Computers in Biology and Medicine 172 (2024) 108284

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Sparse Dynamic Volume TransUNet with multi-level edge fusion for brain tumor segmentation

Zhiqin Zhu [a], Mengwei Sun [a], Guanqiu Qi [b], Yuanyuan Li [a], Xinbo Gao [a], Yu Liu [c,*]

[a] College of Automation, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China
[b] Computer Information Systems Department, State University of New York at Buffalo State, Buffalo, NY 14222, USA
[c] Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China

ARTICLE INFO

Keywords:
Brain tumor segmentation
Transformer
Sparse dynamic
Edge fusion

ABSTRACT

3D MRI Brain Tumor Segmentation is of great significance in clinical diagnosis and treatment. Accurate segmentation results are critical for localization and spatial distribution of brain tumors using 3D MRI. However, most existing methods mainly focus on extracting global semantic features from the spatial and depth dimensions of a 3D volume, while ignoring voxel information, inter-layer connections, and detailed features. A 3D brain tumor segmentation network SDV-TUNet (Sparse Dynamic Volume TransUNet) based on an encoder–decoder architecture is proposed to achieve accurate segmentation by effectively combining voxel information, inter-layer feature connections, and intra-axis information. Volumetric data is fed into a 3D network consisting of extended depth modeling for dense prediction by using two modules: sparse dynamic (SD) encoder–decoder module and multi-level edge feature fusion (MEFF) module. The SD encoder–decoder module is utilized to extract global spatial semantic features for brain tumor segmentation, which employs multi-head self-attention and sparse dynamic adaptive fusion in a 3D extended shifted window strategy. In the encoding stage, dynamic perception of regional connections and multi-axis information interactions are realized through local tight correlations and long-range sparse correlations. The MEFF module achieves the fusion of multi-level local edge information in a layer-by-layer incremental manner and connects the fusion to the decoder module through skip connections to enhance the propagation ability of spatial edge information. The proposed method is applied to the BraTS2020 and BraTS2021 benchmarks, and the experimental results show its superior performance compared with state-of-the-art brain tumor segmentation methods. The source codes of the proposed method are available at https://github.com/SunMengw/SDV-TUNet.

1. Introduction

As intracranial space-occupying lesions, brain tumors originate in the nervous system and subsequently metastasize. Accurate and reproducible tumor quantification and morphology are critical for brain tumor diagnosis, treatment planning, and monitoring treatment response. Brain tumor segmentation aims to localize multiple types of tumor regions from images [1]. Magnetic resonance imaging (MRI) plays a vital role in brain tumor treatment as it enables clinicians to determine key tumor parameters such as location, size, and type. 3D MRI as an imaging technique provides comprehensive and accurate brain structure information. Accurate brain tumor segmentation often requires the utilization of complementary 3D MRI modalities for spatial distribution monitoring and disease progression analysis, including fluid attenuation inversion recovery (FLAIR), T1-weighted (T1), contrast enhanced T1-weighted (T1ce), and T2-weighted (T2) [2]. In particular, brain tumor segmentation typically involves three partitions: enhancing tumor (ET), necrosis and non-enhancing tumor (NCR/NET), and peritumoral edematous/invaded tissue (ED). Fig. 1 illustrates examples of different 3D MRI modalities used for brain tumor segmentation, and the corresponding 3D visualization of the segmentation results. The ground truth (GT) segmentation label, approved by expert neuroradiologists, are shown in Fig. 1(e), in which green, yellow and red indicate ED, ET and NCR/NET regions, respectively [3].

As a key field of medical image segmentation research, brain tumor segmentation is of great significance in clinical practice and has a wide range of applications in quantitative neurological analysis, surgical planning, and functional imaging [4]. Although 3D MRI can accurately represent brain tissue structures, brain tumor segmentation remains challenging due to various factors such as low spatial resolution, presence of noise, and the influence of acquisition artifacts. In

* Corresponding author.
E-mail addresses: zhuzq@cqupt.edu.cn (Z. Zhu), s220333016@stu.cqupt.edu.cn (M. Sun), qig@buffalostate.edu (G. Qi), liyy@cqupt.edu.cn (Y. Li), gaoxb@cqupt.edu.cn (X. Gao), yuliu@hfut.edu.cn (Y. Liu).

https://doi.org/10.1016/j.compbiomed.2024.108284
Received 17 January 2024; Received in revised form 19 February 2024; Accepted 12 March 2024
Available online 15 March 2024

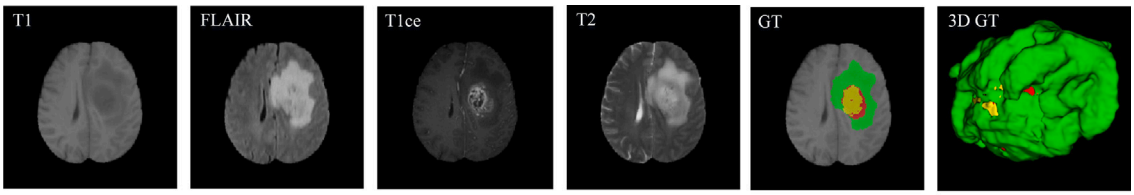

Fig. 1. Examples of a four-modality MRI for brain tumor segmentation and ground truth (GT) 3D visualization. (a) T1 modality. (b) FLAIR modality. (c) T1ce modality. (d) T2 modality. (e) GT segmentation label approved by expert neuroradiologists (the yellow, red and green represent enhancing tumor (ET), necrosis and non-enhancing tumor (NCR/NET), and peritumoral edematous/invaded tissue (ED), respectively). (f) GT segmentation 3D visualization.

recent years, deep learning-based methods have become mainstream methods for brain tumor segmentation. Convolutional neural networks (CNNs) are widely used in medical image segmentation tasks due to their powerful feature representation [5]. Among them, both fully convolutional networks (FCN) [6] and U-Net [7] have shown strong performance in two-dimensional (2D) brain tumor segmentation. However, the limited receptive fields of CNNs hamper their ability to capture extensive spatial dependencies in input images, thus limiting their performance [8]. The concept of dilated convolutional neural networks [9,10], an extension model of CNN, can to some extent enlarge the receptive field of the input, but still has certain limitations in terms of performance enhancement. Inspired by the attention mechanism in natural language processing [11], Transformer as a method capable of global interactive modeling was introduced into brain tumor segmentation [12], and showed clear advantages in complex predictions. 2D networks based on CNN and Transformer have improved segmentation performance. However, in 2D models, continuous slices are typically processed by stacking, which may lead to information loss and boundary artifacts between slices [13]. In contrast, 3D models can process the entire volume, reducing information loss, capturing rich spatial information, and better understanding the overall context of brain structures. To overcome the problem of simply dividing 3D MRI into 2D slices for input and ignoring 3D volume and position information, more and more 3D medical image segmentation methods have been proposed [14,15] to effectively implement local and global feature modeling on 3D volumetric data. However, during the process of segmenting the image into patches and projecting patches into tokens using the standard Vision Transformer (ViT) [16], local structural clues may be compromised. Swin Transformer [17,18] acts as a hierarchical visual Transformer to compute self-attention in an effective shifted window segmentation scheme. A 3D Swin Transformer-based brain tumor segmentation method [19] achieves outstanding results in window information interaction. However, the window segmentation mechanism of Swin Transformer causes blocking artifacts to a certain extent. The shifted window mechanism is inefficient for establishing cross-window connections and cannot fully realize window information interactions [20].

In medical image segmentation, it is crucial to learn global semantic information and detailed local features simultaneously. Existing brain tumor segmentation methods, such as multi-path CNN [21], have explored the fusion of local and global features to enhance segmentation performance by directly integrating global semantic feature paths and local spatial information paths. However, this comes at the cost of high computational complexity. Although some methods [22,23] introduce local feature fusion modules to integrate local and global context information and improve computational efficiency accordingly, they mainly focus on the extraction and fusion of deep semantic features without considering the importance of multi-level edge features for segmentation. Edge features are conducive to obtaining more accurate tumor localization and boundaries [24].

Although CNN- and Transformer-based brain tumor segmentation methods achieve impressive segmentation performance, they primarily focus on the extraction of global semantic features, while ignoring voxel information, inter-layer connections, and detailed features of 3D volumes. Meanwhile, edge information facilitates accurate identification and delineation of tumor locations. The fusion of detailed edge features and global spatial features is critical for improving the quality of brain tumor segmentation. Several segmentation methods focus on the fusion of local and global feature. However, the multi-path fusion methods occupy a large amount of computational memory, and the local feature fusion module ignores the importance of multi-level edge features for segmentation. This paper proposes a 3D brain tumor segmentation network SDV-TUNet (Sparse Dynamic Volume TransUNet) based on an encoder–decoder architecture. The proposed network achieves more accurate segmentation by combining voxel information, inter-layer feature connections, and intra-axis information. In the encoding stage, the proposed continuous self-attention layer based on regular window and shifted window attention integrated with sparse dynamic adaptive strategy works directly on 3D volumes to learn global features along spatial axes of the spatial and depth dimensions. The sparse dynamic strategy uses voxels as the starting point to search for spatially similar regions, and the shifted window mechanism and sparse dynamic strategy realize the effective combination of local tight correlation and long-range sparse correlation of dense prediction tasks. The encoder branch of the edge feature fusion module is also designed to capture local details and edge features. This module uses cascaded and parallel 3D atrous convolutions blocks to fuse detail features from different layers. These fused features are then connected to the decoding stage via skip connections. With the above design, the proposed segmentation framework effectively extracts and fuses the spatial semantic features and edge features of 3D MRI. Experimental results on the BraTS2020 and BraTS2021 benchmarks demonstrate the superior performance of the proposed brain tumor segmentation method.

This paper has three main contributions as follows.

- The proposed brain tumor segmentation method SDV-TUNet can dynamically extract deep 3D spatial features and capture fine-grained spatial edge details. Compared with existing 3D brain tumor segmentation methods, the proposed method achieves sequence-to-sequence prediction by inputting vectors reshaped from volumes into a 3D Transformer backbone network composed of depth dimension extension modeling. The proposed network holistically processes volumetric data through 3D patch segmentation, merging and expansion to achieve feature transfer at different resolutions, effectively integrating global spatial information based on 3D voxels and fusing edge feature information.
- An SD (sparse dynamic) encoder–decoder module is proposed to extract global spatial semantic features for brain tumor segmentation. To address the limitations of 3D segmentation in using intra-axis information and inter-layer contents, multi-head self-attention and sparse dynamic adaptive fusion are integrated into a 3D extended shifted window strategy. The dynamic perception of regional connections and multi-axis information interactions are achieved through local tight correlations and long-range sparse correlations.

- A multi-level edge feature fusion (MEFF) module is proposed to build cascaded and parallel branches of 3D atrous convolutions. This module facilitates the fusion of multi-level local edge information in a layer-by-layer incremental manner. Through skip connections, edge information is connected to the fused attention decoder module, allowing the propagation perception of spatial edge information.

The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 specifies the proposed method. Section 4 discusses the related experiments and analyzes the corresponding experimental results. Section 5 concludes the whole paper.

2. Related work

2.1. CNN-based models

In recent years, deep learning methods have performed well in solving various computer vision problems, including brain tumor segmentation [25,26]. CNNs have had a profound impact on the field of medical imaging due to their ability to learn highly complex representations in a data-driven manner. As one of the main architectures for medical image segmentation, U-Net [7] adopts classical encoder–decoder networks and relies on data augmentation for end-to-end training, especially when segmentation training samples are limited. The development of U-Net [7] has significantly promoted the progress of medical image segmentation algorithms, particularly in the field of brain tumor segmentation. Numerous variants of U-Net, such as U-Net++ [27] and Res-UNet [28], have further enhanced its performance. Liu et al. [29] added a Variational Autoencoder (VAE) decoder to reconstruct the input images, using image fusion as an additional regularization method for feature learning. The CNN-based models perform well in 2D brain tumor image segmentation. However, MRI segmentation methods using slice-by-slice 2D networks ignore 3D sequence information and position information as well as related information in the volume.

Recently, more and more 3D network models have been proposed to exploit 3D spatial information and learn powerful high-dimensional feature representations from 3D MRI data. Compared to 2D networks that can only process individual slices, effectively balancing sparse inter-slice information with dense intra-slice information, 3D networks are capable of considering spatial information along the depth direction. This makes them more effective in understanding the spatial structure of tumors and the relationships with adjacent tissues. When processing volumes, 3D network models better preserve continuous spatial information, reducing the loss of information. This is particularly important for brain tumor segmentation tasks. 3D fully convolutional networks (3D FCN) [30] are widely used for brain volume image segmentation, including 3D U-Net [14], nnU-Net [31]. The nnU-Net framework serves as a general baseline model for 2D and 3D medical image segmentation. Due to its good segmentation performance, various brain tumor segmentation models and their variants are developed based on the nnU-Net framework. CANet proposed by Liu et al. [32] captures sequence information by introducing feature interaction maps, which are combined with the convolutional space to extract discriminative features with context.

2.2. Transformer-based models

Compared with CNN, Transformer has a greater advantage in modeling global interactions, while CNN has difficulty in establishing explicit long-range dependencies due to the limitations of convolution operations. Transformer-based methods have received more and more attention in medical image segmentation, and some representative models have been proposed successively. Chen et al. [12] first proposed TransUNet, a hybrid Transformer-CNN architecture, to explore the potential of Transformer in medical image segmentation. In this architecture, CNN acts as a feature extraction and transformation module, while Transformer performs global context encoding. However, when the image is directly segmented into patches to act as Transformer tokens, the local structure of the 3D volume is ignored. To effectively utilize 3D volumetric data for global interaction modeling between consecutive slices, Wang et al. [15] proposed TransBTS as the first attempt to utilize the Transformer in 3D CNN for 3D MRI brain tumor segmentation. A ViT-based 3D medical image segmentation architecture (UNETR) proposed by Hatamizad et al. [33] consists of a pure Transformer as an encoder to learn the sequential representation of the input quantity, and the encoder is connected to the CNN-based decoder via skip connections to fuse local and global information. Standard ViT-based methods lead to high computational complexity for dense prediction such as semantic segmentation due to their fixed input size. Similar to CNN, Swin Transformer [17] adopts a hierarchical structure to implement feature mapping, which effectively alleviates high computational complexity. This enhancement significantly enhances the potential of Transformer models in medical image segmentation. Zhou et al. [34] proposed a 3D Transformer block-based nnFormer that interleaves convolution and self-attention operations to concatenate encoder and decoder features by skip attention. Swin UNETR [19] uses Transformer-based encoder to learn multi-scale contextual representations and model long-range dependencies. Peiris et al. [35] proposed a lightweight UNet-shaped VT-UNet to segment 3D medical image modalities in a hierarchical manner by encoding local and global clues via a volumetric Transformer.

2.3. Multi-path and local feature fusion models

To obtain more accurate segmentation, most of existing methods realize the interaction of global semantic features and local features by introducing multi-path fusion learning or local information fusion modules. Chandrakar et al. [36] proposed a multipath CNN architecture for brain tumor segmentation and detection, achieving the fusion learning of local and global features. Zhao et al. [37] proposed a deformable multi-path ensemble D-MEM for automatic segmentation of local and global features. Multipath fusion is computationally intensive and calculates all local and global paths for effective feature learning. However, multi-path fusion learning occupies a large amount of computational memory, and its introduction into 3D brain tumor segmentation is not feasible from the perspective of computational resources. A Transformer-based MISSU proposed by Wang et al. [38] introduces self-distillation via a local multi-scale fusion module to extract details from the skip connections in the encoder while simultaneously learning global semantic information and local spatial features. Zhang et al. [39] introduced a multipath feature fusion module and a multichannel feature pyramid module to capture information on small targets. Zhou et al. [40] proposed a technique for lossless feature computation in brain tumor segmentation, using 3D atrous convolutional layers. This approach combines background and lesion information through a coarse convolutional feature pyramid. Although the local information fusion module requires less computational space, it cannot fully capture multi-scale detail features and edge features. The selection and fusion of detail features such as edge and texture play a crucial role in learning image information [41].

Although the aforementioned CNN- and Transformer-based 3D medical image segmentation methods achieve impressive segmentation performance, they primarily focus on learning global spatial features using 3D volumetric data information while ignoring detailed feature representations at different levels and resolutions. The proposed solution not only learns global spatial features along the spatial axes and prioritizes inter-layer feature information, but also emphasizes extracting local features and edge features from multiple types of tumor regions in volume information. The extraction uses a multi-layer approach. Specifically, the inclusion of edge information facilitates

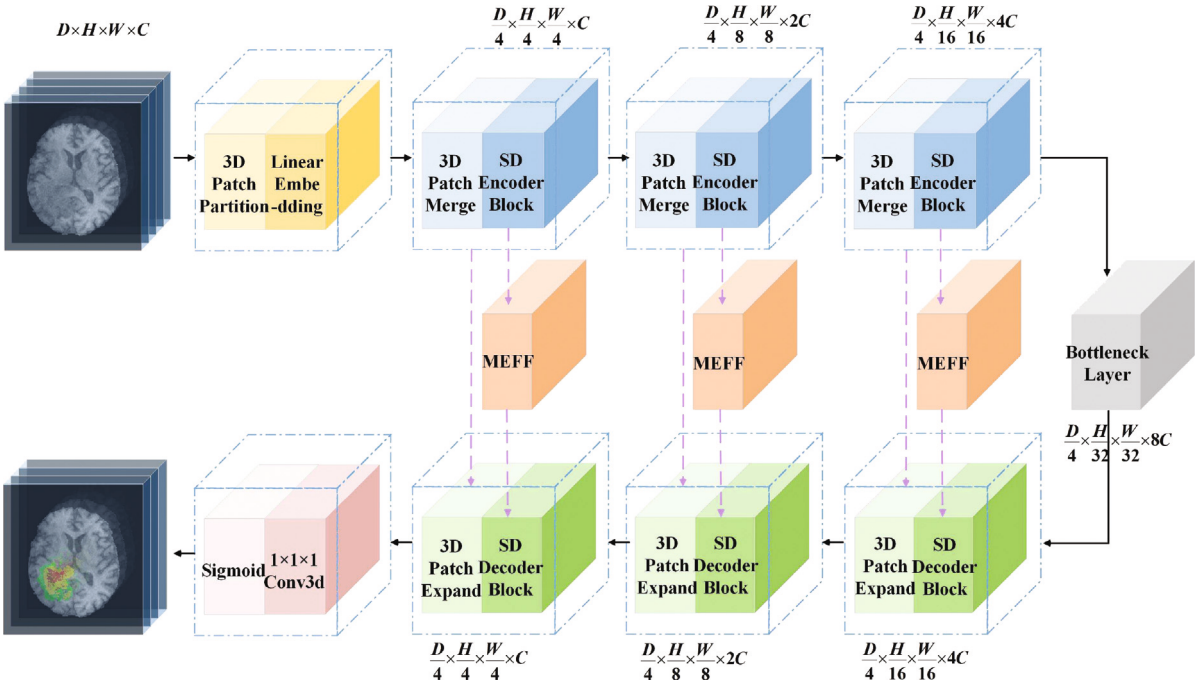

Fig. 2. The overall framework of SDV-TUNet network based on encoder–decoder architecture. It mainly consists of SD encoders, SD decoders and MEFF modules.

accurate identification and delineation of tumor locations. By fusing the extracted detailed edge features with the global spatial features, the segmentation accuracy can be improved, and finally good segmentation performance is achieved.

3. The proposed method

3.1. Overview

As shown in Fig. 2, the proposed SDV-TUNet network architecture is built on the encoder–decoder framework, which primarily consists of an SD encoder–decoder module and an MEFF module. The SD encoder is composed of a 3D patch partition layer, a linear embedding layer, 3D patch merging layers, and SD encoder modules. The SD encoder, using aggregated window attention and sparse dynamic attention mechanisms, extracts global spatial information by progressively downsampling a 3D volume. An encoder branch that receives local features and produces multi-layer edge feature fusion output is introduced to capture detailed local information and complement features. The bottleneck layer connecting the encoder and decoder backbone consists of the encoder module and a 3D patch expanding layer. The structure of the SD decoder is symmetrical to that of the encoder, starting from sequential SD decoder modules that perform patch expansion while fusing multi-layer edge information from the encoder branch. By performing multiple upsampling steps, implicit features containing global spatial information and local information are decoded to progressively generate full-resolution segmentation maps. Skip connections are symmetrically added between the corresponding MEFF modules of the encoder and decoder to facilitate recovery and enhancement to preserve fine-grained details.

3.2. SD encoder

The input to the network is a 3D volume $X \in \mathrm{R}^{D\times H\times W\times C}$ with depth dimension D (number of slices), spatial resolution $H\times W$ and number of channels C (number of modes). The SD encoder consists of a 3D patch partition layer and a linear embedding layer, 3D patch merging layers and SD encoder modules.

3D Patch Partition layer: The 3D patch partition layer creates a series of 3D volume tokens by partitioning the 3D volumes $X \in \mathrm{R}^{D\times H\times W\times C}$ into non-overlapping volume patches with a patch resolution $\left(D', H', W'\right)$. The linear embedding layer after the 3D patch partition layer projects the vectorized 3D tokens $\frac{D}{D'} \times \frac{H}{H'} \times \frac{W}{W'}$ into the latent C dimensional embedding space. To encode patch space information, specific position embeddings are learned and added to patch embeddings to preserve position information.

SD Encoder module: The input tensor containing the spatial information is passed to the SD encoder module after the embedding layer. As the primary objective, globally captured contextual relations, multi-scale layered objects generated by the patch merging process, and high-resolution spatial information encoded by the initial embedding layer are seamlessly integrated. To facilitate this integration, an aggregated attention layer is constructed using a 3D shifted window-based multi-head self attention (MSA) module and a sparse dynamic attention module. This attention layer focuses on both global features and regional associations. The feedforward network consists of a 2-layer multilayer perceptron (MLP) with a Gaussian error linear unit (GELU) in the middle. Layer normalization (Norm) is applied before each attention layer and MLP, and residual connections are incorporated after each module to reduce model complexity and alleviate overfitting.

In contrast to 2D partitioning, the preservation of spatial and multi-axis information from voxels is required in the case of a 3D volume. To achieve this, the 2D shifted window mechanism of Swin Transformer is extended to 3D windows [17]. However, the simple calculation rules and multi-head self-attention of the shifted window partition configuration are not conducive to effective cross-window information interaction and region linkage. Therefore, to efficiently aggregate information within different partition windows, capture intra-axis information and inter-layer connections, sparse dynamic adaptive fusion of expanded voxels is introduced along with efficient computation of rule-based and shifted window-based self-attention. This strategy is able to enhance the interaction between adjacent window features and focus on voxel feature information in similar regions across windows. In two consecutive layers of the self-attention module, the self-attention module of the first layer uses a general window partitioning strategy

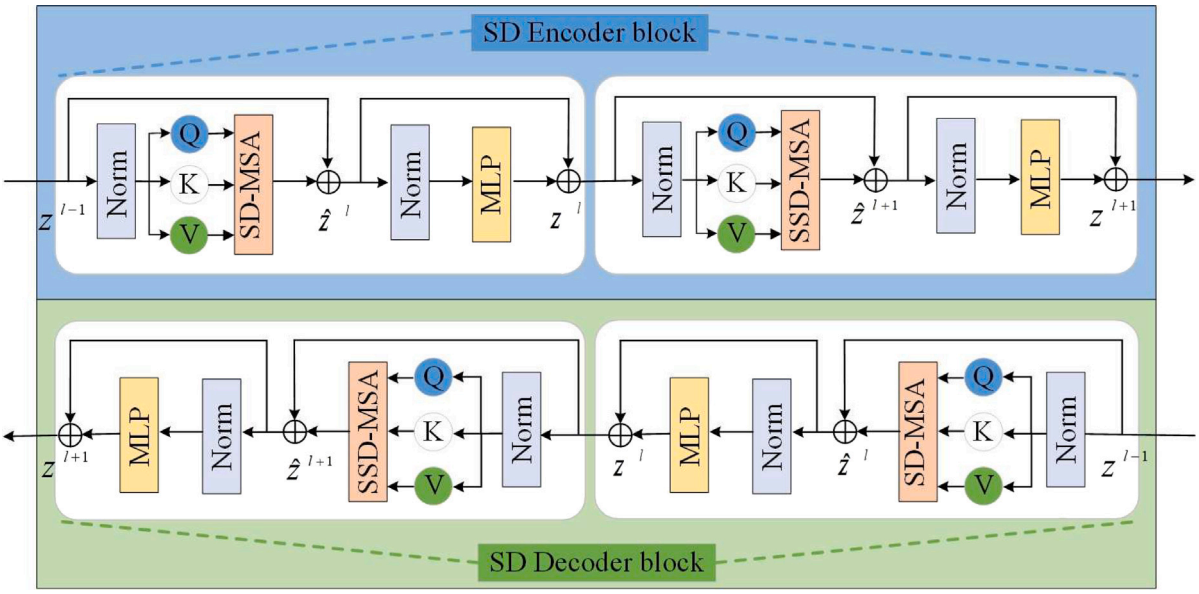

Fig. 3. The SD encoder–decoder with continuous self-attention layers.

that sets the 3D window size to $M \times M \times M$ and arranges the windows to evenly divide the input to obtain $\frac{D'}{M} \times \frac{H'}{M} \times \frac{W'}{M}$ non-overlapping 3D windows in a non-overlapping manner. For the second layer of the self-attention module, the window partition configuration shifts by tokens along the depth axis, height axis, and width axis from the self-attention module of the previous layer. Self-attention is computed into the non-overlapping windows created in the partitioning stage for efficient modeling of token interactions. As shown in Fig. 3, using the shifted window partitioning method, the features are output through two consecutive self-attention layers as follows.

$$\hat{z}^{l} = \text{SD} - \text{MSA}\left(\text{Norm}\left(z^{l-1}\right)\right) + z^{l-1}, \tag{1}$$

$$z^{l} = \text{MLP}\left(\text{Norm}\left(\hat{z}^{l}\right)\right) + \hat{z}^{l}, \tag{2}$$

$$\hat{z}^{l+1} = \text{SSD} - \text{MSA}\left(\text{Norm}\left(z^{l}\right)\right) + z^{l}, \tag{3}$$

$$z^{l+1} = \text{MLP}\left(\text{Norm}\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1}, \tag{4}$$

where $\hat{z}^{l}$ and z^{l} denote the output features of the SD encoder module and the MLP module at layer l, respectively, and SD-MSA and SSD-MSA denote aggregated sparse dynamic attention and multi-head self-attention based on rule and shifted window partitioning configurations, respectively. Sparse dynamic adaptive fusion is introduced at each layer to focus on the most similar regions in the 3D volume, enhancing cross-window connections and 3D volume region links for long-range information interaction.

Fig. 4 illustrates the rule window mechanism, the 3D cyclic shifted window mechanism, and the dynamic adaptive fusion strategy. The 3D volumetric local attention function takes the query (Q), key (K), and value (V) as input. Each query is converted into a weighted sum of values, where the weight is calculated as the normalized dot product between the query and the corresponding key. The multi-head self-attention applied to the 3D window is calculated as follows.

$$\text{Attention}\left(Q, K, V\right) = \text{SoftMax}\left(\frac{QK^{T}}{\sqrt{d}} + B\right)V, \tag{5}$$

where a scalar factor $\sqrt{d}$ is introduced to avoid weight concentration and gradient vanishing [19], for the 3D relative position bias.

To calculate the attention score of a target voxel in sparse dynamic adaptive fusion, attention calculation of non-empty voxels in a 3D volumetric search space centered at this voxel is performed. Specifically, for a non-empty voxel v_i in a 3D volume ω, the non-empty voxels of the 3D body centered on voxels are dynamically searched and adjacent voxels $\left\{v_j, j \in \Omega(i)\right\}$ are obtained based on query sharing of key–value pairs. The query, key and value embedding are computed as follows.

$$Q_i = L_q\left(\omega\left(v_i\right)\right), K_j = L_k\left(\omega\left(v_j\right)\right), V_j = L_v\left(\omega\left(v_j\right)\right), \tag{6}$$

where ω denotes the input 3D volume, L_q, L_k and L_v are the linear projection layers of Q, K and V. Operating at the voxel level, the embeddings for queries, keys, and values transform voxel feature vectors and generate corresponding representations through learned weight matrices in linear projection layers, achieving feature space transformation. For the position embedding P, in order to avoid the influence of bias due to the real coordinate scale of the 3D world, it is calculated from the relative position of voxels in the volume as follows.

$$P_j = L_p\left(v_j - v_i\right), \tag{7}$$

Then, the sparse dynamic attention is calculated as follows.

$$\text{Attention}\left(v_i\right) = \sum_{j \in \Omega(i)} \text{SoftMax}\left(Q_i\left(K_j + P_j\right)/\sqrt{d}\right)\left(V_j + P_j\right). \tag{8}$$

3D Patch Merging layer: After being converted by the SD encoder module, the patch tokens enter the 3D patch merging layer to generate hierarchical feature representations. Among them, the 3D patch merging layer is responsible for downsampling and dimensionality increase of feature maps. To generate a feature map with a hierarchical structure, adjacent 3D tokens are spliced along the spatial axes in a non-overlapping manner via the patch merging layer to reduce the number of tokens. Patch merging divides 2 × 2 adjacent 3D patches into groups and cascades them. The resulting vector is projected by a linear map into a space of doubled token channel dimension. The input is transformed from $\frac{D}{D'} \times \frac{H}{H'} \times \frac{W}{W'} \times C$ to $\frac{D}{D'} \times \frac{H}{2H'} \times \frac{W}{2W'} \times 2C$.

3.3. SD decoder

The network has a highly symmetric encoder and decoder backbone, and the feature information after patch merging enters the decoder through the bottleneck layer, and the SD decoder consists of SD decoder modules, 3D patch expanding layers, and a classifier layer.

SD Decoder module: The SD decoder consists of a three-stage encoder module and 3D patch expanding layers that outputs features with different layer resolutions. In the SD decoder module, two consecutive

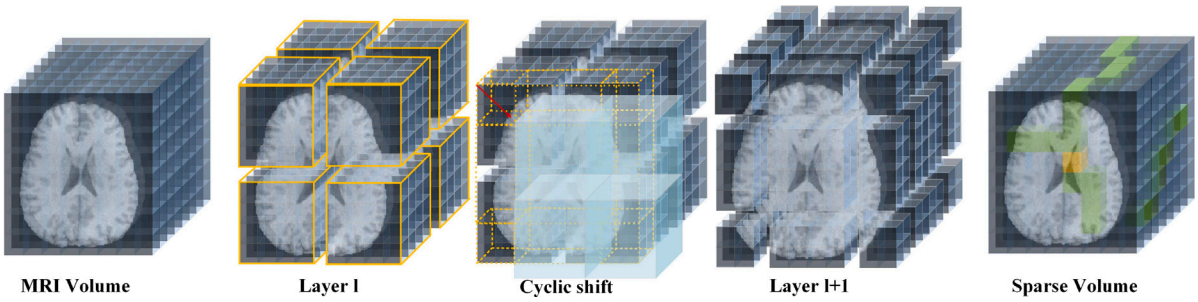

Fig. 4. Overview of shifted window mechanism and sparse dynamic adaptive fusion.

self-attention layers are connected. The self-attention layer fuses multi-head self-attention based on rules and shifted windows with regional dynamic sparsity, respectively, and its feedforward network is a 2-layer MLP with a GELU in the middle. In the SD decoder output feature mapping A_i at the ith stage, the skip connection and 3D patch expansion operations achieve an implicit combination of global features G extracted via the bottleneck layer, feature representations B_i of the encoder module in the symmetric stage, and local edge features F_i obtained from the encoder branches constructed by the MEFF module.

$$A_1 = G \otimes B_{s-1} \otimes F_{s-1}; A_i = A_{i-1} \otimes B_{s-i} \otimes F_{s-i}, i = 2,3 \quad s = 5, \tag{9}$$

where, s represents the total number of global feature extraction stages and i represents the decoder stage. Skip connections cascade the global spatial features extracted by the encoder and the local edge features of the branch with the decoder to merge the outputs at different resolutions. The spatial information lost during the patch merging operation can be recovered, capturing semantic and fine-grained information to predict more accurate feature outputs.

3D Patch Expanding layer: In contrast to the 3D patch merging layer in the SD encoder, the 3D patch expanding layer in the decoder performs an upsampling process on the extracted spatial features. In order to construct an output with the same spatial resolution as the input features, sequence patches need to be recombined to restore the feature dimensions to the original input size. First, a linear mapping is applied to patch expansion to increase the input feature dimension to twice the original dimension. After patch expansion at the bottleneck layer, the input tokens are reshaped along the spatial axes to enhance the feature representation by expanding the resolution of the input features to twice the original and reducing the feature dimensions to a quarter. After going through the three-stage decoder, the connected global spatial features and edge features are fed into the last remaining block to produce the final voxel segmentation prediction using the classifier layer consisting of $1 \times 1 \times 1$ convolutional layer and a sigmoid activation function.

3.4. Multi-level edge feature fusion module

Shallow features have higher resolution and contain rich spatial information. In the context of 3D segmentation tasks, the overlapping area of the perceptual field corresponding to each voxel point is tiny, providing more accurate segmentation target position and edge feature information. While using the UNet variant structure to extract spatial features, the gradual increase in network depth, multi-level feature scale transformation, and abstraction of semantic information enable the network to capture rich semantic information from deep features. However, this process results in partial loss of local feature details with smaller perceptual domains.

As shown in Fig. 5, the MEFF module is proposed to effectively combine low-level detail features with spatial edge information and top-level global spatial features with rich semantic information. The MEFF module constructs bottom-up rather than top-down multi-level fusion features in the form of a feature pyramid. However, unlike the traditional feature pyramid structure [42], MEFF eliminates the top-level feature extraction layer to prevent invalid information from propagating to subsequent stages, specifically for capturing edge detail features. In addition, the pooling layer is removed to minimize computational complexity and prevent the loss of local feature information. Adding the MEFF module to shallow and intermediate layer features enables local features to learn fine-grained edge details for segmentation. The MEFF module acts as an encoder branch to pass multi-layer local fusion output features to the decoder via skip connections.

First, cascaded and parallel branches of the 3D atrous convolution blocks are designed in the MEFF module to convolve multi-level local features B with the spatial and depth dimensions to obtain richer edge detail features. A linear map of the local output feature B is obtained by two $1 \times 1 \times 1$ convolutional layers for classification regression and batch normalization. The features of different layers interact in a layer-by-layer manner, i.e., the output features of one layer are added to the output features of the next layer, and the outputs of multiple branches are fused to obtain richer edge features in an incremental manner.

$$F_1 = f\left(B_1\right), \tag{10}$$

$$F_i = f\left(B_{i-1}\right) \otimes W_\alpha^i + f\left(B_i\right) \otimes W_\beta^i, i = 2,3 \tag{11}$$

where F_i denotes local edge features which is obtained from the encoder branches constructed by the MEFF module, $f(B_i)$ denotes the output features of different layers after the cascaded convolution branch, $\otimes$ denotes feature layer addition, W_α^i denotes the 3D $3 \times 3 \times 3$ convolutional layer, and W_β^i denotes the 3D $1 \times 1 \times 1$ convolutional layer. The 3D $1 \times 1 \times 1$ convolutional layer is used to achieve cross-channel interaction of aggregated multi-scale features while changing dimensions.

3.5. Loss function

The proposed segmentation network is trained using a combined loss function of cross-entropy loss and Dice loss computed in a voxel-wise manner. Specifically, the output of each stage in the decoder is passed to the final expansion block where a combined loss function is applied. The cross-entropy loss function expressed as follows evaluates the accuracy of voxel classification.

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J} y_{i,j} \log\left(\hat{y}_{i,j}\right), \tag{12}$$

where N denotes the total number of voxels, J denotes the number of categories, $y_{i,j}$ and $\hat{y}_{i,j}$ denote the GT and the prediction results of class j on voxel i, respectively.

The Dice loss function given in Eq. (13) is used to calculate the similarity of two sets.

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i}. \tag{13}$$

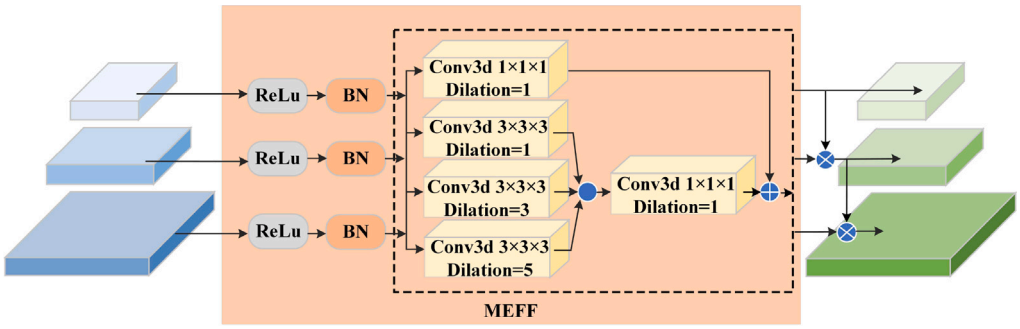

Fig. 5. The proposed multi-layer edge feature fusion module.

where y_i and $\hat{y}_i$ denote the GT and the prediction results respectively.

As given in Eq. (14), the hybrid cross-entropy-Dice loss function combines the cross-entropy loss function and the Dice loss functions for evaluating the accuracy of voxel classification and the similarity of segmentation results, respectively.

$$Lseg = \gamma \cdot L_{CE} + \lambda \cdot L_{\text{Dice}} \tag{14}$$

where γ and λ are two weighting factors used to balance the influence of the two loss functions, $\gamma = 1 - \lambda$. Experiments are conducted within a predefined range, optimizing model performance by adjusting the values of weight parameters. The increase or decrease in experimental performance determines the proportion of the weight parameters. Based on a vast amount of experimental results, the optimal value for γ is 0.3, and accordingly, λ's optimal value is 0.7. With a reasonable setting of the overall loss function, the segmentation performance of the predicted images can gradually approach that of the GT images, from the whole to the boundaries.

4. Experiment

4.1. Dataset and implementation details

In the experiment, the training and testing datasets of the proposed model were obtained from the BraTS2020 and BraTS2021 datasets [43] of the Brain Tumor Segmentation (BraTS) Challenge. The BraTS Challenge, organized by the International Conference on Medical Image Computing and Computer Assisted Intervention Society (MICCAI), serves to evaluate state-of-the-art algorithms in brain tumor segmentation. 3D MRI datasets provided by MICCAI are important public benchmarks. The BraTS datasets consist of multi-parametric MRI scans of brain tumors acquired from multiple institutions in standard clinical settings. These scans were approved by expert neuroradiologists accompanied by voxel-wise truth segmentation labels. The dataset is expanded, removed, or replaced each year to provide more diverse and representative data with challenges. BraTS2020 and BraTS2021 consist of 369 and 1251 annotated subject samples, respectively. We divide samples into 80%, 15% and 5% for training, validation and testing sets, respectively. Each case includes 3D MRI scans in four different modalities: fluid attenuation inversion recovery (FLAIR), T1-weighted (T1), contrast enhanced T1-weighted (T1ce), and T2-weighted (T2). The annotations include three tumor subregions: enhancing tumor (ET), necrosis and non-enhancing tumor (NCR/NET), and peritumoral edematous/invaded tissue (ED). These tumor subregions are clustered into three regions for segmentation evaluation: whole tumor (WT = NCR/NET + ED + ET), tumor core (TC = NCR/NET + ET), and enhancing tumor (ET). In this paper, the performance evaluation was carried out using two commonly-used metrics in medical image segmentation: Dice score and 95% Hausdorff distance (HD95). The Dice score is used to measure the degree of overlap between the predicted segmentation results and the true labels, which can be calculated as:

$$\text{Dice}(X,T) = \frac{2\sum_{i=1}^{I} X_i T_i}{\sum_{i=1}^{I} X_i + \sum_{i=1}^{I} T_i} \tag{15}$$

where X_i and T_i denote the GT and the prediction values for voxel i respectively. The larger the Dice score, the better the segmentation performance. To calculate the distance between segmentation boundaries, the HD95 is utilized. Hausdorff distance(HD) metric can be calculated as:

$$\text{HD}\left(X',T'\right) = \max\left\{\max_{x'\in X' t'\in T'} \left\|x'-t'\right\|, \max_{t'\in T'} \min_{x'\in X'} \left\|t'-x'\right\|\right\} \tag{16}$$

where X' and T' denote ground truth and prediction surface point sets respectively. The HD95 uses the 95th percentile of the distances between ground truth and prediction surface point sets.

Implementation details. The proposed model was implemented using the PyTorch framework. The training process was conducted on one NVIDIA GeForce RTX 4090 GPU. The Adam optimizer was used to train the model. The initial learning rate was set to 0.0001, and a step learning rate policy was employed with a decay rate of 0.9 per iteration. The batch size was set to 4, and the total number of epochs was set to 300.

During the data preprocessing, various data augmentation techniques were dynamically applied to improve the model's generalization ability, including random rotation and scaling, Gaussian noise, additional brightness enhancement, and gamma scaling. After applying data augmentation techniques, the number of training samples used in the Brats2020 and Brats2021 datasets are 295 and 1000, respectively. To reduce computational complexity, a volume cropping strategy was employed, where the non-zero voxels in the brain tumor MRI images were cropped. Since the pixel intensities in MRI were qualitative, voxel normalization was performed by calculating the mean and standard deviation. The 3D image volumes in the dataset, initially sized at $240 \times 240 \times 155$, were randomly cropped to $128 \times 128 \times 128$ voxels. Image sequences were randomized with a 50% probability of adding horizontal flips in the axial, coronal and sagittal planes to increase the robustness of the model to different orientation and viewing angle changes.

4.2. Comparison with state-of-the-art methods

In order to evaluate the effectiveness of the proposed brain tumor segmentation network, SDV-TUNet was compared with state-of-the-art brain tumor segmentation methods. These methods were evaluated on the BraTS2020 and BraTS2021 benchmark datasets. The evaluated methods included 2D and 3D approaches based on encoder–decoder architectures, Transformer-based segmentation methods, and CNN-based methods. Brief descriptions of these methods are given in Table 1. Due to the closed source nature of many existing brain tumor segmentation methods, and to avoid introducing potential bias by retraining models, this paper refers to relevant literature and directly compares the results obtained by the proposed method and quantitative evaluations reported in these publications. This practice is common in brain tumor segmentation research. The quantitative evaluation results on the BraTS2020 and BraTS2021 benchmark datasets are presented in Tables 2 and 3, respectively. The best performing values are highlighted in bold. Figs. 6

Table 1
A brief description of the methods used for performance comparisons in the experiments.

Method	Brief description
nnU-Net [44]	Propose an nnU-Net variant based on region training and data enhancement for brain tumor segmentation.
TransBTS [15]	Propose an encoder–decoder structure consisting of Transformer and U-Net.
Point-Unet [45]	Design a U-Net for fine classification segmentation of the input point cloud.
U-Net++ [27]	Add a series of nested and dense jump connections for segmentation based on U-Net.
Swin-BTS [46]	Propose an encoder structure based on 3D Swin Transformer.
ACMINet [47]	Propose an aligned cross-modal interaction network for segmenting brain tumors and tissue regions from MRI.
nn-UNet [48]	Propose an nn-UNet network with the addition of an axial attention encoder.
Swin UNETR [19]	Design a UNet-shaped network with Swin Transformer as the encoder and CNN as the decoder.
3D ResUNet [49]	Propose a multimodal brain tumor segmentation using 3D ResUNet deep neural network architecture.
H. Peiris [50]	Propose a method for training 3D brain tumor segmentation task based on adversarial learning.
VT-UNet [35]	Propose a UNet-shaped Volume Transformer for multimodal medical image segmentation.
CKD-TransBTS [51]	Propose a clinical knowledge-driven model for brain tumor segmentation.
Segtransvae [52]	Propose a Transformer based on an encoder–decoder architecture combined with a Variational Auto-Encoder (VAE) branch.

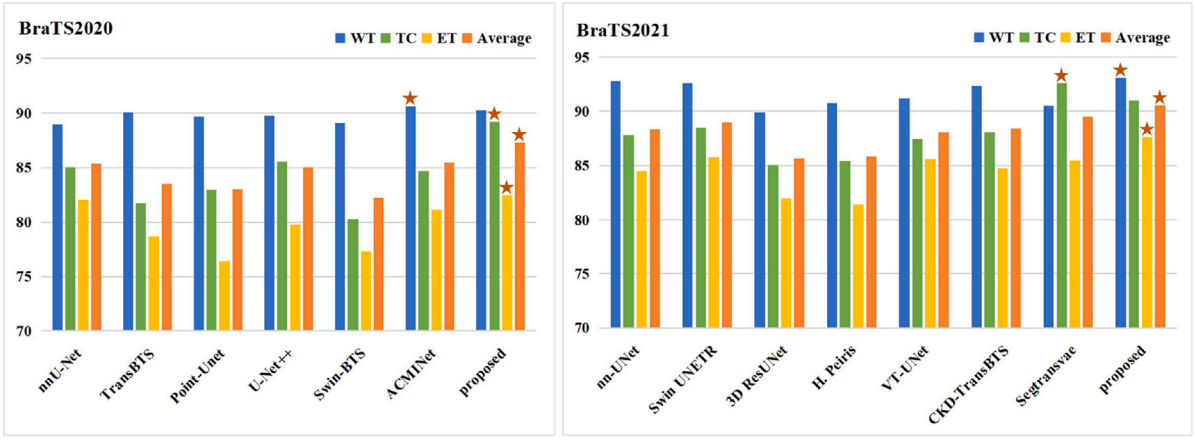

Fig. 6. Performance comparison of different brain tumor segmentation methods on the BraTS2020 and BraTS2021 benchmarks in terms of Dice. The best performing method is marked with an asterisk in each case.

Table 2
Objective evaluation results of different brain tumor segmentation methods on the BraTS2020 benchmark.

Method	WT		TC		ET		Average	
	Dice	HD	Dice	HD	Dice	HD	Dice	HD
nnU-Net [44]	88.95	8.498	85.06	17.337	82.03	17.805	85.35	14.547
TransBTS [15]	90.09	4.964	81.73	9.769	78.73	17.947	83.52	10.893
Point-Unet [45]	89.67	–	82.97	–	76.43	–	83.02	8.260
U-Net++ [27]	89.77	6.299	85.57	5.483	79.83	4.328	85.06	5.370
Swin-BTS [46]	89.06	8.560	80.30	15.780	77.36	26.840	82.24	17.060
ACMINet [47]	**90.61**	4.450	84.70	8.630	81.13	17.500	85.48	10.193
Proposed	90.22	**4.032**	**89.20**	**3.302**	**82.48**	**2.297**	**87.30**	**3.210**

Table 3
Objective evaluation results of different brain tumor segmentation methods on the BraTS2021 benchmark.

Method	WT		TC		ET		Average	
	Dice	HD	Dice	HD	Dice	HD	Dice	HD
nn-UNet [48]	92.75	**3.470**	87.81	7.623	84.51	20.730	88.36	10.610
Swin UNETR [19]	92.60	5.831	88.50	3.770	85.80	6.016	88.97	5.206
3D ResUNet [49]	89.90	4.300	85.03	9.890	81.96	17.890	85.63	10.693
H. Peiris [50]	90.77	5.369	85.39	8.563	81.39	21.830	85.85	11.921
VT-UNet [35]	91.20	10.030	87.41	6.590	85.59	6.230	88.07	7.520
CKD-TransBTS [51]	92.33	4.230	88.07	4.390	84.76	3.160	88.39	3.927
Segtransvae [52]	90.52	3.570	**92.60**	5.840	85.48	2.890	89.53	4.100
Proposed	**93.10**	3.584	90.99	**3.278**	**87.64**	**2.576**	**90.58**	**3.146**

and 7 show the performance of different brain tumor segmentation methods in terms of the Dice and HD95 metrics, respectively. In each case, the best performing method is marked with an asterisk in the corresponding column.

The proposed method exhibits clear advantages over state-of-the-art segmentation methods, as indicated by the results presented in the above tables and figures. In terms of average Dice scores on the BraTS2020 and BraTS2021 benchmarks, the proposed method outperforms other comparable methods by margins ranging from 1.82% to 5.06% and 1.05% to 4.95%, achieving scores of 87.30% and 90.58%, respectively. Additionally, the average HD95 metric for the proposed method decreases to 3.210 and 3.146 on the BraTS2020 and BraTS2021 benchmarks, respectively. Notably, the proposed method shows the most substantial improvement in segmenting the tumor core compared with TransBTS which is the first attempt to utilize Transformers in 3D CNN for 3D MRI brain tumor segmentation. It also demonstrates superior segmentation results in the ET region, which contains fewer voxels. Furthermore, compared with Swin UNETR, a 3D brain tumor segmentation network that utilizes hierarchical Swin Transformer as an encoder, and VT-UNet, which employs volume Transformers for encoding and decoding, the proposed network demonstrates significant improvements in tumor core segmentation and enhancing tumor region, resulting in more competitive HD95 results. By analyzing objective evaluation metrics, the proposed method exhibits the capability to extract global features from both spatial and deep dimensions while enhancing the processing capacity of local and edge information. Consequently, the proposed method achieves sharper segmentation boundaries.

Furthermore, utilizing 2D brain tumor segmentation, which involves decomposing a 3D MRI volume into 2D slices and passing them to a segmentation model, may result in loss of volumetric information. Compared with 2D brain tumor segmentation methods, analyzing Dice and HD95 segmentation performance evaluation metrics for 3D brain tumor segmentation networks can quantitatively verify the efficient use of volumetric and spatial information in 3D data. The overall processing

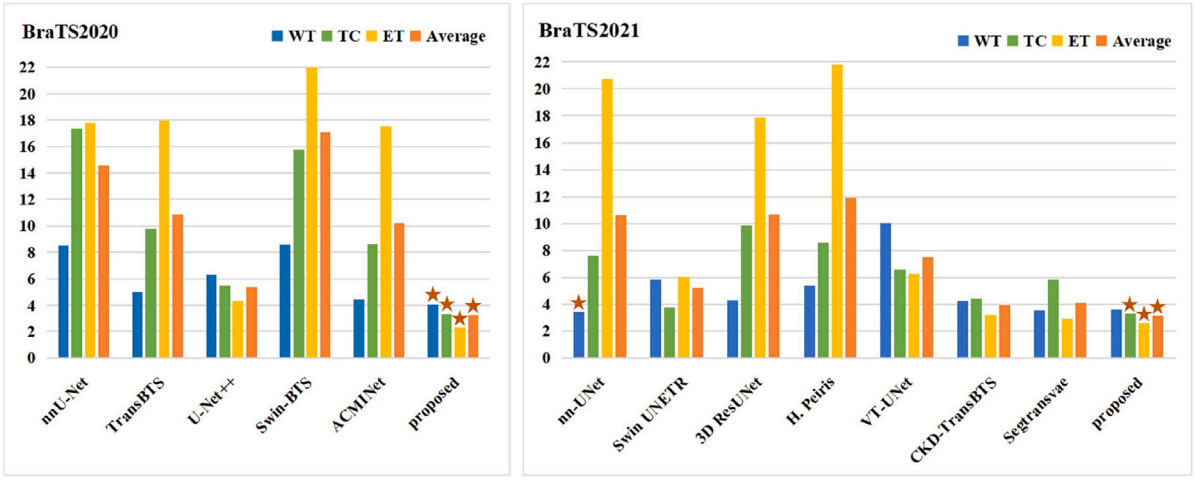

Fig. 7. Performance comparison of different brain tumor segmentation methods on the BraTS2020 and BraTS2021 benchmarks in terms of HD95. The best-performing method is marked with an asterisk in each case.

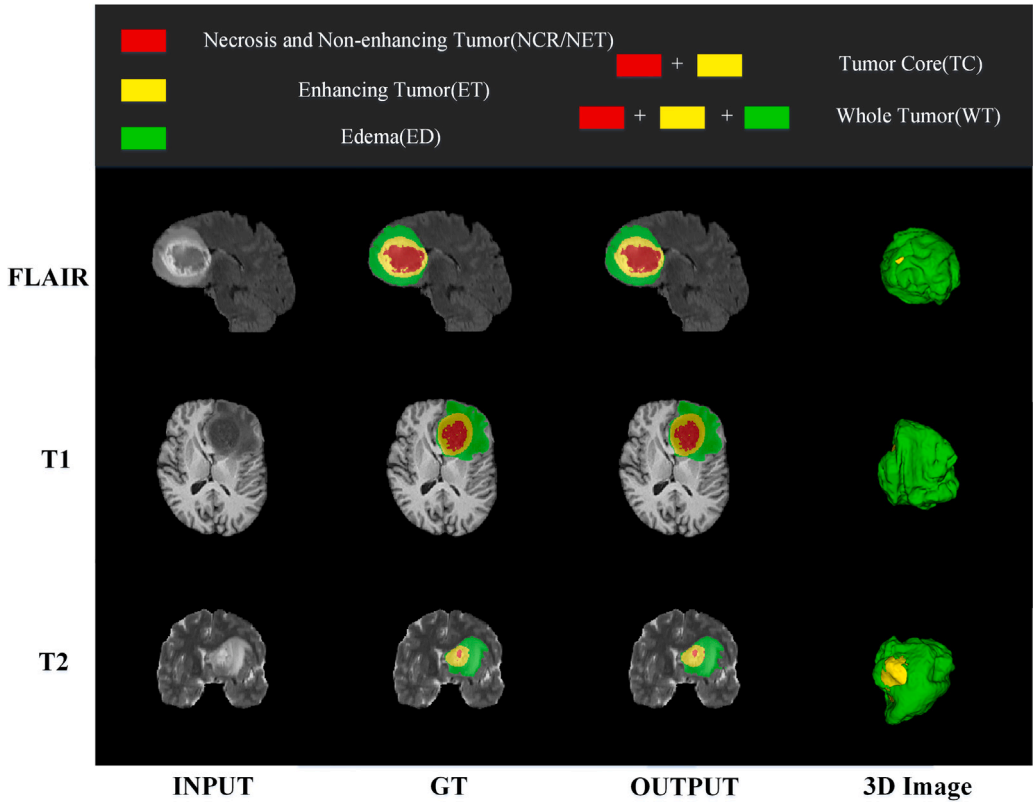

Fig. 8. 2D and 3D segmentation effects of different modality imaging of 3D segmentation network. Green, yellow and red indicate the ED, ET and NCR/NET regions, respectively.

of volumetric data and the extraction of 3D MRI voxel features play a crucial role in the performance of the proposed model. Therefore, it is of great value to improve the structure of the brain tumor image segmentation network in terms of spatial and volume information. Fig. 8 shows the visual comparison between the 2D and 3D views of MRI brain tumor segmentation results versus actual images. The original images are selected from the training set of BraTS2021. The red region represents necrosis and non-enhancing tumor, the yellow region indicates the enhancing tumor, and the green region represents the peritumoral edematous/invaded tissue. The proposed method achieves accurate and robust brain tumor segmentation. Fig. 9 shows a visual comparison of the brain tumor segmentation results obtained by different methods. By referring to the ground truth (GT), the proposed method obtains more accurate segmentation results compared with other methods, especially in terms of tumor edge segmentation and capturing independent voxel points.

4.3. Ablation study

To verify and evaluate the proposed model, as well as the effectiveness of each component, ablation study was conducted to analyze the segmentation effect of different components. These included the SD encoder–decoder module, which incorporated a sparse dynamic adaptive strategy, and the MEFF module. Three combinations of different components was compared to illustrate the effects of different components from quantitative and qualitative perspectives, respectively. To

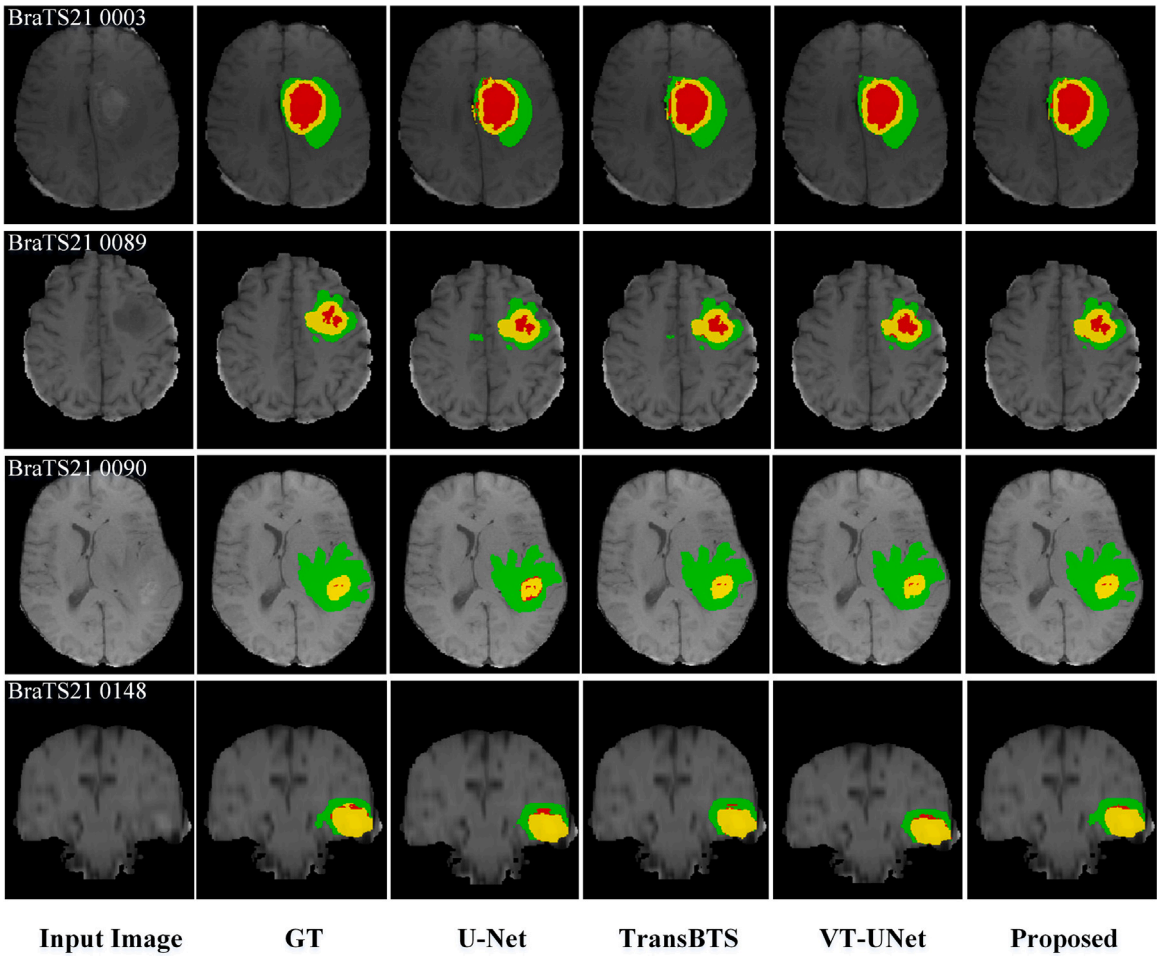

Fig. 9. Comparison of visual results of brain tumor segmentation results obtained by different methods.

Table 4
Objective evaluation results of the BraTS2020 baseline ablation study.

	WT		TC		ET		Average	
	Dice	HD	Dice	HD	Dice	HD	Dice	HD
Baseline	89.61	4.478	86.74	3.770	80.45	3.067	85.60	3.772
Baseline+SD	89.90	4.256	88.17	3.478	82.07	2.691	86.71	3.475
Baseline+SD+MEFF	**90.22**	**4.032**	**89.20**	**3.302**	**82.48**	**2.297**	**87.30**	**3.210**

Table 5
Objective evaluation results of the BraTS2021 baseline ablation study.

	WT		TC		ET		Average	
	Dice	HD	Dice	HD	Dice	HD	Dice	HD
Baseline	91.17	4.741	88.20	4.446	84.38	3.520	87.92	4.236
Baseline+SD	92.37	3.930	90.19	3.582	86.40	2.776	89.65	3.429
Baseline+SD+MEFF	**93.10**	**3.584**	**90.99**	**3.278**	**87.64**	**2.576**	**90.58**	**3.146**

fully verify the proposed model and the impact of adding components, ablation experiments were performed on BraTS2020 and BraTS2021 datasets. The quantitative results are presented in Tables 4 and 5, where the changes in the values of Dice metric and HD95 metric visually indicate the enhanced performance of the components to a certain extent. The qualitative analysis is shown in Fig. 10.

In these experiments, the Swin Transformer-based encoder–decoder network [19] was used as the baseline model, and then different combinations of modules were gradually added to verify their effectiveness. Specifically, the effects of the following models were compared.

Baseline model: An encoder–decoder network using Swin Transformer block acts as the baseline model for 3D brain tumor segmentation. As shown in Tables 4 and 5, the baseline model, when directly applied to the dataset, exhibits the worst test results in terms of Dice and HD95 metrics across different regions.

Baseline model + SD encoder–decoder: The encoder and decoder using Swin Transformer block is changed to SD encoder–decoder module with a sparse dynamic adaptive strategy, which extracts global spatial semantic feature of brain tumor segmentation based on the baseline model. The effectiveness of SD encoder–decoder is evaluated by comparing the baseline model with the baseline model + SD encoder–decoder. As shown in the second row of Tables 4 and 5, the model with the addition of SD encoder–decoder with sparse dynamic adaptive strategy achieved significant performance improvement compared with the base model. In particular, the Dice metrics improved by 1.43% and 1.62% for TC and ET on the BraTS2020 dataset, and by 1.20%, 1.99%, and 2.02% for WT, TC, and ET on the BraTS2021 dataset, respectively. The comparison of the quantitative results in the first and second rows of the tables shows the advantages of using SD encoder–decoder for global interaction modeling and extraction of global spatial features. Additionally, combined with the qualitative results as shown in the second column of Fig. 10, the global semantic segmentation is more accurate and can accurately segment long-range independent voxel points compared with the baseline model in the first column, which demonstrates the effectiveness of the SD encoder–decoder module, achieving the perception interaction of regional connections.

Baseline model + SD encoder–decoder + MEFF: The proposed final model includes an SD module with a sparse dynamic adaptive

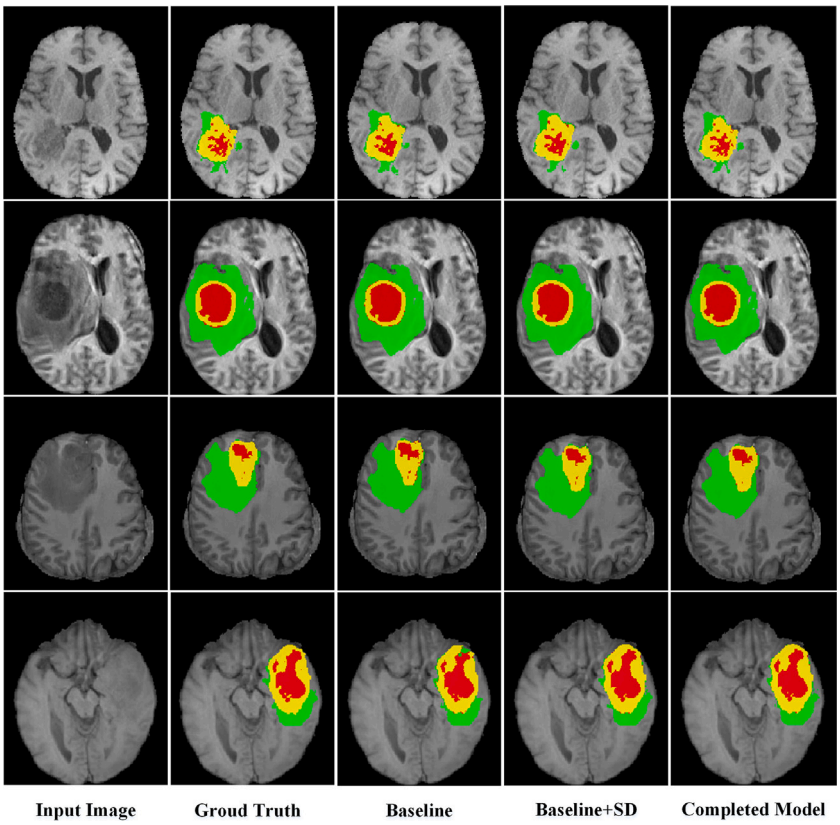

Fig. 10. Visual comparison of the segmentation results of different models in the ablation study.

strategy and an MEFF module as the encoder branch for edge feature extraction. The MEFF module was introduced to comprehensively capture local details and edge features. The effectiveness of the edge detection module can be verified by comparing the baseline model + SD encoder–decoder with the final model (the MEFF module cannot be used alone without the global semantic segmentation network). The quantitative results show that the addition of the MEFF module as a branch of the encoder improves the segmentation results in general. The boundaries are sharper. The Dice metrics are a bit improved and HD95 is reduced compared with both the baseline model + SD encoder–decoder. The Dice metrics of TC improved by 1.03% on the BraTS2020 dataset and ET improved by 1.24% on the BraTS2021 dataset. From the qualitative analysis, the complete model in Fig. 10 has sharper and less disturbing segmentation boundaries compared with the segmentation without the MEFF module added, which can achieve stable and accurate segmentation compared with ground truth.

4.4. Visual analysis of the proposed module

Fig. 11 provides a visualization of the segmentation results of different models with added components, including the SD encoder–decoder and MEFF modules. The model without SD encoder–decoder and MEFF modules can relatively accurately segment the edema region (shown in green region). Although the proposed model can describe the general outline of brain tumors, it cannot effectively learn the global spatial features and process details. The presence of interference in the necrosis and non-enhancing tumor (shown in red region) indicates unstable segmentation results. The addition of SD encoder–decoder eliminates some unnecessary interference noise, the segmentation in the global range is more accurate, and some relative independent long-range voxel points can be accurately segmented. The complete model with SD encoder–decoder and MEFF module achieves significantly more accurate and more stable segmentation compared with the baseline model. Moreover, it performs well in capturing local detail edges compared with the model with only an SD encoder–decoder module. The proposed model with SD encoder–decoder and MEFF modules obtains almost the same segmentation results as ground truth due to tight and sparse combination and local-to-global feature modeling.

5. Conclusion

This paper proposes a deep learning-based 3D brain tumor segmentation network SDV-TUNet to improve segmentation accuracy by fully integrating intra-axis information, inter-layer feature connections, and voxel information. The proposed method consists of the SD encoder–decoder module and the multi-level edge feature fusion module called MEFF (multi-level edge feature fusion). The SD encoder–decoder module is used to extract global spatial semantic features in brain tumor segmentation. The encoding stage incorporates multi-head self-attention and sparse dynamic adaptive fusion into a 3D extended shifted window strategy, enabling effective integration of local tight correlations and long-range sparse dependencies for dense prediction tasks. The MEFF module is applied to capture local details and edge features. It uses cascaded and parallel 3D atrous convolution blocks to fuse detailed features from different layers in a layer-by-layer incremental manner, connecting them to the decoding stage via skip connections. Experimental results demonstrate the effectiveness of the proposed key components. Moreover, the proposed method outperforms state-of-the-art methods on the BraTS benchmark. In future work, the proposed 3D segmentation method will conduct lightweight model design, weigh computational resources and segmentation performance, optimize and extend to other medical image segmentation tasks and the analysis of video images with spatial and temporal dimensions.

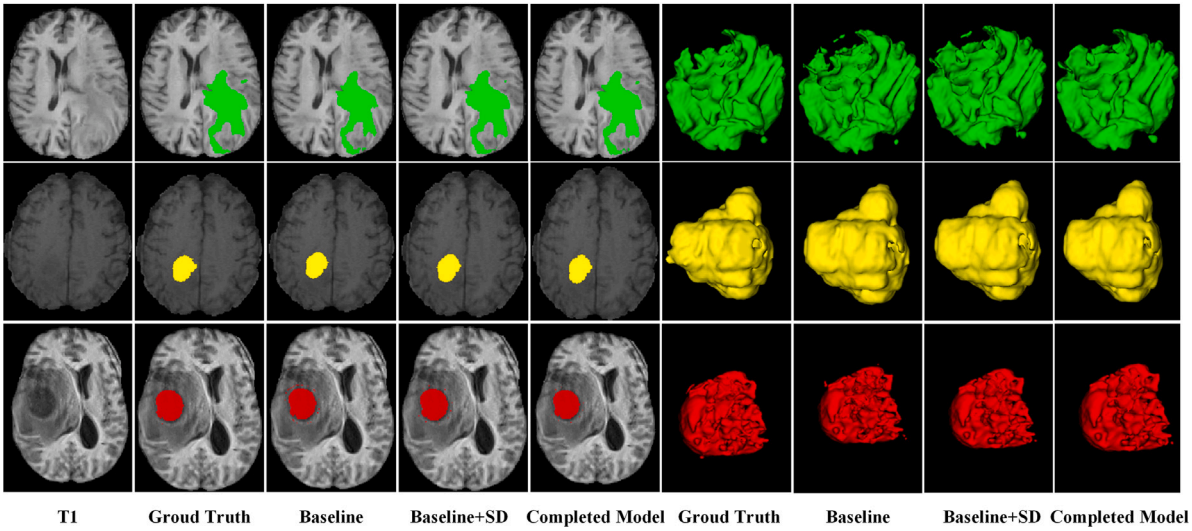

Fig. 11. Comparison of 2D and 3D visual effects of the model segmentation results with the addition of different components. One specific categories are shown for better visualization. Green, yellow and red indicate the ED, ET and NCR/NET regions, respectively.

CRediT authorship contribution statement

Zhiqin Zhu: Methodology. **Mengwei Sun:** Writing – original draft, Software. **Guanqiu Qi:** Writing – review & editing. **Yuanyuan Li:** Investigation. **Xinbo Gao:** Data curation. **Yu Liu:** Validation.

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.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grants 62176081, 62276037 and U23A20294), Cooperation Project Between Undergraduate Universities in Chongqing and Institutions Affiliated to the Chinese Academy of Sciences (No. HZ2021018), Special key project of Chongqing technology innovation and application development: CSTB2022TIAD-KPX0039, and Development Fund of Key Laboratory of Chongqing University Cancer Hospital (cquchkfjj005).

References

[1] F. Piccialli, V. Di Somma, F. Giampaolo, S. Cuomo, G. Fortino, A survey on deep learning in medicine: Why, how and when? Inf. Fusion 66 (2021) 111–137.
[2] S. Mo, M. Cai, L. Lin, R. Tong, Q. Chen, F. Wang, H. Hu, Y. Iwamoto, X.-H. Han, Y.-W. Chen, Multimodal priors guided segmentation of liver lesions in MRI using mutual information based graph co-attention networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI, 2020, pp. 429–438.
[3] Y. Liu, Y. Shi, F. Mu, J. Cheng, X. Chen, Glioma segmentation-oriented multi-modal MR image fusion with adversarial learning, IEEE/CAA J. Autom. Sin. 9 (8) (2022) 1528–1531.
[4] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[5] A. Wadhwa, A. Bhardwaj, V.S. Verma, A review on brain tumor segmentation of MRI images, Magn. Reson. Imaging 61 (2019) 247–259.
[6] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[7] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI, Springer, 2015, pp. 234–241.
[8] Z. Wang, T. Li, J.-Q. Zheng, B. Huang, When cnn meet with vit: Towards semi-supervised learning for multi-class medical image semantic segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 424–441.
[9] W. Wang, D. Qin, S. Wang, Y. Fang, Y. Zheng, A multi-channel UNet framework based on SNMF-DCNN for robust heart-lung-sound separation, Comput. Biol. Med. (2023) 107282.
[10] H. Zhao, X. Qiu, W. Lu, H. Huang, X. Jin, High-quality retinal vessel segmentation using generative adversarial network with a large receptive field, Int. J. Imaging Syst. Technol. 30 (3) (2020) 828–842.
[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[12] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, 2021, arXiv preprint arXiv:2102.04306.
[13] Y. Liu, Y. Shi, F. Mu, J. Cheng, C. Li, X. Chen, Multimodal MRI volumetric data fusion with convolutional neural networks, IEEE Trans. Instrum. Meas. 71 (2022) 1–15.
[14] F. Wang, R. Jiang, L. Zheng, C. Meng, B. Biswal, 3D u-net based brain tumor segmentation and survival days prediction, in: International MICCAI Brainlesion Workshop, MICCAI, Springer, 2019, pp. 131–141.
[15] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, J. Li, Transbts: Multimodal brain tumor segmentation using transformer, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI, Springer, 2021, pp. 109–119.
[16] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al., A survey on vision transformer, IEEE Trans. Pattern Anal. Mach. Intell. 45 (1) (2022) 87–110.
[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[18] J. Zhang, Q. Qin, Q. Ye, T. Ruan, ST-unet: Swin transformer boosted U-net with cross-layer feature enhancement for medical image segmentation, Comput. Biol. Med. 153 (2023) 106516.
[19] 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, MICCAI, Springer, 2021, pp. 272–284.
[20] X. Chen, X. Wang, J. Zhou, Y. Qiao, C. Dong, Activating more pixels in image super-resolution transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22367–22377.
[21] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, I.B. Ayed, HyperDense-net: a hyper-densely connected CNN for multi-modal image segmentation, IEEE Trans. Med. Imaging 38 (5) (2018) 1116–1126.
[22] K. Kamnitsas, C. Ledig, V.F. Newcombe, J.P. Simpson, A.D. Kane, D.K. Menon, D. Rueckert, B. Glocker, Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, Med. Image Anal. 36 (2017) 61–78.
[23] Y. Liu, H. Li, J. Cheng, X. Chen, MSCAF-net: a general framework for camouflaged object detection via learning multi-scale context-aware features, IEEE Trans. Circuits Syst. Video Technol. 33 (2023) 4934–4947.
[24] 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.

[25] X. Zhou, Y. Chen, Z. Wu, A.A. Heidari, H. Chen, E. Alabdulkreem, J. Escorcia-Gutierrez, X. Wang, Boosted local dimensional mutation and all-dimensional neighborhood slime mould algorithm for feature selection, Neurocomputing (2023) 126467.
[26] B. Shi, J. Chen, H. Chen, W. Lin, J. Yang, Y. Chen, C. Wu, Z. Huang, Prediction of recurrent spontaneous abortion using evolutionary machine learning with joint self-adaptive sime mould algorithm, Comput. Biol. Med. 148 (2022) 105885.
[27] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imaging 39 (6) (2019) 1856–1867.
[28] Z. Zhang, Q. Liu, Y. Wang, Road extraction by deep residual u-net, IEEE Geosci. Remote Sens. Lett. 15 (5) (2018) 749–753.
[29] Y. Liu, F. Mu, Y. Shi, X. Chen, Sf-net: A multi-task model for brain tumor segmentation in multimodal mri via image fusion, IEEE Signal Process. Lett. 29 (2022) 1799–1803.
[30] H.R. Roth, H. Oda, X. Zhou, N. Shimizu, Y. Yang, Y. Hayashi, M. Oda, M. Fujiwara, K. Misawa, K. Mori, An application of cascaded 3D fully convolutional networks for medical image segmentation, Comput. Med. Imaging Graph. 66 (2018) 90–99.
[31] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods 18 (2) (2021) 203–211.
[32] Z. Liu, L. Tong, L. Chen, F. Zhou, Z. Jiang, Q. Zhang, Y. Wang, C. Shan, L. Li, H. Zhou, Canet: Context aware network for brain glioma segmentation, IEEE Trans. Med. Imaging 40 (7) (2021) 1763–1777.
[33] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H.R. Roth, D. Xu, Unetr: Transformers for 3d medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 574–584.
[34] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, Y. Yu, Nnformer: Interleaved transformer for volumetric segmentation, 2021, arXiv preprint arXiv:2109.03201.
[35] 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, MICCAI, Springer, 2022, pp. 162–172.
[36] M.K. Chandrakar, A. Mishra, Brain tumor detection using multipath convolution neural network (CNN), Int. J. Comput. Vis. Image Process. (IJCVIP) 10 (4) (2020) 43–53.
[37] J. Zhao, Q. Li, X. Li, H. Li, L. Zhang, Automated segmentation of cervical nuclei in pap smear images using deformable multi-path ensemble model, in: 2019 IEEE 16th International Symposium on Biomedical Imaging, ISBI 2019, IEEE, 2019, pp. 1514–1518.
[38] N. Wang, S. Lin, X. Li, K. Li, Y. Shen, Y. Gao, L. Ma, MISSU: 3D medical image segmentation via self-distilling TransUNet, IEEE Trans. Med. Imaging (2023) 1–1.
[39] Y. Zhang, Z. Bai, Y. You, X. Liu, X. Xiao, Z. Xu, Multi-path feature fusion and channel feature pyramid for brain tumor segmentation in MRI, in: International Conference on Image and Graphics, Springer, 2023, pp. 26–36.
[40] Z. Zhou, Z. He, Y. Jia, AFPNet: A 3D fully convolutional neural network with atrous-convolution feature pyramid for brain tumor segmentation via MRI images, Neurocomputing 402 (2020) 235–244.
[41] Y. Li, W.-G. Cui, H. Huang, Y.-Z. Guo, K. Li, T. Tan, Epileptic seizure detection in EEG signals using sparse multiscale radial basis function networks and the Fisher vector approach, Knowl.-Based Syst. 164 (2019) 96–106.
[42] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
[43] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F.C. Kitamura, S. Pati, et al., The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification, 2021, arXiv preprint arXiv:2107.02314.
[44] F. Isensee, P.F. Jäger, P.M. Full, P. Vollmuth, K.H. Maier-Hein, nnU-net for brain tumor segmentation, in: International MICCAI Brainlesion Workshop, MICCAI, Springer, 2021, pp. 118–132.
[45] N.-V. Ho, T. Nguyen, G.-H. Diep, N. Le, B.-S. Hua, Point-unet: A context-aware point-based neural network for volumetric segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI, Springer, 2021, pp. 644–655.
[46] 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) (2022) 797.
[47] Y. Zhuang, H. Liu, E. Song, C.-C. Hung, A 3D cross-modality feature interaction network with volumetric feature alignment for brain tumor and tissue segmentation, IEEE J. Biomed. Health Inf. 27 (1) (2022) 75–86.
[48] H.M. Luu, S.-H. Park, Extending nn-UNet for brain tumor segmentation, in: International MICCAI Brainlesion Workshop, MICCAI, Springer, 2021, pp. 173–186.
[49] L. Pei, Y. Liu, Multimodal brain tumor segmentation using a 3D ResUNet in BraTS 2021, in: International MICCAI Brainlesion Workshop, MICCAI, Springer, 2021, pp. 315–323.
[50] 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, MICCAI, Springer, 2021, pp. 171–181.
[51] J. Lin, J. Lin, C. Lu, H. Chen, H. Lin, B. Zhao, Z. Shi, B. Qiu, X. Pan, Z. Xu, et al., CKD-TransBTS: clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation, IEEE Trans. Med. Imaging (2023) 1–1.
[52] Q.-D. Pham, H. Nguyen-Truong, N.N. Phuong, K.N. Nguyen, C.D. Nguyen, T. Bui, S.Q. Truong, Segtransvae: Hybrid cnn-transformer with regularization for medical image segmentation, in: 2022 IEEE 19th International Symposium on Biomedical Imaging, ISBI, IEEE, 2022, pp. 1–5.