Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluatio
DOI: 10.54254/2755-2721/2025.PO25359

# *EffTransUNet: One Method for Medical Image Tasks Based on TransUNet*

**Zhuohang Chen**

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

***Abstract.*** With the rapid advancement of deep learning technologies, their applications have expanded across various fields, particularly in medical image analysis, where segmentation remains a critical task. This study proposes a modified version of the TransUNet model, named EffTransUNet, to enhance segmentation performance on the Synapse multi-organ segmentation dataset and the LeftAtrium dataset. Additionally, the model is applied to the Brain Tumor dataset for image classification. Experimental results show that EffTransUNet achieves an accuracy of 91.33% on the LeftAtrium dataset and a classification accuracy of 99.69% on the Brain Tumor dataset. These findings demonstrate that the proposed model effectively improves segmentation performance and accurately classifies brain tumour MRI images, indicating good generalization ability.

***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, including medicine. In particular, CNNs are commonly applied to medical image segmentation tasks. This paper focuses on improving the performance of the TransUNet model by modifying its parameters and structure to achieve higher segmentation accuracy and expanding its application to other medical imaging tasks.

Previous studies have proposed architectures such as U-Net and EfficientNet for medical image segmentation [1,2], as well as encoder-decoder structures for improved feature extraction [3,4]. However, these models often struggle to capture global features in medical images, and their segmentation accuracy still needs improvement.

The contributions of this paper are as follows:

(a) Higher Segmentation Accuracy:

(i) Parameter variation of TransUNet: Accuracy is slightly improved by modifying the parameters.

(ii) Structural modification: We propose EffTransUNet, which increases accuracy by at least 2.8% on multi-organ datasets.

(b) Complete New Dataset Tasks:

EffTransUNet is applied to the LeftAtrium medical segmentation dataset and achieves an accuracy above 91%. It is also used on the Brain Tumor Data for classification tasks, achieving a

DOI: 10.54254/2755-2721/2025.PO25359

classification accuracy of 99.69%.

Code: https://github.com/next293/EffTransUNet

## 2. Related work

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

(a) CNNs: Studies [1,3,6] proposed, respectively, U-Net, RescueNet, and DeepMRSeg, all of which adopt encoder-decoder architectures. U-Net is insufficient for handling complex tasks. RescueNet uses a training method to label large-scale data, but struggles with multi-task problems. DeepMRSeg performs well in complex scenarios but is limited in small-sample situations. Study [2] proposed EfficientNet, which is effective for general image tasks but lacks the ability to flexibly capture global context.

(b) CNNs with Transformer: Studies [4,7-9] introduced various Transformer-based improvements. Studies [7, 8] proposed models that combine Transformers with CNNs, but these are limited to processing local information and suffer from poor interpretability. Studies [4,9] proposed TransUNet and Swin-Unet, respectively, which can process both global and local information. However, they are less effective for detecting small targets. TransUNet is used as the baseline in this paper.

(c) Attention Mechanisms: Studies [10-12] proposed Convolutional Block Attention Module (CBAM), Coordinate Attention (CA) and Squeeze-and-Excitation Networks (SE), respectively. SE focuses on channel information while ignoring spatial dimensions. CA incorporates spatial information but struggles with complex tasks. CBAM combines channel and spatial attention but incurs a high computational cost.

## 3. Methodology

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

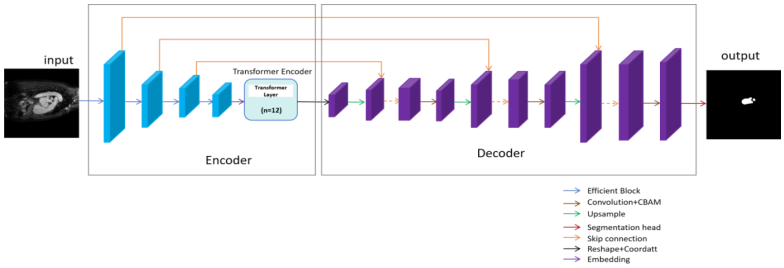


Figure 1. The overall flow of EffTransUNet. In this paper, the input is a 3-channel image

DOI: 10.54254/2755-2721/2025.PO25359

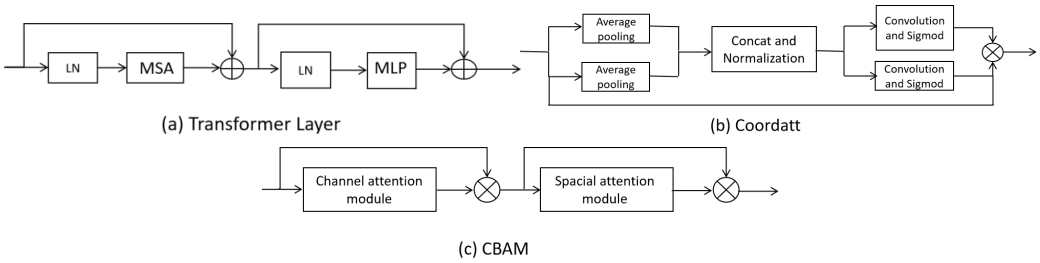


Figure 2. Details of EffTransUNet. (a) shows the structure of the Transformer layer. (b) and (c) show the details of the attention mechanisms added to the decoder

### 3.1. Overview of EffTransUNet

In the experiments, a medical image x ( $x \in R^{H \times W \times 3}$ ) is input into the network. The input is a 3-channel RGB image. The goal is to generate predicted label maps corresponding to medical images whose sizes are $H \times W$ . This task is accomplished by combining Transformer and CNN architectures. The network is divided into two parts: encoder and decoder. The encoder includes EfficientNet and Transformer components. In the decoder, CBAM and CA mechanisms are incorporated. The final part of the network is the segmentation head, which generates the segmentation output with the same resolution as the input.

### 3.2. EffTransUNet encoder

The TransUNet encoder combines EfficientNet and Transformer, forming a hybrid encoder structure that leverages CNNs and Transformers. The first step of image processing involves changing the shape of the image from the input $3 \times 224 \times 224$ to $7\ 68 \times 7 \times 7$ using EfficientNet. Then, the output is sequentialized and passed into the Transformer encoder. The input is converted into a vector, with the form $49 \times 768$ .

#### 3.2.1. EfficientNet

EfficientNet is used for feature extraction. It adopts a compound scaling strategy that simultaneously optimises the depth, width and resolution of the network. Its core module is the Mobile Inverted Bottleneck, which mainly consists of Depthwise convolution and the Squeeze-and-Excitation Network [2].

#### 3.2.2. Transformer layers

There are a total of 12 Transformer layers in the hybrid encoder structure. Each Transformer encoder consists of multiple layers of Multihead Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks, as shown in Figure 2 (a). The structure of the Transformer layer can be represented by the following equations (1) and (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.

$$x\prime = input + MSA(LN(input)) \tag{1}$$

$$output = x\prime + MLP(LN(x\prime)) \tag{2}$$

DOI: 10.54254/2755-2721/2025.PO25359

The 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 upsampling layers and attention mechanisms. Its main function is to restore the output to the original image resolution. CA is incorporated to provide additional positional information to the features. The upsampling layers and the segmentation head together form the main body of the decoder. Each upsampling layer consists of one upsampling operation, one skip connection, two convolution operations and a CBAM module. Skip connections fuse features from the encoder and decoder. The features are then processed by CBAM to focus on the effective information of channels and spaces. Finally, the result is output by the segmentation head.

As shown in 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. After convolution, the features 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 a channel attention module and a spatial attention module. The channel attention module applies global average pooling and global maximum pooling to the feature maps. Spatial attention performs average and maximum pooling across all channels at each spatial location, then uses convolution to generate the spatial attention weights [10].

### 3.4. Difference between EffTransUNet and TransUNet

EffTransUNet is a network that modifies the feature extraction component of the original TransUNet architecture. In this paper, EfficientNet is used to replace ResNet in TransUNet. Specifically, EffTransUNet-B3, EffTransUNet-B4 and EffTransUNet-B5 use EfficientNet-B3, EfficientNet-B4 and EfficientNet-B5, respectively, as the feature extraction network to replace ResNet.

Additionally, CA is added to the model after the encoder, taking the encoder's output as input. CBAM is introduced into each upsampling layer. The main advantage of this new structure lies in the compound scaling strategy of EfficientNet combined with the attention mechanisms. This design allows the model to extract more effective features compared to TransUNet and enhances its focus on important spatial and channel information.

## 4. Experiments

### 4.1. Experiment setup

The experiments were conducted using an NVIDIA GPU with 16GB of memory. The model was implemented using the Python language.

#### 4.1.1. Dataset

Three datasets were used in the experiments:

(a) Synapse multi-organ segmentation dataset: A nine-class 3D medical image segmentation dataset, also used in the original model [4].

DOI: 10.54254/2755-2721/2025.PO25359

(b) LeftAtrium: A binary classification 3D medical image segmentation dataset [13].

(c) Brain Tumor Data: A four-class medical image dataset used for classification tasks related to brain tumour classification [14].

#### 4.1.2. Algorithm

The performance of TransUNet and the proposed EffTransUNet was compared across different segmentation datasets. TransUNet consists of ResNet, a Transformer encoder and a U-Net decoder. It loads pre-trained ImageNet weights and is trained for 150 epochs [4].

EffTransUNet, by contrast, integrates EfficientNet, a Transformer encoder and a decoder enhanced with CBAM and CA modules. The algorithm uses the Adamw optimiser. During training, EfficientNet parameters are frozen for 300 epochs for pre-training, then unfrozen, and the full model is trained for an additional 150 epochs.

#### 4.1.3. Evaluation metric

In this experiment, the model's performance is evaluated using two metrics:

(a) Dice Coefficient: A function that measures set similarity. A higher Dice Coefficient indicates better segmentation performance.

(b) Hausdorff Distance: A metric for measuring the similarity between two point sets. A lower Hausdorff Distance value indicates better experimental results [16].

#### 4.1.4. Variation of parameters in TransUNet

To improve TransUNet's segmentation accuracy, three parameters were adjusted using the synapse multi-organ segmentation dataset:

(a) Epochs: The number of training epochs was increased from 101 to 200 to determine the optimal number for peak accuracy.

(b) Learning rate: This controls the step size of the model when updating parameters.

(c) Batch size: Defines the number of samples 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 below:

(a) Epoch: As shown in Figure 3 (a), the Dice score reaches its highest value of 0.7848 when the number of epochs is 105.

(b) Learning rate: The learning rates were { 0.2, 0.1, 0.05, 0.01, 0.001 }, and the Dice metrics were { 0.7572, 0.7733, 0.7785, 0.7788, 0.7328 }. The model achieves its best performance at a learning rate of 0.01. When the learning rate exceeds 0.01, the Dice score gradually decreases. This may be because a learning rate that is too low results in slow convergence, while a rather that is too high may cause the model to fail to 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 the batch size exceeds 24, performance decreases. This may be due to larger batch sizes causing slow convergence, potentially resulting in underfitting or the model entering a suboptimal solution.

DOI: 10.54254/2755-2721/2025.PO25359

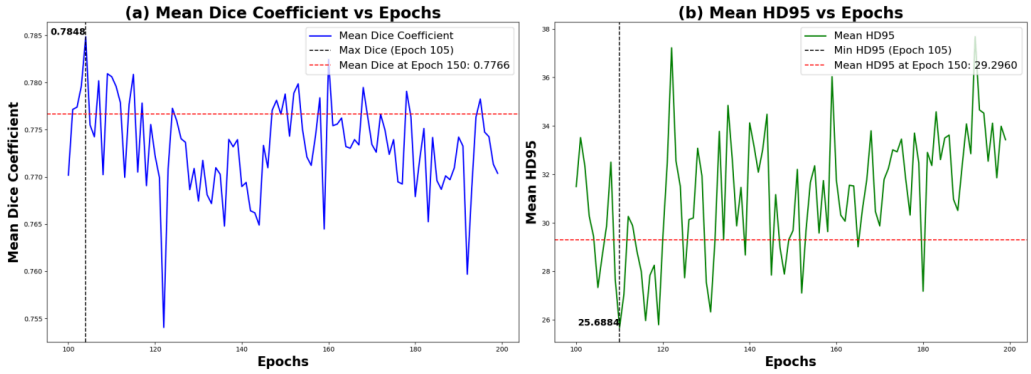


Figure 3. The diagram shows the change in metrics by changing epoch

### 4.2.2. Results of different models

After conducting the experiments, the performance 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, respectively. On the LeftAtrium dataset, the scores were 0.9009 and 0.8777 under the same conditions.

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

·Synapse multi-organ segmentation dataset

Experimental results of EffTransUNet and TransUNet are shown in Table 1. When the learning rate is 0.01, the Dice score of EffTransUNet-B4 is 0.8006 (2.8% higher than TransUNet). EffTransUNet-B3 achieves a Dice score of 0.8134, which is significantly better than TransUNet's score of 0.7328.

The improved accuracy of EffTransUNet can be attributed to EfficientNet's compound scaling strategy. It can obtain rich semantic and spatial information by uniformly and co-ordinately scaling the depth, width and input resolution of the network, while ResNet only improves performance from a single dimension. The lower segmentation accuracy of TransUNet, at a learning rate of 0.001, is likely due to insufficient convergence during training or getting stuck in a local optimum, caused by the small step size during training.

Table 1. Performance comparison of EffTransUNet and TransUNet on multi-organ dataset

| 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 |

DOI: 10.54254/2755-2721/2025.PO25359

Table 2. Performance comparison of EffTransUNet and TransUNet on LeftAtrium

| 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 multi-organ segmentation experiments is that, after unfreezing the parameters of EfficientNet, the model was trained for an additional 200 epochs. The results are shown in Table 2. When the learning rate is 0.01, the Dice score of EffTransUNet-B4 reaches 0.9133. When the learning rate is 0.001, the Dice score of EffTransUNet-B3 reaches 0.9113, which is 3.8% higher than that of TransUNet. EffTransUNet performs better than TransUNet in binary classification problems, likely because the model can extract both detailed information and global context information.

· Brain Tumor Data

EffTransUNet was also applied to the brain tumour image classification task. In this task, skip connections were removed. The model achieved an accuracy of 99.69%, indicating that EffTransUNet has potential as a foundational model for multi-task learning and transfer learning. This also demonstrates the robustness of the algorithm, showing that 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. Experimental results show that EffTransUNet achieves higher segmentation accuracy than TransUNet based on the two segmented datasets. The model was also applied to a classification dataset, achieving high classification accuracy. EffTransUNet successfully completes both medical image segmentation and medical image classification tasks. The results confirm that EffTransUNet is a robust model with the potential for solving multi-task learning problems and complete transfer learning tasks.

## References

[1] Ronneberger, O., Fischer, P. and 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. and 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. and 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., ... and Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv: 2102.04306.

[5] He, K., Zhang, X., Ren, S. and 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. and Girirajan, B. (2022). Optimal DeepMRSeg based tumor segmentation with GAN for brain tumor classification. Biomedical Signal Processing and Control, 74, 103537.

DOI: 10.54254/2755-2721/2025.PO25359

[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. and 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. and 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. and 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. and 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. and 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. and 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. and Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. IEEE Transactions on pattern analysis and machine intelligence, 15(9), 850-863.
[17] Chen, Z. EffTransUNet: One Method for Medical Image Tasks Based on TransUNet. https: //github.com/next293/EffTransUNet/blob/main/code/technical%20report.pdf.