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

Ali Hatamizadeh[1], Vishwesh Nath[1], Yucheng Tang[2], Dong Yang[1], Holger R. Roth[1], and Daguang Xu[1]

[1] NVIDIA
[2] Vanderbilt University
ahatamizadeh@nvidia.com

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

1 Introduction

There are over 120 types of brain tumors that affect the human brain [28]. As we enter the era of Artificial Intelligence (AI) for healthcare, AI-based intervention

for diagnosis and surgical pre-assessment of tumors is at the verge of becoming a necessity rather than a luxury. Elaborate characterization of brain tumors with techniques such as volumetric analysis is useful to study their progression and assist in pre-surgical planning [17]. In addition to surgical applications, characterization of delineated tumors can be directly utilized for the prediction of life expectancy [33]. Brain tumor segmentation is at the forefront of all such applications.

Brain tumors are categorized into primary and secondary tumor types. Primary brain tumors originate from brain cells, while secondary tumors metastasize into the brain from other organs. The most common primary brain tumors are gliomas, which arise from brain glial cells and are characterized into low-grade (LGG) and high-grade (HGG) subtypes. High grade gliomas are an aggressive type of malignant brain tumors that grow rapidly and typically require surgery and radiotherapy and have poor survival prognosis [41]. As a reliable diagnostic tool, Magnetic Resonance Imaging (MRI) plays a vital role in monitoring and surgery planning for brain tumor analysis. Typically, several complimentary 3D MRI modalities, such as T1, T1 with contrast agent (T1c), T2 and Fluid-attenuated Inversion Recovery (FLAIR), are required to emphasize different tissue properties and areas of tumor spread. For instance, gadolinium as the contrast agent emphasizes hyperactive tumor sub-regions in the T1c MRI modality [15].

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.

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].

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.

2   Related work

In the previous BraTS challenges, ensembles of U-Net shaped architectures have achieved promising results for multi-modal brain tumor segmentation. Kamnitsas *et al.* [22] proposed a robust segmentation model by aggregating the outputs of various CNN-based models such as 3D U-Net [12], 3D FCN [27] and Deep Medic [23]. Subsequently, Myronenko *et al.* [31] introduced SegResNet, which utilizes a residual encoder-decoder architecture in which an auxiliary branch is used to reconstruct the input data with a variational auto-encoder as a surrogate task. Zhou *et al.* [43] proposed to use an ensemble of different CNN-based networks by taking into account the multi-scale contextual information through an attention block. Zhou *et al.* [21] used a two-stage cascaded approach consisting of U-Net models wherein the first stage computes a coarse segmentation prediction which will be refined by the second stage. Furthermore, Isensee *et al.* [19] proposed the nnU-Net model and demonstrated that a generic U-Net architecture with minor modifications is enough to achieve competitive performance in multiple BraTS challenges.

Transformer-based models have recently gained a lot of attraction in computer vision [14,42,25] and medical image analysis [11,16]. Chen *et al.* [11] introduced a 2D U-Net architecture that benefits from a ViT in the bottleneck of the network. Wang *et al.* [39] extended this approach for 3D brain tumor segmentation. In addition, Xie *et al.* [40] proposed to use a ViT-based model with deformable transformer layers between its CNN-based encoder and decoder by processing the extracted features at different resolutions. Different from these approaches, Hatamizadeh *et al.* [16] proposed the UNETR architecture in which a ViT-based encoder, which directly utilizes 3D input patches, is connected to a CNN-based decoder. UNETR has shown promising results for brain tumor segmentation using the MSD dataset [1]. Unlike the UNETR model, our proposed Swin UNETR architecture uses a Swin transformer encoder which extracts feature representations at several resolutions with a shifted windowing mechanism
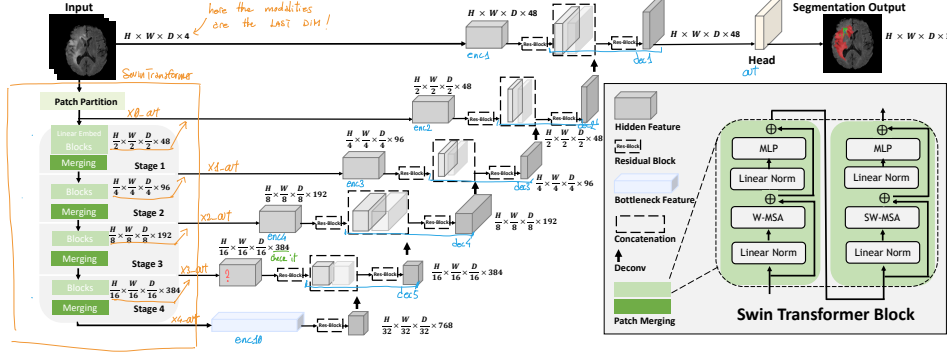
Fig. 1. Overview of the Swin UNETR architecture. The input to our model is 3D multi-modal MRI images with 4 channels. The Swin UNETR creates non-overlapping patches of the input data and uses a patch partition layer to create windows with a desired size for computing the self-attention. The encoded feature representations in the Swin transformer are fed to a CNN-decoder via skip connection at multiple resolutions. Final segmentation output consists of 3 output channels corresponding to ET,WT and TC sub-regions.

for computing the self-attention. We demonstrate that Swin transformers [25] have a great capability of learning multi-scale contextual representations and modeling long-range dependencies in comparison to ViT-based approaches with fixed resolution.

3   Swin UNETR

3.1   Encoder

We illustrate the architecture of Swin UNETR in Fig. 1. The input to the Swin UNETR model $\mathcal{X} \in \mathbb{R}^{H \times W \times D \times S}$ is a token with a patch resolution of (H', W', D') and dimension of $H' \times W' \times D' \times S$. We first utilize a patch partition layer to create a sequence of 3D tokens with dimension of $\lceil \frac{H}{H'} \rceil \times \lceil \frac{W}{W'} \rceil \times \lceil \frac{D}{D'} \rceil$ and project them into an embedding space with dimension C. The self-attention is computed into non-overlapping windows that are created in the partitioning stage for efficient token interaction modeling. Fig. 2 shows the shifted windowing mechanism for subsequent layers. Specifically, we utilize windows of size $M \times M \times M$ to evenly partition a 3D token into $\lceil \frac{H'}{M} \rceil \times \lceil \frac{W'}{M} \rceil \times \lceil \frac{D'}{M} \rceil$ regions at a given layer l in the transformer encoder. Subsequently, in layer $l+1$, the partitioned window regions are shifted by $(\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor)$ voxels. In subsequent

layers of l and $l+1$ in the encoder, the outputs are calculated as

$$
\begin{aligned}
\hat{z}^l &= \text{W-MSA}(\text{LN}(z^{l-1})) + z^{l-1} \\
z^l &= \text{MLP}(\text{LN}(\hat{z}^l)) + \hat{z}^l \\
\hat{z}^{l+1} &= \text{SW-MSA}(\text{LN}(z^l)) + z^l \\
z^{l+1} &= \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1}.
\end{aligned}
\tag{1}
$$

Here, W-MSA and SW-MSA are regular and window partitioning multi-head self-attention modules respectively; $\hat{z}^l$ and $\hat{z}^{l+1}$ denote the outputs of W-MSA and SW-MSA; MLP and LN denote layer normalization and Multi-Layer Perceptron respectively. For efficient computation of the shifted window mechanism, we leverage a 3D cyclic-shifting [25] and compute self-attention according to
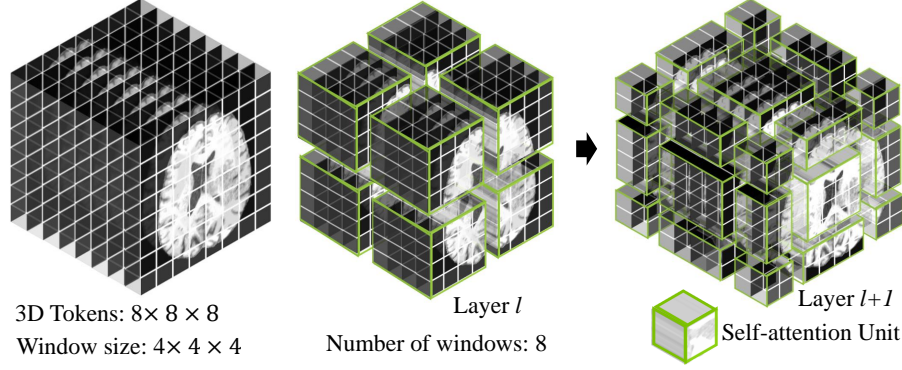
$$
\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V.
\tag{2}
$$

In which Q, K, V denote queries, keys, and values respectively; d represents the size of the query and key.

The Swin UNETR encoder has a patch size of $2 \times 2 \times 2$ and a feature dimension of $2 \times 2 \times 2 \times 4 = 32$, taking into account the multi-modal MRI images with 4 channels. The size of the embedding space C is set to 48 in our encoder. Furthermore, the Swin UNETR encoder has 4 stages which comprise of 2 transformer blocks at each stage. Hence, the total number of layers in the encoder is $L = 8$. In stage 1, a linear embedding layer is utilized to create $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}$ 3D tokens. To maintain the hierarchical structure of the encoder, a patch merging layer is utilized to decrease the resolution of feature representations by a factor of 2 at the end of each stage. In addition, a patch merging layer groups patches with resolution $2 \times 2 \times 2$ and concatenates them, resulting in a $4C$-dimensional feature embedding. The feature size of the representations are subsequently reduced to $2C$ with a linear layer. Stage 2, stage 3 and stage 4, with resolutions of $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}$, $\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}$ and $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$ respectively, follow the same network design.

3.2   Decoder

Swin UNETR has a U-shaped network design in which the extracted feature representations of the encoder are used in the decoder via skip connections at each resolution. At each stage i ($i \in \{0, 1, 2, 3, 4\}$) in the encoder and the bottleneck ($i = 5$), the output feature representations are reshaped into size $\frac{H}{2^i} \times \frac{W}{2^i} \times \frac{D}{2^i}$ and fed into a residual block comprising of two $3 \times 3 \times 3$ convolutional layers that are normalized by instance normalization [37] layers. Subsequently, the resolution of the feature maps are increased by a factor of 2 using a deconvolutional layer and the outputs are concatenated with the outputs of the previous stage. The concatenated features are then fed into another residual block as previously described. The final segmentation outputs are computed by using a $1 \times 1 \times 1$ convolutional layer and a sigmoid activation function.

3D Tokens: 8× 8 × 8

Window size: 4× 4 × 4

Layer l

Number of windows: 8

Layer $l+1$

Self-attention Unit

Fig. 2. Overview of the shifted windowing mechanism. Note that $8 \times 8 \times 8$ 3D tokens and $4 \times 4 \times 4$ window size are illustrated.

Embed Dimension	Feature Size	Number of Blocks	Window Size	Number of Heads	Parameters	FLOPs
768	48	[2,2,2,2]	[7,7,7]	[3,6,12,24]	61.98M	394.84G

Table 1. Swin UNETR configurations.

3.3   Loss Function

We use the soft Dice loss function [30] which is computed in a voxel-wise manner as

$$\mathcal{L}(G,Y) = 1 - \frac{2}{J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I}G_{i,j}Y_{i,j}}{\sum_{i=1}^{I}G_{i,j}^2 + \sum_{i=1}^{I}Y_{i,j}^2}. \tag{3}$$

where I denotes voxels numbers; J is classes number; $Y_{i,j}$ and $G_{i,j}$ denote the probability of output and one-hot encoded ground truth for class j at voxel i, respectively.

3.4   Implementation Details

Swin UNETR is implemented using PyTorch[3] and MONAI[4] and trained on a DGX-1 cluster with 8 NVIDIA V100 GPUs. Table 1 details the configurations of Swin UNETR architecture, number of parameters and FLOPs. The learning rate is set to 0.0008. We normalize all input images to have zero mean and unit standard deviation according to non-zero voxels. Random patches of $128 \times 128 \times 128$ were cropped from 3D image volumes during training. We apply a random axis mirror flip with a probability of 0.5 for all 3 axes. Additionally, we

---

[3] http://pytorch.org/

[4] https://monai.io/
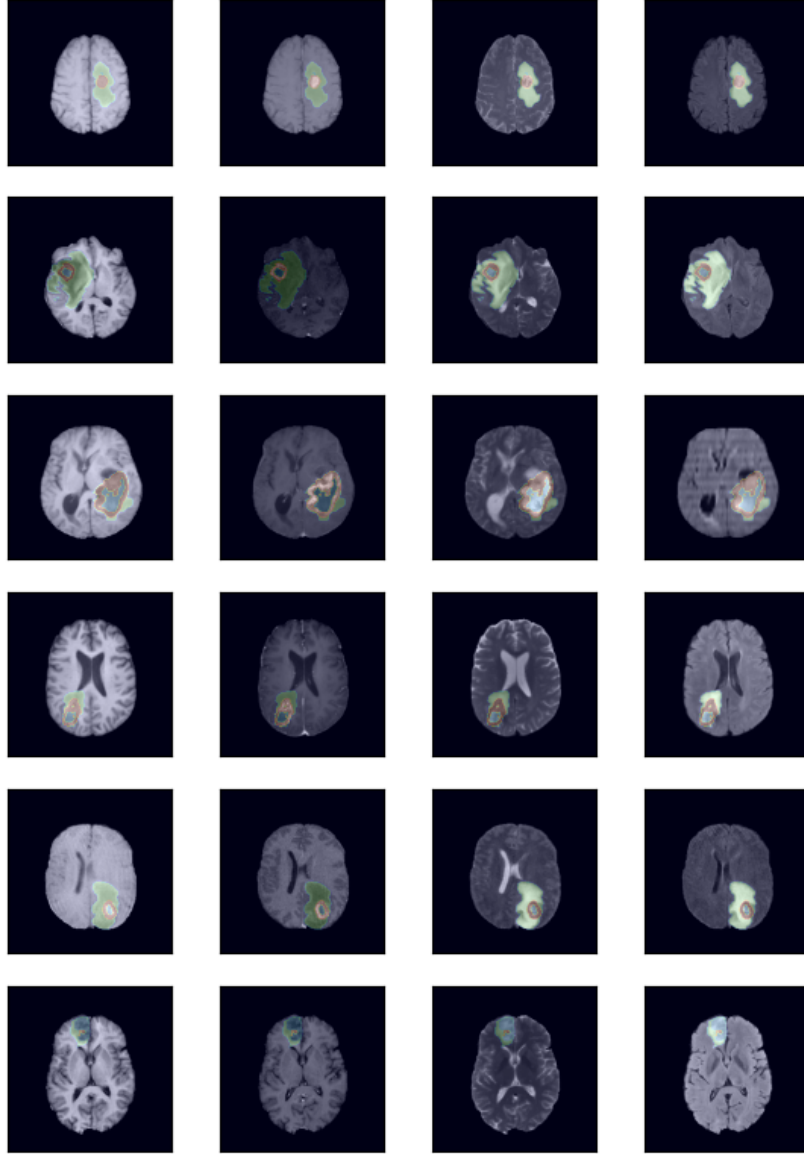
Fig. 3. A typical segmentation example of the predicted labels whic are overlaid on T1, T1c, T2 and FLAIR MRI axial slices in each row. The first two rows depict ∼75th percentile performance based on the Dice score. Rows 3 and 4 depict ∼50th percentile performance while the last two rows are at ∼25th percentile performance. The image intensities are on a gray color scale. The blue, red and green colors correspond to TC, ET and WT sub-regions respectively. Note that all samples have been selected from the BraTS 2021 validation set.

apply data augmentation transforms of random per channel intensity shift in the range $(-0.1, 0.1)$, and random scale of intensity in the range $(0.9, 1.1)$ to input image channels. The batch size per GPU was set to 1. All models were trained for a total of 800 epochs with a linear warmup and using a cosine annealing learning rate scheduler. Fonr inference, we use a sliding window approach with an overlapping of 0.7 for neighboring voxels.

3.5   Dataset and Model Ensembling

The BraTS challenge aims to evaluate state-of-the-art methods for the semantic segmentation of brain tumors by providing a 3D MRI dataset with voxel-wise ground truth labels that are annotated by physicians [6,29,5,3,4]. The BraTS 2021 challenge training dataset includes 1251 subjects, each with four 3D MRI modalities: a) native (T1) and b) post-contrast T1-weighted (T1Gd), c) T2-weighted (T2), and d) T2 Fluid-attenuated Inversion Recovery (T2-FLAIR), which are rigidly aligned, and resampled to a $1 \times 1 \times 1$ mm isotropic resolution and skull-stripped. The input image size is $240 \times 240 \times 155$. The data were collected from multiple institutions using various MRI scanners. Annotations include three tumor sub-regions: the enhancing tumor, the peritumoral edema, and the necrotic and non-enhancing tumor core. The annotations were combined into three nested sub-regions: Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET). Fig. 3 illustrates typical segmentation outputs of all semantic classes. During this challenge, two additional datasets without the ground truth labels were provided for validation and testing phases. These datasets required participants to upload the segmentation masks to the organizers' server for evaluations. The validation dataset, which is designed for intermediate model evaluations, consists of 219 cases. Additional information regarding the testing dataset was not provided to participants.

Our models were trained on BraTS 2021 dataset with 1251 and 219 cases in the training and validation sets, respectively. Semantic segmentation labels corresponding to validation cases are not publicly available, and performance benchmarks were obtained by making submissions to the official server of BraTS 2021 challenge. We used five-fold cross-validation schemes with a ratio of 80:20. We did not use any additional data. The final result was obtained with an ensemble of 10 Swin UNETR models to improve the performance and achieve a better consensus for all predictions. The ensemble models were obtained from two separate five-fold cross-validation training runs.

4   Results and Discussion

We have compared the performance of Swin UNETR in our internal cross validation split against the winning methologies of previous years such as SegResNet [31], nnU-Net [19] and TransBTS [39]. The latter is a ViT-based approach which is tailored for the semantic segmentation of brain tumors.

Table 2. Five-fold cross-validation benchmarks in terms of mean Dice score values. ET, WT and TC denote Enhancing Tumor, Whole Tumor and Tumor Core respectively.

	Swin UNETR				nnU-Net				SegResNet				TransBTS			
Dice Score	ET	WT	TC	Avg.	ET	WT	TC	Avg.	ET	WT	TC	Avg.	ET	WT	TC	Avg.
Fold 1	**0.876**	**0.929**	**0.914**	**0.906**	0.866	0.921	0.902	0.896	0.867	0.924	0.907	0.899	0.856	0.910	0.897	0.883
Fold 2	**0.908**	**0.938**	**0.919**	**0.921**	0.899	0.933	**0.919**	0.917	0.900	0.933	0.915	0.916	0.885	0.919	0.903	0.902
Fold 3	**0.891**	**0.931**	**0.919**	**0.913**	0.886	0.929	0.914	0.910	0.884	0.927	0.917	0.909	0.866	0.903	0.898	0.889
Fold 4	**0.890**	**0.937**	**0.920**	**0.915**	0.886	0.927	0.914	0.909	0.888	0.921	0.916	0.908	0.868	0.910	0.901	0.893
Fold 5	**0.891**	**0.934**	**0.917**	**0.914**	0.880	0.929	**0.917**	0.909	0.878	0.930	0.912	0.906	0.867	0.915	0.893	0.892
Avg.	**0.891**	**0.933**	**0.917**	**0.913**	0.883	0.927	0.913	0.908	0.883	0.927	0.913	0.907	0.868	0.911	0.898	0.891

Table 3. BraTS 2021 validation dataset benchmarks in terms of mean Dice score and Hausdorff distance values. ET, WT and TC denote Enhancing Tumor, Whole Tumor and Tumor Core respectively.

	Dice			Hausdorff (mm)		
Validation dataset	ET	WT	TC	ET	WT	TC
Swin UNETR	0.858	0.926	0.885	6.016	5.831	3.770

Evaluation results across all five folds are presented in Table 2. The proposed Swin UNETR model outperforms all competing approaches across all 5 folds and on average for all semantic classes (e.g. ET, WT, TC). Specifically, Swin UNETR outperforms the closest competing approaches by 0.7%, 0.6% and 0.4% for ET,WT and TC classes respectively and on average 0.5% across all classes in all folds. The superior performance of Swin UNETR in comparison to other top performing models for brain tumor segmentation is mainly due to its capability of learning multi-scale contextual information in its hierarchical encoder via the self-attention modules and effective modeling of the long-range dependencies.

Moreover, it is observed that nnU-Net and SegResNet have competitive benchmarks in these experiments, with nnU-Net demonstrating a slightly better performance. On the other hand, TransBTS, which is a ViT-based methodology, performs sub-optimally in comparison to other models. The sub-optimal performance of TransBTS could be attributed to its inefficient architecture in which the ViT is only utilized in the bottleneck as a standalone attention module, and without any connection to the decoder in different resolutions.

The segmentation performance of Swin UNETR in the BraTS 2021 validation set is presented in Table 3. According to the official challenge results[5], our benchmarks (Team: NVOptNet) are considered as one of the top-ranking methodologies across more than 2000 submissions during the validation phase, hence being the first transformer-based model to place competitively in BraTS challenges. In addition, the segmentation outputs of Swin UNETR for several cases in the validation set are illustrated in Fig. 3. Consistent with quantitative benchmarks, the segmentation outputs are well-delineated for all three sub-regions.

---

[5] https://www.synapse.org/#!Synapse:syn25829067/wiki/612712

Table 4. BraTS 2021 testing dataset benchmarks in terms of mean Dice score and Hausdorff distance values. ET, WT and TC denote Enhancing Tumor, Whole Tumor and Tumor Core respectively.

	Dice			Hausdorff (mm)		
Testing dataset	ET	WT	TC	ET	WT	TC
Swin UNETR	0.853	0.927	0.876	16.326	4.739	15.309

Furthermore, the segmentation performance of Swin UNETR in the BraTS 2021 testing set is reported in Table 4. We observe that the segmentation performance of ET and WT are very similar to those of the validation benchmarks. However, the segmentation performance of TC is decreased by 0.9%.

5   Conclusion

In this paper, we introduced Swin UNETR which is a novel architecture for semantic segmentation of brain tumors using multi-modal MRI images. Our proposed model has a U-shaped network design and uses a Swin transformer as the encoder and CNN-based decoder that is connected to the encoder via skip connections at different resolutions. We have validated the effectiveness of our approach by in the BraTS 2021 challenge. Our model ranks among top-performing approaches in the validation phase and demonstrates competitive performance in the testing phase. We believe that Swin UNETR could be the foundation of a new class of transformer-based models with hierarchical encoders for the task of brain tumor segmentation.

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., van Ginneken, B., et al.: The medical segmentation decathlon. arXiv preprint arXiv:2106.05735 (2021)
2. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
3. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., John Freymann, K.F., Davatzikos, C.: Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. The Cancer Imaging Archive (2017), https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q
4. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., John Freymann, K.F., Davatzikos, C.: Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. The Cancer Imaging Archive (2017), https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF
5. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri

collections with expert segmentation labels and radiomic features. Scientific data **4** (9 2017)

6. Bakas, S., Reyes, M., et Int, Menze, B.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. In: arXiv:1811.02629 (2018)

7. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)

8. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)

9. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)

10. Chen, C., Liu, X., Ding, M., Zheng, J., Li, J.: 3d dilated multi-fiber network for real-time brain tumor segmentation in mri. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 184–192. Springer (2019)

11. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

12. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)

13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)

15. Grover, V.P., Tognarelli, J.M., Crossey, M.M., Cox, I.J., Taylor-Robinson, S.D., McPhail, M.J.: Magnetic resonance imaging: principles and techniques: lessons for clinicians. Journal of clinical and experimental hepatology **5**(3), 246–255 (2015)

16. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., Xu, D.: Unetr: Transformers for 3d medical image segmentation. arXiv preprint arXiv:2103.10504 (2021)

17. Hoover, J.M., Morris, J.M., Meyer, F.B.: Use of preoperative magnetic resonance imaging t1 and t2 sequences to determine intraoperative meningioma consistency. Surgical neurology international **2** (2011)

18. Huo, Y., Xu, Z., Xiong, Y., Aboud, K., Parvathaneni, P., Bao, S., Bermudez, C., Resnick, S.M., Cutting, L.E., Landman, B.A.: 3d whole brain segmentation using spatially localized atlas network tiles. NeuroImage **194**, 105–119 (2019)

19. Isensee, F., Jaeger, P.F., Full, P.M., Vollmuth, P., Maier-Hein, K.: nnu-net for brain tumor segmentation. In: BrainLes@MICCAI (2020)

20. Isensee, F., Jäger, P.F., Full, P.M., Vollmuth, P., Maier-Hein, K.H.: nnu-net for brain tumor segmentation. In: International MICCAI Brainlesion Workshop. pp. 118–132. Springer (2020)

21. Jiang, Z., Ding, C., Liu, M., Tao, D.: Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In: International MICCAI Brainlesion Workshop. pp. 231–241. Springer (2019)

22. Kamnitsas, K., W. Bai, E.F., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., Glocker, B.: Ensembles of multiple models and architectures for robust brain tumour segmentation. In: International Conf. on Medical Image Computing and Computer Assisted Intervention. Multimodal Brain Tumor Segmentation Challenge (MICCAI, 2017). LNCS (2017)

23. Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical image analysis **36**, 61–78 (2017)

24. Liu, D., Zhang, H., Zhao, M., Yu, X., Yao, S., Zhou, W.: Brain tumor segmention based on dilated convolution refine networks. In: 2018 IEEE 16th International Conference on Software Engineering Research, Management and Applications (SERA). pp. 113–120. IEEE (2018)

25. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)

26. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. arXiv preprint arXiv:2106.13230 (2021)

27. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)

28. Louis, D.N., Ohgaki, H., Wiestler, O.D., Cavenee, W.K., Burger, P.C., Jouvet, A., Scheithauer, B.W., Kleihues, P.: The 2007 who classification of tumours of the central nervous system. Acta neuropathologica **114**(2), 97–109 (2007)

29. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E.R., Weber, M.A., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, C., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Raviv, T.R., Reza, S.M.S., Ryan, M.T., Sarikaya, D., Schwartz, L.H., Shin, H.C., Shotton, J., Silva, C.A., Sousa, N., Subbanna, N.K., Szekely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Unal, G.B., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Leemput, K.V.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015)

30. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. IEEE (2016)

31. Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: BrainLes, Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 311–320. LNCS, Springer (2018), https://arxiv.org/abs/1810.11654

32. Myronenko, A., Hatamizadeh, A.: Robust semantic segmentation of brain tumor regions from 3d mris. In: International MICCAI Brainlesion Workshop. pp. 82–89. Springer (2019)

33. Nie, D., Zhang, H., Adeli, E., Liu, L., Shen, D.: 3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. In: International conference on medical image computing and computer-assisted intervention. pp. 212–220. Springer (2016)
34. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? Advances in Neural Information Processing Systems **34** (2021)
35. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
36. Tang, Y., Yang, D., Li, W., Roth, H., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. arXiv preprint arXiv:2111.14791 (2021)
37. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
39. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 109–119. Springer (2021)
40. Xie, Y., Zhang, J., Shen, C., Xia, Y.: Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. arXiv preprint arXiv:2103.03024 (2021)
41. Zacharaki, E.I., Wang, S., Chawla, S., Soo Yoo, D., Wolf, R., Melhem, E.R., Davatzikos, C.: Classification of brain tumor type and grade using mri texture and shape in a machine learning scheme. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine **62**(6), 1609–1618 (2009)
42. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6881–6890 (2021)
43. Zhou, C., Chen, S., Ding, C., Tao, D.: Learning contextual and attentive information for brain tumor segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2018). Multimodal Brain Tumor Segmentation Challenge (BraTS 2018). BrainLes 2018 workshop. LNCS, Springer (2018)