Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images

Abstract. Semantic segmentation of brain tumors is a fundamental medical image analysis task involving multiple MRI imaging modalities that can assist clinicians in diagnosing the patient and successively studying the progression of the malignant entity. In recent years, Fully Convolutional Neural Networks (FCNNs) approaches have become the de facto standard for 3D medical image segmentation. The popular "U-shaped" network architecture has achieved state-of-the-art performance benchmarks on different 2D and 3D semantic segmentation tasks and across various imaging modalities. However, due to the limited kernel size of convolution layers in FCNNs, their performance of modeling long-range information is sub-optimal, and this can lead to deficiencies in the segmentation of tumors with variable sizes. On the other hand, transformer models have demonstrated excellent capabilities in capturing such long-range information in multiple domains, including natural language processing and computer vision. Inspired by the success of vision transformers and their variants, we propose a novel segmentation model termed Swin UNEt TRansformers (Swin UNETR). Specifically, the task of 3D brain tumor semantic segmentation is reformulated as a sequence to sequence prediction problem wherein multi-modal input data is projected into a 1D sequence of embedding and used as an input to a hierarchical Swin transformer as the encoder. The swin transformer encoder extracts features at five different resolutions by utilizing shifted windows for computing self-attention and is connected to an FCNN-based decoder at each resolution via skip connections. We have participated in BraTS 2021 segmentation challenge, and our proposed model ranks among the top-performing approaches in the validation phase.
Code: https://monai.io/research/swin-unetr

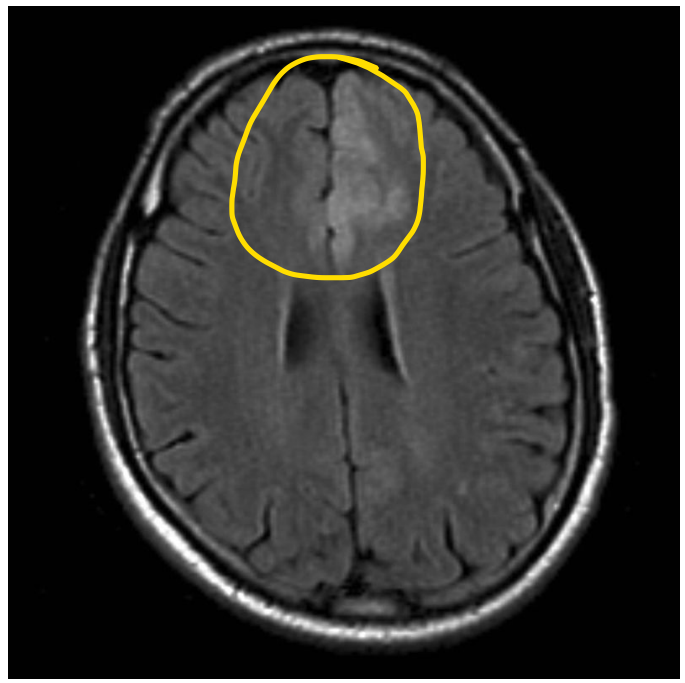
Keywords: Image Segmentation · Vision Transformer · Swin Transformer · UNETR · Swin UNETR · BRATS · Brain Tumor Segmentation

What is a Swin transformer?

Seq2Seq:
transformer combines an encoder and a decoder. the encoder first processes the input sequence into a context representation. then the decoder generates an output sequence

Furthermore, automated medical image segmentation techniques [18] have shown prominence for providing an accurate and reproducible solution for brain tumor delineation. Recently, deep learning-based brain tumor segmentation techniques [31,21,32,20] have achieved state-of-the-art performance in various benchmarks [7,35,2]. 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. Although several efforts [24,10] have tried to address this limitation by increasing the receptive field of the convolutional kernels, the effective receptive field is still limited to local regions.

ker size



Tumor boundary is not defined by sharp edges, but by global intensity patterns across distant regions.

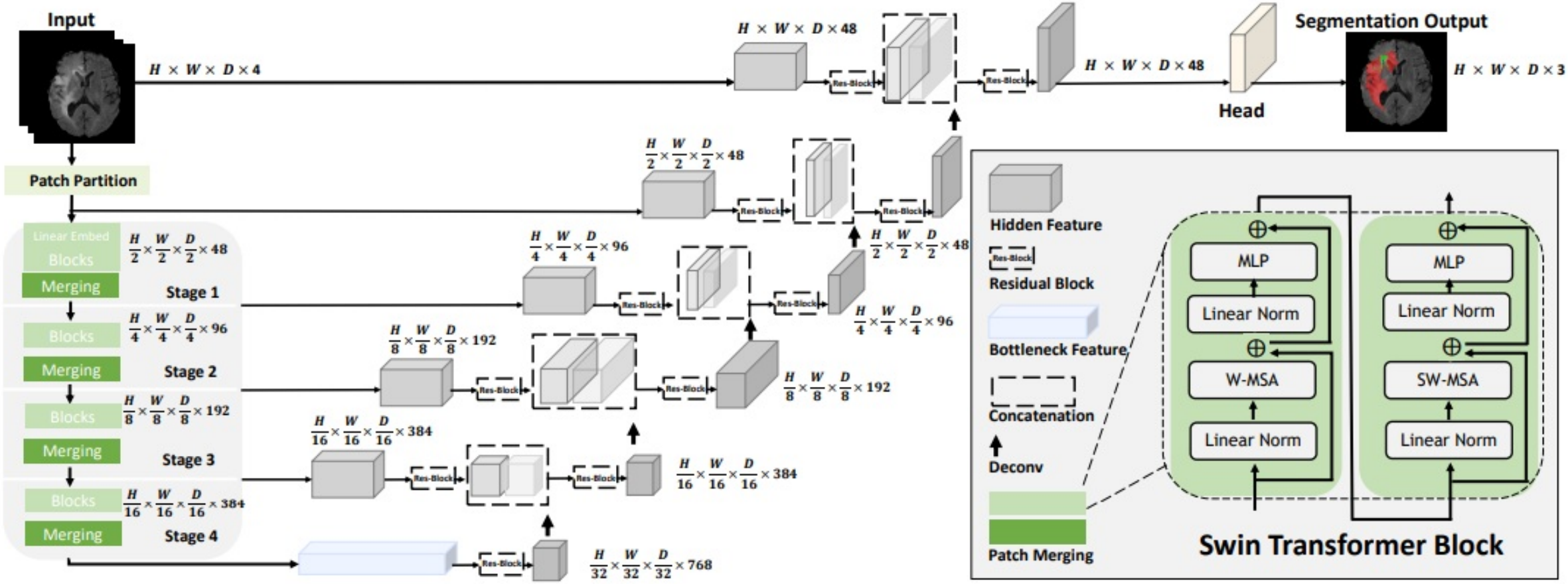
Recently, transformer-based models have shown prominence in various domains such as natural language processing and computer vision [38,13,14]. In computer vision, Vision Transformers [14] (ViT)s have demonstrated state-of-the-art performance on various benchmarks. Specifically, self-attention module in ViT-based models allows for modeling long-range information by pairwise interaction between token embeddings and hence leading to more effective local and global contextual representations [34]. In addition, ViTs have achieved success in effective learning of pretext tasks for self-supervised pre-training in various applications [9,8,36]. In medical image analysis, UNETR [16] is the first methodology that utilizes a ViT as its encoder without relying on a CNN-based feature extractor. Other approaches [40,39] have attempted to leverage the power of ViTs as a stand-alone block in their architectures which otherwise consist of CNN-based components. However, UNETR has shown better performance in terms of both accuracy and efficiency in different medical image segmentation tasks [16].

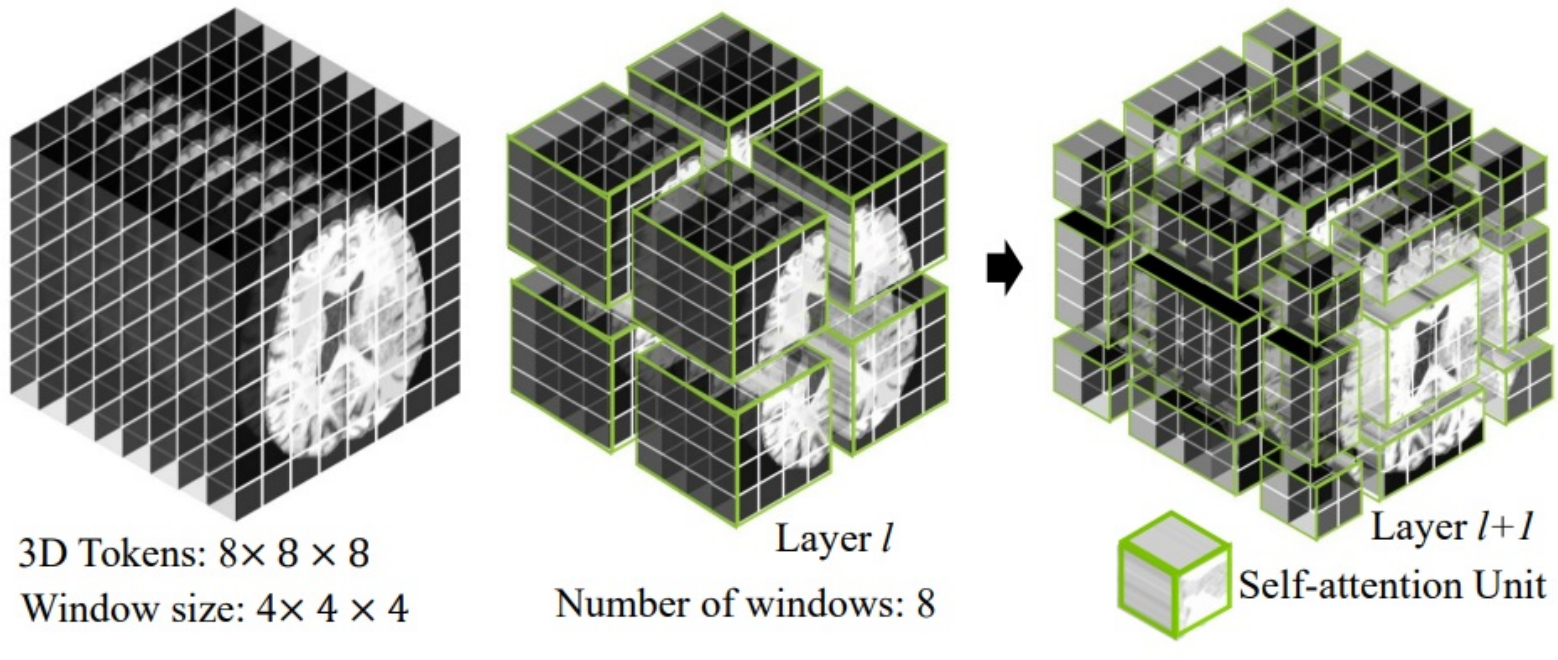
We let each part of the image decide which other parts are important to look at before making a decision.

A previous paper to this one?

this is a must read paper

Recently, Swin transformers [25,26] have 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. In this work, we propose a novel architecture termed Swin UNEt TRansformers (Swin UNETR), which utilizes a U-shaped network with a Swin transformer as the encoder and connects it to a CNN-based decoder at different resolutions via skip connections. We validate the effectiveness of our approach for the task of multi-modal 3D brain tumor segmentation in the 2021 edition of the Multi-modal Brain Tumor Segmentation Challenge (BraTS). Our model is one of the top-ranking methods in the validation phase and has demonstrated competitive performance in the testing phase.


Input
$H \times W \times D \times 4$
Patch Partition
Linear Embed
Blocks
Merging
Stage 1
Stage 2
Stage 3
Stage 4
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 48$
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 96$
$\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 192$
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 384$
$\frac{H}{32} \times \frac{W}{32} \times \frac{D}{32} \times 768$
Res-Block
$H \times W \times D \times 48$
$H \times W \times D \times 3$
Head
Segmentation Output
Hidden Feature
Residual Block
Bottleneck Feature
Concatenation
Deconv
Patch Merging
MLP
Linear Norm
W-MSA
SW-MSA
Swin Transformer Block


3D Tokens: 8× 8 × 8
Window size: 4× 4 × 4
Layer *l*
Number of windows: 8
Layer *l*+1
Self-attention Unit