Performance of Transformer-based Swin-UnetR Model On Semantic Segmentation of Multi-regional Volumetric Medical Images

Shadid Yousuf
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology (BUET)

K M Fahim Asif
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology (BUET)

Galib Mahmud Sharif
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology (BUET)

Tamim Hasan Bhuiyan
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology (BUET)

***Abstract*—Medical image segmentation plays as important role in computer aided diagnosis and treatment. Although various deep learning architectures have been deployed to aid the segmentation task, most of the the studies are conducted to extract the region of interests from 2D image slices. On the other hand, 3D segmentation techniques are capable of capturing the spatial contexts in all three dimensions, thus offering more delineation and accuracy.In recent years, transformer based architectures, for their robustness in capturing global contexts, are proven useful in various computer vision tasks including medical image segmentation. In this study, we deploy the state-of-the-art transformer based model Swin U-NetR to evaluate its performance on two tasks : 1. Brain Tumor Segmentation and 2. Spleen Segmentation. For the brain tumor segmentation task, we propose a transfer learning-based approach to optimize the model for two publicly available brain tumor datasets, namely, BraTS 2021 and Medical Segmentation Decathlon (MSD). For the spleen segmentation, we deploy several models to determine the best performing model to replicate state-of-the-art results.**

I. Introduction

The advent of 3D medical imaging technologies, such as MRI (Magnetic Resonance Imaging) and CT (Computed Tomography), has revolutionized the field of diagnostics, enabling detailed visualization of anatomical structures and pathological conditions in three dimensions. However, the voluminous data generated by these modalities pose significant challenges for manual analysis and interpretation. Automated segmentation of 3D medical images emerges as a critical solution to this problem, facilitating precise delineation of anatomical regions and pathological lesions, thus significantly enhancing diagnosis, treatment planning, and monitoring of disease progression.

3D medical image segmentation involves the partitioning of volumetric images into segments or regions corresponding to different tissues or structures. This process is paramount for quantitative analysis in clinical practice and medical research, including volumetric measurement, structure localization, and computer-assisted surgery planning. The complexity of 3D medical images, characterized by the variability in anatomy among patients, the presence of noise, and the subtle differences between different types of tissues, necessitates the development of robust and accurate segmentation algorithms.

Recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs) and Transformer models, have shown promising results in addressing these challenges. Models like U-Net [1] and its variants have become the de facto standard due to their effectiveness in capturing spatial hierarchies and context, essential for accurate segmentation. These advances are mainly due to the powerful feature extraction capabilities of Convolutional Neural Networks (CNN)s. However, the limited kernel size of CNN-based techniques restricts their capability of learning long-range dependencies that are critical for accurate segmentation of tumors that appear in various shapes and sizes

Transformers, initially prominent in fields like natural language processing and computer vision, have recently made significant strides in medical image analysis. Notably, UNETR [2] marks a pioneering approach by integrating a Vision Transformer as its core encoder, moving away from traditional CNN-based feature extraction mechanisms. This model has demonstrated superior performance, enhancing both accuracy and efficiency across a spectrum of medical image segmen-

tation challenges. Furthermore, UNETR++, with its advanced efficient hybrid hierarchical structure, has achieved remarkable improvements in segmentation precision and computational efficiency, evidenced by its reduced parameters, FLOPS, and quicker inference speeds [3].

Recently, Swin transformer [4] has been proposed as a hierarchical vision transformer that computes self-attention in an efficient shifted window partitioning scheme. As a result, Swin transformers are suitable for various downstream tasks wherein the extracted multi-scale features can be leveraged for further processing. Within the domain of 3D medical image analysis, innovative architectures such as Swin UNETR and Swin Unet3D, referenced in studies [5,6] and [7], have been introduced, leveraging the Swin transformer technique. These advanced frameworks have successfully set new benchmarks, achieving unparalleled performance in the field of 3D medical image segmentation. This underscores the efficacy of Swin transformer methodologies in enhancing the precision and efficiency of segmentation tasks, marking a significant advancement in medical imaging technology.

In the context of this project, we embarked on the task of brain tumor segmentation by initially selecting three distinct models: Unet3D, UnetR, and Swin UNetR. These models were rigorously trained on the BRATS 2021 dataset, followed by a comparative analysis grounded in quantitative evaluations. Our findings indicated that Swin UNetR outperformed its counterparts, showcasing superior results. Building upon this insight, we further refined the Swin UNetR model through extended training over a greater number of epochs. The enhanced model's performance was subsequently assessed on the MSD Brain dataset, providing valuable insights into its efficacy.

Additionally, in a separate analysis, we utilized the Unet model for segmenting the MSD spleen dataset. This investigation yielded benchmark-setting results, highlighting the effectiveness of the Unet model in spleen segmentation tasks. This multifaceted research not only underscores the prowess of Swin UNetR in brain tumor segmentation but also showcases the adaptability of Unet in handling different segmentation challenges

.

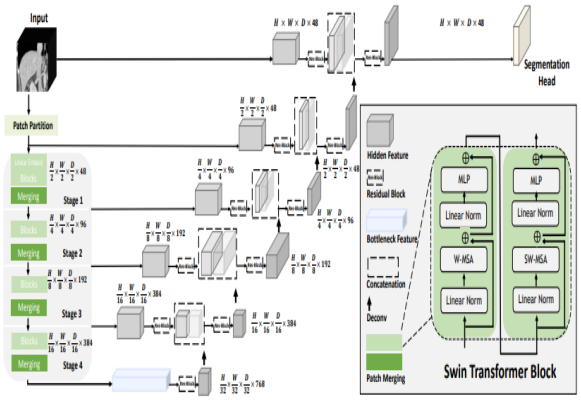
II. Methodology

A. Model Selection

For the Brain tumor segmentation task, at first we chose 3D U-Net, U-NetR and Swin U-NetR for a comparative analysis between the performances of the models.

3D U-Net is a Convolutional Neural Network (CNN)-based architecture that extends the original U-Net model for handling volumetric images. Unlike the original 2D U-Net, which operates on 2D images, the 3D U-Net performs 3D convolutional operations to capture spatial context and structural information in volumetric data. The convolutional filters used in the 3D U-Net slide across three dimensions (width, height, and depth) of the input volume to extract spatial features.

The CNN based architectures, although capable of capturing local contexts efficiently, fails to capture global contexts or features. In order to address this limitation, transformer-based architerctures like U-NetR and Swin U-NetR have been proposed where CNN based encoders are replaced by stack of transformers. The transformers have skip connections with the CNN-based decoder block.



0.5

Fig. 1. Swin U-NetR Architecture

In the trial-and-error phase, we trained all three models for 5000 iterations on the same dataset to conclude that Swin U-NETR is the the best performing model with the highest Mean DICE score. We decided to select Swin U-NETR as our baseline model to train on the data and infer on another validation data.

B. Loss Function

For the Brain tumor segmentation task, the loss function is a combination of soft dice loss and cross-entropy loss. The mathematical expression of the loss function is given by-

$$\mathcal{L}(G,Y) = 1 - \frac{2}{J}\sum_{j=1}^{J}\left(\frac{\sum_{i=1}^{I} G_{i,j}Y_{i,j}}{\sqrt{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} Y_{i,j}^2}}\right) - \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} G_{i,j}\log Y_{i,j}.$$

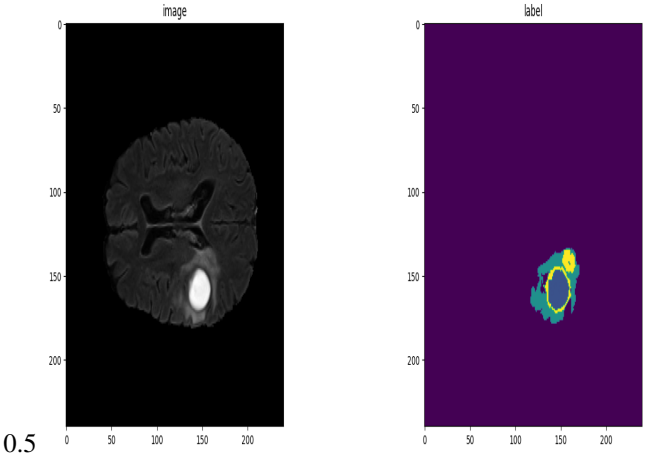
Here, I is the number of voxels, J is the number of classes, $Y_{i,j}$ and $G_{i,j}$ are output and one-hot encoded ground truth for class j at voxel i, respectively.

III. Experiment

A. Datasets

For brain tumor segmentation, we trained the model on BraTS 2021 [ref] and evaluated the performance of the trained model on MSD Task 01 : Brain Tumor [ref] dataset.

BraTS 2021: The BraTS 2021 dataset contains 1470 samples of multimodal mpMRI data along with their ground truths given in NIFTI format (.nii). The modalities are - FLAIR,T1w,T1gd,T2w, respectively. The training dataset contains 1251 samples and the validation dataset contains 219 samples. As per the BraTS challenge instruction, the labels were mapped into 3 classes (Enhanced Tumor, Tumor Core and Whole Tumor). The input channel is 4.



0.5

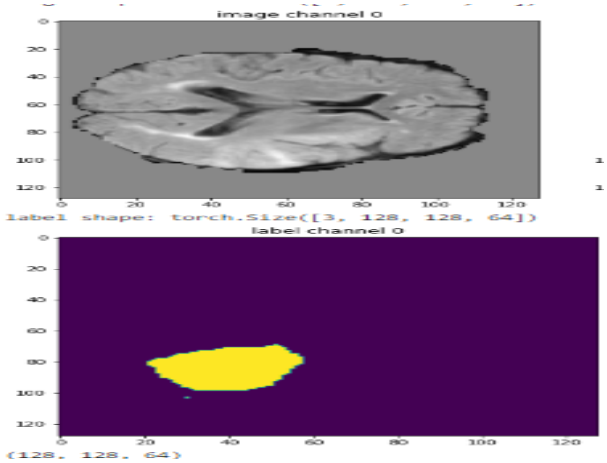
Fig. 2. A sample image from BraTS dataset and corresponding label

The voxel dimensions of the region of interest was set to (128,128, 64) as per the model input shape requirement, with the voxel spacing given as 1×1×1 mm³. As a preprocessing step, Z-score normalization is. applied to the images. In order to prevent overfitting of the model, 50% of the training data was augmented with random intensity scaling and shifting, and random flipping of the images along each of the three axes.

MSD Task01 (Brain Tumor): The Medical Segmentation Decathlon (MSD) dataset contains 2633 3D images with annotations from the following body parts : brain, heart, liver, hippocampus, prostrate, lung, pancreas, hepatic vessel, spleen and colon. Our area of interest is in "Task 01 : Brain Tumor" which contains 750 (484 Training+266 Testing) volumes. Only validation dataset is used in testing, while whole of the dataset is utilized while finetuning. The annotations provided are for edema (label 1), non-enhancing solid core(label 3), and enhancing tumor (label 2). The labels were mapped into the BraTS labels as follows - i. Label 2 = Enhancing Tumor (ET) ii. Label 1 + Label 2 + Label 3 = Whole Tumor (WT) iii. Label 2 + Label 3 = Tumor Core

The modalities and voxel dimensions are same as BraTS dataset. Therefore the preprocessing steps are same as before.

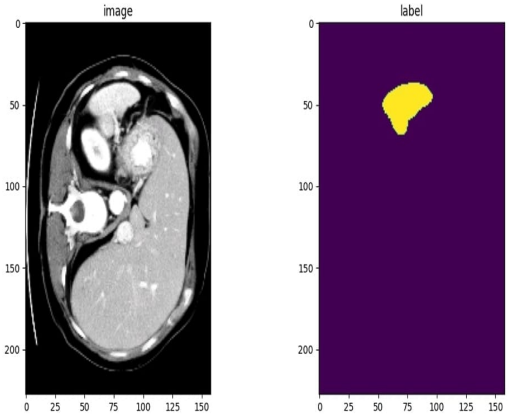
MSD Task09 (Spleen): The MSD spleen dataset comprises 61 3D volumes, with 41 designated for training and 20 for testing, containing MRI data in NIfTI format. The voxel dimensions of the images are adjusted to (228, 158, 113) to meet model requirements, with a single input channel. The data preprocessing pipeline involves loading CT images and labels from NIFTI files. Subsequent steps include adjusting spacing using pixdim=(1.5, 1.5, 2). Intensity normalization is



0.5

Fig. 3. A sample image from MSD Brain Tumor dataset and corresponding label

achieved through ScaleIntensityRanged, scaling values to [0, 1] within the range [-57, 164]. CropForegroundd is utilized to remove zero borders, concentrating on the valid body area. RandCropByPosNegLabeld randomly crops patch samples based on positive/negative ratios, ensuring negative sample centers are within the valid body area. Lastly, RandAffined efficiently applies rotations, scaling, shearing, and translations using a PyTorch affine transform. The resulting preprocessed data is tailored for the model, facilitating accurate spleen segmentation in CT scans.



0.5

Fig. 4. A sample image from MSD Spleen dataset and corresponding label

B. Evaluation Metric

The metric used to evaluate the accuracy of the segmentation task is DICE score, which is given by -

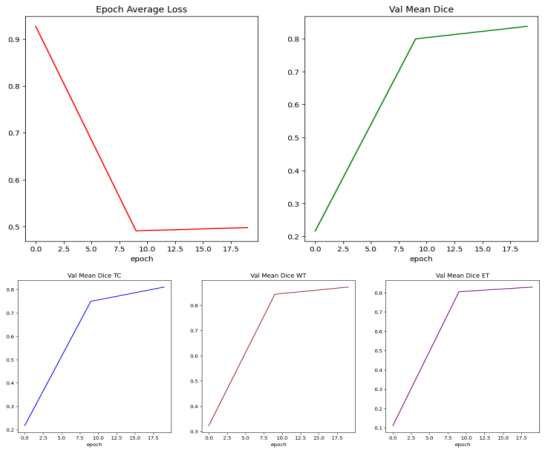
$$\mathcal{D}\rangle\rfloor\rceil(G,P) = \frac{2\sum_{i=1}^{I} G_{i,j}P_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} P_{i,j}^2}$$

where G_i and P_i represent ground truth and prediction value of voxel i, respectively.

C. *Implementation Details*

The segmentation task is implemented using PyTorch and MONAI frameworks. The models are all pre-trained or fine-tuned on Kaggle using NVIDIA TESLA G100 GPU.

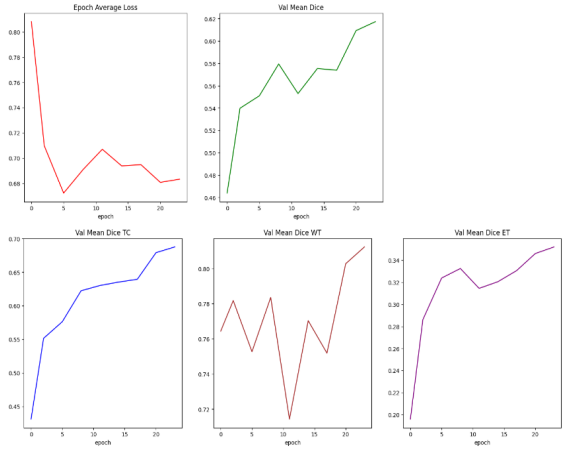
For pretraining the model on BraTS dataset, the batch size is set to 2. Foreground of the images are cropped to a dimension of [128,128,64] to focus on the foreground of the images. AdamW optimizer is utilized with an initial learning rate of 0.0001. For the given batch size, the training time of the Swin U-NetR model was roughly four hours for 12,525 iterations(25 epochs). The weights are updated in every 3 epochs, and the best performing model parameters are saved for inference. Sliding window inference is used with batch size set to 4 and overlap portion set to 0.5.



0.5

Fig. 5. Learning Curve and Validation Curves for Pretraining

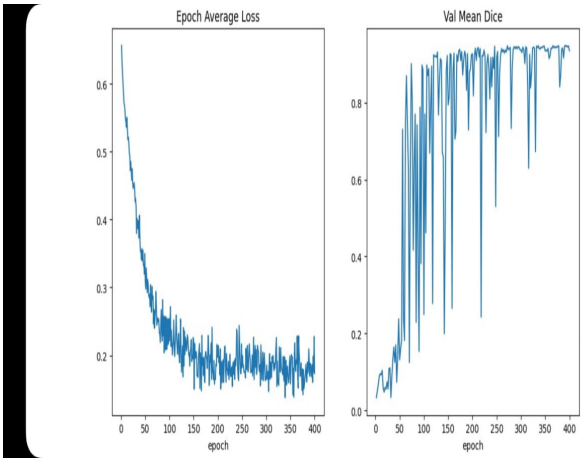
The pretrained model is further fine-tuned on the MSD dataset with all the identical parameters. All layers are kept trainable in the finetuning state, that is, no layer is frozen for fine-tuning.



0.5

Fig. 6. Learning Curve and Validation Curves for Finetuning

For the spleen segmentation task, the batch size is set to 2 and Adam optimizer is used. The training time is around 1 hour and 50 minutes for 6400 iterations (400 epochs). As before, sliding window inference is used for inference on test dataset.



0.5

Fig. 7. Learning curve and validation curve for spleen segmentation

IV. RESULT

A. *Quantitative Evaluations*

Our trained Swin U-NetR model achieves 85.66% average DICE score in BraTS 2021 dataset, with the DICE scores 83.43%, 87.87% and 85.69% for Tumor Core (TC), Whole Tumor (WT) and Enhanced Tumor (ET), respectively.

The weights of the trained model is then used on MSD Brain Tumor Dataset. Although the DICE score obtained for WT is higher than the benchmark score (73.91%), the model performs poorly on the other two tumor types. In an attempt to optimize the model for both of these datasets used, the model is fine-tuned on MSD dataset.

The fine-tuned model results in a trade-off between an enhanced performance in MSD dataset and a reduced performance on the BraTS 2021 dataset.

The following table summarizes the scores obtained in each of the above procedures compared to the benchmark [ref].

TABLE I
MODEL PERFORMANCE ON BRAIN TUMOR SEGMENTATION

Model	**BraTS 2021**				**MSD**		
Description	***WT***	***ET***	***TC***	***Avg.***	***WT***	***ET***	
Benchmark [ref]	87.6%	92.9%	91.4%	90.6%			
Swin U-NetR Pretrained on BraTS 2021	87.87%	85.69%	83.43%	85.66%	73.91%	11.16%	2
Swin U-NetR Fine-tuned on MSD	70.11%	56.47%	45.36%	57.31%	81.24%	35.20%	6

[a]Performace Summary of different models

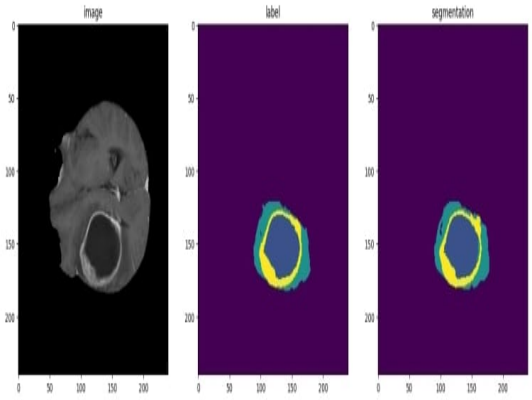
In case of Spleen Segmentation, both 3D U-Net and Swin U-NetR are applied to determine the best performing model. 3D U-Net achieves accuracy closer to benchmark in this regard. The following table summarizes the performances of different models.

TABLE II
MODEL PERFORMANCE ON SPLEEN SEGMENTATION

Model	DICE Score
Benchmark [ref]	97.43%
Swin U-NetR	
3D U-NetR	95.75%

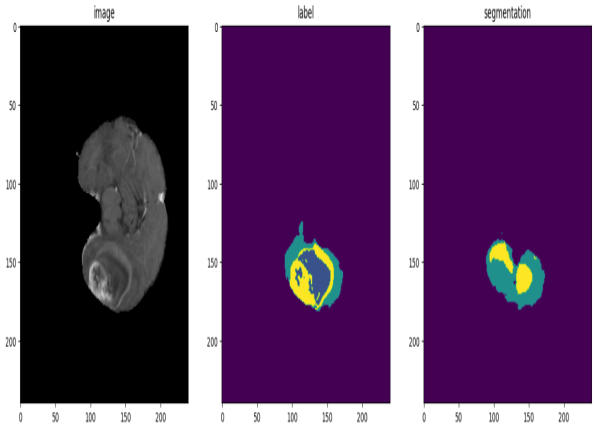
B. *Qualitative Results*

The figures below give us qualitative estimation of segmentation performance of different methods.



0.5

Fig. 8. Pretrained model testing on BraTS dataset

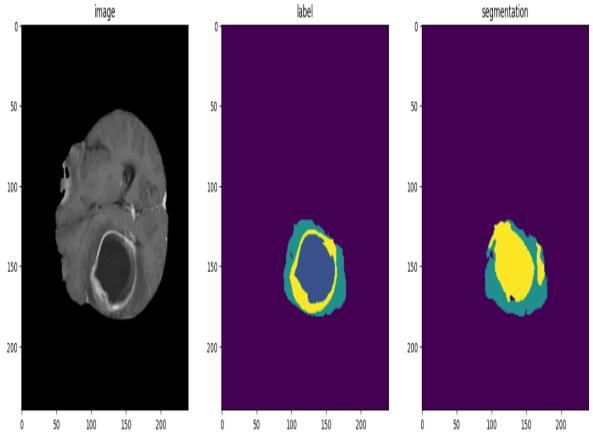


0.5

Fig. 9. Pretrained model testing on MSD Brain Tumor dataset

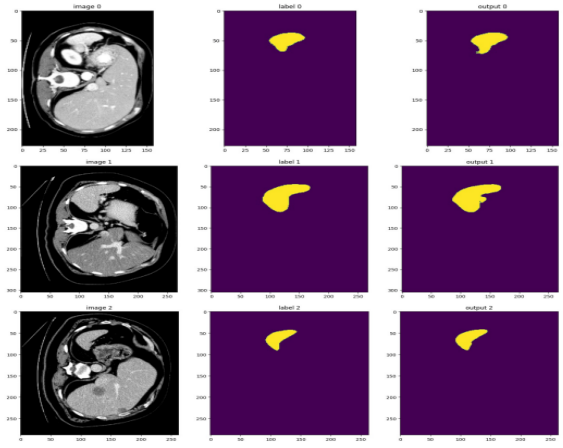
V. CONCLUSION

This study proves the effectiveness of Transformers for capturing global contextual features in the domain of medical



0.5

Fig. 10. Finetuned model testing on BraTS dataset



0.5

Fig. 11. Spleen Segmentation Inference

images. We showed that Swin U-NetR model was close to generate state-of-the art results for MSD Spleen Segmentation dataset. The model's performance on BraTS 2021 dataset upon training was also impressive and close to benchmark.

For the model pretrained on BraTS 2021 dataset, We evaluate the generizability of the model by testing it on another multi-site dataset, namely, MSD. The DICE score obtained for Whole Tumor(WT) was quite close to benchmark meand DICE than the other two labels. Upon inspecting the annotations provided in the MSD dataset, we conclude that the number of containing prominent edema is far more than enhancing and non-enhancing tumors, leading to the model's accuracy for just one label.

In order to introduce generalizabilty to our model, we finetune the model on MSD dataset which sees considerable degrade in performance on the original dataset. There is a scope for further improving the performace by freezing some of the layers of the pretrained mode, then finetuning on the rest of the layers.

VI. REFERENCES

REFERENCES

[1] O. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d U-net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 424–432. DOI: https://doi.org/10.48550/arXiv.1606.0665010.48550/arXiv.1606.06650

[2] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. Roth, and D. Xu, "UnetR: Transformers for 3d medical image segmentation," *arXiv preprint* arXiv:2103.10504 (2021). DOI: https://doi.org/10.48550/arXiv.2103.1050410.48550/arXiv.2103.10504

[3] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "UNETR++: Delving into Efficient and Accurate 3D Medical Image Segmentation," *arXiv* 2212.04497 (2022). DOI: https://doi.org/10.48550/arXiv.2212.0449710.48550/arXiv.2212.04497

[4] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[5] A. Hatamizadeh *et al.*, "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," *arXiv* 2201.01266 (2022). DOI: https://doi.org/10.48550/arXiv.2201.0126610.48550/arXiv.2201.01266

[6] Y. Tang *et al.*, "Self-supervised pre-training of swin transformers for 3d medical image analysis," *arXiv preprint* arXiv:2111.14791, 2021. DOI: https://doi.org/10.48550/arXiv.2111.1479110.48550/arXiv.2111.14791

[7] Y. Cai *et al.*, "Swin Unet3D: A three-dimensional medical image segmentation network combining vision transformer and convolution," *BMC Med. Informat. Decis. Making*, vol. 23, no. 1, pp. 33, Feb. 2023. DOI: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-023-02129-z10.1186/s12911-023-02129-z