Computers in Biology and Medicine 186 (2025) 109662

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

VcaNet: Vision Transformer with fusion channel and spatial attention module for 3D brain tumor segmentation

Dichao Pan [a], Jianguo Shen [a,b,*], Zaid Al-Huda [c], Mohammed A.A. Al-qaness [a,b]

[a] College of Physics and Electronic Information Engineering, Zhejiang Normal University, Jinhua, 321004, China
[b] Zhejiang Institute of Optoelectronics, Jinhua, 321004, China
[c] Stirling College, Chengdu University, Chengdu, Sichuan, 610106, China

ARTICLE INFO

Keywords:
Brain tumor segmentation
Vision Transformer
Convolutional neural networks
Multi-scale feature fusion
Channel and spatial attention module

ABSTRACT

Accurate segmentation of brain tumors from MRI scans is a critical task in medical image analysis, yet it remains challenging due to the complex and variable nature of tumor shapes and sizes. Traditional convolutional neural networks (CNNs), while effective for local feature extraction, struggle to capture long-range dependencies crucial for 3D medical image analysis. To address these limitations, this paper presents VcaNet, a novel architecture that integrates a Vision Transformer (ViT) with a fusion channel and spatial attention module (CBAM), aimed at enhancing 3D brain tumor segmentation. The encoder of VcaNet employs a 3D enhanced convolution (ENCO) module to capture local volumetric features, while a Vision Transformer and multi-scale feature fusion module are incorporated in the bottleneck to capture global dependencies. Additionally, a CBAM is introduced in the decoder to further improve the integration of local and global features, enhancing segmentation accuracy. Extensive experiments on the two public BraTS Datasets demonstrate that VcaNet outperforms existing models, particularly in handling the complex spatial structures of brain tumors. This approach provides valuable insights for improving brain tumor segmentation, and its performance in 3D tasks surpasses that of 2D models, laying a foundation for future advancements in medical imaging.

1. Introduction

Brain tumors, characterized by their heterogeneous nature and high morbidity and mortality rates, pose a significant challenge to patient health [1]. Once developed, they cause increased intracranial pressure and compression of brain tissues, which in turn triggers severe central nervous system damage, resulting in seizures, poor memory, or physical dysfunction in patients. If the tumor can be detected early before it turns into a malignant tumor, the survival cycle of the patient can be significantly improved, and the cost of treatment can be reduced [2]. Traditionally, neurosurgeons relied on manual analysis of brain lesion images to determine tumor characteristics and surgical planning, a process that was time-consuming and prone to subjective errors. With the rapid advancements in artificial intelligence (AI) and machine learning (ML), automated approaches to tumor diagnosis and surgical planning are becoming increasingly important. Machine learning methods, particularly deep learning, have proven to be highly effective in medical imaging applications, offering improvements in both accuracy and efficiency. Techniques such as voxel-based analysis and predictive modeling have been employed to study tumor progression and assist in preoperative planning [3]. The application of machine learning in brain tumor segmentation has led to more accurate and reproducible results compared to traditional methods, which are often constrained by human limitations.

Among machine learning models, Convolutional Neural Networks (CNNs) have emerged as leading algorithms for image content analysis, showing exceptional performance in classification, segmentation, and detection tasks. U-Net [4] uses an encoder-decoder architecture with a skip connection, which is capable of making high-resolution pixel-level predictions while preserving image details, and high-precision pixel-level prediction, which is particularly suitable for medical image segmentation tasks. Subsequent research around U-Net and its unique encoder-decoder architecture has resulted in the derivation of several U-Net variants, including U-Net++ [5], and V-Net [6], which further enhance image segmentation performance. Despite the impressive performance of CNN-based segmentation methods before, the performance of the convolutional kernel is slightly weak in the extraction of long-range and global contextual information due to its limited effective

* Corresponding author. College of Physics and Electronic Information Engineering, Zhejiang Normal University, Jinhua, 321004, China.
E-mail addresses: pandc561@zjnu.edu.cn (D. Pan), shenjianguo@zjnu.cn (J. Shen), zaid@stir.ed.cn (Z. Al-Huda), alqaness@zjnu.edu.cn (M.A.A. Al-qaness).

https://doi.org/10.1016/j.compbiomed.2025.109662
Received 28 September 2024; Received in revised form 4 January 2025; Accepted 6 January 2025
Available online 14 January 2025
0010-4825/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

receptive field. For brain tumors with varying shapes and sizes, global semantic information is crucial, and the limitations of convolutional operations pose new challenges to researchers.

The attention mechanism, initially developed for natural language processing [7], has inspired significant advancements in image segmentation and classification. Many current studies have successfully integrated attention mechanisms into existing CNN models, enhancing their capabilities. The Attention UNet [8] incorporates an attention gate module at the end of each layer of jump connections, which reduces redundant jump connections and enhances the extraction of image feature details.

The Transformer [9] model has revolutionized natural language processing and other fields with its powerful attention mechanism and flexible architecture. Originally developed for modeling long-range dependencies in sequence-to-sequence tasks, the Transformer effectively captures relationships between arbitrary elements within an input sequence, addressing the inherent limitations of RNN [10] and CNN models, and significantly enhancing performance in natural language processing. The Vision Transformer (ViT) [11], which adapts the Transformer architecture for computer vision applications, has demonstrated its capability in handling complex visual tasks. ViT innovatively divides the input image into small blocks (patches), each of which is a fixed-size image region. These image blocks are linearly mapped into vectors, which serve as input sequences to the Transformer encoder. These vectors are modeled to learn the global correlations between the image chunks, resulting in better performance than CNN. For example, TransUnet [12] is an excellent work on medical image segmentation using ViT modules in recent years, which combines the respective advantages of Unet and Transformer by introducing a multi-layer ViT module in the encoder part for modeling the global context information in the image, while retaining the efficient spatial recovery capability of U-Net in the decoder. This approach achieves a balance between segmentation performance and computational efficiency. However, the stacking of modules in multiple layers and the need to use weights that have been pre-trained on large-scale datasets pose challenges to the number of covariates of the model and the migration of the ViT modules.

This paper proposes **VcaNet**, a novel method that leverages the strengths of 3D CNNs and Vision Transformers for 3D brain tumor segmentation. This method uses the traditional encoder-decoder architecture, we add one more layer to the original structure layer, and utilize the designed ENCO (Enhanced Convolution) module to perform multiple convolution downsampling on the input 3D image to enhance the extraction of detailed image features, so that the model can obtain rich local context information and connect to the decoder with the same resolution through jump connections. decoder with the same resolution. The decoder still uses the ENCO module and a channel and spatial self-attention module (CBAM) [13] module is added between the decoder and the upsampling, which consists of a channel and spatial attention mechanism that adaptively extracts and enhances the most important channel and spatial features in the image, and inputs into the CBAM module after fusing the upsampling output features with the encoder features passed over the hopping connection to enhance the cross-scale feature representation to make the model more focused on important regions. After experiments, it is found that the resulting method is not sufficient for detailed feature extraction. After comparing the experiments and the inspiration provided by the TransBTS method, we propose an MSCTrans (Multiscale Feature Extraction Transformer) module in the bottleneck layer section at the bottom of the model, which combines the Multiscale Feature Extraction (MSC) module combined with the ViT module. Deeply separable convolutions with three different kernel sizes are first used to fuse features at different scales. The resulting feature map is then divided into image patches, which are linearly mapped into individual vectors (markers) and fed into the ViT module for global feature modeling. The implemented ViT module eliminates the need to pre-train the weights on large-scale datasets and can be learned from scratch with good segmentation results. Compared to other existing medical image segmentation models incorporating ViT, VcaNet has better accuracy while ensuring no excessive computational complexity, and achieves better segmentation results in all regions of brain tumors. Finally, extensive experiments on the BraTS 2020 and 2021 datasets validate the effectiveness of VcaNet.

This work presents several key contributions as follows.

1. A novel image segmentation algorithm is proposed, utilizing an encoder-decoder architecture integrated with the Vision Transformer.
2. We redesign the bottleneck layer and propose an MSCTrans module, which combines the local feature extraction of the MSC module with the advantages of the ViT module in global context modeling. The proposed ViT module can be trained from scratch without the need of pre-training the weights and achieves good results.
3. We design the ENCO module and introduce the CBAM module in the downsampling and upsampling phases, which enhances the extraction of detailed features and fuses the multidimensional features through skip connections.
4. The proposed VcaNet achieves superior segmentation performance on the BraTS2020 and BraTS2021 datasets when compared to other CNN-based and Transformer models.

The remainder of the paper is structured as follows: Section II provides a comprehensive overview of existing medical image segmentation methods, including those employing 3D CNNs and Transformers. Section III describes the overall architecture and each component of our model in detail. Section IV describes the datasets, experimental environments, and hyper-parameters used for the experiments, analyzes and compares the results of the individual experiments, and finally summarizes them at the end of this paper.

2. Related work

This section reviews existing approaches to medical image segmentation, focusing on both convolution-based and transformer-based methods, including 3D CNN models like 3D U-Net and V-Net, and transformer-based models like TransBTS and Swin-UNet.

2.1. Convolution-based works

3D convolutional neural networks (CNNs) offer a significant advantage over traditional 2D CNNs by capturing feature information from an image in all three dimensions simultaneously. This capability enables a more comprehensive understanding and segmentation of complex medical image structures. 3D CNNs have found widespread application in various medical imaging domains, including brain MRI, CT scans, and ultrasound images. 3D U-Net [14] was the first model to propose the use of 3D CNNs for medical image segmentation, and it's a 3D version of U-Net, which performed well on the BraTS dataset. Following this, V-Net was introduced in the same year, utilizing a novel Dice loss function to address the common issue of imbalanced samples in medical images, specifically the disproportion between foreground and background voxels.

In 2018, the Attention U-Net was proposed, integrating an attention mechanism into the 3D U-Net. This design includes an attention gate in the upsampling stage, which adaptively focuses on relevant regions, thereby improving segmentation performance. 3D ResUNet [15] combines the residual network and U-Net structures. By introducing residual connections, 3D ResUNet effectively addresses the gradient vanishing problem common in deep networks while preserving model depth, resulting in excellent performance in segmenting brain tumor CT images. AGU-Net [16] proposes a method to introduce an attention-gate mechanism in the network, which allows the feature fusion to focus on important regions adaptively, thus improving the ability to handle multi-scale and multimodal features in medical images. nnU-Net [17],

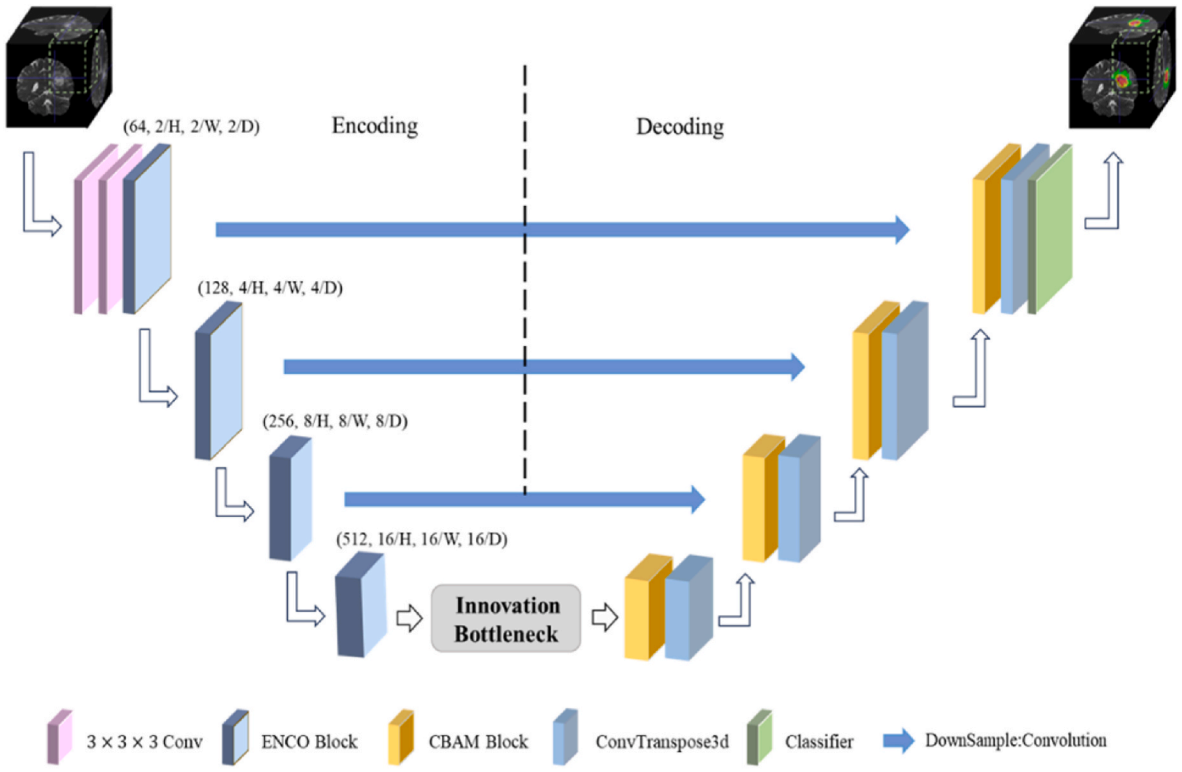


Fig. 1. The overall architecture of the VcaNet.

proposed by Isensee et al., innovatively automates the adjustment of network architecture, hyperparameters, and preprocessing steps based on the specific dataset. This approach has achieved state-of-the-art (SOTA) results in numerous medical image segmentation tasks. Zhu et al. [18] proposed a novel method using multi-modality spatial information enhancement and boundary shape correction to improve segmentation accuracy and boundary delineation.

2.2. Transformer-based works

In recent years, deep learning models combining Transformer and CNN have witnessed significant advancements in medical image segmentation, which can capture both local and global contextual information simultaneously. TransBTS [19] is the first attempt to embed a Transformer into the bottleneck layer of the encoding-decoding structure, and at the same time, combines with 3D CNN in both spatial and depth dimensions for local and global features modeling and performs well on the MRI 3D brain tumor image segmentation. TransBTSv2 [20] is an improved version of TransBTS, which mainly incorporates some optimization techniques based on the original Transformer module, such as the improvement of the multi-head attention mechanism, layer normalization, etc. Additionally introduces a deformable convolution at the jump junction to better capture image information at different scales. BiTr-Unet [21] distinguishes itself from TransBTS by incorporating two ViT layers into the skip connection portion, further enhancing global feature modeling. Zhu et al. [22] proposed a model that combines deep semantic features with edge information using a Swin Transformer [23] and CNN-based edge detection, effectively enhancing multimodal MRI information fusion through graph convolution for improved segmentation accuracy. SwinBTS [24], on the other hand, combines the strengths of Swin-Transformer and TransBTS, employing the Swin Transformer as both encoder and decoder, which continuously captures the local and global context information of the image during downsampling and upsampling. LMIS [25] is a lightweight segmentation network that integrates a novel backbone feature extraction and a multi-scale feature interaction guidance framework, significantly reducing computational complexity while maintaining high segmentation accuracy. UNETR [26] adopts a different approach, employing ViT to act as an encoder and a convolutional layer to act as a decoder to construct the network, which significantly improves the accuracy and robustness of medical image segmentation.

However, the use of numerous ViT layers results in a large number of model parameters and high hardware requirements. Hatamizadeh et al. proposed a new network model, Swin UNETR [27], based on the UNETR model by using Swin Transformer blocks instead of ordinary ViT blocks. By leveraging the Swin Transformer's moving-window self-attention computation mechanism, Swin UNETR achieves significant performance gains while reducing the number of parameters and computational requirements. VT-UNet [28] presents a lightweight model that incorporates a self-attention layer in the encoder to capture both local and global information. In the decoder, a window-based self-attention module, cross-attention module, and Fourier coding enable the model to focus on specific regions during training. WS-MTST [29] is the first end-to-end weakly supervised segmentation model proposed for brain tumor segmentation, eliminating the need for labor-intensive pixel-level annotations. Instead, it relies on a well-designed loss function and a contrastive learning pre-training process to accurately segment brain tumor subregions.

3. Methodology

In this section, we describe the methodology utilized to develop the VcaNet model for 3D brain tumor segmentation. The architecture is built on a robust encoder-decoder structure, leveraging both local feature extraction and global context modeling to enhance segmentation accuracy.

3.1. Overall architecture

The overall architecture of the VcaNet is illustrated in Fig. 1.

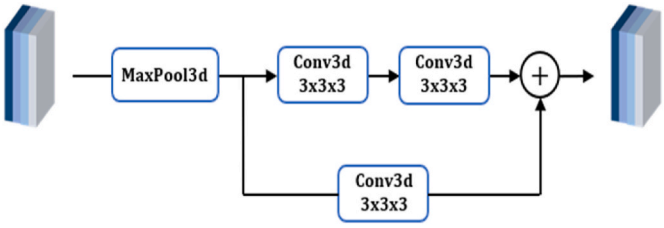


Fig. 2. The architecture of the ENCO Block.

Specifically, it takes as input an MRI scan image $X \in R^{C\times H\times W\times D}$, where $H \times W$ represents the spatial resolution, D is the depth dimension (number of slices), and C denotes the number of channels (modalities). Given the proven effectiveness of encoder-decoder architectures in medical image segmentation tasks, VcaNet adopts this structure as its foundation.

First, we utilize the designed Enhanced Conv (ENCO) module to gradually extract the local features from the input image during the downsampling process, generating a compact feature map containing rich spatial and depth information, followed by a multiscale feature extraction module that fuses the features from different scales through depth-separable convolutions with different convolution kernel sizes. A ViT encoder then processes the extracted features to model long-range dependencies in the global space. Afterward, we iteratively overlay the upsampling and deconvolution layers and introduce the Channel and Spatial Attention Mechanism Module (CBAM) in the upsampling layer to fuse the features from the downsampled portion passed by the jump connections, gradually producing high-resolution segmentation prediction maps. The following sections provide a detailed description of each component within VcaNet.

3.2. Network encoder

Due to the computational complexity of the traditional Transformer scales quadratically with the number of tokens (i.e., the sequence length), flattening the input image into a sequence for the Transformer leads to significantly increased computational demands, which heavily strain both hardware resources and training time. Therefore, ViT is proposed to split the image into image blocks of the same size (16×16) and then reshape each image block into a token so that the length of the sequence will be significantly reduced to 162. However, for 3D volumetric data, directly using ViT as an encoder and splitting the image into multiple image blocks at the data input stage would make ViT unable to perform the computation across the spatial and depth dimensions of the image spatial and depth dimensions to extract local contextual information of the image, thus affecting the model's ability to segment on local feature details. To address this issue, an ENCO module was designed, consisting of two stacked $3 \times 3 \times 3$ convolution blocks (with stride $= 2$) combined with residual concatenation to facilitate feature fusion from low-level to high-level, as shown in Fig. 2.

nnFormer's [30] approach demonstrates the role of convolutional downsampling in improving the model's performance. Convolutional downsampling generates a hierarchical representation that can model object concepts at multiple scales, progressively encoding the input object into a low-resolution/high-level feature representation $X \in R^{C\times\frac{H}{16}\times\frac{W}{16}\times\frac{D}{16}}$ ($C = 512$),i.e., 1/16 of the input dimensions H, W and D. This process involves applying a 3D convolution with a $3 \times 3 \times 3$ kernel and a stride of 2, followed by BatchNorm for normalization and ReLU for activation. The residual connections are then combined to fuse the features after the two convolutional downsampling operations. Following this, the multi-scale feature extraction module further refines the features, which are fused using depth-separable convolutions with varying kernel sizes. Finally, the output is fed into the ViT module to model the global context information.

3.3. Enhanced bottleneck layer design

In the bottleneck layer of the encoder-decoder architecture, a novel MSCTrans block is introduced by integrating the multi-scale feature extraction (MSC) module with the Vision Transformer (ViT) module, as shown in Fig. 3.

This combination improves the model's ability to extract features and capture contextual information at various scales, enhancing the accuracy and robustness of volumetric data segmentation. Firstly, the multi-scale feature extraction module effectively captures feature information at multiple scales by combining depth-separable convolution [31] operations with different receptive fields without significantly increasing the number of model parameters. Specifically, we used, at each encoding stage, three different scales of depth-separable convolution: $1 \times 1 \times 1$, $3 \times 3 \times 3$ and $5 \times 5 \times 5$, for the input feature map $X \in R^{C\times\frac{H}{16}\times\frac{W}{16}\times\frac{D}{16}}$, the outputs obtained after passing them through the three convolution kernels are respectively, as an example, is shown in Eq. (1).

$$X_3 = DSConv_{3\times3\times3}(X), X_5 = DSConv_{5\times5\times5}(X), X_7 = DSConv_{7\times7\times7}(X). \quad (1)$$

To fuse these features efficiently, three learnable weight parameters α、β and γ are introduced. Each weight parameter is initialized to 1, i.e., at the beginning of training, each convolutional layer contributes the same amount. As training proceeds, these three weight parameters are adjusted by backpropagation to optimize the model performance, the

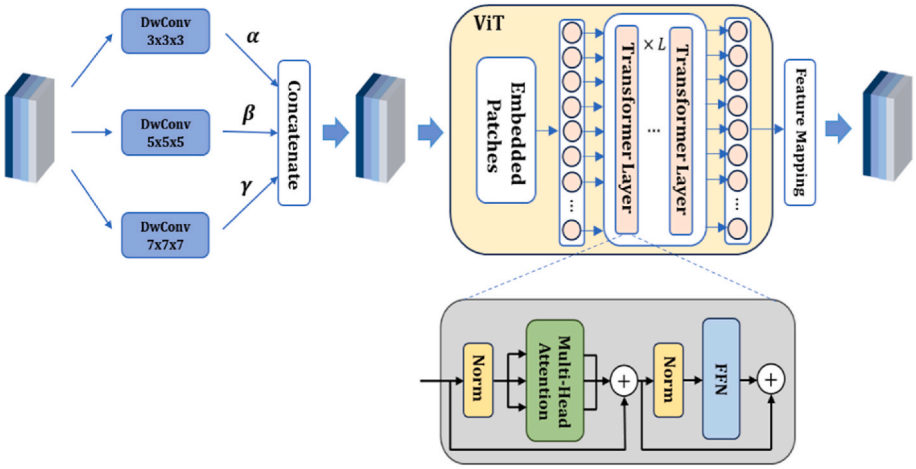


Fig. 3. The architecture of the MSCTrans block.

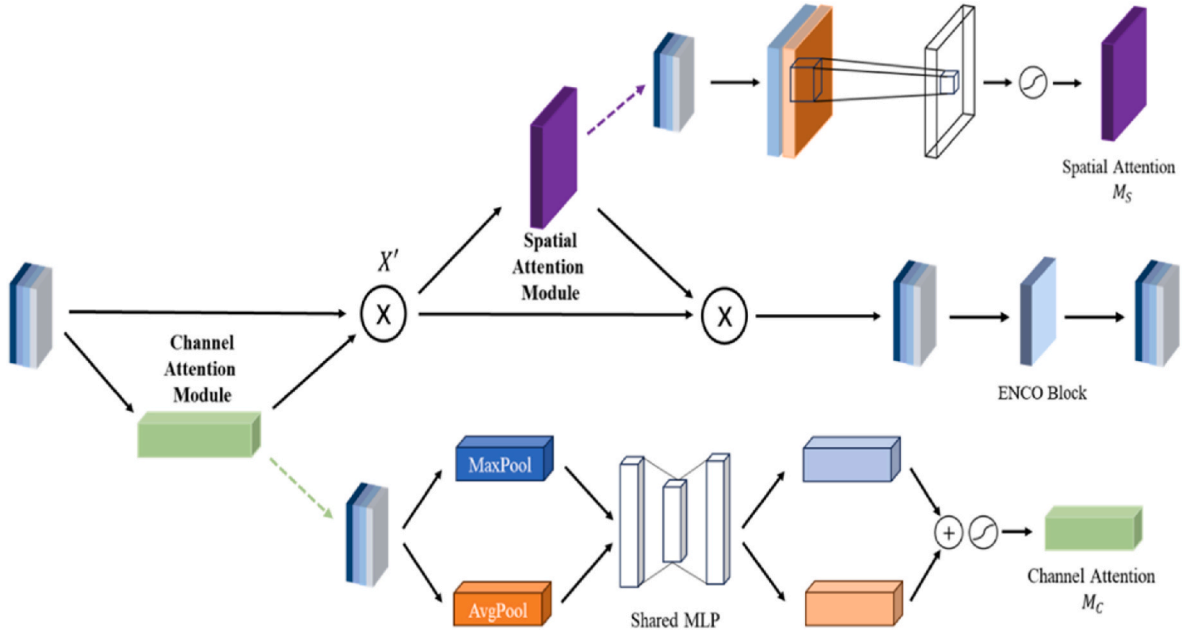


Fig. 4. The architecture of the upsampling block.

weight parameters are normalized to ensure that they sum up to 1. The results are weighted and fused according to the normalized weights, and the feature fusion is performed by the following Eq. (2):

$$X_{multi} = \alpha \cdot X_3 + \beta \cdot X_5 + \gamma \cdot X_7. \tag{2}$$

After a multi-scale feature extraction process, we introduced the Vision Transformer (ViT) module for capturing long-range dependencies in volumetric data. Since the Transformer requires a sequence as input, the ViT module first divides the passed high-dimensional feature map $X_{multi} \in R^{C \times \frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}}$ by a Patch Embedding layer into fixed-size 3D patches and mapping each patch to a high-dimensional embedding space by linear projection, as shown in Eq. (3):

$$Patch(X) = Conv3d(X, kernel_{size} = (p,p,p), stride = (p,p,p)), \tag{3}$$

where p represents the patch size. The obtained embedding sequence $X_{patch} \in \mathbb{R}^{N \times d}$ is then fed into the ViT layers for long-range dependency modeling. The Transformer encoder comprises L layers of a standard Transformer structure. Each layer consists of a Multi-Headed Attention (MHA) and a Feedforward Network (FFN), which is computed as shown in Eqs. (4) and (5):

$$Z_\ell = MHA(LN(Z_{\ell-1})) + Z_{\ell-1}. \tag{4}$$

$$Z_\ell = FFN\left(LN\left(Z_\ell\right)\right) + Z_\ell. \tag{5}$$

There LN represents the layer normalization, and Z_ℓ denotes the output of the Transformer encoder of the $\ell - th$ layer. With the ViT module, VcaNet can capture the contextual information of the volumetric data on the global scale, which enhances the understanding and segmentation of brain tumors with complicated structures.

3.4. Network decoder

Similarly, to restore the image to its original three-dimensional dimensions ($H \times W \times D$) to generate high-resolution segmentation prediction results, we introduced a 3D CNN decoder for feature upsampling and segmentation prediction at pixel-level (as shown on the right side of Fig. 1).

Feature Mapping: A feature mapping module is incorporated into the decoder part to align the input dimensions with the subsequent 3D CNN decoder. The sequence data output from the Vision Transformer (ViT) is projected back into a four-dimensional feature map. Specifically, the output matrix $\mathcal{Z}_L \in \mathbb{R}^{d \times N}$ is reshaped into a feature map of dimensions $d \times \frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$. With this feature mapping module, it is possible to obtain the same feature-mapping $Z \in \mathbb{R}^{C \times \frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}}$ in the same dimensions as the encoder portion X at the beginning of the decoder.

Upsampling with Attention module: Following the reshape operation, a series of cascading upsampling and transposed convolutions are performed on the obtained feature mapping Z to gradually recover the segmentation prediction map at the original input image resolution. Given the VcaNet's application to the brain tumor segmentation task, the final output has a size of $4 \times H \times W \times D$. Additionally, to further enhance the selective and spatial attention capability of the feature expression, the Convolutional block attention module (CBAM) is introduced at this upsampling stage, as shown in Fig. 4.

Features from both the encoder and decoder, passed through skip connections, are fused, and then the CBAM module enhances the feature map with two sub-modules, channel and spatial attention. The first attention submodule uses average and maximum pooling operations on the input feature map in the channel dimension to generate channel-level attention weights, as defined in Eq. (6):

$$\begin{aligned} M_c(X) = {} & \sigma(Conv3d(Relu(Conv3d(AvgPool(X))))) \\ & + \sigma(Conv3d(Relu(Conv3d(MaxPool(X))))), \end{aligned} \tag{6}$$

where σ denotes the Sigmoid function, and $M_c(X)$ denotes channel attention weights. The spatial attention sub-module generates spatial-level attention weights by computing average and maximum pooling in spatial dimension, concatenating the results, and passing them through a convolutional layer, as shown in Eq. (7):

$$M_s(X) = \sigma(Conv3d([AvgPool(X); MaxPool(X)])). \tag{7}$$

Ultimately, the CBAM module enhances the input feature map by Eq. (8):

$$M_{CBAM} = M_c(X) \cdot X \cdot M_s(X). \tag{8}$$

The introduction of the CBAM module after feature fusion allows the model to automatically adjust the important information in the feature map, increasing the ability to recover spatial details and thus improving

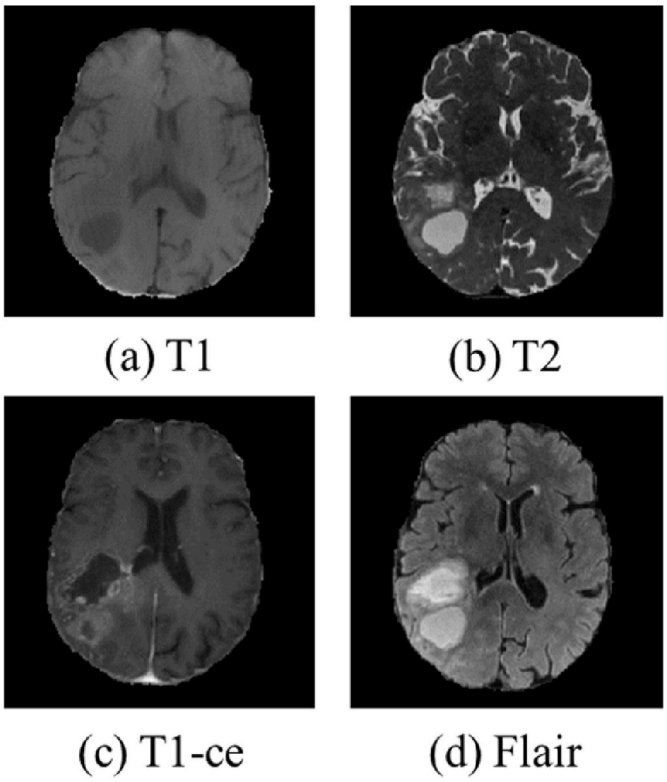


Fig. 5. Visualization of one case from the BraTS2021 training data.

the fineness of the segmentation results.

4. Experiments

In this section, we present a detailed overview of the experimental setup and results to evaluate the effectiveness of the proposed VcaNet model. The experiments were designed to assess both segmentation performance and computational efficiency across multiple datasets and metrics.

4.1. DataSet

The experiment utilize the BraTS 2020 [32] and BraTS 2021 [33] datasets, two brain tumor MRI image datasets provided by the International Multimodal Brain Tumor Segmentation Challenge (BraTS Chanllenge) [34]. It has been successfully held for eleven years, which is a prominent competition in medical image processing. The BraTS dataset is divided into training, validation, and test sets. All tumor regions are manually annotated by multiple experts following a standardized protocol. The official challenge provides labeled data for the training set, while the labels for the validation and test sets are withheld and used exclusively for online performance evaluation of the models. The BraTS 2020 dataset comprises 369 training cases and 125 validation cases [35], and each case contains MRI scans of four modalities, namely T1-weighted (T1), post-contrast T1-weighted (T1-ce), T2-weighted (T2), and Fluid-Attenuated Inversion Recovery (Flair), as shown Fig. 5.

Each modality's volume is $240mm \times 240mm \times 155mm$, and is aligned to the same space. The dataset labels were divided into four: background (Label 0), necrotic and non-enhancing tumor regions (Label 1), peri-tumoral edema (Label 2), and GD-enhancing tumor (Label 4). The purpose of the segmentation task is to segment the enhanced tumor region (ET, Label 4), regions of the tumor core (TC, Label 1, 4), and the whole tumor region (WT, Label 1, 2, 4) from a multimodal brain MRI image. The BraTS 2021 consists of 1251 training cases and 219 validation cases, with the validation set labels of the samples are not disclosed. All data in BraTS 2020 and BraTS 2021 are the same except for the case number.

4.2. Implementation details

VcaNet was implemented using PyTorch and trained from scratch on an NVIDIA GeForce RTX 3090 GPU with 24 GB of memory. The model was trained for 200 epochs with a batch size of 4. Stochastic Gradient Descent (SGD) was employed as the optimizer, and the learning rate was adjusted using a cosine decay strategy, with an initial learning rate of 0.004 and a minimum learning rate of 0.002. The decay was applied over 10 epochs.

The following data augmentation techniques were applied.

1. Volumes were cropped to a fixed size of $160 \times 160 \times 128$ by removing unnecessary background regions.
2. Each image in the dataset is rotated $90°, 180°,$ or $270°$ in a counterclockwise direction, followed by random flipping of the image in the axial, coronal, and sagittal planes.
3. Gaussian noise was added to the input data with the noise variance range $[0, 0.1]$, and the probability of adding noise is set to 0.5.

The training process utilized a combined loss function, incorporating both Dice loss and cross-entropy loss, as described in Eqs. (9)–(11):

$$\mathscr{L}_{dl}(G,Y) = 1 - \frac{2}{J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I} G_{i,j}Y_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} Y_{i,j}^2}, \tag{9}$$

$$\mathscr{L}_{ce}(G,Y) = -\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} G_{i,j}\log Y_{i,j}, \tag{10}$$

$$\mathscr{L}(G,Y) = \mathscr{L}_{dl}(G,Y) + \mathscr{L}_{ce}(G,Y), \tag{11}$$

where I represent the voxels number, J represents the classes number, $Y_{i,j}$, $G_{i,j}$ represent the probabilistic output of the $j-th$ class on the $i-th$ voxel and the ground truth of the unique heat encoding, respectively.

4.3. Evaluation metrics

In this experiment, the evaluation metrics used Dice Score and 95% Hausdorff Distance (HD) given by the official BraTS Challenge. These metrics are employed to assess the accuracy of our model's segmentation performance, which are defined as shown in Eq.(12)–(14):

$$Dice(G,Y) = \frac{2\sum_{i=1}^{I} G_i Y_i}{\sum_{i=1}^{I} G_i + \sum_{i=1}^{I} Y_i}, \tag{12}$$

$$HD95(G',Y') = 0.95 \times max\{hd(G',Y'), hd(Y',G')\}, \tag{13}$$

$$hd(G',Y') = \max_{g'\in G'}\left\{\min_{y'\in Y'}\|g'-y'\|\right\}, hd(Y',G') = \max_{y'\in Y'}\left\{\min_{g'\in G'}\|y'-g'\|\right\}. \tag{14}$$

In the equations, G_i and Y_i represent groundtruth and predicted value at voxel i, respectively, G', Y', respectively denote the set of voxel points in the groundtruth and predicted value segmentation map, g' and y', respectively denote a voxel point in these two value segmentation maps.

4.4. Main results

In this section, we present a detailed evaluation of the proposed VcaNet model, emphasizing its segmentation performance on the BraTS 2020 and BraTS 2021 datasets. The results are compared with various state-of-the-art models to highlight VcaNet's effectiveness in both segmentation accuracy and computational efficiency.

Table 1
Comparison results on the BraTS 2020 test set.

Method	Dice Score(%)				95 % HD (mm)			
	ET	TC	WT	AVG.	ET	TC	WT	AVG.
3D-UNet [14]	70.63 ± 0.284	73.70 ± 0.128	85.84 ± 0250	76.72	34.30	18.86	10.93	21.36
V-Net [6]	68.97	77.90	86.11	77.66	43.52	16.15	14.49	24.72
ResUNet [36]	71.63	76.47	82.46	76.85	37.42	13.11	12.34	20.95
UNETR [26]	71.18 ± 0.297	75.85 ± 0.100	88.30 ± 0.226	78.44	34.46	10.63	8.18	17.75
VT-Unet [28]	76.45 ± 0.267	80.39 ± 0.107	88.73 ± 0.218	81.86	28.99	14.76	9.54	17.76
TransBTS [19]	78.23 ± 0.272	81.73 ± 0.075	88.78 ± 0.174	82.91	29.83	**9.77**	**5.60**	15.06
VcaNet	**78.72 ± 0.227**	**83.35 ± 0.234**	**90.57 ± 0.084**	**84.21**	**15.83**	10.08	10.61	**12.17**

Table 2
Comparison results of model parameter szie and computation complexity on the BraTS 2020 test set.

Method	Dice Score(%)				Param. (M)	GFlops
	ET	TC	WT	AVG.		
3D-UNet [14]	70.63	73.70	85.84	76.72	16.90	586.71
V-Net [6]	68.97	77.90	86.11	77.66	69.30	765.90
Cascaded U-Net [37]	73.60	81.00	90.80	82.80	1173.22	33.80
TransUNet [12]	78.42	78.37	89.46	82.80	105.28	1205.76
TransBTS [19]	78.23	81.73	88.78	82.91	32.99	333.00
UNeXt [38]	76.49	81.37	88.70	82.19	1.47	0.45
VcaNet	**78.72**	**83.35**	**90.57**	**84.21**	79.32	1140.67

4.4.1. *BraTS 2020 dataset*

We divided all the training samples of 369 cases into the training, validation, and test sets according to the ratio of 8:1:1, 295 cases, 37 cases, and 37 cases, respectively. The VcaNet's performance in segmentation was evaluated using the Dice score and 95 % Hausdorff distance metrics provided by the BraTS Challenge. These metrics were used to compare the performance of VcaNet with other excellent segmentation models. On this dataset, the sum of Dice Scores of VcaNet on ET, TC, and WT categories are 78.72%, 83.35%, 90.57%, and the 95% Hausdorff distances are 15.83 mm, 10.08 mm, and 10.61 mm, respectively. To further evaluate the efficiency of VcaNet, we compared it with several state-of-the-art models, all trained on the same dataset division. Table 1 summarizes the segmentation results, where VcaNet demonstrated superior performance in all three tumor categories and showed a significant advantage in the Dice score.

Additionally, Table 2 provides a comparison of model parameter sizes and computational complexity across various models. This table highlights VcaNet's balance between accuracy and computational efficiency by displaying the parameter sizes (in millions) and GFlops. In comparison to traditional methods like 3D-UNet and more recent approaches like TransBTS, VcaNet achieves better Dice scores while maintaining a reasonable model size and computational cost. This is due to the integration of the ViT module into the bottleneck layer, which enhances global context extraction without compromising the preservation of local details.

4.4.2. *BraTS 2021 dataset*

We also evaluated VcaNet on the BraTS2021 dataset, and the comparison can be seen in Table 2. We also divided the training samples of 1251 cases into training, validation, and test sets according to the ratio of 8:1:1 and trained the VcaNet directly with the hyperparameters set on BraTS2020. Table 3 compares the experimental data with other models that perform well on this data.

As a result, VcaNet also achieves great segmentation results on this dataset. The Dice Scores of VcaNet on ET, TC, and WT categories are 83.25 %, 89.52 %, and 92.03 %, and the 95 % Hausdorff distances are 18.17 mm, 6.74 mm, and 4.53 mm, which outperform most of the 3D CNN and Transformer Based Segmentation Methods.

We similarly plotted a histogram of Dice scores on the two segmented test datasets to facilitate a more intuitive comparison of our experimental results, and the five-pointed stars represent the best Dice scores achieved in that region, as shown in Fig. 6, where our VcaNet achieves high-performance segmentation in test samples from both datasets.

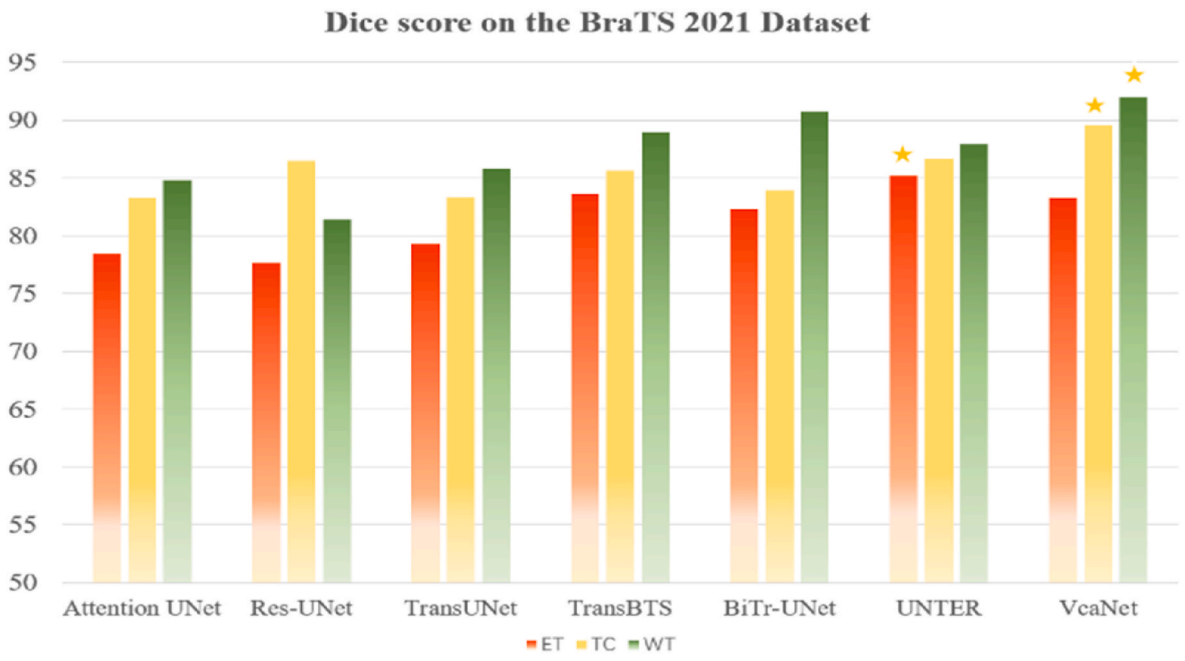
4.4.3. *Visual comparison*

In Fig. 7, we also present a visual comparison of the segmentation outcomes from other techniques, including 3D-UNet, Attention U-Net, TransBTS, and our VcaNet, for qualitative interpretation. The brain tumor's cross-sectional image is displayed in the first row of Fig. 7, its sagittal image is displayed in the second row, and its coronal image is displayed in the third row. To make the experiment easier to observe and compare, we display all three interface images and adjust the cutoff point's coordinates to (141, 150, 96). The figure displays the ET region labeled in red, the TC region labeled in yellow, and the WT region labeled in green. The segmentation in the WT region is the best for all models, according to the experimental data, but there are major variations in the ET and TC sectors. This is because these two regions have more complex edge details, which challenges the models in their ability to extract complex details. Our VcaNet demonstrates a strong segmentation performance in the TC region, and at the same time, in the ET region, it can maintain the same or even higher level of segmentation performance than the other models at the same or even higher level.

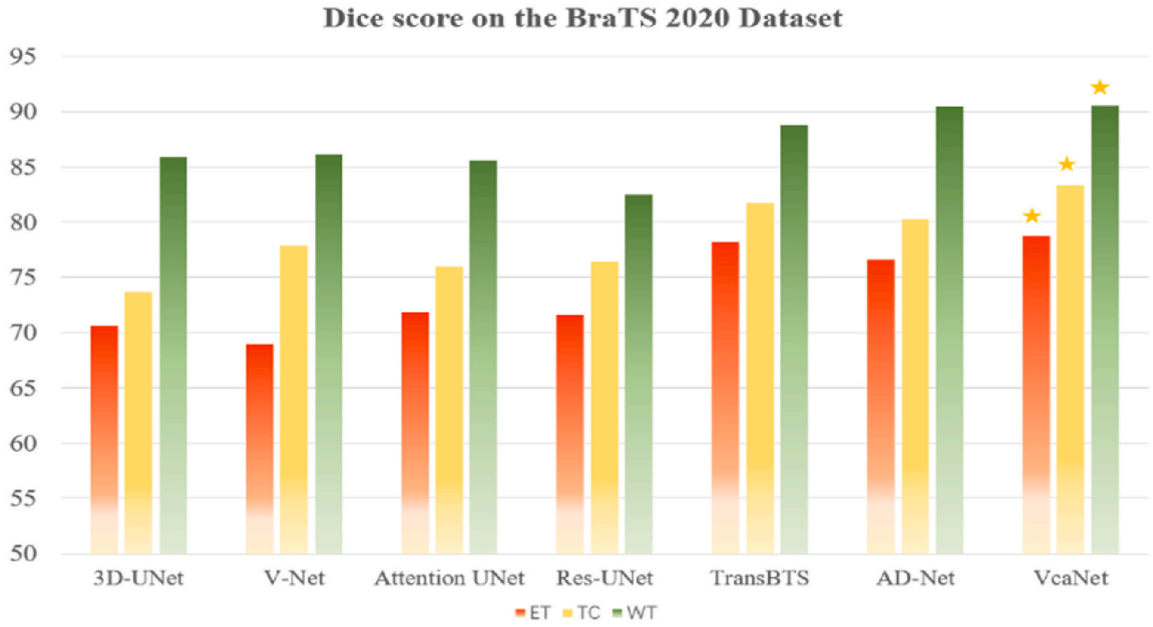
Additionally, shown in Fig. 8, the 3D volume segmentation results generated by VcaNet are compared with the ground truth to evaluate segmentation accuracy. The results demonstrate that the model produces 3D segmentations closely aligned with the ground truth, indicating high accuracy in the segmentation performance.

Table 3
Comparison results on the BraTS 2021 test set.

Method	Dice Score(%)				95 % HD (mm)			
	ET	TC	WT	AVG.	ET	TC	WT	AVG.
Attention UNet [8]	78.46 ± 0.317	83.23 ± 0.264	84.79 ± 0.245	82.16	18.65	12.33	16.15	15.71
ResUNet [36]	77.68	86.45	81.42	81.85	37.45	13.27	12.35	21.02
TransUNet [12]	79.28 ± 0.294	83.34 ± 0.201	85.78 ± 0.104	82.80	26.38	15.24	10.34	17.32
SwinBTS [24]	83.21 ± 0.222	84.75 ± 0.227	91.83 ± 0.078	86.60	16.03	14.51	8.19	12.91
UNETR [26]	84.50 ± 0.283	88.20 ± 0.198	91.60 ± 0.089	87.95	12.23	7.73	7.78	9.26
SwinUNETR [27]	**84.90 ± 0.284**	88.60 ± 0.203	91.02 ± 0.092	**88.50**	**11.09**	6.89	7.33	**8.44**
VcaNet	83.25 ± 0.230	**89.52 ± 0.175**	**92.03 ± 0.081**	88.26	18.16	**6.75**	**4.53**	9.81



(a) Comparison of results on BraTS 2020



(b) Comparison of results on BraTS 2021

Fig. 6. Histogram comparing Dice scores on BraTS 2021 and BraTS 2020.

4.5. Discussion

The results presented in this study demonstrate the significant improvement of VcaNet over previous models, particularly in handling the complex and variable structures of brain tumors. The integration of the Vision Transformer (ViT) within the bottleneck layer, alongside the Enhanced Convolutional (ENCO) and Channel and Spatial Attention Mechanism (CBAM) modules, allowed for more effective feature extraction and segmentation accuracy across all three tumor subregions: Enhancing Tumor (ET), Tumor Core (TC), and Whole Tumor (WT). These improvements were consistent across both the BraTS 2020 and BraTS 2021 datasets, as evidenced by the substantial increase in Dice scores and reduction in 95 % Hausdorff distances when compared to other state-of-the-art models.

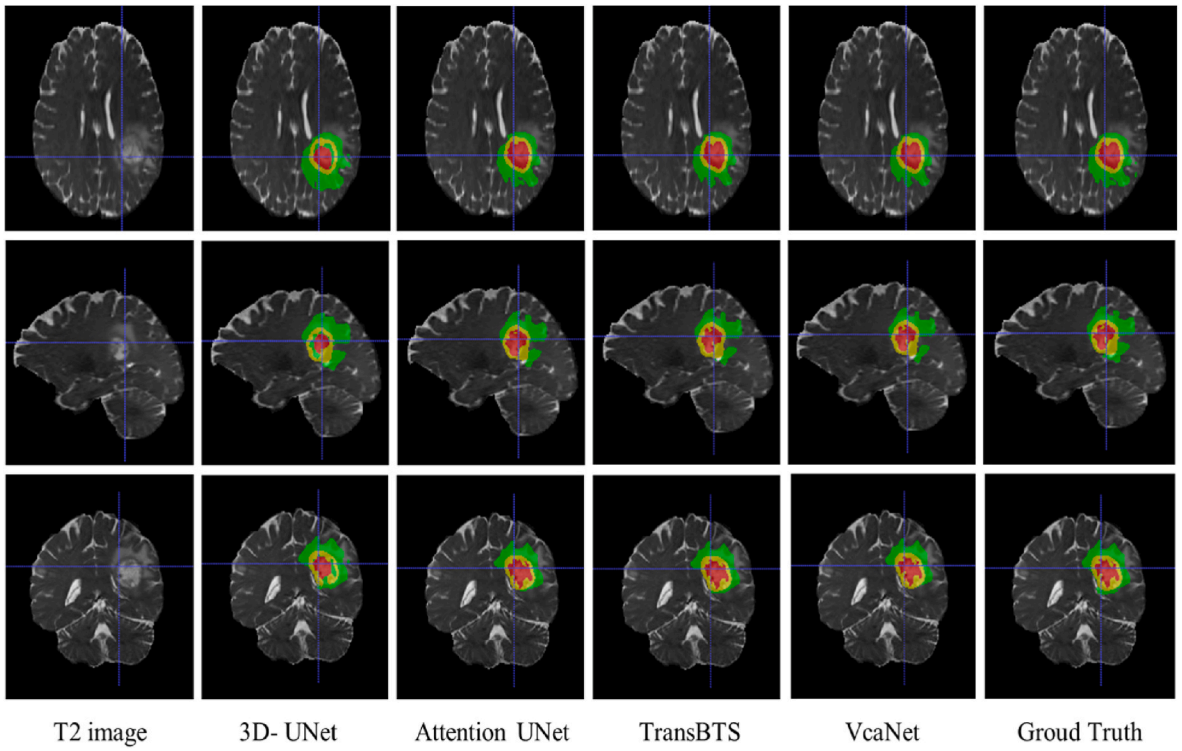


Fig. 7. The results of the segmentation of MRI brain tumors are visual comparison.

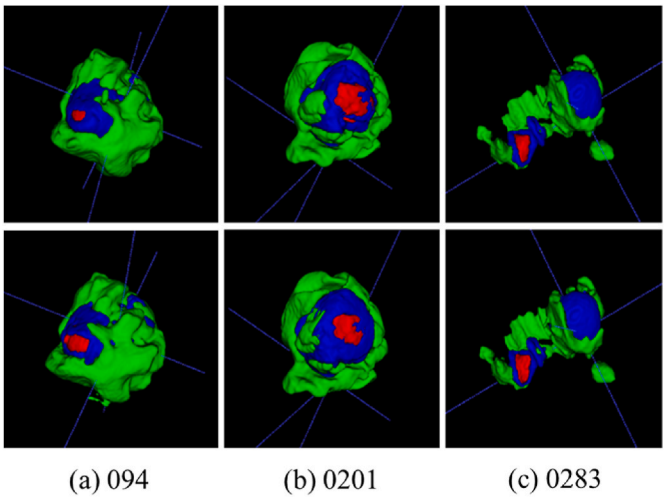


Fig. 8. Segmentation results of VcaNet. VcaNet results are shown in the first row. The ground truth is shown in the second raw. Additionally, numbers refer to the indices of samples.

The enhanced performance of VcaNet, especially compared to models like TransBTS, can be attributed to its ability to balance local feature extraction and global contextual modeling. The ENCO module enabled the capture of high-resolution local features in the early stages of the encoder, while the multi-scale feature extraction (MSC) module combined with the ViT effectively captured long-range dependencies in the data. This approach addresses a key limitation of convolutional neural networks (CNNs), which struggled with long-range dependencies due to their limited receptive field.

Moreover, the CBAM module, introduced during the upsampling process, further improved segmentation accuracy by focusing attention on the most critical regions of the image. This was particularly beneficial in the segmentation of tumor edges, where local feature details often play a critical role. As observed in the qualitative comparison in Fig. 7, VcaNet produced more precise tumor boundaries compared to other models, particularly in challenging cases involving the Enhancing Tumor region.

However, some limitations remain. One significant challenge is the computational complexity associated with the Vision Transformer (ViT). While ViT significantly enhances the model's ability to capture global dependencies, its quadratic scaling with the input sequence length makes it computationally expensive. This can limit its applicability in real-time clinical environments where hardware resources are constrained. Future research will focus on optimizing the ViT module or exploring more lightweight transformer variants to make VcaNet more computationally efficient.

Another limitation is the model's intermittent failure to accurately segment certain tumor regions, particularly in cases with highly irregular tumor morphologies or when the tumor regions demonstrate low contrast relative to the surrounding brain tissue. As shown in Fig. 9, some failure cases indicate that VcaNet struggled with under-segmentation in the Tumor Core (TC) region and over-segmentation in the Whole Tumor (WT) region. These errors are likely due to the model's sensitivity to low-contrast areas, where even the attention mechanisms fail to effectively differentiate between tumor and non-tumor tissues. Further refinement of the attention modules and the inclusion of more diverse training data could help mitigate these errors.

4.6. Ablation studies

We conducted a sufficient ablation study on VcaNet on the BraTS 2021 training samples to validate the model. The three main ablation experiments are as follows. (1) We investigated ablation experiments on module implementation details in the bottleneck layer. (2) We investigate the effect of introducing the CBAM attention module in the upsampling phase on the segmentation performance. (3) We study the Transformer effect at various model scales (i.e., depth L).

Bottleneck block: We use VcaNet as our baseline network and modify the implementation details of the modules in the bottleneck

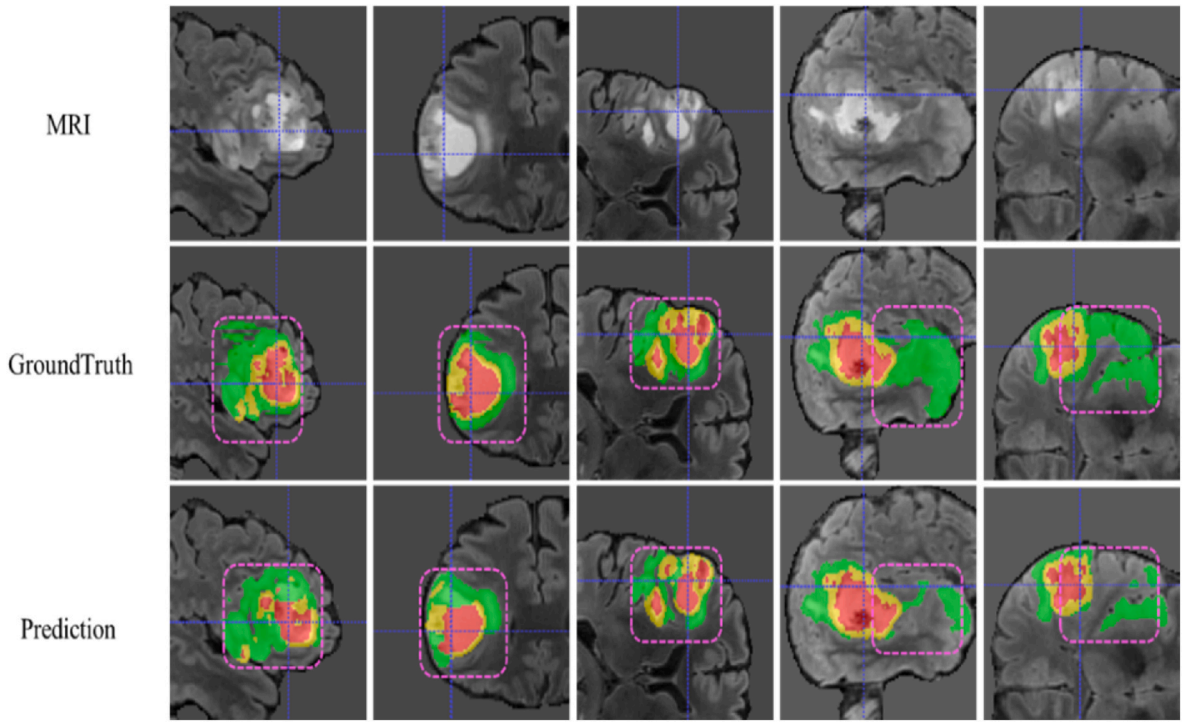


Fig. 9. Some failure segmentation cases of VcaNet.

Table 4
Ablation study results for each module.

Method	Dice Score(%)			
	ET	TC	WT	AVG.
Baseline	78.24	81.33	87.96	82.51
Baseline + ViT	82.63	87.47	91.97	87.36
Baseline + MSC + ViT	**83.25**	**89.52**	**92.03**	**88.26**

Table 5
Ablation study of each module.

Method	Dice Score(%)			
	ET	TC	WT	AVG.
Baseline	81.41	87.33	88.97	85.90
Baseline + SE	80.16	82.09	88.96	83.74
Baseline + CBAM	**83.25**	**89.52**	**92.03**	**88.26**

layer to obtain a series of variants of VcaNet for ablation experiments. First, we take the bottleneck layer based on 3D-UNet without any modification as our baseline, and add the MSC module and the ViT module sequentially to the baseline, and denote them as Baseline + MSC and Baseline + MSC + ViT, respectively. Table 4 displays the outcomes of the ablation experiments conducted on each module.

The results in Table 3 demonstrate that replacing the original bottleneck layer's 3D CNN with the ViT module significantly enhances VcaNet's ability to model global contextual information. This modification effectively compensates for the loss of spatial and depth details. In contrast, when the MSC module is absent, the model prioritizes the global information in the image while disregarding the local and detailed information in each image, which lowers the performing model's segmentation.

Upsampling block: We use the model without the CBAM attention module introduced in the up-sampling stage as the base model and introduce the SE attention module and the CBAM module for comparison in order to examine the impact of the CBAM attention module introduction on the model segmentation performance. Table 5 displays the outcomes of the experiment. The model's segmentation performance achieves its peak when the CBAM module is introduced during the upsampling step. In contrast, the SE attention module, which is also frequently utilized in the field of image segmentation, has no effect on the model's performance. The CBAM module, through the two sub-modules of channel attention and spatial attention, effectively fuses the encoder features of the encoder part passed by the skipping connection and the decoder features of the decoder part. This fusion process is used for further feature extraction, which is conducive to recovering the detailed information lost during the downsampling process and enhances VcaNet's ability to predict segmentation. As shown in Fig. 10, a comparison of the visualization of the ablation experiments is given.

Transformer Scale: Considering the computational and memory overhead of the model, we conducted an ablation experiment on the factor of the total amount of Transformer layers (depth L), which is the main factor affecting the Transformer scale. To investigate the impact of the Transformer scale on the model segmentation performance, the total amount of Transformer layers in the bottleneck layer is set at 4, 6, and 8 in this research. Table 6 displays the outcomes of the experiment. When $L = 6$, the network achieves optimal performance. A lower number of layers negatively impacts the global modeling ability of the Transformer. However, increasing the number of layers leads to excessive computational and memory demands, making the model difficult to train, which ultimately degrades segmentation performance.

5. Conclusion

In this study, we introduced VcaNet, a novel architecture for 3D brain tumor segmentation that effectively combines the strengths of Convolutional Neural Networks (CNNs) and Vision Transformers (ViT). The incorporation of the Enhanced Convolutional (ENCO) module enabled the extraction of high-resolution local features crucial for accurate segmentation, while the Vision Transformer, integrated within the bottleneck layer, facilitated the modeling of long-range dependencies and global contextual information. The Channel and Spatial

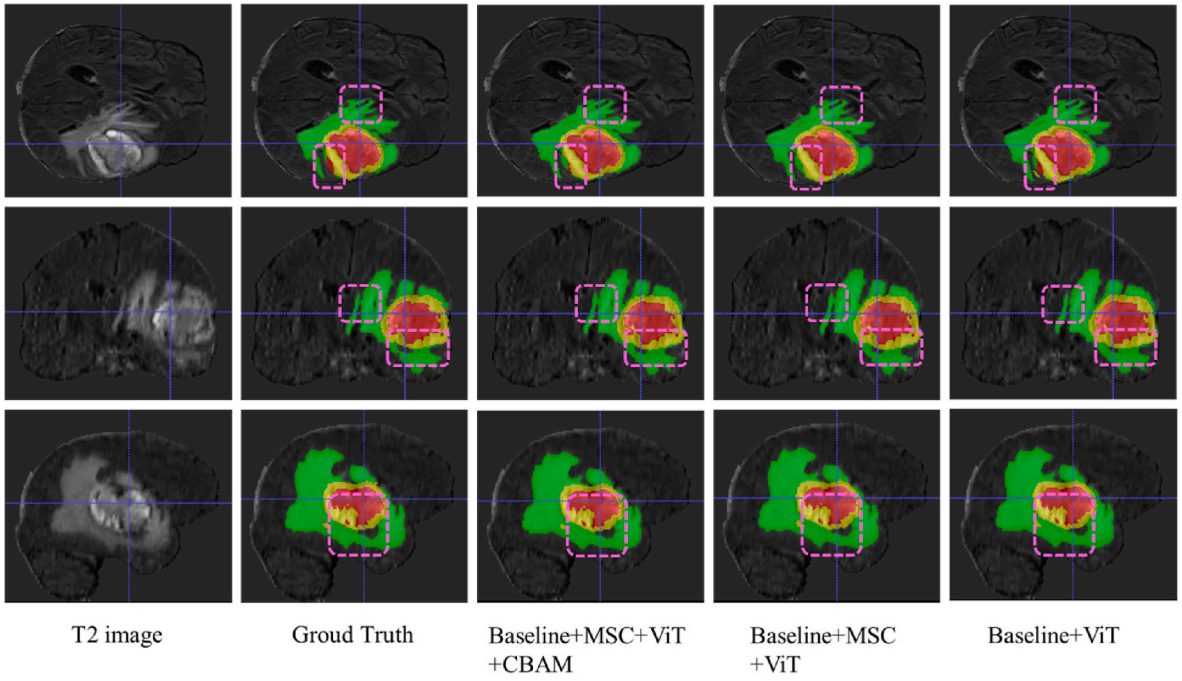


Fig. 10. Some segmentation cases of ablation experiments.

Table 6
Ablation study on Transformer.

Depth(L)	Dice Score(%)			
	ET	TC	WT	AVG.
4	82.60	86.68	91.47	86.92
6	**83.25**	**89.52**	**92.03**	**88.26**
8	82.18	88.36	91.45	87.33

Attention Mechanism (CBAM) further refined the feature integration process by enhancing the balance between local and global features during upsampling, ensuring the preservation of critical spatial details. Extensive experiments on the BraTS 2020 and 2021 datasets validated the robustness of VcaNet, demonstrating its superior performance over existing models, particularly in segmenting complex and irregular brain tumor morphologies. These findings highlight the model's potential to significantly improve the accuracy of medical image analysis by addressing the limitations inherent in traditional CNN architectures, making a substantial contribution to the field of 3D medical image segmentation.

While VcaNet shows promising advancements, there are some challenges and limitations. The high computational demand resulting from the integration of CNN and ViT architectures presents obstacles for deploying the model in resource-constrained environments, potentially limiting its scalability and real-time clinical applications. Additionally, the model demonstrated some inconsistencies in segmenting specific tumor subregions, such as under-segmentation of the Tumor Core and over-segmentation of the Whole Tumor, likely due to low-contrast regions in the MRI scans. Future research should focus on optimizing the computational efficiency of VcaNet, potentially through lightweight attention mechanisms or more efficient transformer variants. Moreover, further refinement of the attention modules and the inclusion of more diverse and representative training data could mitigate segmentation errors. These directions would enhance VcaNet's applicability for clinical use, especially in real-time tumor diagnosis and segmentation systems, providing the medical community with a powerful tool for more precise brain tumor analysis.

CRediT authorship contribution statement

Dichao Pan: Writing – original draft, Visualization, Software, Resources, Methodology, Data curation. **Jianguo Shen:** Writing – review & editing, Validation, Supervision, Resources, Investigation, Formal analysis. **Zaid Al-Huda:** Writing – review & editing, Validation, Investigation, Formal analysis, Conceptualization. **Mohammed A.A. Al-qaness:** Writing – review & editing, Validation, Investigation, Formal analysis, Conceptualization.

Ethics statement

We declare that we did not use human or animals for our paper. We used public datasets as described in the paper. So, no ethical approval is needed.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Z. Zhu, S. Mengwei, Q. Guangiu, Y. Li, X. Gao, Y. Liu, Sparse Dynamic Volume TransUNet with multi-level edge fusion for brain tumor segmentation, Comput. Biol. Med. (2024) 108284.
[2] Y. Xu, K. Yu, G. Qi, Y. Gong, X. Qu, L. Yin, P. Yang, Brain tumour segmentation framework with deep nuanced reasoning and Swin-T, IET Image Process. 18 (6) (2024) 1550–1564.
[3] Z. Yu, L. Xiang, L. Jiaxin, C. Weigiang, T. Zhiri, G. Daoying, HSA-net with a novel CAD pipeline boosts both clinical brain tmor MR image classification andsegmentation, Comput. Biol. Med. 170 (2024) 108039.
[4] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.
[5] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: a nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, Springer, 2018, pp. 3–11.

[6] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision (3DV), Ieee, 2016, pp. 565–571.
[7] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv: arXiv:1409.0473 (May 19, 2016) [Online]. Available: http://arxiv.org/abs/1409.0473. (Accessed 17 August 2024).
[8] J. Schlemper, et al., Attention gated networks: learning to leverage salient regions in medical images, Med. Image Anal. 53 (2019) 197–207.
[9] A. Vaswani, et al., "Attention is all you need,", arXiv: arXiv:1706.03762 (Aug. 01, 2023) [Online]. Available: http://arxiv.org/abs/1706.03762. (Accessed 17 August 2024).
[10] R.M. Schmidt, Recurrent neural networks (RNNs): a gentle introduction and overview, arXiv: arXiv:1912.05911 (Nov. 23, 2019) [Online]. Available: http://arxiv.org/abs/1912.05911. (Accessed 17 August 2024).
[11] L. Yuan, et al., Tokens-to-token vit: training vision transformers from scratch on imagenet, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 558–567.
[12] J. Chen, et al., TransUNet: transformers make strong encoders for medical image segmentation, arXiv: arXiv:2102.04306 (Feb. 08, 2021) [Online]. Available: http://arxiv.org/abs/2102.04306. (Accessed 17 August 2024).
[13] 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.
[14] 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 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, Springer, 2016, pp. 424–432.
[15] F.I. Diakogiannis, F. Waldner, P. Caccetta, C. Wu, ResUNet-a: a Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data, Jan. 15, 2020, https://doi.org/10.1016/j.isprsjprs.2020.01.013.
[16] D. Maji, P. Sigedar, M. Singh, Attention Res-UNet with Guided Decoder for semantic segmentation of brain tumors, Biomed. Signal Process Control 71 (Jan. 2022) 103077, https://doi.org/10.1016/j.bspc.2021.103077.
[17] F. Isensee, et al., nnU-Net: self-adapting framework for U-Net-Based medical image segmentation, arXiv: arXiv:1809.10486 (Sep. 27, 2018) [Online]. Available: http://arxiv.org/abs/1809.10486. (Accessed 17 August 2024).
[18] Z. Zhu, Z. Wang, G. Qi, N. Mazur, P. Yang, Y. Liu, Brain tumor segmentation in MRI with multi-modality spatial information enhancement and boundary shape correction, Pattern Recogn. 153 (2024) 110553.
[19] W. Wenxuan, C. Chen, D. Meng, Y. Hong, Z. Sen, L. Jiangyun, Transbts: multimodal brain tumor segmentation using transformer, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 109–119.
[20] J. Li, et al., TransBTSV2: towards better and more efficient volumetric segmentation of medical images, arXiv: arXiv:2201.12785 (May 17, 2022) [Online]. Available: http://arxiv.org/abs/2201.12785. (Accessed 17 August 2024).
[21] Q. Jia, H. Shu, Bitr-unet: a cnn-transformer combined network for mri brain tumor segmentation, in: International MICCAI Brainlesion Workshop, Springer, 2021, pp. 3–14.
[22] Z. Zhu, X. He, G. Qi, Y. Li, B. Cong, Y. Liu, Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI, Inf. Fusion 91 (2023) 376–387.
[23] Z. Liu, et al., Swin transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[24] Y. Jiang, Y. Zhang, X. Lin, J. Dong, T. Cheng, J. Liang, SwinBTS: a method for 3D multimodal brain tumor segmentation using Swin transformer, Brain Sci. 12 (6) (Jun. 2022) 797, https://doi.org/10.3390/brainsci12060797.
[25] Z. Zhu, K. Yu, G. Qi, B. Cong, Y. Li, Z. Li, X. Gao, Lightweight medical image segmentation network with multi-scale feature-guided fusion, Comput. Biol. Med. 182 (2024) 109204.
[26] A. Hatamizadeh, et al., Unetr: transformers for 3d medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 574–584.
[27] 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.
[28] H. Peiris, M. Hayat, Z. Chen, G. Egan, M. Harandi, A robust volumetric transformer for accurate 3D tumor segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 162–172.
[29] H. Chen, J. An, B. Jiang, L. Xia, Y. Bai, Z. Gao, WS-MTST: weakly supervised multi-label brain tumor segmentation with transformers, IEEE J. Biomed. Health Inform. 27 (12) (Dec. 2023) 5914–5925, https://doi.org/10.1109/JBHI.2023.3321602.
[30] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, Y. Yu, nnformer: interleaved transformer for volumetric segmentation, ArXiv Prepr. ArXiv210903201 (2021).
[31] F. Chollet, Xception: deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.
[32] B.H. Menze, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging 34 (10) (Oct. 2015) 1993–2024, https://doi.org/10.1109/TMI.2014.2377694.
[33] U. Baid, et al., The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification, arXiv: arXiv:2107.02314 (Sep. 12, 2021) [Online]. Available: http://arxiv.org/abs/2107.02314. (Accessed 17 August 2024).
[34] S. Bakas, et al., Advancing the Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features, Sci. Data 4 (1) (Sep. 2017) 170117, https://doi.org/10.1038/sdata.2017.117.
[35] S. Bakas, et al., Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge, arXiv: arXiv:1811.02629 (Apr. 23, 2019) [Online]. Available: http://arxiv.org/abs/1811.02629. (Accessed 17 August 2024).
[36] Z. Zhang, Q. Liu, Y. Wang, Road extraction by deep residual U-net, IEEE Geosci. Remote Sens. Lett. 15 (5) (May 2018) 749–753, https://doi.org/10.1109/LGRS.2018.2802944.
[37] D. Lachinov, E. Vasiliev, V. Turlapov, Glioma segmentation with cascaded UNet, in: International MICCAI Brainlesion Workshop, Springer, 2018, pp. 189–198.
[38] J.M.J. Valanarasu, V.M. Patel, Unext: mlp-based rapid medical image segmentation network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 23–33.