# Transformers Effectiveness in Medical Image Segmentation: A Comparative Analysis of UNet-based Architectures

Tagne Poupi Theodore Armand
Institute of Digital Anti-Aging Healthcare
Inje University
Gimhae, Republic of Korea
poupiarmand2@gmail.com

Subrata Bhattacharjee
Department of Computer Engineering
Inje University
Gimhae, Republic of Korea
subrata_bhattacharjee@outlook.com

Heung-Kook Choi
Department of Computer Engineering
Inje University
Gimhae, Republic of Korea
cschk@inje.ac.kr

Hee-Cheol Kim
Institute of Digital Anti-Aging Healthcare
Inje University
Gimhae, Republic of Korea
heeki@inje.ac.kr

***Abstract*— Medical image segmentation is a crucial task in healthcare as it helps in the accurate diagnosis and treatment of various medical conditions. UNet-based architectures have been widely used for medical image segmentation due to their ability to produce high-quality segmentations. However, there is a need to improve the performance of these architectures to enhance their effectiveness in medical image segmentation further. One promising approach is using transformers, which have shown great potential in improving the performance of various deep learning models. This research compares four UNet-based architectures (UNet, UNetR, Trans-UNet, and Swin-UNet) with and without transformers to evaluate their effectiveness in medical imaging using four independent datasets. The findings of this study will be valuable in advancing the field of medical image segmentation and contributing to the optimization of U-net-based architectures.**

***Keywords— transformers, segmentation, UNet, medical image segmentation***

## I. Introduction

Medical image segmentation is a process used in healthcare to separate different structures or regions of interest in medical images. It is critical in clinical analysis and diagnosis, providing doctors with the necessary information for disease diagnosis and treatment planning [1]. Medical image segmentation is crucial to healthcare systems because it improves diagnostic efficiency and accuracy, helps doctors detect a condition early, and makes more accurate diagnoses. By accurately identifying and separating different structures and regions, medical image segmentation provides a reliable basis for clinical diagnosis and pathology research, outputting the region of interest to the doctor [2]. Segmentation approaches are used to perform lesion and organ segmentation, which are essential to computer-aided disease diagnosis and have been improved by using deep convolutional neural networks, which have shown excellent results. However, medical image segmentation poses a significant challenge for networks in balancing local and global information, resulting in unstable segmentation outcomes. Therefore, several modifications have been proposed on U-Net, such as UNet++, SAUNet++, and CLAW U-NET, to improve segmentation accuracy [2-4]. Accurate medical image segmentation is important in healthcare to effectively diagnose and treat various medical conditions, making it necessary even in cases where incomplete overlap occurs across multiple phases. Multi-phase images can enhance medical image segmentation, which is beneficial for computer-aided disease diagnosis and achieving high accuracy in lesion segmentation. Despite the improvements made by existing segmentation models, there is still a need for further improvement in medical image segmentation.

UNet-based architectures are widely utilized in medical image segmentation applications, including CT scans, MRI, histopathology, and endoscopy of the prostate, breast, colon, abdomen, brain, lungs, etc. The UNet architecture has gained popularity in segmentation because it captures both high- and low-frequency information, even though it has a considerable semantic gap between the encoder and the decoder. The UNet-based architecture employs a contracting path and an expanding path. The contracting path is a typical convolutional network consisting of repeated convolution applications, followed by a rectified linear unit (ReLU) and a max-pooling operation. The expanding path enables precise localization using transposed convolutions. On the other hand, the UNet-based transformer models employ a transformer and adapt it into a UNet-like architecture, showing improved segmentation performance compared to other segmentation models [5]. However, U-Net-based architectures have not been extensively studied in relation to some cancers, such as lung cancer [3].

Transformers are becoming increasingly popular in medical image segmentation, particularly in UNet-based architectures. UNet Transformers (UNetR) introduces a novel architecture that treats segmentation as a sequence-to-sequence prediction problem, using transformers to perform better [6]. Additionally, researchers have proposed a Convolutional Neural Network (CNN)-based transformer architecture (Trans-UNet) that integrates self-attention into a convolutional neural network to enhance medical image segmentation, replacing the standard design of UNet with a hybrid CNN-transformer as an encoder [7]. Swin-UNet is a pure transformer-based U-shaped architecture that leverages Swin Transformer's success to achieve more accurate 2D medical image segmentation [8].

In this research, we carried out a comparative analysis to evaluate the performance and effectiveness of three transformer-based architectures, namely UNetR, Trans-UNet, and Swin-UNet, over the original UNet architecture proposed by Ranneberger et al. [9] using four medical imaging datasets. We analyze and discuss the results after training and testing the various models with our datasets.

## II. Datasets description

This research uses four datasets containing histopathological and endoscopy images.

Dataset 1 contains retinal images obtained as a combination of three different sub-datasets. A sub-dataset, DRIVE [10] was obtained from a diabetic retinopathy (DR) screening program in the Netherlands with over 40 retinal images accompanied by segmentation annotation for two classes describing the presence or absence of the DR. The second sub-dataset, CHASE_DB1 [11], contains retinal vessel segmentation data collected from fourteen children on both eyes. The last dataset, STARE [12], comprises 20 retinal fundus images with 50% pathology presence. These three datasets were combined to obtain the first dataset of our experiment. The combination was due to the limited amount of data from distinct datasets. The second dataset (Kvasir-SEG) contains polyp endoscopic images extracted from colonoscopy videos. Kvasir includes 1,000 polyp images collected from the polyp class in the Kvasir-SEG dataset [13].

Additionally, we considered a histopathological dataset containing colorectal cancer images for this analysis. The NCT-CRC-HE dataset [14] consists of 100,000 non-overlapping image patches from hematoxylin & eosin (H&E) stained histological images of human (CRC) and normal tissue. All images are of size 224×224 and classified into nine classes. For this experiment, we considered two classes: normal colon mucosa and colorectal adenocarcinoma epithelium. Lastly, we used the Multi-organ Nucleus Segmentation dataset (MoNuSeg) [15]. These data were collected from several hospitals and contain a variety of cancer types of data consisting of 30 tissue images, each of size 1000×1000 pixels, having 21,623 hand-annotated nuclear boundaries from 7 different organs, specifically breast, liver, kidney, prostate, bladder, colon, and stomach. Figure 1 shows sample data from all the datasets.

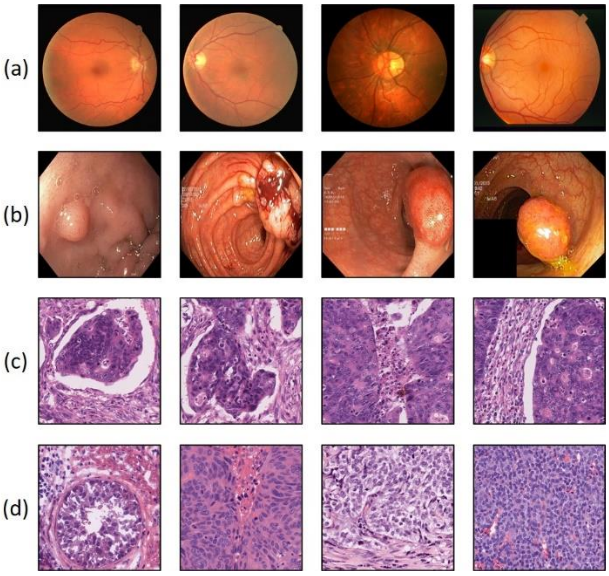

Figure 1: Dataset Samples. (a) Dataset 1: DRIVE+CHASE-DB1+STARE; (b) KvasirSeg Dataset; (c) NCT-CRC-HE Dataset; (d) MoNuSeg Dataset.

## III. Research design

We trained and tested four state-of-the-art algorithms using the datasets presented above. This section briefly portrays each model we considered for analysis, followed by the evaluation metrics used for comparative analysis.

### *A. The UNet architecture*

The UNet architecture, proposed by Ranneberger et al. (see Figure 2), is a type of CNN specifically designed to segment biomedical images. It comprises an encoder and a decoder pathway, interconnected through skip connections. This structure enables the network to effectively capture high-level semantic and low-level spatial information, leading to improved segmentation performance. The input image is passed through a series of convolutional layers in the encoder pathway, followed by max-pooling layers, which progressively downsample the input and extract higher-level features. On the other hand, the decoder pathway consists of a series of up-convolutional layers, which increase the spatial resolution of the feature maps and enable the network to reconstruct the segmented output image. Skip connections bridge the encoder and decoder pathways by transferring feature maps from the encoder to the corresponding layers in the decoder. This allows the network to leverage the high-resolution features from the earlier layers for more accurate segmentation to generate optimal output.

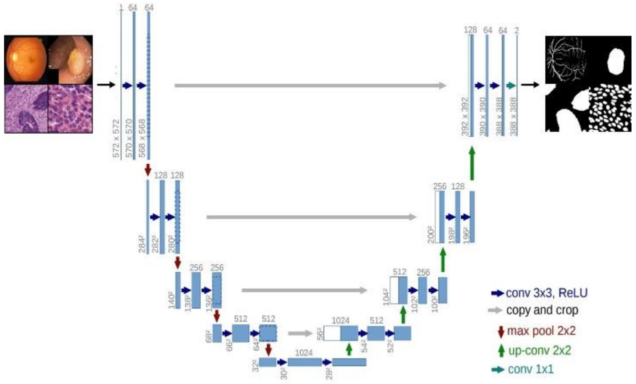

Figure 2: UNet architecture [9].

## B. *UNet Transformers (UNetR) architecture*

UNet Transformers (UNetR) is a novel deep learning architecture proposed by Hatamizadeh et al. designed to tackle various image segmentation and medical imaging tasks. The UNetR model is an innovative combination of the popular UNet architecture and the powerful transformer model. The UNet architecture, originally developed for biomedical image segmentation, has been widely adopted in various image-processing tasks due to its robustness and effectiveness. Its symmetric encoder-decoder structure enables the model to capture high-level semantic information and fine-grained local details. The encoder captures the context of the input image while the decoder reconstructs the segmented output from the encoded features. Conversely, transformers have gained immense popularity in natural language processing and computer vision, demonstrating state-of-the-art (SOTA) performance on various tasks. The key component of the Transformer architecture is the self-attention mechanism, which allows the model to weigh different input features according to their relevance. Integrating transformers into the UNet created an efficient and powerful model for various image segmentation and analysis tasks.

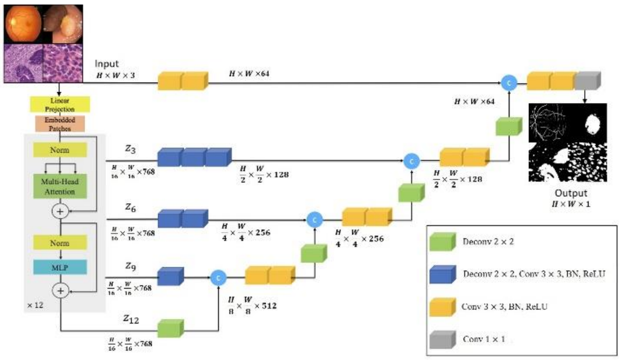
Input
$H \times W \times 3$
$H \times W \times 64$
Linear Projection
Embedded Patches
Norm
Multi-Head Attention
Norm
MLP
× 12
$z_3$
$\frac{H}{16} \times \frac{W}{16} \times 768$
$\frac{H}{2} \times \frac{W}{2} \times 128$
$z_6$
$\frac{H}{16} \times \frac{W}{16} \times 768$
$\frac{H}{4} \times \frac{W}{4} \times 256$
$z_9$
$\frac{H}{16} \times \frac{W}{16} \times 768$
$\frac{H}{8} \times \frac{W}{8} \times 512$
$z_{12}$
$\frac{H}{16} \times \frac{W}{16} \times 768$
$\frac{H}{4} \times \frac{W}{4} \times 256$
$\frac{H}{2} \times \frac{W}{2} \times 128$
$H \times W \times 64$
Output
$H \times W \times 1$
Deconv 2 × 2
Deconv 2 × 2, Conv 3 × 3, BN, ReLU
Conv 3 × 3, BN, ReLU
Conv 1 × 1

Figure 3: UNet Transformers (UNetR) architecture [6].

## C. *The transformer architecture Trans-UNet*

The Trans-UNet model is a popular transformer-based medical image segmentation model with SOTA results. In Trans-UNet, a modified transformer encoder is used as the backbone of the model to extract features from medical images. The extracted features are then passed through a UNet-style decoder to perform segmentation. Using transformers as encoders in Trans-UNet has several advantages over traditional encoder-decoder architectures. Firstly, transformers can capture long-range dependencies in the image, which is crucial for accurate segmentation. Secondly, transformers can better preserve spatial information in the image, resulting in sharper and more precise segmentation masks. Figure 4 below illustrates the Trans-UNet architecture. An encoding block consisting of CNN and transformer layers is used for down-sampling. The CNN layers use feature extractors to generate a feature map tokenized into a 2D embedding shape by linear projection and fed into the transformer layers. The upsampling process is straightforward, as the CNN-Transformer encoder is run by a 3×3 convolution layer with ReLU activation, upsampled, and then concatenated with the output of the third-level CNN feature extractor until the output is generated.

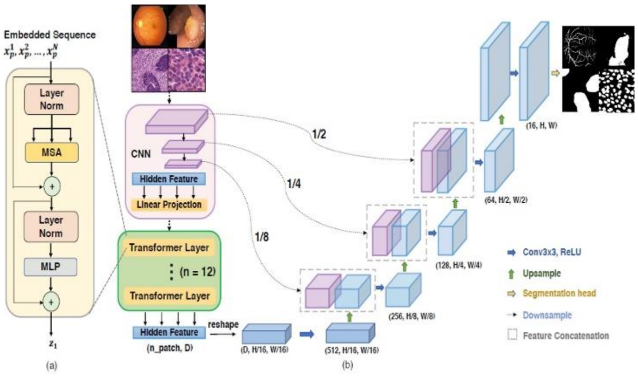

Figure 4: Trans-UNet architecture [7].

## D. *Swin-Unet architecture*

The Swin-Unet architecture proposed by Cao et al. for medical image segmentation is a U-Net-like pure transformer. This model is based on the swin transformer, a hierarchical transformer with shifted windows. The swin transformer has been shown to outperform previous transformer-based models on various computer vision tasks. The Swin-Unet model proposed leverages the strengths of the swin transformer to achieve SOTA performance on medical image segmentation tasks. The architecture comprises a series of swin transformer blocks in the encoder and decoder pathways (See figure 5). The swin transformer blocks process the tokenized image patches to extract high-level features and perform non-linear transformations. Like the original U-Net architecture, the Swin-Unet model employs skip connections between the encoder and decoder pathway. The skip connections allow the network to transfer high-resolution features from the encoder to the corresponding decoder layers, which helps to improve the segmentation performance. The Swin-Unet architecture is an innovative extension of the U-Net architecture and represents a significant advancement in transformer-based models for medical image segmentation.

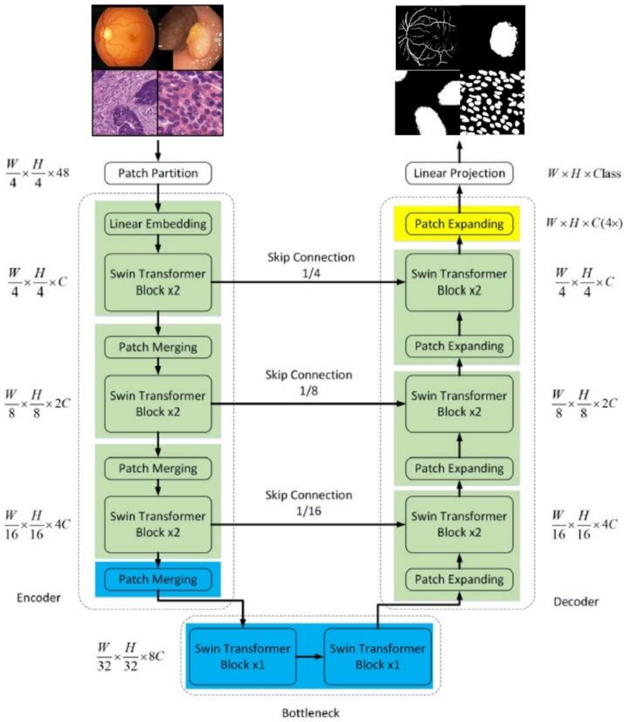
$\frac{W}{4}\times\frac{H}{4}\times 48$
Patch Partition
Linear Embedding
$\frac{W}{4}\times\frac{H}{4}\times C$
Swin Transformer Block x2
Patch Merging
$\frac{W}{8}\times\frac{H}{8}\times 2C$
Swin Transformer Block x2
Patch Merging
$\frac{W}{16}\times\frac{H}{16}\times 4C$
Swin Transformer Block x2
Patch Merging
Encoder
Skip Connection 1/4
Skip Connection 1/8
Skip Connection 1/16
Linear Projection
$W\times H\times Class$
Patch Expanding
$W\times H\times C(4\times)$
Swin Transformer Block x2
$\frac{W}{4}\times\frac{H}{4}\times C$
Patch Expanding
Swin Transformer Block x2
$\frac{W}{8}\times\frac{H}{8}\times 2C$
Patch Expanding
Swin Transformer Block x2
$\frac{W}{16}\times\frac{H}{16}\times 4C$
Patch Expanding
Decoder
$\frac{W}{32}\times\frac{H}{32}\times 8C$
Swin Transformer Block x1
Swin Transformer Block x1
Bottleneck

Figure 5: Swin-Unet architecture [8].

### E. *Evaluation Metrics*

To evaluate the segmentation performance of distinct models, two standard evaluation metrics are used to compare the methods. The evaluation metrics include Dice Coefficient (DC) and Intersection over Union (IoU) or Jaccard Coefficient. Four different values, namely true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN), are used to calculate the DC and IoU. These evaluation metrics are computed in equations (1) and (2).

$$Dice = \frac{2\times TP}{2\times TP+FP+FN} \quad (1)$$

$$IoU = \frac{TP}{TP+FP+FN} \quad (2)$$

## IV. Results and Discussion

In this section, we present a comprehensive analysis of the performance of Transformer-based models on four different medical image segmentation datasets. Our study explores how well the latest models handle other challenges in medical image datasets. We want to determine how effective these models are in dealing with the different eventual problems in various medical images. We compare different Transformer-based models on various medical image datasets to understand their strengths and possible limitations. By looking at the results side by side, we want to figure out how well these models can work with different types of medical images and if they can accurately outline the areas we are interested in.

Tables I-IV show the comparative analysis of segmentation models using four datasets: DRIVE + CHASE-DB1 + STARE, Kvasir-SEG, NCT-CRC-HE, and MoNuSeg. Table I indicates that Trans-UNet outperforms other methods by giving overall DC and IoU of 0.716 and 0.573, respectively. Similarly, on the NCT-CRC-HE dataset in Table III, the Trans-UNet outperforms other methods on DC and IoU by achieving 0.835 and 0.728, respectively. For the Kvasir-SEG dataset in Table II, the base model UNet outperformed other SOTA methods on all metrics. On the other hand, based on the results of MoNuSeg in Table IV, we can observe that the Swin-UNet achieved better performance than other datasets and outperformed previous SOTA methods on all metrics. To visualize the performance improvement of each method, we plot the comparison results, shown in Figure 6. Figure 7 shows the visualization results of medical image segmentation using distinct deep learning models.

TABLE I. RESULTS OF THE DATASET 1

| Method | DC | IoU | Test loss |
|---|---|---|---|
| UNet | 0.6190 | 0.4656 | 0.1567 |
| UNetR | 0.6106 | 0.4596 | 0.1112 |
| **Trans-UNet** | **0.7167** | **0.5732** | **0.0892** |
| Swin-Unet | 0.6126 | 0.4460 | 0.1515 |

TABLE II. RESULTS OF THE KVASIRSEG DATASET

| Method | DC | IoU | Test loss |
|---|---|---|---|
| **UNet** | **0.7312** | **0.5880** | **0.1936** |
| UNetR | 0.4422 | 0.2880 | 0.3418 |
| Trans-UNet | 0.4021 | 0.2580 | 0.4015 |
| Swin-Unet | 0.5173 | 0.3523 | 0.3238 |

TABLE III. RESULTS OF THE NCT-CRC-HE DATASET

| Method | DC | IoU | Test loss |
|---|---|---|---|
| UNet | 0.8162 | 0.7056 | 0.2194 |
| UNetR | 0.7184 | 0.5721 | 0.3722 |
| **Trans-UNet** | **0.8364** | **0.7281** | **0.2147** |
| Swin-Unet | 0.8363 | 0.7257 | 0.2504 |

TABLE IV. RESULTS OF THE MONUSEG DATASET

| Method | DC | IoU | Test loss |
|---|---|---|---|
| UNet | 0.9203 | 0.8525 | 0.0961 |
| UNetR | 0.6235 | 0.4534 | 0.2939 |
| Trans-UNet | 0.8461 | 0.7335 | 0.1289 |
| **Swin-Unet** | **0.9547** | **0.9133** | **0.0432** |

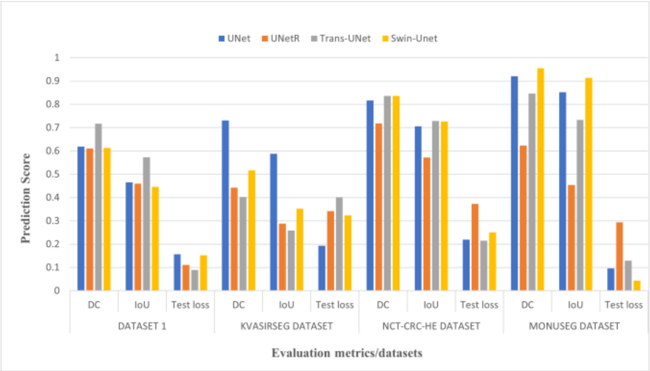

Figure 6: Comparative results for various segmentation models/datasets.

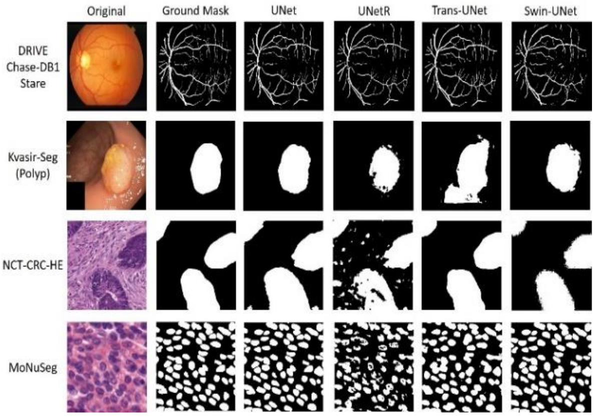

Figure 7: Visualization of segmentation results.

The performance of UNet, UNetR, Trans-UNet, and Swin-UNet depends on various factors, and its effectiveness can vary across different image datasets. There is no guarantee that a particular architecture will always perform better for all datasets. However, before choosing a model architecture, it is essential to conduct thorough experimentation and validation on the specific dataset and task at hand. Comparing the performance of multiple architectures and tuning hyperparameters is part of finding the best model for a particular application. Moreover, interpreting and understanding the results through visual inspection of segmentation outputs can provide valuable insights into the model's behavior on specific datasets.

## V. Conclusion

This paper investigates the effectiveness of some UNet transformer-based models in medical image segmentation tasks. We used four independent datasets for the comparative analysis, considering the Dice and Jaccard coefficients as the most informative metrics to appreciate the model's effectiveness. After experiments, we concluded that the performance of UNet, UNetR, Trans-UNet, and Swin-UNet depends on various factors, and its efficacy can vary across different image datasets. There is no guarantee that a particular architecture will always perform better for all datasets; therefore, choosing architectures and models requires thorough experimentation and validation on the specific dataset.

## Acknowledgment

This research was supported both by a grant from Korea Health Technology R&D Project through the Korea Health Industry Development Institute validation (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant No: HI21C0977), and by the MSIT (Ministry of Science ICT), Korea, under the National Program for Excellence in SW, supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation) in 2022.

## References

[1] Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., & Nandi, A. K. (2022). Medical image segmentation using deep learning: A survey. *IET Image Processing*, *16*(5), 1243-1267.

[2] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4* (pp. 3-11). Springer International Publishing.

[3] Xiao, Hanguang, et al. "SAUNet++: an automatic segmentation model of COVID-19 lesion from CT slices." *The Visual Computer* 39.6 (2023): 2291-2304.

[4] Chang, Yao, et al. "Transclaw u-net: Claw u-net with transformers for medical image segmentation." *arXiv preprint arXiv:2107.05188* (2021).

[5] Xu, G., Wu, X., Zhang, X., & He, X. (2021). Levit-unet: Make faster encoders with transformer for medical image segmentation. *arXiv preprint arXiv:2107.08623*.

[6] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., ... & Xu, D. (2022). Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 574-584).

[7] 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*.

[8] 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.

[9] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *In Proceedings of Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference*, pages 234-241, 2015.

[10] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004.

[11] A. Carballal et al., "Automatic multiscale vascular image segmentation algorithm for coronary angiography," *Biomed. Signal Process. Control.*, vol. 46, pp. 1–9, 2018.

[12] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEETrans.Med. Imag.*, vol. 19, no. 3, pp. 203–210, Mar. 2000

[13] Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., de Lange, T., Johansen, D., & Johansen, H. D. (2020). Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26* (pp. 451-462). Springer International Publishing.

[14] Kather, J. N., Halama, N., & Marx, A. (2018). 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo10*, *5281*.

[15] Kumar, N., Verma, R., Anand, D., Zhou, Y., Onder, O. F., Tsougenis, E., ... & Sethi, A. (2019). A multi-organ nucleus segmentation challenge. *IEEE transactions on medical imaging*, *39*(5), 1380-1391.