DEEP LEARNING FOR BRAIN TUMOR SEGMENTATION

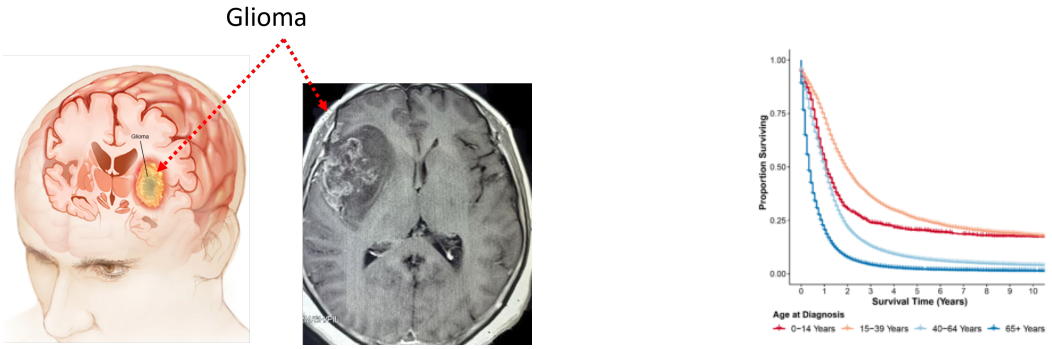
Veronica Mata, Karthik Dinesh
Department of Data Science
University of Rochester
Rochester, NY 14623, USA
{vmata,kdinesh}@ur.rochester.edu

ABSTRACT

Brain tumors are life-threatening, particularly gliomas, which are familiar forms of brain tumors, that cause brain hemorrhage resulting in death. MRI scans, which are the most reliable tools in detecting and staging the cancers, however, have intrinsic variation resulting in difficulty in diagnosis. Computational tools are needed to accurately identify the brain cancer sub-regions. We came up with a novel architecture called V-Net Transformer which used the advantages of the V-Net encoder-decoder-skip connections and the transformers to perform cancer sub-region identification, particularly, whole tumor, tumor core, and enhancing tumor segmentation. Experimental evaluations show that the average dice coefficient score is 74.6 which is comparable with the existing architectures.

1 INTRODUCTION

Gliomas, the most common form of brain tumors, shown in Fig. 1 (a), present complex diagnostic and treatment challenges, largely due to their diverse and unpredictable nature evident in MRI scans. The BraTS challenge has been pivotal in providing a dataset Menze et al. (2014); Bakas et al. (2017; 2018) for benchmarking brain tumor segmentation methods in MRI scans, which is vital for accurate medical analyses. It emphasizes the importance of segmenting gliomas, characterized by their heterogeneity in both appearance and structure, to improve treatment precision and predict patient outcomes. This segmentation task, crucial for medical image analysis, also contributes to survival prediction (shown in Fig. 1 (b)) and enables precision treatments through the examination of machine learning techniques.



(a) (b)

Figure 1: Glioma affecting the brain region and survival chart (Image credits, https://www.ncbi.nlm.nih.gov/books/NBK441874/,https://www.naaccr.org/the-leading-cause-of-death-due-to-brain-tumors-in-the-united-states/,https://www.mayoclinic.org/diseases-conditions/glioma/symptoms-causes/syc-20350251)

Motivated by this we took up a brain segmentation task for this project. In the realm of medical imaging, particularly in the segmentation of brain tumors using MRI scans, deep learning has made

significant strides. However, traditional Convolutional Neural Network (CNN) approaches such as the UNet architecture, while yielding promising results, are limited by their convolutional operations' narrow receptive fields. This restricts their ability to capture long-range dependencies within the data, a critical factor in medical image analysis.

Hence, we came up with a proposal shown in Fig. 2. Our methodology comprises a loop that includes data pre-processing, data modality fusion (encompassing early, intermediate, and late fusion stages), and the application of a deep learning model. This is followed by a testing loop with tasks focused on the segmentation of the whole tumor, the tumor core, and the enhancing tumor. The final stage involves the evaluation of test data to validate the model's performance.

We successfully implemented our proposal by introducing the V-Net Transformer architecture for three segmentation problems. This innovative solution integrates the expansive contextual comprehension afforded by transformers with the detailed feature extraction capabilities of CNNs. The V-Net Transformer distinguishes itself by fusing multi-modal MRI data at the earliest stages of processing. This fusion strategy allows the architecture to attain a comprehensive understanding of the tumor's characteristics from multiple MRI modalities. We further describe the model in detail in the following sections.

The report is organized as follows. In Section 2, we talk about the related work. We will describe the dataset and the model architecture in detail in Section 3. Next, we describe the data preprocessing, hyperparameter selection, and experimental results in Section 4 followed by a conclusion in Section 5.

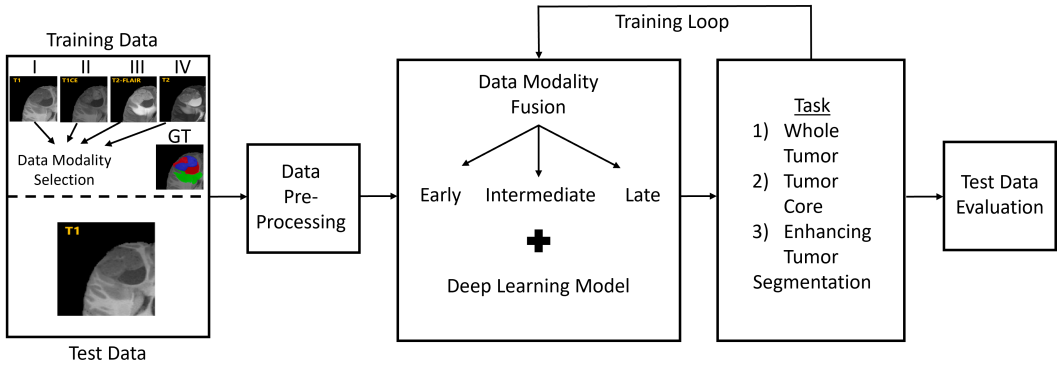


Figure 2: Proposal statement (Images from Kazerooni et al. (2023))

2 Related Work

In the literature, several architectures have aimed to solve the BraTS challenge by solving the three segmentation tasks. Particularly, the combination of U-Net and transformers has been tried in different ways. The Trans-Unet model Chen et al. (2021) uses convolution layers initially in the encoder and makes use of a stack of 12 transformers as bottlenecks followed by de-convolution on the decoder side. The UNETR model Hatamizadeh et al. (2022) uses transformers on the encoder and bottleneck and normal de-convolutions on the decoder side. The UNETR++ Shaker et al. (2022) is similar to UNETR but makes use of attention blocks in the decoder. Swin-Unet Cao et al. (2022) is all transformer architecture but processes the MRI volume as 2D slices and does not make use of the 3D structure. The recent segment anything model (SAM) Kirillov et al. (2023) has become a strong foundational model for image segmentation. An extension of this model for MRI images called SAM3D Bui et al. (2023) uses an encoder-decoder style architecture but does not use any skip connections.

3 METHODS

3.1 DATASET

The 2020 BraTS Challenge offered an extensive dataset Menze et al. (2014); Bakas et al. (2017; 2018) of MRI volumes. This dataset featured four unique MRI modalities, accompanied by corresponding segmentation masks, with each modality emphasizing a different aspect of the tumor. An example of data from one participant is illustrated in Fig. 3. This figure displays the four distinct data modalities/channels in a left-to-right sequence, along with their respective mask labels. The various MRI modalities are described in detail below.

1. FLAIR MRI: These images reveal the full extent of the tumor.
2. T2: T2 shows the anatomy and structure of the brain and can distinctly reveal areas of the tumor that are more solid.
3. T1ce or T1Gd: These images highlight active tumor structures, around their structures there are cystic or necrotic components.
4. T1: T1 provides good anatomical details of the brain. In these images, fat appears bright, and water or fluid appears dark. This contrast is useful for visualizing the normal structures of the brain, such as the gray and white matter differentiation.

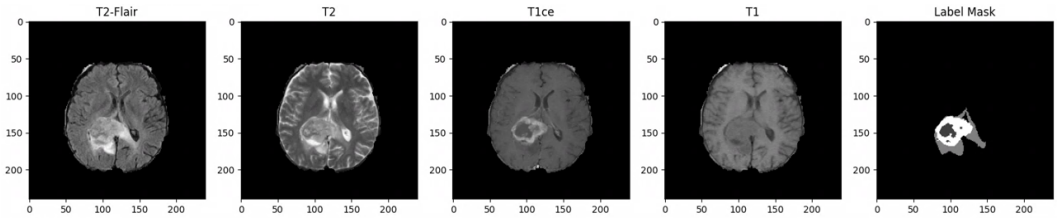


Figure 3: Dataset example showing four image modalities and the ground truth mask.

In the tumor classification tasks, we delineate the distinct classes our algorithm is trained to identify: 'Whole Tumor' in yellow, 'Tumor Core' in red, and 'Enhancing Tumor' in blue, as shown in Fig. 4, each color signifying a specific segment within the MRI data for precise segmentation.

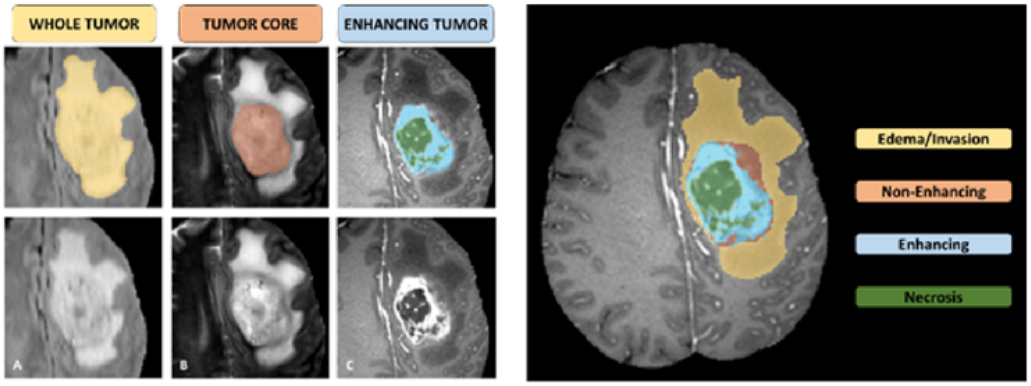


Figure 4: Whole tumor, tumor core, and enhancing tumor segments in a MRI image.

3.2 MODEL

In this project, we develop our network architecture inspired by the two concepts. The first one is the U-Net architecture Ronneberger et al. (2015) and the other is the transformer Vaswani et al. (2017). We know that U-Net architecture was developed for biomedical image segmentation. The U-Net

architecture is successful because of two main structural innovations, one is the encoder-bottleneck-decoder structure, and the other is the skip connections that connect encoder and decoder at each level. When the image passes through the encoder, at each level, different types of features are extracted and passed to the next level. At each level, the image resolution gets lower and the depth becomes higher. Then comes the bottleneck where the transition happens. In the decoder part, the resolution starts getting higher and the information from the encoder part is passed on to the decoder at each level, which helps the decoder to decode the relation between the image and the segmented output. Further research in the U-Net has resulted in the removal of the bottleneck resulting in a V-shaped structure and this is called V-Net. Milletari et al. (2016) Next, we all know that the innovation of transformers has created a paradigm shift in the field of natural language processing and subsequently computer vision. They completely remove the necessity of convolution layers and comprise the attention mechanism and feedforward layers. Also, they have proved to be an excellent feature extractor. Hence, in our model, we take the structural and functional advantage of V-Net and transformers. The model architecture is shown in Fig. 5 which is called the V-Net transformer.

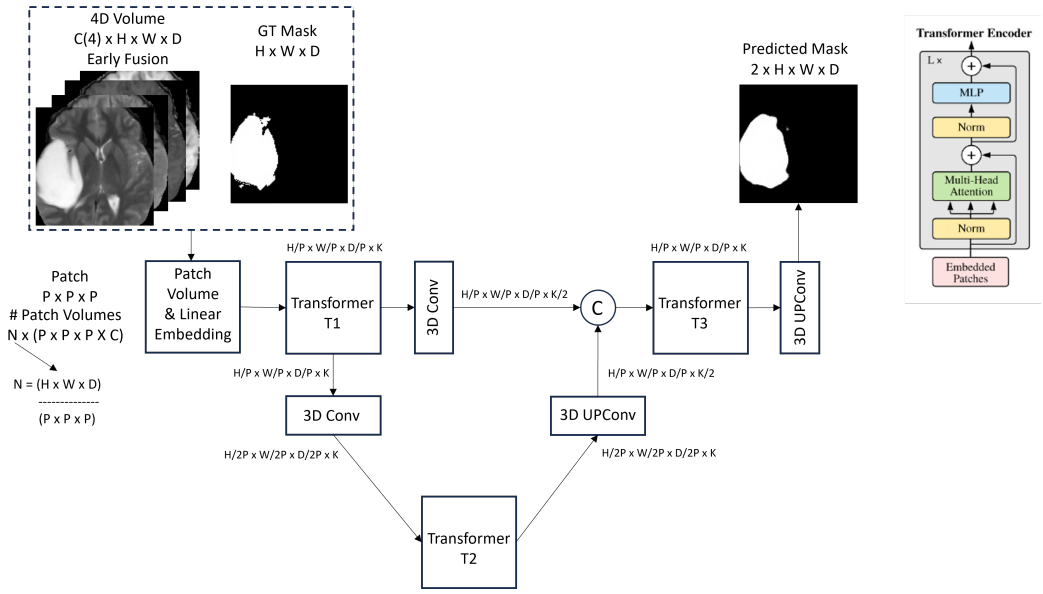


Figure 5: Network architecture of the proposed V-Net Transformer

We can observe that the input to the model is a volume $C \times H \times W \times D$ where C represents the number of channels, H, W, and D represents the height, width, and depth of the volume, respectively. In our project, we perform an early fusion of all image modalities described previously. The mask of the ground truth has a size of $H \times W \times D$. The volume is then divided into a set of patches of size $P \times P \times P$ and each of the patch volumes goes through a linear embedding. A patch size of P results in a total of N, P^3C volumes, where $N = \frac{H}{P} \times \frac{W}{P} \times \frac{D}{P}$. The linear embedding converts $N \times P^3C$ into $N \times K$ where K represents the embedding dimension. This $N \times K$ input goes to the encoder part of the model where it passes through the first transformer T1 and the output has the same size as that of the input. Next to reduce the resolution this output from T1 is passed through a 3D convolution which results in half the resolution of the input. This smaller resolution passes through the transformer T2. Since we have only 1 layer in our proposed V-Net transformer, the output of the T2 transformer goes to the decoder part. Hence to match the size of the volume at output of T1, we pass the output of T2 through a 3D up-convolution which doubles the resolution and halves the number of channels. The output of T1 is passed through 3D convolution which just halves the number of channels retaining the resolution. Now, the output of T1 and T2 are concatenated and passed through the final transformer T3, the output of which goes through another 3D convolution to obtain a prediction mask of size $2 \times H \times W \times D$. The transformer encoder used in the model is also shown in Fig. 5 which is the same as the one used in the original transformer architecture.

4 EXPERIMENTS AND RESULTS

4.1 DATA PREPROCESSING

To address overfitting and enhance the robustness of our model, we incorporated a series of data augmentation techniques during the training phase. These augmentations simulate various transformations that medical images may undergo, thus enriching our dataset:

1. Rotation: Each image and corresponding label were rotated by an angle of -0.4 degrees. The image was interpolated using a bilinear method to preserve smooth gradients, and the label was interpolated using the nearest neighbor method to maintain the integrity of class boundaries.
2. Gaussian Noise: Random Gaussian noise was introduced tot he images with a probability of 0.6.This noise follows a Gaussian distribution with mean of zero, adding a layer of complexity and variability to the training data.
3. Intensity Normalization: To standardize intensity values across the dataset, we performed channel-wise normalization on the images where only non-zero intensity values were considered, ensuring that the mode is not biased by variations in contrast or brightness.

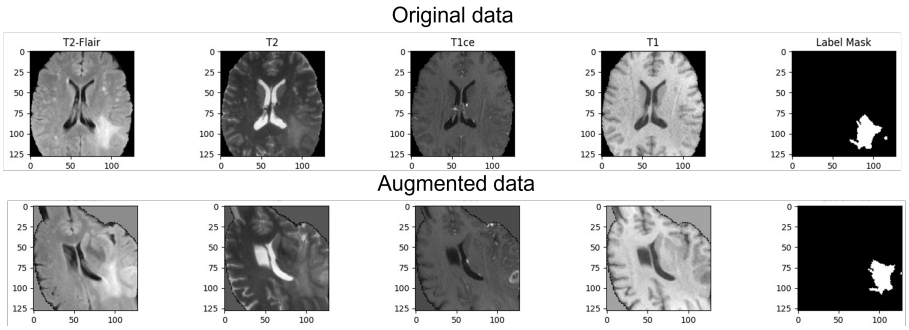


Figure 6: Example augmentation (Rotation shown in bottom row. Top row shows normal images.)

An example of using the augmentation for the dataset is shown in Fig. 6 where we have shown the original data in the top row and in the bottom row we show the example of applying the rotation on the top row.

4.2 HYPERPARAMETERS

The BRaTS dataset comprised data from 369 participants. We used 70% of the data (295) for training and 30% of the data (74) for the testing purpose. The original volume size was $240 \times 240 \times 155$. We cropped the image to a size of $128 \times 128 \times 128$. Hence, $H = W = D = 128$. The number of channels was $C = 4$ and the patch size was chosen as $P = 8$. A batch size of 4 was used. We used a softmax dice loss function and Adam optimizer with a learning rate of 0.0001. The model was run for 30 epochs. synapstaThe training loss for the 30 epoch is shown in Fig. 7. We can observe the reduction in training loss as the epoch increase.

4.3 RESULTS

We use the Dice Coefficient Score (DCS) as the metric to evaluate our model and compare it with a few other architectures. The DCS comparison results are shown in Fig. 8. We can observe that the proposed model is fair competitive with the other models both in terms of the DCS and number of parameters. We can see that for the whole tumor, tumor core, and enhancing tumor segmentation, the DCS is 84.7, 73.9, and 65.3, respectively. For the overall average DCS, our model performs better than two other models and is closer to one model. In terms of the number of parameters, our model has a smaller number of parameters in comparison with other models.

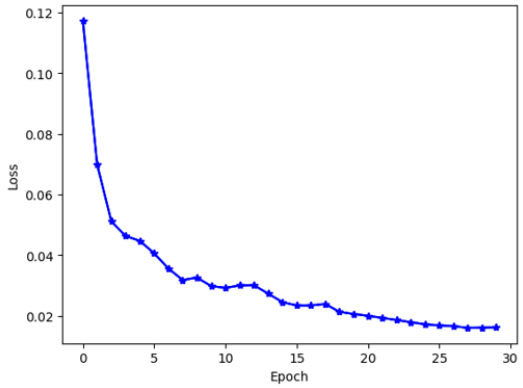


Figure 7: Training loss vs the epoch curve.)

Methods	Params	Average	Whole Tumor	Enhancing Tumor	Tumor Core
		Dice Coefficient Score			
TransUNet	96.07M	64.4	70.6	54.2	68.4
UNETR	92.46M	71.1	78.9	58.5	76.1
nnFormer	150.5M	86.4	91.3	81.8	86.0
INETR++	42.65M	77.7	91.2	78.5	78.4
V-Net Transformer (Proposed)	45.7M	74.6	84.7	65.3	73.9

Figure 8: Dice coefficient score and network parameter comparison of the proposed model with few other existing models.

We also show the visualization of the three types of segmentation as shown in Fig 9. The visualization seems to be consistent with the results shown in Fig. 8. We can observe that the whole tumor has the best predicted segmentation mask followed by the tumor core and then the enhancing tumor.

5 Conclusion

In this project, we proposed a V-Net Transformer architecture to identify cancer regions and subregions in the form of whole tumor, tumor core, and enhancing tumor segmentation. The architecture uses the best features from the V-Net (encoder-decoder-skip connections) and transformers (contextual feature extraction). We got an average dice coefficient score of 74.6 and a highest score of 84.7 for the whole tumor segmentation. We get this fair good score by using the proposed model with lower parameters to train (45.7 million parameters).

References

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4 (1):1–13, 2017.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, 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.

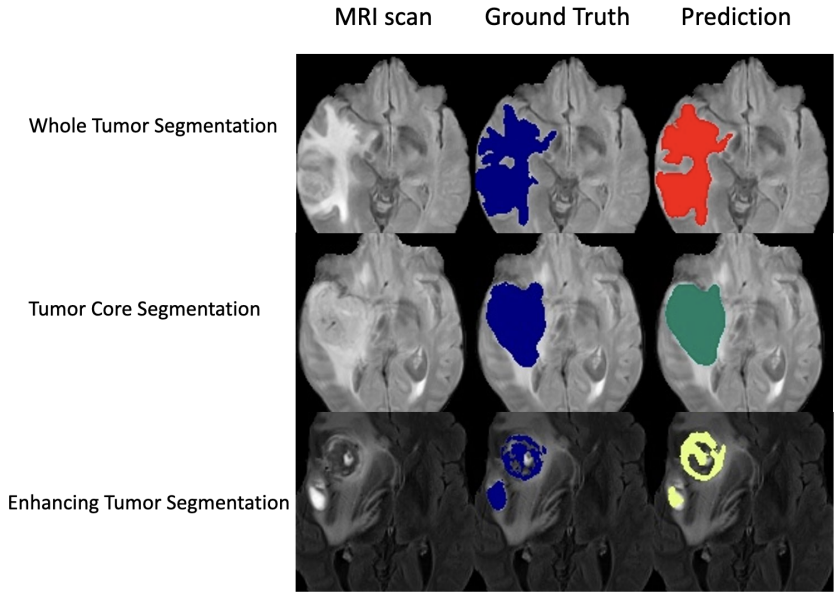


Figure 9: Side-by-side evaluation of V-Net model.It showcases across three distinct segmentation tasks: Whole Tumor, Tumor Core, and Enhancing Tumor. From left to right, each row displays the original MRI scan, the expert-annotated ground truth, and the algorithm's prediction, with the segmentation regions color-coded—blue for ground truth and contrasting colors for predictions (red for Whole Tumor, green for Tumor Core, and yellow for Enhancing Tumor).

Nhat-Tan Bui, Dinh-Hieu Hoang, Minh-Triet Tran, and Ngan Le. Sam3d: Segment anything model in volumetric medical images. *arXiv preprint arXiv:2309.03493*, 2023.

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. In *European conference on computer vision*, pp. 205–218. Springer, 2022.

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 preprint arXiv:2102.04306*, 2021.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 574–584, 2022.

Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, Debanjan Haldar, Zhifan Jiang, Syed Muhammed Anwar, Jake Albrecht, Maruf Adewole, Udunna Anazodo, Hannah Anderson, et al. The brain tumor segmentation (brats) challenge 2023: Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds). *ArXiv*, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34 (10):1993–2024, 2014.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Unetr++: delving into efficient and accurate 3d medical image segmentation. *arXiv preprint arXiv:2212.04497*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.