Deep Learning-Based 3D Brain Tumor Segmentation: A Novel Approach for Accurate and Efficient Diagnosis

Navodita Mathur, Yushui Han, Pengyu Chen
School of Computing and Information Science
University of Pittsburgh

1. Introduction

1.1. Problem Statement

Brain tumor1 image segmentation plays a crucial role in medical diagnosis and treatment planning, offering insights into tumor characteristics and assisting clinicians in making informed decisions. This task involves partitioning magnetic resonance imaging (MRI) scans of the brain into different regions, such as tumor core, peritumoral edema, and healthy brain tissue. Accurate segmentation is essential for quantifying tumor size, assessing treatment response, and guiding surgical interventions.

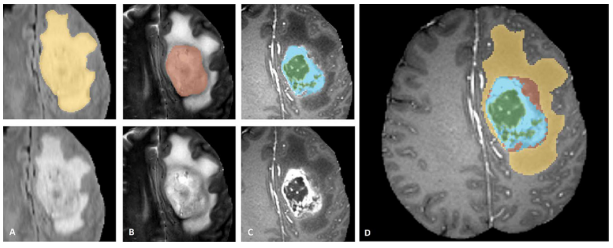

Figure 1. Manual annotation through expert raters. Shown are image patches with the tumor structures that are annotated in the different modalities (top left) and the final labels for the whole dataset (right). Image patches show from left to right: the whole tumor visible in FLAIR (A), the tumor core visible in T2 (B), the enhancing tumor structures visible in T1c (blue), surrounding the cystic/necrotic components of the core (green) (C). Segmentations are combined to generate the final labels of the tumor structures (D): edema (yellow), non-enhancing solid core (red), necrotic/cystic core (green), enhancing core(blue).[16]

The proposed deep learning model for brain tumor segmentation is of profound interest due to its potential to fundamentally transform clinical practice. By automating the segmentation process, it offers the promise of expedited and precise identification of tumor regions within neuroimaging data. This not only streamlines treatment planning procedures but also mitigates inter-observer variability, thus enhancing the reliability and consistency of tumor delineation. Furthermore, the integration of advanced deep learning techniques in medical image analysis underscores the interdisciplinary nature of this research endeavor, offering compelling avenues for collaboration between computer science and medical domains.

1.2. Challenge

Brain tumor segmentation presents significant challenges due to the complex and heterogeneous nature of tumors, variability in imaging data, and inherent limitations of medical imaging technology. Tumors exhibit diverse characteristics in terms of size, shape, and appearance, making it difficult to accurately delineate tumor boundaries. Moreover, MRI scans, the primary modality for brain imaging, often suffer from noise, artifacts, and limited resolution, further complicating segmentation tasks. Additionally, tumors may overlap with normal brain structures, leading to ambiguity in segmentation. Inter-observer variability and clinical variability add another layer of complexity, requiring segmentation algorithms to be robust and adaptable. Overcoming these challenges demands the development of sophisticated algorithms that can handle tumor heterogeneity, noise, and variability in imaging data while providing accurate and reliable segmentation results to assist clinicians in diagnosis and treatment planning. Moreover, the imperative to achieve high levels of accuracy while preserving computational efficiency underscores the demand for innovative methodologies capable of balancing computational complexity with clinical utility.

1.3. Importance

The proposed research holds profound significance within the realm of neuro-oncology, offering tangible benefits for both patients and healthcare providers. By advancing the state-of-the-art in brain tumor segmentation, the proposed deep learning model stands to catalyze paradigm shifts in clinical decision-making and patient management. Rapid and accurate tumor delineation not only expedites treatment planning processes but also empowers clinicians with quantitative insights into tumor burden and spatial

distribution. Furthermore, the proposed model's potential to reduce diagnostic turnaround times and enhance treatment efficacy holds implications for resource optimization and improved patient outcomes. Thus, by addressing the pressing clinical need for reliable and efficient tumor segmentation methodologies, this research contributes to the overarching goal of advancing precision medicine in neuro-oncology.

1.4. Data to Use

The RSNA-ASNR-MICCAI BraTS 2021 challenge utilizes multi-institutional pre-operative baseline multi-parametric magnetic resonance imaging (mpMRI) scans, and focuses on the evaluation of state-of-the-art methods for (Task 1) the segmentation of intrinsically heterogeneous brain glioblastoma sub-regions in mpMRI scans.

MRI scans are essential imaging modalities in medical diagnosis, particularly for brain imaging. Among the four main types—T1-weighted, T1-weighted with gadolinium contrast, T2-weighted, and FLAIR—each provides unique information about brain anatomy and pathology. T1-weighted scans offer detailed anatomical information, while gadolinium-enhanced T1-weighted scans enhance lesion visibility. T2-weighted scans highlight tissue water content differences, and FLAIR scans suppress cerebrospinal fluid signal, aiding in lesion detection near CSF-filled spaces. Together, these scans enable comprehensive assessment and diagnosis in neurological conditions.

BraTS 2021 Dataset Training Set consists of scans from 1250 patients spread across 4 MRI Scans: T1, T1Gd, T2 and FLAIR. Fused Expert Segmentation of 4 Classes:
1. Non-Tumor (Not Shown) [0]
2. Edema: Green [1]
3. Neurotic and Non-enhancing Tumor: Yellow [2]
4. GD-Enhancing Tumor: Blue [3]
And when using it, we combined these three labels(back ground did not change) as three new labels tc, wt and et. And tc means tumor core, wt means whole tumor, et means enhancing tumor. These label is transformed from the original one.

2. Related Work

In the early stages of brain tumor segmentation research, traditional image processing techniques such as thresholding, region growing, and edge detection were commonly used[8]. While these methods were straightforward, they often struggled with handling noise, variability in tumor appearance, and complex tumor shape. This study[1] uses a filter with different wavelet bands for noise reduction and to enhance the region of interest (ROI), along with PF clustering for segmentation purposes.

2.1. Machine-learning based Approaches

With the advancement of machine learning techniques, researchers began turning to supervised learning algorithms to tackle the task of brain tumor segmentation. These algorithms, including support vector machines (SVMs), random forests (RF), k-nearest neighbors (KNN), and linear discriminant analysis (LDA), offer the capability to learn discriminative features directly from the data, thereby improving segmentation accuracy.These methods showed improvements over traditional techniques by learning discriminative features directly from the data. In a study by [25], various machine learning algorithms, namely KNN, RF, SVM, and LDA, were employed to classify MR brain image features. The research concluded that SVM, with an accuracy rate of 90%, outperformed other algorithms in classifying brain image features effectively. Similarly, in another study [12], multimodal features such as texture, morphological, entropy-based, Scale Invariant Feature Transform (SIFT), and Elliptic Fourier Descriptors (EFDs) were extracted from a brain tumor imaging database. Robust machine learning techniques, including Support Vector Machine (SVM) with polynomial, Radial Base Function (RBF), and Gaussian kernels, as well as Decision Tree (DT) and Naïve Bayes, were employed to detect tumors based on these features. The study found that Naïve Bayes followed by Decision Tree yielded the highest detection accuracy, particularly when considering entropy, morphological, SIFT, and texture features.

Brain tumor detection involves identifying the presence or absence of a tumor within medical images, typically using binary classification techniques. Classification aims to categorize tumors into different types or grades based on their characteristics, assisting in treatment planning and prognosis. Segmentation, on the other hand, involves delineating the boundaries of tumors within images, providing detailed spatial information essential for surgical planning and monitoring tumor progression.

2.2. Convolution-based Segmentation Methods

The advent of deep learning, particularly convolutional neural networks (CNNs), revolutionized medical image segmentation, including brain tumor segmentation. CNNs handle 3D data typically extended 2D convolutional operations to 3D by using 3D convolutional kernels. These 3D kernels operate across the three spatial dimensions of the input volume, allowing the network to capture spatial relationships in 3D space. Pooling layers such as max pooling or average pooling were also extended to 3D to

reduce spatial dimensions and extract relevant features from volumetric data. This article[13] presents a deep convolutional neural network (CNN) to segment brain tumors in MRIs. The network architecture consists of multiple neural network layers connected in sequential order with the feeding of convolutional feature maps at the peer level. This paper[4] introduces a two-pathway model for brain tumor image segmentation, integrating average and max pooling layers to capture diverse features. Additionally, it incorporates 1×1 convolutional kernels and a fully connected Conditional Random Field (FCRF) mixture model to enhance segmentation accuracy by leveraging global context information.

Models like U-Net[19], introduced by Ronneberger et al. in 2015, became popular for their ability to capture spatial information and hierarchical features through encoder-decoder architectures with skip connections. [6] uses the UNet architecture, one of the deep learning networks, as a hybrid model with a pre-trained DenseNet121 architecture for the segmentation process, leading to improved performance as compared to the then commonly used models."Automatic Brain Tumor Segmentation using Cascaded Anisotropic Convolutional Neural Networks"[22] proposes a novel approach to brain tumor segmentation leveraging cascaded anisotropic convolutional neural networks (CNNs). This method aims to enhance the segmentation accuracy by incorporating anisotropic convolutions, which adapt to the varying resolutions of medical images. The cascaded architecture iteratively refines the segmentation results, leading to improved performance in delineating brain tumor boundaries in medical imaging data.

2.3. Transformer-based Segmentation Methods

CNNs, however, have limited ability to capture long-range dependencies and global context in images, which can be crucial for tasks such as image captioning or dense prediction tasks like brain tumor segmentation. More recently, transformers, initially designed for natural language processing tasks, have been adapted to medical image analysis, including brain tumor segmentation. Models like Vision Transformer (ViT) and Swin Transformer have shown promise in capturing spatial dependencies and extracting meaningful features from medical images. This paper[20] introduced ViTBIS which leverages the Vision Transformer (ViT) architecture, originally designed for natural image processing, to effectively segment biomedical images by transforming the input images into sequences of patches and processing them through self-attention mechanisms, thereby capturing both local and global features for accurate segmentation.

2.4. Hybrid Segmentation Methods

Researchers have also explored hybrid architectures that combine the strengths of CNNs and transformers for brain tumor segmentation. These models leverage the feature extraction capabilities of CNNs and the attention mechanisms of transformers to achieve state-of-the-art performance. TransUnet[5] and TransBTS[23] are a kind of hybrid model in combining CNN and Transformer, using successive convolutional layers and Transformer in the encoder for feature extraction and transposed convolution for upsampling operations in the decoder to recover spatial resolution for semantic segmentation. UnetR[9], on the other hand, used ViT layers as encoders and convolutional layers as decoders to build the network. The method achieved excellent performance on several tasks, but the model resulted in a large number of parameters due to a large number of ViT layers used. In order to reduce the number of parameters, there have been several attempts since. VT-Unet [18] is a lightweight model for segmenting 3D medical images in a hierarchical manner. It introduces two self-attention layers in the encoder to capture local and global information. This model also introduces window-based self-attention, cross-attention modules, and Fourier position coding in the decoder part to significantly improve accuracy and efficiency. Cotr [24] designed a deformable transformer encoder, which focuses on only a small portion of the key location feature information, which also greatly reduces the computational complexity and spatial complexity. The experimental results show that this method has a significant improvement in effectiveness compared to other transformer and CNN combination methods. In SwinBTS architecture[14], the authors leverage the use of Swin Transformer to initially extract the image features while using the CNN as the backbone network in both its encoders and decoders. Similarly, Swin-UNETR[10] combines the Swin Transformer with the U-Net architecture, utilizing the Swin Transformer for feature extraction and the U-Net structure for segmentation.

Various other architectures have employed Swin transformers due to shift windows mechanism, which have shown promising results as compared to vision transformers, particularly in the field of medical image segmentation. DS-TransUNet[15] benefits from the self-attention computation in swin transformer and the designed dual-scale encoding, which can effectively model the non-local dependencies and multiscale contexts for enhancing the semantic segmentation quality of varying medical images. In this article[7], a novel method called Swin Pyramid Aggregation network (SwinPA-Net) is proposed by combining two designed modules, named dense multiplicative connection (DMC) module and local pyramid attention (LPA) module, with Swin Transformer to learn more powerful and robust

features.

The U-Net structure, while successful, still faces challenges in segmentation performance due to the semantic gap between encoding and decoding stages. This gap can hinder feature fusion, as low-level features crucial for edge segmentation and deep-level features essential for object recognition may not be effectively explored.

Inspired by swin-unet 3D[3][2], and Swin-UNETR is our model which incorporates an attention mechanism and spatial squeeze and excitation to enhance the local features, and at the same time working to decrease this gap between encoder and decoder. While the Swin Transformer backbone is effective at capturing long-range dependencies, further improvements may be necessary, especially in scenarios where spatial relationships across distant regions are critical for accurate segmentation. Spatial Attention can enhance the model's ability to capture such dependencies by facilitating direct communication between tokens across spatial dimensions.

Continued advancements in these models aim to provide clinicians with more accurate and reliable tools for diagnosing and treating brain tumors, ultimately leading to improved patient outcomes and quality of life.

3. Proposed Method

3.1. Architecture

The architecture is shown in the figure 3. It is defined by the contracting path and the expanding path. On the contracting path, it can be seen that it has swin transformers. This combination captures contextual information and downsamples the feature maps.

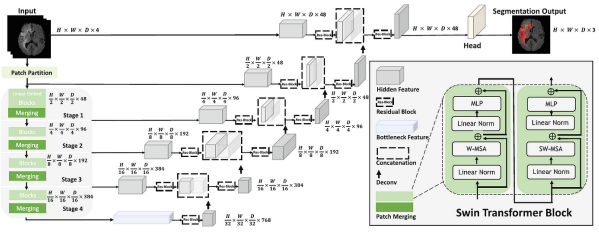

Figure 2. Swin-UNETR Architecture[10]

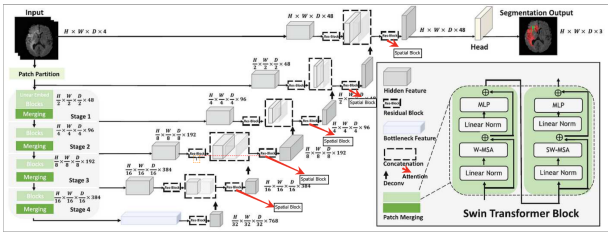

Figure 3. Attention Swin-UNETR Architecture

3.1.1 Encoder

The encoder comprises of Swin transformer that blocks that initially divides the input data into non-overlapping 2×2×2 patches and uses a patch partition layer to create windows of the desired size for computing self-attention. Considering multi-modal MRI images with 4 channels, the encoder will have a patch size of 2×2×2 and a feature dimension of 32 (2×2×2×4=32). The size of the embedding space C is set to 48, and the encoder is divided into 4 stages, each containing 2 transformer blocks, making the total number of layers in the encoder L=8. At the end of each stage, a patch merging layer will be utilized to decrease the resolution of feature representations by a factor of 2 and group patches with a 2×2×2 resolution and concatenate them, resulting in a 4C-dimensional feature embedding. Subsequently, the feature size of the representations will be reduced to 2C with a linear layer. Each stage also has a CNN based basic block comprising of 2 convolution layers before being given to decoder.

3.1.2 Decoder

The decoder has 2 residual blocks consisting of 3*3*3 convolution layers in addition to Attention block and Squeeze and Excitation block. This is done in several stages comprising of: 1. Increasing the resolution of feature maps by a factor of 2 using a deconvolutional layer. 2. Employing attention to the output with features from the previous stage to enhance feature integration across resolutions. 3. Feeding the reshaped features into a residual block comprising two 3×3×3 convolutional layers and squeeze and excitation block.

The final segmentation outputs are computed by using a 1×1×1 convolutional layer and a sigmoid activation function.

3.1.3 Swin transformer

The architecture has Swin transformer4 as its backbone. It is composed of multiple stages, each containing 2 Transformer blocks. Unlike traditional Transformers, which process images globally and require extensive computational resources, the Swin Transformer adopts a shifted window mechanism to reduce computation. This mechanism allows

each token to attend only to nearby tokens within a local window, significantly reducing the number of pairwise interactions. Furthermore, the Swin Transformer employs patch partitioning, dividing each input patch into smaller local windows to capture fine-grained spatial information and long-range dependencies.

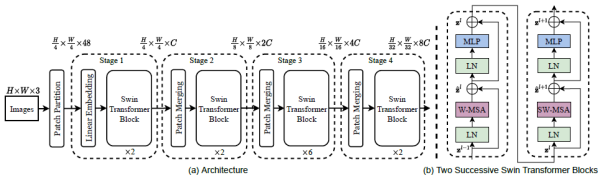

Figure 4. Swin Transformer[21]

Within each stage, the model processes images at multiple resolutions, enabling it to capture both global context and fine details across different scales.

3.1.4 Attention

Attention mechanism similar to that in attention-unet is employed to capture long-range dependencies and relationships within an image instead of concatenation. The architecture employed here is relatively simple, as shown in Fig 5. It operates on two sets of feature maps known as the ”query” tensor (g) and the ”key” tensor (x). These tensors represent different levels of abstraction or spatial information within the network. It computes attention weights based on the similarity between features in the ”query” and ”key” tensors, using an attention mechanism implemented with convolutional layers and activation functions, sigmoid and leaky ReLU.The attended feature map is combined with the original input tensor to produce the final output of the AttentionBlock. It allows the network to selectively attend to relevant information while suppressing irrelevant features, improving its ability to capture spatial dependencies.

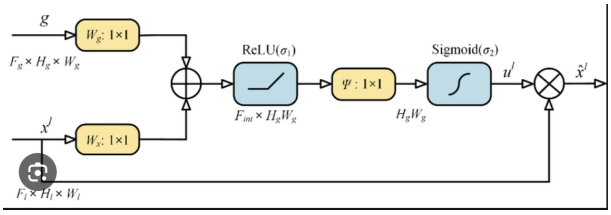

Figure 5. Attention[17]

3.1.5 Squeeze and Excitation Block

This aims to capture both spatial and channel-wise dependencies within feature maps, thereby improving the discriminative ability of the network. The structure of the SE building block is depicted in fig. 6. Initially, it has 2 blocks comprising of 3*3*3 convolution layer. The features obtained U are first passed through a squeeze operation, which produces a channel descriptor by aggregating feature maps across their spatial dimensions (H × W). The function of this descriptor is to produce an embedding of the global distribution of channel-wise feature responses, allowing information from the global receptive field of the network to be used by all its layers. The aggregation is followed by an excitation operation, which takes the form of a simple self-gating mechanism that takes the embedding as input and produces a collection of per-channel modulation weights. These weights are applied to the feature maps U to generate the output of the SE block. The output is then combined with the input from the bottom layer to generate the final output of the layer.

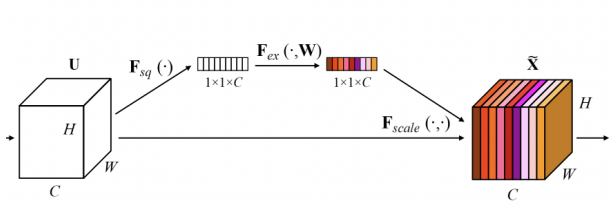

Figure 6. Squeeze and Excitation Block[11]

4. Evaluation

The evaluation considers the predicted mask against the ground truth mask for each pixel in the image. In binary mask problems, this means that 1 is equivalent to foreground while 0 is background. This yields a percentage of pixels that were correctly classified. However, multi-category problems consider all categories. In the BRATS Challenge, we consider voxels as categories and evaluate the performance using the metrics described below.

4.1. Metrics

4.1.1 Dice Coefficient

The Dice Coefficient, or F1 Score, measures the similarity between the two sets. It is calculated using the following formula:

$$DiceCoefficient = \frac{2|A \cap B|}{|A| + |B|}. \quad (1)$$

Where intersection is the number of pixels that are common to both the predicted and ground truth sets and the total number of pixels in both sets is the sum of pixels in the

predicted set and the ground truth set.

The Dice Coefficient ranges from 0 to 1, with 0 indicating no overlap between the sets (complete dissimilarity) and 1 indicating a perfect match (complete similarity). Higher Dice Coefficient values correspond to better segmentation performance. In the context of image segmentation, the sets being compared are often the pixels predicted by a model (the segmentation mask) and the true segmentation mask (the ground truth). The Dice Coefficient provides a measure of how well the predicted segmentation aligns with the actual segmentation.

When applied to multi-class segmentation tasks, the Dice coefficient is calculated for each class separately, and then averaged to obtain a single value representing the average performance across all classes. This metric is often referred to as the average Dice coefficient or mean Dice coefficient (mDSC).

To calculate the average Dice coefficient across N classes, we compute the Dice coefficient for each class and then take the average. The formula for calculating the mDSC is:

$$mDSC = \frac{1}{N} \sum_{i=1}^{N} DSC_i \tag{2}$$

Where, DSC_i is the Dice coefficient for class i and N is the total number of classes.

This average Dice coefficient provides a comprehensive assessment of the segmentation model's performance across all classes, taking into account both the accuracy and the spatial overlap of the predicted and ground truth segmentations.

4.2. Loss

The soft Dice loss is a variant of the Dice coefficient used as a loss function in training neural networks for segmentation tasks. It is particularly common in medical image segmentation due to its effectiveness in handling class imbalance, which is often prevalent in medical data. It penalizes deviations between predicted and ground truth probabilities, encouraging the model to produce probability distributions that align better with the ground truth while considering all classes simultaneously. It provides a differentiable loss function that can be optimized using gradient descent methods during training.

It is calculated as follows:

$$SoftDiceLoss = 1 - \frac{2 \times \sum_{i=1}^{N} p_i \times g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2} \tag{3}$$

Where, p_i represents the predicted probability (or value) for class i, g_i represents the ground truth label for class i, and N is the total number of classes.

By using soft Dice loss as the optimization objective, neural networks can learn to produce segmentation outputs that optimize spatial overlap and accuracy simultaneously, which is crucial for tasks like medical image segmentation where accurate delineation of structures is vital.

5. Implementation

The model was trained for about 180 epochs. The input image size of each mode is 240 × 240 × 155, which has been aligned and resampled to a 1 × 1 × 1 mm isotropic resolution and skull-stripped. The labels include a background (Label 0) and three tumor categories, namely necrotic and non-enhancing tumors (Label 1), peritumoral edema (Label 2), and enhancing tumors (Label 4). The three categories were combined into three nested sub- regions: whole tumor (WT, Labels 1, 2, 4), tumor core (TC, Labels 1, 4), and enhancing tumor (ET, Label 4).

3 separate mechanisms attention, self-attention, and spatial and channel attention mechanisms were tested on a part of dataset to test the performance and efficiency, keeping the number of parameters in mind, before a combination of attention and spatial and channel attention was chosen for the entire data to be trained on.

The model was trained in 2 parts. For the first 120 epochs, the learning rate was 1e-4 before decreasing to 1e-5.

5.1. Augmentation

In order to improve the model generalization and mitigate over-fitting, we used some augmentation techniques as below:

2.1 Foreground Cropping:

Crop the foreground region from the image and label.

2.2 Random Spatial Cropping:

Randomly crop a region of interest (ROI) from the image and label with a fixed size.

2.3 Random Flipping:

Randomly flip the image and label along the spatial axes (x, y, z) with a probability of 0.5.

2.4 Random Intensity Scaling:

Randomly scale the intensity of the image with a factor sampled from a uniform distribution within the range [-0.1, 0.1].

2.5 Random Intensity Shifting:

Randomly shift the intensity of the image with an offset sampled from a uniform distribution within the range [-0.1, 0.1].

5.2. Hyperparameters

learning rate:1e-4
roi = (128, 128, 128)
batch_size = 2
sw_batch_size = 4 (which means slide window batch)
weight_decay=1e-5
lr_scheduler.CosineAnnealingLR (To balance the speed and performance)

5.3. Post Processing

Given the initial predictions and pairwise potentials, CRF post-processing performs inference to refine the predictions. The goal is to find the labeling configuration that maximizes the overall compatibility with both the initial predictions and the pairwise potentials. The output of CRF post-processing is the refined predictions, which have been adjusted to better align with the underlying structure of the input data.

6. Results

6.1. Results with Proposed Architecture

This is the current saved best model's performance on test data:
Accuracy: dice_tc: 0.9172699, dice_wt: 0.9375892, dice_et: 0.84364265
Which tc means tumor core, wt means whole tumor, et means enhancing tumor. These label is transformed from the original ones in order to compare with the baseline because they did the same thing for step training.

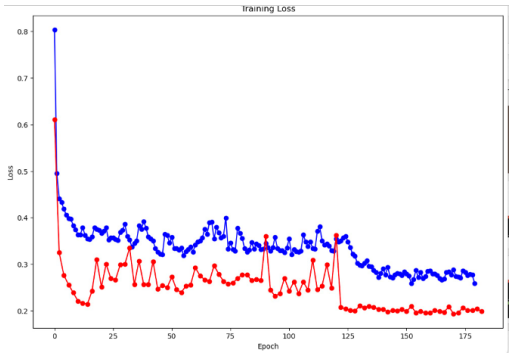

Figure 7. Loss

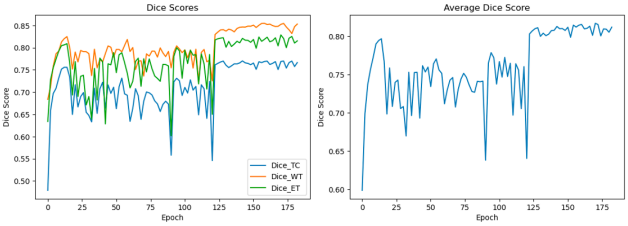

Figure 8. Performance measurement

6.2. Prediction

This is an example of prediction using the trained model:

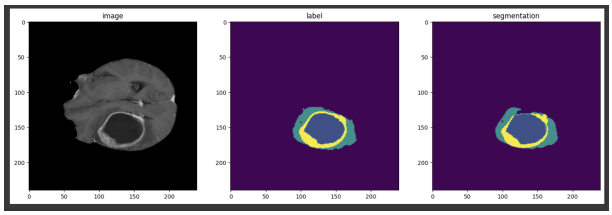

Figure 9. The left one is original slice, middle is ground truth, right is our prediction

6.3. After post-processing

Table 1. Performance after post-processing

	TC	WT	ET
Original dice	0.7741752	0.86161256	0.84151834
Dice after CRF	0.7741761	0.8550863	0.8415188

We can see from the table that implementing CRF can increase the dice of central tumor and enhancing tumor. So CRF is useful in these two labels.

7. Comparison

nn-UNet is based on the U-Net architecture, which consists of an encoder-decoder structure with skip connections. It has demonstrated strong performance in various medical imaging tasks due to its ability to capture spatial information effectively. It excels in tasks where detailed local features are crucial, such as organ delineation and lesion segmentation. However,it requires extensive data augmentation and preprocessing techniques to handle data variability and improve generalization. Additionally, processing 3D volumetric data with these models can be computationally expensive, especially for large-scale datasets.

In contrast, Swin-UNetR adopts the Swin Transformer architecture, which replaces conventional convolutional

layers with self-attention mechanisms and tokenization strategies. This architecture allows Swin-UNetR to capture global context and handle long-range dependencies more efficiently. Swin-UNetR processes image patches using parallelizable self-attention mechanisms, potentially offering advantages in computational efficiency and scalability, particularly for large-scale datasets. Moreover, Swin-UNetR has shown promising results with relatively less data augmentation due to its ability to capture long-range dependencies more effectively.

The performance of proposed model was compared against Swin-UNETR and NNUnet

7.1. NNunet

NNunet is a framework which is not focussed on one model development but on preprocessing and postprocessing of images. Using this framework we can know how important these are. Training loss, duration and adaptive learning rate are shown in Figure 10:
We know it uses:
1. preprocessing: resize to voxel space = [1,1,1]
2. augmentation:crop, mirror and gamma correction.
3. postprocessing: first try the eliminate smaller part of segmentation results in each patch before combine them. If this improve the dice generally on validate set, this method will be used for this label.
Final result: average dice: 0.9087

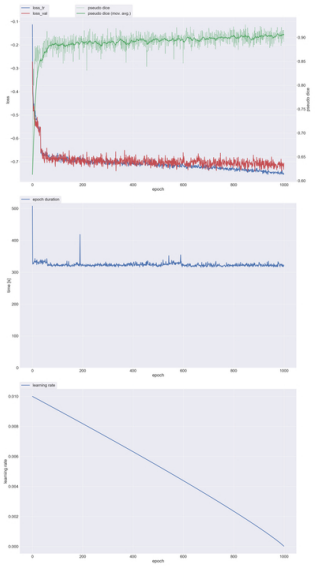

Figure 10. Loss,duration and learning rate of NN-unet

7.2. Swin-UNETR

Swin-UNETR 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. This was trained for 800 epochs using 5 fold cross-validation and using 5 ensemble Swin-UNETRs. Our experiment using Swin-UNETR was trained for 180 epochs using fold 1 resulting in 0.79 dice score.

7.3. Comparision

Table 2. Model Performance

Model	Average Dice Score
NNUnet	0.9087
SwinUNETR	0.913
Attention SwinUNETR	0.8167

8. Conclusion and discussion

1. We did not beat the SOTA(Swin UNETR), the reason maybe the attention should be implemented in a better position. Our attention may payed attention to less informative parts
2. Postprocessing is useful in improving performance.Then testing more post-processing ways like morphological operations may be helpful in improving the performance.
3. Adding attention requires more time to train but leads to faster convergence.

9. Future work

We plan to use more efficient transformers such as ELSA swin transformer to help in scaling up the capacity without compromising the efficiency and performance of the model. We further plan to improve performance using cross-validation to further generalize the model.

References

[1] Javaria Amin, Muhammad Sharif, Mudassar Raza, Tanzila Saba, and Muhammad Almas Anjum. Brain tumor detection using statistical and machine learning method. *Computer Methods and Programs in Biomedicine*, 177:69–79, 2019. 2

[2] Yimin Cai, Yuqing Long, Zhenggong Han, Mingkun Liu, Yuchen Zheng, Wei Yang, and Liming Chen. Swin unet3d: a three-dimensional medical image segmentation network combining vision transformer and convolution. *BMC Medical Informatics and Decision Making*, 23, 2023. 4

[3] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet:

Unet-like pure transformer for medical image segmentation, 2021. 4

[4] Jie Chang, Luming Zhang, Naijie Gu, Xiaoci Zhang, Minquan Ye, Rongzhang Yin, and Qianqian Meng. A mix-pooling cnn architecture with fcrf for brain tumor segmentation. *Journal of Visual Communication and Image Representation*, 58:316–322, 2019. 3

[5] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation, 2021. 3

[6] Necip Cinar, Alper Ozcan, and Mehmet Kaya. A hybrid densenet121-unet model for brain tumor segmentation from mr images. *Biomedical Signal Processing and Control*, 76: 103647, 2022. 3

[7] Hao Du, Jiazheng Wang, Min Liu, Yaonan Wang, and Erik Meijering. Swinpa-net: Swin transformer-based multiscale feature pyramid aggregation network for medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2022. 3

[8] Nelly Gordillo, Eduard Montseny, and Pilar Sobrevilla. State of the art survey on mri brain tumor segmentation. *Magnetic Resonance Imaging*, 31(8):1426–1438, 2013. 2

[9] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation, 2021. 3

[10] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, 2022. 3, 4

[11] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019. 5

[12] Lal Hussain, Sharjil Saeed, Imtiaz Ahmed Awan, Adnan Idris, Malik Sajjad Ahmed Nadeem, and Qurat-Ul-Ain Chaudhry. Detecting brain tumor using machines learning techniques based on different features extracting strategies. *Current medical imaging reviews*, 15 6:595–606, 2019. 2

[13] Sajid Iqbal, M. Usman Ghani, Tanzila Saba, and Amjad Rehman. Brain tumor segmentation in multi-spectral mri using convolutional neural networks (cnn). *Microscopy Research and Technique*, 81(4):419–427, 2018. 3

[14] Yun Jiang, Yuan Zhang, Xin Lin, Jinkun Dong, Tongtong Cheng, and Jing Liang. Swinbts: A method for 3d multimodal brain tumor segmentation using swin transformer. *Brain Sciences*, 12(6), 2022. 3

[15] Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, Guangming Lu, and David Zhang. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 71:1–15, 2022. 3

[16] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. 1

[17] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018. 5

[18] Himashi Peiris, Munawar Hayat, Zhaolin Chen, Gary Egan, and Mehrtash Harandi. A robust volumetric transformer for accurate 3d tumor segmentation, 2022. 3

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 3

[20] Abhinav Sagar. Vitbis: Vision transformer for biomedical image segmentation, 2022. 3

[21] Sik-Ho Tsang. Review: Swin transformer, Feb 22, 2022. 5

[22] Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. *Automatic Brain Tumor Segmentation Using Cascaded Anisotropic Convolutional Neural Networks*, page 178–190. Springer International Publishing, 2018. 3

[23] Wenxuan Wang, Chen Chen, Meng Ding, Jiangyun Li, Hong Yu, and Sen Zha. Transbts: Multimodal brain tumor segmentation using transformer, 2021. 3

[24] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation, 2021. 3

[25] Gökalp Çınarer and Bülent Gürsel Emiroğlu. Classificatin of brain tumors by machine learning algorithms. In *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–4, 2019. 2