LATUP-Net: A Lightweight 3D Attention U-Net with Parallel Convolutions for Brain Tumor Segmentation

Ebtihal J. Alwadee*, Xianfang Sun, Yipeng Qin and Frank C. Langbein

School of Computer Science and Informatics, Cardiff University, Cardiff, CF24 4AG, UK

ARTICLE INFO

Keywords:
Brain Tumor Segmentation
Lightweight
Parallel Convolutions
Attention
U-Net
Deep Learning
Grad-CAM

ABSTRACT

Early-stage 3D brain tumor segmentation from magnetic resonance imaging (MRI) scans is crucial for prompt and effective treatment. However, this process faces the challenge of precise delineation due to the tumors' complex heterogeneity. Moreover, energy sustainability targets and resource limitations, especially in developing countries, require efficient and accessible medical imaging solutions. The proposed architecture, a Lightweight 3D ATtention U-Net with Parallel convolutions, LATUP-Net, addresses these issues. It is specifically designed to reduce computational requirements significantly while maintaining high segmentation performance. By incorporating parallel convolutions, it enhances feature representation by capturing multi-scale information. It further integrates an attention mechanism to refine segmentation through selective feature recalibration. LATUP-Net achieves promising segmentation performance: the average Dice scores for the whole tumor, tumor core, and enhancing tumor on the BraTS 2020 dataset are 88.41%, 83.82%, and 73.67%, and on the BraTS 2021 dataset, they are 90.29%, 89.54%, and 83.92%, respectively. Hausdorff distance metrics further indicate its improved ability to delineate tumor boundaries. With its significantly reduced computational demand using only 3.07 M parameters, about 59 times fewer than other state-of-the-art models, and running on a single V100 GPU, LATUP-Net stands out as a promising solution for real-world clinical applications, particularly in settings with limited resources. Investigations into the model's interpretability, utilizing gradient-weighted class activation mapping and confusion matrices, reveal that while attention mechanisms enhance the segmentation of small regions, their impact is nuanced. Achieving the most accurate tumor delineation requires carefully balancing local and global features. The code is available at `https://qyber.black/ca/code-bca`.

1. Introduction

Brain tumors, particularly gliomas, represent some of the most lethal forms of cancer due to their inherent complexity, and high degree of variability among patients. Gliomas, classified into high-grade and low-grade gliomas, encompass various tumoral structures, including the enhancing tumor, necrotic tumor core, and edema [1]. Magnetic resonance imaging (MRI) is the standard imaging technique for diagnosing gliomas. MRI's central role in glioma imaging underscores the importance of precise segmentation for comprehensive tumor treatment and sparing healthy tissue, thereby influencing diagnosis, treatment planning, and prognostication. The subjectivity and labor-intensiveness of manual segmentation underscore the necessity for more efficient and accurate automatic segmentation techniques [2].

In response to the limitations of manual tumor segmentation, advances have been made towards fully automatic methods, specifically employing deep learning. Convolutional neural networks (CNNs) have demonstrated significant achievements in several computer vision domains, including, but not limited to, image segmentation, classification, and object detection. They are commonly utilized in medical imaging due to their characteristics of gaining local receptive fields, allowing weight sharing, and reducing parameters by pooling [3, 4]. However, despite the laudable progress in deep-learning-based strategies, brain tumor segmentation remains challenging due to tumor heterogeneity, anatomical variations, inconsistencies in imaging protocols, diverse and irregular tumor appearances, imaging artifacts, and inherent uncertainties in medical images. In addition, current deep architectures, particularly those employing 3D convolutions, have complex structures, high computing costs, and risk of overfitting, particularly with limited data [5]. These issues are especially significant in real-world applications like MRI image segmentation [3].

The increasing reliance on medical images and clinician-to-patient ratio challenges in low- and middle-income countries necessitate high precision, intelligent automated systems for timely and accurate medical diagnoses [6]. Moreover, energy sustainability targets and resource limitations, especially in geographic regions with limited computational resources, favor solutions requiring less computational resources. Hence, advances in medical image segmentation are geared towards exploring and developing models that are not only efficient but also resource-conscious, making them suitable for a wide range of computing environments [7].

Deep neural networks (DNNs) extract features that contain both, essential and redundant, information, ensuring a comprehensive understanding of input images. For example, inception blocks [8] utilize parallel convolutional paths with varying kernel sizes and a max-pooling operation to capture features at different scales and abstraction levels starting from the same input. However, each path in an inception

*Corresponding author

✉ AlwadeeEJ@cardiff.ac.uk (E.J. Alwadee); SunX2@cardiff.ac.uk (X. Sun); QinY16@cardiff.ac.uk (Y. Qin); frank@langbein.org (Frank C. Langbein)

ORCID(s): 0009-0004-3398-4023 (E.J. Alwadee); 0000-0002-6114-0766 (X. Sun); 0000-0002-1551-9126 (Y. Qin); 0000-0002-3379-0323 (Frank C. Langbein)

block operates independently from the start, without a shared feature extraction stage. This may lead to redundancy and an increase in the number of parameters, as each path may learn similar features. While this redundancy may enrich the feature set for DNN training, there is a trend towards reducing it towards more resource-efficient approaches.

Our proposed architecture, the LATUP-Net (Lightweight 3D ATtention U-Net with Parallel convolutions), is a novel U-Net variant with fewer parameters and comparable results to state-of-the-art architectures. It balances network depth and width, which is crucial for achieving optimal segmentation performance in lightweight models. Our work introduces an innovative parallel convolutions (PC) approach, derived from the inception block [8], within the first encoder block to tackle the challenge of feature redundancy, and efficiently harness diverse features at varying scales and orientations. After an initial shared embedded convolutional layer, the model processes the data through parallel paths with distinct kernel sizes followed by immediate pooling, capturing multi-scale spatial features. Compared to the inception block [8], this reduces feature redundancy and parameter count by leveraging a shared feature extraction stage. It simplifies the parallel block's paths and reduces convolutional depth before concatenation, but maintains diverse features while keeping the parameter count lower than that of inception blocks. This strategy addresses computational and memory constraints common in 3D image processing tasks and meets the demand for more resource-efficient models in real-world applications, where limitations often exist.

This paper further bridges the gap between the need for efficient models and the allure of attention mechanisms. We propose a lightweight model that significantly reduces computational requirements while maintaining high segmentation performance. We also investigate the intricate dynamics of attention mechanisms, which leads us to conclude that the attention mechanism effectively highlights tumor features but may overlook broader contextual information essential for accurate segmentation. LATUP-Net's performance is evaluated using the per-sample Dice similarity coefficient (DSC) and the Hausdorff distance (HD95). Results show comparable performance to state-of-the-art models while reducing the number of model parameters from over 181.57 million in the state-of-the-art nnU-Net [9] to 3.07 million. This offers benefits in terms of sustainability and accessibility.

The contributions of our work are:

- We propose LATUP-Net, an efficient, novel 3D U-Net variant, which incorporates parallel convolutions and attention mechanisms to achieve high brain tumor segmentation performance at highly reduced computational costs. Empirically, we show that LATUP-Net achieves significantly better segmentation accuracy than U-Net with half as many parameters and considerably fewer parameters than other state-of-the-art models with comparable performance.
- We demonstrate the efficacy of parallel convolutions in capturing features at different scales and orientations, leading to a richer representation of the input data and enhanced segmentation results. The parallel convolutions significantly improve efficiency and segmentation accuracy by incorporating a shared embedded convolution and a subsequent max-pooling operation following each parallel path.
- We engage in a detailed exploration of attention mechanisms. The crux of our investigation lies in assessing whether integrating attention invariably enhances segmentation or if there are scenarios where local context captured by convolutions suffices. We integrate an attention mechanism in the model and interpret its behavior using gradient-weighted class activation mapping (Grad-CAM) [10] and confusion matrix analysis. This reveals a nuanced trade-off: while attention effectively emphasizes salient features within tumor regions, it leads to an oversight of the broader contextual information essential for accurate segmentation. A more balanced approach that considers local details and global contextual cues could improve the model's segmentation performance, giving us a fuller picture of the tumor's structure.

The subsequent sections of this paper are structured as follows: Section 2 presents an overview of previous research conducted on the segmentation of brain tumors. Section 3 explains the LATUP-Net architecture, encompassing its essential elements and their impact on performance. Section 4 provides a comprehensive analysis of the experimental configuration, encompassing the datasets used and the evaluation metrics employed. Section 5 presents and analyses the obtained results. Finally, Section 6 concludes the paper and provides suggestions for further research.

2. Related Work

To analyze existing methods for brain tumor segmentation, we consider three perspectives: convolutional neural network (CNN) capacity, lightweight models, and attention mechanisms

2.1. CNN Models Capacity

Brain tumor segmentation using CNNs has been extensively studied in the literature, with most methods employing either 2D or 3D convolutions. Initially, 2D models dominated the field, where CNNs processed individual 2D slices derived from 3D MRI scans. However, these 2D slices inherently lack the volumetric context present in full 3D MRIs, leading to the potential loss of important semantic features. This issue is compounded by the fact that the resolution within the plane of 2D slices is often higher than that across slices, and the presence of small gaps between slices further exacerbates the loss of spatial continuity. To capture 3D feature information, 3D CNNs emerged as the preferred approach for analyzing MRI images of brain tumors, addressing the limitations inherent in 2D slice-based analysis [11].

While 3D convolutions make better use of spatial information, they also require more computational power and memory. To address this issue, Chen *et al.* [12] developed a memory-efficient solution, while preserving most of the volumetric information, by introducing a decoupled 3D U-Net model. It relies on separating a 3D convolution into sequential 2D and 1D convolutions and creating three parallel branches of these separated convolutions, one for each orthogonal view (axial, sagittal, and coronal) for the 1D convolution direction. The suggested model achieved competitive results while demonstrating high efficiency when tested on the BraTS 2018 dataset. While local and global features are necessary for making decisions during segmentation, low-level feature gradients (such as those containing information about boundaries, edges, lines, or dots) converge to zero as one proceeds deeper into the model. To address this, Wang *et al.* [13] proposed a TransBTS architecture that effectively embeds a transformer into a 3D U-Net model. To begin with, local feature maps are extracted using a 3D CNN encoder. The extracted features are transmitted through a transformer to capture global features. Afterward, the decoder incorporates the local and global features during the upsampling process to produce the segmentation result. Though this model requires more computational resources, it has shown promising results on the BraTS 2019 and 2020 datasets, achieving competitive or higher Dice score performance compared with state-of-the-art 3D models.

Traditional U-Net architectures perform exceptionally well on semantic segmentation tasks [14]. Nevertheless, such structures lack strategies to extract global feature information [15, 16]. To address this, the inception module [8] and a densely connected module [17] were added to the U-Net architecture by Zhang *et al.* [18]. Each inception module in the network uses 1×1, 3×3, and 5×5 convolution kernels to acquire multi-scale information. Their method performs admirably in segmenting images of lung tissue, blood arteries, and brain tumors. However, while this approach improves the representation power of the model, it also increases the number of parameters, which heightens the risk of overfitting and limits its effectiveness in scenarios with limited training data. Thus, given the constraints of low resource funding in hospitals, it becomes crucial to strike a balance between processing efficiency and network size through the deployment of lightweight networks.

2.2. Lightweight Models

U-Net variants have demonstrated satisfactory segmentation results for medical images. However, 3D networks require significantly more GPU memory than 2D networks with the same convolutional network structure and depth. Consequently, hardware requirements limit improvements in segmentation. Researchers have proposed a series of lightweight models to reduce network complexity and overcome hardware limitations. By reducing the number of network parameters and achieving highly accurate segmentation, Chen *et al.* [19] created a dilated multi-fiber (DMF) network that replaces convolutions with dilated convolutions of varying sizes as the fundamental unit. Although dilated convolutions, which modify the convolution's field of view by introducing gaps in the convolutional kernel, can capture features at various scales, they do not necessarily increase the diversity of features captured. Luo *et al.* [20] proposed a lightweight hierarchical decoupled convolution (HDC) unit by replacing 3D convolutions with pseudo-3D convolutions. However, the model's final segmentation precision is not very good, despite its ability to explore multi-scale, multi-view spatial contexts rapidly with a large reduction in computing complexity. In addition, Magadza *et al.* [21] utilized depth-wise separable convolutions to reduce computational complexity without sacrificing performance. However, this method cannot handle diverse and fundamental features in multiple, independent directions and orientations.

In summary, while traditional 3D U-Net models and their variants offer high segmentation accuracy, they are often resource-intensive and prone to overfitting in limited data scenarios. In contrast, lightweight models like the DMF network [19], HDC unit [20], and depth-wise separable convolutions [21], although less precise, provide a viable solution for environments with computational constraints. Our proposed lightweight 3D network addresses these concerns by employing a specific version of parallel convolutions which enhances feature extraction and segmentation performance with significantly fewer parameters, offering a balanced solution between computational efficiency and segmentation precision. We further incorporate an attention mechanism into our model, which addresses the overfitting issue typically associated with complex models.

2.3. Attention Mechanism

Traditional U-Nets give equal importance to all features within the feature maps. Given the notable class imbalance in brain tumor segmentation, some features are more crucial than others for accurate results. Attention mechanisms have emerged as effective tools to emphasize these crucial features and downplay less significant ones. Generally, attention mechanisms are bifurcated into two main types: channel attention and spatial attention. Channel attention enables the network to adaptively weigh the importance of different channels based on specific features in the image. This can potentially prioritize channels that are crucial for tumor detection [22]. Spatial attention, instead, fine-tunes the spatial feature maps adaptively, allowing the network to concentrate on specific regions with significant features [23]. In this study, we explore various lightweight attention mechanisms, all recognized for their capacity to enhance model expressiveness and boost overall performance.

The attention mechanisms explored include Squeeze-and-Excitation (SE) [22], the Convolutional Block Attention Module (CBAM) [24], Efficient Channel Attention

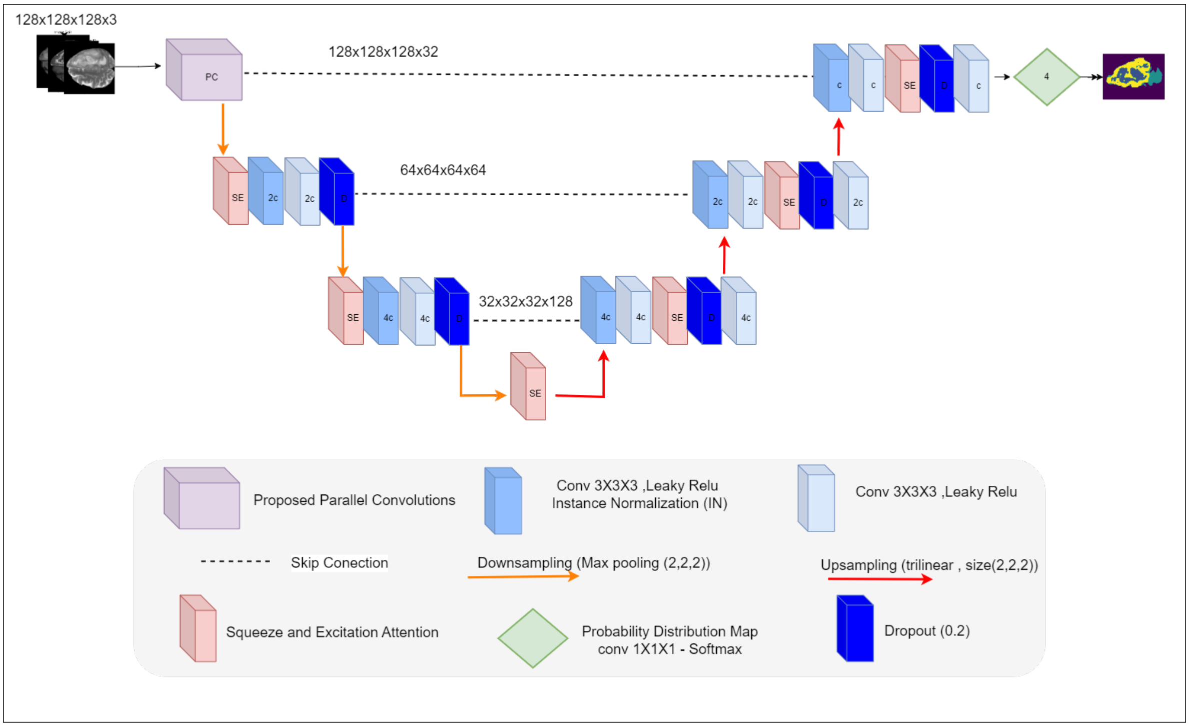

128x128x128x3
PC
128x128x128x32
c
c
SE
D
c
4
SE
2c
2c
D
64x64x64x64
2c
2c
SE
D
2c
SE
4c
4c
D
32x32x32x128
4c
4c
SE
D
4c
SE
Proposed Parallel Convolutions
Conv 3X3X3 ,Leaky Relu
Instance Normalization (IN)
Conv 3X3X3 ,Leaky Relu
Skip Conection
Downsampling (Max pooling (2,2,2))
Upsampling (trilinear , size(2,2,2))
Squeeze and Excitation Attention
Probability Distribution Map
conv 1X1X1 - Softmax
Dropout (0.2)

Figure 1: The LATUP-Net architecture.

(ECA) [25], and Residual Squeeze-and-Excitation (RSE) [26]. We further introduce a modified variant of SE where the fully connected (dense) layers are replaced by a 3D convolutional layer. We also experiment with combined mechanisms such as fusing CBAM and SE. The motivation behind combining CBAM and SE is to leverage the strengths of both. CBAM's ability to focus on pertinent spatial and channel features and SE's capacity to recalibrate channel-wise features may enhance the model's ability to capture complex interdependencies in the data. Another approach is designed to exploit the strengths of convolutions while adhering to the principles of the SE mechanism. Replacing the dense layer in SE with a 3D convolution layer (SE-3D) is aimed at maintaining the spatial information of the input tensor and capturing local spatial correlations, while simultaneously maintaining the ability to recalibrate channel-wise features.

3. The LATUP-Net Architecture

Here, we explain the components of our LATUP-Net architecture, illustrated in Figure 1, a lightweight variant of the original U-Net [11] with fewer parameters intended for the semantic segmentation of 3D brain tumors. Moreover, LATUP-Net utilizes multi-scale parallel convolutions (see Section 3.1) and channel attention on multi-modal data fusion (see Section 3.2).

Our encoder consists of three down-sampling blocks with 32, 64, and 128 filters, respectively. Only the first encoder block contains our parallel convolution block. The remaining two encoder blocks consist of a squeeze and excitation attention block [22] followed by two consecutive convolutions with instance normalization [27], and LeakyReLU activation with a negative slope of 0.1. The resulting tensor is passed through a dropout layer at a rate of 0.2. For each encoder block, an identity skip connection is added to map the encoder blocks onto their corresponding decoder blocks. All convolutions have a kernel size of $3 \times 3 \times 3$, and down-sampling is achieved by $2 \times 2 \times 2$ max-pooling to reduce the spatial resolution of the feature maps.

The decoder takes the feature maps of the encoder and doubles their spatial resolution using 3D trilinear up-sampling. It has three up-sampling blocks, each consisting of two 3×3×3 convolutions with 128, 64, and 32 filters respectively, followed by instance normalization, and LeakyReLU. A squeeze and excitation attention block has been added between the two convolutions. This is followed by a dropout layer, implemented with a rate of 0.2 to mitigate overfitting by randomly deactivating a portion of the neural connections during training. After the dropout layer, an additional 3×3×3 convolutional layer is incorporated. This layer is pivotal for refining the feature representations post-dropout, ensuring the restoration of spatial dimensions, and enhancing the network's ability to learn detailed, spatially coherent features essential for accurate segmentation.

The last two blocks of the encoder and decoder use an L2 regularizer. Finally, a $1 \times 1 \times 1$ convolutional layer with softmax activation is applied to the output of the decoder, which generates a probability map for each voxel indicating its likelihood of belonging to one of the tumor region classes to be segmented.

In line with findings by Rajamani *et al.* [28], integrating an attention block at the network's bottleneck—specifically, an SE block in our case—facilitates the capture of longer-range dependencies within the lowest-resolution activation maps. This adjustment boosts performance with only a slight increase in model complexity.

3.1. Parallel Convolutions (PC)

CNNs have demonstrated efficacy in feature extraction. However, indiscriminate augmentation of network layers might precipitate overfitting and computational overheads [29]. A balance between network depth and width remains paramount. Our strategy employs parallel convolutions with varying kernel sizes, drawing inspiration from the inception model [30]. This design allows the network to capture features at different scales, yielding a more efficient model with enhanced segmentation performance.

To improve the representation power of the network, which is a key factor in improving its accuracy and reliability, parallel convolutional layers can be added to different encoder and decoder blocks. However, adding parallel convolutions to all blocks may result in overfitting due to the large number of learnable parameters and limited training data. Therefore, it is added only to the first block of the encoder to extract the most fundamental and diverse features from the input data. Parallel convolutions can capture these features at different scales and orientations. This way, we improve representation power while reducing the risk of overfitting.

The proposed PC block (see Figure 2) is designed to process the input through a series of convolutional layers with different kernel sizes, each aiming to capture features at various spatial scales. Initially, the input passes through a shared embedded layer of a single $3 \times 3 \times 3$ convolution, which extracts a preliminary set of features from the input data. Following this, the features are processed in parallel through three distinct paths: one continues directly from the initial $1 \times 1 \times 1$ convolution, another passes through an additional $3 \times 3 \times 3$ convolution, and the third through a $5 \times 5 \times 5$ convolution. Convolutional layers with smaller kernel sizes, such as $1 \times 1 \times 1$ or $3 \times 3 \times 3$, are adept at detecting local patterns like edges and textures. Layers with larger kernels, like $5 \times 5 \times 5$, are suited for identifying broader spatial patterns and hierarchical structures within the data, thereby providing an extended receptive field. Each path then concludes with a max-pooling operation, reducing the dimensionality and computational load of the subsequent layers. The outputs of these parallel paths are concatenated, combining the multi-scale features into a unified feature map that is richer and more informative than what could be obtained from any single path.

This approach contrasts with the inception block [30], which typically includes multiple parallel paths starting from the same input, each with different combinations of convolutions and sometimes pooling operations, without a shared embedded convolution. The inception block aims to capture multi-scale information by applying various-sized convolutions (e.g., $1 \times 1 \times 1$, $3 \times 3 \times 3$, $5 \times 5 \times 5$) in parallel and then merging their outputs. However, each path in an inception block operates independently from the start, without a shared feature extraction stage. This increases the number of parameters and may lead to redundancy, as each path may learn similar features.

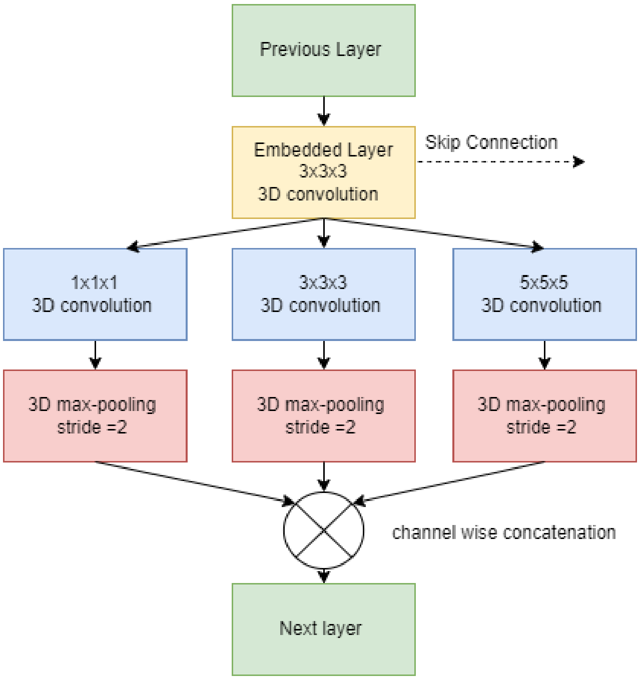


Figure 2: Proposed parallel convolutions.

Our proposed PC block contributes to making the model lightweight in several ways. Firstly, the shared embedded layer ensures that all paths operate on a common set of features, reducing redundancy and the need for each path to learn from scratch. This decreases the number of parameters compared to having multiple independent paths, as seen in inception blocks [30]. Secondly, by limiting each path to a single convolution and a pooling operation after the shared convolution, the model avoids the parameter growth associated with stacking multiple convolutions in each path. This streamlined approach enables efficient multi-scale feature extraction without the complexity and parameter overhead typically associated with more elaborate multi-path designs. Consequently, this design choice balances capturing diverse spatial features and maintaining a compact, efficient model architecture.

3.2. SE Attention Block

Based on rigorous evaluation of attention mechanisms in Section 5, we incorporate an SE block [22] into our final model. This block is recognized for its efficiency and lightweight nature. The SE block is composed of two distinct operations: Squeeze and Excitation. In the squeeze phase, input images of size $H \times W \times D \times C$ are transformed to a $1 \times 1 \times 1 \times C$ format through a Global Average Pooling (GAP) layer, which compresses the spatial resolutions,

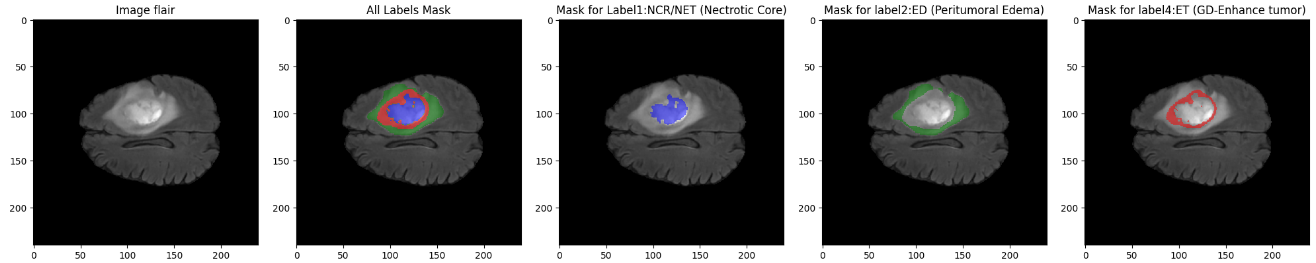


Image flair
All Labels Mask
Mask for Label1:NCR/NET (Nectrotic Core)
Mask for label2:ED (Peritumoral Edema)
Mask for label4:ET (GD-Enhance tumor)
0
50
100
150
200
0
50
100
150
200

Figure 3: MRI scan of a brain tumor featuring ground truth segmentation masks: blue represents necrotic and non-enhancing tumor areas (NCR/NET, Label 1), green highlights edema regions (ED, Label 2), and red indicates areas of enhancing tumor (ET, Label 4).

retaining only channel-centric information for the subsequent excitation operation. The excitation phase employs a series of layers, beginning with a fully connected layer complemented by a reduction factor r. This is then subjected to ReLU activation, succeeded by another fully connected layer, culminating in a sigmoid activation to produce the final output of the Excitation operation. A scaling transformation is executed to assimilate the channel-specific data, yielding an output enriched with channel-level information.

4. Data and Implementation Details

4.1. Data and Pre-processing

The proposed model is trained and validated using the Brain Tumor Segmentation (BraTS) benchmark datasets: BraTS 2020 [1, 31, 32] and BraTS 2021 [1]. BraTS 2021, is a superset of BraTS 2020, encompasses 1,251 patients, which include high-grade gliomas (HGG) and low-grade gliomas (LGG). The BraTS 2020 dataset contains 369 patients, of which 76 have been diagnosed with LGG, with the remainder having HGG.

The BraTS 2020 and BraTS 2021 datasets are used for efficiency and sustainability, with initial testing on the smaller dataset allowing for algorithm refinement and model selection before deployment on the larger dataset. It conserves computational resources and time, and aligns with iterative development best practices, where initial testing on a subset of data can provide quick feedback, critical for early-stage model tuning and optimization [33].

Each dataset consists of 3D scans with 155 individual "slices" or images, each having 240×240 pixels. These scans capture four MRI modalities–T2, T1, T1ce, and FLAIR–crucial for brain tumor segmentation, offering distinct insights. T1-weighted images illustrate anatomical structures, distinguishing between gray and white matter. T2-weighted images aid in visualizing edema, by emphasizing water content. T1ce images, with contrast enhancement, focus on blood vessel imaging, a key component in identifying active tumor growth. Conversely, FLAIR images suppress cerebrospinal fluid signals, illuminating anomalies in intensity and subtle lesions indicative of tumor expansion. Ground truth segmentation masks are meticulously annotated by one to four expert neuroradiologists per case. The scan has been segmented into 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) as exemplified in Figure 3. Following common practices in the literature, we classify these into three main tumor regions for segmentation: whole tumor (WT) encompassing NCR/NET, ED, and ET (Labels 1, 2, 4), tumor core (TC) consisting of NCR/NET and ET (Labels 1, 4), and the sole ET (Label 4).

The BraTS datasets have been meticulously preprocessed by their developers, including co-registration to a consistent anatomical template, interpolation to a unified resolution (1 mm^3), and skull stripping. However, MRI scan intensity values often display inconsistencies and may fluctuate due to many factors. To mitigate this, we normalize the intensity range to the interval $[0, 1]$ using the min-max scaler [34]. This adjustment not only enhances data consistency but also optimizes it for deep learning algorithms.

We further crop the images to a standard size of $128 \times 128 \times 128$ voxels, centered on the MRI scans. Preliminary tests suggested that including T1 images with T1ce, T2, and FLAIR only marginally improves segmentation results. Since T1ce is essentially a contrast-enhanced derivative of T1, and T1 mainly contributes to identifying a small fraction of the edema, which FLAIR can effectively detect as well, we chose to exclude T1 from our inputs to conserve computational resources.

4.2. Implementation Details

We implement our network via Keras in Tensorflow 2.15. Computations are executed on a single NVIDIA V100 16GB GPU. However, it's important to note that our architecture is versatile and can be run on other graphics cards, including consumer-grade GPUs. For training, we employ the ADAM optimizer [35], setting the learning rate to 1×10^{-4}. Training proceeds with a batch size of 1, a choice primarily dictated by GPU memory constraints. We use a constant number of 200 epochs. To mitigate the risk of overfitting and enhance the model's generalization capabilities, L2 regularization is applied to the convolutional kernel parameters with a factor of 0.02.

During preliminary experimentation, we evaluated the `ReduceLROnPlateau` callback for dynamic learning rate adjustments. However, observations indicated a predisposition

Table 1
Hyperparameters for the LATUP-Net model.

Hyperparameter	Value
Input size	$128 \times 128 \times 128 \times 3$
Batch size	1
Hidden layer activation	Leaky ReLU($\alpha = 0.1$)
Optimizer	ADAM ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Learning rate	1×10^{-4}
Number of epochs	200
Loss function	Weighted Dice score Loss (see Section 4.3)
Dropout	0.2
Regularization	L2 (factor 0.02)
Output layer activation	Softmax
Output size	$128 \times 128 \times 128 \times 4$

towards overfitting when employed. As such, it was excluded from final training (see Section 5.1.1).

The specific hyper-parameter settings we adopted during model training are detailed in Table 1. Detailed results with a discussion are in Section 5. The source code is available at [36] with final models and analysis results at [37].

4.3. Loss Function

Loss function selection is a critical factor in contemporary deep-learning network designs, especially in the field of brain tumor segmentation. Recent studies indicate that no single popular loss function consistently offers superior performance across various segmentation tasks [38].

Compound loss functions, which combine two or more types of loss functions, have emerged as the most robust and competitive in different scenarios [38]. In our experiments, we aim to enhance segmentation performance and address the severe class imbalance in the BraTS datasets by combining Dice loss with Binary Cross Entropy (BCE) [39] and Dice loss with focal loss [40]. However, based on our experiments, these compounded loss function approaches did not significantly outperform Dice loss alone. Therefore, to boost the segmentation performance and solve the class imbalance problem, the loss function used during the final training process is the Weighted Dice score Loss (WDL).

The Dice score loss for each class i, corresponding to the network output channels BG, NCR/NET, ED, and ET, is

$$DSL_i = 1 - \frac{2\sum_n \left(y_{i,n}^{\text{true}} \odot y_{i,n}^{\text{pred}}\right) + \epsilon}{\sum_n \left(y_{i,n}^{\text{true}}\right)^2 + \sum_n \left(y_{i,n}^{\text{pred}}\right)^2 + \epsilon}. \quad (1)$$

y_i^{true} and y_i^{pred} represent the ground truth and predicted segmentation masks for class i, respectively; n iterates over all elements of y_i^{true} and y_i^{pred}; $\odot$ signifies point-wise multiplication; and ϵ is a negligible constant introduced to avoid division by zero. In our experiments, we set $\epsilon = 0.00001$. Note that here the network output masks are the original BG, NCR/NET, ED, and ET regions in the ground truth (see Section 4.1).

The WDL weights w_i for class i are computed according to the ENet paper [41],

$$w_i = \frac{1}{\log(C + \frac{c_i}{T})} \quad (2)$$

where $C = 1.02$, c_i is the voxel count for class i, and T is the total count of voxels across all classes. This formula ensures that classes with fewer voxels receive higher weights to balance the loss during the training process. Here we use the WT, TC and ET regions instead of the individual output channels above to compute the weights and we get $w_{\text{WT}} = 1.64$, $w_{\text{TC}} = 2.55$, and $w_{\text{ET}} = 3.40$.

Overall this givens our Weighted Dice score Loss (WDL),

$$\begin{aligned} \text{WDL} = {} & w_{\text{WT}} \cdot (DSL_{\text{NCR/NET}} + DSL_{\text{ED}} + DSL_{\text{ET}}) \\ & + w_{\text{TC}} \cdot (DSL_{\text{NCR/NET}} + DSL_{\text{ET}}) \\ & + w_{\text{ET}} \cdot DSL_{\text{ET}}, \end{aligned} \quad (3)$$

which is equivalent to

$$\begin{aligned} \text{WDL} = {} & (w_{\text{WT}} + w_{\text{TC}} + w_{\text{ET}}) \cdot DSL_{\text{ET}} \\ & + (w_{\text{WT}} + w_{\text{TC}}) \cdot DSL_{\text{NCR/NET}} \\ & + w_{\text{WT}} \cdot DSL_{\text{ED}}. \end{aligned} \quad (4)$$

In this expression, the dice score loss (DSL) for each class $i \in \{\text{NCR/NET}, \text{ED}, \text{ET}\}$ is computed separately and weighted according to the importance of the corresponding tumor region (WT, TC, and ET) and then summed up to compute the total weighted Dice score loss for the segmentation task. This loss function links the clinical relevance of each tumor region with the network's output channels, ensuring that the segmentation process prioritizes the most clinically significant areas which is crucial for achieving optimal segmentation performance. Using the ENet weights helps in addressing the class imbalance by assigning higher weights to smaller but more important tumor regions.

It is also important to note that while the network output channels include the background class (BG), it is not included in the loss and weights calculation. We found its presence improves the segmentation performance of the model. We also explored the use of different output channels, weights, manual adjustment of the weights. However, it did not yield satisfactory results.

4.4. Evaluation Metrics

We measure the effectiveness of the proposed model using the Dice similarity coefficient (DSC), and the 95^{th} percentile Hausdorff distance (HD95). DSC and HD95 are widely accepted as the primary performance evaluation metrics in image segmentation tasks. The DSC quantifies the spatial overlap between the ground truth and the predicted segmentation region. HD95 calculates the 95^{th} percentile of the distances between the points in the ground truth and the predicted set. This is akin to the conventional symmetric Hausdorff distance but reduces the impact of outliers by focusing on the 95^{th} percentile. HD95 in particular indicates the accuracy of boundary prediction, and a lower score reveals the model's precision in delineating the tumor margins.

The metrics are defined as

$$\mathrm{DSC} = \frac{2|P \cap T|}{|P| + |T|}, \tag{5}$$

$$\mathrm{HD95} = \max\left(\max_{p \in P_{95\%}} \min_{t \in T} \|p - t\|, \max_{t \in T_{95\%}} \min_{p \in P} \|t - p\| \right), \tag{6}$$

where p and t are voxel coordinates for the predicted and ground truth regions respectively; P is the predicted region, and T is the ground truth region. The corresponding 95^{th} percentile regions are represented by $P_{95\%}$ and $T_{95\%}$.

We employ different metrics for evaluating our model based on the data partitioning approach used. For model selection (see Section 5.1, and Section 5.2), we employ an 80/20 training-testing holdout split, and we report the mean and standard deviation of the per-sample (per-patient) metrics to gain insights into the model's performance under controlled conditions. Once the optimal model (LATUP-Net) is determined, and for the comparison to other state-of-the-arts models (see Section 5.4), five-fold cross-validation is applied to consider any data dependencies and ensure a comprehensive evaluation of the model's performance across varied data scenarios. During cross-validation, we calculate the mean and standard deviation of the mean DSC and HD95 across all folds to assess the model's robustness and general performance.

5. Experimental Results and Discussion

This section covers four main analyses. Firstly, we analyze the performance of some variants of our architecture and the influence of learning rate optimization on the segmentation performance. Then, we investigate which attention mechanism gives the best performance. We also evaluate the effectiveness of the attention mechanism. Finally, we compare the performance of our LATUP-Net to the state-of-the-art on BraTS 2020 and 2021.

5.1. Overall Performance Analysis

To find the best variant of our proposed model and training, we compare the following architectures:

U-Net: The baseline U-Net model trained using Dice loss.

Inception: Modified U-Net model with inception module trained using Dice loss.

PC: Modified U-Net model with parallel convolutions trained using Dice loss.

PC + SE: Modified U-Net model with parallel convolutions and channel attention trained using Dice loss.

PC + WDL: Modified U-Net model with parallel convolutions trained using weighted Dice score Loss.

PC + SE + WDL: Modified U-Net model with parallel convolutions, and attention trained using weighted Dice score Loss.

The models are initially selected based on a single 80/20 split using the same training hyperparameters (see Table 1) on the BraTS 2020 dataset. Table 2 shows the segmentation results of the test set from these training runs. The table also lists the parameter counts, allowing for a comparison of model complexity. The performance of these models is assessed by employing two key metrics: per-sample DSC and HD95.

In the inception model, we modify the baseline U-Net by adding the inception module to its first block. We observe that the number of parameters decreased by 2 M. The model outperforms the U-Net with a DSC increment by 4.31, 3.78, and 8.63 and a decrement in the HD95 by 7.63, 8.8, and 4.44 for WT, TC, and ET respectively. However, despite the reduction in the parameter counts, using the inception block led to the model requiring a lot of memory, thereby increasing its memory consumption during training. This memory-intensive behavior of the model might attributed to the inception block design.

The PC model has 2.59 M fewer parameters than U-Net and exhibits rapid convergence during the initial training epochs, leading to shorter training time and reduced need for computational resources. It also obtains a significant improvement in the segmentation performance compared to U-Net with a DSC increment by 4.91, 7.06, and 9.58 and HD95 decrement by 13.51, 9.75, and 6.48 for WT, TC and ET respectively.

When comparing our proposed parallel convolutions (PC) with the inception module, PC offers a strategic advantage. By replacing the conventional $1 \times 1 \times 1 \times 1$ convolution with a $3 \times 3 \times 3 \times 3$ convolution, PC achieves a more spatially compact and contextually rich representation, while effectively reducing memory consumption. In contrast, the inception module with 0.59 M more parameters faced a surge in model memory usage during training. Furthermore, our design's judicious positioning of pooling operations optimally condenses feature map dimensions, ensuring efficient memory usage without compromising capturing features. The PC model outperformed the inception model with 0.6, 3.28, and 0.95 DSC increment and 5.88, 0.95, and 2.08 HD95 decrement for WT, TC, and ET respectively which proves PC's ability to segment difficult tumor regions such as TC.

Integrating the attention mechanism, specifically SE attention with the PC model, leads to an increment in the model parameters but slightly improves the performance on the test data. It enhances the model's ability to segment small regions, as evidenced by 0.39 and 1.63 increments in the DSC for WT and ET, respectively. In contrast, the DSC performance of TC detection decreased slightly by 0.93. At the same time, the benefits of the parallel convolutions are maintained, despite a slight trade-off for certain tumor regions. Several lightweight attention modules are compared in Section 5.2.

Table 2
Comparative analysis of model architectures for brain tumor segmentation using the BraTS 2020 dataset: This table illustrates the mean and standard deviation (indicated by ±) for the per-sample Dice similarity coefficient (DSC) and the 95^{th} percentile Hausdorff distance (HD95). Results are segmented into whole tumor (WT), tumor core (TC), and enhancing tumor (ET) categories, based on an 80/20 train/test set split. Additionally, the table includes the number of trainable parameters for each model.

Model	DSC (%)			HD95 (mm)			Parameters
	WT	TC	ET	WT	TC	ET	
U-Net	83.22 ± 9.29	77.13 ± 18.34	60.65 ± 27.44	18.50 ± 22.17	15.38 ± 22.91	19.34 ± 30.24	5.65 M
Inception	87.53 ± 7.49	80.91 ± 18.64	69.28 ± 29.05	10.87 ± 15.72	6.58 ± 7.78	14.94 ± 29.06	3.65 M
PC	88.13 ± 7.14	84.19 ± 16.74	70.23 ± 28.40	4.99 ± 4.00	5.63 ± 6.71	12.86 ± 26.39	**3.06** M
PC + SE	88.52 ± 7.10	83.26 ± 17.18	71.86 ± 27.02	5.98 ± 7.88	5.51 ± 5.20	12.96 ± 26.55	3.07 M
PC + WDL	**89.58 ± 5.70**	**85.35 ± 15.49**	73.21 ± 27.18	**4.78 ± 4.41**	5.25 ± 6.38	11.65 ± 26.33	**3.06** M
PC + SE + WDL	88.72 ± 6.33	84.71 ± 15.65	**74.49 ± 25.98**	5.76 ± 4.41	**5.15 ± 6.28**	**10.65 ± 25.33**	3.07 M

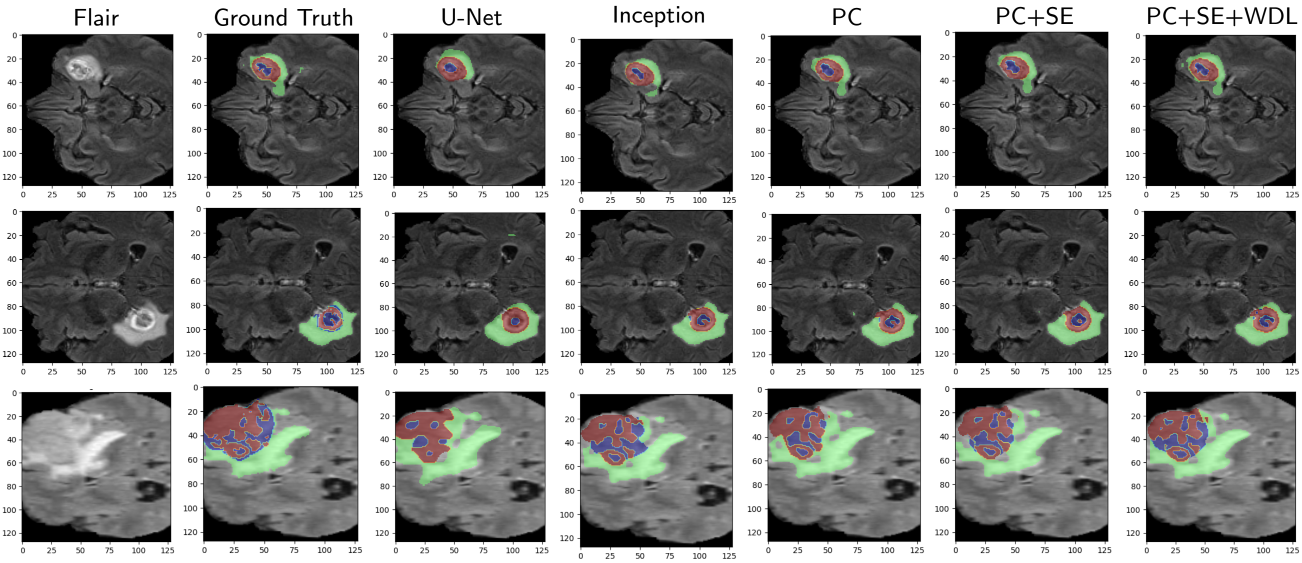


Figure 4: Segmentation results from the test set that are typical of those produced by the various networks. The results for a single patient from each network are shown in each row. The enhancing tumor is depicted in blue, the necrotic and non-enhancing tumor in red, and the edema in green (after extracting the distinct regions from the partially overlapping segmentation results).

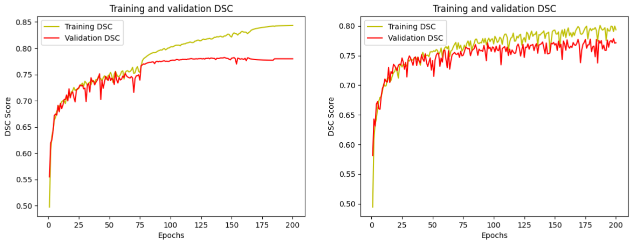
Notably, our investigations reveal a significant performance enhancement upon employing the WDL with the PC model. Specifically, the DSC for WT, TC, and ET increased by 1.45, 1.16, and 2.98 respectively and HD95 reduced by 0.21, 0.38, and 1.21 respectively.

This improvement underscores the utility of the WDL in case of a segmentation region size imbalance. However, we notice that when adding the attention mechanism, the result of WT and TC decreased slightly in both DCS and HD95, which leads us to check whether attention is needed. This is further investigated in Section 5.3.

Figure 4 depicts qualitative comparisons between the various networks in segmenting the distinct tumor regions. The qualitative results demonstrate that our LATUP-Net which consists of PC and SE trained with WDL, outperformed all other models, consistent with the quantitative results in Table 2.

5.1.1. The Influence of Learning Rate Decay on the Segmentation Performance

Training neural networks necessitates careful control of convergence and prevention of overfitting. Traditional models typically utilize learning rate schedulers alongside stochastic gradient descent. However, with the introduction of advanced optimizers such as ADAM, which integrates momentum and regularization, there has been a shift to strategies like `ReduceLROnPlateau`. This approach adjusts the learning rate in response to training plateaus by monitoring validation loss, akin to early stopping criteria [42]. Despite the recognized benefits of `ReduceLROnPlateau` in contemporary optimization scenarios, our experimentation yielded nuanced insights. While training models with and without learning rate decay, a small performance improvement was observed, but the associated learning curves exhibit clear signs of overfitting (see Figure 5(a)). In contrast, models



(a) (b)

Figure 5: Modified U-Net with PC: Dice score curves for all tumor regions (a) with and (b) without learning rate decay.

trained without learning rate decay did not display such overfitting tendencies (see Figure 5(b)). These observations are likely attributed to our model's already minimal learning rate. Consequently, in our final model configuration, we opted to forgo the learning rate decay to circumvent the observed overfitting tendencies.

5.2. Comparison of Different Attention Mechanisms

In our approach so far, we have used the SE attention mechanism. In this section, we explore whether alternative attention mechanisms (see Section 2.3) may yield improved performance within the confines of our model. For a uniform evaluation, we incorporate all attention mechanisms after the parallel convolutions (PC). It is also important to note that the model trained using dice loss and the results presented here are derived from configurations that do not include the weighted Dice score loss (WDL). This choice was made to isolate the impact of the attention mechanisms themselves, allowing for a clear comparison of their effectiveness in the absence of the WDL's influence.

Table 3 summarizes our findings, detailing the performance metrics of residual SE, CBAM, ECA, and their combinations. While all mechanisms enhance the model's accuracy, SE stands out as the most effective.

CBAM, despite its comprehensive design, might be limited by its complexity, potentially impacting efficiency. Similarly, while ECA aims for computational efficiency, it might fall short in capturing intricate inter-channel dynamics as effectively as SE. SE's distinct capability in identifying enhanced tumors, along with its reduced parameters, underscores its utility. Our results affirm the significance of SE in tumor detection but also suggest that the performance of other attention mechanisms could be context-dependent, varying with the dataset or task.

Analyzing the standard deviation values, the variance in DSC scores among methods appears minimal. Given this minimal divergence, the definitive superiority of certain attention mechanisms can be debated. In such scenarios, streamlined models with fewer parameters, such as PC+SE with 3.07 M parameters, offer competitive or superior performance.

The HD95 metric presents more noticeable differences across methodologies. This metric sheds light on spatial discrepancies in segmentation outcomes, possibly arising from the mechanisms' ability to capture spatial nuances, outliers, irregular tumor shapes, or unique dataset attributes.

5.3. Evaluating the Effectiveness of the Attention Mechanism

To understand the impact of the integrated attention mechanism and to determine if it operates as intended in our proposed architecture, we conduct two primary experiments: gradient-weighted class activation mapping (Grad-CAM) visualizations [10] and confusion matrix analysis.

5.3.1. Visual Interpretation using Grad-CAM

In investigating the efficacy of attention mechanisms in brain tumor detection, the question arises whether the local context is sufficient for accurate segmentation. Traditional convolutions inherently capture the local context without necessitating attention. Given the qualitative nature of our experiments and the challenges of visualizing every data point, it is crucial to strategically select representative samples for thorough analysis. To interpret and delve deeper into the model's behavior, Grad-CAM is applied to three distinct samples that are the same for both models, with attention (LATUP-Net) and without attention(PC + WDL). These samples are selected based on their loss value with some representative slice from the test set and represent the best-performing, the median, and the worst-performing w.r.t. the WDL for comparative analysis.

Grad-CAM [10] was applied to the layer preceding the softmax activation used for generating the predictions. This layer was particularly selected because it directly contributes to the decision-making process, offering a transparent view into how the model weighs different regions in its segmentation task. The visualizations from Grad-CAM, as shown in Figure 6, provide insights into the regions the models deemed most important in making their predictions. The images reveal that the attention mechanism seems to focus too intently on tumor regions, potentially overlooking broader contexts. Notably, while attention appears to work as designed by emphasizing certain regions, it might not always align with the most accurate predictions, especially with areas like the necrotic region being overemphasized. Additionally, the model sometimes misinterprets the texture of regions, resulting in potential misclassifications. This is particularly noticeable between necrotic and enhancing regions, which may have similar textures.

5.3.2. Confusion Matrix Analysis and Implications

To provide quantitative evidence supporting the findings from Grad-CAM visualizations, we perform a confusion matrix analysis, as shown in Figure 7. This matrix details how pixels from one class are misclassified as those from another, offering insights into the LATUP-Net model behavior. It is important to note that, there were patients with no enhanced tumor (ET) class in their ground truth. These cases have not been considered for the confusion matrix, as including them would distort the percentages. A primary observation is the misclassification between the necrotic and edema

Table 3
Efficiency analysis of various attention mechanisms for brain tumor segmentation on the BraTS 2020 test set. This table illustrates the mean and standard deviation (indicated by ±) for the per-sample Dice similarity coefficient (DSC) and the 95^{th} percentile Hausdorff distance (HD95). Results are segmented into whole tumor (WT), tumor core (TC), and enhancing tumor (ET) categories, based on an 80/20 train/test set split. Additionally, the table includes the number of trainable parameters for each model.

Method	DSC (%)			HD95 (mm)			Parameters
	WT	TC	ET	WT	TC	ET	
PC+SE	88.52 ± 7.10	83.26 ± 17.18	**71.86 ± 27.02**	5.98 ± 7.38	**5.51 ± 5.20**	**12.96 ± 26.55**	**3.07 M**
PC+CBAM	89.38 ± 6.53	**84.36 ± 15.94**	70.01 ± 29.39	**4.78 ± 4.19**	5.51 ± 5.65	14.12 ± 28.27	3.45 M
PC+CBAM+SE	88.91 ± 6.31	83.87 ± 15.63	70.25 ± 28.33	5.24 ± 5.07	5.58 ± 5.60	12.92 ± 26.42	3.26 M
PC+ (SE-3D)	89.05 ± 6.56	83.91 ± 16.97	69.83 ± 29.63	5.38 ± 6.89	5.72 ± 7.37	13.75 ± 28.14	**3.07 M**
PC+Residual SE	**89.60 ± 6.50**	84.06 ± 17.38	70.79 ± 29.42	5.40 ± 7.46	5.93 ± 8.16	13.19 ± 26.34	**3.07 M**
PC+ECA	84.47 ± 6.97	80.12 ± 18.21	63.19 ± 28.15	6.81 ± 4.01	7.08 ± 7.05	15.03 ± 28.13	3.10 M

regions, a claim also derived from the qualitative visuals (see Section 5.3.1). The model perceives textural similarities between these areas, further complicated by nested tumor structures. Additionally, the necrotic region is notably vulnerable to misclassification, perhaps due to its texture and position within the tumor's structure. When examining the standard deviation in Figure 8, combined with the average misclassification percentages, the classification performance variability is clear. This variability might spotlight outliers or deviations in model predictions, with classes like the necrotic region and enhancing region showing significant misclassification variability. The variations imply that certain samples heavily influence averages. In conclusion, the attention mechanism, while offering nuanced insights, can overly prioritize local features, causing the model to overlook important topological relationships recognized by human experts. A balance between local and global contexts seems important.

5.4. Comparison with State-of-the-Art Models

We compare our LATUP-Net model with the state-of-the-art models on the BraTS 2020 and 2021 datasets. We conduct five-fold cross-validation resulting in the performance results shown in Figures 9 and 10 of the test sets. We compare these results with other high-performing models and directly refer to related publications to obtain the evaluation results of the corresponding methods, which is a common approach in the study of brain tumor segmentation, as the source code of many existing brain tumor segmentation methods has not been released, and to avoid the bias introduced by model retraining.

5.4.1. Comparison with BraTS 2020 Results

We compare our LATUP-Net model with various state-of-the-art models on the BraTS 2020 dataset. The evaluation focuses on three key metrics: the Dice similarity coefficient (DSC), the 95^{th} percentile Hausdorff distance (HD95), and the number of model parameters. For our LATUP-Net model, we present the average performance across five-fold cross-validation to ensure a robust and comprehensive assessment. In contrast, for the state-of-the-art models, we report the results as they are presented in their respective original publications, which may include their best-split results or cross-validation outcomes. Detailed results of this comparative analysis are compiled in Table 4.

Our LATUP-Net model demonstrates significant improvements in HD95 for all tumor regions (whole tumor, tumor core, and enhancing tumor), suggesting accurate predictions of tumor boundaries. This precision is of paramount importance in medical imaging since accurate boundaries can greatly impact clinical decision-making. However, it is worth noting that in the final evaluation of the five-fold cross-validation, the HD95 has been ignored for images that do not have clear boundaries, since having unclear, ambiguous 'edges' from which to measure the distances may mislead the results.

Our model surpasses several others in DSC measurements for whole tumor segmentation. Specifically, it outperforms Raza *et al.* [43], Ballestar *et al.* [44], and Messaoudi *et al.* [45] by 1.81, 4.2, and 7.73 respectively.

Although a deeper analysis is needed to determine the exact mechanisms behind this efficiency, the results from the earlier model comparison section provide concrete evidence of the effectiveness of our approach. Specifically, our adoption of parallel convolutions in the first encoder block appears to play a crucial role. This is evidenced by the PC model having 2.59 M fewer parameters than U-Net, yet achieving rapid convergence during the initial training epochs. This leads to shorter training times and a reduced need for computational resources.

Brain tumor segmentation, particularly the tumor core and enhancing tumor, poses significant challenges. While models like nnU-Net exhibit strong results, our LATUP-Net model showcases similar scores. It is critical to differentiate

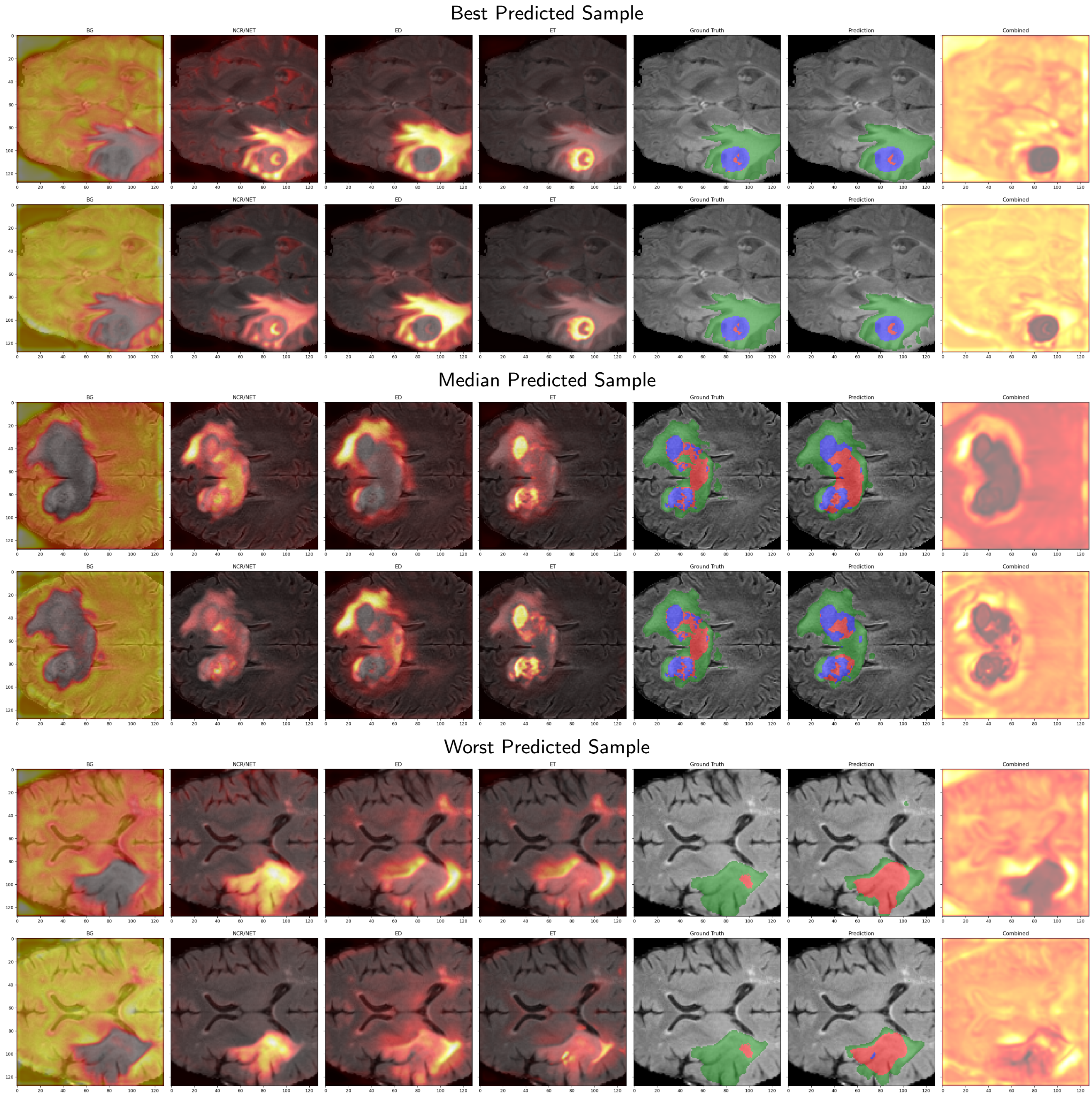


Figure 6: Visual interpretations of model predictions using Grad-CAM: For each of the three samples selected from the test set, representing the best, median, and worst predicted cases, both the standard and attention-enhanced models. From left to right in each row we show the GradCAM for the BG, NEC/NET, ED, ET output channel, the ground truth and the prediction, and finally the combined GradCAM for all output channels. The first row visualizes the standard model (PC + WDL), and the second row visualizes the attention-enhanced model (LATUP-Net) per case.

between DSC and HD95 scores. Although our HD95 scores indicate accurate boundary predictions, our DSC for the enhancing tumor (ET) of 73.67% is not the highest, signaling areas of improvement in volumetric overlap with the ground truth.

Concerning the number of parameters, our model is remarkably efficient, with only 3.07 M parameters, a stark contrast to the 181.56 M parameters in nnU-Net. This efficiency reduces computational costs and the time required for processing, which is crucial for real-time clinical applications, while the performance remains comparable and even better in some aspects (see the HD95 results). Even if Although Raza *et al.* [43] has a parameter count closer to ours, they do not achieve a performance comparable to ours, highlighting the efficiency of our architecture.

Table 4
Comparison of the performance and complexity of different methods on the BraTS 2020 test set. WT – whole tumor, TC – tumor core, ET – enhancing tumor. Results highlighted in red indicate the best result, while those in blue represent the second best. The symbol - indicates values not provided in the original paper.

Study	Image Size	WT		TC		ET		Parameters
		DSC (%)	HD95 (mm)	DSC (%)	HD95 (mm)	DSC (%)	HD95 (mm)	
3D U-Net Baseline	128 × 128 × 128	83.58	18.50	82.19	15.38	68.76	19.34	5.65 M
Isensee *et al.* [46]	128 × 128 × 128	88.95	8.498	85.06	17.337	82.03	17.805	181.56 M
Tang *et al.* [47]	80 × 96 × 64	89.29	4.62	78.97	10.07	70.30	34.30	-
Ballestar *et al.* [44]	64 × 64 × 64	84.21	20.40	75.03	12.92	61.75	48.76	-
Wang *et al.* [13]	128 × 128 × 128	90.09	4.96	81.73	9.76	78.73	17.94	32.99 M
Messaoudi *et al.* [45]	192 × 160 × 108	80.68	-	75.20	-	69.59	-	-
Raza *et al.* [43]	128 × 128 × 128	86.60	-	83.57	-	80.04	-	30.47 M
LATUP-Net (proposed)	128 × 128 × 128	88.41	3.19	83.82	4.24	73.67	3.97	3.07 M

Table 5
Comparison of the performance and complexity of different methods on the BraTS 2021 test set. WT – whole tumor, TC – tumor core, ET – enhancing tumor. Results highlighted in red indicate the best result, while those in blue represent the second best. The symbol - indicates values not provided in the original paper.

Study	Image Size	WT		TC		ET		Parameters
		DSC (%)	HD95 (mm)	DSC (%)	HD95 (mm)	DSC (%)	HD95 (mm)	
Peiris *et al.* [48]	128 × 128 × 128	90.77	5.37	85.39	8.5	81.38	21.83	-
Akbar *et al.* [49]	128 × 128 × 128	89.07	11.78	80.73	21.17	78.02	25.8	-
Jia *et al.* [50]	128 × 128 × 128	92.53	3.45	87.96	5.86	84.80	14.17	17.91 M
Li *et al.* [51]	128 × 128 × 128	90.18	6.15	81.61	16.65	76.89	30.21	-
Ma *et al.* [9]	128 × 128 × 128	92.59	3.80	87.86	9.20	82.17	21.09	-
Hatamizadeh *et al.* [52]	128 × 128 × 128	92.6	5.83	88.5	3.77	85.8	6.01	61.98 M
Roth *et al.* [53]	128 × 128 × 128	90.6	4.54	83.5	10.11	79.2	16.61	-
LATUP-Net (proposed)	128 × 128 × 128	90.29	3.03	89.54	2.44	83.92	3.06	3.07 M

5.4.2. *Comparison with BraTS 2021 Results*

Evaluating our LATUP-Net model on the more recent BraTS 2021 dataset shows promising results, comparable to state-of-the-art models (see Table 5). Specifically, our model achieves robust performance in segmenting the tumor core, with a DSC of 89.54%. This surpasses many of the other models listed. However, it's notable that Hatamizadeh *et al.* [52] displayed competitive DSC values, especially for WT and ET categories. Our model holds an advantage in terms of fewer parameters, making it computationally more efficient. The patterns observed in this dataset align closely with those from BraTS 2020, suggesting consistent performance.

Our model's DSC for the whole tumor and the enhanced tumor stand at 90.29% and 83.92%, respectively. When we mention these scores are in line with leading models, we specifically refer to Hatamizadeh *et al.* [52], Jia *et al.* [50], and Ma *et al.* [9], as shown in Table 5.

Considering HD95 values, our model consistently demonstrates accuracy in determining tumor boundaries. For the whole tumor, tumor core, and enhanced tumor, the HD95 scores are 3.03 mm, 2.44 mm, and 3.06 mm, respectively. These figures, especially for the enhanced tumor, indicate that our model's predictions closely align with actual tumor boundaries.

A significant advantage of the LATUP-Net model is its efficiency. With only 3.07 M parameters, it is much lighter than Hatamizadeh *et al.*'s model [52] with 61.98 M and Jia *et al.*'s model [50] with 17.91 M. For models without a specified parameter count in the table, while we do not have the exact numbers, it is commonly understood in the field that such models typically have a considerable number of parameters due to their complex architectures.

In conclusion, the LATUP-Net model presents a fine balance between performance and efficiency, demonstrating the potential for practical clinical applications where both, accuracy and computational efficiency, are paramount.

6. Conclusion and Future Directions

In this work, we have unveiled the LATUP-Net network, an enhanced U-Net variant for 3D brain tumor segmentation designed to be lightweight in its computational demand. This model substantially decreases the number of parameters needed while maintaining, and in some aspects surpassing, the segmentation performance of state-of-the-art methods. With 3.07 M parameters, about 59 times fewer parameters than the state-of-the-art nnU-Net with 181.56 M parameters, LATUP-Net underscores an advancement where efficient modeling coupled with parallel convolutions can lead to a significant reduction in overfitting risk and more judicious use of computational resources.

Our model demonstrates an impressive ability to delineate tumor boundaries with high accuracy, as evidenced by its performance in Hausdorff distance (HD95) measurements. These achievements indicate the model's potential

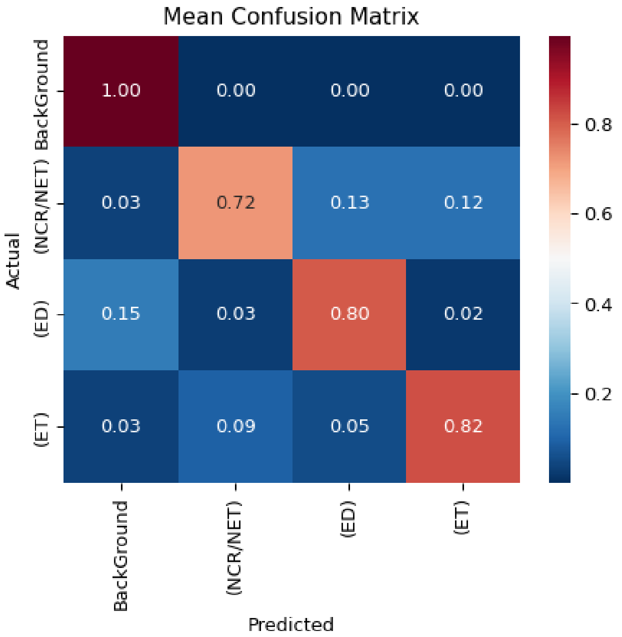


Figure 7: Confusion matrix illustrating the pixel-wise misclassification rates between different tumor regions, normalized by the number of actual class instances in each row and predicted class. These values are averaged over all samples in the dataset to provide a comprehensive overview of the LATUP-Net model's performance.

utility in clinical settings, where precise segmentations are integral to formulating effective treatment plans.

A pivotal aspect of our research is incorporating attention mechanisms, which refine our model's capability to focus on salient features within MRI scans. Our comparative analysis across different attention mechanisms, such as SE, CBAM, and ECA, reveals that while all contribute to accuracy improvements, SE provides a balance between performance and parameter efficiency, particularly in delineating enhanced tumors. However, the enhancements brought by attention are found to be nuanced. The slight underperformance in Dice score coefficients for enhancing tumor segmentation suggests that attention mechanisms do not unilaterally enhance performance across all regions. This is corroborated by gradient-weighted class activation mapping (Grad-CAM) and confusion matrix analyses. These investigations highlight scenarios where attention mechanisms seem to focus too narrowly on local features, occasionally at the expense of contextual understanding, leading to potential misclassification between regions with texturally similar features. The attention-enhanced model, while showing promise in segmenting small regions, also illustrates that there are instances where traditional convolutions may suffice and that the features they capture can be integral to achieving precise segmentations.

The LATUP-Net model stands as a testament to the possibility of achieving state-of-the-art performance with a fraction of the computational cost, highlighting a promising direction for medical image analysis research and the development of practical, accessible tools for brain tumor segmentation. Future work includes adapting the architecture to other medical imaging segmentation tasks, and refining the balance between attention and convolutional features, especially to enhance our model's sensitivity to enhancing tumor regions and to reduce the variations in segmentation performance across different regions.

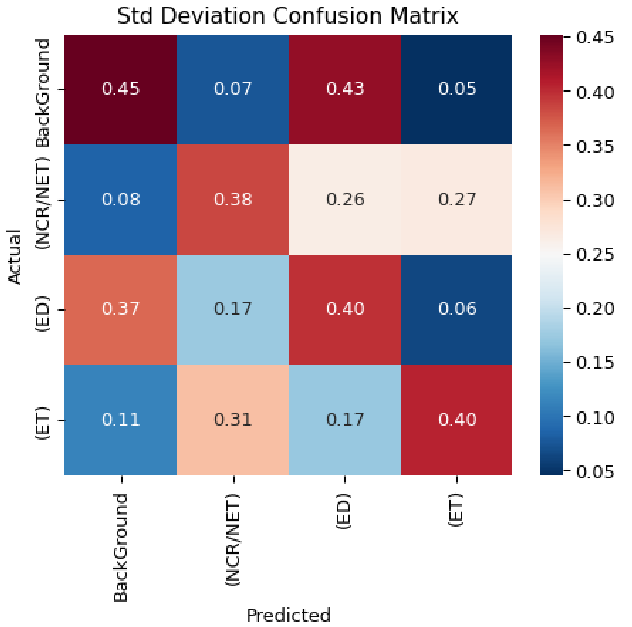


Figure 8: Confusion matrix of the standard deviation of misclassification rates for each tumor region from Figure 7 of the LATUP-Net model.

CRediT authorship contribution statement

Ebtihal J. Alwadee: Conceptualization, Methodology, Software, Investigation, Writing - Original Draft. **Xianfang Sun:** Supervision, Writing - Review & Editing. **Yipeng Qin:** Supervision, Writing - Review & Editing. **Frank C. Langbein:** Software, Data Curation, Supervision, Writing - Review & Editing.

References

[1] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Transactions on Medical Imaging 34 (2014) 1993–2024.

[2] A. Işın, C. Direkoğlu, M. Şah, Review of MRI-based brain tumor image segmentation using deep learning methods, Procedia Computer Science 102 (2016) 317–324.

[3] M. Wieczorek, J. Siłka, M. Woźniak, S. Garg, M. M. Hassan, Lightweight convolutional neural network model for human face detection in risk situations, IEEE Transactions on Industrial Informatics 18 (2021) 4820–4829.

[4] M. Woźniak, J. Siłka, M. Wieczorek, Deep neural network correlation learning mechanism for ct brain tumor detection, Neural Computing and Applications (2021) 1–16.

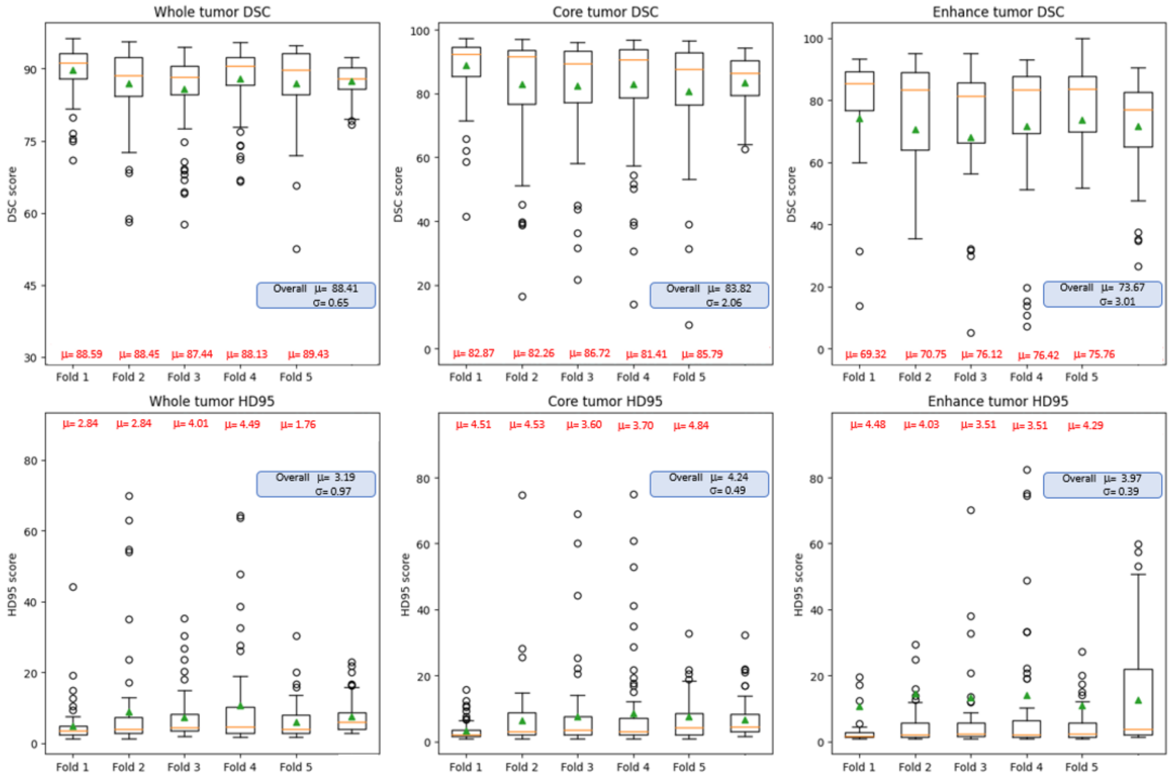


Figure 9: Boxplots of the DSC and HD95 metrics measured per sample (patient) on the BraTS 2020 five-fold cross-validation results with mean μ and standard deviation σ. The orange line within each boxplot represents the median of the data. The green triangles represent the mean, and the circles denote the outliers. We also show the average distributions over all five folds.

[5] R. V. K. D R Sarvamangala, Convolutional neural networks in medical image understanding: a survey, Evolutionary Intelligence 15 (2022) 1–22.

[6] J. Beutel, Handbook of Medical Imaging, volume 3, Spie Press, 2000.

[7] T. Ma, K. Wang, F. Hu, Lmu-net: lightweight u-shaped network for medical image segmentation, Medical & biological engineering & computing (2023) 1–10.

[8] H. Li, A. Li, M. Wang, A novel end-to-end brain tumor segmentation method using improved fully convolutional networks, Computers in Biology and Medicine 108 (2019) 150–160.

[9] J. Ma, J. Chen, Nnunet with region-based training and loss ensembles for brain tumor segmentation, in: International MICCAI Brainlesion Workshop, Springer, 2021, pp. 421–430.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[11] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: learning dense volumetric segmentation from sparse annotation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI: 19th International Conference, Proceedings, Part II 19, Springer, 2016, pp. 424–432.

[12] W. Chen, B. Liu, S. Peng, J. Sun, X. Qiao, S3D-UNet: separable 3D U-Net for brain tumor segmentation, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, Part II 4, Springer, 2019, pp. 358–368.

[13] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, J. Li, TransBTS: Multimodal brain tumor segmentation using transformer, in: Medical Image Computing and Computer Assisted Intervention–MICCAI: 24th International Conference, Proceedings, Part I 24, Springer, 2021, pp. 109–119.

[14] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI: 18th International Conference, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[15] Y. Liu, X. Zhang, G. Cai, Y. Chen, Z. Yun, Q. Feng, W. Yang, Automatic delineation of ribs and clavicles in chest radiographs using fully convolutional DenseNets, Computer Methods and Programs in Biomedicine 180 (2019) 105014.

[16] S. L. Oh, E. Y. Ng, R. San Tan, U. R. Acharya, Automated beat-wise arrhythmia diagnosis using modified U-Net on extended electrocardiographic recordings with heterogeneous arrhythmia types, Computers in Biology and Medicine 105 (2019) 92–101.

[17] Z. Liu, Y.-Q. Song, V. S. Sheng, L. Wang, R. Jiang, X. Zhang, D. Yuan, Liver CT sequence segmentation based with improved U-Net and graph cut, Expert Systems with Applications 126 (2019) 54–63.

[18] Z. Zhang, C. Wu, S. Coleman, D. Kerr, DENSE-INception U-Net for medical image segmentation, Computer Methods and Programs in Biomedicine 192 (2020) 105395.

[19] C. Chen, X. Liu, M. Ding, J. Zheng, J. Li, 3D dilated multi-fiber network for real-time brain tumor segmentation in MRI, in: Medical Image Computing and Computer Assisted Intervention–MICCAI: 22nd International Conference, Proceedings, Part III 22, Springer, 2019, pp. 184–192.

[20] Z. Luo, Z. Jia, Z. Yuan, J. Peng, HDC-Net: hierarchical decoupled convolution network for brain tumor segmentation, IEEE Journal of Biomedical and Health Informatics 25 (2020) 737–745.

[21] T. Magadza, S. Viriri, Brain tumor segmentation using partial depthwise separable convolutions, IEEE Access 10 (2022) 124206–124216.

[22] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

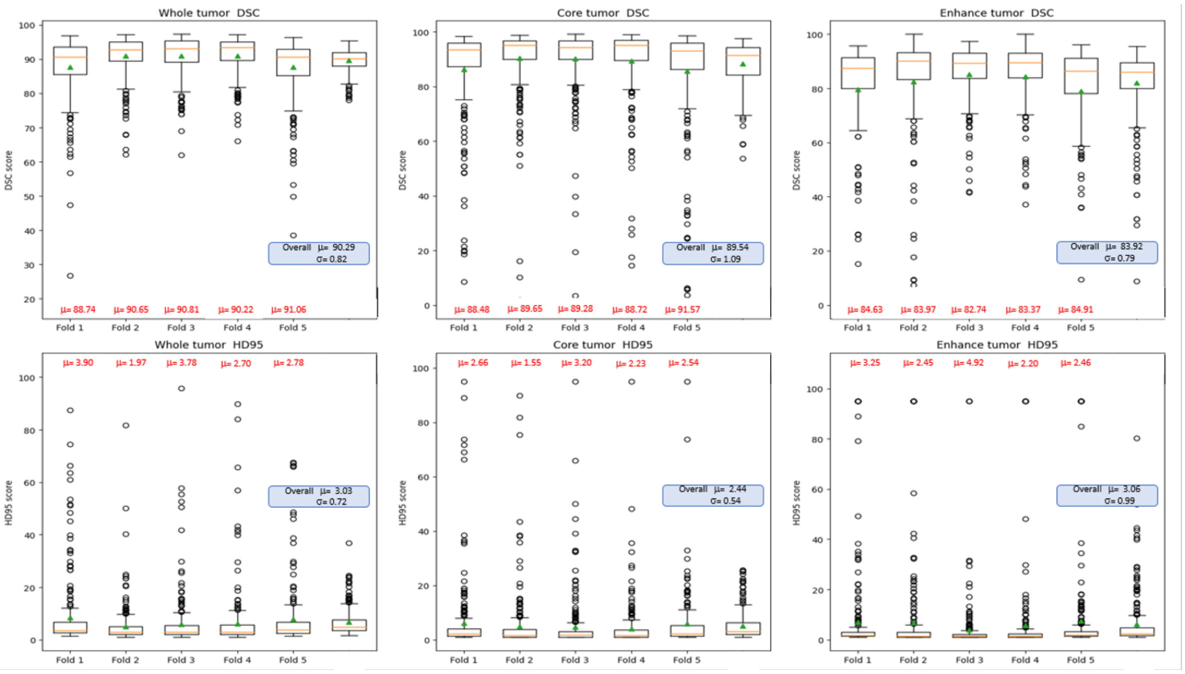


Figure 10: Boxplots of the DSC and HD95 metrics measured per sample (patient) on the BraTS 2021 five-fold cross-validation results with mean μ and standard deviation σ. The orange line within each boxplot represents the median of the data. The green triangles represent the mean, and the circles denote the outliers. We also show the average distributions over all five folds.

[23] A. G. Roy, N. Navab, C. Wachinger, Concurrent spatial and channel ‘squeeze & excitation’in fully convolutional networks, in: Medical Image Computing and Computer Assisted Intervention–MICCAI: 21st International Conference, Proceedings, Part I, Springer, 2018, pp. 421–429.

[24] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, CBAM: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

[25] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11534–11542.

[26] J. Gu, X. Sun, Y. Zhang, K. Fu, L. Wang, Deep residual squeeze and excitation network for remote sensing image super-resolution, Remote Sensing 11 (2019) 1817.

[27] D. Ulyanov, A. Vedaldi, V. Lempitsky, Instance normalization: The missing ingredient for fast stylization, arXiv:1607.08022 (2016).

[28] K. T. Rajamani, P. Rani, H. Siebert, R. ElagiriRamalingam, M. P. Heinrich, Attention-augmented u-net (aa-u-net) for semantic segmentation, Signal, image and video processing 17 (2023) 981–989.

[29] X. Chen, Research on algorithm and application of deep learning based on convolutional neural network, Zhejiang Gongshang University (2014).

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[31] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, C. Davatzikos, Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features, Scientific data 4 (2017) 1–13.

[32] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, et al., Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge, arXiv:1811.02629 (2018).

[33] S. Raschka, Model evaluation, model selection, and algorithm selection in machine learning, arXiv:1811.12808 (2018).

[34] S. Patro, K. K. Sahu, Normalization: A preprocessing stage, arXiv:1503.06462 (2015).

[35] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980 (2014).

[36] E. Alwadee, F. C. Langbein, Bca - brain cancer segmentation python package, version 1.0, software, 2024. URL: `https://qyber.black/ca/code-bca`.

[37] E. Alwadee, F. C. Langbein, BCa segmentation results: LATUPNet, version 1.0. software and data, 2024. URL: `https://qyber.black/ca/results-bca-latup`.

[38] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, C. Davatzikos, Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection, The Cancer Imaging Archive 286 (2017).

[39] S. A. Taghanaki, Y. Zheng, S. K. Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, G. Hamarneh, Combo loss: Handling input and output imbalance in multi-organ segmentation, Computerized Medical Imaging and Graphics 75 (2019) 24–33.

[40] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (2018) 318–327.

[41] A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, Enet: A deep neural network architecture for real-time semantic segmentation, arXiv:1606.02147 (2016).

[42] T. Vo, P. Dave, G. Bajpai, R. Kashef, N. Khan, Brain tumor segmentation in mri images using a modified u-net model, in: 2022 IEEE International Conference on Digital Health (ICDH), 2022, pp. 29–33.

[43] R. Raza, U. I. Bajwa, Y. Mehmood, M. W. Anwar, M. H. Jamal, dResU-Net: 3d deep residual U-Net based brain tumor segmentation from multimodal mri, Biomedical Signal Processing and Control 79 (2023) 103861.

[44] L. M. Ballestar, V. Vilaplana, Brain tumor segmentation using 3D-CNNs with uncertainty estimation, arXiv:2009.12188 (2020).

[45] H. Messaoudi, A. Belaid, M. L. Allaoui, A. Zetout, M. S. Allili, S. Tliba, D. Ben Salem, P.-H. Conze, Efficient embedding network for 3D brain tumor segmentation, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Part I 6, Springer, 2021, pp. 252–262.

[46] F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth, K. H. Maier-Hein, nnU-Net for brain tumor segmentation, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Part II 6, Springer, 2021, pp. 118–132.

[47] J. Tang, T. Li, H. Shu, H. Zhu, Variational-autoencoder regularized 3D MultiResUNet for the BraTS 2020 brain tumor segmentation, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Part II 6, Springer, 2021, pp. 431–440.

[48] H. Peiris, Z. Chen, G. Egan, M. Harandi, Reciprocal adversarial learning for brain tumor segmentation: a solution to brats challenge 2021 segmentation task, in: International MICCAI Brainlesion Workshop, Springer, 2021, pp. 171–181.

[49] A. S. Akbar, C. Fatichah, N. Suciati, Unet3D with multiple atrous convolutions attention block for brain tumor segmentation, in: International MICCAI Brainlesion Workshop, Springer, 2021, pp. 182–193.

[50] H. Jia, C. Bai, W. Cai, H. Huang, Y. Xia, HNF-NetV2 for brain tumor segmentation using multi-modal mr imaging, in: International MICCAI Brainlesion Workshop, Springer, 2021, pp. 106–115.

[51] Z. Li, Z. Shen, J. Wen, T. He, L. Pan, Automatic brain tumor segmentation using multi-scale features and attention mechanism, in: International MICCAI Brainlesion Workshop, Springer, 2021, pp. 216–226.

[52] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, D. Xu, Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, in: International MICCAI Brainlesion Workshop, Springer, 2021, pp. 272–284.

[53] J. Roth, J. Keller, S. Franke, T. Neumuth, D. Schneider, Multi-plane unet++ ensemble for glioblastoma segmentation, in: International MICCAI Brainlesion Workshop, Springer, 2021, pp. 285–294.

Ebtihal J. Alwadee received her BSc (Hons) in Information Systems from King Khalid University, Saudi Arabia, in 2013 and her MSc in Computer Science from California State University, Fullerton, USA, in 2020. She is currently a PhD candidate at Cardiff University, with research interests in visual computing, healthcare, deep learning, medical image segmentation, and explainable AI.

Yipeng Qin received a BSc degree in electrical engineering from Shanghai Jiao Tong University, China, and a PhD degree from the National Centre for Computer Animation (NCCA), Bournemouth University, UK. He was a Postdoctoral Research Fellow with the Visual Computing Center (VCC), King Abdullah University of Science and Technology (KAUST), Saudi Arabia. He is currently a Lecturer at the School of Computer Science and Informatics, Cardiff University, UK. His current research interests include deep learning, computer vision, computer graphics, and human–computer interaction (HCI), with a focus on generative modeling and visual content creation.

Xianfang Sun received a PhD degree from the Institute of Automation, Chinese Academy of Sciences, in 1994. He is currently a Senior Lecturer at the School of Computer Science, Cardiff University, UK. His main research interests include computer vision, computer graphics, pattern recognition, and artificial intelligence.

Frank C. Langbein received his Mathematics degree from Stuttgart University, Germany, in 1998 and a PhD from Cardiff University, UK, in 2003. He is currently a senior lecturer at the School of Computer Science and Informatics, Cardiff University, where he is a member of the Visual Computing Research Section and leads the Quantum Control Research Group. He co-leads Qyber, a research network in quantum control and computational intelligence. His research interests include control, machine learning, and geometry applied in quantum technologies, visual computing, geometric modeling, and healthcare. He is a member of the Institute of Electrical and Electronics Engineers (IEEE) and the American Mathematical Society (AMS).