PURE VERSUS HYBRID TRANSFORMERS FOR MULTI-MODAL BRAIN TUMOR SEGMENTATION: A COMPARATIVE STUDY

G. Andrade-Miranda [⋆] *V. Jaouen* [⋆,◇] *V. Bourbonne* [⋆,•,○] *F. Lucia* [⋆,•,○] *D. Visvikis* [⋆] *P.-H. Conze* [⋆,◇]

[⋆] LaTIM UMR 1101, Inserm, Brest, France [◇] IMT Atlantique, Brest, France
[•] Brest University Hospital Centre, Brest, France [○] University of Western Brittany, Brest, France

ABSTRACT

Vision Transformers (ViT)-based models are witnessing an exponential growth in the medical imaging community. Among desirable properties, ViTs provide a powerful modeling of long-range pixel relationships, contrary to inherently local convolutional neural networks (CNN). These emerging models can be categorized either as hybrid-based when used in conjunction with CNN layers (CNN-ViT) or purely Transformers-based. In this work, we conduct a comparative quantitative analysis to study the differences between a range of available Transformers-based models using controlled brain tumor segmentation experiments. We also investigate to what extent such models could benefit from modality interaction schemes in a multi-modal setting. Results on the publicly-available BraTS2021 dataset show that hybrid-based pipelines generally tend to outperform simple Transformers-based models. In these experiments, no particular improvement using multi-modal interaction schemes was observed.

***Index Terms*—** Vision Transformers, tumor segmentation, multi-modality, hybrid CNN-Transformers models

1. INTRODUCTION

Deep segmentation models derived from U-Net [1] have significantly transformed the field of medical image segmentation due to their ability to learn complex local representations at multiple spatial scales in a data-driven fashion [2, 3, 4]. After having established new benchmark performances on image classification tasks, Vision Transformers (ViT) models [5] have recently emerged as the most popular option to replace or complement convolutional neural networks (CNN) for computer vision applications [6]. Their popularity are now also rapidly growing in medical image analysis [7], especially for medical image segmentation with an exponential growth of related publications in the last year [8]. Depending on the type of encoder used in these models, two categories can be identified: pure Transformers-based made of ViT layers only and hybrid-based models composed of both CNN and ViT layers. The first category exploits the global context modeling capability of Transformers to effectively encode the relationships between spatially distant voxels. However, as anatomical structures can substantially vary in scale, they cannot be properly modelled using a set of fixed sub-regions of the image [9]. For this reason, hybrid architectures combining the global context modeling ability of Transformers with the CNN inductive bias are also popular [10, 11]. CNN layers are used as multi-scale feature extractors, while Transformers capture long-term dependencies among features that would be potentially lost with purely convolutional models. Recently, hierarchical ViT such as Swin Transformers [12, 13] have also been introduced to overcome these challenges by extracting features at different resolutions while saving the linear computational complexity with respect to image size.

The advantages of pure Transformer-based models over hybrid approaches in medical imaging are not yet clear as of today. Moreover, the exploitation of cross-modality correlations in these models did not receive much attention, leaving aside potentially meaningful information for segmentation purposes. The goal of this paper is therefore to provide insights into the performances of various hybrid and Transformers-based networks in the context of multi-modal tumor segmentation with BraTS, a popular segmentation challenge. The rest of the paper is organized as follows. Sect.2 introduces both ViT and CNN-ViT architectures. Sect.3 presents implementation details as well as the evaluation strategy. Sect.4 explains and discusses the results we obtained. Conclusions and perspectives are given in Sect.5.

2. METHODS

2.1. CNN backbone

The hybrid models we consider in this work are made of ResNet alike encoders at the shallower levels (Fig.1*a*) to capture compact features at multiple scales, while deeper levels are encoded with Transformer blocks to model long-range dependencies in a more global manner. Given an input $X \in \mathbb{R}^{C \times H \times W \times D}$ with spatial resolution $H \times W$,

This work benefited from state aid managed by the National Research Agency under the Future Investment Program bearing the reference ANR-17-RHUS-0005 (FollowKnee project). This work was also partially funded by France Life Imaging (grant ANR-11-INBS-0006).

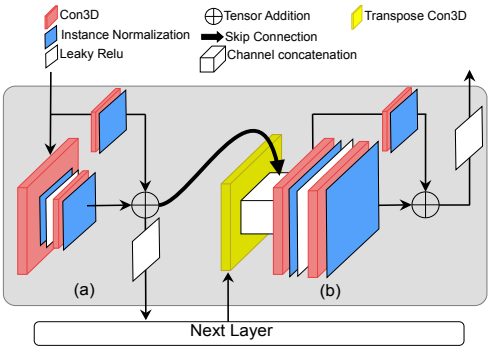

Fig. 1. Basic building blocks used to build the CNN backbone encoder: (*a*) down- and (*b*) up-sampling ResNet blocks.

D as depth dimension (# of slices) and C channels (# of modalities), down-sampling blocks gradually encode input images into a low-resolution/high-level feature representation $\mathcal{F} \in \mathbb{R}^{F \times \frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}}$ which is $\frac{1}{8}$ of input dimensions and where F represents the number of encoded feature maps [11]. An extra max-pooling operation is applied to further reduce the burden of computational complexity as input of the Transformer blocks. Such backbone can be extended to a multi-encoder-based framework, where independent encoders learn intra-modality feature representations meanwhile the Transformer block carries out the inter-modality fusion operation.

2.2. Transformer block

Since Transformers operate on 1D sequences, we reshape the input $X \in \mathbb{R}^{C \times H \times W \times D}$ into a sequence of flattened uniform non-overlapping patches $x_v \in \mathbb{R}^{N \times C \times P^3}$, where C can either represent the number of modalities or features maps F, (P, P, P) is the size of each patch and $N = HWD/P^3$ is the resulting number of patches, which is also the effective input length of the Transformer. Then, we distinguish three types of Transformer blocks: vanilla ViT, single-stream ViT (Fig.2*a*) and multiple-stream ViT (Fig.2*b*).

The vanilla ViT block (ViT_v) follows the design proposed in the original ViT paper [5]. First, a linear layer is used to project the flattened patches into a K-dimensional embedding space. Then, a 1D learnable patch position embedding $E_{pos} \in \mathbb{R}^{N \times k}$ is added to the projected embeddings to retain positional information. L Transformer blocks are then further stacked, comprising h multi-head attention and multi-layer perceptron (MLP) sub-layers. On the other hand, single-stream ViT (ViT_s) layers collectively operate on a concatenation of images modalities [14] by adding an extra modal-type embedding to the learnable token-position and projected embeddings. This design enables early and unconstrained fusion of cross-modal information. In multiple-stream ViT (ViT_m) [15], modalities are first handled by independent Transformer blocks. The resulting representations are then fed to a co-attention Transformer layer where intra-modal interactions are alternated with inter-modal interactions.

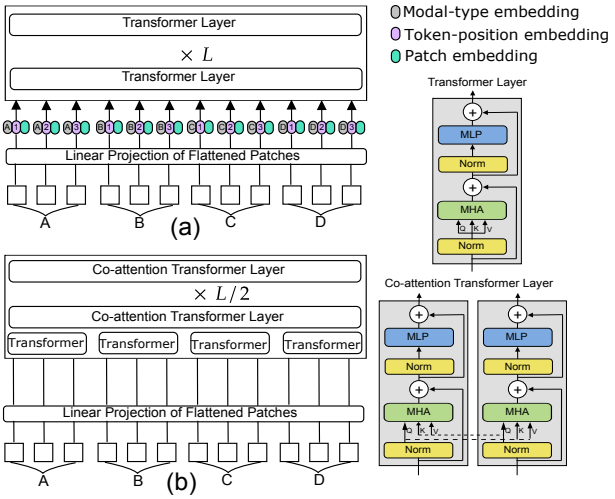

Fig. 2. Two possible modality interaction schemes: (*a*) single-stream (ViT_s), (*b*) multiple-stream (ViT_m) interactions.

2.3. Multi-level feature aggregation decoder

Decoding towards the segmentation mask space is handled using CNN up-sampling with multi-level feature aggregation. The building decoding block is shown in Fig.1*b*. For hybrid models, we first project back the output of the Transformer block in the layer L, $z_L \in \mathbb{R}^{N \times k}$ to $k \times \frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$. Then, we apply a transpose 3D convolution to go back to the original output of the CNN encoder $k \times \frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}$. The up-sampling blocks reduce the channel dimension of the decoder to match with the dimension of the skip-connections coming from the encoding part, decreasing computational complexity at the same time. Lastly, we perform a progressive feature up-sampling to gradually recover the full pixel-level segmentation segmentation $Y \in \mathbb{R}^{H \times W \times D}$. On the other hand, the ViT-based architecture follows the same decoder proposed in [9] that only differs from the aforementioned one in the operation needed to reshape the features from the multiple resolutions of the Transformers-based encoder.

2.4. Hybrid and Transformers-based architectures

The networks presented in this work follow a U-shaped design in which the extracted feature representations from encoder layers are fused with their decoder counterparts by concatenation for finer segmentation masks with richer spatial details through skip-connections. The proposed models are denoted as CNN+ViT_τ-B/P, MCNN+ViT_τ-B/P and ViT_τ-B/P, where the subscript τ represents the type of ViT block, B indicates the base ViT model configuration, P the patch size and M stands for multi-encoder based CNN. From the different possible configurations, we derive a total of seven mod-

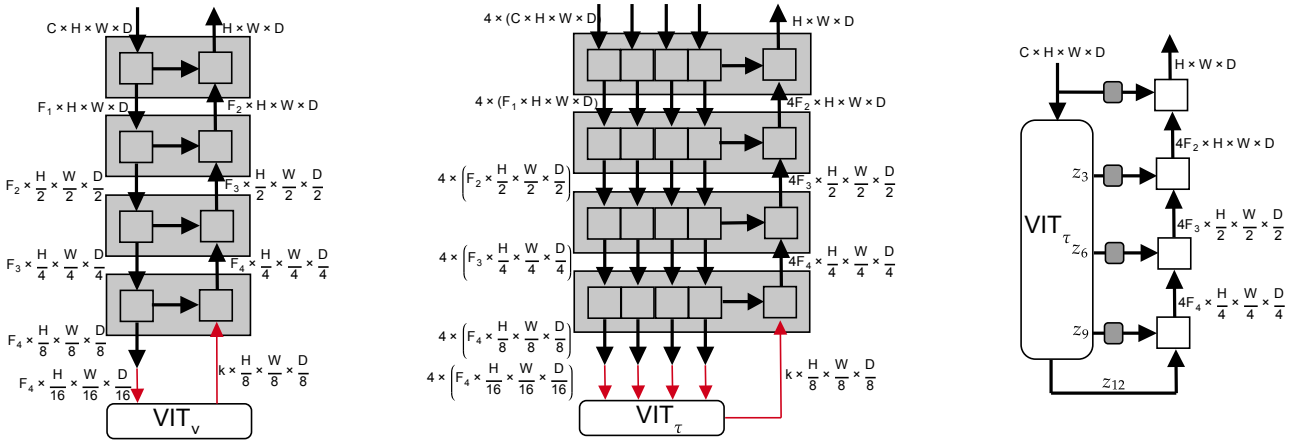

Fig. 3. Hybrid CNN-Transformers models: (*left*) CNN+ViT$_v$-B/P, (*middle*) MCNN+ViT$_\tau$-B/P, (*right*) ViT$_\tau$-B/P.

Hybrids	Transformers-based
CNN+VIT$_v$-B/1	VIT$_v$-B/16
MCNN+VIT$_v$-B/1	VIT$_s$-B/16
MCNN+VIT$_s$-B/1	VIT$_m$-B/16
MCNN+VIT$_m$-B/1	UNETR[9]

Table 1. Hybrids and Transformers-based architectures employed for comparisons in brain tumor segmentation.

els: four hybrids and three Transformers-based. The different models are illustrated in Fig.3 and summarized in Tab.1.

3. EXPERIMENTS

3.1. Imaging datasets

The magnetic resonance (MR) images used for this work are from the BraTS2021 training dataset [16, 17, 18]. The related BraTS challenge focuses on the evaluation of state-of-the-art methods for the segmentation of intrinsically heterogeneous brain glioblastomas. The training set contains 1251 subjects. The 3D MR scans include four modalities: native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2) and T2 fluid attenuated inversion recovery (T2-FLAIR). Annotations include Gadolinium-enhancing tumor region (ET), peritumoral edematous/invaded tissues and the necrotic tumor core that are combined to obtain three sub-regions: tumor core (TC), whole tumor (WT) and enhanced tumor (ET). All 3D volumes are skull-striped and resampled to 1mm^3 isotropic resolution with an input image size of $240 \times 240 \times 155$ voxels.

3.2. Implementation details

We implemented our models using PyTorch and MONAI[1]. Hybrid and Transformers-based models were trained using Nvidia A6000 and Titan RTX GPUs. All models were trained for a total of 150 epochs, with a batch size of 2 and NovoGrad

[1]https://monai.io/

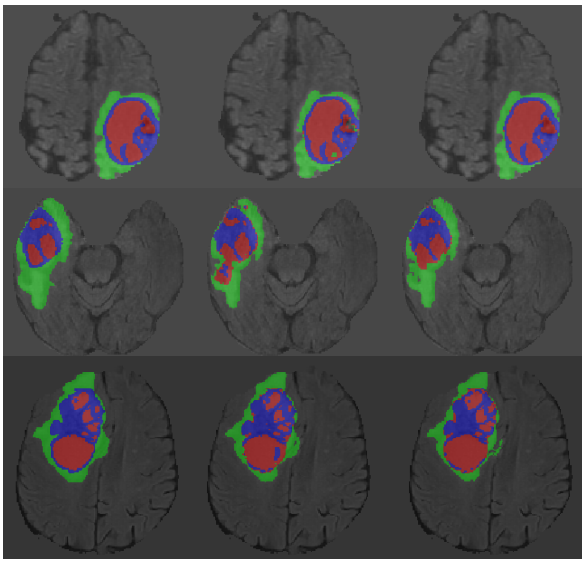

Fig. 4. Visual segmentation results: (*left*) ground truth, (*middle*) UNETR [9] and (*right*) MCNN+ViT$_v$-B/P. The green, red and blue colors respectively correspond to WT, TC and ET glioblastoma sub-regions.

as optimizer (learning rate $= 0.002$, weight decay $= 0.05$). The training objective was the sum of Dice and cross-entropy losses. The CNN feature maps F were set to 16, 32, 64 and 128 for all experiments. We used different patch resolutions as inputs to the Transformer blocks depending on the model implemented. For hybrid models, we used a patch size P of $1 \times 1 \times 1$. Meanwhile, the patch size was set to $16 \times 16 \times 16$ for Transformers-based models. The employed Transformer blocks followed the ViT base configuration [5] with $L = 12$ layers, an hidden dimension of $k = 768$, a MLP $= 3072$ and $h = 12$ heads. For inference, we used a sliding window approach with an overlap portion of 0.5. We did not use any pre-trained weights, neither for the CNN nor the ViT blocks. For a fair comparison, we followed both pre-processing and

Hybrid models (CNN-ViT)																
	CNN+ViT$_v$-B/1				MCNN+ViT$_v$-B/1				MCNN+ViT$_s$-B/1				MCNN+ViT$_m$-B/1			
DSC	ET	WT	TC	Avg.	ET	WT	TC	Avg.	ET	WT	TC	Avg.	ET	WT	TC	Avg.
Fold1	0.893	0.918	0.900	0.904	0.886	**0.925**	0.912	0.908	0.889	0.924	0.906	0.907	**0.895**	0.920	**0.920**	**0.911**
Fold2	0.875	0.916	0.913	0.901	**0.893**	**0.930**	0.912	**0.912**	0.882	0.905	0.911	0.899	0.876	0.914	**0.920**	0.903
Fold3	0.879	0.909	0.919	0.902	0.878	**0.929**	0.924	0.910	**0.890**	0.928	0.931	0.916	0.889	0.927	**0.937**	**0.918**
Fold4	0.901	0.919	0.923	0.914	**0.906**	0.920	0.934	0.920	0.905	**0.922**	**0.940**	**0.922**	0.904	0.918	0.923	0.915
Fold5	0.853	0.920	0.893	0.889	**0.877**	**0.928**	0.915	**0.907**	0.870	0.916	**0.917**	0.901	0.867	0.920	0.909	0.898
Avg.	0.880	0.916	0.909	0.902	**0.888**	**0.926**	0.919	**0.911**	0.887	0.919	0.921	0.909	0.886	0.920	**0.922**	0.909
Transformers-based models (ViT)																
	ViT$_v$-B/16				ViT$_s$-B/16				VIT$_m$-B/16				UNETR [9]			
Fold1	0.854	**0.893**	0.833	0.860	0.836	0.867	0.778	0.827	0.843	0.851	0.781	0.825	**0.862**	0.889	**0.867**	**0.873**
Fold2	0.840	**0.902**	0.827	0.856	0.838	0.850	0.811	0.833	0.841	0.870	0.815	0.842	**0.857**	0.896	**0.862**	**0.872**
Fold3	0.833	**0.889**	0.824	0.848	0.814	0.849	0.810	0.824	0.839	0.846	0.815	0.833	**0.847**	0.887	**0.862**	**0.865**
Fold4	0.874	0.898	0.839	0.870	0.852	0.844	0.810	0.835	0.839	0.852	0.782	0.824	0.875	**0.903**	**0.889**	**0.889**
Fold5	0.836	0.901	**0.859**	**0.866**	0.804	0.866	0.791	0.820	0.805	0.860	0.792	0.819	**0.840**	**0.905**	0.854	**0.866**
Avg.	0.848	**0.896**	0.836	0.860	0.829	0.855	0.800	0.828	0.833	0.856	0.797	0.829	**0.856**	**0.896**	**0.867**	**0.873**

Table 2. Five-fold cross-validation benchmark in terms of averaged Dice score for hybrid and Transformers-based models. ET, WT and TC respectively denote enhancing tumor, whole tumor and tumor core. Best results are in bold.

data augmentation strategies employed in nnU-Net [2] for all the implemented models.

3.3. Evaluation of predicted segmentation

We first splited the BraTS2021 training phase data with a 95:5 ratio to get both training and test sets. Over the new training set, we used 5-fold cross-validation with fixed 80:20 split for all experiments and evaluated the performance of our models using averaged Dice (DSC) scores. The evaluation was carried out using eight different models: four hybrids-based, three Transformer-based and UNETR [9] as baseline. The results of the cross-validation are summarized in Tab.2 meanwhile Fig.4 illustrates qualitative segmentation results for the best hybrid and Transformer-based models, over the test set.

4. RESULTS AND DISCUSSION

Results provided in Tab.2 show that MCNN+ViT$_v$-B/1 achieves the highest overall Dice score (0.911). However, its MCNN+ViT$_s$-B/1 and MCNN+ViT$_m$-B/1 variants are not far (0.909 for both). Concerning the tumor sub-regions, MCNN+ViT$_v$-B/1 is only outperformed by its multi-modal variants for TC by 0.002 and 0.003 respectively, which indicates no particular improvements from the use of modality interaction schemes. For all folds, multi-encoder based frameworks outperformed the single encoder approach by around 0.009. Another interesting finding is the fact that all the hybrids models outperformed the Transformers-based ones by approximately 0.038. This suggests that ViT blocks learn more helpful cross-modal representation when they rely on previous image feature extraction through a CNN backbone. The highest overall Dice score in the Transformers-based model was obtained with UNETR (0.873), closely followed by ViT$_v$-B/16 (0.860). Both models have the same encoder but they slightly differ in the decoding process as well as for skip-connections. This difference may be due to the design of the decoder, for which further studies out of the scope of the present work should be performed. Surprisingly, ViT$_s$-B/16 and ViT$_m$-B/16 are the worst ranked models with Dice scores of 0.828 and 0.829. This could be linked to the lack of multi-scale information, making it hard for the modality interaction schemes to discover meaningful cross-modal representations.

In these experiments, the hybrid MCNN+ViT$_v$-B/1 model achieved the best result. Learning modality-specific encoding using multi-encoder-based frameworks improved delineation performance (Fig.4) but at the expense of a higher computational cost, while pure ViT$_s$-B/16 and ViT$_m$-B/16 performed worse, suggesting that current modality interaction strategies may in fact be detrimental to model robustness when implemented in a pure Transformers-based fashion.

5. CONCLUSION

In this work, we compared a variety of ViT and hybrid ViT-CNN architectures in the context of multi-modal tumor segmentation using the BraTS dataset. The major conclusion of this preliminary work is that it seems better to take advantage of hybrid methods exploiting the local inductive bias from CNN encoders at shallower levels but also self-attention mechanisms able to capture longer-range dependencies among features, rather than using pure Transformers-based architectures. While single and multiple cross-modal interaction models did not achieve the highest performances in these experiments, future works will help us to better understand this issue, with the objective of achieving parameter-efficient architectures in a multi-modal context. The presented results are preliminary, and more experiments are indeed needed to generalize our conclusions to other datasets. Investigating pre-trained models, robustness to varying data regimes (from low- to high-) and hierarchical Transformers should also deserve more in-depth investigations.

6. REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.

[2] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2020.

[3] Pierre-Henri Conze, Ali Emre Kavur, Emilie Cornec-Le Gall, Naciye Sinem Gezer, Yannick Le Meur, M Alper Selver, and François Rousseau, "Abdominal multi-organ segmentation with cascaded convolutional and adversarial deep networks," *Artificial Intelligence in Medicine*, vol. 117, pp. 102109, 2021.

[4] Andrei Iantsen, Vincent Jaouen, Dimitris Visvikis, and Mathieu Hatt, "Squeeze-and-excitation normalization for brain tumor segmentation," in *International MICCAI Brainlesion Workshop*, 2020, pp. 366–373.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, 2020, pp. 213–229.

[7] Eunji Jun, Seungwoo Jeong, Da-Woon Heo, and Heung-Il Suk, "Medical Transformer: Universal brain encoder for 3D MRI analysis," *arXiv preprint arXiv:2104.13633*, 2021.

[8] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu, "Transformers in medical imaging: A survey," *arXiv preprint arXiv:2201.09873*, 2022.

[9] 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 *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 574–584.

[10] Jieneng Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[11] Wenxuan Wang, Chen Chen, Meng Ding, Jiangyun Li, Hong Yu, and Sen Zha, "TransBTS: Multimodal brain tumor segmentation using Transformer," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2021.

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin Transformer: Hierarchical vision Transformer using shifted windows," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.

[13] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu, "Swin UNETR: Swin Transformers for semantic segmentation of brain tumors in MR images," *arXiv preprint arXiv:2201.01266*, 2022.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional Transformers for language understanding," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[15] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee, "VilBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems*, 2019, vol. 32.

[16] Ujjwal Baid, Satyam Ghodasara, Michel Bilello, Suyash Mohan, Evan Calabrese, Errol Colak, Keyvan Farahani, and Jayashree Kalpathy-Cramer et al., "The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021.

[17] 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*, vol. 34, no. 10, pp. 1993–2024, 2014.

[18] 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*, vol. 4, pp. 170117, 2017.