# Medical Image Segmentation Based on TransUnet and A$^2$B

Zijian Liu *, Yufan Liu

School of Automation, Nanjing University of Aeronautics and Astronautics, Nanjing, China

* Corresponding Author Email: zijian_liu@nuaa.edu.cn

**Abstract.** The objective of medical picture segmentation technology is to accurately delineate and partition regions affected by pathology, which is of great significance to medical diagnosis and treatment. Thanks to deep learning's quick development, prior efforts mainly utilize the convolutional neural networks equipped with powerful feature representation capability. In this paper, to make the segmentation performance a petter appearance, we introduce a new model in medical image segmentation technology, TransUNet, which combines Transformer and U-Net and uses labeled and unlabeled images for training. To enhance the feature quality and boost the accuracy and robustness of the segmentation results, we further introduce the A2B (Attention in Attention Block) attention mechanism in the TransUNet backbone network. At the same time, we use the CNN-Transformer hybrid model and conduct ablation experiments to verify the improvement effect of the performance indicators. The empirical findings show that TransUNet has accomplished success in tasks requiring medical picture segmentation and can help doctors and researchers more accurately identify and locate key structures.

**Keywords:** Medical image segmentation, Transformer, Attention in attention.

## 1. Introduction

One of the most crucial technologies for medical diagnosis and therapy is the application of medical imaging. Doctors can acquire information of body's internal image and structure by using different imaging technologies, like X-ray, supersonic wave, computed tomographic scan (CT), magnetic resonance imaging (MRI) and so on. Medical image segmentation technology is a significant field within medical image processing. Its purpose is to extract and segment certain structures and regions of lesions, hence contributing to the broader domain of computer vision tasks. In the actual diagnosis, medical images are often affected by many factors, like fog, noise, artifacts and so on. Images with low quality may lead to inaccuracy of diagnosis, which will increase the difficulty of doctors' judgment and interpretation. Actually, rapid and exact analysis is important in the clinical environment. Doctors need to acquire and interpret the result of medical images quickly to diagnose and make treatment plans. Therefore, efficient image analysis and automation algorithm are necessary.

In recent years, scientific researchers have exploited various advanced deep learning models to increase the precision and effectiveness of image segmentation in computer vision tasks. Parmar et al. put forward the transformer, which is used to design sequence-to-sequence prediction by using the attention mechanism to catch the correlation among different locations, and then achieve the result of fine semantic segmentation [1]. However, Transformer models may lack low-level details since they mainly focus on global information and semantic representation, which leads to the limitation of positioning capability. Yuan raised the application of U-Net neural network in medical image segmentation [2], which adds cascades to maintain the original features between the encoder and the decoder. Actually, considering the inherent locality of convolution operation, U-Net cannot capture the relevant global context information to some image segmentation tasks. In particular, it shows limitations and poor robustness when dealing with large-scale structures or targets with long-range dependencies. In comparison to these deep learning models, TransUNet is a semi-supervised image segmentation model that combines Transformer and U-Net, which can be trained using labeled and unlabeled images. The unlabeled image is reconstructed by the generator and discriminator network and compared with the original image during training, this may aid in enhancing the model's resilience and capacity for generalisation.

Based on the segmentation task of medical images in the TransUnet backbone network, this paper connects the $A^2B$ (Attention in Attention Block) attention mechanism at the jump and combines the advantages of TransUnet and $A^2B$ to increase the segmentation results' reliability and accuracy. By combining these two models, doctors and researchers can identify and locate key structures more accurately. In this approach, a hybrid model combining CNN and Transformer is employed instead of using a standalone Transformer as the encoder. Here, we choose ResNetV2 with excellent performance. Two sets of ablation experiments were performed to compare whether the performance indicators were improved.

## 2. Related Work

### 2.1. Attention in Attention Block

The attention mechanism in human vision may be thought of as a technique to distribute computer resources to the components with the most information, and the attention process in deep learning is similar to this [3]. The use of this technique is often seen in many computer vision applications such as image captioning [4-5] and picture classification [6], typically comprises a threshold function to build a feature mask. Previous research [6] It has been demonstrated that low-resolution space contains valuable high-frequency components and abundant low-frequency components, and does not use the network's attention mechanism as a result, treating all characteristics equally. The network may be able to focus more on high-frequency features thanks to the attention mechanism. Typically, the high-frequency components represent regions that are rich in edges, textures, and other characteristics [7]. Therefore, adding an attention mechanism to deep learning can enhance network performance by allowing the network to focus on more crucial feature data.

As mentioned above, image information includes low-frequency information and valuable high-frequency components. The use of the attention mechanism can help the network pay attention to more valuable feature information. Despite this, many attention models [7-8] employ fixed attention features regardless of the content of the image. However, at various points in the neural network, the attention layer's efficacy varies. As a result, it was additionally suggested by $A^2B$ on the basis of this [9] that it can develop and serve as an internal branch of the attention weight. The input features define the weight in a dynamic manner [10]. The attention network's capacity has been substantially boosted by the structure $A^2B$, and the parameter overhead is minimal. It is easily adaptable to other low-cost methods with high potential value.

### 2.2. Transformer

Transformer has established a technical level in many NLP problems and was first introduced by [11] for machine translation. Additionally, it has achieved a superior rank in numerous NLP tasks. We have made some changes to the converter to make it appropriate for computer vision jobs as well. For example, Child et al. [12] proposed the sparse transform generator, which uses a scalable global self-attention approximation method. Recently, Wang [13] proposed sparse converter and visual converter (ViT) [13], which reached the most advanced level in ImageNet classification directly implementing converters with global self-attention.

Recently, Vision Transformer (ViT) [12] has successfully classified full-size images using ImageNet by using converters that use global self-attention, the desired outcome may be achieved. This research builds on [12] by combining Transformer with CNN, optimizing Transformer using modules, and creating a novel deep learning framework for medical picture segmentation.

## 3. Proposed Method

The most common approach is to decode the picture back to its original spatial resolution after promptly training CNN (for instance, UNet) to encode it into a high-level feature representation. In

contrast to currently used techniques, our approach provides an encoder with a Transformer-designed self-attention mechanism, whose whole architecture is depicted in Fig. 1.

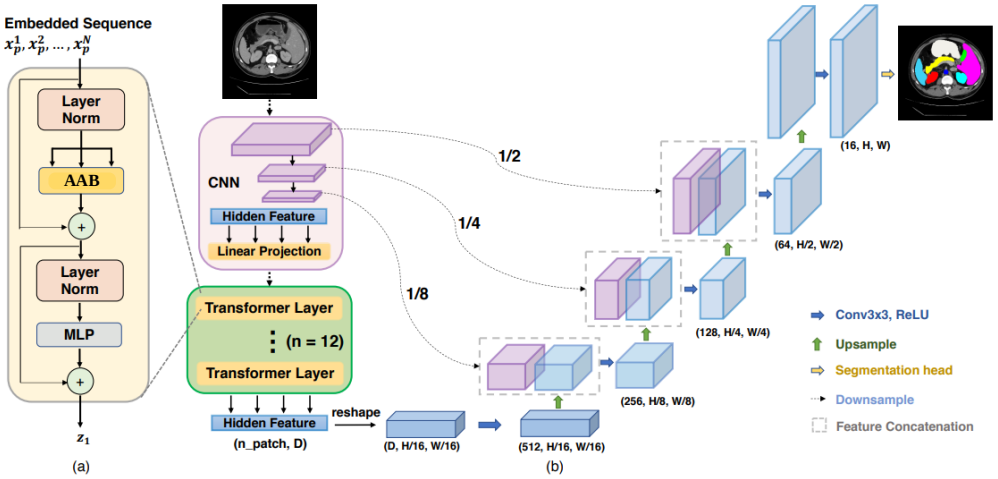



**Figure 1.** The framework of proposed method

### 3.1. Network Structure

The encoder in the approach suggested combines a CNN and Transformer model. A CNN, namely ResNetV2, is utilised to extract features and produce an input feature map because to its superior performance. Blocks taken from the CNN feature maps are applied block embedding rather than the source pictures. The performance of the hybrid CNN-Transformer encoder surpasses that of a pure Transformer encoder attributable to this design's two upsides: 1) This method enables the use of interim high-resolution CNN feature maps throughout the decoding process; and 2) it gives two advantages over other designs.

In the Transformer encoder, the multi-head attention mechanism of the Transformer is improved by including the attention mechanism of the $A^2B$ module, hence enhancing the acquisition of global traits all throughout the learning stage. The following experiments also demonstrate that the AAB module significantly enhances the network's performance.

### 3.2. Attention Calculation

Here, we balance attention branches and non-attention branches while automatically deleting some unimportant attention features using a learnable dynamic attentiveness module. Each dynamic attention module, whose unique structure is shown in Fig. 2, uses weighted summing for controlling the dynamic weighted inputs from cognitive and non-attentional branches.

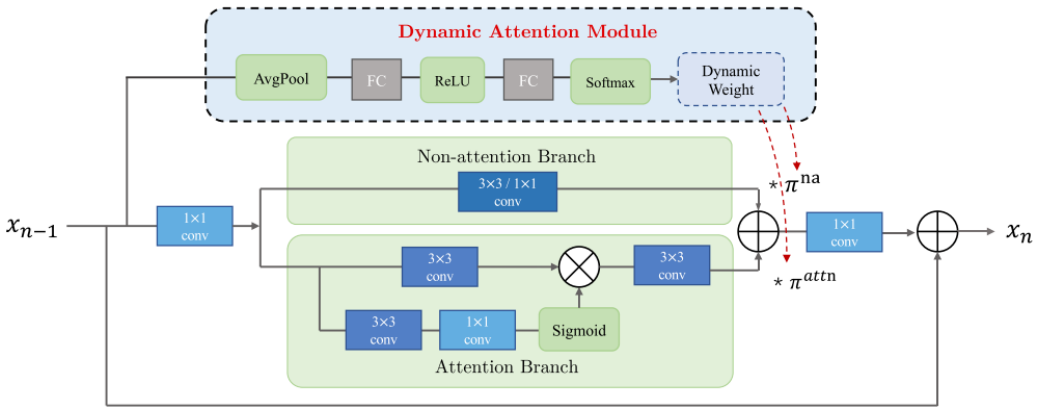



**Figure 2.** The architecture of attention in attention block

Correspondingly, we can also use the mathematical meaning to describe:

$$x_{n+1} = f_{1\times1}(\pi_n^{na} \times x_n^{na} + \pi_n^{attn} \times x_n^{attn}) \quad (1)$$

Among this, $x_n^{na}$ is an outcome of the non-attention branch, and $x_n^{\text{attn}}$ is the exportation of the attention branch.

$x_{n+1} = f_{1\times1}(\pi_n^{na} \times x_n^{na} + \pi_n^{attn} \times x_n^{attn})$ provides the $1 \times 1$ kernel convolution. $\pi_{na}$ and $\pi_{\text{attn}}$ are the non-attention branch alongside the attention branch weights that the network determines based on the input attributes, rather than the two fixed values set artificially. In order to calculate the dynamic weight, there are:

$$\pi_n = f_{da}(x_n) \tag{2}$$

Where $f_{da}(x_n)$ is a keen interest module that changes. Fig. 2 illustrates the dynamic attention module in further detail. It first compresses the data it receives employing global average pooling. The proposed architecture consists of two completely linked layers, each using the Rectified Linear Unit (ReLU) activation function make up the connecting layer. We broaden the receptive area with global combining networks, allowing the dynamic attention module to pick up details from the entire image. The weightings of the two branches undergo modifications in response to variations in the input attributes. As studied in [7], constraint dynamic weight can promote the learning of dynamic attention module. Specifically, there can be abundance sum-to-one constraint $\pi_n^{na} + \pi_n^{attn} = 1$. The summation of dynamic weights imposes a restriction that has the potential to decrease the size of the kernel space. The procedure of learning $\pi$ is significantly facilitated. The multi-head attention mechanism of the Transformer is enhanced by leveraging the attention mechanism. An overview of the attention block's structure $A^2B$ is shown in Fig. 2.

In general, the attention block $A^2B$ In the subsequent experiments, we also verified that the attention block $A^2B$ can effectively improve the index and accuracy of image segmentation.

## 4. Experimentation

### 4.1. Data Set

The Transformer model incorporates many self-attention methods and a substantial number of parameters, necessitating significant computational resources and memory all through both training and inference stages. The size of the data set is important for Transformer. By processing CT slices instead of the entire CT sequence, the size of the data set can be expanded. MLP-based Transformer takes up a lot of graphics storage space. Therefore, Transformer does not significantly increase the size of the weight file, so it is more suitable for 2D images. Therefore, when dealing with CT slices in the experiment, we segment the whole image according to the approximate maximum absolute value of CT slices and make 420 samples. The training set consists of 95% (400) of the samples, whereas the test set comprises the remaining 5% (20) of the data. Fig. 3 illustrates a collection of the training sets and test sets.

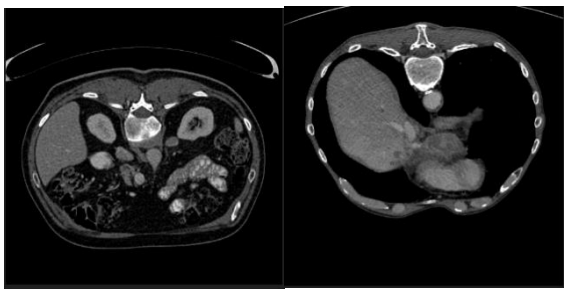


**Figure 3.** Examples of training and test data

### 4.2. Experiment Setting

For encoders based solely on Transformers, we utilize ViT with 12 Transformer layers. In the case of hybrid encoder design, ResNetV2 and ViT are combined; they are referred to as "RV2-ViT" in this work. Unless otherwise stated, Both the feedback resolution and patch size P are configured as

224x224 and 16, respectively. So as to get the maximum level of detail, we use a series of four 2-fold Upsampling blocks consecutively inside the Contextual Upsampling Pipeline (CUP). The model is trained using the Stochastic Gradient Descent (SGD) optimizer, including a learning rate of 0.01, a momentum of 0.9, and a weight decay of 1 e-4. The default batch size is predetermined to be 24, and the number of training iterations is fixed at 100 generations. The tests conducted in this publication are run on 16.0 GB RAM, 64-bit operating system, Intel (R) Core (TM) i7-10750 H CPU @ 2.60 GHz 2.59 GHz processor based on x64, GPU of NVIDIA GeForce RTX 2060.

### 4.3. Ablation Experiment

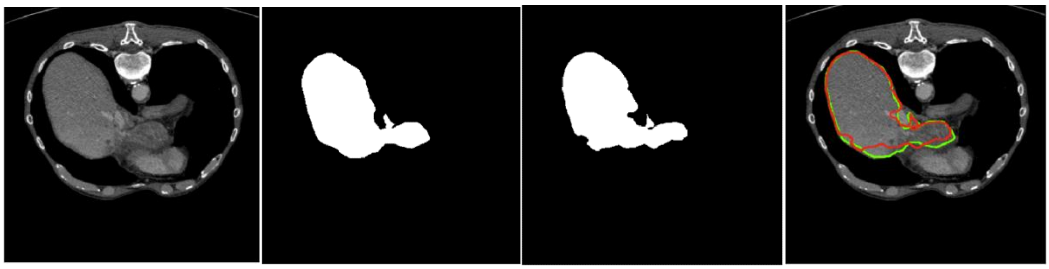
To ascertain the efficacy of the hypothesised Transformer + CNN model in enhancing network performance, and to evaluate the impact of including the A2B attention mechanism module within the Transformer encoder on network performance. In this paper, the ablation experiment of this section is set up. The experiment is carried out in the same environment. The findings obtained from the experimental study are as follows in Tabel 1. According to the data, we can know that in the A2B+Transformer+CNN model, a variety of indicators have reached the optimal results, among which the performance in IoU has increased by 0.0116, which has achieved great improvement. Based on the empirical evidence, it can be inferred that the model indicated in this research study attains the highest possible standard of performance.

**Table 1.** Results of ablation experiment (the best data is blacked out)

| Model | Dice | IoU | Acc | Recall | Pre |
|---|---|---|---|---|---|
| $A^2B$+Transformer+CNN | **0.9897** | **0.9796** | **0.9992** | **0.9877** | 0.9917 |
| Transformer+CNN | 0.9836 | 0.9680 | 0.9988 | 0.9750 | **0.9926** |
| CNN | 0.9805 | 0.9617 | 0.9984 | 0.9719 | 0.9892 |

### 4.4. Visual Analysis

To qualitatively assess the effectiveness of this paper's method for segmenting medical images, we visualized different segmentation results, as shown in Fig. 4. Fig. 4a is the original image to be processed cut out from the CT image. Fig. 4b is the ground truth label while Fig. 4c is the segmented prediction of our proposed method. We also compare the disparity between the ground truth consequence and our projected consequence result in Fig. 4d, where the green range is the standard result and the red range is the prediction target. Through the visual image, we can intuitively see that the difference between the trained model segmentation result and the true value is small.



(a) Input (b) Ground truth (c) Prediction (d) Comparison

**Figure 4.** Visualization of segmentation results

## 5. Conclusion

U-Net has emerged as a widely used and very successful approach in several medical picture segmentation tasks. However, the U-Net architecture sometimes exhibits constraints in effectively capturing distant relationships, mostly attributed to the inherent spatial proximity of convolutional operations. The transformer architecture has gained prominence as a model created for sequence-to-sequence prediction, offering an intrinsic global self-attention mechanism. Nevertheless, the restricted positioning ability may be attributed to the absence of comprehensive low-level details.

Based on the segmentation task of medical images in the TransUnet backbone network, the $A^2B$ (Attention in Attention Block) attention mechanism is connected at the jump, and the advantages of TransUnet and $A^2B$ are combined to strengthen the accuracy and robustness of the segmentation results. By combining these two models, doctors and researchers can more accurately identify and locate key structures.

## Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

## References

[1] Parmar, N., A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran. "Image Transformer." In Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholm, SWEDEN, Jul 10-15 2018.

[2] Yuan Meifang, Yang Yi, Zhao Biao et al. Research on automatic segmentation of thyroid based on dense residual U-net neural network in localized CT images [J]. Beijing Biomedical Engineering, 2022, 41 (01): 42-48.

[3] Zhang Xiaoyue, Wang Yongxiong, Zhang Jiapeng, etc. Early gastric cancer screening algorithm based on activation layer pre-compressed excitation residual network [J]. Journal of Southern Medical University, 2021, 41 (11): 1616-1622.

[4] Wang Fang, Qiao Ruiping. SPAM: Spatial segmentation attention module in deep convolutional neural network for image classification [J / OL]. Journal of Xi'an Jiaotong University, 2023 (09): 1-8 [2023-08-30].

[5] Fang Yunhua. Research on the neural mechanism of the correlation between visual attention function and TCM constitution and age based on functional magnetic resonance technology [D]. Fujian University of Traditional Chinese Medicine, 2017.

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7132–7141, 2018.

[7] Liu Kewen, Ma Yuan, Xiong Hongxia, etc. Medical image super-resolution reconstruction method based on residual channel attention network [J]. Progress in laser and optoelectronics, 2020, 57 (02): 161-169.

[8] Wang Jing.Research on attention mechanism for pathological image super-resolution [D]. Beijing Jiaotong University, 2022.

[9] Wang Dongfei. Application of convolutional neural network based on channel attention in image super-resolution reconstruction [J]. Broadcast and television technology, 2018, 45 (06): 63-66.

[10] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In Proceedings ofthe IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11030–11039, 2020.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin.Attention Is All You Need.arXiv, 2017.

[12] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth16x16 words: Transformers for image recognition at scale. In: ICLR, 2021.

[13] Wang Jinxiang, Fu Lijun, Yin Pengbin, etc. Medical image segmentation based on CNN and Transformer [J]. Computer system application, 2023, 32 (04): 141-148.