Automatic Brain Tumor Segmentation: Exploring Beyond U-Net

Anusheh Chaudry
Stanford University
Stanford, CA
anusheh@stanford.edu

Sarah Barragan
Stanford University
Stanford, CA
sabarrag@stanford.edu

Abstract

Brain tumor segmentation is crucial for diagnosing and treating neuro-oncological conditions. In this research, we evaluated three algorithms: the baseline 3D U-Net employing a 3D convolutional neural network, the 3D Attention U-Net enhancing the 3D U-Net with attention mechanisms to selectively focus and suppress features, and the DubTransBTS utilizing a transformer-based approach to capture long-range dependencies and complex spatial relationships. Using a dataset of brain MRI scans gathered from the BraTS 2020 challenge, the baseline 3D U-Net demonstrated strong spatial information capture and high accuracy across tumor regions with Dice scores of 0.782 for TC, 0.745 for ET, and 0.87 for WT. The 3D Attention U-Net further improved segmentation by focusing on relevant features, especially in distorted areas, achieving Dice scores of 0.786 for TC, 0.796 for ET, and 0.881 for WT, however, the DubTransBTS model struggled to capture tumor details effectively.

1. Introduction

Brain tumors are known to be one of the leading causes of cancer, therefore it is extremely important to determine the most accurate methods of detecting these tumors while they are still in their initial stages of development. Brain tumors occur when there is sudden and unregulated growth of brain tissue within the skull. These tumors can either be benign or malignant, where malignant tumors are known to grow much more rapidly [14]. The United States is one of the countries that feels the impact of brain tumors most significantly, with 86,970 cases of benign and malignant brain tumors diagnosed in 2019 [24]. The biopsy procedure is one method used by radiologists to diagnose brain tumors, however this process only has a 49.1% accuracy and introduces many risks such as inflammation and severe bleeding. Other methods of detection, such as Cerebrospinal fluid analysis and X-ray analysis include similar risks for patients [20].

Given this, image modalities have been becoming more popular due to their increased accuracy rates and decreased levels of risk. One of the most prominent methods for capturing medical imaging data has been magnetic reasoning imaging (MRI), which is a non-invasive procedure that provides radiologists with the information necessary to diagnose any brain abnormalities. Image segmentation is an essential step for brain tumor analysis of MRI images [17]. In the present scenario, radiologists perform tumor segmentation manually. This manual segmentation is a very time-consuming and tedious task with the results being very dependent on human expertise [3]. On the other hand, Computer-Aided Diagnosis (CAD) methods allow for the detections of brain tumors without human intervention [28].

With machine learning and deep learning applications becoming more prevalent in the medical imaging field, the CAD process has immensely improved. However, automatic segmentation of medical images remains a difficult task as medical images are complex in nature and rarely have any simple linear feature [15]. The contrasts, intensities, shapes, location and boundaries of the brain tissue vary from person to person therefore, although a number of algorithms have been proposed in the field of medical image segmentation, medical image segmentation continues to be a complex and challenging problem.

One stand out algorithm in particular is the U-Net [19] model which has been widely accepted for medical image segmentation over the past eight years. The traditional U-shaped encoder decoder structure composed of CNN layers has made great strides in feature enhancement, feature fusion and model generalization for brain MRI segmentations [29]. However, with its CNN based framework, U-Net is limited to the local receptive field of the convolutional filter making it unable to capture long-range dependencies which are crucial for proper interpretation and accurate diagnosis [8]. Therefore we seek improvement upon this model by going beyond a CNN based framework through our proposed Attention U-Net [9] and DubTransBTS methods.

Given an input of multimodal brain MRI scan our 3D U-Net, 3D Attention U-Net, and DubTransBTS models will output a segmented brain MRI image matching the same

dimension as the input. Through the Attention U-Net we will explore the application of attention gates in the decoder of the traditional U-Net architecture in an attempt for more precise and relevant feature extraction. We then go further to explore the use of Vision Transformers in the traditional U-Net architecture for not only more relevant feature extraction due to applied attention but global feature modeling from transformer layers through the DubTransBTS.

In the following, we introduce a brief overview of previous work for biomedical image segmentation using the U-Net architecture and its room for improvement using attention and Vision Transformers.

2. Literature Review

In recent years, deep convolutional neural networks (CNNs) have achieved considerable progress in tackling medical image segmentation due to their powerful feature extraction capabilities that allow for precise localization.

2.1. U-Net Architecture

A breakthrough in the use of CNNs for automatic image segmentation came when Ronneberger et al. proposed U-Net at the MICCAI conference in 2015 [19]. Their iconic encoder-decor based method consists of convolutional blocks that have a U-shaped dimension variation made up of a downsampling (analysis) path from the input layer of the encoder to the input layer of the decoder and an upsampling (synthesis) path to the output. Furthermore, the features in the downsampling path are concatenated to the corresponding features in the expanding path to recover the detailed image information lost during the downsampling process. Since the introduction of traditional U-Net architecture there have been several modifications attempted.

In attempts to improve feature fusion and increase precision, the U-Net++ [30] proposed by Zhou Z. et al. This method takes the original U-Net designs and adds a series of nested and dense hopping paths between the encoder and decoder network to narrow the semantic gap between the encoding and decoding subnetwork of feature maps. However, this model does not express enough information on multiple scales and the network parameters are numbers and complex. In attempts to improve, Huang et al, proposed U-Net 3++ [11] which utilizes full-scale skip links instead of nested dense skip connections and adopots full-scale deep supervision. Such changes allow for the integration of multiscale features and produce more precise location perception and segmentation maps with clarified boundaries with reduced parameters. However, although proven effective, this model suffers from a locality-heavy learning scheme. Trying to boost the reasoning of convolutional layers through residual structures Diakogiannis et al proposed the ResU-Net [6] which includes a residual block between every two convolutional blocks on both the encoder and decoder side as well as skip connections between residual blocks with the same dimension between the encoder and decoder. Despite improving flow of information between different layers to better identify correlation between different locations in the image, progress achieved was minimal across modifications of the U-Net structure [12].

Thus, while the U-Net architecture has been crucial to the development of automatic segmentation, its intrinsic fully convolutional structure makes it suboptimal to relate local features to global representations [5]. Additionally, the U-Net methods described above are either very-time consuming or require heavy computations due to the large number of parameters required.

2.2. Attention and Vision Transformers

To address the limitations of the U-Net architecture and improve the segmentation performance, researchers have explored the use of 3D Attention U-Net and transformers in medical image segmentation tasks. The 3D Attention U-Net extends the original U-Net architecture by incorporating attention mechanisms which allow the model to focus on relevant regions of the input while suppressing irrelevant information [7]. By integrating attention mechanisms into the U-Net architecture, the 3D Attention U-Net can effectively capture spatial dependencies and long-range dependencies in 3D medical images helping to improve the model's ability to localize and segment brain tumor regions accurately. The attention modules in the network enable the model to assign higher weights to informative features and suppress noise or irrelevant features, enhancing the overall segmentation performance [9].

In addition to the 3D Attention U-Net, transformers have also shown promising results in medical image segmentation. Transformers were originally introduced for natural language processing tasks but have recently been adapted for computer vision applications. Transformers are based on self-attention mechanisms that allow the model to capture global dependencies between different regions of the input. In the context of medical image segmentation, transformer-based models can effectively capture long-range dependencies in the image, allowing for better integration of contextual information using Vision Transformer [21]. TransBTS is another Transformer-based network for 3D biomedical image segmentation that combines a U-Net architecture with downsampling, upsampling, and skip connections. The unique aspect of TransBTS is the connection between the encoder and decoder, which includes a linear projection layer, a patch embedding layer for sequence data transfer, ViT layers for extracting long-range dependencies, and a feature mapping layer for fitting the sequence data to the 3D CNN decoder. This arrangement allows ViT to compute long-range dependencies with a global

receptive field at the end of the 3D CNN encoder [25]. While the combination of CNN and Transformer, feature embedding, and feature mapping has been found to be successful, the CNN encoder and decoder could still be further refined [4].

ViTs have the potential to enhance global reasoning in automatic image segmentation for global feature modeling beyond the traditional U-Net architecture which prioritizes localization [21]. Despite these advantages, it's worth noting that the adoption of 3D Attention U-Net and transformers in medical image segmentation is still relatively new, and further research is needed to explore their full potential. Nonetheless, the initial results are promising, and these models show great potential for improving the accuracy and efficiency of brain tumor segmentation, ultimately assisting in more effective medical treatments.

3. Methods

3.1. Baseline: 3D U-Net

We started by constructing a basic 3D U-Net model as our baseline method using the codebase at [22]. We represent the input as a fourth order tensor $I \in R^{C \times H \times W \times D}$ where $C = 4$, $H = 128$, $W = 128$, and $D = 128$. We take $C = 4$, since our inputs contain 4 MR modalities described in Sec. 4. The input traverses two paths successively: the encoder (analysis) path which extracts relevant features and the decoder (synthesis) path which takes the extracted features and reconstructs a segmented image. Each path consists of 4 resolution layers.

In the encoder path each layer consists of double 3 x 3 x 3 convolutional layers with stride 1 and padding of 1. After each convolutional layer we apply 3D group normalization, rather than batch normalization implemented in the traditional U-Net model, followed by a ReLU activation layer. Furthermore, after each double convolution we perform 2 x 2 x 2 max pooling with strides of two in each dimension to effectively act as a feature extractor and downsample the input image. More specifically, at each downsampling step we double the number of feature channels and reduce the spatial dimension in half until we reach the bottleneck.

In the decoder path each layer consists of the upsampling of feature maps by 2 x 2 x 2 up-convolutions with a stride of two in each dimension and two subsequent 3 x 3 x 3 convolutional layers each followed by a ReLU activation layer. At each up-sampling the number of channels is reduced to half while the up-convolution increases the width and height of the image. To increase the dimension of the image after each up-convolution, a concatenation from the same level layer in the encoder path of the feature map is added. More specifically, these long skip connections are made from the equal-resolution feature map in the encoder path to the decoder to provide the necessary high-resolution features in the decoding path. At the very last layer, a 1 x 1 x 1 convolution is performed to reduce the feature map to 4 channels and generate a final segmented image with 3 classes. The output image is dimension $I \in R^{C \times H \times W \times D}$ where $C = 4$, $H = 128$, $W = 128$, and $D = 128$ equivalent to the input.

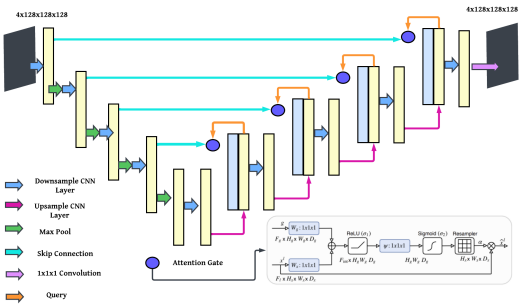


Figure 1. 3D Attention U-Net architecture diagram.

3.2. 3D Attention U-Net

The attention-based 3D U-Net model further extends the 3D U-Net model by incorporating attention mechanisms to focus the model for better and more relevant feature extraction. It takes as input a multimodal MRI scan represented as a fourth order tensor $I \in R^{C \times H \times W \times D}$ and outputs a final segmented image with dimension equivalent to the input. To leverage attention mechanisms, the model adopted the codebase for the renowned Attention U-Net architecture [18], celebrated for its spatial attention capabilities and widespread adoption, primarily by replacing all 2D convolutions and max pooling to their 3D variants. Batch normalization was replaced with group normalization, since group norm provides normalization across spatial dimensions rather than batch dimensions, making it more suitable for irregularly shaped inputs. While channel attention, which assigns importance to different channels based on learned weights, is conceptually simpler, it was deemed less impactful in this context.

In the attention blocks of the Attention U-Net, raw skip connections were replaced by attention-based mechanisms. These blocks received inputs denoted as G and X, representing the output of the coarser layer and the corresponding layer in the encoder, respectively. Employing point-wise 1 x 1 x 1 convolutions, the inputs served as the query and key components in the attention mechanism. Subsequently, the outputs underwent addition, Rectified Linear Unit (ReLU) activation, another point-wise convolution with a single filter, sigmoid activation (representing the value), and subsequent upsampling. Finally, the upsampled output was element-wise multiplied with the skip connection.

To introduce novelty, the attention-based 3D U-Net model incorporated self-attention, combining spatial and self-attention concepts. Diverging from the conventional

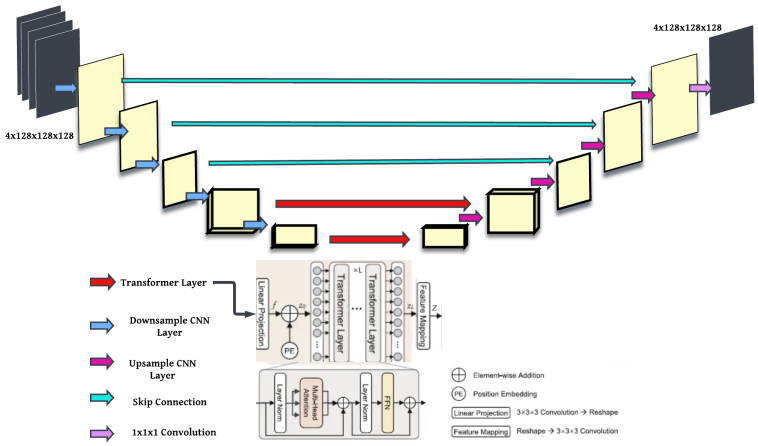


Figure 2. DubTransBTS architecture diagram.

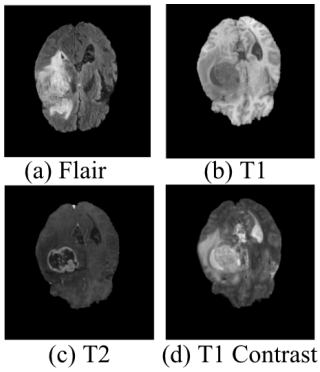


Figure 3. Datset images

approach of obtaining G from a coarser scale prior to upscaling, G was obtained after the upscaling process. This modification enabled attention computation to exclusively rely on the current layer and the skip connection. Consequently, the need for a separate upsampler at the end of the attention block was obviated, leading to enhanced network efficiency. The architectural diagram of this novel module, illustrating the spatial and self-attention components and their connections within the attention-based 3D U-Net model, is presented in the corresponding Fig. 1.

3.3. DubTransBTS (Double Transformer Brain Tumor Segmentation)

In our efforts to overcome the limitation of locality in MRI image segmentation finally, we propose a third method utilizing Vision Transformers [16] for global feature modeling, the DubTransBTS. This method contains the main characteristics of TransBTS (Transformer BrainTumor Segmentation) [25], whose codebase we used [26], with the encoder-decoder structure of the traditional U-Net architecture for precise feature extraction using attention (extension from Attention U-Net) and enhanced global reasoning with Transformer layers.

The encoder of the proposed DubTransBTS consists of 3D CNN downsample layers similar those of the traditional U-Net with the addition of two 3D Convolutional Block Attention Modules (CBAMs) [10, 27] at last two layers to improve performance of the CNN model for adaptive feature selection by computing attention maps. The network takes as input a multimodal MRI scan represented as a fourth order tensor $I \in R^{C \times H \times W \times D}$. The input first goes through an initial convolutional layer with stride 1 in each dimension to increase the channel dimension of the feature map by 4C. This layer is followed by four 3 x 3 x 3 convolutional layers with stride 2 in each dimension extracting the features from the input image. At the fourth and fifth layer each consist of 3D CBAM blocks. Given their intermediate feature map F at each layer, 3D CBAM produces a channel attention map $M_c \in \mathbb{R}^{C \times 1 \times 1 \times 1 \times 1}$ and a 3D spatial attention map $M_s \in \mathbb{R}^{1 \times H \times W \times D}$. The computation is as follows:

$$F\prime = M_c(F) \otimes F, \tag{1}$$

$$F\prime\prime = M_s(F') \otimes F\prime \tag{2}$$

where $\otimes$ is element-wise multiplication between maps and $F\prime\prime$ represent the refined feature map outputted by the CBAM block.

Just as the original 3D U-Net, the decoder consists of 3D CNN upsample layers. There are four 3 x 3 x 3 convolutional layer with stride 2 in each dimensions and one final 1 x 1 x 1 convolutional layer to decrease the dimension of the feature map to $C \times H \times W \times D$ to construct the output segmentation image of the same dimension as the input.

DubTransBTS models similar skip connections to the original 3D U-Net to fuse the encoder features with decoder counterparts richer spatial details for reconstruction. Hence, the output of the first three 3D CNN layers in the encoder is concatenated to the input of the last three 3D CNN layers in the decoder.

However, unlike the first three layers, the output of the fourth and fifth layers are passed through a feature embedding, transformer layers, and a feature mapping layer to then be concatenated to the input of their corresponding upsample layer in the decoder. More specifically, just as introduced by the original TransBTS, the intermediate feature maps of the fourth and fifth layer undergo a linear projection (3 x 3 x 3 convolutional layer) to increase its channel dimension to d. The spatial and depth dimensions are then flattened to one dimension N resulting in a feature map f with dimension $d \times N$ thought of as N d-dimensional tokens for the input of the Transformer block. We replicate

the learned positional embedding added directly to the tokens as utilized in TransBTS :

$$z_0 = f + PE = W \times F + PE, \tag{3}$$

where W is the linear projection, PE is the positional embeddings and z_0 is the feature embeddings used as input for the Transformer block [Lui]. Following feature embedding the sequence of tokens is passed through traditional Transformer layers with Multi-Head Attention block and a Feed Forward Network. We implement four transformer layers where output of the lth layer is computed as:

$$z_l\prime = MHA(LN(z_{l-1})) + z_{l-1}, \tag{4}$$

$$z_l = FFN(LN(z_l\prime)) + z_l\prime, \tag{5}$$

where LN is traditional layer normalization and Z_{l-1} is the previous transformer layer output.

We must transfer the transformer layer output from sequence data to intermediate feature representation for the 3D CNN upsample layer in the decoder. This is done by reshaping Z_4 with dimension $d \times N$ to $Z_4 \in R^*d \times M \times M \times M$ to then be passed through a 3 x 3 x 3 convolutional block to revert the channel dimension d back to its original value prior to the being passed through the transformer layers as an intermediate feature map.

3.4. Loss

To Train each network, we minimize the Binary Cross Entropy (BCE) Dice loss between the prediction and the true label. The BCE Dice loss is Binary Cross Entropy loss with the addition of dice loss. Binary Cross Entropy loss is one of the most commonly used loss functions for the task of image segmentation. This loss examines each pixel individually comparing the class predictions to our ground truth label as follows:

$$\mathcal{L}_{BCE}(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}), \tag{6}$$

where $y, \hat{y}$ are $\{0, 1\}^N$ where $\hat{y}$ refers to the predicted value and y refers to the ground truth label. However, Cross Entropy is based on minimizing pixel-wise error thus, in our study where our input has far less pixels in our dataset belonging to the tumors than surrounding area(i.e. class imbalance) there will be an over-representation of the larger space surrounding the tumor resulting in poorer quality segmentation of small objects being the tumors [23].

Dice loss is inspired by the Dice Coefficient score which is an evaluation metric used to evaluate the results of image segmentation tasks by calculating the similarity between two images see Sec. 5. The Dice loss can therefore be defined as:

$$\mathcal{L}_{DSC} = 1 - DSC, \tag{7}$$

or

$$DiceLoss(X, Y) = -\frac{2 * sum(X * Y}{sum(X) + |Y|}, \tag{8}$$

where DSC is the Dice Coefficient and X is the feature map of predicted probabilities for each pixel and Y is the ground truth segmentation. Dice loss is introduced to medical image segmentation to address this issue of data imbalance since it ignores a large number of background pixels when calculating the intersection and ratio across images [23]. However, ignoring the background pixels takes away information that could be used to promote feature learning [13].

As a result, the combination of the Dice and BCE loss functions allows for the network to not completely ignore the association with background pixels in the process of focusing on the foreground pixel learning (i.e. tumor region). The BCE Dice loss is defined as follows:

$$BCEDiceLoss = DiceLoss + BCELoss. \tag{9}$$

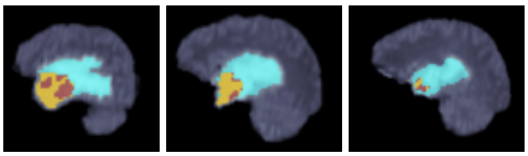
4. Dataset

The BraTS 2020 is a dataset that utilizes multi-institutional pre-operative 3D MRI scans and primarily focuses on the segmentation of intrinsically heterogeneous brain tumors, namely gliomas. The dataset was provided by the 2020 BraTS Multimodal Brain Tumor Segmentation Challenge [1]. The dataset is divided into three groups: training, validation, and testing. The training dataset contains scans from 369 diffuse glioma patients, the validation contains 125 cases and the testing comprises a total of 166 cases of patients with diffuse gliomas. Each dataset consists of multimodal scans describing four different sequences: native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) volumes for patients with either high-grade gliomas or low-grade. All of which were acquired with different clinical protocols in addition to varying scanners across a total of nineteen institutions. All images in the dataset have been segmented manually by 1 - 4 raters using the same annotation protocol to produce 3 labels: (1) non-enhancing tumor core (NCR/NET), (2) peritumoral edema (ED), and (4), GD-enhancing tumor (ET) as can be seen in Fig. 3. Their annotations were approved by expert neuroradiologoists.

The imaging data has been preprocessed to be coregistered to the same anatomical template as well as interpolated to the same $1mm^3$ voxel resolution. Furthermore, the MRI scans have been skull-stripped. The dataset images have dimension 240 x 240 x 155 but given that most of which accounts for background we cropped the images to the size of 128 x 128 x 128 to reduce blank space. The images are combined across the 4 modalities described above to represent a 128 x 128 x 128 voxel tile of the image with

Model	$Dice_{TC}$	$Dice_{ET}$	$Dice_{WT}$
3D U-Net	0.782	0.745	0.87
3D Attention U-Net	0.786	0.796	0.881
DubTransBTS	0.592	0.569	0.635

Table 1. Results. Dice Score.



(a) (b) (c)

Figure 4. Segmentation output example a) 3D Attention U-Net b) 3D U-Net c) DubTransBTS

4 channels. We normalized the pixel intensities of each image to be zero mean and unit variance to account for the varying scans across institutions and minimize the effects of artifacts and the proneness of medical images to be corrupted by noise. Lastly, applied data augmentation techniques including random horizontal flips with probability 0.5 and random brightness contrast with probability 0.2 to reduce the potential for overfitting.

5. Experiments & Discussion

5.1. Metrics

The accuracy of segmentation is determined using Dice Score and IoU, also known as the jaccard index [2].

Dice Score: The Dice score is another widely used metric for evaluating the performance of image segmentation models, including brain tumor segmentation and is our primary metric used as seen in Tab. 1. It quantifies the similarity between the predicted and ground truth segmentation masks by computing the ratio of twice the intersection to the sum of the prediction and ground truth areas, focusing only on the segmentation classes and not the background class. The pixels are classified as True Positive (T P), False Negative (F N) and False Positive (FP). The Dice score ranges from 0 to 1, where a value closer to 1 indicates a higher degree of agreement between the predicted and ground truth segmentations.

$$Dice = \frac{2 * TP}{(TP + FP) + (TP + FN)}, \tag{10}$$

IoU: IoU (Intersection over Union) is a commonly used evaluation metric for image segmentation models, including brain tumor segmentation. It measures the overlap between the predicted and ground truth segmentation masks by calculating the ratio of their intersection to their union. The pixels are classified as True Positive (TP), False Negative (FN) and False Positive (FP). IoU scores range from 0 to 1, with higher scores indicating better segmentation accuracy.

$$IoU = \frac{TP}{(TP + FP + FN)}, \tag{11}$$

5.2. 3D U-Net

The baseline 3D U-Net model was trained and evaluated on the BraTS 2020 dataset, which consists of multi-institutional 3D MRI scans of brain tumors. The model utilized a combination of preprocessing techniques, including co-registration, voxel interpolation, skull stripping, and data normalization. Training was conducted for 30 epochs with a learning rate of 5e-4 using the Adam optimizer and a batch size of 1. The model employed the BCE Dice loss function for segmentation.

The evaluation results showed high Dice scores, indicating strong overlap between predicted and ground truth tumor regions: WT = 0.87, TC = 0.782, ET = 0.745. IOU scores also demonstrated good agreement: WT = 0.78, TC = 0.662, ET = 0.622. These scores highlight the model's ability to accurately segment tumor regions, which can be attributed to the 3D U-Net architecture's suitability for volumetric segmentation tasks, leveraging multi-modal information, and employing data augmentation techniques to mitigate overfitting. To view segmentation results see Fig. 4 (b).

However, variations in imaging protocols and manual annotations by raters may impact the model's performance on unseen data. Nonetheless, the results obtained by the baseline model lay a foundation for further advancements and exploration in brain tumor segmentation using more advanced algorithms.

5.3. 3D Attention U-Net

During the experimentation with the 3D Attention U-Net model, three different training configurations were evaluated to fine-tune the model's performance. Each configuration varied in parameters such as batch size, number of epochs, learning rate (LR), and the choice of the Adam optimizer. The first training configuration involved a batch size of 4, 150 epochs, LR of 0.0002, and the Adam optimizer with a LR of 0.0002. In the transition from configuration 1 to configuration 2 of the 3D Attention U-Net model, a decrease in the number of epochs was accompanied by an increase in the learning rate, which aimed to balance the trade-off between training time and the risk of overshooting the optimal solution. By reducing the number of epochs from 150 to 100, the training process became shorter, allowing faster iterations over the dataset and the higher learning rate of 0.0003 likely facilitated more significant updates to the model's parameters during each iteration. This combi-

nation of fewer epochs and a higher learning rate sought to expedite the training process while maintaining the potential for the model to converge to a satisfactory solution. However, the transition from configuration 2 to configuration 3 followed a different strategy. Here, the number of epochs was increased back to 150, while the learning rate was decreased to 0.0001. Increasing the number of epochs provided more opportunities for the model to learn from the data, allowing it to capture complex patterns and improve segmentation accuracy. The lower learning rate of 0.0001 helped to fine-tune the model's parameter updates, ensuring a more controlled and precise convergence towards the optimal solution as seen in Fig. 5.

Among the three training configurations, configuration 3 demonstrated the highest performance in terms of Dice scores and IoU scores for the 3D Attention U-Net model. Configuration 3 obtained a Dice score of 0.786 for TC, 0.796 for ET, and 0.881 for WT, along with IoU scores of 0.671 for TC, 0.634 for ET, and 0.792 for WT. To see segmentation results see Fig. 4 (a). To train configuration 3, a lower learning rate (LR) of 0.0001 was employed over 150 epochs, allowing the model to converge gradually and effectively fine-tune its weights. Additionally, configuration 3 utilized the Adam optimizer with the same LR, which proved to be a suitable optimization strategy for this specific training setup.By comparing the results of the 3D Attention U-Net model with the baseline 3D U-Net model, it becomes evident that the attention mechanism played a crucial role in improving segmentation performance and enhancing accuracy in capturing tumor sub-regions by focusing on relevant tumor regions while suppressing irrelevant background information. The higher Dice scores indicate better overlap between predicted tumor regions and ground truth annotations, highlighting the model's ability to accurately delineate tumor boundaries.

The incorporation of attention mechanisms allows the 3D Attention U-Net model to selectively amplify relevant regions and actively suppress irrelevant regions enabling the model to capture intricate details and handle complex or irregular tumor shapes, resulting in improved segmentation performance compared to the baseline 3D U-Net model. It eliminates the need for external object localization and improves the model's sensitivity and accuracy to foreground pixels, making it computationally more efficient.However, it is important to note that the attention mechanism may introduce additional noise or uncertainty during the training process and could struggle to effectively capture subtle tumor details, which may lead to reduced segmentation performance in some cases. Nevertheless, the attention mechanism proves beneficial when specific tumor subregions require focused attention for accurate segmentation or when dealing with tumors that exhibit complex or irregular shapes.

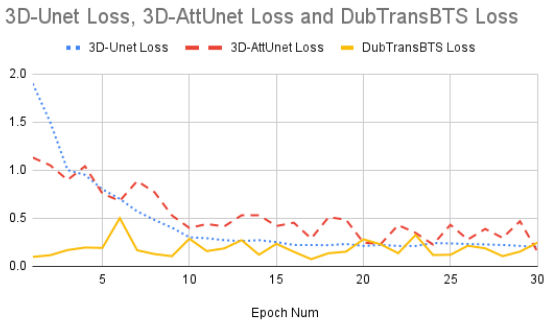


Figure 5. Loss Curve.

5.4. DubTransBTS

During the experimentation with the DubTransBTS model, two different training configurations were evaluated to fine-tune the model's performance. Each configuration varied in parameters such as batch size, number of epochs, learning rate (LR), and the choice of the Adam optimizer. The first training configuration involved a batch size of 4, 500 epochs, LR of 0.0003, and the Adam optimizer with a LR of 0.0003. The learning rate decays by each iteration with a power of 0.9 for better convergence. In the transition from configuration 1 to configuration 2, an increase in the number of epochs and a decrease in the learning rate were observed. This adjustment was made in response to the lower dice scores achieved in configuration 1 compared to the baseline 3D U-Net and 3D Attention U-Net models. Increasing the number of epochs from 500 to 1000 allowed for a more extensive training period, giving the model additional iterations to refine its learned representations. By doing so, the model could potentially capture more intricate patterns and improve segmentation accuracy over time. The decision to decrease the learning rate from 0.0003 to 0.0002 in configuration 2 aimed to provide a more controlled and fine-grained adjustment of the model's parameters during the training process, since a higher learning rate can lead to larger updates in parameter values, causing the model to potentially overshoot the optimal solution and hinder convergence. By reducing the learning rate, the model's parameter updates became smaller and more precise, enabling a more stable and gradual convergence towards the optimal solution and seeking to address the potential issue of suboptimal convergence observed in configuration 1 as seen in Fig. 5.

The results obtained for the DubTransBTS model, which employs a transformer-based approach, indicate that further refinement and longer training duration may be required to achieve comparable performance to the baseline 3D U-Net and 3D Attention U-Net models. During the hyperparameter tuning process, configuration 2 exhibited the best per-

formance among the tested configurations, but the achieved Dice scores remained lower compared to the other models. configuration 2 was trained for 1000 epochs with a batch size of 4 and a learning rate of 0.0002. The Dice scores obtained were 0.569 for ET, 0.635 for WT, and 0.592 for TC. To view segmentation result see Fig. 4 (c).

The lower Dice scores in the DubTransBTS model could be attributed to the need for a high number of epochs to capture and learn complex patterns in the brain tumor segmentation task. Transformers are known for their ability to model long-range dependencies and capture intricate relationships in sequential data. However, brain tumor segmentation requires the model to comprehend and extract fine-grained details, which might necessitate an extended training period to allow the model to converge and learn these features effectively.

It is important to note that the DubTransBTS model's performance might improve with an even longer training duration or alternative strategies such as pre-training on a large dataset or using transfer learning from a related task. Transformers have demonstrated exceptional performance in various natural language processing tasks but might require substantial computational resources and longer training times to achieve similar levels of success in medical image segmentation.

6. Conclusion

In this report, we conducted an extensive evaluation of three algorithms for brain tumor segmentation: the baseline 3D U-Net, the 3D Attention U-Net, and the DubTransBTS model based on a transformer approach. The highest-performing algorithms were the baseline 3D U-Net and the 3D Attention U-Net, which achieved competitive results in terms of Dice scores and IoU scores for tumor segmentation. The baseline 3D U-Net demonstrated strong segmentation capabilities, effectively capturing spatial information and achieving high accuracy across different tumor regions. The 3D Attention U-Net improved upon the baseline model by incorporating attention mechanisms, enabling better focus on relevant features and enhancing segmentation accuracy, particularly for challenging areas with irregular tumor boundaries. However, the DubTransBTS model, although promising in its architecture, did not achieve comparable results. It showed potential for learning complex spatial relationships but struggled to capture the intricate details of brain tumors. This may be due to the need for longer training periods and a higher number of epochs to fully exploit the benefits of the transformer-based approach.

Moving forward there are several avenues for further exploration. Firstly, investigating alternative architectures for the DubTransBTS model, such as variations of the transformer-based approach, may help improve its performance and bridge the gap with the other models. Additionally, exploring advanced training strategies, such as transfer learning or ensembling, could lead to more robust and accurate tumor segmentation. On top of this, there could be further incorporation of additional data augmentation techniques and preprocessing steps in order to better generalize and handle the inherent variability in medical images. Furthermore, considering the integration of clinical information, such as patient demographics or other imaging modalities, may provide valuable contextual information for improving the segmentation performance. Lastly, evaluating the models on larger and more diverse datasets, including data from multiple institutions, would help assess their robustness and generalizability in real-world clinical settings.

7. Contributions

Anusheh Chaudry and Sarah Barragan designed, modified, and trained the 3D U-Net, 3D Attention U-Net, and DubTransBTS models. Anusheh Chaudry and Sarah Barragan implemented the evaluation metrics. Anusheh Chaudry and Sarah Barragan wrote the paper. Public code was used at https://github.com/ozan-oktay/Attention-Gated-Networks, https://github.com/Wenxuan-1119/TransBTS, and https://github.com/johschmidt42/PyTorch-2D-3D-UNet-Tutorial.

References

[1] Brats 2020 challenge. http://braintumorsegmentation.org/, 2020. 5

[2] Taha AA and Hanbury A. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):29, 2015. 6

[3] Jyotismita Chaki. Brain mri segmentation using deep learning: Background study and challenges. *Brain Tumor MRI Image Segmentation Using Deep Learning Techniques*, pages 1–12, 2022. 1

[4] Bingzhi Chen, Yishu Liu, Zheng Zhang, Guangming Lu, and Adams Wai Kin Kong. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. 2022. 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. *ArXiv*, abs/2102.04306, 2021. 2

[6] Foivos I. Diakogiannis, Francois Waldner, Peter Caccetta, and Chen Wu. Resunet: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS: Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020. 2

[7] Ozan Oktay et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv*, 1804.03999, 2018. 2

[8] Xiyao Fu, Zhexian Sun, Haoteng Tang, Eric M. Zou, Heng Huang, Yong Wang, and Liangm Zhan. 3d bi-directional transformer u-net for medical image segmentation. *Frontiers in Big Data*, 5, 2023. 1

[9] Maryann M. Gitonga. Multiclass mri brain tumor segmentation using 3d attention-based u-net. *Strathmore University*, 2023. 1, 2

[10] Theophraste Henry, Alexandre Carre, Marvin Lerousseau, Theo Estienne, Charlotte Robert, Nikos Paragios, and Eric Deutsch. Brain tumor segmentation with self-ensembled, deeply-supervised 3d u-net neural networks: a brats 2020 challenge solution. *ArXiv*, abs/2011.01045, 2020. 4

[11] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yenwei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. pages 1055–1059, 2020. 2

[12] Fabian Isenee, Paul F. Jaeger, Simon A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:203–211, 2021. 2

[13] Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, pages 1–7, 2020. 5

[14] Zhao L. and Jia K. Multiscale cnns for brain tumor segmentation and diagnosis. *Comput. Math. Methods Med*, 8356294:2016, 2016. 1

[15] Matthew Lai. Deep learning for medical image segmentation. *CoRR*, abs/1505.02000, 2015. 1

[16] Qianying Liu, Chaitanya Kaul, Jun Wang, Christos Anagnostopoulos, Roderick Murray-Smith, and Fani Deligianni. Optimizing vision transformers for medical image segmentation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 4

[17] Mabray M.C. and Cha S. Advanced mr imaging techniques in daily practice. *Neuroimaging Clin*, 26(4):647–666, 2016. 1

[18] Ozan Oktay. Attention-gated-networks. https://github.com/ozan-oktay/Attention-Gated-Networks, 2018. 3

[19] Olaf Ronneberger, Phillip Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medial Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015. 1, 2

[20] Kasraeian S., Allison D.C., Ahlmann E.R., Fedenko A.N., and Menendez L.R. A comparison of fine-needle aspiration, core biopsy, and surgical biopsy in the diagnosis of extremity soft tissue masses. *Clin Orthopaedics Rel Res*, 468(11):2992–3002, 2010. 1

[21] Tummala S, Kadry S, Bukhari SAC, and Rauf HT. Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. *Curr Oncol*, 29(10):7498–7511, 2022. 2, 3

[22] Johannes Schmidt. Pytorch-2d-3d-unet-tutorial. https://github.com/johschmidt42/PyTorch-2D-3D-UNet-Tutorial, 2023. 3

[23] Jun Su, Zhe Liu, Jing Zhang, Victor S. Sheng, Yuqing Song, Yan Zhu, and Yi Liu. Dv-net: Accurate liver vessel segmentation via dense connection model with d-bce loss function. *Knowledge-Based Systems*, 232, 2021. 5

[24] Hollon T.C., Pandian B., Adapa A.R., Urias E., Save A.V., Khalsa S.S.S., Eichberg D.G., D'Amico R.S., Farooq Z.U., and Lewis S. Near real time intraoperative brain tumor diagnosis using stimulated raman histology and deep neural networks. *Nat Med*, 26(1):52–58, 2020. 1

[25] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li. Transbts: Multimodal brain tumor segmentation using transformer. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, vol 12901:109–119, 2021. 3, 4

[26] Wenxuan. Transbts. https://github.com/Wenxuan-1119/TransBTS, 2021. 4

[27] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 4

[28] Chen Y., Shao Y., Yan J., Yuan T.F., Qu Y., Lee E., and Wang S. A feature-free 30-disease pathological brain detection system by linear regression classifier. *CNS Neurol Disorders-Drug Targets (Formerly Current Drug Targets-CNS Neurological Disorders)*, 16(1):5–10, 2017. 1

[29] Xiao-Xia Yun, Le Sun, Yuhan Fu, Ruiliang Lu, and Yanchun Zhang. U-net-based medical image segmentation. *Journal of Healthcare Engineering*, 2022:1–16, 2022. 1

[30] Zongwei Zhou, Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Net++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11, 2018. 2