Fast Brain Tumor Segmentation with Model Optimization

Mostafizur RAHMAN [1], Wenmin WANG [2], Jiawei WANG [3], Yu WANG [1,]*

[1] School of Computer and Artificial Intelligence, Beijing Technology and Business University, 100048, Beijing, China
[2] School of Computer Science and Engineering, Macau University of Science and Technology, 999078, Macao, China
[3] Dept. of Neurosurgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, PR China

mostafiz.btbu@gmail.com, wmwang@must.edu.mo, jwwang_beijing@126.com, wangyu@btbu.edu.cn

Submitted February 1, 2025 / Accepted April 28, 2025 / Online first June 10, 2025

Abstract. *Segmenting brain tumors is important for effective diagnosis and treatment planning. Conventional 3D segmentation models achieve high accuracy but are computationally intensive, often limiting real-time applicability. In this study, pseudo-3D convolutions, which consist of spatial and depthwise convolutions, are used in place of traditional 3D convolutions. Adaptive Dilated Multi-Fiber (DMF) units dynamically extract multi-scale features and parallel Multi-Fiber (MF) units combine them with weighted sum. Efficient Channel Attention (ECA) and Cross Attention improve feature selection and fusion in decoder and encoder. Structured pruning reduces superfluous parameters and Quantization Aware Training (QAT) increases the speed of inference with the model converted to INT8 precision. Combination of Dice Loss and Boundary Loss enhances the precision of tumor boundaries. The framework has been evaluated on the BraTS 2021 data validation set and achieved high Dice scores of Whole Tumor 91.85%, Tumor Core 88.52%, and Enhancing Tumor 85.55%, with Hausdorff95 values of 2.58 mm, 3.53 mm, and 3.65 mm. Our proposed model requires only 3.57M parameters and 21.26 GFLOPs, achieving an inference time of 0.016 seconds per 3D volume while maintaining precision alongside efficiency to clinical application.*

Keywords

Brain tumor segmentation, pseudo-3D convolutions, adaptive dilation, computational efficiency, inference time

1. Introduction

In a clinical context, brain tumor segmentation is an integral part of one's diagnosis, treatment strategy, and monitoring of a neurological disease. Providing accurate segmentation helps the necessary precautions and information needed for surgery and radiotherapy as well as progression tracking and evaluation of the treatment results [1], [2]. Decision making accuracy and speed are both vital factors while determining the desired outcome within a patient [3]. There is a fundamental need for optimization between utmost getting care to patients with brain tumor and computational resources.

Delineating three specific tumor regions on multimodal MRI scans is effective in brain tumor segmentation. The Enhancing Tumor (ET) refers to the actively growing tumor region visible on contrast-enhanced scans, providing crucial insights into aggressive tumor behavior [4]. The Tumor Core (TC) region includes both the surrounding edema as well as the necrotic and enhancing tumor regions, this is important for the understanding of the tumor central structure and progression. Whereas the Whole Tumor (WT) region represents the entire tumor structure which provides an overview of the extent and impact on surrounding brain tissue. These segmentation regions are helpful during treatment planning and clinical decision making and have shown rapid progress in automated segmentation and expert label validation [5].

It is important to note that there are significant computational implications while processing volumetric 3D MRI data for brain tumors segmentation. The most common 3D CNNs, like the 3D U-Net [6] and V-Net [7], are built upon multiple frameworks of 3D convolutions to accomplish this. Indeed, these architectures exhibit highly accurate segmentation. However, the multi-resolution designs lead to increased memory and computation requirements, rendering such approaches nonviable for coarse-grained or real-time clinical applications [8]. Ensemble models, as well as optimized 3D U-Net, have proven to achieve better segmentation, yet their implementation comes at the price of increased computational overhead and time expenditure. This limits the usability of these methods in decision-making sensitive situations, which are time constrained [9], [10].

It is an understatement that a balance has to be struck between segmentation accuracy and computational speed when it comes to brain tumor segmentation. Simplified architectures designed to enhance speed risk compromising segmentation precision, while highly accurate models may be too slow for practical use [11]. Developing optimized models that effectively balance these factors is essential for

DOI: 10.13164/re.2025.0393

advancing brain tumor segmentation in clinical practice, ensuring both accuracy and efficiency [12].

2. State of the Art

Recent studies have proposed strategies to address these challenges. For instance, 3D-ESPNet which extends ESPNet a fast and efficient network based on point-wise convolution for 2D semantic segmentation to 3D medical volumetric imaging data describes a significant decrease in the volumetric reasoning computational costs [13]. Analogously, SD-UNet uses separating 3D convolutions, splits each convolution into three parallel branches with the aim of decreasing the learnable parameters, thus increasing the effectiveness of the model [14]. While these lightweight architectures have made strides in computational efficiency, their segmentation performance often falls short of state-of-the-art models.

Real-time inference has been achieved during the 3D brain MRI segmentation process with high levels of accuracy. This is just one of many benefits brought on by multi-branch sharing networks [15]. Improved segmentation performance has been reported [16] due to the use of cascaded convolutional networks which have been used to optimize memory expenditure, model complexity, and the field of view. There is, however, a need for further development in the lightweight, real-time, and highly efficient models that can be used in clinical practices. These models are clearly quite different from the traditional ones employed in today's setting which are overly complex, inefficient, and tedious.

Efficiency in computation accompanied by a potential increase in feature representation and segmentation performance can be achieved through the use of attention mechanism and parallel convolution integration as seen in LATUP-Net and other extension models [17]. Additionally, research in segmentation tasks can focus on enhancing the ensemble deep learning models that have already shown an increase in classification accuracy with brain tumor grading [18].

Development of real time clinical applications requires further research even with advancements made so far. The goal is to complete ongoing efforts which involves creating models that achieve optimal accuracy metrics in segmentation while also being low in complexity to facilitate practical use in clinical settings with time constraints.

The DMFNet, proposed by Chen et al. [19], introduced a novel approach to 3D brain tumor segmentation by utilizing multi-fiber units with group convolutions to achieve computational efficiency while maintaining high segmentation accuracy. The application of dilated convolutions created an improved multi-scale feature representation, and a multiplexer moved information around efficiently through grouped residual units. With such few parameters and the ability to process volumes in milliseconds while maintaining powerful performance on the BraTS 2018 dataset, DMFNet proved its strong real-time capabilities, however, it suffered from severe drawbacks. Among it, reliance on expensive full 3D convolutions and inability to adapt dilation rates to different scales of lesions was the biggest one.

Inspired by the work of Chen et al., in this study DMFNet is built upon to propose a novel framework that achieves superior segmentation accuracy and computational efficiency. Previous efforts such as Liu et al. [20] and Wang et al. [21] introduced early use of dilated pseudo-3D convolutions and multi-direction fusion techniques for brain tumor segmentation, which laid the foundation for designing computationally efficient yet accurate models. First, full 3D convolutions in the residual units and multi-fiber design are replaced with pseudo-3D convolutions, which decompose operations into spatial 3×3×1 and depthwise 1×1×3 convolutions, significantly reducing FLOPs while retaining volumetric context. Lightweight attention mechanisms, including Efficient Channel Attention (ECA) [22], and Cross Attention are incorporated to enhance feature interactions and refine encoder-decoder feature fusion. Structured pruning techniques are applied to remove redundant parameters at the channel and layer levels [23], creating a lightweight model without sacrificing accuracy. Additionally, Quantization-Aware Training (QAT) [24] is employed to convert the model to INT8 precision, further accelerating inference.

To integrate Dice Loss with Boundary Loss was necessary to increase the segmentation accuracy and aid in the detection of small lesions alongside their boundaries. Auxiliary segmentation heads were introduced to inspire Multi-Scale Feature Learning during training. The fixed dilation strategies have been substituted with dynamic feature selection mechanisms that pick out the relevant features, while sub pixel convolution replaces the trilinear interpolation equipped in the decoder to enhance spatial detail recovery.

Evaluated on the BraTS 2021 and 2020 dataset, the proposed model demonstrates superior performance, achieving higher Dice scores, reduced Hausdorff Distance (HD95), and an inference time of 0.016 seconds per 3D volume. This work solves the real-time segmentation bottlenecks of computing tackling the issues of segmentation accuracy and computational efficiency and making it easier to use deep learning models in clinical settings. Our main contributions are the following:

(1) In the proposed framework the conventional 3D convolutions are replaced with pseudo-3D convolutions, decomposing operations into spatial 3×3×1 and depthwise 1×1×3 convolutions, significantly reducing computational overhead while retaining volumetric context.

(2) In adaptive DMF unit relevant features are dynamically selected through adaptive weighting mechanisms, improving multi-scale representation and segmentation precision.

(3) In parallel MF units the features from parallel fibers are efficiently aggregated while the adaptability is maintained via weighted summation.

(4) In ECA modules feature weighting is enhanced, enabling the network to focus on critical regions with minimal computational cost.

(5) Structured pruning techniques eliminate redundant parameters at both the channel and group levels, reducing model complexity and memory usage.

(6) QAT converts the model to INT8 precision, accelerating inference without sacrificing accuracy.

(7) The integration of Dice Loss and Boundary Loss improves boundary precision and enhances the detection of small lesions.

These contributions tackle the issue of segmentation of brain tumors in a practical way which is crucial for clinical use.

The rest of the article is organized as follows. Section 3 covers the methodology, including data preprocessing and the proposed architecture, which is detailed with an overview of the framework and its key components, such as pseudo-3D convolutions, multi-fiber designs, and DMF units. Section 4 presents the experimental setup, including the dataset, training settings, and evaluation metrics, followed by results and analysis covering ablation studies, quantitative results, and qualitative analysis. In Sec. 5 the key findings and future directions are concluded, followed by acknowledgments.

3. Method

3.1 Data Preprocessing

Our preprocessing pipeline for brain tumor segmentation was designed to ensure consistency and standardization across MRI data. The initial scans underwent a series of preprocessing techniques to get them ready for segmentation. The first step was the application of N4ITK bias field correction, which was used to reduce intensity inhomogeneities which stemmed from the magnetic field fluctuations during the image capturing process.

Let I represent the original voxel intensity and B the estimated bias field. The corrected intensity, denoted as I_c, was computed

$$I_c(x,y,z) = \frac{I(x,y,z)}{B(x,y,z)}. \tag{1}$$

The various stages of our preprocessing pipeline are shown in Fig. 1. Specifically, (a) shows MRI modalities after applying N4ITK bias field correction, (b) depicts the original and cropped MRI slices along with their segmentation masks, and (c) visualizes the normalization process with voxel intensity mapping.

The original MRI scans consist of 155 × 240 × 240 voxels, representing 155 slices per scan with each slice measuring 240 × 240 pixels. Regardless of these previously stated issues, the intensity values of the MRI scan are often inconsistent on account of the differing acquisition settings. The developers of the BraTS datasets have carried out extensive pre-processing operations which include: interpolating to a standardized resolution of 1 mm^3, putting through co-registration to an anatomy template, and skull stripping to get rid of the areas that are not of the brain.

To ensure compatibility with the deep learning model, all the imaging modalities and their corresponding segmentation masks were resampled to a uniform spatial resolution of 128 × 128 × 128 voxels. This process was done using image modality nearest interpolation for image modalities and trilinear interpolation for segmentation masks. Let I_r represent the resampled intensity, and f_x, f_y, f_z denote the coordinate mapping functions. The resampling process can be expressed as

$$I_r(x',y',z') = I\left(f_x(x'), f_y(y'), f_z(z')\right) \tag{2}$$

where (x', y', z') are the new coordinates in the resampled space.

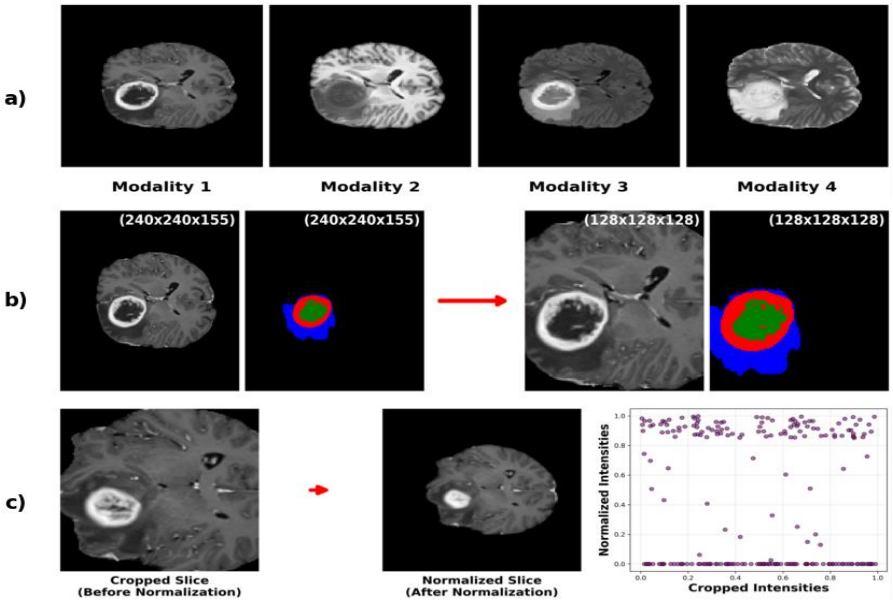

Fig. 1. MRI preprocessing steps including (a) MRI modalities after N4ITK bias correction; (b) Original vs. cropped MRI and segmentation masks, and (c) MRI slices with intensity mapping before and after normalization.

To reduce variability across datasets, intensity normalization was applied by scaling voxel intensities to the range [0,1] as suggested by Patro et al. [25]. Let I_n denote the normalized intensity, and I_{min} and I_{max} represent the minimum and maximum intensity values respectively within the modality and is calculated by

$$I_n = \frac{I_r - I_{min}}{I_{max} - I_{min}} \quad (3)$$

which ensuring that all intensities are normalized for consistent input to the model. Finally, the normalized modalities were combined into a single 4D array, denoted as A, with dimensions $C \times H \times W \times D$ where C is the number of channels, and H, W, D are the spatial dimensions. The preprocessed data, including A and the resized segmentation mask, was saved in HDF5 format with GZIP compression for efficient storage and loading.

3.2 Proposed Architecture Overview

3.2.1 Overview of the Framework

Figure 2 shows the overview of the proposed brain tumor segmentation framework, which integrates pseudo-3D convolutions, attention mechanisms, pruning strategies, QAT, and advanced loss functions to achieve high efficiency and accuracy. The process begins with the input stage, which accepts 4-channel 3D medical images of dimensions $4 \times 128 \times 128 \times 128$. This input is passed to the initial pseudo-3D convolution layer. The initial pseudo-3D convolution layer replaces traditional 3D convolutions with efficient pseudo-3D convolutions, combining spatial $3 \times 3 \times 1$ and depthwise $1 \times 1 \times 3$ convolutions, reducing the output dimensions to $32 \times 64 \times 64 \times 64$ with a stride of 2.

The encoder stage comprises two sets of DMF units, ECA modules and pruning for computational optimization. The first set of DMF units reduces the dimensions to $128 \times 32 \times 32 \times 32$, where pruning reduces the number of groups from $g = 16$ to $g = 8$. The second set of DMF units further reduces the dimensions to $256 \times 16 \times 16 \times 16$, with additional pruning applied. Each DMF unit utilizes pseudo-3D convolutions, and ECA modules enhance feature weighting. The encoder stage is followed by the bottleneck stage, represented as a DMF unit with adaptive dilation, dynamically adjusting dilation rates to capture multi-scale features. The bottleneck module significantly reduces the tensor dimensions from $384 \times 8 \times 8 \times 8$ to $256 \times 8 \times 8 \times 8$.

The decoder stage reconstructs the segmentation map by progressively upsampling the feature maps while integrating skip connections from the encoder. The first stage in the decoder is a MF unit with cross-attention, outputting dimensions of $256 \times 8 \times 8 \times 8$. This is followed by three upsampling blocks, each doubling the spatial resolution. The first block outputs dimensions of $256 \times 16 \times 16 \times 16$ with a skip connection from the second set of DMF units. The second block outputs dimensions of $256 \times 32 \times 32 \times 32$ with a skip connection from the first set of DMF units. The final block outputs dimensions of $128 \times 64 \times 64 \times 64$ with a skip connection from the initial pseudo-3D convolution layer.

The final pseudo-3D convolution layer reduces the dimensions to $32 \times 128 \times 128 \times 128$ using a $1 \times 1 \times 1$ convolution. This is followed by the output stage, which includes an upsampling block and a softmax layer, both with dimensions $4 \times 128 \times 128 \times 128$. The final output represents segmented tumor regions, categorized into ET, TC, and WT classes.

Apart from these steps, QAT is used to convert the model to INT8 precision which improves inference speed while maintaining optimal accuracy. Combination of Dice Loss and Boundary Loss is used to improve boundary precision and enhance the detection of small lesions. Fixed dilation rates within the bottleneck are replaced by adaptive weighting which allows the decoder to select relevant features for improved multi-scale feature representation. Additionally, the decoder replaces trilinear interpolation with sub-pixel convolution, which enhances segmentation map detail.

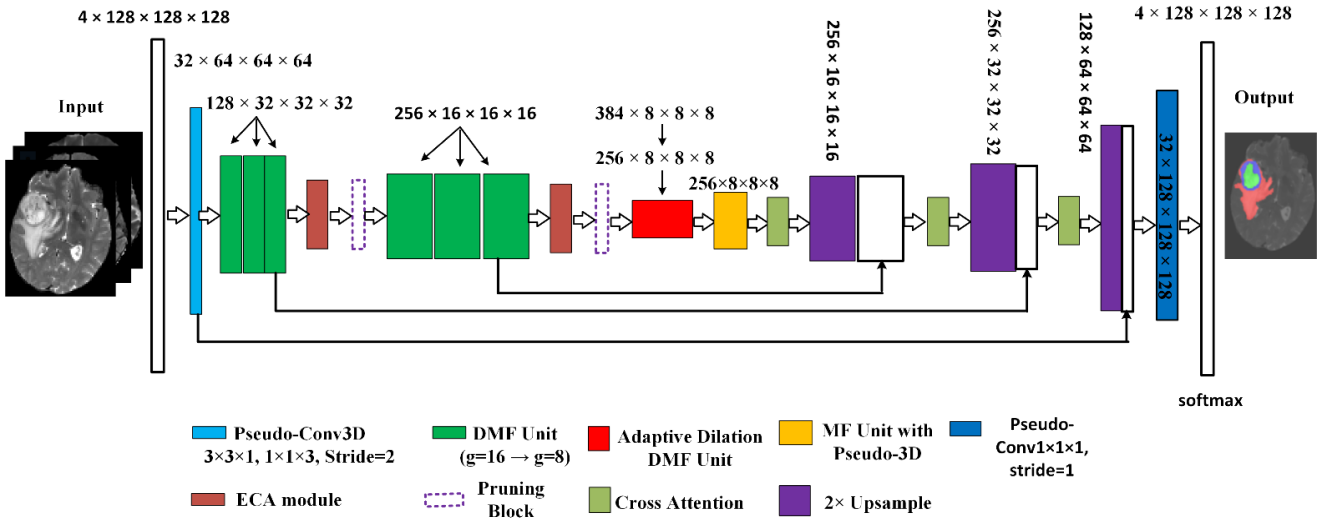

Fig. 2. Proposed brain tumor image segmentation architecture.

In this architecture pseudo-3D convolutions, lightweight attention mechanisms, pruning, QAT, and refined loss functions are integrated in order to boost segmentation efficiency as well as accuracy.

3.2.2 Key Components of the Proposed Architecture

(1) Residual unit with pseudo-3D convolutions

The residual unit with pseudo-3D convolutions, as illustrated in Fig. 3(a), is designed to enhance computational efficiency while preserving feature integrity. The unit begins with an input channel C_{in}, which undergoes a spatial convolution operation $3 \times 3 \times 1$, producing an intermediate output channel C_{mid}. This is followed by a depthwise convolution operation $1 \times 1 \times 3$, resulting in the final output channel C_{out}. A residual shortcut connection directly links the input C_{in} to the output C_{out}, combining both pathways through an addition operation. This structure allows the gradient and features to flow efficiently, and since we prefer pseudo-3D convolutions over 3D convolutions, the computational expenses are kept minimal.

(2) Multi-fiber design with grouped residual units

As shown in Fig. 3(b), three parallel fibers are incorporated, each representing a residual unit enhanced by pseudo-3D convolutions. In each fiber, the input channels are divided by the number of fibers, g, for computational efficiency. A spatial convolution $3 \times 3 \times 1$ processes the input C_{in}/g, producing intermediate channels C_{mid}/g which is followed by a depthwise convolution $1 \times 1 \times 3$, resulting in output channels C_{out}/g. A residual shortcut connects the input C_{in}/g directly to the output C_{out}/g in each fiber, ensuring feature preservation. As a result of allocating the channels over fibers, the overall parameter count comes down by g which optimizes the computational cost without diminishing the output performance.

(3) MF Unit

The MF unit with enhanced residual and adaptive fusion in Fig. 3(c) describes an architecture that is capable of collecting features from multiple parallel fibers and incorporating them into a single output in such a way that adaptability is achieved through weighted summation. Every fiber processes the input data independent of each other which is powered by a pseudo 3D spatial convolution $3 \times 3 \times 1$ and also a depthwise $1 \times 1 \times 3$ convolving. All the fibers' outputs are collected and summed together by the weighted summation block dynamically and in real time.

In the weighted summation block, learnable scalar weights w_1, w_2, and w_3 are applied to the outputs of the fibers F_1, F_2, and F_3 enabling the model to adjust the contribution of each fiber based on the task requirements. Mathematically, the output of the block is represented by

$$F_o = w_1 \cdot F_1 + w_2 \cdot F_2 + w_3 \cdot F_3 \tag{4}$$

where F_i denotes the feature map from the i-th fiber. This adaptive fusion mechanism provides effective feature aggregation without incurring remarkably high computational costs, thereby augmenting the unit's efficiency in extracting intricate spatial and depthwise features.

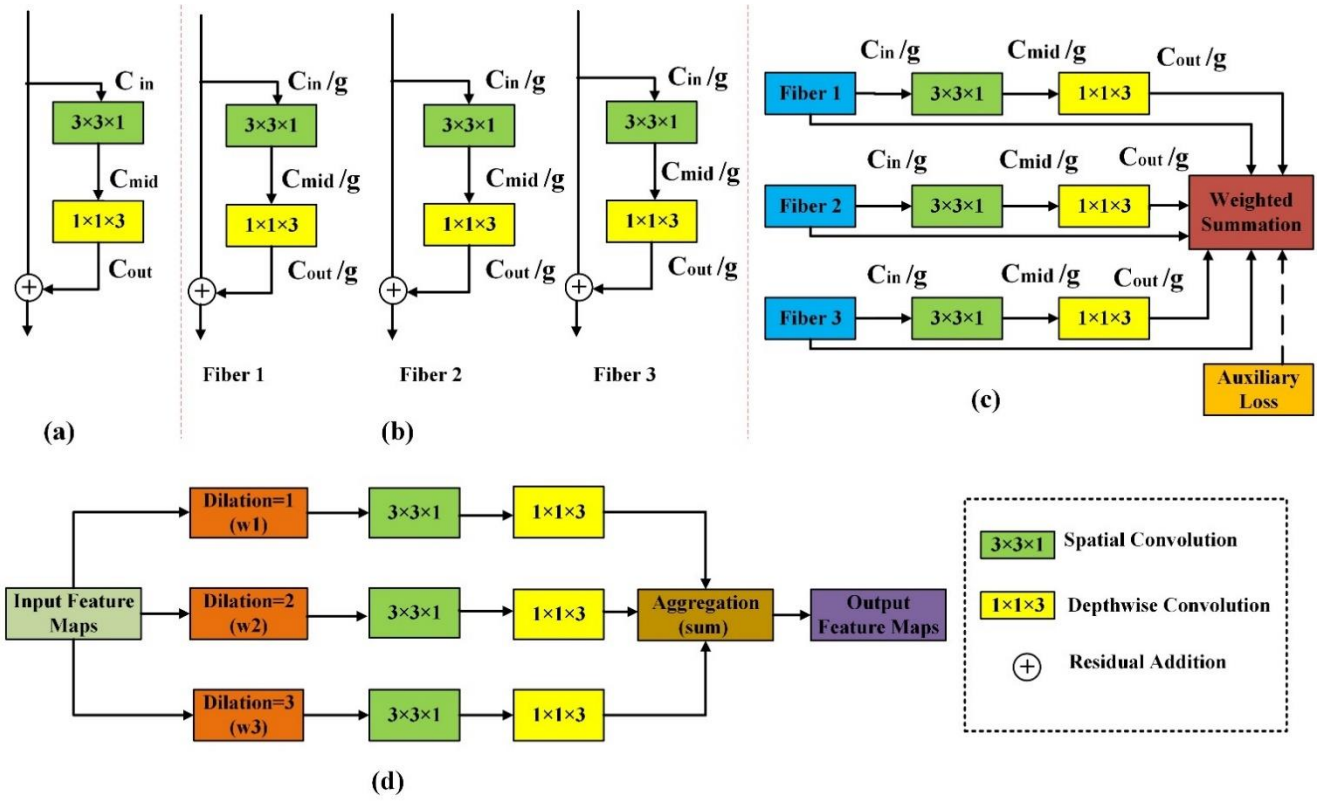

Fig. 3. Built blocks in the proposed architecture including: (a) Residual unit with pseudo-3D convolutions; (b) Multi-fiber design with grouped pseudo-3D residual units; (c) MF unit with adaptive fusion, and (d) DMF unit with adaptive weighting.

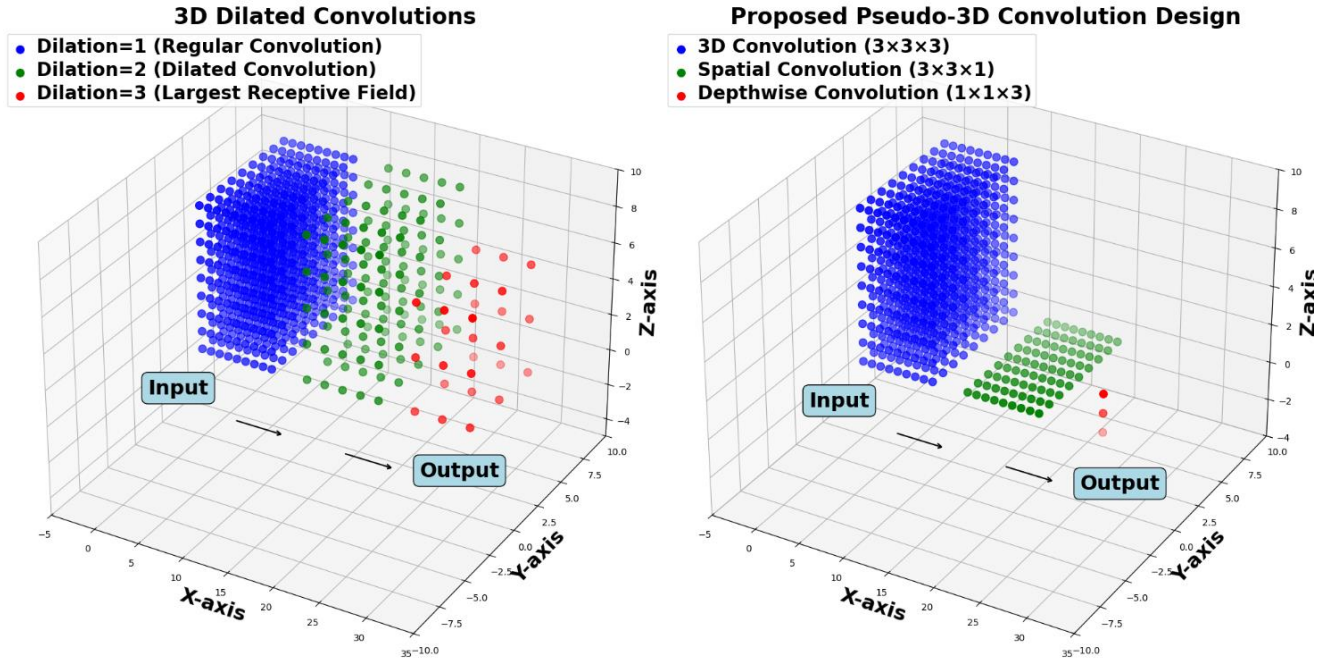

Fig. 4. Comparison of 3D dilated and pseudo-3D convolution in the proposed pipeline.

Furthermore, the MF unit has an auxiliary loss branch that emanates from the output of the weighted summation block. This branch facilitates exit loss and provides gradient intermediate supervision during training, thus ameliorate convergence.

(4) DMF unit

The DMF unit with adaptive weighting, as shown in Fig. 3(d), captures multi-scale features through parallel branches with distinct dilation rates $d = 1$, $d = 2$, and $d = 3$. Each branch processes the input using pseudo-3D convolutions, consisting of a spatial convolution $3 \times 3 \times 1$ followed by a depthwise convolution $1 \times 1 \times 3$, and applies adaptive weights w_1, w_2, and w_3 to adjust the contribution of each branch. The outputs of the branches are aggregated at a sum node, producing refined multi-scale features directed to the output feature maps. This design efficiently combines multi-scale information while reducing computational overhead.

(5) Schematic of 3D dilated and pseudo-3D convolution operations

To better understand the architectural improvements in our model, it is important to visualize how convolution operations differ in structure and computation. The differences between conventional 3D dilated convolutions and the proposed pseudo-3D convolution design are illustrated in Fig. 4.

It can be seen from Fig. 4 that the left panel illustrates 3D dilated convolutions with varying dilation rates $d = 1$, $d = 2$, and $d = 3$, which increase the receptive field to capture multi-scale contextual information. Every dilation factor corresponds to grids whose points are arranged proportionate to the dilation factor. The grids intensify the phenomenon that higher dilation rates lead to coarser sampling and broader spatial coverage meaning that dilated convolutions can leverage both local textures and global information.

The right panel of Fig. 4 demonstrates the decomposition of a standard 3D convolution $3 \times 3 \times 3$ into two sequential operations, specifically a spatial convolution $3 \times 3 \times 1$ followed by a depthwise convolution $1 \times 1 \times 3$. This design of pseudo-3D convolution avoids the high computational cost by splitting the expensive 3D operation into less expensive 2D and 1D operations. The grids show how spatial features are first processed in a plane, and then the volumetric dimension is processed depthwise. Such designing achieves the application of these 3D convolutional neural networks without incurring the high cost of FLOPs, and is therefore useful for volumetric medical imaging tasks.

4. Experimental Results and Analysis

4.1 Experimental Setup

To check how well the algorithm works for the proposed brain tumor segmentation task, a set of experiments were set up to compare the results obtained with different algorithms developed for the task. The experiments were performed on a high-performance system featuring an Intel Core i9-14900K processor with 24 cores running at a frequency of 6.0 GHz, an Nvidia GeForce RTX 3090 GPU with 24 GB VRAM, 1 TB HDD, 500 GB SSD, and 48 GB RAM. The software platform utilized included Python 3.9, CUDA 11.8, and PyTorch 2.0.0, alongside other relevant Python function libraries. Furthermore, this architecture was built and tested with Keras running on TensorFlow 2.15 for model building purposes.

4.2 Dataset and Training Settings

In this study, the proposed model is trained and validated using the Brain Tumor Segmentation (BraTS) benchmark datasets, including BraTS 2020 [26] and BraTS 2021 [27], which are widely used in the medical image analysis community for evaluating brain tumor segmentation algorithms. BraTS 2021, an extension of the BraTS 2020 dataset, comprises a total of 1251 patient cases covering both high-grade gliomas (HGG) and low-grade gliomas (LGG). In comparison, BraTS 2020 includes MRI data from 369 patients, with 76 cases diagnosed as LGG and the remaining cases classified as HGG. An example of multi-modal MRI scans with expert-labeled ground truth segmentation is shown in Fig. 5(a).

We can see from Fig. 5(a) that the multi-modal MRI scans provide complementary information for tumor characterization, with each modality highlighting different tissue properties. Both datasets consist of 3D MRI scans, each comprising 155 slices with an original resolution of 240×240 pixels per slice. The scans include four imaging modalities: T2, T1, T1ce, and FLAIR, each offering distinct diagnostic insights. T1-weighted images capture anatomical structures, differentiating gray and white matter. T2-weighted images emphasize areas with high water content, aiding in the visualization of edema. T1ce images, enhanced with contrast agents, highlight blood vessels and regions of active tumor growth. FLAIR images suppress cerebrospinal fluid signals, allowing better identification of subtle lesions and anomalies linked to tumor expansion.

Ground truth segmentation masks, annotated by one to four expert neuroradiologists, include four primary classes: background (BG, Label 0), necrotic and non-enhancing tumor (NCR/NET, Label 1), edema (ED, Label 2), and enhancing tumor (ET, Label 4). Figure 5(b) shows an example of a typical MRI scan including labeled ground truth segmentation masks, where green represents necrotic and non-enhancing tumor regions (NCR/NET, Label 1), red highlights edema regions (ED, Label 2), and yellow denotes enhancing tumor areas (ET, Label 4). These classes are commonly grouped into three main tumor regions for segmentation including: the whole tumor (WT), which encompasses NCR/NET, ED, and ET (Labels 1, 2, 4), the tumor core (TC), which includes NCR/NET and ET (Labels 1, 4), and the enhancing tumor (ET, Label 4).

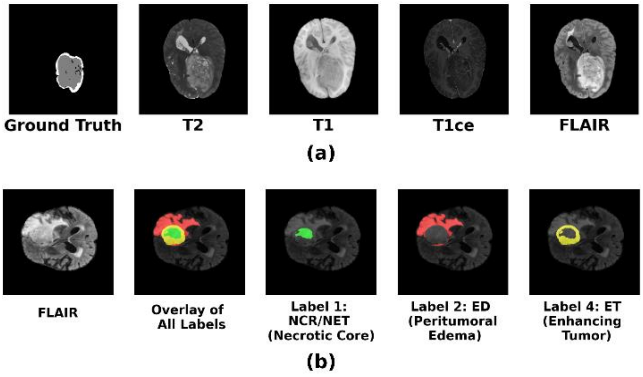

Fig. 5. Visual examples of the dataset including: (a) Multi-modal MRI scans and expert-labeled ground truth for a patient. (b) Ground truth segmentation on an MRI scan: green denotes necrotic and non-enhancing tumor regions (NCR/NET, label 1), red represents edema (ED, label 2), and yellow highlights enhancing tumor regions (ET, label 4).

Since there was no official validation or test data available for BraTS 2020 and 2021, the proposed model was validated with five-fold cross validation. The data was further split in $4:1$ ratio for training and validation, thus ensuring a reliable evaluation. Training was conducted using the Adam optimizer with an initial learning rate set to 1×10^{-4}. Due to GPU memory limitation, the model was trained with a batch size of 2 and gradient accumulation was implemented to mimic a higher batch size. This kept computational burden under control while not playing a detriment to model performance. For achieving better model generalization with respect to the training dataset and preventing overfitting, L_2 regularization was employed to the parameters of the convolutional kernels with a magnitude of 0.02. This factor was chosen after thorough testing over the selection of the model.

A total of 400 epochs were used to train the model for this task, while segmentation performance was optimized with a hybrid loss function of Dice Loss and Boundary Loss. To improve inference speed, the model was first converted to INT8 precision QAT. Collapsible structured pruning was then performed to remove unnecessary parameters in the model and improve performance. Along with pseudo-3D convolutions and ECA modules, these alterations ensured great segmentation performance of the model without sacrificing speed.

Data augmentation techniques were applied to enhance model robustness, such as the random flip along specified spatial axes, random shifts in the depth, height and width, and insertion of Gaussian noise with a standard deviation of 0.01. These augmentations assisted with accurate segmentation by achieving the desired degree of detail preservation while improving the model performance in unseen data.

4.3 Evaluation Metrics

The proposed segmentation algorithm is used in segmenting the ET, TC, and WT regions from multimodal MRI images of the patients organized hierarchically with TC containing ET and WT containing TC. To assess the performance of the model for these regions, we calculated several metrics, such as Dice coefficient, sensitivity, specificity, 95% Hausdorff distance (HD95), Giga Floating-Point Operations (GFLOPs), and inference time. Each measure examines an angle of the model's capability which makes the evaluation framework more comprehensive.

The Dice coefficient is the primary metric used to assess the overlap between the predicted P and ground truth T tumor regions and is calculated by

$$\text{Dice} = \frac{2 \cdot |P \cap T|}{|P| + |T|} \tag{5}$$

where $|P \cap T|$ denotes the true positive pixels, and $|P|$ and $|T|$ are the total predicted and actual pixels. A higher Dice score indicates greater accuracy in segmentation.

Since the segmentation performance is evaluated separately for each tumor subregion including ET, TC, and WT the task can be considered a set of binary segmentation problems. In this context, the Dice coefficient used in our evaluation is mathematically equivalent to the F_1-score, commonly used in classification tasks. The F_1-score is given by

$$F_1\text{-score} = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{6}$$

Thus, the reported Dice scores for ET, TC, and WT also represent the corresponding F_1-score for these tumor subregions.

Sensitivity, also referred to as the recall rate, measures the model's ability to detect tumor regions and is defined by

$$\text{Sensitivity} = \frac{|P \cap T|}{|T|}. \tag{7}$$

High sensitivity indicates effective tumor detection. Specificity measures the ability of the model to correctly identify non-tumor regions while minimizing false positives and is given by

$$\text{Specificity} = \frac{|P^{\mathrm{C}} \cap T^{\mathrm{C}}|}{|T^{\mathrm{C}}|} \tag{8}$$

where $|P^{\mathrm{C}} \cap T^{\mathrm{C}}|$ denotes the true negatives, and $|T^{\mathrm{C}}|$ represents the total number of actual background pixels. Higher specificity indicates fewer false alarms in segmentation. The HD95 at the 95th percentile evaluates the boundary alignment between the predicted P and ground truth T tumor regions and is calculated by

$$\mathrm{HD95}(A,B) = \max\left(\sup_{a \in A} \inf_{b \in B} d(a,b), \sup_{b \in B} \inf_{a \in A} d(a,b) \right) \tag{9}$$

where A and B denote the surfaces of T and P, respectively, a and b are points on their surfaces, and d (a, b) calculates the Euclidean distance between points. GFLOPs assess the computational complexity of the model by calculating the number of floating-point operations required for a single input and is given by

$$\text{GFLOPs} = \sum_{i=1}^{L} C_i \cdot H_i \cdot W_i \cdot K_i^2 \tag{10}$$

where L is the number of layers, C_i is the number of input channels, H_i and W_i are the spatial dimensions respectively, and K_i is the kernel size. Lower GFLOPs indicate a more efficient model with reduced computational overhead.

Furthermore, the model's speed was determined by the time needed to segment a single 3D volumetric input. It is measured in seconds and is an important criterion to judge the potential of the model for real-time clinical use.

4.4 Ablation Studies

The findings of the ablation study displayed in Tab. 1 further support the effectiveness of the proposed methodology illustrating the impact of each element in the system under consideration with respect to the whole system effectiveness on the BraTS 2021 benchmark dataset. Beginning from the baseline of 3D convolutions, pseudo-3D convolutions greatly restrain FLOPs and parameters, enhancing efficiency and accuracy by simplifying operations on grouped residuals units further subdivide the channels and so optimize the computations while MF units increase the parameters slightly because of the adaptive fusion.

Adaptive DMF units improve the multi-scale representation for an elemental increase in FLOPs and the ECA modules weight the features more effectively on computation cost. Structured pruning saves a lot of redundant parameters and improved efficiency through ECA modules, and further speeds the inference to 0.016 s per volume with QAT. The combination of Dice + Boundary Loss improves the training segmentation parameter accuracy while relieving constraints on parameters or FLOPs optimization. Emphasizing these features shows that this framework is clinically applicable due to its sufficiency in accuracy, speed and efficiency.

4.5 Quantitative Results

To evaluate the effectiveness of the proposed model, we present a detailed comparison of its performance with the existing segmentation models for brain tumors in BraTS 2021 and BraTS 2020. This is presented in Tab. 2 and Tab. 3, with emphasis on metrics such as Dice and Hausdorff95 distances.

Method	Mean Dice (%)			Mean HD95(mm)			Params (M)	FLOPs	Inference Time
	WT	TC	ET	WT	TC	ET			
3D Convolutions	86.12	81.42	78.73	8.89	10.54	12.23	5.86	42.31G	0.034s
+ Pseudo-3D	87.23	82.87	78.65	8.21	11.98	10.67	3.52	32.24	0.027s
+ Grouped Residual	86.78	83.23	79.12	9.92	10.34	11.12	2.99	25.79	0.023s
+ MF Units	87.27	83.82	81.85	9.35	10.76	12.75	3.29	28.55	0.025s
+ Adaptive DMF Units	88.62	84.02	82.10	7.10	9.34	9.12	3.78	29.74	0.026s
+ ECA Modules	89.76	86.32	82.34	7.82	8.87	8.16	3.97	28.35	0.022s
+ Structured Pruning	89.62	88.02	84.45	6.65	7.15	6.70	3.57	24.32	0.017s
+ QAT (INT8 Precision)	91.15	88.02	84.73	4.58	6.53	5.65	3.57	21.26	0.016s
+ Dice + Boundary Loss	91.85	88.52	85.55	2.58	3.53	3.65	3.57	21.26	0.016s

Tab. 1. Ablation study results of the proposed framework on BraTS 2021.

Model Name	Published Year	Dice (%)			HD95 (mm)		
		WT	TC	ET	WT	TC	ET
Cao et al. [28]	2024	91.00	89.70	**85.60**	6.15	9.98	11.23
Liu et al. [29]	2024	89.87	86.48	78.65	6.42	5.88	4.39
Jiang et al. [30]	2022	91.83	84.75	83.21	3.65	14.51	16.03
Zhou et al. [31]	2024	89.47	89.24	83.62	10.09	7.41	11.86
Hou et al. [32]	2023	**92.7**	88.70	85.40	3.510	5.771	13.983
Sun et al. [33]	2024	85.3	86.10	78.10	2.66	**1.51**	**2.82**
Al-Fakih et al. [34]	2024	86.6	**89.90**	85.50	22.70	9.78	9.60
Liu et al. [35]	2022	91.6	86.80	83.30	5.945	7.567	19.27
Proposed	------	91.85	88.52	**85.57**	**2.583**	3.537	3.657

Tab. 2. Comparison of the proposed model with state-of-the-art methods on BraTS 2021 data validation set.

Model Name	Published Year	Dice (%)			HD95 (mm)		
		WT	TC	ET	WT	TC	ET
Alwadee et al. [36]	2025	88.41	83.82	73.67	3.19	4.24	3.97
Magadza et al. [37]	2023	91.2	84.8	79.2	4.41	6.20	29.31
Isensee et al. [38]	2021	91.2	85.1	79.9	3.69	7.82	23.50
Yang et al. [39]	2023	**95.30**	**94.53**	**90.53**	2.20	**1.59**	**1.32**
Pan et al. [40]	2025	90.57	83.35	78.72	10.61	10.08	15.88
Diao et al. [41]	2024	91.10	86.34	79.41	5.20	5.85	3.30
Zhao et al. [42]	2023	92.00	84.00	75.00	**1.04**	2.88	3.19
Gao et al. [43]	2025	92.29	85.71	77.14	1.152	2.942	2.375
Proposed	------	90.63	87.16	84.31	2.9762	2.433	2.104

Tab. 3. Comparison of the proposed model with state-of-the-art methods on BraTS 2020 data validation set.

Model	Dice (%)			HD95 (mm)		
	WT	TC	ET	WT	TC	ET
Fold1	90.12	85.33	83.74	3.683	4.117	4.026
Fold2	90.52	84.96	83.24	3.174	5.332	4.167
Fold3	89.78	85.12	84.63	5.086	5.231	5.675
Fold4	91.06	86.32	84.08	4.342	**3.473**	3.752
Fold5	90.77	86.92	85.12	3.657	3.778	3.769
Ensemble	**91.85**	**88.52**	**85.55**	**2.583**	3.537	**3.657**

Tab. 4. Five-fold cross-validation results on BraTS 2021 benchmark dataset.

Model	Dice (%)			HD95 (mm)		
	WT	TC	ET	WT	TC	ET
Fold1	89.84	85.54	83.71	3.675	4.086	4.021
Fold2	90.25	84.92	83.96	4.132	3.122	3.436
Fold3	90.02	84.72	82.72	3.074	2.765	2.331
Fold4	90.14	85.77	84.12	4.772	3.564	4.164
Fold5	89.97	84.87	83.24	3.121	3.012	4.232
Ensemble	**90.63**	**87.16**	**84.31**	**2.976**	**2.433**	**2.104**

Tab. 5. Five-fold cross-validation results on BraTS 2020 benchmark dataset.

It can be seen from Tab. 2 and Tab. 3 that the proposed model demonstrates strong performance on both BraTS 2021 and BraTS 2020 validation sets, achieving high Dice scores and superior HD95 values across all tumor subregions. On BraTS 2021, it outperforms methods such as Cao et al. [28], Liu et al. [29], and Zhou et al. [31], maintaining a balanced segmentation accuracy with significantly lower HD95 values, indicating superior boundary precision. While Hou et al. [32] reported a slightly higher Dice score for WT, the proposed model achieved the lowest HD95 values, demonstrating improved boundary delineation. Similarly, on BraTS 2020, the model outperforms Magadza et al. [37], Pan et al. [40], and Gao et al. [43], achieving superior HD95 values, which signifies more accurate tumor boundary segmentation. Although Yang et al. [39] reported the highest Dice scores, the proposed model effectively balances segmentation accuracy, computational efficiency, and boundary precision.

The superior performance of the proposed model can be attributed to several architectural innovations. For instance, the switching out of traditional 3D convolutions for

pseudo 3D convolutions which drastically decreases computational costs while maintaining volumetric context. The multiscale representation and segmentation precision is further enhanced by the adaptive DMF unit. Feature aggregation is done in parallel with MF units which increases efficiency while also maintaining a high degree of flexibility. The model is able to concentrate on more relevant areas of the tumor while expending far less energy due to the ECA feature weighting method. Lastly, the combination of Dice Loss and Boundary Loss widens the multi scale representation and performs exceedingly well by increasing the boundary accuracy and lesion detection which explains the models advanced HD95 scores. Lastly, Tables 4 and 5 show the five-fold cross validation results which demonstrate the efficiency of the model on the BraTS 2021 and 2020.

The five-fold cross-validation results presented in Tab. 4 and Tab. 5 demonstrate the consistent and reliable performance of the proposed model on the BraTS 2021 and BraTS 2020 datasets. The ensemble results consistently outperformed the individual folds, with Dice scores higher than the average of the five folds and HD95 values lower than the average. For instance, on BraTS 2021, the ensemble Dice score for WT was 91.85% compared to an average of 90.45% across the five folds, while the HD95 value for WT was reduced to 2.583 mm from an average of 3.988 mm. Similarly, on BraTS 2020, the ensemble results showed a Dice score improvement and reduced HD95 values, further confirming the robustness of the proposed approach. Ensemble models always perform better due to the pooling of various fold predictions where the strengths are complementary so as to minimize the weaknesses of folds.

The ensemble procedure was done by first computing the softmax probability maps for the separate models and then picking the maximum probability class for every voxel. In segmentation, noisy and conflicting predictions are prevalent. This form of integration actively enhances spatial contiguity and reduces overlapping classifications. The ensemble technique reduces the chances of overfitting by blending in predictions from different training subsets, thus enhancing performance.

The efficiency and accuracy of the framework has been further corroborated as seen in Tabs. 6 and 7. An efficiency comparison of other models using parameters such as FLOPs, parameter number and inference time is done in Tab. 8, showing the efficiency of the suggested model relative to segmentation accuracy.

The proposed model achieves significant improvements over existing methods on the BraTS 2021 and BraTS 2020 validation datasets, which can be seen in Tabs. 6 and 7. On BraTS 2021, it achieved the highest Mean Dice scores and superior boundary precision, with 91.85% for WT and 88.52% for TC, along with lower HD95 values, such as 2.58 mm for WT. The model was able to sustain best scores like those of the Swin UNETR and TransBTS models for the methods like segmentation and boundary precision. In the same manner, it obtained strong segmentation performance along with excellent boundary precision on BraTS 2020, with HD95 values of 3.37 mm for WT and 5.98 mm for TC, indicating its ability to seamlessly shift between segmentation and boundary separation metrics across datasets.

It can be seen from Tab. 8 that the superior performance of the provided model emanates from its new architecture. Pseudo-3D convolutions reduce computational overhead while retaining volumetric context for efficient

Model	Mean Dice (%)			Mean Sensitivity (%)			Mean Specificity (%)			Mean HD95 (mm)		
	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
3D U-Net [6]	88.53	86.73	82.23	90.23	86.17	82.05	99.85	99.72	99.67	6.23	9.73	9.67
TransBTS [44]	88.75	85.51	83.42	90.31	84.17	82.33	99.87	99.68	99.85	5.73	9.82	5.44
TransUNet [45]	89.23	83.31	81.54	92.66	85.73	84.76	99.82	99.77	99.62	6.10	7.43	9.14
Swin UNETR [46]	91.05	88.11	84.63	94.32	89.56	87.45	99.87	**99.82**	99.72	6.42	7.17	9.76
Proposed	**91.85**	**88.52**	**85.55**	**94.88**	**90.76**	**89.56**	**99.95**	99.72	**99.97**	**2.58**	**3.53**	**3.65**

Tab. 6. Comparison of segmentation results of different models on BraTS 2021.

Model	Mean Dice (%)			Mean Sensitivity (%)			Mean Specificity (%)			Mean HD95 (mm)		
	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
3D U-Net [6]	87.76	83.22	80.21	90.21	85.42	81.77	99.81	99.72	99.85	5.17	9.33	9.45
TransBTS[44]	88.10	86.22	82.23	91.78	89.34	85.75	99.87	99.84	99.81	4.97	9.06	8.41
TransUNet[45]	88.66	84.54	80.22	92.77	89.32	82.86	99.92	99.86	99.81	6.67	8.13	9.42
Swin UNETR [46]	**91.10**	86.51	83.17	94.32	88.67	87.76	**99.95**	99.87	99.87	5.63	6.25	7.34
Proposed	90.63	**87.16**	**84.31**	**94.88**	**90.86**	**89.66**	99.94	**99.89**	**99.97**	**3.37**	**5.98**	**4.52**

Tab. 7. Comparison of segmentation results of different models on BraTS 2020.

Model	Parameters (M)	FLOPs (G)
3D U-Net [6]	16.21	1670
VcaNet [40]	79.32	1140.67
Swin UNETR [46]	61.98	394.84
TransBTS [44]	32.99	333
DResU-Net [47]	30.47	374.04
HNF-Netv2 [48]	17.91	449.79
DMFNet [19]	3.88	27.04
LATUP-Net [36]	**3.07**	**15.79**
Proposed	3.57	21.26

Tab. 8. Comparison of model efficiency in terms of FLOPs and parameters based on BraTS benchmark dataset.

feature extraction. Adaptive DMF units are capable of building multi-scale features and MF units perform further feature weighting by fusion which is done adaptively. For efficient resource spending, ECA modules concentrate on the prominent areas of feature weighting. Structured pruning and QAT result in a model which is super lightweight and optimized for boosted inference with INT8 precision while retaining accuracy. Dice and Boundary Loss functions increase segmentation accuracy with exact boundary delineation, and result in the model having 3.57M parameters, 21.26 GFLOPs, and an inference time of 0.016 seconds per 3D volume on an NVIDIA RTX 3090 GPU. This makes the model applicable for real-time clinical use.

It is worth mentioning that LATUP-Net [36] which has 3.07M parameters and 15.79 GFLOPS has higher inference latency at 212 milliseconds and renders lesser results on BraTS 2020. In the same breath, DMFNet [19] also has comparatively lower segmentation accuracy for ET and TC subregions on BraTS 2018, despite improving to 3.88M parameters and 27.04 GFLOPS with an inference latency of 0.019 seconds. On the contrary, the proposed model has enhanced segmentation accuracy on BraTS 2020. It shows that better balance between computational efficiency, inference speed, and accuracy can be achieved and maintained.

4.6 Qualitative Analysis

To facilitate a more intuitive comparison of segmentation performance, we present the visualization results of the competing methods alongside our proposed model on the BraTS 2021 and 2020 datasets, as shown in Fig. 6, and the 3D segmentation results of our proposed model in Fig. 7.

The output segmentation images from the proposed model differ substantially from other models as depicted in Fig. 6. TransUNet displays minor estimation errors together with incorrect prediction areas in the final rows of Fig. 6. A probable cause for this situation is the inability of these models to use completely the combined information which exists across various modalities during brain tumor segmentation. The proposed model demonstrates superior segmentation outcomes which lead to excellent WT and TC segmentation results together with competitive ET segmentation results. The proposed improvements stem from two key elements that replace traditional 3D convolutions with pseudo-3D convolutions to reduce computational expenses and maintain volumetric context and the implementation of Dice Loss and Boundary Loss for enhancing precision of edges and small lesion detection.

In Fig. 7, the first two rows of 3D visualizations are from BraTS 2021 cases, including BraTS2021_00216, 00266, 00336, 00789, 00816, and 01324. The third row, for comparison, is taken from BraTS 2020 cases, specifically BraTS20_Training_047, 053, and 338. The results illustrate the model's ability to effectively segment WT, TC, and ET regions in 3D, leveraging volumetric context to enhance segmentation quality.

5. Conclusion and Future Directions

The research findings demonstrate that the designed framework delivers superior brain tumor segmentation results with high efficiency across the BraTS 2021 and 2020 datasets. Pseudo-3D convolutions, together with adaptive DMF units, MF units, ECA modules, and structured pruning, improve multi-scale representation, minimize computational load, and speed up inference to 0.016 seconds per 3D volume. The lightweight system architecture features 3.57M parameters and 21.26 GFLOPs while maintaining its performance level for segmentation operations. QAT serves to improve the model's performance by making it clinically applicable quickly while maintaining precision in segmentation results.

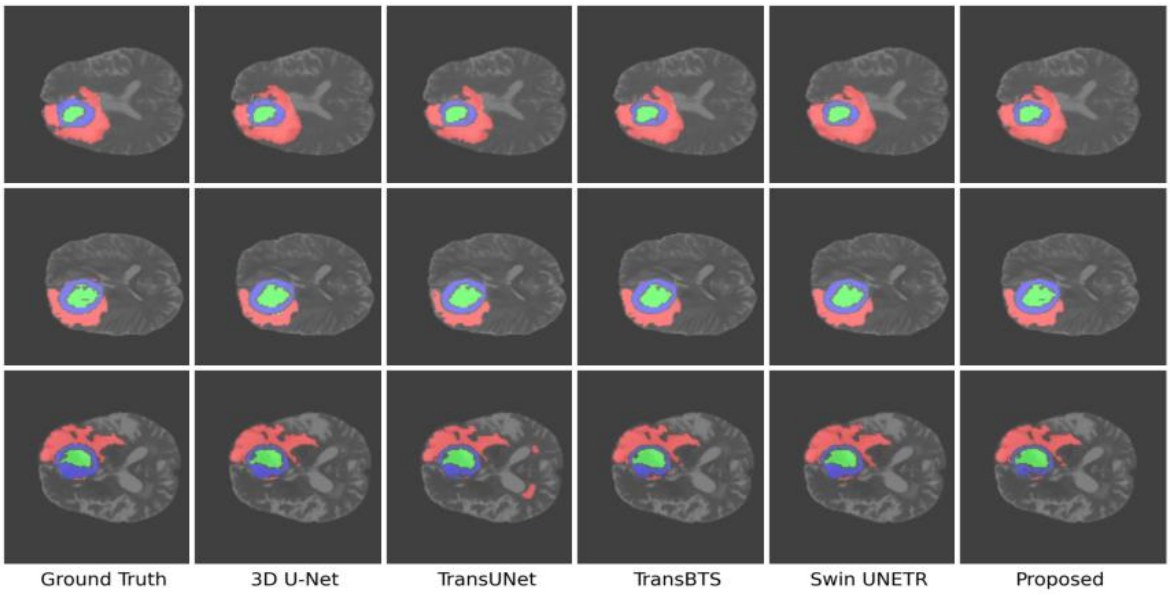

Fig. 6. Visualization and comparison of segmentation results from various methods on the BraTS 2021 dataset. Overlay colors: Red for WT, green for TC, blue for ET.

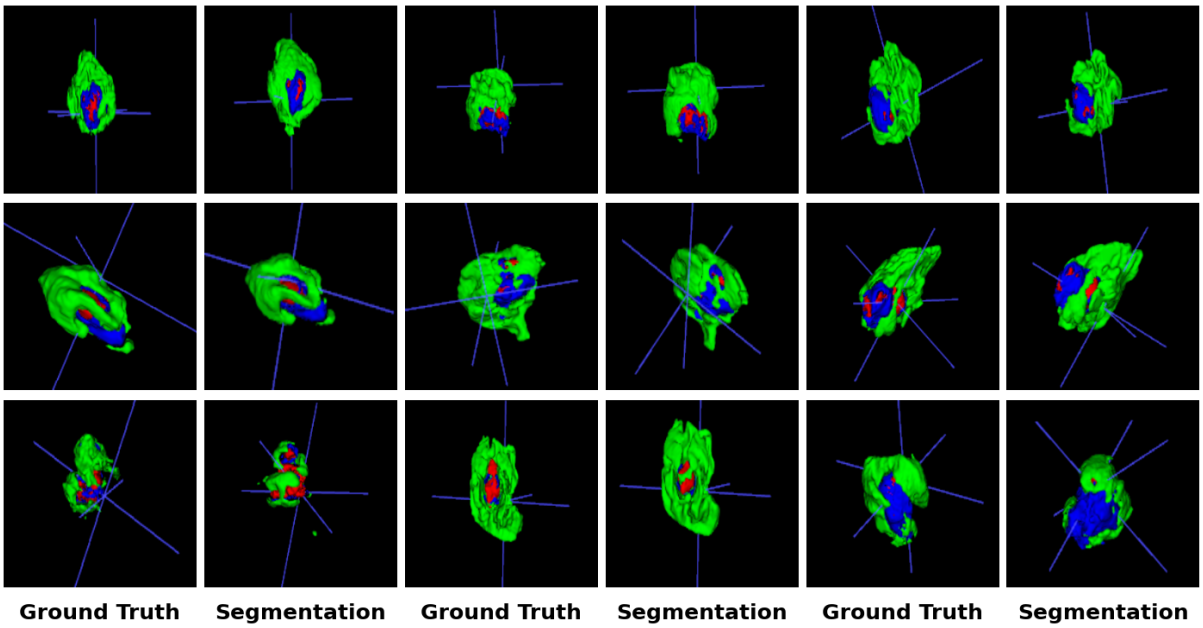

Fig. 7. 3D segmentation results of some samples from the BraTS 2021 and 2020 datasets. Overlay colors: red for WT, green for TC, blue for ET.

The BraTS domain-specific dataset restricts the evaluation of this framework across various medical images, while its evaluation across different types of cases needs additional research. The model maintains good accuracy levels while providing efficient performance; yet, its practical deployment requires more attention to MRI variation and scanner inconsistency issues. Upcoming research will solve the present constraints by performing diverse testing on wider datasets and implementing the approach in liver, knee, and cardiac imaging procedures. The functionality of generalization will be optimized by applying domain adaptation methods and self-supervised learning procedures to enhance robustness. The framework demonstrates high potential for medical use in preoperative planning, radiotherapy treatment, and automated diagnostic systems because of its efficient segmentation performance. To support reproducible research, the mathematical model presented in this work, including the architectural framework and relevant formulations, has been made publicly available at https://github.com/Rahman768/RMM.

Acknowledgments

This work is supported by Joint Project of Beijing Natural Science Foundation and Beijing Municipal Education Commission (No. KZ202110011015). The authors would like to thank CSC (China Scholarship Council) for funding to make this research possible lively and successful.

References

[1] ZAKERI, Y., KARASFI, B., JALALIAN, A. A review of brain tumor segmentation using MRIs from 2019 to 2023 (statistical information, key achievements, and limitations). *Journal of Medical and Biological Engineering*, 2024, vol. 44, p. 1–26. DOI: 10.1007/s40846-024-00860-0

[2] AGGARWAL, M., TIWARI, A. K., SARATHI, M., et al. An early detection and segmentation of brain tumor using deep neural network. BMC *Medical Informatics and Decision Making*, 2023, vol. 23, p. 1–12. DOI: 10.1186/s12911-023-02174-8

[3] MAAS, B., ZABEH, E., ARABSHAHI, S. QuickTumorNet: Fast automatic multi-class segmentation of brain tumors. In *Proceedings of the 2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*. Italy, May 2021, p. 81–85. DOI: 10.1109/NER49283.2021.9441286

[4] DE VERDIER, M. C., SALUJA, R., GAGNON, L., et al. The 2024 Brain Tumor Segmentation (BraTS) challenge: Glioma segmentation on post-treatment MRI. 10 pages. [Online] Cited 2024-05-28. Available at: https://arxiv.org/abs/2405.18368. DOI: 10.48550/arXiv.2405.18368

[5] BAKAS, S., AKBARI, H., SOTIRAS, A., et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 2017, vol. 4, no. 1, p. 1–13. DOI: 10.1038/sdata.2017.117

[6] ÇIÇEK, Ö., ABDULKADIR, A., LIENKAMP, S. S., et al. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Proceedings of the 19 th Medical Image Computing and Computer-Assisted Intervention (MICCAI 2016)*. Athens (Greece), 2016, Part II, p. 424–432. DOI: 10.1007/978-3-319-46723-8_49

[7] MILLETARI, F., NAVAB, N., AHMADI, S. A., et al. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*. Stanford (CA, USA), 2016, p. 565–571. DOI: 10.1109/3DV.2016.79

[8] CASAMITJANA, A., PUCH, S., ADURIZ, A., et al. 3D convolutional neural networks for brain tumor segmentation: A comparison of multi-resolution architectures. In *Proceedings of the Second International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Athens (Greece), 2016, p. 150–161. DOI: 10.1007/978-3-319-55524-9_15

[9] FENG, X., TUSTISON, N. J., PATEL, S. H., et al. Brain tumor segmentation using an ensemble of 3D U-Nets and overall survival prediction using radiomic features. *Frontiers in Computational*

Neuroscience, 2020, vol. 14, p. 1–12. DOI: 10.3389/fncom.2020.00025

[10] LIU, D., SHENG, N., HAN, Y., et al. SCAU-net: 3D self-calibrated attention U-Net for brain tumor segmentation. *Neural Computing and Applications*, 2023, vol. 35, no. 33, p. 23973–23985. DOI: 10.1007/s00521-023-08872-8

[11] GAMAL, A., BEDDA, K., ASHRAF, N., et al. Brain tumor segmentation using 3D U-Net with hyperparameter optimization. In *Proceedings of the 2021 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*. Giza (Egypt), 2021, p. 269–272. DOI: 10.1109/NILES53778.2021.9600556

[12] LIU, L., XIA, K. BTIS-Net: Efficient 3D U-Net for brain tumor image segmentation. *IEEE Access*, 2024, vol. 12, p. 133392 to 133405. DOI: 10.1109/ACCESS.2024.3460797

[13] NUECHTERLEIN, N., MEHTA, S. 3D-ESPNet with pyramidal refinement for volumetric brain tumor image segmentation. In *Proceedings of Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018. Held in Conjunction with MICCAI 2018*. Granada (Spain), 2018, Part II, p. 245–253. DOI: 10.1007/978-3-030-11726-9_22

[14] CHEN, W., LIU, B., PENG, S., et al. S3D-UNet: Separable 3D U-Net for brain tumor segmentation. In *Proceedings of Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018. Held in Conjunction with MICCAI 2018*. Granada (Spain), 2018, Part II, p. 358–368. DOI: 10.1007/978-3-030-11726-9_32

[15] ALI, S., KHURRAM, R., REHMAN, K. U., et al. An improved 3D U-Net-based deep learning system for brain tumor segmentation using multi-modal MRI. *Multimed Tools and Applications*, 2024, vol. 83, p. 85027–85046. DOI: 10.1007/s11042-024-19406-2

[16] AHMED, S. F., RAHMAN, F. S., TABASSUM, T., et al. 3D U-Net: Fully convolutional neural network for automatic brain tumor segmentation. In *Proceedings of the 2019 22nd International Conference on Computer and Information Technology (ICCIT)*. Dhaka (Bangladesh), 2019, p. 1–6. DOI: 10.1109/ICCIT48885.2019.9038237

[17] WU, Q., PEI, Y., CHENG, Z., et al. SDS-Net: A lightweight 3D convolutional neural network with multi-branch attention for multimodal brain tumor accurate segmentation. *Mathematical Biosciences and Engineering*, 2023, vol. 20, no. 9, p. 17384–17406. DOI: 10.3934/mbe.2023773

[18] SANKAR, M., BAIJU, B. V., PREETHI, D., et al. Efficient brain tumor grade classification using ensemble deep learning models. *BMC Medical Imaging*, 2024, vol. 24, no. 1, p. 1–22. DOI: 10.1186/s12880-024-01476-1

[19] CHEN, C., LIU, X., DING, M., et al. 3D dilated multi-fiber network for real-time brain tumor segmentation in MRI. In *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI 2019)*. Shenzhen (China), 2019, p. 314–322. DOI: 10.1007/978-3-030-32248-9_21

[20] LIU, S. A., XU, H., LIU, Y., et al. Improving brain tumor segmentation with dilated pseudo-3D convolution and multi-direction fusion. In *Proceedings of the 26th International Conference on MultiMedia Modeling (MMM 2020)*. Daejeon (South Korea), 2020, p. 727–738. DOI: 10.1007/978-3-030-37731-1_59

[21] WANG, K., LI, B., TAO, R. Pseudo-3D fully convolutional DenseNets for brain tumor segmentation. In *Proceedings of the Tenth International Conference on Digital Image Processing (ICDIP 2018)*. Shanghai (China), 2018, p. 1462–1468. DOI: 10.1117/12.2502863

[22] WANG, Q., WU, B., ZHU, P., et al. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle (WA, USA), 2020, p. 11531–11539. DOI: 10.1109/CVPR42600.2020.01155

[23] LI, H., KADAV, A., DURDANOVIC, I., et al. Pruning filters for efficient convnets. In *International Conference on Learning Representations (ICLR)*. Toulon (France), 2016, p. 1–13. ArXiv Preprint. DOI: 10.48550/arXiv.1608.08710

[24] JACOB, B., KLIGYS, S., CHEN, B., et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City (UT, USA), 2018, p. 2704–2713. DOI: 10.1109/CVPR.2018.00286

[25] PATRO, S. G. K., SAHU, K. K. Normalization: A preprocessing stage. *International Advanced Research Journal in Science, Engineering and Technology*, 2015, vol. 2, no. 3, p. 20–22. DOI: 10.17148/IARJSET.2015.2305

[26] MENZE, B. H., JAKAB, A., BAUER, S., et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 2014, vol. 34, no. 10, p. 1993 to 2024. DOI: 10.1109/TMI.2014.2377694

[27] BAID, U., GHODASARA, S., MOHAN, S., et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint, 2021. DOI: 10.48550/arXiv.2107.02314

[28] CAO, Y., SONG, Y. New approach for brain tumor segmentation based on Gabor convolution and attention mechanism. *Applied Sciences*, 2024, vol. 14, no. 11, p. 1–16. DOI: 10.3390/app14114919

[29] LIU, L., XIA, K. BTIS-Net: Efficient 3D U-Net for brain tumor image segmentation. *IEEE Access*, 2024, vol. 12, p. 133392 to 133405. DOI: 10.1109/ACCESS.2024.3460797

[30] JIANG, Y., ZHANG, Y., LIN, X., et al. SwinBTS: A method for 3D multimodal brain tumor segmentation using swin transformer. *Brain Sciences*, 2022, vol. 12, no. 6, p. 1–15. DOI: 10.3390/brainsci12060797

[31] ZHOU, Z., WANG, P., YU, X., et al. GSFormer: A gated skip connection and feature fusion transformer-based neural network for 3D MRI brain tumor segmentation. In *Proceedings of the 2024 International Conference on Intelligent Computing and Data Mining (ICDM)*. Chaozhou (China), 2024, p. 67–71. DOI: 10.1109/ICDM63232.2024.10762256

[32] HOU, Q., PENG, Y., WANG, Z., et al. MFD-Net: Modality fusion diffractive network for segmentation of multimodal brain tumor image. *IEEE Journal of Biomedical and Health Informatics*, 2023, vol. 27, no. 12, p. 5958–5969. DOI: 10.1109/JBHI.2023.3318640

[33] SUN, J., HU, M., WU, X., et al. MVSI-Net: Multi-view attention and multi-scale feature interaction for brain tumor segmentation. *Biomedical Signal Processing and Control*, 2024, vol. 95, Part A, p. 1–14. DOI: 10.1016/j.bspc.2024.106484

[34] AL-FAKIH, A., SHAZLY, A., MOHAMMED, A., et al. FLAIR MRI sequence synthesis using squeeze attention generative model for reliable brain tumor segmentation. *Alexandria Engineering Journal*, 2024, vol. 99, p. 108–123. DOI: 10.1016/j.aej.2024.05.008

[35] LIU, D., SHENG, N., HE, T., et al. SGEResU-Net for brain tumor segmentation. *Mathematical Biosciences and Engineering*, 2022, vol. 19, no. 6, p. 5576–5590. DOI: 10.3934/mbe.2022261

[36] ALWADEE, E. J., SUN, X., QIN, Y., et al. LATUP-Net: A light-weight 3D attention U-Net with parallel convolutions for brain tumor segmentation. *Computers in Biology and Medicine*, 2025, vol. 184, p. 1–16. DOI: 10.1016/j.compbiomed.2024.109353

[37] MAGADZA, T., VIRIRI, S. Efficient nnU-Net for brain tumor segmentation. *IEEE Access*, 2023, vol. 11, p. 126386–126397. DOI: 10.1109/ACCESS.2023.3329517

[38] ISENSEE, F., JÄGER, P. F., FULL, P. M., et al. nnU-Net for brain tumor segmentation. In *Proceedings of Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop (BrainLes 2020)*. Held in Conjunction with

MICCAI 2020. Lima (Peru), 2020, Part II, p. 118–132. DOI: 10.1007/978-3-030-72087-2_11

[39] YANG, H., ZHOU, T., ZHOU, Y., et al. Flexible fusion network for multi-modal brain tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2023, vol. 27, no. 7, p. 3349–3359. DOI: 10.1109/JBHI.2023.3271808

[40] PAN, D., SHEN, J., AL-HUDA, Z., et al. VcaNet: Vision transformer with fusion channel and spatial attention module for 3D brain tumor segmentation. *Computers in Biology and Medicine*, 2025, vol. 186, p. 1–12. DOI: 10.1016/j.compbiomed.2025.109662

[41] DIAO, Y., FANG, H., YU, H., et al. Multimodal invariant feature prompt network for brain tumor segmentation with missing modalities. *Neurocomputing*, 2025, vol. 616, p. 1–13. DOI: 10.1016/j.neucom.2024.128847

[42] ZHAO, J., XING, Z., CHEN, Z., et al. Uncertainty-aware multi-dimensional mutual learning for brain and brain tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2023, vol. 27, no. 9, p. 4362–4372. DOI: 10.1109/JBHI.2023.3274255

[43] GAO, T., HU, W., CHEN, M., et al. MSDMAT-BTS: Multi-scale diffusion model and attention mechanism for brain tumor segmentation. *Biomedical Signal Processing and Control*, 2025, vol. 104, p. 1–12. DOI: 10.1016/j.bspc.2025.107505

[44] WANG, W., CHEN, C., DING, M., et al. TransBTS: Multimodal brain tumor segmentation using transformer. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2021)*. Strasbourg (France), 2021, Part I, p. 109–119. DOI: 10.1007/978-3-030-87199-4_11

[45] CHEN, J., LU, Y., YU, Q., et al. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint, 2021, p. 1–13. DOI: 10.48550/arXiv.2102.04306

[46] HATAMIZADEH, A., NATH, V., TANG, Y., et al. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *Proceedings of Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2021)*. Held in conjunction with MICCAI 2021.Virtual conference, Cham (Switzerland), 2021, p. 272–284. DOI: 10.1007/978-3-031-08999-2_22

[47] RAZA, R., IJAZ BAJWA, U., MEHMOOD, Y., et al. dResU-Net: 3D deep residual U-Net based brain tumor segmentation from multimodal MRI. *Biomedical Signal Processing and Control*, 2023, vol. 79, no. 1, p. 1–12. DOI: 10.1016/j.bspc.2022.103861

[48] JIA, H., BAI, C., CAI, W., et al. HNF-Netv2 for brain tumor segmentation using multi-modal MR imaging. In *Proceedings of Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2021)*. Held in conjunction with MICCAI 2021. Virtual conference, Cham (Switzerland), 2021, vol. 12963, p. 106–115. DOI: 10.1007/978-3-031-09002-8_10

About the Authors ...

Mostafizur RAHMAN was born in 1986. He completed his B.Sc. in Electrical and Electronic Engineering (EEE) in 2009 and his M.Sc. in EEE in 2010, both at Islamic University, Kushtia, Bangladesh. He later received an additional M.Sc. in Computer Science from the same university in 2017. He is currently a Ph.D. candidate at Beijing Technology and Business University, Beijing, China. His research interests include pattern recognition, image processing, and computer vision.

Wenmin WANG received Ph.D. degree from Harbin Institute of Technology in 1989. Thereafter, he worked as an Associate Professor until 1991. He then gained overseas industrial experience for 18 years. Invited to return to China, he served from 2009 as a Professor at the School of Electronic and Computer Engineering, Peking University. Since 2019, he has been a Professor at the School of Computer Science and Engineering, Macau University of Science and Technology. His research interests include computer vision and multimedia processing.

Jiawei WANG, born in 1985, earned his M.D. and Ph.D. degree in 2013 from Nanjing University. He completed his postdoctoral fellowship from 2015 to 2017 at the Beijing Institute for Functional Neurosurgery at Xuanwu Hospital, Capital Medical University. Afterward, he joined the Department of Neurosurgery at the Cancer Hospital, Chinese Academy of Medical Sciences, where he currently serves as an Associate Chief Physician. His research is centered on the study of brain tumors, brain functions, and the application of multimodal brain imaging techniques.

Yu WANG (corresponding author) was born in 1977. She received her Ph.D. degree from the University of Science and Technology Beijing in 2009. She was engaged in scientific research as a post-doctoral in the Beijing Key Laboratory of Multidimensional and Multiscale Computing Photography, Tsinghua University from 2009 to 2011. She is now a Professor and doctoral supervisor of Beijing Technology and Business University. Her research interests include pattern recognition, image processing and computer vision.