediction	Error
B4 (lr=0.01)	(LeftAtrium)			
B3 (lr=0.001)	(LeftAtrium)			
B4 (lr=0.01)	(multi-organ)			
B3 (lr=0.001)	(multi-organ)	