Comparative Analysis of Transformer Integration in U-net Networks for Enhanced Medical Image Segmentation

Zixin Hao*

Faculty of Engineering and IT, University of Melbourne, Melbourne, Australia

*Corresponding author: zixhao1@student.unimelb.edu.au

Abstract. Transformer is popular in Natural Language Processing (NLP) and is a cornerstone of large models. Transformer has been used by researchers to address the limitations of Convolutional Neural Networks (CNNs) in medical picture segmentation models. Through an extensive literature review and case studies, this paper comparatively analyzes the performance of different models in this field, summarizes different methods of integrating transformers into U-net, and points out existing gaps and challenges. Research has found that the Transformer model can significantly improve the accuracy and efficiency of medical image analysis. The paper discusses the advantages, disadvantages, innovations, performance, and complexity of various models in detail, and shows how to enhance performance by integrating the Transformer structure into the U-net network. In particular, the paper also analyzes the advantages of Transformers that are most suitable for integration into the encoder part and highlights the balance that needs to be made between improving performance and computational cost. The conclusion shows that although there is no perfect model, optimal performance and efficiency can be achieved by selecting different combinations of Transformer and U-net according to the actual situation. It can be seen from the networks' performance that the mixed use of a U-shaped convolutional network and Transformer module has good development prospects and high research significance.

Keywords: Transformer, U-net, Medical image segmentation, Computer vision.

1. Introduction

Modern healthcare is based heavily on medical imaging, which provides vital information about a wide range of illnesses and ailments [1]. It is crucial to be able to interpret these images correctly since they have a direct impact on the effectiveness of treatment plans and diagnoses. In this context, the role of image segmentation becomes particularly significant.

Image segmentation is a key process in computer vision. It splits a digital image into several segments or pixel groups. This makes the image simpler and transforms it into a format that's easier to understand and analyze. It is a critical step in image processing that enables the identification of objects and boundaries within images. So it is critical in medical imaging because it enables the precise identification and separation of different structures and tissues in the image, thus providing critical information for the diagnosis and treatment planning of diseases. For example, in magnetic resonance imaging (MRI) or computed tomography (CT) images, medical segmentation helps identify tumors, lesions, or other abnormal structures.

Historically, a variety of methodologies, from straightforward thresholding techniques to more intricate clustering-based algorithms, have been used to drive image segmentation [2]. Early systems' shortcomings were usually ascribed to their reliance on threshold-based methods or manual intervention, which required far more expert knowledge beforehand and were not adaptable to illumination or image quality variations.

With the advent of machine learning, and particularly deep learning, the domain of image segmentation experienced a notable shift and evolution. CNN became the cornerstone of modern image segmentation techniques, especially in complex scenarios like medical imaging. Deep learning algorithms automatically extract features, overcoming the need for expert medical knowledge in traditional image segmentation. They are also flexible and can be adapted to different tasks using transfer learning.

The U-net architecture, introduced in 2015, holds the highest popularity in medical picture segmentation tasks [3]. Numerous enhanced iterations of the U-shaped network have demonstrated strong performance in numerous medical picture segmentation tasks since its proposal. To make up for the shortcomings of convolutional networks, some scholars thought of introducing transformers that have made breakthroughs in the field of NLP, so Vision Transformer (ViT) came into being [4]. The ViT has revolutionized computer vision by overcoming some limitations of traditional CNNs. ViT excels in capturing long-range dependencies in images and dynamically adjusts its receptive field, enhancing image processing efficiency. Its parameter efficiency and generalization capabilities across various tasks mark a significant advancement. Following ViT, notable variants like the Swin Transformer and Data-efficient Image Transformers (DeiT) have emerged [5, 6]. Swin Transformer reduces computational demands while maintaining long-range dependency capture, and DeiT shows ViT's effectiveness on smaller datasets. These developments in ViT and its adaptations demonstrate great potential, especially in high-precision fields like medical image analysis.

The impressive features of ViT have caught the medical field's interest, leading to many new variations. With so many developments, it's hard to stay updated. Thus, exploring the newest Transformer-based U-net networks for medical image segmentation is important.

The research approach includes a thorough review of current literature and case studies, with a focus on the latest advancements in the field. This paper will compare different models to help readers understand the strengths and weaknesses of the latest applications of Transformers in U-net networks. This study aims to summarize the present state of research, pinpoint existing gaps and challenges, and emphasize how these models could enhance the precision and efficiency of medical image analysis. This will provide readers with valuable insights into integrating U-net architectures with Transformer structures.

2. U-net and Transformer

This section briefly introduces the structures and characteristics of the U-net and the Transformer network, which are considered the cornerstone for future research in the field.

U-net is a neural network framework designed for image segmentation tasks. It maintains spatial accuracy and achieves high accuracy in segmentation. On the other hand, U-net is suitable for processing small data sets, making it effective for medical image segmentation. So, this network serves as the cornerstone for most subsequent research in medical segmentation. Because of its U-shaped design, U-net generates segmentation maps by using a symmetric decoder for segmentation map production and an encoder for feature extraction. Known for its skip connections, it retains older characteristics to reveal finer details.

Transformer is a revolutionary deep learning model that improves information processing and the capture of long-distance dependencies by using attention mechanisms to process sequence data [4]. Originally applied in the field of NLP, it has become the most influential model in this field and the cornerstone of large models such as GPT-3. The transformer uses an encoder-decoder structure, including a self-attention mechanism, feed-forward neural network, and residual connection. Multi-head attention enables the model to simultaneously concentrate on various locations, making it excellent at handling long-distance dependencies and sequence tasks.

3. Transformer-based U-network

Transformers and U-net (TransUNet) for medical image segmentation were introduced in the aftermath of Transformer's NLP achievements and the successful use of ViT for image classification [7]. The transformer's suitability for handling moderately small medical datasets has been refined via research and development. One popular tactic is to combine Transformers with U-shaped networks. Combining the best features of both architectures, this combination opens up a promising new direction for medical picture analysis.

In this section, this study will conduct a comparative analysis and classification of significant research based on this network architecture, focusing on studies published after 2022.

3.1. Transformer in Encoder

TransUNet is the first U-shaped network that applies a Transformer to the field of medical image segmentation. To extract glocal context, TransUNet gets the input sequence by encoding the tokenized image patches. The encoded features are upsampled by the decoder and combined with the CNN feature maps at high resolution. This structure takes advantage of Transformer's global self-attention mechanism and solves U-net's shortcomings in modeling long-range dependency. However, the positioning capability is limited, and the segmentation efficiency needs to be improved.

An Efficient Hierarchical Hybrid Transformer by He et al. introduces a novel deep learning model that effectively combines CNNs and Transformers [8]. It features a hierarchical hybrid structure with multi-scale channel attention, efficiently capturing both local and long-range features. This model excels in computational efficiency and is particularly effective for medical image segmentation tasks, demonstrating superior performance on various 2D and 3D medical imaging tasks while maintaining lower computational complexity.

3.2. Transformer in Encoder and Decoder

The PCAT-UNet architecture, designed for retinal vessel segmentation, effectively merges the advantages of Convolutional Neural Networks (CNNs) and Transformers [9]. It comprises an encoder-decoder structure, convolutional branches, skip connections, and side output layers. The key components of this architecture are the PCAT block and the FGAM (Feature Grouping Attention Module). The PCAT block, or Patch Convolution Attention Transformer, integrates CNN's local feature extraction with the Transformer's global information capture. It utilizes Cross Patch Convolutional Attention (CPCA) and Intra Patch Convolutional Attention (IPCA) for multi-scale feature extraction. The FGAM enhances channel feature maps by extracting varied scale features using small-scale group convolutions. This module is instrumental in merging local and global channel attention effectively.

A study proposed D-Former, an innovative dilated Transformer-based framework for 3D medical picture segmentation [10]. It employs a U-shaped encoder-decoder design, combining Global Scope Modules (GSM) and Local Scope Modules (LSM), effectively exploring long-range dependencies by expanding the attention range. D-Former also incorporates dynamic position encoding to flexibly learn important positional information within the input sequence, reducing model parameters and computational complexity. Experimental results on the Synapse and ACDC datasets demonstrate that D-Former achieves state-of-the-art semantic segmentation performance while reducing computational costs.

3.3. Transformer in Other Places

3.3.1. Transition

Multiscale transunet ++ (MS-TransUNet++) is a dense hybrid model that combines CNN and Transformer, mainly composed of the encoder, intermediate nodes, and decoder [11]. The model first uses CNN to extract features, then maps these features into sequential patterns through positional encoding in the final stage of the encoder, taking advantage of the long-range dependence and adaptability of the Transformer. The decoder adopts a pure CNN structure and upsamples the feature map through stepwise transposed convolution. In addition, as shown in Fig. 1, MS-TransUNet++ adopts nested and dense skip connections, combined with multi-scale combination strategies (including 1×1 and 3×3 convolutions) to capture more diverse local features. WANG B and DONG P also designed a new loss function, and experiments proved that this method achieved more accurate medical image segmentation. Relative to the fixed ratio in standard TransUNet, MS-TransUNet++'s positional encoding and input tokens are better suited to dense prediction tasks.

3.3.2. Output block

Towards Boundary-Aware Lightweight Transformer (BATFormer) introduced by Lin et al., a novel transformer model addressing the limitations of conventional CNNs and transformers in medical image segmentation [12]. The model is distinctive for its boundary-aware design and computational efficiency. It features a Cross-Scale Global Transformer (CGT) module for capturing global dependencies across multiple small-scale features, thus enhancing long-range dependence while consuming less computing power. Additionally, the Boundary-Aware Local Transformer (BLT) module in BATFormer flexibly creates windows surrounding the edges of objects for precise shape preservation. This unique approach to dynamic and adaptive window partitioning is a pioneering effort in the field. BATFormer outperforms current methods based on convolutional neural networks and those using transformer or mixed models in analyzing a range of 2D and 3D medical imaging datasets that are openly accessible.

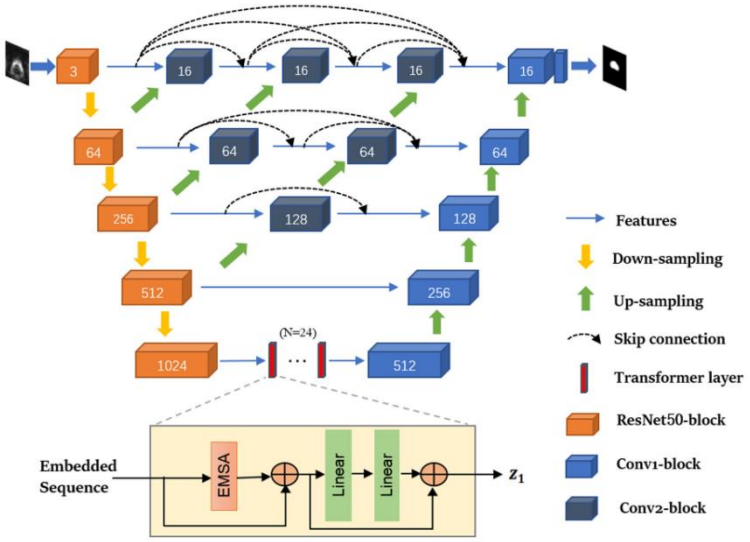

Fig. 1 The overview architecture of MS-TransUNet + + [11]

As shown in Table 1, summarizes the main tasks faced by the above model, the data set used and the position of the Transformer integrated into U-net.

Table 1. Overview of models based on transformers for segmenting medical images

Transformer Location	Model	Segmentation Task	Dataset	Year
Encoder	TransUNet[7]	Abdominal multiple organ or heart segmentation	BCV[13]/ACDC[14]	2021
	H2Former[8]	Multiple tasks	ACDC[14] / IDRiD[15]/ Kvasir-SEG[16] / Skin Lesion[17] / Synapse[18]	2023
Encoder and Decoder	PCAT-UNet[9]	Retinal vascular segmentation	DRIVE[19]/STARE[20]/ CHASE_DB1[21]	2022
	D-Former[10]	Abdominal multiple organ or heart segmentation	BCV[13]/ACDC[14]	2023
Transition	MS-TransUnet++[11]	Prostate or liver tumour segmentation	PROMISE12[22]/LiTS[23]	2022
Outside Module	BATFormer[12]	Heart segementation	ACDC[14] / ISIC 2018[24]	2023

4. Discussion

This paper has introduced some recent researches in the last section and categorized these networks based on the position in the transformer is applied.

There is a notable trend in segmentation models, with a clear preference for incorporating the Transformer predominantly in the encoder position, as opposed to other positions. The possible reason is that the advantages of the transformer are most suitable to be integrated into the features of the encoder, and it is difficult to take advantage of it in other positions to capture contextual connections. Incorporating a transformer into both the encoder and decoder segments of a model can lead to increased complexity and reduced efficiency. The dimensionality of the feature map at the transition position is the lowest, so the multi-layer Transformer will not have a large load, but the feature extraction and fusion capabilities are limited. In general, any improvement has its advantages and disadvantages. We should start from the specific tasks and choose the corresponding method according to the required characteristics to integrate the transformer into U-Net.

Table 2 is a comparative analysis of different 3D medical image segmentation models on the Automated Cardiac Diagnosis Challenge (ACDC) dataset, with the Dice score being the criterion for evaluation. The dataset is a benchmark in medical image analysis that provides cardiac MRI scans for the automated assessment of cardiac structures and functions. It is instrumental for analyzing specific heart components such as the right ventricle (RV), myocardium (MYO), and left ventricle (LV). In this context, the Dice coefficient serves as the evaluation matrix, gauging the similarity between segmentation results and the ground truth. Its value ranges from 0 to 1, with 1 denoting perfect overlap and thus, ideal segmentation accuracy.

Table 2. Performance comparison of different segmentation methods on ACDC. The evaluation matrix is DICE Score

Models	Average	RV	MYO	LV
TransUNet	89.71	88.86	84.54	95.73
BATFormer	92.80	92.55	90.55	95.30
H2Former	92.40	91.31	90.12	95.76
D-Former	92.29	91.33	89.60	95.93

From Table 2, the contenders include TransUNet, BATFormer, H2Former, and D-Former. BATFormer emerges as the frontrunner in terms of average Dice scores, closely followed by H2Former, while D-Former trails slightly behind, though the margin is minimal.

Delving into the segmentation of distinct cardiac structures, D-Former shows performance on par with BATFormer in delineating the RV and LV but exhibits a slight lag in segmenting the MYO. This insight leads to the conclusion that while D-Former may not top the list in average Dice scores, its ability to segment specific cardiac structures is on a level playing field with the leading models, marking its efficacy in the domain of medical image segmentation.

In addition to focusing on the performance differences among models for different characters, researchers should also focus on the complexity of the models and seek a balance between them. In Table 3, the paper compares the parameter numbers and floating point operations (FLOPs) among various methods.

Table 3. Analysis of parameters and FLOPs for TransUNet, H2Former, and D-Former.

Models	Parameters	FLOPs
	Input size of 512 × 512	
TransUNet	109.54M	56.66G
H2Former	33.71M	33.56G
	Input size of 64 × 128 × 128	
D-Former	44.26M	54.46G

TransUNet (Encoder) has the highest number of parameters at 109.54M, which suggests it is the most complex model in terms of the capacity to learn from data. It also has the highest FLOPs at 56.66G, indicating it requires the most computation for a single forward pass. However, H2Former (Encoder) presents significantly fewer parameters at 33.71M compared to TransUNet, which may imply a less complex model with potentially faster inference times due to lower computational requirements, evidenced by its FLOPs of 33.56G. D-Former (Encoder and Decoder), despite its 3D input, has 44.26M parameters, which is more than H2Former but less than TransUNet. Its FLOPs are 54.46G, which is close to TransUNet, suggesting a higher computational cost than H2Former but comparable to TransUNet, potentially due to its 3D processing capability.

There is a trade-off between model complexity (number of parameters) and computational efficiency (FLOPs). TransUNet, while being the most complex, may not be the most efficient in terms of computation. H2Former seems to offer a balance between complexity and efficiency. D-Former, although it deals with 3D data, manages to maintain a moderate number of parameters and a high computational demand, highlighting the additional complexity introduced by 3D image processing.

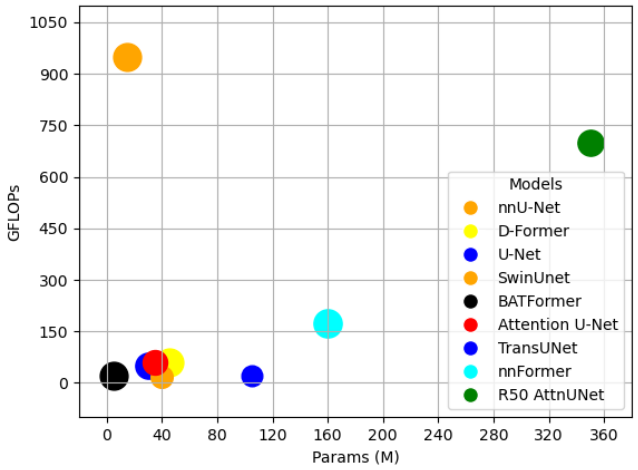

Fig. 2 Comparison of model parameters and computational complexity across various models on the ACDC dataset [12]

In Fig. 2, the diameter of each circle represents the performance level of the respective model, with larger circles denoting better performance. "Params" refers to the number of parameters within the model, quantified in millions (M). "FLOPs" stands for floating-point operations per second, a metric used to gauge the computational complexity of models, and "GFLOPs" is an abbreviation for Giga Floating-point Operations per Second [12].

This paper analyzes the effectiveness of models based on CNN in comparison with those utilizing transformer and hybrid methods, to assess the deployment feasibility of both approaches. Using the ACDC dataset, the graph compares the computational complexity and performance of many machine learning models. Every circle is a model, and the bigger the circle, the better the model's performance is shown by its size. With a label of "Params (M)," the horizontal axis displays the total number of model parameters in millions. The computational complexity of the model is represented by the vertical axis, "GFLOPs," which is expressed in giga floating-point operations per second. There are models with fewer parameters, such as nnU-Net and R50 AttUNet, that show high GFLOPs, indicating better performance. Conversely, models such as TransUNet have fewer GFLOPs, potentially indicating less efficiency, even with a higher number of parameters.

5. Conclusion

This study focuses on the combined application of Transformer and U-shaped network in the field of medical image segmentation. It is found that this hybrid framework can address the challenge of

dispersed focal regions and significant variances in shapes, thereby improving segmentation performance. In particular, combining the multi-scale local spatial feature extraction of the U-shaped network with the global information processing of the Transformer is conducive to achieving improved equilibrium in performance for tasks involving the segmentation of medical images. Since the scale of medical image data sets is generally small, it is often difficult to use a Transformer alone to maximize its advantages, but integrating it into a U-shaped network structure can make full use of limited sample information. Integrating Transformers in different ways will have different effects. No model is perfect. It is necessary to balance performance and complexity when researchers select the most appropriate model for a specific task.

This review provides a clear and cutting-edge understanding of research in this field by analyzing the application of the latest Transformer and U-shaped networks in medical image segmentation. It provides a reference for subsequent researchers in this field.

In terms of prospects for future research, although current research has made certain progress in the field of medical image segmentation, there are still some shortcomings. For example, use the performance of a Transformer on large data sets to explore semi-supervised or unsupervised learning methods to improve the processing efficiency and quality of medical images; Secondly, how to further improve processing efficiency and accuracy while keeping the complexity of the model under control, remains a challenge. The research selected in this article is relatively new, but the scope and number of studies need to be broadened and improved to obtain a more comprehensive perspective.

References

[1] Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized medical imaging and graphics, 2007, 31(4-5): 198-211.

[2] Patil D D, Deore S G. Medical image segmentation: a review. International Journal of Computer Science and Mobile Computing, 2013, 2(1): 22-27.

[3] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.

[5] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[6] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention. International conference on machine learning. PMLR, 2021: 10347-10357.

[7] Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306, 2021.

[8] He A, Wang K, Li T, et al. H2Former: An Efficient Hierarchical Hybrid Transformer for Medical Image Segmentation. IEEE Transactions on Medical Imaging, 2023.

[9] Chen D, Yang W, Wang L, et al. PCAT-UNet: UNet-like network fused convolution and transformer for retinal vessel segmentation. PloS one, 2022, 17(1): e0262689.

[10] Wu Y, Liao K, Chen J, et al. D-former: A u-shaped dilated transformer for 3d medical image segmentation. Neural Computing and Applications, 2023, 35(2): 1931-1944.

[11] Wang B, Wang F, Dong P, et al. Multiscale transunet++: dense hybrid u-net with transformer for medical image segmentation. Signal, Image and Video Processing, 2022, 16(6): 1607-1614.

[12] Lin X, Yu L, Cheng K T, et al. BATFormer: Towards Boundary-Aware Lightweight Transformer for Efficient Medical Image Segmentation. IEEE Journal of Biomedical and Health Informatics, 2023.

[13] Sheth I, Braga P H M, Sujit S, et al. RelationalUNet for Image Segmentation. International Workshop on Machine Learning in Medical Imaging. Cham: Springer Nature Switzerland, 2023: 320-329.

[14] Bernard O, Lalande A, Zotti C, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?. IEEE transactions on medical imaging, 2018, 37(11): 2514-2525.

[15] Porwal P, Pachade S, Kamble R, et al. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. Data, 2018, 3(3): 25.

[16] Jha D, Smedsrud P H, Riegler M A, et al. Kvasir-seg: A segmented polyp dataset. MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26. Springer International Publishing, 2020: 451-462.

[17] Gutman D, Codella N C F, Celebi E, et al. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1605.01397, 2016.

[18] Landman B, Xu Z, Igelsias J, et al. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. 2015, 5: 12.

[19] Staal J, Abràmoff M D, Niemeijer M, et al. Ridge-based vessel segmentation in color images of the retina. IEEE transactions on medical imaging, 2004, 23(4): 501-509.

[20] Hoover A D, Kouznetsova V, Goldbaum M. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Transactions on Medical imaging, 2000, 19(3): 203-210.

[21] Owen C G, Rudnicka A R, Mullen R, et al. Measuring retinal vessel tortuosity in 10-year-old children: validation of the computer-assisted image analysis of the retina (CAIAR) program. Investigative ophthalmology & visual science, 2009, 50(5): 2004-2010.

[22] Litjens G, Toth R, Van De Ven W, et al. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Medical image analysis, 2014, 18(2): 359-373.

[23] Bilic P, Christ P, Li H B, et al. The liver tumor segmentation benchmark (lits). Medical Image Analysis, 2023, 84: 102680.

[24] Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data, 2018, 5(1): 1-9.