3D CATBraTS: Channel Attention Transformer for Brain Tumour Semantic Segmentation

Rim El Badaoui
School of Computer Science and Engineering
University of Westminster
London, United Kingdom
R.Elbadaoui1@westiminster.ac.uk

Ester Bonmati Coll
School of Computer Science and Engineering
University of Westminster
London, United Kingdom
E.Bonmaticoll@westiminster.ac.uk

Aleka Psarrou
School of Computer Science and Engineering
University of Westminster
London, United Kingdom
A.Psarrou1@westiminster.ac.uk

Barbara Villarini
School of Computer Science and Engineering
University of Westminster
London, United Kingdom
B.Villarini@westiminster.ac.uk

***Abstract*—Brain tumour diagnosis is a challenging task yet crucial for planning treatments to stop or slow the growth of a tumour. In the last decade, there has been a dramatic increase in the use of convolutional neural networks (CNN) for their high performance in the automatic segmentation of tumours in medical images. More recently, Vision Transformer (ViT) has become a central focus of medical imaging for its robustness and efficiency when compared to CNNs. In this paper, we propose a novel 3D transformer named 3D CATBraTS for brain tumour semantic segmentation on magnetic resonance images (MRIs) based on the state-of-the-art Swin transformer with a modified CNN-encoder architecture using residual blocks and a channel attention module. The proposed approach is evaluated on the BraTS 2021 dataset and achieved quantitative measures of the mean Dice similarity coefficient (DSC) that surpasses the current state-of-the-art approaches in the validation phase.**

***Index Terms*—CNN, Transformers, ViT, Semantic Segmentation**

I. Introduction

A brain tumour is an abnormal growth of cells in the brain that can be cancerous (malignant) or non-cancerous (benign). There are over 100 types of brain tumours, which significantly vary in shape and size depending on their type and other factors namely the age and gender of the patient and at which stage the tumour is. Primary brain tumours are categorised into 2 grades: Low-Grade Glioma (LGG) and High-Grade Glioma (HGG). LGGs are benign brain tumours that tend to grow slowly, while HGGs are malignant tumours that grow fast and can damage brain tissue [1][2][3]. Characterisation of brain tumours in terms of shape, size, location and volumetric analysis is essential for surgical planning, progression prediction and life expectancy [4]. Automated medical image segmentation techniques can produce elaborated information imperative for accurately representing and analysing brain tumours.

The past years have seen the rapid development of Convolutional Neural Networks (CNNs) such as Deep Residual Learning (ResNet)[5] in healthcare for automated segmentation of tumours in the brain and other organs [6]. CNNs dominate the field because of their good performance and accurate predictions and for their ability to extract features and find patterns that are difficult to identify by traditional approaches[7].

The field of automated medical imaging with artificial intelligence (AI) is maturing, and we are witnessing a rapid development of promising algorithms that tend to be more robust and outperform traditional CNNs. The Attention-based Transformer Network is one of these trending frameworks that has introduced a new approach to solving Natural Language Processing (NLP) and Computer Vision tasks[8]. In 2017, Vaswani et al. introduced the first transformer neural network for machine translation[9]. Transformers are neural networks that rely on self-attention to handle global/long-range dependencies. After the success of the self-attention approach, the Brain Research team from Google applied minor modifications to the transformer and used it for image recognition. The refined version of the transformer is now known as Vision Transformer (ViT)[10].

Inspired by the success of the Vision Transformer, novel segmentation methods in medical imaging have been recently published. In particular, the Swin UNEt TRansformers (Swin UNETR), a ViT-CNN-based network, has been proposed for the 3D semantic segmentation of brain tumours on multi-modal magnetic resonance images (MRIs) achieving state-of-the-art results in terms of segmentation accuracy[11][12].

In this paper, we propose a novel model named Channel Attention Transformer for a 3-Dimensional MultiModal Brain Tumour Segmentation (3D CATBraTs). 3D CATBraTs is a modified version of the Swin UNETR following the encoder-decoder architecture. There are several important areas where this study makes an original contribution, which are:

- We introduce a novel network using ViT for a 3D segmentation of brain tumours on multi-modal MRIs named 3D CATBraTs.

- We propose a modified CNN-encoder architecture using residual blocks and a channel attention module.
- We show, through evaluation, that our model performs better compared to the top-performing models in the Brain Tumour Segmentation Challenge 2021 (BraTS 2021) validation phase [13], Swin UNETR, and the SegResNet [14].

The rest of the paper is organised as follows. Section II presents the two out-performing approaches in brain tumour segmentation that will be used in the comparison with our method in the results and evaluation section. Section III describes the methodology of the proposed brain tumour segmentation method. Section IV presents the quantitative results and findings of our approach. Finally, section V provides the conclusion of the proposed work.

II. Related Work

During the past decade, there has been a huge advancement in deep learning (DL) neural networks for computer vision. In healthcare, these models were mostly dedicated to developing an AI computer-aided diagnosis tool. Many diverse approaches are used for this task, some of which are based on CNNs, and others follow the Vision Transformers (ViTs) approach [16]. We are mainly concerned with two of the top-performing models in automated medical image segmentation, namely Swin UNETR and SegResNet, which are presented in this section.

CNNs are DL neural networks used for image classification and other computer vision tasks. They consist of multiple layers composed of convolutional layers, pooling layers, and fully connected layers [17]. On the other hand, ViTs are a type of Transformer-based architecture originally developed for NLP [18]. They have been adapted for computer vision by transforming the image into a sequence of patches and processing them using the Transformer architecture. Unlike CNNs, ViTs do not have any convolution or pooling layers and rely on self-attention mechanisms to capture relationships between the patches [19].

Swin Transformer (Swin-T) is a neural network designed for a range of computer vision tasks that was introduced as a primary ViT. Liu et al. proposed a hierarchical structure using shifted windows for a robust pixel-level prediction[15]. Swin-T presents different layers: the Patch Partition layer, which splits an RGB image into non-overlapping patches; the Linear Embedding layer, which takes the patches from the Patch Partition layer and applies a linear transformation on an n-dimensional concatenated features (raw pixel RGB values); the Patch Merging layer, which reduces the number of patches as the network gets deeper; and The Swin-T Block which was designed with a self-attention module based on shifted windows. Fig. 1 shows the architecture of a Swin-T. Following the success of Swin-T, much attention has been drawn to it with an increasing interest in its application for medical image segmentation combined with state-of-the-art CNN models such as the Swin UNETR.

Swin UNETR is based on the Swin-T merged with a CNN-based encoder-decoder network. Swin UNETR has a U-shaped architecture with an encoder and a decoder using residual blocks. The architecture was specifically designed for a 3D semantic segmentation of brain tumours[11]. The first step of the network creates partitions of the input image using the Swin Transformer to be fed to a four stages encoder. The encoded feature representations are then sent to a CNN-based decoder through skip connections at multiple resolutions [20][15]. The results of the network were compared to competing networks, including SegResNet, nnU-Net, and TransBTS. Interestingly, Swin UNETR outperformed all the other models proving that transformers with CNNs can produce better segmentation than traditional CNN-based networks.

SegResNet[1] is a CNN that follows the encoder-decoder architecture [14]. Myronenko et al. used a Variational AutoEncoder (VAE)[21] branch to the encoder endpoint to optimise the loss when using a small dataset. In SegResNet, the encoder blocks have the same architecture as the ResNet blocks, where downsampling is applied. After downsampling, the output of the encoder is fed to the decoder to perform upsampling. The structure of the decoder is similar to the encoder part while the output image size of this phase is the same as the original image size [5]. SegResNet has been applied to various semantic segmentation tasks and has achieved competitive performance compared to other state-of-the-art methods.

Swin UNETR and SegResNet are two top-ranking DL networks that have shown promising results for 3D semantic segmentation of brain tumours in the BraTS 2021 Challenge. In this paper, we compared our results to those obtained with these two state-of-the-art models to evaluate the performance of our proposed 3D CATBraTs.

III. Methodology

We designed our model by modifying the encoder of the Swin UNETR model using channel-wise attention Res blocks. Fig. 2 provides the overall architecture of the proposed model. Fig. 3 shows the architecture of the first level of the encoder. All the levels in the encoder follow the same construction. In more detail, the proposed network consists of several components, which are:

- Input: The input to the network is a brain MRI scan acquired using 4 different MRI modalities.
- Swin Transformer: is the main transformer component of our model; it splits the input image into non-overlapping patches using a patch-splitting module. The Swin Transformer processes the sequence of tokens and extracts features from the input image [15]. This will generate blocks of merged patches and features that will be fed into the encoders block as shown in Fig. 2.
- Down-sampling: This phase is responsible for extracting high-level features from the previous blocks' output. Our encoder consists of five modified Res blocks which

[1]https://docs.monai.io/en/stable/_modules/monai/networks/nets/segresnet.html#SegResNet

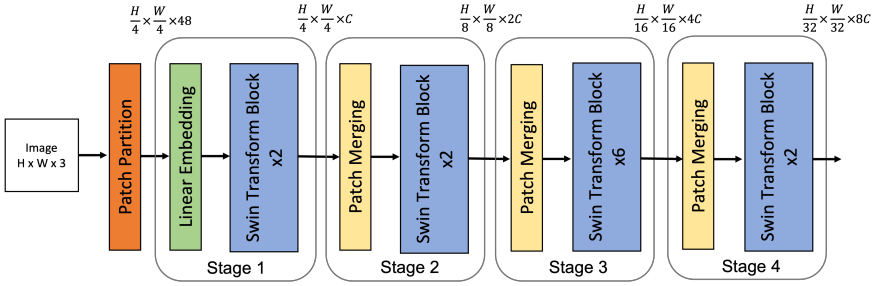

Fig. 1. Architecture of a Swin Transformer, which consists of 4 stages [15].

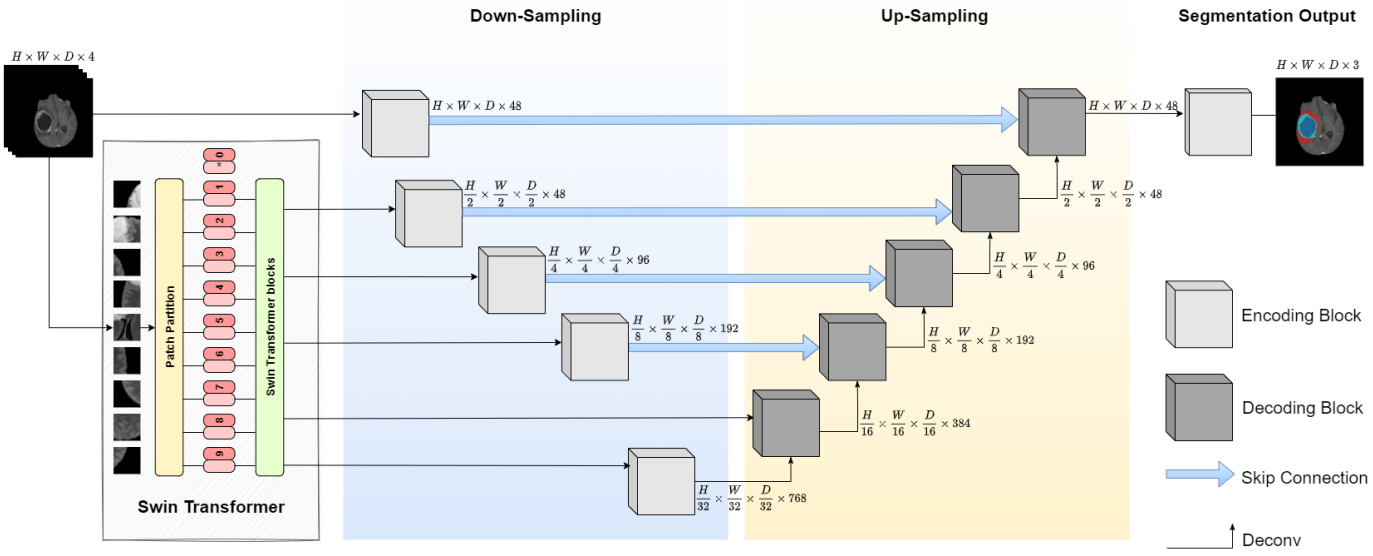

Fig. 2. Overview of Enhanced Swin UNETR architecture. The 3D input image is flattened into patches. The Swin Transformer blocks extract features from the patches and output blocks of merged patches and feature transformation. The output of the transformer is fed to a CNN encoder. The encoder consists of our modified version of the Res blocks. The decoder uses features generated by the encoder using skip connections. The final stage outputs 3 channels, denoted ET, WT, and TC, for Enhancing Tumour, Whole Tumour, and Tumour Core, respectively.

are enhanced with a Squeeze-and-Excitation channel-attention layer. The architecture of the first encoding block is represented in Fig. 3. The block consists of several convolutional and activation layers. The first layer is a 3D convolutional layer (Conv3D) with kernel size 3x3x3 and padding equals 1. This layer is responsible for performing convolution over the input. Then it is followed by a 3D Batch Normalisation layer (BatchNorm3d) and an activation layer such as LeakyReLU[22]. The LeakyReLU is concatenated to Conv3D and BatchNorm3d layers. The summation of these layers with BatchNorm3d of a Conv3D is given as input to the channel attention (ChannelSE) layer which is followed by LeakyReLU to output the feature maps. ChannelSE is a channel-wise attention layer that is formed of 2 layers: the global average pooling layer and a fully connected (FC) layer [23]. Feature maps hold different information learned from the input; for example, feature maps learning edges are more important than feature maps learning background representations. Hence, it is crucial to let the model decide which channel to focus on. With the channel-attention layer, we scale each channel based on its importance by dynamically utilising the dependencies between features, which will eventually improve the overall accuracy of the model. The output of the first block is $H \times W \times D \times 48$ while the last block is $\frac{H}{32} \times \frac{W}{32} \times \frac{D}{32} \times 768$

- Skip connections: A skip connection from the corresponding encoder layer helps to preserve low-level features in the image [24].
- Up-sampling: this phase uses the features generated by the encoder using skip connections and up-samples it back to the original resolution. This phase consists of five Res blocks [20]. The input of each of the decoding blocks is the output of the previous blocks and the output of the related encoding block with the exception of the first decoder block, which takes as input the output of the last encoder block and the last stage output of the

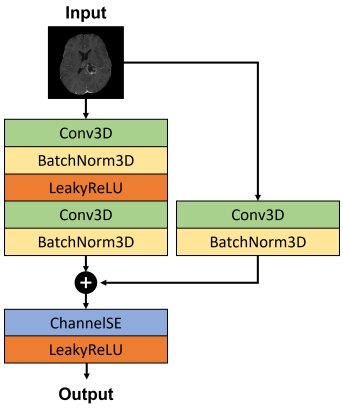

Fig. 3. 3D CATBraTs Encoding Block. ChannelSE is a channel-attention layer.

Swin-T.

- Output: The final output is the segmentation of the Enhancing Tumour (ET), Tumour Core (TC), and Whole Tumour (WT) [25].

A. Dataset

We used the BraTS 2021[2] dataset for training and evaluating the model [26] [25] [27]. The dataset includes multi-parametric MRI (mpMRI) scans of 1251 cases diagnosed with brain tumours. The MRI scans were acquired using four different contrast MRI modalities: native T1, post-contrast T1-weighted (T1Gd), T2-Weighted (T2), and fluid-attenuated inversion recovery (T2-Flair). All images were manually labelled and reviewed by neuro-radiologists to identify the ground truth location of the tumour on each of the modalities. The labelled regions are the Enhancing Tumour (ET), tumour necrosis (NCR) or tumour core (TC), and the whole tumour (WT), which includes both TC and ED (peritumoral edematous/invaded tissue). The MICCAI BraTS 2021 dataset is a high-quality dataset where the images are pre-processed by performing normalisation, histogram normalisation, interpolation to a uniform isotropic resolution, and skull stripping. The original height and width of the images are 240×240 with 155 slices for each of the modalities. We divided the dataset into 1001 cases used for training and 250 cases for validation.

B. Implementation Details

We implemented our model in Python using MONAI[3][28]. All models were trained and evaluated on an Acer Predator core i9 processor, 32 GB RAM, NVIDIA GeForce RTX 3080 10 GB. We used AdamW as an optimiser with an initial learning rate of 5e-5, progressively decreasing using the CosineAnnealingLR scheduler [29]. Regarding the loss function, we used the Dice loss function that can be computed as the complement of the Dice similarity coefficient with respect to 1. This loss function is useful in handling imbalanced data. Feature size was set at 48, and the dropping rate was at zero. We trained all models for 100 epochs on cropped MRIs with image size $128 \times 128 \times 96$ and batch size of 1.

[2]https://www.synapse.org/#!Synapse:syn27046444/wiki/616992

[3]https://docs.monai.io/en/stable/api.html

IV. Results and Discussion

In order to assess the 3D CATBraTs neural network, we performed a quantitative analysis of the proposed method as well as of Swin UNETR [11] and SegResNet [14]. We trained the models using the same hyper-parameters, such as learning rate and batch size and then fine-tuned them until we obtained an optimal model. We used the BraTS 2021 dataset to evaluate the segmentation performance of each of the methods in the evaluation phase. In terms of the evaluation metric, we used the mean Dice similarity coefficient (DSC), which is widely used in various medical imaging tasks. DCS measures the overlap between two masks and can be calculated using

$$DSC(X,Y) = \frac{2|X \cap Y|}{|X| + |Y|}$$

where X and Y are the masks from ground truth and predicted by a model [30].

Table I provides an overview of the results. The SegResNet scored the lowest Dice coefficient compared to Swin UNETR and to our approach. SegResNet achieved a mean DSC of 0.706, and 0.634, 0.704, and 0.780 for ET, TC, and WT, respectively. On the other hand, the mean DSC of Swin UNETR was 0.776, while our approach attained 0.834. Further, the proposed method recorded a notable improvement in each of the classes in comparison with the other architectures, achieving a DSC of 0.826 (ET), 0.799 (TC), and 0.876 (WT) whereas Swin UNETR achieved 0.780 (ET), 0.731 (TC), and 0.817 (WT). The overall validation DSC and the average loss plots of our model are shown in Fig. 5.

The results indicate that our method outperformed the competing method by 5.8%, 4.6%, 6.8%, and 5.9% for mean DSC of ET, TC, and WT, respectively. This comparison provides important insights into CNN transformers showing that they can outrank native CNNs. Furthermore, some visual results are provided in Fig.4 where examples of MRI slices from three different patients are displayed. The figure shows the comparison of the ground truth of the brain tumour for the three categories and the corresponding segmentation obtained with the proposed automated segmentation method.

TABLE I
VALIDATION RESULTS OF OUR MODEL, SEGRESNET, AND SWIN UNETR MODELS ON THE BRATS 2021 DATASET.

	DSC			
Network	Mean±std	ET	TC	WT
SegResNet	0.706 ± 0.317	0.634	0.704	0.780
Swin UNETR	0.776 ± 0.069	0.780	0.731	0.817
Our Model	**0.834** ± 0.133	**0.826**	**0.799**	**0.876**

V. Conclusion

This paper proposes a novel Swin-T model based on CNN encoder-decoder architecture for 3D segmentation of brain

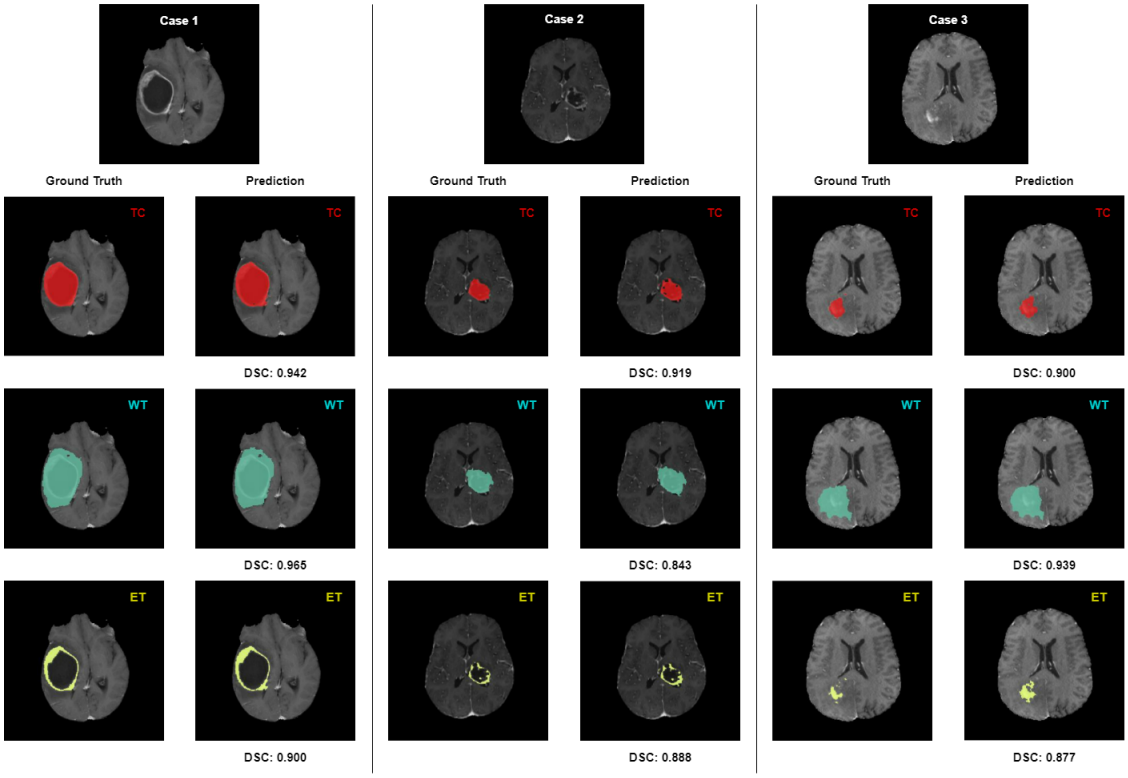

Fig. 4. Examples of images from three different cases showing the brain tumour categorised as Tumour Core (TC), Whole Tumour (WT), and Enhancing Tumour (ET). Ground truth and the predictions of the proposed model are shown with DSC.

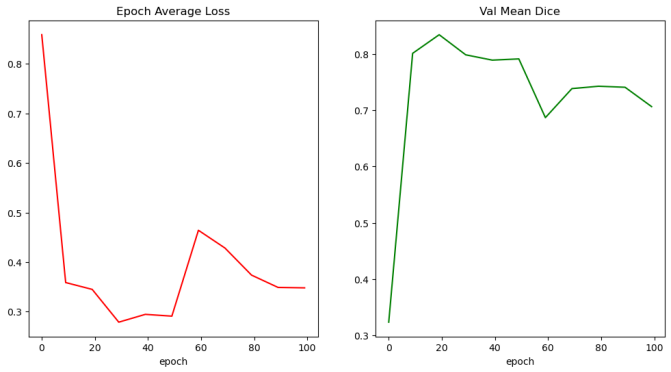

Fig. 5. The plot to the left is for the average training loss obtained by our method. The plot to the right shows the overall validation DSC.

tumours on multi-modal MRIs named 3D CATBraTs. The proposed model was inspired by the state-of-the-art Swin UNETR. The proposed modification on the encoder based on a channel-attention Res block showed increased performance in terms of tumour segmentation accuracy. We validated our results on the BraTS 2021 dataset, and we found that our method outperformed current state-of-the-art approaches in the validation phase.

ACKNOWLEDGEMENT

This research received no external funding.

REFERENCES

[1] A. Osborn, D. Louis, T. Poussaint, L. Linscott, and K. Salzman, "The 2021 world health organization classification of tumors of the central nervous system: What neuroradiologists need to know," *American Journal of Neuroradiology*, vol. 43, no. 7, pp. 928–937, Jul. 2022, ISSN: 0195-6108, 1936-959X. DOI: 10.3174/ajnr.A7462.

[2] S. Vidyadharan, B. V. V. S. N. Prabhakar Rao, Y. Perumal, K. Chandrasekharan, and V. Rajagopalan, "Deep learning classifies low- and high-grade glioma patients with high accuracy, sensitivity, and specificity based on their brain white matter networks derived from diffusion tensor imaging," *Diagnostics*, vol. 12, no. 12, p. 3216, Dec. 19, 2022, ISSN: 2075-4418. DOI: 10.3390/diagnostics12123216.

[3] L. M. DeAngelis, "Brain tumors," *New England Journal of Medicine*, vol. 344, no. 2, pp. 114–123, Jan. 11, 2001, Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJM200101113440207, ISSN: 0028-4793. DOI: 10.1056/NEJM200101113440207.

[4] D. Nie, H. Zhang, E. Adeli, L. Liu, and D. Shen, "3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens,*

Greece, October 17-21, 2016, Proceedings, Part II 19, Springer, 2016, pp. 212–220.

[5] K. He, X. Zhang, S. Ren, and J. Sun, *Deep Residual Learning for Image Recognition*, arXiv:1512.03385 [cs], Dec. 2015.

[6] B. Villarini, H. Asaturyan, S. Kurugol, O. Afacan, J. D. Bell, and E. L. Thomas, "3d deep learning for anatomical structure segmentation in multiple imaging modalities," in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 2021, pp. 166–171. DOI: 10.1109/CBMS52027.2021.00066.

[7] J. Gu, Z. Wang, J. Kuen, *et al.*, "Recent advances in convolutional neural networks," en, *Pattern Recognition*, vol. 77, pp. 354–377, May 2018, ISSN: 0031-3203. DOI: 10.1016/j.patcog.2017.10.013.

[8] S. Tuli, I. Dasgupta, E. Grant, and T. L. Griffiths, *Are Convolutional Neural Networks or Transformers more like human vision?* arXiv:2105.07197 [cs], Jul. 2021.

[9] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention Is All You Need*, arXiv:1706.03762 [cs], Dec. 2017.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, arXiv:2010.11929 [cs], Jun. 2021.

[11] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, *Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images*, arXiv:2201.01266 [cs, eess], Jan. 2022. DOI: 10.48550/arXiv.2201.01266.

[12] V. P. Grover, J. M. Tognarelli, M. M. Crossey, I. J. Cox, S. D. Taylor-Robinson, and M. J. McPhail, "Magnetic Resonance Imaging: Principles and Techniques: Lessons for Clinicians," en, *Journal of Clinical and Experimental Hepatology*, vol. 5, no. 3, pp. 246–255, Sep. 2015, ISSN: 09736883. DOI: 10.1016/j.jceh.2015.08.001.

[13] U. Baid, S. Ghodasara, S. Mohan, *et al.*, *The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification*, arXiv:2107.02314 [cs], Sep. 2021.

[14] A. Myronenko, *3D MRI brain tumor segmentation using autoencoder regularization*, arXiv:1810.11654 [cs, q-bio], Nov. 2018.

[15] Z. Liu, Y. Lin, Y. Cao, *et al.*, *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*, arXiv:2103.14030 [cs] version: 2, Aug. 2021.

[16] J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, and S. K. Zhou, "Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives," en, *Medical Image Analysis*, p. 102 762, Jan. 2023, ISSN: 1361-8415. DOI: 10.1016/j.media.2023.102762.

[17] K. O'Shea and R. Nash, *An Introduction to Convolutional Neural Networks*, arXiv:1511.08458 [cs], Dec. 2015.

[18] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," en, *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–41, Jan. 2022, ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3505244.

[19] O. Moutik, H. Sekkat, S. Tigani, *et al.*, "Convolutional Neural Networks or Vision Transformers: Who Will Win the Race for Action Recognitions in Visual Data?" en, *Sensors*, vol. 23, no. 2, p. 734, Jan. 2023, Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 1424-8220. DOI: 10.3390/s23020734.

[20] J. C. Ye and W. K. Sung, *Understanding Geometry of Encoder-Decoder CNNs*, arXiv:1901.07647 [cs, stat], May 2019.

[21] D. P. Kingma and M. Welling, *Auto-Encoding Variational Bayes*, arXiv:1312.6114 [cs, stat], Dec. 2022.

[22] B. Xu, N. Wang, T. Chen, and M. Li, *Empirical evaluation of rectified activations in convolutional network*, Nov. 27, 2015. arXiv: 1505.00853[cs,stat].

[23] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, *Squeeze-and-excitation networks*, May 16, 2019. arXiv: 1709.01507[cs].

[24] F. Liu, X. Ren, Z. Zhang, X. Sun, and Y. Zou, *Rethinking Skip Connection with Layer Normalization in Transformers and ResNets*, arXiv:2105.07205 [cs], May 2021.

[25] S. Bakas, H. Akbari, A. Sotiras, *et al.*, "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," en, *Scientific Data*, vol. 4, no. 1, p. 170 117, Sep. 2017, Number: 1 Publisher: Nature Publishing Group, ISSN: 2052-4463. DOI: 10.1038/sdata.2017.117.

[26] B. H. Menze, A. Jakab, S. Bauer, *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015, Conference Name: IEEE Transactions on Medical Imaging, ISSN: 1558-254X. DOI: 10.1109/TMI.2014.2377694.

[27] S. Bakas, M. Reyes, A. Jakab, *et al.*, *Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge*, arXiv:1811.02629 [cs, stat], Apr. 2019.

[28] M. J. Cardoso, W. Li, R. Brown, *et al.*, *Monai: An open-source framework for deep learning in healthcare*, 2022. arXiv: 2211.02701 [cs.LG].

[29] I. Loshchilov and F. Hutter, *SGDR: Stochastic Gradient Descent with Warm Restarts*, arXiv:1608.03983 [cs, math], May 2017.

[30] J. Bertels, T. Eelbode, M. Berman, *et al.*, "Optimizing the dice score and jaccard index for medical image segmentation: Theory & practice," in vol. 11765, 2019, pp. 92–100. DOI: 10.1007/978-3-030-32245-8_11. arXiv: 1911.01685[cs,eess].