3D Medical Axial Transformer: A Lightweight Transformer Model for 3D Brain Tumor Segmentation

Cheng Liu[1] 9645269447@edu.k.u-tokyo.ac.jp

Hisanori Kiryu[1] kiryu-h@k.u-tokyo.ac.jp

[1] *Kiryu Lab, The University of Tokyo, Tokyo, Japan*

Editors: Under Review for MIDL 2023

Abstract

In recent years, Transformer-based models have gained attention in the field of medical image segmentation, with research exploring ways to integrate them with established architectures such as Unet. However, the high computational demands of these models have led most current approaches to focus on segmenting 2D slices of MRI or CT images, which can limit the ability of the model to learn semantic information in the depth axis and result in output with uneven edges. Additionally, the small size of medical image datasets, particularly those for brain tumor segmentation, poses a challenge for training transformer models. To address these issues, we propose 3D Medical Axial Transformer (MAT), a lightweight, end-to-end model for 3D brain tumor segmentation that employs an axial attention mechanism to reduce computational demands and -distillation to improve performance on small datasets. Results indicate that our approach, which has fewer parameters and a simpler structure than other models, achieves superior performance and produces clearer output boundaries, making it more suitable for clinical applications.

Keywords: Deep learning, 3D brain tumor segmentation, 3D Transformer, axial attention, self-distillation

1. Introduction

Medical image segmentation is a key component in computer-aided diagnosis and a fundamental procedure in medical image processing (Doi, 2007). It helps clinicians make more accurate diagnoses and treatment decisions by segmenting organs or tumors in medical scans. With the development of convolutional neural networks (CNNs), Unet (Ronneberger et al., 2015) emerged as a popular medical image segmentation network with its simple U-shaped structure and innovative skip connections design. Many variations of Unet have been developed, including V-Net (Milletari et al., 2016), Res-Unet (Zhang et al., 2018), H-Dense-Unet (Li et al., 2018) and 3D-Unet (Çiçek et al., 2016) for 3D medical image segmentation. Specifically, the nnUnet (Isensee et al., 2021) model has demonstrated state-of-the-art performance in a wide range of tasks, including brain tumor segmentation. CNNs have achieved success in medical image segmentation, but struggle with long-range dependencies and processing global context (Vaswani et al., 2017). This is especially problematic in brain tumor images, where doctors need to combine information from multiple regions for diagnosis (Valanarasu et al., 2021).

Transformer-based models, utilizing self-attention mechanisms, have become popular in natural language processing and set state-of-the-art benchmarks in recent years (Devlin et al., 2018; Brown et al., 2020). These models are able to efficiently compute dependencies

© 2023 C. Liu & H. Kiryu.

between sequential inputs, even when distant from each other, addressing the problem of long-range dependencies which traditional convolutional models struggle with. In computer vision, the Vision Transformer (ViT) (Dosovitskiy et al., 2020) adapted the Transformer for image classification tasks with successful results. Subsequently, Swin-Transformer (Liu et al., 2021) has combined the sliding window concept of CNNs with self-attention mechanisms to form the Transformer using shifted window methods, capable of handling downstream tasks such as classification, detection, and segmentation.

In the field of medical image segmentation, Trans-UNet (Chen et al., 2021) pioneered the use of transformer-based models by integrating them into the traditional Unet architecture. Another notable approach is Medical Transformer (MedT)(Valanarasu et al., 2021), which uses gated axial attention for efficient semantic feature learning. Additionally, Swin-Unet (Cao et al., 2021) leveraged the efficient structure of the Swin-Transformer to achieve superior performance. However, the high computational cost of Transformer models with 3D inputs has limited their application to mainly 2D slices. UNETR(Hatamizadeh et al., 2022) is an exception, achieving good results on 3D medical images, but at the cost of high GPU memory consumption. To mitigate this, CoTr (Xie et al., 2021) proposed dividing 3D images into smaller blocks, but this also sacrifices semantic information. AFTer-Unet (Yan et al., 2022) utilized several neighboring axial slices as a 3D input image, reducing resource overhead but limiting global dependencies and requiring extra pre-processing.

In this study, we propose the 3D Medical Axial Transformer (MAT) model for efficient 3D brain tumor segmentation. Building on the success of MedT (Valanarasu et al., 2021), we introduce three one-dimensional gated axial attention mechanisms within the Transformer module to decompose the three-dimensional attentions, reducing GPU resource consumption and memory cost. Unlike AFTer-Unet, we focus on the entire depth axis rather than on neighboring axial slices. Additionally, we use a self-distillation training technique to improve performance on small datasets by using the model's output on the previous mini-batch as a soft target. Our main contributions are: (1) an end-to-end model that eliminates the need for pre-processing and allows for the direct use of 3D images, (2) a 3D self-attention module utilizing axial attention mechanism, which effectively reduces resource consumption, and (3) the use of self-distillation with a warm-up schedule, a novel approach in medical image processing, which helps the transformer module learn information with small-scale datasets.

2. Methodology

The proposed 3D Medical Axial Transformer (MAT) model utilizes a Transformer architecture with axial attention mechanisms in the encoder component, in combination with CNNs. The encoder is connected to a decoder, composed of CNNs, through skip connections at various resolutions. It is noteworthy that, due to the constraints imposed by the 3D input on the mini-batch size, group normalization (GN) is employed as a regularization technique in MAT, as opposed to batch normalization. The overall architecture of the model is depicted in Figure 1(a). In this section, we present a thorough description of the encoder and decoder of the proposed MAT model.

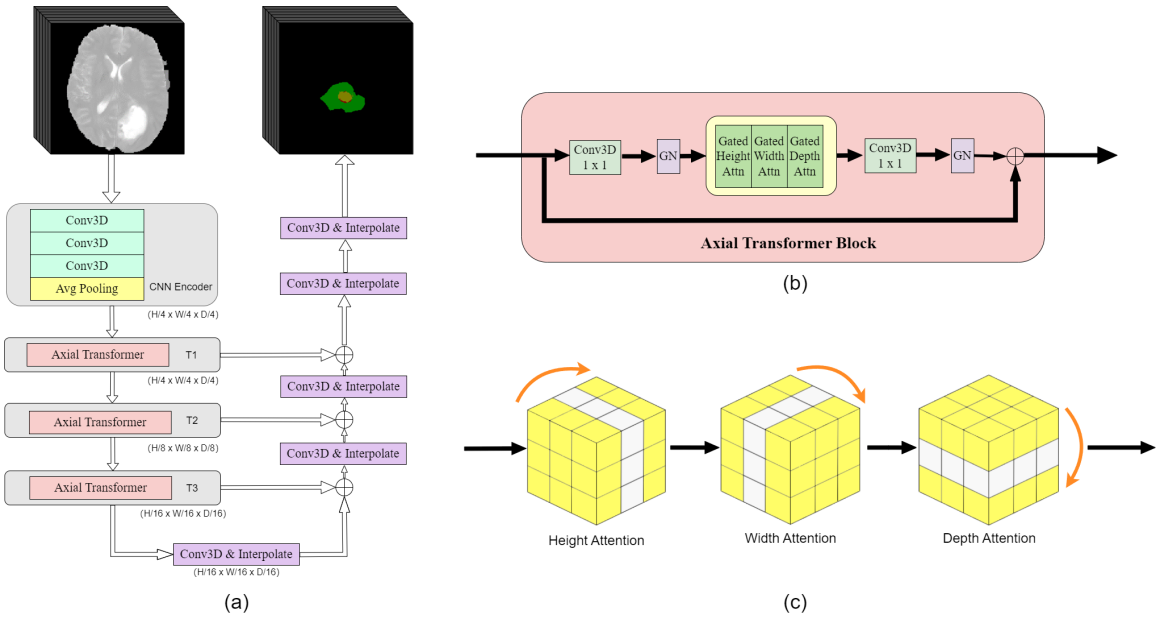

Figure 1: (a) Architecture overview of MAT (b) designs of Axial Transformer Block (c) Schematic diagram of the Axial Attention calculation method

2.1. CNN Encoder

MAT utilizes a minimal number of CNN blocks to extract features before passing the input to the Transformer encoder, following the design of MedT (Valanarasu et al., 2021). This approach allows the Transformer encoder to better learn semantic information and reduces the number of model parameters. The CNN encoder consists of one (7×7×7) convolutional block and two (5×5×5) convolutional blocks. The initial convolutional block (7×7×7) with a stride of 2 improves feature extraction, as established in (Simonyan and Zisserman, 2014), with a group normalization layer added between the convolutions. In contrast to MedT, an average pooling layer of size (2×2×2) with stride 2 is implemented between the CNN encoder and the Transformer encoder to extract further features while conserving memory resources.

Suppose the input to the encoder is a 3D image $x \in \mathbb{R}^{C \times D \times H \times W}$, where (D × H × W) is the image resolution and C is the channel of the input (e.g. C = 4 channels for MRI). After the CNN encoder, the output should be $x_{CNN} \in \mathbb{R}^{C^{cnn} \times \frac{D}{4} \times \frac{H}{4} \times \frac{W}{4}}$, where C^{cnn} is the number of channels output from the last convolution layer.

2.2. Axial Transformer Encoder

After using CNNs to extract shallow image features, we introduce Transformer encoders to facilitate the learning of deep semantic information. To reduce computational complexity, we implement the Transformer blocks using 3D axial attention, known as Axial Transformer, which allows for self-attention computation in all three dimensions, resulting in a comprehensive 3D modeling of the images, as illustrated in Figure 1(b). This design choice

effectively reduces resource consumption while enabling the model to effectively process and understand the full context of the medical images.

2.2.1. Axial Attention

In this work, we propose the use of three axial attention mechanisms in MAT for efficient 3D self-attention computation. This approach decomposes the 3D attention calculation into three 1D calculations in the three dimensions, reducing computational complexity, as illustrated in Figure 1(c). We also incorporate a learnable positional bias term and relative positional encoding in the self-attention module to effectively capture positional information in the 3D medical images. This approach builds upon previous research (Ho et al., 2019; Wang et al., 2020) that have shown the effectiveness of axial attention in capturing semantic information in 3D images. As an example, the output of the height axial attention is illustrated as Eq (1), Where R^Q, R^K, and R^V are the relative position encoding for queries, keys, and values respectively.

$$y_{\text{height }(i,j,k)} = \sum_{h=1}^{H} \text{Softmax}\left(Q^T_{(i,j,k)} K_{ihk} + Q^T_{(i,j,k)} R^Q_{ihk} + K^T_{ihk} R^K_{ihk}\right)\left(V_{ihk} + R^V_{ihk}\right) \tag{1}$$

To address the challenge of training the Transformer on small datasets and ensuring adequate position encoding, we incorporate a gating mechanism, as proposed in previous research (Valanarasu et al., 2021). The gated axial attention on the height axis is as follows.

$$y_{\text{height }(i,j,k)} = \sum_{h=1}^{H} \text{Softmax}\left(Q^T_{(i,j,k)} K_{ihk} + G^Q Q^T_{(i,j,k)} R^Q_{ihk} + G^K K^T_{ihk} R^K_{ihk}\right)\left(G^V_1 V_{ihk} + G^V_2 R^V_{ihk}\right) \tag{2}$$

Where G^Q, G^K, G^V_1, and G^V_2 are gates for queries, keys, and values, respectively. These gates act as learnable parameters that effectively control the final output of attention. In general, the gating mechanism limits the output of poor position encoding and gives higher weights to those that have been learned relatively well (Valanarasu et al., 2021).

2.2.2. Architecture

Our Axial Transformer block is based on the traditional Transformer design and features 3D convolutional layers, group normalization, and axial attention for height, width, and depth, as depicted in Figure 1(b). The output of the last GN layer is connected to the input via a skip connection (Vaswani et al., 2017; He et al., 2016).

Medical images have been found to require higher accuracy rather than a complex method for processing, based on research of various datasets (Isensee et al., 2021). To address this, we have chosen to use a lightweight model with fewer blocks to avoid issues such as overfitting or difficulty in practical application. We use 3 modules for axial attention computation and divide it into three stages: T1, T2, and T3, with an average pooling layer (strides = 2) behind each of the modules to reduce feature map resolution, except for T1 to prevent loss of important semantic information at an early stage.

The input is assumed to be $x_{CNN} \in \mathbb{R}^{C^{cnn} \times \frac{D}{4} \times \frac{H}{4} \times \frac{W}{4}}$ (the output of the CNN encoder). Each stage expands the number of channels by a factor of two through multi-headed atten-

tion, so the output of each stage turns to be $\left(2C^{cnn} \times \frac{D}{4} \times \frac{H}{4} \times \frac{W}{4}\right)$, $\left(4C^{cnn} \times \frac{D}{8} \times \frac{H}{8} \times \frac{W}{8}\right)$ and $\left(8C^{cnn} \times \frac{D}{16} \times \frac{H}{16} \times \frac{W}{16}\right)$.

2.3. CNN Decoder

Our CNN decoder is designed using the Unet architecture (Ronneberger et al., 2015), with a structure largely symmetric to the Axial Transformer encoder. It comprises five CNN-interpolation blocks, where the first block has, a stride of 1, while the remaining blocks have a stride of 2. The ReLU activation function is employed, and the image resolution is gradually increased by a factor of two with each block on the last four blocks. A skip connection is utilized to connect the encoder and decoder to restore details lost during downsampling. Instead of the traditional four skip connections in the up-sampling blocks, we used only three, discarding the topmost one as suggested by (Guo et al., 2022) to avoid negative impact from noisy shallow layer semantic information.

Let the final output of Axial Transformer encoder be $\mathrm{X}_{\text{encoder}}$ with a resolution of $\left(\frac{D}{16} \times \frac{H}{16} \times \frac{W}{16}\right)$, then the output after five CNN-interpolation blocks will be $\left(\frac{D}{16} \times \frac{H}{16} \times \frac{W}{16}\right)$, $\left(\frac{D}{8} \times \frac{H}{8} \times \frac{W}{8}\right)$, $\left(\frac{D}{4} \times \frac{H}{4} \times \frac{W}{4}\right)$, $\left(\frac{D}{2} \times \frac{H}{2} \times \frac{W}{2}\right)$ and $(D \times H \times W)$, which means that the resolution reduction is completed. At the last, a convolutional layer is used for classification, thus completing the downstream task of image segmentation.

2.4. Self-distillation for Regularization with warm-upped DLB

To address the challenge of training Transformer-based models on small datasets, we propose the use of self-distillation, a technique that utilizes the model's own output as a soft target for learning (Zhang et al., 2019). This approach allows for the optimization of the model directly through the training schedule, without the need for extensive modification of the architecture or the use of a large teacher model like traditional knowledge distillation (Hinton et al., 2015). (Bhat et al., 2021; Gani et al., 2022) has shown that self-distillation can be effective in improving performance on small datasets, and may also act as a regularization-like effect to aid in model training(He et al., 2022). Recently, (Shen et al., 2022) proposed a new method of implementing self-distillation, named DLB, where the model from the previous iteration is used as the teacher for the current iteration. The loss function for this method is formulated as:

$$\mathcal{L}_{LB} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{T}^2 \cdot D_{KL}\left(p_i^{\mathcal{T},t-1} \| p_i^{\mathcal{T},t}\right) \tag{3}$$

Where n is the number of samples in a batch, $\mathcal{T}$ is the temperature of the distillation, and p_i is the predicted distribution of each sample (Shen et al., 2022).

To date, there is a lack of literature on the application of self-distillation to the task of medical image segmentation. We propose to integrate the DLB loss as an additional term to the primary loss function in our model architecture, which is a combination of weighted cross-entropy loss and Dice loss, commonly used in medical image segmentation tasks. The overall loss function for our model is presented in Eq (4).

$$\mathcal{L}_{MAT} = 0.4 \cdot \mathcal{L}_{CE} + 0.6 \cdot \mathcal{L}_{\text{Dice}} + \alpha \cdot \mathcal{L}_{LB} \tag{4}$$

Due to the Transformer models tending to converge more slowly and being more susceptible to instability during the early stages of training, the use of DLB may instead lead to the problem of each iteration affecting each other and eventually being difficult to converge. To mitigate these issues, we implement a dynamic warm-up schedule for the distillation coefficient α and temperature $\mathcal{T}$. Specifically, the value of α is set to 0 for the first 50 training epochs, before being set to 1. In the following epoch, $\mathcal{T}$ is linearly increased from 1 to 2, with the aim of placing more emphasis on the distribution of negative samples during later stages of training.

3. Experiments

3.1. Setup

3.1.1. Dataset

We conducted experiments using brain tumor datasets from BraTS:

BraTS is brain tumor datasets from the MICCAI Brain Tumor Segmentation Challenge for each year, which provides multimodal 3D brain MRI scans annotated by physicians, including ground truth segmentations of tumor regions. The dataset includes 4 MRI modalities per case (T1, T1c, T2, and FLAIR), with annotations of four tumor subregions - peritumoral edema, necrotic core, enhancing, and non-enhancing tumor core. These annotations are consolidated into three nested sub-regions - whole tumor (WT), tumor core (TC), and enhancing tumor (ET). The BraTS2018 (Menze et al., 2014a; Bakas et al., 2017c, 2018) dataset was collected from 19 institutions and includes both low-grade and high-grade gliomas. In this study, we focus on high-grade gliomas, as it allows us to better demonstrate the advantages of our model on small datasets. The HGG group contains 210 samples, randomly split into 180 training cases and 30 validation cases for validation. In comparison, BraTS2021 (Bakas et al., 2017c; Baid et al., 2021; Menze et al., 2014b; Bakas et al., 2017a,b) has a larger dataset and we randomly split it with 1200 cases for training sets and 200 for validation sets, providing a means to compare the performance of our method with larger datasets. We report the mean Dice score for the three tumor sub-regions as well as the 95% Hausdorff Distance for both BraTS2018 and BraTS2021. Results for BraTS2021 are included in the Appendix section due to space constraints.

3.1.2. Evaluation Metrics

As in previous studies (Valanarasu et al., 2021; Yan et al., 2022), we also used the Dice score and the 95% Hausdorff Distance to assess the overall accuracy of segmentation as well as the surface accuracy. The Dice and Hausdorff Distance are defined as follows:

$$\text{Dice}(T, P) = \frac{2\sum_{i=1}^{I} T_i P_i}{\sum_{i=1}^{I} T_i + \sum_{i=1}^{I} P_i} \tag{5}$$

$$HD\left(T', P'\right) = \max\left\{\max_{t' \in T'} \min_{p' \in P'} \left\|t' - p'\right\|, \max_{p' \in P'} \min_{t' \in T'} \left\|p' - t'\right\|\right\} \tag{6}$$

Where T_i and P_i denote the ground truth and predicted values of voxel i, while T' and P' denote the set of surface points of the ground truth and predicted values, respectively.

HD95 is the output of HD multiplied by 95% in order to eliminate the effect of a very small subset of the outliers.

3.1.3. Implementation details

To optimize the training process of our model, we re-sized all images in the BraTS 2018 and 2021 datasets to dimensions of (160×224×224) and (128×160×160). This was done to ensure consistency in the dimensions of the images and to fill any empty slices in cases where the depth of the image was insufficient. Data augmentation, including random flips and rotations, was applied with a probability of 50% to increase the model's fitting ability. The AdamW optimizer (Loshchilov and Hutter, 2018) with a weight decay of 10-5 was utilized with a warm-up schedule for the learning rate (Gotmare et al., 2018), where the learning rate grows linearly from 0 to 10-3 for the first 10 epochs, followed by cosine anneal (Loshchilov and Hutter, 2016) to complete the learning rate decay. In the Transformer encoder, we employed 1, 2, and 4 blocks for stage T1, T2, and T3, respectively. For the sake of fairness, in our experiments, all models were trained for an equal number of 250 epochs.

3.2. Results on BraTS2018

Table 1: Dice scores and HD95 of different methods on the BraTS2018 dataset (validation).

Metrics	WT		TC		ET		Mean	
	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95
nnUnet(3D)	85.61	4.33	78.72	6.59	70.23	4.91	78.19	5.28
Trans-Unet(2D)	84.42	4.91	75.12	7.18	73.97	5.07	77.84	5.72
Swin-Unet(2D)	90.64	6.01	80.78	7.07	75.54	5.98	82.32	6.35
AFTer-Unet(3D)	91.93	4.15	87.15	6.76	81.58	**3.91**	86.89	4.94
MAT(3D)	**93.05**	**4.06**	**87.91**	**6.09**	**82.81**	4.02	**87.92**	**4.72**

We compared our proposed 3D Medical Attention Transformer (MAT) model to other medical image segmentation models on the BraTS2018 dataset in Table 1. We selected several established models such as nnUnet (Isensee et al., 2021), Trans-Unet (Chen et al., 2021), Swin-Unet (Cao et al., 2021), and AFTer-Unet (Yan et al., 2022) for fair comparison. These models have shown state-of-the-art results in various medical image segmentation tasks and are widely used in the field.

The results in Table 1 show that all models achieved good segmentation ability in the Whole Tumor (WT) region, likely due to the clear distinction between tumor and non-tumor regions and larger relative segmentation volume. However, for the Enhancing Tumor (ET) region, which is considered the most challenging to segment, a high Dice score can only be achieved with 3D models that incorporate the Transformer module. The inclusion of the depth axis and long-range dependencies improves the model's understanding of the overall image, leading to better judgments. 3D models significantly outperformed 2D models in terms of HD95 metrics across all regions, likely due to the problematic jagged segmentation edges that arise when combining 2D slices into 3D images. MAT leverages 3D convolution to extract features and inputs feature maps with richer semantic information into the Axial

Transformer module, achieving true global long-range dependency modeling, which resulting in superior performance. Overall, MAT achieved a mean Dice of 87.92% and a mean HD95 of 4.72 on the BraTS2018, surpassing AFTer-Unet in nearly all metrics.

3.3. Qualitative results on BraTS2018

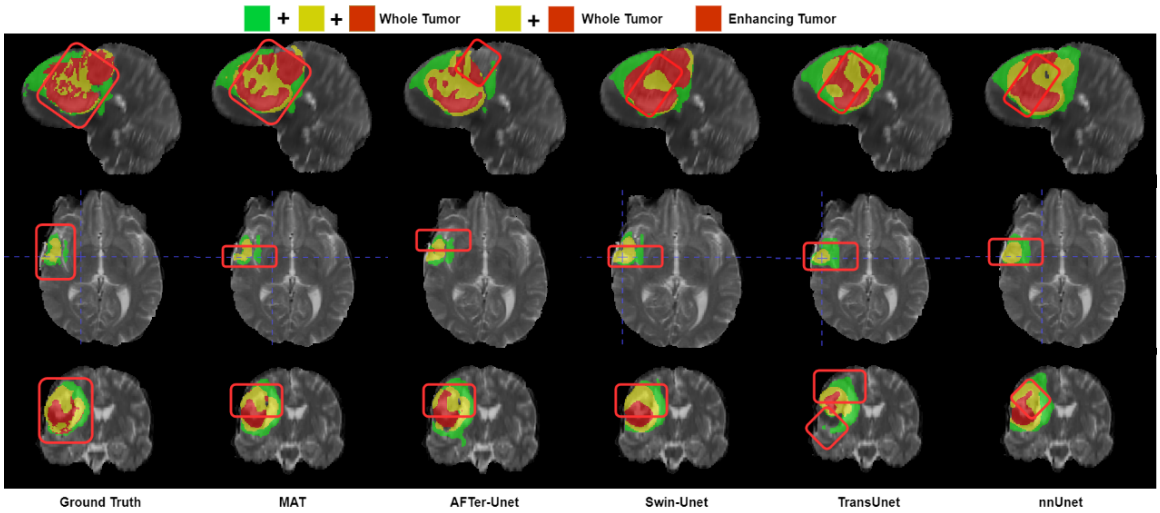

Figure 2: Qualitative results of different methods on the BraTS2018 dataset. The first row is sagittal, the second row is axial, and the third row is the coronal view. Red boxes identify the parts that demonstrate the validity of our method.

Figure 2 presents qualitative results of different models on the BraTS2018 dataset. We presented slices of images from three views (sagittal, axial, and coronal) to facilitate the comparison of 3D segmentation results. The most challenging regions to segment in brain tumor images are TC (yellow part + red part) and ET (red part) due to their small distinction from other tumors and small volume. The design of MAT's 3D Axial Transformer module with DLB enables it to learn global information, resulting in excellent segmentation ability for ET as seen in sagittal and coronal views. Additionally, as shown in the axial view (second row) and sagittal view (first row), MAT's predictions of gaps in tumors are more accurate due to its pixel-level long-distance dependence in three dimensions, in contrast to other models. Overall, MAT demonstrates a superior ability for multi-class tumor segmentation compared to other models, for both overall tumor region and local gaps.

4. Conclusion

We propose 3D Medical Axial Transformer(MAT), a 3D end-to-end framework for brain tumor image segmentation that utilizes the Axial Transformer and self-distillation techniques. The design of MAT enables efficient learning of semantic information while maintaining a lightweight architecture, making it suitable for clinical applications. Our experiments on brain tumor datasets demonstrate the superiority of MAT over previous related methods.

References

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, and J. Kirby. Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. *The Cancer Imaging Archive*, 2017a. doi: 0.7937/K9/TCIA.2017.KLXWJJ1Q.

S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, and J. Kirby. Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. *The Cancer Imaging Archive*, 2017b. doi: 10.7937/K9/TCIA.2017.GJQ7R0EF.

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, 2017c.

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.

Prashant Bhat, Elahe Arani, and Bahram Zonooz. Distill on the go: Online knowledge distillation in self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2678–2687, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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. *arXiv preprint arXiv:2105.05537*, 2021.

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.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4-5):198–211, 2007.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Hanan Gani, Muzammal Naseer, and Mohammad Yaqub. How to train vision transformer on small-scale datasets? *arXiv preprint arXiv:2210.07240*, 2022.

Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243*, 2018.

Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *arXiv preprint arXiv:2209.08575*, 2022.

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*, pages 574–584, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Yangji He, Weihan Liang, Dongyang Zhao, Hong-Yu Zhou, Weifeng Ge, Yizhou Yu, and Wenqiang Zhang. Attribute surrogates learning and spectral tokens pooling in transformers for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9119–9129, 2022.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018.

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, 2014a.

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, 2014b.

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)*, pages 565–571. IEEE, 2016.

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*, pages 234–241. Springer, 2015.

Yiqing Shen, Liwu Xu, Yuzhe Yang, Yaqian Li, and Yandong Guo. Self-distillation from the last mini-batch for consistency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11943–11952, 2022.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 36–46. Springer, 2021.

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.

Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2020.

Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 171–180. Springer, 2021.

Xiangyi Yan, Hao Tang, Shanlin Sun, Haoyu Ma, Deying Kong, and Xiaohui Xie. After-unet: Axial fusion transformer unet for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3971–3981, 2022.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722, 2019.

Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.

Appendix A. Model parameters and memory cost

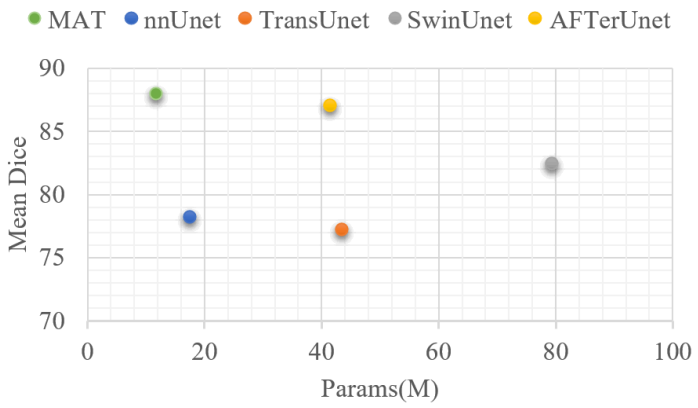

Figure 3: Comparison of other medical image segmentation models with our proposed model in terms of the number of parameters and mean Dice score.

To fully illustrate the benefits of our proposed method, we have compared the number of parameters used by nnUnet (Isensee et al., 2021), Trans-Unet (Chen et al., 2021), Swin-Unet (Cao et al., 2021), AFTer-Unet (Yan et al., 2022), and MAT with the final mean Dice score at BraTS2018. As shown in Fig 3, our 3D axial attention algorithm significantly reduces the model's complexity while still maintaining excellent segmentation performance, effectively highlighting the superiority of our model architecture. Additionally, our MAT model has a lower computational overhead compared to other Transformer-based models and can be fully trained on a single Tesla T4 (16GB) GPU.

Appendix B. Results on BraTS2021

Table 2: Dice scores and HD95 of different methods on the BraTS2021 dataset (validation).

Metrics	WT		TC		ET		Mean	
	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95
nnUnet(3D)	92.14	7.33	89.56	3.94	83.67	4.02	88.46	5.10
Trans-Unet(2D)	91.73	7.92	85.47	6.02	81.25	4.68	86.15	6.21
Swin-Unet(2D)	**93.51**	7.51	90.64	5.61	**85.34**	4.18	89.83	5.77
AFTer-Unet(3D)	93.47	**6.95**	90.48	3.76	85.31	**3.29**	89.75	**4.67**
MAT(3D)	93.21	7.13	**91.91**	**3.56**	85.05	3.61	**90.06**	4.77

In Table 2, we present a comparison of our proposed MAT model with other existing medical image segmentation models on the BraTS2021 dataset. As before, we compared MAT against nnUnet (Isensee et al., 2021), Trans-Unet (Chen et al., 2021), Swin-Unet (Cao et al., 2021), and AFTer-Unet (Yan et al., 2022). However, it's worth noting that unlike BraTS2018, the BraTS2021 dataset contains a larger amount of MRI data. Given that our

model is designed to be lightweight, it may face challenges in results when dealing with larger datasets such as BraTS2021.

Despite this, the results show that MAT's performance is comparable to other larger models. Notably, MAT performed significantly better than the other models in the TC (tumor core) region, with a Dice score improvement of 1.27% and a reduction of 0.20 in HD95 score. This highlights MAT's strong ability to learn with a limited number of parameters. Furthermore, on average, MAT is the best model under the Dice metric, and only slightly below the AFTer-Unet at 0.10 according to HD95.

Appendix C. Ablation study of warm-upped DLB

Table 3: Ablation study of DLB with warm-up schedule Method on Brain Tumor Segmentation Using BraTS2018 and BraTS2021 Datasets through Mean Dice Score.

Mean Dice	**BraTS 2018**		**BraTS 2021**	
	Training	Validation	Training	Validation
MAT without DLB	85.97	86.03	89.51	89.93
MAT with original DLB	77.13	79.22	82.76	82.92
MAT with warm-upped DLB	**87.65**	**87.92**	**90.02**	**90.06**

To evaluate the efficacy of the proposed warm-upped DLB method, ablation experiments were conducted on the BraTS2018 and BraTS2021 datasets, as reported in Table 3. In these experiments, the parameters $\mathcal{T}$ and α of the original DLB were set to 3 and 1, respectively, as per the sets in (Shen et al., 2022).

The results in Table 3 show that the warm-upped DLB method the performance of the MAT model on the BraTS2018 dataset, with an increase of 1.68% and 1.89% for the training and validation sets, respectively. Furthermore, the results indicate that the absence of warm-up led to a decrease in performance, highlighting the importance of this strategy. Additionally, while the warm-upped DLB still resulted in the best performance on the BraTS2021 dataset, the improvement was smaller, suggesting that this method is more effective for small-scale datasets.

Appendix D. The output of axial attention for depth axis and width axis

$$y_{\text{depth }(i,j,k)} = \sum_{d=1}^{D} \text{Softmax}\left(Q^T_{(i,j,k)}K_{djk} + Q^T_{(i,j,k)}R^Q_{djk} + K^T_{djk}R^K_{djk}\right)\left(V_{djk} + R^V_{djk}\right) \tag{7}$$

$$y_{\text{width }(i,j,k)} = \sum_{w=1}^{W} \text{Softmax}\left(Q^T_{(i,j,k)}K_{ijw} + Q^T_{(i,j,k)}R^Q_{ijw} + K^T_{ijw}R^K_{ijw}\right)\left(V_{ijw} + R^V_{ijw}\right) \tag{8}$$