Medical Image Segmentation using Transformer and Unet Architecture

SADIAH AlANSAR (id:21-45612-3)a, SYEDA SARAH FERDOUS (id: 20-44176-2)a,

SADAF AKHTER (id: x21 -44658-1)a, A. S. M. SAYEM (id: 20-44115-2)a

*aDepartment of Computer Sciences, American International University-Bangladesh*

# Abstract

Medical image segmentation is a challenging task in healthcare, but significant progress has been achieved by leveraging deep learning techniques. Hybrid model integration, and deep learning's versatility as driving advancements in medical image segmentation and it is important to improve accuracy, generalization, and applicability of medical image segmentation methods. In this paper, we represent TransFuse architecture [4], which combines CNNs and Transformers. Transformers' role in medical image segmentation like TransUNet [5], and Vision Transformers (ViT) [6] are discussed. The CNN branch employs U-Net for local-global feature extraction. BiFusion module is used also to selectively fuses features through a series of steps, incorporating local and global information. We believe that using the complete details of the input image could help the model to identify important and distinct characteristics more effectively. This could result in a better segmentation outcome.

*Keywords:* ***Deep learning, Computer Vision, Object detection, NN, CNN***

# Introduction

In recent years, the advancement of medical image analysis has witnessed the merging of deep learning techniques, opening a new way for accurate and efficient image segmentation. One of them is integration of Transformer architectures and Convolutional Neural Networks (CNNs). Merging them is used to tackle the complex challenges inherent in medical image segmentation tasks. The general objective of our paper is to improve medical image segmentation tasks. With the use of Transformer and Unet architecture the medical image segmentation task is more better and refined. More specific objective of the paper is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze by the use of global location and context at the same time. Which generate the question “How is this going to perform better than ResNet for image segmentation?”

U-Net helps in retaining the spatial features of the image in better segmentation results. The merger of Transformer architectures and CNNs, enables the integration of both structural and contextual information for enhanced accuracy. Meanwhile, the classical “U-Net" architecture remains a foundation in the field, demonstrating the effectiveness of changed design principles in addressing challenges unique to biomedical image segmentation. It shows how a fusion of cutting-edge technologies and foundational architectures can drive progress and empower accurate diagnosis and treatment in healthcare.

# Literature Review

The segmentation of medical images is a difficult task in modern healthcare. manipulating the power of deep learning, some progress has been achieved in enhancing the efficiency of medical image segmentation. This comprehensive review focuses on providing an in-depth analysis of powerful papers in this field, highlighting their achievements, methodologies, and the challenges they address.

The U-Net architecture introduced by Ronneberger et al. has become a foundation in medical image segmentation [1]. It employs a shrink-expanding framework along with skip connections, U-Net captures both global context and intricate local details. The paper focuses on its use in biomedical image analysis and its impact on addressing segmentation challenges, especially with limited training data.

Bi et al.'s innovative approach for dermoscopic image segmentation utilizes class-specific learning through step-wise integration [2]. This strategy effectively addresses the class imbalance challenge by repetitively focusing on different lesion categories. In order to achieve correct segmentation of different skin lesions, the study highlights the importance of progressive refinement.

Al-masni et al. tackle skin lesion segmentation by maintaining full-resolution information throughout their deep convolutional network [3]. This unique approach enables the capture of intricate lesion details critical for accurate diagnosis. The paper highlights the architectural decisions essential to solve medical image segmentation complexities.

Zhang et al.'s TransFuse addresses a fusion of Transformers and Convolutional Neural Networks (CNNs) to enhance medical image segmentation [4]. By teaming the global contextual awareness of Transformers and the local feature extraction capabilities of CNNs, TransFuse achieves higher segmentation accuracy. The paper makes progress in the concept of hybrid architectures in medical imaging.

Chen et al.'s TransUNet employs Transformers as encoders for medical image segmentation, manipulating their ability to capture long-range dependencies [5]. This integration enhances contextual understanding, complementing standard encoders. The paper also demonstrates Transformers' effectiveness and adaptability in addressing medical image segmentation challenges.

Dosovitskiy et al. introduce the Vision Transformer (ViT) architecture, demonstrating the use of Transformers in image recognition tasks [6]. While not being limited to on medical images, this work has sparked a surge of interest in applying Transformers to various visual tasks, including medical image segmentation.

Isensee et al. present an automated approach to designing deep learning methods for biomedical image segmentation [7]. By improving architecture and hyperparameters, the model development has rather been streamlined by this method. The paper highlights the significance of adapted architecture design in achieving optimal segmentation performance.

The papers that have been reviewed indicates the remarkable advancements has been provided in medical image segmentation through deep learning techniques. The accomplishments that have been made highlight the significance of architectural innovation, hybrid model integration, and the versatility of deep learning paradigms. Meanwhile other substantial progress, challenges such as data scarcity, domain adaptation, and robustness to variations persist has been achieved. In the future, research will certainly build upon these foundations to further maximize the accuracy, generalization, and applicability of medical image segmentation methods.

# Proposed Method

**A diagram of a block diagram

Description automatically generated**

The architecture, TransFuse proposed by [4] is designed to combine the capabilities of Convolutional Neural Networks (CNNs) and Transformers for enhanced medical image segmentation. As illustrated in Figure 1, TransFuse comprises two parallel branches: the CNN branch and the Transformer branch, which are later combined using a BiFusion module.

In the original TransFuse architecture, the authors used a combination of Resnet 34 and Resnet 50 for the CNN branch. The authors of this paper proposed a change to the use of U-net in the CNN branch. A U-Net architecture focuses on gradually increasing the receptive field and encoding features from local to global levels. It is particularly effective for retaining local details while capturing broader context. Features extracted from the U-Net branch are preserved for later fusion [1].

The Transformer branch follows the classic encoder-decoder architecture. The input medical images are divided into patches which are linearly embedded and enhanced with positional embeddings to incorporate spatial information. The Transformer encoder, consisting of multiheaded self-attention (MSA) and Multi-layer Perceptron (MLP) layers, processes these patches to capture global context. Progressive upsampling is then employed to reconstruct feature maps of various scales [4].

The main innovation of TransFuse lies in the proposed BiFusion Module. This module selectively fuses the features extracted from both the CNN and Transformer branches. It combines self-attention and multi-modal fusion mechanisms, effectively integrating local and global information. The fused feature representations are generated using a series of steps [4].

Firstly, a channel attention mechanism is applied to enhance global information from the Transformer branch. Then the U-net features are processed through a convolution operation to capture multi-scale contextual information. Spatial Attention filters are then used to emphasize local details and suppress noise in low-level CNN features. The fine-grained interactions between U-Net and Transformer features are modeled using the Hadamard product. The Hadamard product, also known as the element-wise or entry-wise product, is a mathematical operation defined on matrices or vectors of the same dimensions.

After this, the interaction features are combined with the attended features using a Residual block, which in turn effectively captures global and local contexts. To generate the final segmentation, the fused features from each branch are combined using the attention-gated skip-connection (AG) mechanism. This results in a comprehensive and refined feature representation that is used for segmentation prediction.

For the loss function, the network is trained end-to-end using a combination of the weighted Intersection over Union (IoU) loss and binary cross-entropy loss. Boundary pixels are assigned larger weights to enhance boundary delineation. The segmentation prediction is generated by a simplified head that resizes feature maps to the original resolution and applies convolution layers to produce class-specific segmentation maps [4].

Deep supervision is employed by additionally supervising the Transformer branch and the first fusion branch, as outlined in previous works. This approach helps improve gradient flow during training, enhancing convergence and segmentation accuracy.

# Results

# TO EVALUATE THE EFFECTIVENESS OF TRANSFUSE, THE AUTHORS OF THIS PAPER USED THE CVC-CLINICDB DATASET WHICH WAS AVAILABLE FOR PUBLIC USE ON KAGGLE. FOR PRE-PROCESSING, THE RESOLUTION OF THE TRAINING IMAGES IS RESIZED TO 352X352. TRANSFUSE WAS BUILT IN THE PYTORCH FRAMEWORK AND TRAINED USING THE FREE T4 GPU OF GOOGLE COLLABORATORY. FOR COMPILING THE MODEL, THE ADAM OPTIMIZER IS USED WITH A LEARNING RATE OF 1E-4. THE MODEL IS TRAINED FOR 30 EPOCHS WITH A BATCH SIZE OF 16 FOR EACH EPOCH. AS DIFFERENT MEDICAL IMAGE SEGMENTATION TASKS SERVE DIFFERENT DIAGNOSES OR OPERATIVE PURPOSES, THE AUTHORS FOLLOW THE COMMONLY USED EVALUATION METRICS FOR EACH OF THE SEGMENTATION TASKS TO QUANTITATIVELY ANALYZE THE RESULTS.

# 

# Discussion

TransFuse is a method which is invented for segmenting medical images that combines different types of neural networks. It uses a U-Net architecture to capture both small and large details in the images, while also using a Transformer architecture to process the images as a whole.The unique part of TransFuse is the BiFusion module, which combines features from both types of networks. This module uses self-attention and fusion mechanisms to bring together local and global information. It also uses a Residual block to capture the overall context of the images. Another important part of TransFuse is the attention-gated skip-connection mechanism, which helps to refine the features for better segmentation predictions.During training, TransFuse uses a combination of weighted Intersection over Union (IoU) loss and binary cross-entropy loss. It also uses deep supervision to improve the flow of information through the networks. Here we have use CVC-ClinicDB dataset for segmentation. The model achieved promising results in terms of segmentation accuracy, as per the commonly used evaluation metrics for medical image segmentation tasks.

# 

# Conclusion

In this work, we presented a new TransFuse method which combine with U-Net and Convolutional Neural network method and BiFusion module for Medical image segmentation. The integration of CNNs and Transformers in the TransFuse architecture has shown promising results for medical image segmentation. The use of U-Net allows for the preservation of local details and capturing broader context, while the Transformer branch captures global context. The innovation of the BiFusion module effectively combines local and global information through self-attention and multi-modal fusion mechanisms. Overall, TransFuse presents a novel approach for enhancing the accuracy and generalization of medical image segmentation methods.

# References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.

[2] L. Bi, J. Kim, E. Ahn, A. Kumar, D. Feng, M. Fulham, "Step-wise Integration of Deep Class-Specific Learning for Dermoscopic Image Segmentation," in IEEE Transactions on Medical Imaging, vol. 39, no. 6, pp. 2016-2026, 2020.

[3] M. A. Al-masni, M. A. Al-antari, M.-T. Choi, S.-M. Han, T.-S. Kim, "Skin Lesion Segmentation in Dermoscopy Images via Deep Full Resolution Convolutional Networks," in IEEE Access, vol. 8, pp. 91560-91571, 2020.

[4] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation," in IEEE Transactions on Medical Imaging, vol. 40, no. 4, pp. 1395-1405, 2021.

[5] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," in IEEE Transactions on Medical Imaging, vol. 40, no. 10, pp. 2839-2850, 2021.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proceedings of the International Conference on Learning Representations (ICLR), 2021.

[7] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Automated Design of Deep Learning Methods for Biomedical Image Segmentation," in Nature Communications, vol. 12, no. 1, pp. 2917, 2021.

[8] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges," in IEEE Access, vol. 9, pp. 18629-18645, 2021.

# Contribution

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Sadiah Alansar** | **SYEDA SARAH FERDOUS** | **SADAF AKHTER** | **A. S. M. SAYEM** | **Contribution (%)** |
|  | *21-45612-3* | *20-44176-2* | *21-44658-1* | *20-44115-2* |
| Conceptualization | 25% | 25% | 25% | 25% | 100 % |
| Data curation | 25% | 25% | 25% | 25% | 100 % |
| Formal analysis | 25% | 25% | 25% | 25% | 100 % |
| Investigation | 25% | 25% | 25% | 25% | 100 % |
| Methodology | 25% | 25% | 25% | 25% | 100 % |
| Implementation | 25% | 25% | 25% | 25% | 100 % |
| Validation | 25% | 25% | 25% | 25% | 100 % |
| Theoretical derivations | 25% | 25% | 25% | 25% | 100 % |
| Preparation of figures | 25% | 25% | 25% | 25% | 100 % |
| Writing – original draft | 25% | 25% | 25% | 25% | 100 % |
| Writing – review & editing | 25% | 25% | 25% | 25% | 100 % |