Biomedical Signal Processing and Control 95 (2024) 106484

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

MVSI-Net: Multi-view attention and multi-scale feature interaction for brain tumor segmentation

Junding Sun [a,*], Ming Hu [a], Xiaosheng Wu [a], Chaosheng Tang [a], Husam Lahza [b], Shuihua Wang [a,c,d], Yudong Zhang [a,b,d,**]

[a] School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454003, China
[b] Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[c] Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, China
[d] School of Computing and Mathematical Sciences, University of Leicester, Leicester LE1 7RH, UK

ARTICLE INFO

Keywords:
MRI brain tumor segmentation
Attention mechanism
Multi-scale feature extraction
U-Net architecture

ABSTRACT

Brain tumor segmentation using MRI remains a challenging task due to the high incidence and complexity of gliomas. The irregular variations in tumor location, size, shape, and unclear edge contours of diverse tumor categories contribute to subpar segmentation accuracy. To address these issues, we propose MVSI-Net, a novel MRI brain tumor segmentation method that integrates a multi-view attention mechanism and multi-scale feature interaction into the UNet architecture. Our approach proposes a multi-view attention mechanism that captures global and local features from three different perspectives: channel, content, and position. This mechanism facilitates the localization of the target region and enhances feature representation in lesion areas. Additionally, we design a multi-scale feature interaction module that selectively extracts valuable information from multiple receptive fields of varying sizes, promoting cross-dimensional interaction. As a result, our method enables precise segmentation of the edge contours of different tumor categories. To evaluate the performance of MVSI-Net, we conducted experiments on three widely used datasets: BraTs 2019, BraTs 2020, and BraTs 2021. The experimental results demonstrate that our proposed method outperforms similar approaches in brain tumor segmentation accuracy. In conclusion, our study presents a novel and effective MRI brain tumor segmentation method that addresses the challenges posed by gliomas. However, our model still has certain limitations. Firstly, the model has not been applied in clinical experiments, and there may be challenges in terms of accuracy in certain complex cases. Secondly, further exploration is required to assess the model's generalization capability beyond specific medical image datasets. Moving forward, we plan to address these limitations in future research.

1. Introduction

Brain tumors [1,2] are abnormal cells that grow in the brain and can be classified as either benign or malignant. Benign tumors can be cured through surgery, while malignant tumors, also known as brain cancer, are one of the deadliest types of cancer and can directly cause patient deaths [3]. Therefore, accurate segmentation of brain tumors is of great significance for patient diagnosis and surgical treatment. Magnetic Resonance Imaging (MRI) is a non-invasive imaging technique that can generate high-quality brain images without damage or skull artifacts, providing comprehensive information for brain tumor diagnosis [4]. In the field of medical image analysis for brain tumors, MRI is commonly used to obtain four modalities (T1, T2, T1CE, and FLAIR) of images. Each modality provides different information for brain tumor imaging [5]. Fig. 1 displays the four modalities of images and the manually segmented results by experts for the same patient. Specifically, T1 imaging observes brain anatomical structures, T1CE imaging assesses the enhanced region of the tumor, T2 imaging evaluates the whole tumor region, and FLAIR identifies peritumoral edema. In clinical practice, doctors typically rely on the patient's condition and the four modalities generated by MRI to diagnose brain tumors. However, manual segmentation methods rely heavily on doctors' prior knowledge, resulting

* Corresponding author.
** Corresponding author at: School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454003, China.
E-mail addresses: sunjd@hpu.edu.cn (J. Sun), hm@home.hpu.edu.cn (M. Hu), wuxs@hpu.edu.cn (X. Wu), tcs@hpu.edu.cn (C. Tang), hlahza@kau.edu.sa (H. Lahza), shuihuawang@ieee.org (S. Wang), yudongzhang@iee.org (Y. Zhang).

https://doi.org/10.1016/j.bspc.2024.106484
Received 8 September 2023; Received in revised form 28 December 2023; Accepted 20 May 2024
Available online 25 May 2024

in variability in annotations among experts. Additionally, manual segmentation is time-consuming when dealing with a large volume of MRI brain tumor data. Therefore, accurate automatic tumor segmentation methods play a crucial role in clinical diagnosis.

Due to the complexity of brain tissue structure and the similarity of specific tissue structures, distinguishing between normal and pathological tissues is challenging. Furthermore, brain tumors exhibit irregularities in spatial position, shape, and size, as well as significant variations. Additionally, there is an issue of class imbalance as the pixel quantities within different tumor regions are not uniform. The boundaries of brain tumors are often blurred, making automatic segmentation of brain tumors a challenging task.

Classic segmentation methods for MRI brain tumors include atlas-based methods [6], threshold-based methods [7], and region-growing methods [8]. Sun et al. [9] proposed a Boltzmann method that utilizes the mean and covariance of pixels surrounding the lesion to extract higher-level features. Liu et al. [10] classified brain structures such as the thalamus, basal ganglia, and pallidum using probabilistic atlases, which provide probability values for each pixel. Jiang et al. [11] performed a rough segmentation of brain tumors by considering the gray-scale distribution of pixels through statistical analysis. Sharma et al. [12] applied Otsu's method in a differential evolution algorithm to obtain skull-stripped images. Meier et al. [13] proposed a brain tumor segmentation method that utilizes contextual information as distinguishing features, using a decision forest as the classifier to label seven categories. Zhao et al. [14] extended the method of multiscale feature extraction to achieve better detection results. Although these methods have the advantages of fast computation and ease of development, their segmentation performance is relatively low. They are sensitive to pixel correlations, contrast variations, and noise, making it difficult to handle small-sized images.

In recent years, deep learning has rapidly advanced in clinical medicine and has been successfully applied in various areas of medical image analysis. The use of deep learning methods addresses the biases caused by manual processing and overcomes the limitations of traditional segmentation methods, such as low accuracy and poor segmentation results. With the development of deep learning, Convolutional Neural Networks (CNN) have been widely adopted in brain tumor segmentation due to their excellent feature learning capabilities [15]. Long et al. [16] transformed fully connected layers into convolutional layers and proposed a pixel-level semantic segmentation network called FCN, which can classify each pixel end-to-end and retain spatial information by producing output results with the same size as the input image. However, FCN suffers from the issue of imprecise edge segmentation. To improve the accuracy of brain tumor edge segmentation, Shen et al. [17] introduced a tree-structured multi-task FCN to enhance the contextual relationships among pixels in the original FCN architecture. Nevertheless, the limited training data in medical image analysis often leads to unsatisfactory segmentation results for lesion regions. To address this issue, Ronneberger et al. [18] proposed the UNet network, composed of an encoder, a decoder, and skip connections, which can learn feature representations well with a small dataset. As a result, UNet has been widely applied in the field of biomedical image segmentation. Dong et al. [19] pioneered applying UNet to brain tumor segmentation. They employed skip connections to concatenate high-level semantic information with low-level semantic information and considered detailed information during image restoration to achieve higher segmentation performance. The outstanding performance of UNet in semantic segmentation has inspired many researchers to study it, resulting in numerous variations of UNet. Aghalari et al. [20] addressed the issues of irregular tumor shape and blurry boundaries by designing a dual-path residual network based on UNet, which can extract local and global features simultaneously. Ibtehaz et al. [21] rethought the U-shaped encoder-decoder structure and proposed the MultiResUNet network, which replaces 5×5 and 7×7 convolutions with a series of 3×3 convolutions to improve operational efficiency and reduce computational complexity. FAN et al. [22] introduced a multi-scale attention network for tumor segmentation, which integrates adaptive global and local dependencies by incorporating self-attention mechanisms. NAHIAN et al. [23] provided a comprehensive overview of UNet and its variants. To enhance segmentation performance, more advanced techniques such as dense connection blocks, dilated convolutions, attention mechanisms, and multi-scale feature extraction have been integrated into the U-shaped structure. However, many improved architectures focus solely on deepening the network or stacking other methods, neglecting the potential connections between the network structure and modules. Based on the progressive feature extraction and restoration from shallow to deep layers, the addition of modules with specific functions at different key positions in the network is necessary.

In response to the abovementioned issues, we propose MVSI-Net: an MRI brain tumor segmentation method that integrates a multi-view attention (MVA) mechanism and multi-scale feature interaction (MSI). Because shallow networks only capture low-level edge and texture information, their contribution to the segmentation results is limited. On the other hand, deep networks can learn more semantic information, achieving target localization and better representing the entire lesion region. Therefore, we add MVA to the last two layers of the encoder. Additionally, while the skip connections in UNet combine high-level semantic information with shallow texture information, the fusion of shallow and deep networks introduces unnecessary non-lesion information, leading to additional computational complexity. Thus, we apply MVA in the last two layers of the decoder to extract lesion-related information and remove redundancy.

Furthermore, the bottleneck layer contains abundant target information and is crucial in feature extraction. We incorporate MSI at the bottleneck layer to obtain the required scale features and better segment different tumor boundary regions. However, our model also has certain limitations regarding clinical experimentation and generalization. In the future, we aim to apply the model in clinical experiments further to improve its adaptability and accuracy in different complex cases. In summary, our research makes the following main contributions:

(1) Considering the different features extracted at different positions in the network, we propose the multi-view attention (MVA) mechanism and incorporate it in the last two layers of the encoder and decoder. MVA allows for focusing on regions of interest,

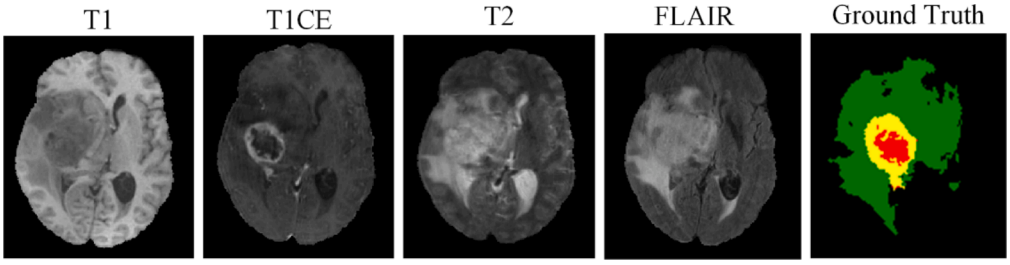

Fig. 1. Four modal maps and label maps of the 001 sample in the BraTs 2020 training set.

eliminating redundant information, and achieving target localization while enhancing the capability to represent features in the lesion region.

(2) To better utilize high-level semantic information in the deep layers of feature extraction, we propose the multi-scale feature interaction (MSI) module at the bottleneck layer. MSI enables comprehensive learning of feature representation from different receptive fields, facilitating precise segmentation of tumor boundaries for different tumor categories.

(3) This paper proposes a new, fully automatic method, MVSI-Net, for MRI brain tumor segmentation. This method combines attention mechanism and cross-dimensional feature interaction to learn tumor information, which can better focus on the lesion area and realize MRI brain tumor precise segmentation.

(4) Experiments were conducted on three publicly available datasets, BraTs 2019, BraTs 2020, and BraTs 2021, to evaluate the performance of our model. The results demonstrate that our proposed method achieved Dice scores of 86.4 %, 88.4 %, and 80.1 % for segmenting the whole tumor (WT) region, tumor core (TC) region, and enhancing tumor (ET) region, respectively, on the BraTs 2019 dataset. Similarly, on the BraTs 2020 dataset, our method achieved Dice scores of 87.6 %, 87.5 %, and 81.7 % for segmenting the WT, TC, and ET regions, respectively. On the BraTs 2021 dataset, our method achieved Dice scores of 85.3 %, 86.1 %, and 78.1 % for segmenting the WT, TC, and ET regions, respectively. These scores surpassed those of comparable methods in segmentation accuracy.

2. Related work

2.1. Attention mechanism

In recent years, attention-based methods have become increasingly popular in deep learning. They have been widely applied in various computer vision domains due to their ability to capture long-range dependencies and exhibit effective adaptability. The attention mechanism enables the network to focus on regions of interest and eliminate redundant information.

Recently, researchers have proposed numerous attention-based approaches for brain tumor image segmentation. Maji et al. [24] proposed an attention-guided decoder network for brain tumor segmentation. Based on the baseline ResUNet architecture, an attention gate was introduced in the decoder to focus on the lesion region and reduce the possibility of false positives. Zhang et al. [25] presented a novel attention-gate Resunet model known as AGResU-Net. This network integrated residual modules and attention gates into the original UNet architecture, adding a series of attention gate units in the skip connections to highlight salient feature information while eliminating irrelevant and noisy redundant information. Although both methods incorporated a series of attention gates in the decoder and skip connections, they overlooked the varying expressive abilities of features at different network positions. The excessive use of attention mechanisms introduces additional computational complexity and storage overhead. Zhou et al. [26] proposed a multi-modal segmentation network guided by three attention fusion modules to improve segmentation results, taking into consideration the correlation between different modalities in MRI imaging. While there is a strong correlation between different modalities, the approach only places the three attention modules at the highest level of feature extraction without utilizing attention mechanisms to facilitate the interaction between different modalities at other stages. To emphasize the correlation between salient features and channels, Huang et al. [27] designed a coarse-to-fine cross-channel attention module that effectively extracts edge texture information of tumor regions. Cheng et al. [28] proposed a modified cascaded network structure that utilizes a dual attention mechanism to fully exploit the relationship between spatial and channel dimensions for long-range contextual modeling, thereby improving brain tumor segmentation results. However, the attention mechanisms employed in the aforementioned methods only allocate weights in the channel or spatial dimension, without establishing long-range dependencies based on feature correlations from different perspectives in local and global contexts.

To solve the above problems, according to the different attention levels of various network locations to focal regions, inspired by the existing attention mechanisms [29,30], we designed a multi-view attention (MVA) mechanism, which was added to the last two layers of the encoder and decoder, respectively, extracted global and local features from the three perspectives of channel, content, and location. To achieve accurate localization of the target area and enhance the feature extraction ability of the focal region.

2.2. Multi-Scale features

It is well known that utilizing multi-scale feature extraction improves the segmentation performance of small target lesion regions in medical image segmentation. To obtain different receptive fields, various methods have been proposed by researchers to capture multi-scale semantic information.

Xiao et al. [31] designed a multi-view hierarchical segmentation network (MVHS-Net) for brain tumor segmentation based on the improved 3D UNet. They proposed multi-view fusion convolution and multi-view hierarchical segmentation block, which can effectively capture multi-scale and multi-view information and reduce redundant features. Mazher et al. [32] presented a multi-view, multi-scale 3D fetal brain segmentation method that incorporates axial, coronal, and sagittal views of MRI through an effective fusion approach, utilizing different depths and multi-scale information to comprehensively extract fetal brain features. Although the two methods above can extract multi-scale feature information through different receptive fields, 3D networks introduce higher parameter, computational, and memory requirements than 2D convolutions. Additionally, 3D networks heavily rely on the available data, and the limited availability of MRI brain tumor data does not satisfy the demands of 3D networks. Lu et al. [33] proposed a lightweight Ghost Spatial Pyramid (GSP) module based on the Ghost module. The GSP module, serving as a shared encoder in the network, aggregates detailed features from multiple scales of receptive fields at a low computational cost to enhance the capability of multi-scale feature extraction. SURIZA et al. [34] utilized a single encoder to extract features from four modalities and introduced a multi-modal fusion network based on a Bi-directional Feature Pyramid Network (Bi-FPN). Bi-FPN models the correlations among different modalities in a cascaded manner between the four encoders, effectively leveraging the complex relationships among modalities. Both approaches employ feature pyramids to extract information from multiple scales of receptive fields. However, the different scale features in the feature pyramid often contain redundant information, resulting in repeated use of feature information and increasing memory overhead and computational complexity. Rehman et al. [35] proposed BU-Net for the segmentation of various subregions of brain tumors. They introduced residual dilated skip connections into the existing UNet architecture to extract multi-scale features by increasing the effective receptive field. Wang et al. [36] employed a spatial DFP module composed of multiple parallel dilated convolutions to extract multi-scale image features and enhance the neural network's ability to extract and utilize multi-scale features. These two methods utilize multiple dilated convolutions to extract multi-scale features without incorporating attention mechanisms to extract useful feature information from each scale selectively. The use of multiple different-sized receptive fields can generate redundant information, leading to additional computational complexity.

To address the above problems, we propose the multi-scale feature interaction (MSI) module utilizing 2D convolutions. By combining multi-scale feature extraction with attention mechanisms, the MSI

module selectively extracts useful feature information from multiple receptive fields of varying sizes, eliminating redundant information. Moreover, it facilitates interplay among different dimensions, allowing for comprehensive learning of the feature representation capabilities of various information. As a result, our approach achieves precise segmentation of tumor boundaries for different tumor categories.

3. Methodology

In this section, we provide a detailed description of the proposed MVSI-Net, which includes the multi-view attention (MVA) mechanism and the multi-scale feature interaction (MSI) module.

3.1. Multi-View attention (MVA) mechanism

In a clinical setting, doctors first roughly locate the area of brain tumors and then carefully observe the surrounding tissues for accurate labeling. Due to the classification of brain tumors into whole tumor region, enhancing tumor region, and core tumor region, detailed delineation of these three categories is required. Previous works mostly employed convolutional neural networks (CNNs) based on UNet architectures for feature extraction, as conventional convolution operations fail to segment different categories of brain tumors precisely. To address this issue, many methods introduced attention mechanisms, but most only considered channel attention, spatial attention, or a simple combination of both, especially neglecting the importance of local contextual information for small target segmentation while focusing on global contextual information during feature extraction. Inspired by channel attention [29] and global self-attention [30], we propose a multi-view attention mechanism, MVA, to capture global and local information of input feature maps from the perspectives of content, position, and channel. The overall architecture of MVA is illustrated in Fig. 2. This module comprises multi-scale channel attention (MSCA) and content and position attention (CPA) in tandem, where content and position attention are connected in parallel.

The MSCA aggregates local and global contextual information using Global Average Pooling (GAP) and *PWConv* (1 × 1 convolution). After channel attention, the feature maps with global and local information enter the CPA. In this process, the content attention selectively focuses on the pixel information of interest based on all pixels' content. At the same time, position attention focuses on the content of each pixel itself and considers the relative spatial position of the pixel's surroundings to generate attention weights for each pixel. Channel, content, and position attention synergistically collaborate to achieve more precise segmentation results. Specifically, channel attention informs the content and position attention on which channels should be focused, and the content attention informs the channel and position attention of the importance of its content so that they can selectively focus on it. The position attention guides the channel and content attention to locate the positions within the channels based on the relevant content information.

Based on previous literature observations, it has been found that shallow networks only capture low-level edge and texture information. Still, they consume significant computational resources and make minimal contributions to the segmentation results [37]. In contrast, deep networks can learn more abundant semantic information, and the features they extract better represent the entire lesion region. Therefore, we incorporate the proposed MVA into the last two layers of the encoder. Although deeper layers extract features closer to the lesion region, some details may be overlooked as the image size decreases. While the skip connections in UNet partially address this issue, combining shallow and deep networks introduces unnecessary non-lesion features, leading to additional computational complexity. To mitigate this, we adopt MVA in the last two layers of the decoder, which can combine deep high-level semantic information with shallow low-level texture details and filter out redundant information.

Let the input image $x \in R^{C\times H\times W}$, after MVA the output is $Y \in R^{C\times H\times W}$, calculated as follows:

$$Y = Y_1 \otimes Y_2. \tag{1}$$

Among them, Y_1 is the output result after MSCA, and Y_2 is the output result after CPA. $\otimes$ denotes element-wise multiplication, $Y_1 \in R^{C\times H\times W}$, $Y_2 \in R^{C\times H\times W}$.

The calculation formula of Y_1 is as follows:

$$Y_1 = x \otimes \sigma[L(x) \oplus G(x)], \tag{2}$$

where, $\oplus$ denotes addition, σ is the *sigmoid* activation function. $L(x)$ is

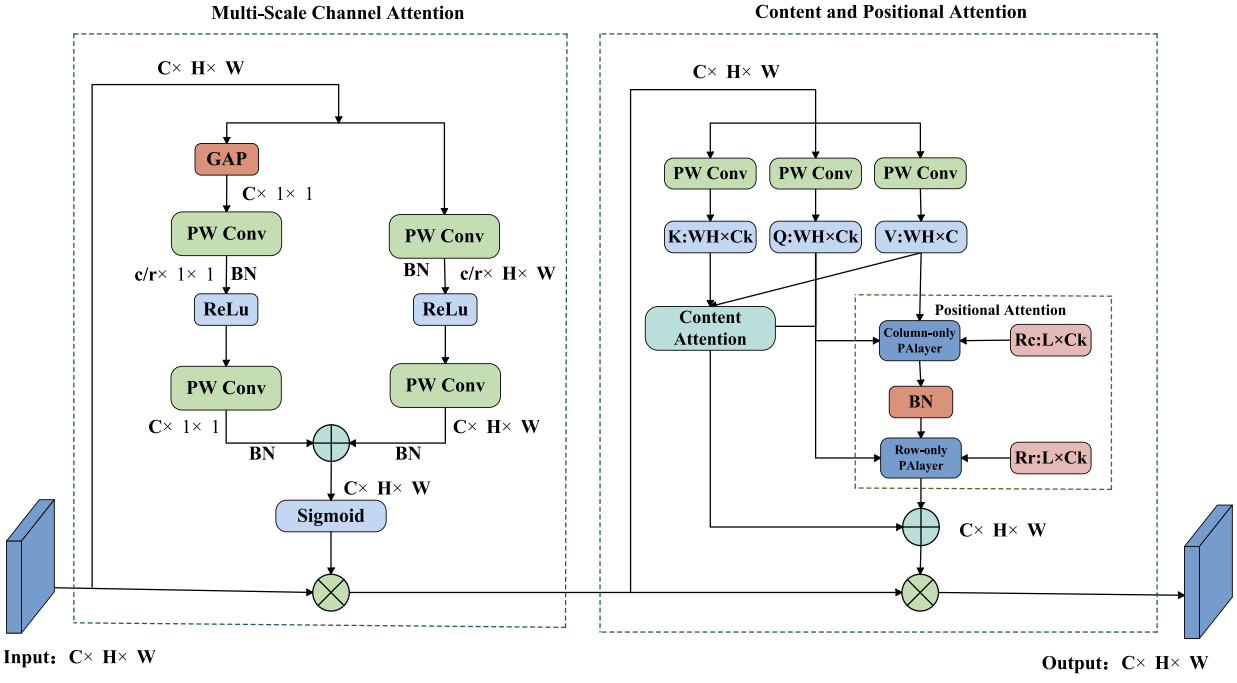

Fig. 2. Multi-view attention (MVA) mechanism.

the output result after the local information feature extraction of the input image, and $G(x)$ is the output result after the global information feature extraction of the input image.

$L(x)$ and $G(x)$ are calculated as follows:

$$L(x) = \beta\{PWConv_2[\delta(\beta(PWConv_1(x)))]\}, \tag{3}$$

$$G(x) = \beta\{\omega_2[\delta(\beta(\omega_1(g(x))))]\}. \tag{4}$$

The dimension of $PWConv_1$ is $\frac{c}{r} \times c \times 1 \times 1$ and the dimension of $PWConv_2$ is $c \times \frac{c}{r} \times 1 \times 1$. $g(x)$ is the output result of input image x after GAP, δ is the relu activation function, ω_1 and ω_2 respectively represent the ascending and descending dimensions through *PWConv*.

The calculation formula of Y_2 is as follows:

$$Y_2 = CA \oplus PA = Q(\rho(K^T)V) \oplus (q_{ab}R^{rT})\beta[(q_{ab}R^{cT})V^c_{ab}], \tag{5}$$

where *CA* is content attention and *PA* is location attention. *Q*, *K*, and *V* are the key, query, and value matrices generated after three *PW Conv* (1 × 1 convolution), respectively, K^T is the transposed matrix of *K*, and ρ represents the softmax normalization operation applied to each pixel, respectively. $R^c \in R^{L \times C_k}$ is the relative position embedding along the columns, which is used as the key to focus on each input pixel of the column to get the output result, R^{cT} is the transpose matrix of R^c. Similarly, $R^r \in R^{L \times C_k}$ is the relative position embedding along the rows. q_{ab} denotes the query at pixel (a, b) and $V^c_{ab} \in R^{L \times C}$ is the matrix consisting of the values of the L-column neighborhood at pixel (a,b). β is the Batch Normalization.

The calculation formula of Q, K, and V is as follows:

$$K = Q = PWConv_k(Y_1), V = PWConv(Y_1), \tag{6}$$

where, $K = Q \in R^{C_k \times H \times W}$, $V \in R^{C \times H \times W}$. The pseudocode for MVA is shown in Table 1.

3.2. Multi-Scale feature interaction (MSI) module

With the increasing depth of convolutional layers, the receptive field in the feature extraction process becomes larger. However, the size of feature maps becomes smaller. In the deep layers of feature extraction, known as the bottleneck, a significant amount of rich target information exists in the feature maps. At the same time, the traditional dual convolutional methods fail to capture feature information from multiple perspectives effectively. Therefore, we propose a multi-scale feature interaction (MSI) module to obtain the required scale features and better segment the boundary regions of different tumor categories. The architecture of MSI, as shown in Fig. 3, comprises a multi-scale feature extraction (MSFE) module and a cross-dimensional feature interaction (CDFI) module. Firstly, MSI utilizes the MSFE to extract feature information at different receptive fields and focuses on the lesion area in feature maps of diverse sizes. Then, the CDFI is applied dimension transformation on the features extracted based on different receptive field sizes, interacting with information from various types of target regions and achieving more precise segmentation of the boundary contours of different categories of brain tumors. More details about the MSI module are presented below.

Table 1
Pseudocode for the MVA.

Algorithm 1: The Procedure of MVA

Input: The size of image X is $C \times H \times W$
Output: The Y size of the feature map after MVA is $C \times H \times W$
i : attention through different perspectives;
for i in range $(0, 2)$ do
if $i = 0$ do
Using branch 1 to extract global features $G(X) = F_{PW_2}\left(X^{GAP}_{PW_1}\right)$
Using branch 2 to extract local features $L(X) = F_{PW_2}(X_{PW_1})$
Generating Attention Weights $F(X) = Sigmoid\{Add[L(X), G(X)]\}$
Obtaining results for multiscale channel attention $Y_1 = Multiply[X, F(X)]$
end if
if $i = 1$ do
Based on Q, K, V generates content attention $CA = Q[Softmax(K^T)V]$
Based on Q, K, V generates position attention $PA = (q_{ab}R^{rT})\beta[(q_{ab}R^{cT})V^c_{ab}]$
Generating Attention Weights $Y_2 = CA + PA$
end if
end for
Through MVA the output $Y = Multiply(Y_1, Y_2)$

3.2.1. Multi-Scale feature extraction (MSFE) module

The MSFE captures useful feature information from multiple receptive fields. It comprises a multi-branch (MB) module and two Large Kernel Attention (*LKA*) modules [38]. The MB generates feature information with different receptive fields by using convolutional kernels and dilated convolutions of varying sizes to form a multi-branch structure capturing multi-scale features. Based on the multi-scale features captured by the MB, the two *LKA* modules focus on the tumor lesion regions of interest in different receptive fields.

In the segmentation of MRI brain tumor images, we adopted three-classification segmentation, where different lesion regions have different sizes. A comprehensive representation of multi-scale features is conducive to the model's perception of data features [39]. Therefore, the MSFE utilizes the multi-branch (MB) module to extract features of multiple sizes. As shown in Fig. 3, the MB comprises three parallel branches, each composed of standard and dilated convolution. The standard convolution with different sizes of convolutional kernels generates different receptive fields. We substitute $1 \times n$ and $n \times 1$ convolution for the original $n \times n$ convolution to reduce parameter complexity. Following the standard convolution, we employ dilated convolution with different dilation rates. The dilated convolution captures feature information of larger receptive fields without adding additional parameters and generates higher-resolution feature maps.

Since the usable information in feature maps of various sizes varies, so do their contributions to segmentation accuracy. To do this, we embed two Large Kernel Attention (*LKA*) modules at the end of the multi-branch (MB) module, as shown in Fig. 4. The figure observes that the *LKA* consists of Depthwise Convolution (*DWConv*), Depthwise Dilated Convolution (*DW-D-Conv*), and 1 × 1 Convolution (*PWConv*). It fully utilizes the advantages of large kernel convolution and self-attention, establishing long-range dependencies on feature maps of different sizes. According to input features, it selectively judges useful information and dynamically selects the regions of interest in the lesion area, generating attention weights.

For the input image $x \in R^{C \times H \times W}$, the output results after passing through the multi-branch (MB) module are denoted as x_i (where $i = 1, 2, 3, 4$). MSFE utilizes the first *LKA* to model remotely from the cascaded results y_1 of two different receptive fields, namely feature maps x_1 and x_2. The output result is denoted as LKA^1_{output}, and its calculation formula is as follows:

$$LKA^1_{output} = Att_{y_1} \otimes y_1, \tag{7}$$

$$Att_{y_1} = PWConv\{DWDConv[DWConv(y_1)]\}. \tag{8}$$

The two multi-branch output results x_1 and x_2 are reduced by *PWConv* after cascading, and the output result is $y_1 \in R^{C \times H \times W}$. $LKA^1_{output} \in R^{C \times H \times W}$, and Att_{y_1} represents the attention weight generated by *LKA*.

Next, the second *LKA* is applied to the result y_2 after the cascade of LKA^1_{output} and multi-branch x_3, further establishing the remote dependency. Similar to the above operation, the output result after the second *LKA* is LKA^2_{output}, which is calculated as follows:

$$LKA^2_{output} = Att_{y_2} \otimes y_2, \tag{9}$$

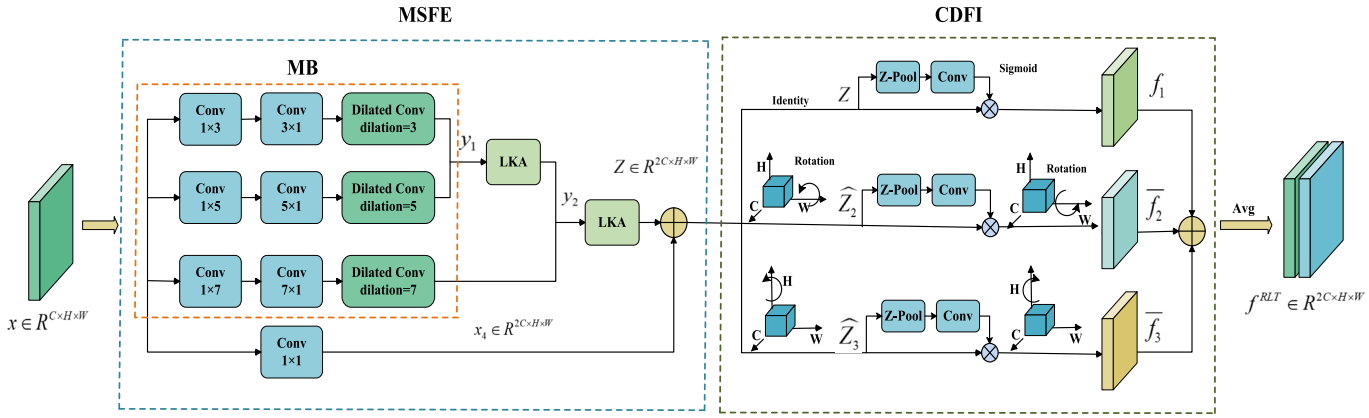

Fig. 3. Multi-scale feature interaction (MSI) module.

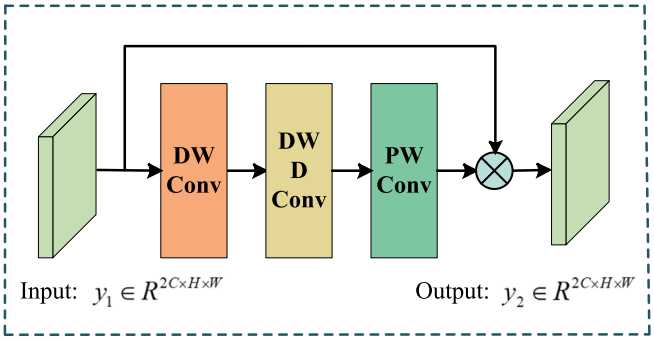

Fig. 4. Large kernel attention (*LKA*).

$$Att_{y2} = PWConv\{DWDConv[DWConv(y_2)]\}, \tag{10}$$

$$y_2 \in R^{2C\times H\times W},\ LKA^2_{output} \in R^{2C\times H\times W}$$

Finally, to avoid neglecting useful information in the input image, the final output $Z \in R^{2C\times H\times W}$ is calculated by mapping originally from residual connections. The calculation formula is as follows:

$$Z = LKA^2_{output} \oplus PWConv(x). \tag{11}$$

3.2.2. Cross-Dimensional feature interaction (CDFI) module

MSI employs a cross-dimensional feature interaction (CDFI) module to address the issue of unclear edge contours in the brain tumor segmentation of different categories. Accurate feature extraction is crucial for improving the accuracy of small-scale medical image segmentation, and using attention mechanisms can focus on regions of interest while preserving global contextual information [40]. Previous attention mechanisms have primarily focused on dimensionality reduction, which is redundant for capturing local dependencies between channels. However, CDFI highlights the importance of dimensionality transformation for feature extraction. Therefore, the feature maps generated from the MSFE conduct cross-dimensional information interactions by using the CDFI to extract feature information of target regions of different categories and achieve segmentation of edge contours for different types of brain tumors.

CDFI captures the cross-dimensionality between the spatial and channel dimensions of the input features to construct global contextual features. Specifically, the output feature maps from the MSFE enter three branches of CDFI: (1) capturing the cross-dimensionality between the channel dimension *C* and the spatial dimension *W*; (2) capturing the cross-dimensionality between the channel dimension *C* and the spatial dimension *H*; (3) a branch similar to CBAM to build spatial attention. Finally, the outputs of the three branches are averaged for aggregation. CDFI uses three branches to capture the interdependencies between the input tensors (*C, H*), (*C, W*), and (*H, W*) by employing cross-dimensionality interaction, extracting spatial and channel information from different dimensions.

The final output of the MSI module after MSFE and CDFI is $f^{MSI} \in R^{2C\times H\times W}$, and the calculation formula is as follows:

$$f^{MSI} = \frac{1}{3}\{z\sigma[\psi_1(\hat{z})] \oplus \overline{\hat{z}_2\sigma[\psi_2(\hat{z}_2^*)]} \oplus \overline{\hat{z}_3\sigma[\psi_3(\hat{z}_3^*)]}\}. \tag{12}$$

Among them, $\hat{z}_2$ and $\hat{z}_3$ are respectively the output results of the input image *z* after it is rotated 90° counterclockwise along the *H* and *W* axes after passing through the CDFI module; $\hat{z}$ and $\hat{z}_2^*$ and $\hat{z}_3^*$ are respectively the output results of *z*, $\hat{z}_2$ and $\hat{z}_3$ after $z-pool$. $\psi_1(\hat{z})$ represents the output result after standard convolution with convolution kernel size k and Batch Normalization, $\sigma(\psi_1(\hat{z}))$ is the generated attention weight, and the attention weight is multiplied point by point with the input image *Z* to get $z\sigma(\psi_1(\hat{z}))$. Since the previous rotation was 90° counterclockwise, to recover the same feature map as the input dimension in the output result, the input image needs to be rotated 90° clockwise to recover the input dimension after obtaining the result of multiplying the input image with the attention weight point by point.

To simplify the above formula, the following formula is obtained:

$$f^{MSI} = \frac{1}{3}(z\omega_1 \oplus \overline{\hat{z}_2\omega_2} \oplus \overline{\hat{z}_3\omega_3}) = \frac{1}{3}(f_1 \oplus \overline{f_2} \oplus \overline{f_3}), \tag{13}$$

where ω_1, ω_2, and ω_3 are the three attention weights generated after passing through the three branches in the CDFI, $\overline{f_2}$ and $\overline{f_3}$ represent a clockwise rotation of 90° to remain the same size as the original input shape ($C \times H \times W$). The pseudocode for MSI is shown in Table 2.

3.3. Overview of the MVSI-Net

Due to the limited representation of lesion areas in previous studies using single-modal images, the singular imaging modality in medical imaging has proven insufficient. Multi-modal images provide rich and diverse information, enabling a comprehensive and integrated depiction of lesion areas. Furthermore, in multi-modal MRI brain tumor segmentation, commonly employed fusion methods include input-level fusion and decision-level fusion. Specifically, in decision-level fusion, each modality image undergoes feature extraction using a separate encoder, resulting in independent outputs that fail to establish internal relationships between different modalities of the same patient, and it incurs significant computational costs. Moreover, previous MRI brain tumor segmentation approaches have neglected the varying semantic information extracted at different positions within the network. As the network deepens, the feature maps become increasingly compact while

containing richer semantic information. Many methods treat all encoders involved in the feature extraction process equally without placing sufficient emphasis on the expressive power of deep network features. Similarly, in the process of feature restoration, previous methods have largely treated different positions within the decoder equally without placing sufficient emphasis on the shallow-level edge texture information and the redundant information brought by skip connections. Accordingly, we propose a novel network, MVSI-Net, to address the aforementioned issues in MRI brain tumor segmentation. The overall architecture of MVSI-Net is illustrated in Fig. 5.

Firstly, we employ input-level fusion, which concatenates the four modalities in each MRI brain tumor sample into a single image with four channels and then input it into the network. Secondly, MVSI-Net is an improvement on the UNet architecture, additionally due to significant variations in the location, shape, and size of brain tumors among different patients and substantial changes in tumor lesions at different stages. Therefore, given the different positions of the network have differential impacts on feature extraction and restoration, we propose the multi-view attention (MVA) mechanism and the multi-scale feature interaction (MSI) module, which are respectively placed at specific positions in the baseline, considering multi-view and multi-scale information to extract discriminative features for different tumor categories and achieve superior brain tumor segmentation performance. Similar to the existing UNet model, MVSI-Net consists of an encoder, a bottleneck, and a decoder. The encoder and decoder comprise four blocks, each containing two 3×3 convolutions. However, unlike UNet, we add the MVA separately to the last two blocks of the encoder and decoder. Besides, the feature extraction at the bottleneck consists of a multi-scale feature interaction (MSI) module.

4. Experiments and results

In this section, we validate the effectiveness of the proposed method and compare it with other methods on the BraTs dataset.

4.1. Experimental setup and dataset

The code for this experiment was implemented using Python 3.8.13 environment and PyTorch 1.10.0 framework. The experiments were conducted on a server with an NVIDIA RTX 8000 GPU and 64 GB RAM.

Table 2
Pseudocode for the MSI module.

Algorithm 2: The Procedure of MSI

Input: The size of image X is $C \times H \times W$
Output: The f^{MSI} size of the feature map after MSI is $2C \times H \times W$
Split X into 3 parts the channel-wise;
for i in range $(0, 3)$ do
Calculate features in different scale receptive field
$X_i \in R^{H \times W \times C} = F^{MB}(X_i, d = 2i + 1)$
end
for X_i do
Concatenate X_0 and X_1 to get y_1 into LKA
$LKA^1_{output} \in R^{C \times H \times W} = Multiply(Att_{y_1}, y_1)$
Concatenate LKA^1_{output} and X_2 to get y_2 into LKA
$LKA^2_{output} \in R^{2C \times H \times W} = Multiply(Att_{y_2}, y_2)$
Connect the original input X and LKA^2_{output} via residuals to obtain the output of MSFE
$Z^{MSFE} \in R^{2C \times H \times W} = Add\left(LKA^2_{output}, X_{1\times 1}\right)$
Split Z^{MSFE} into 3 parts the channel-wise;
for j in range $(0, 3)$ do
Perform feature interactions of different dimensions
$f_j \in R^{2C \times H \times W} = F^{CDFI}\left(Z_j^{MSFE}, ((C,H),(C,W),(H,W))\right)$
end
Through CDFI obtain the output of MSI
$f^{MSI} \in R^{2C \times H \times W} = 1/3Add(f_1, \overline{f_2}, \overline{f_3})$
end

We used standard segmentation loss functions such as Binary Cross Entropy Loss and Adam optimizer to optimize all models. The learning rate was set to 1e-4, the batch size was set to 16, and the number of epochs was set to 200.

To evaluate the effectiveness of MVSI-Net, we trained and tested the model using three publicly available datasets: BraTs 2019, BraTs 2020, and BraTs 2021. The Multi-modal Brain Tumor Segmentation (BraTs) Challenge datasets comprise a training set and a test set, where the test set does not provide ground truth labels. Therefore, we only used the training sets for our experiments. The training set consists of high-grade gliomas (HGG) and low-grade gliomas (LGG). The BraTs 2019 training set contains 335 3D MRI images, including 259 HGG and 76 LGG samples. The BraTs 2020 training set contains 369 3D MRI images, including 293 HGG and 76 LGG samples. The BraTs 2021 training set contains 1251 3D MRI images. We divided the training sets of the three datasets into training and validation sets with a 7:3 ratio. Each sample in the dataset includes four modalities: T1-weighted (T1W), T1-weighted with contrast-enhancement (T1CE), T2-weighted (T2W), and fluid-attenuated inversion recovery (FLAIR). The size of each sample is 240 $\times$ 240 $\times$ 155. Fig. 6 illustrates the four MRI modalities and corresponding expert segmentation results for different patients' brains.

The BraTs datasets consist of four tissue labels: normal tissue (label 0), non-enhancing tumor core and necrotic core on T2 MRI (label 1), peritumoral edema (label 2), and enhancing tumor on T1CE MRI (label 4). In experiments, we performed a three-classification segmentation. We combined labels 1, 2, and 4 to represent the whole tumor (WT), combined labels 1 and 4 to represent the tumor core (TC), and used label 4 to represent the enhancing tumor (ET). As shown in Fig. 6, the green region represents WT, the yellow region represents TC, and the red region represents ET.

4.2. Data preprocessing

Due to the large memory consumption, longer training cycles of 3D data, and the high requirements for the devices used, we preprocess the data into 2D images for training purposes. Each sample in the BraTs dataset contains four Magnetic Resonance Imaging (MRI) modalities. Since each modality has different physical imaging methods and contrast, we apply the Z-Score method to standardize each modality's image, i.e., normalize the mean and standard deviation to 0 and 1, respectively.

After standardizing each sample, we also cropped the background area that occupied a large proportion of the images. The original size of each modality image in BraTs is 240 $\times$ 240 $\times$ 155, and we crop to 160 $\times$ 160 $\times$ 155. Then, we slice the 3D MRI images into 2D MRI images along the axial position. Each sliced sample has a size of 160 $\times$ 160 $\times$ 1, resulting in 155 slices. To alleviate the class imbalance problem, we discard the slices that do not include lesion areas. Furthermore, since BraTs consist of multi-modal images, we concatenate the slices of the four different modalities to form a 4-channel image. The preprocessed image size is 160 $\times$ 160 $\times$ 4, and the images are finally saved in the npy format.

4.3. Evaluation metrics

To validate the effectiveness of the proposed method in this study, we utilized standard evaluation metrics in brain tumor segmentation, including the Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), Precision (PPV), and Sensitivity (SE). These metrics are computed as follows:

$$DSC = \frac{2TP}{2TP + FP + FN}, \tag{14}$$

$$HD(X, Y) = \max\left\{ \underset{x \in X}{\max}\underset{y \in Y}{\min}||x - y||, \underset{y \in Y}{\max}\underset{x \in X}{\min}||y - x|| \right\}, \tag{15}$$

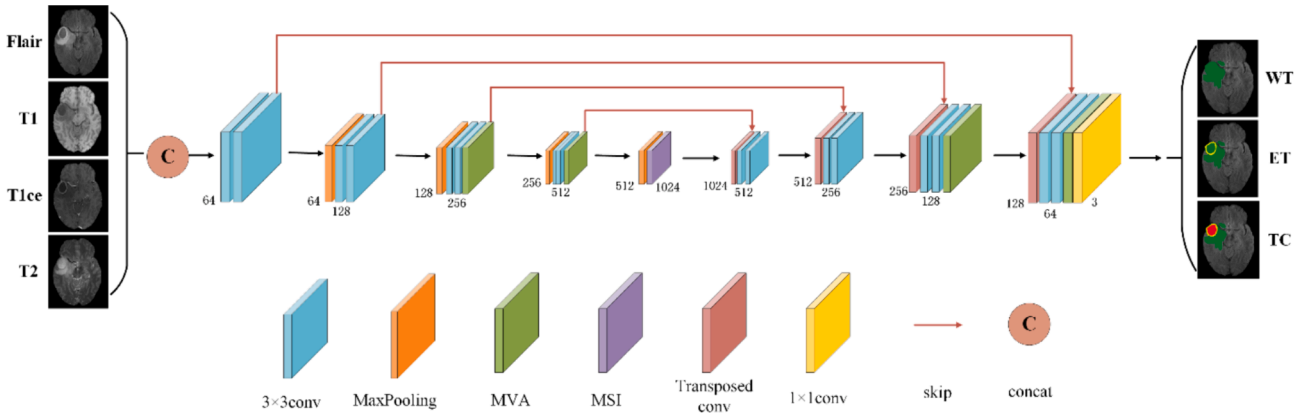

Fig. 5. MVSI-Net architecture.

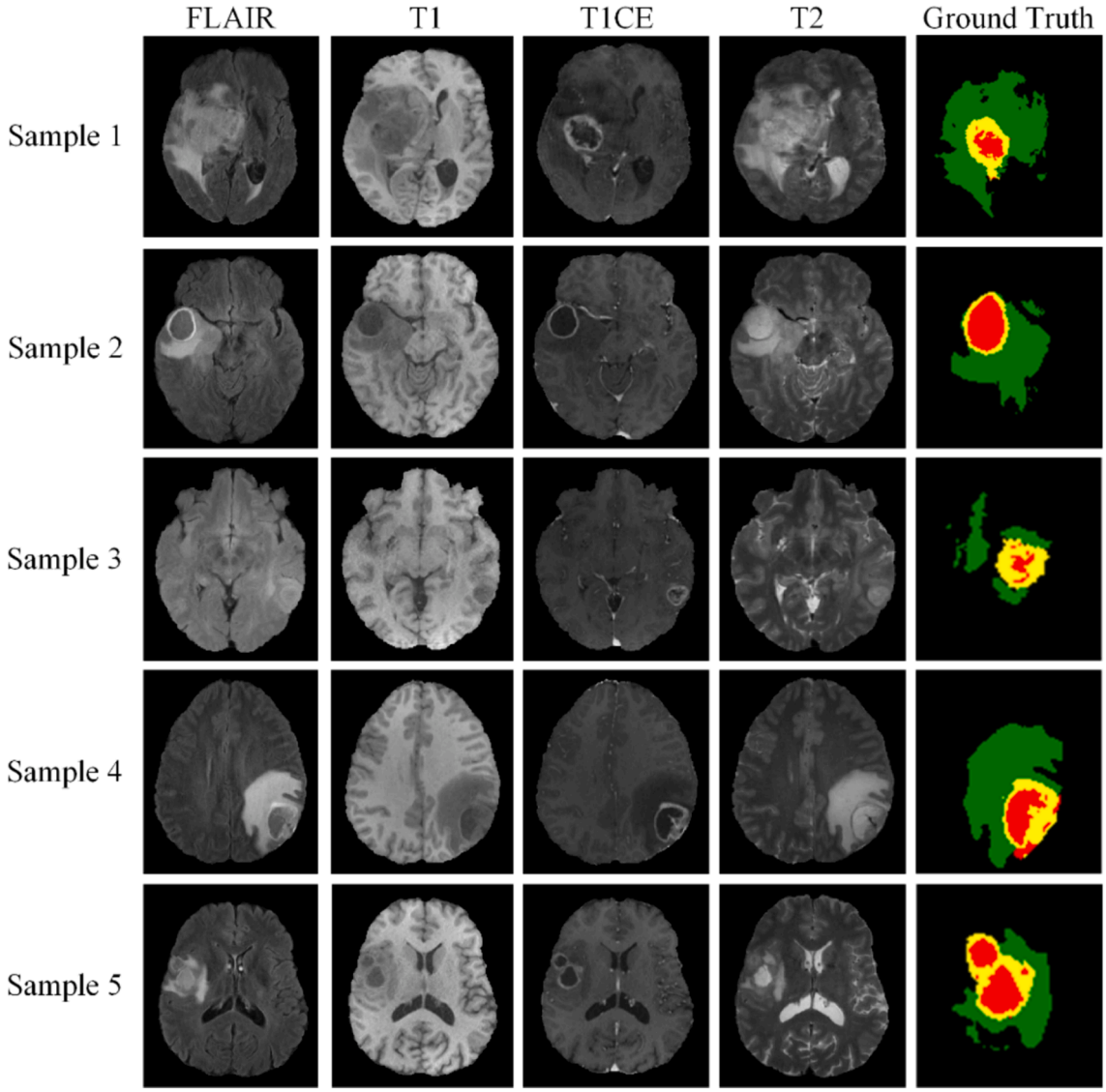

Fig. 6. Taking a sample from the BraTs 2020 training dataset as an example, display axial plots of the four modalities and labels.

$$PPV = \frac{TP}{TP + FP}, \tag{16}$$

$$SE = \frac{TP}{TP + FN}. \tag{17}$$

DSC computes the similarity between the segmentation result and the ground truth labels, ranging from 0 to 1. A higher value indicates a better segmentation result for brain tumors. HD is sensitive to the boundaries of the segmentation, making it suitable for measuring the distance between the segmentation result and the ground truth labels. A smaller distance suggests a closer proximity between the segmentation result and the ground truth. PPV describes the proportion of all correctly segmented tumor pixels in the segmentation results to the real tumor pixels. Sensitivity measures the proportion of detected tumor pixels in the segmentation result among all the true tumor pixels.

In addition, TP represents true positives, meaning the test value and the ground truth label are positive samples. TN represents true negatives, indicating the test value and the ground truth labels are negative samples. FP represents false positives, where the test value is classified as a positive sample, and the ground truth label is a negative sample. FN represents false negatives, where the test value is classified as a negative sample, and the ground truth label is a positive sample. In this context, x denotes the ground truth labels of the samples in dataset X, and y represents the prediction results in the test set Y.

4.4. Ablation studies

In this section, we conducted ablation studies on the BraTS 2019, BraTS 2020, and BraTS 2021 datasets to validate the effectiveness of the proposed brain tumor segmentation method. To evaluate the segmentation performance of the proposed modules, we individually added MVA and MSI to the UNet and conducted experiments with the following specific configurations:

(1) Baseline model UNet: U-shaped structure based on an encoder and a decoder.
(2) UNet + MVA: MVA added to the last two layers of the encoder and decoder in the baseline UNet.
(3) UNet + MSI: MSI was added at the bottleneck of the baseline UNet.
(4) UNet + MVA + MSI: MVA was added to the last two layers of the encoder and decoder, and MSI was added at the bottleneck of the baseline UNet.

The ablation results on the Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) are shown in Table 3. The results of ablation experiments in Precision and Sensitivity are shown in Table 4. From the observation of Table 3 and Table 4, it can be seen that adding MVA and MSI to the baseline UNet model improves the segmentation performance of brain tumors. Moreover, the proposed model achieves better segmentation performance on all three datasets. The experiments confirm that MVA and MSI enhance the segmentation performance, and as the network depth increases, the segmentation accuracy also improves.

To observe the changes in the Dice Similarity Coefficient (DSC) after applying the MVA and MSI modules to all three datasets more clearly, bar charts were plotted as shown in Fig. 7, Fig. 8, and Fig. 9.

After comparing the Dice Similarity Coefficients, it was found that on the BraTS 2019 dataset, the DSC values for the WT, TC, and ET categories after adding the MVA module to the baseline UNet were 0.860, 0.867 and 0.791, respectively. Similarly, after adding the MSI module to the baseline UNet, the DSC values for the WT, TC, and ET categories were 0.856, 0.855, and 0.793, respectively. On the BraTS 2020 dataset, the DSC values for the WT, TC, and ET categories after adding the MVA module to the baseline UNet were 0.869, 0.863, and 0.798, respectively. Likewise, after adding the MSI module to the baseline UNet, the DSC values for the WT, TC, and ET categories were 0.825, 0.890, and 0.775, respectively. On the BraTS 2021 dataset, the DSC values for the WT, TC, and ET categories after adding the MVA module to the baseline UNet were 0.837, 0.849, and 0.763, respectively. Similarly, after adding the MSI module to the baseline UNet, the DSC values for the WT, TC, and ET categories were 0.827, 0.840, and 0.758, respectively.

4.5. Comparison with other methods

(1) Experimental results on the BraTS 2019 dataset.

The experiment was conducted on the BraTS 2019 dataset. To evaluate the effectiveness of MVSI-Net in brain tumor segmentation, we compared the proposed method with state-of-the-art methods on various metrics such as DSC, HD, Precision, and Sensitivity. These methods include UNet [18], UNet2+ [41], UNet3+ [42], DeepResUnet [43], ResUnet-a [44], ResUnet++ [45], DoubleUnet [46], MultiResUnet [21], AttentionUnet [47], and DenseUnet [48]. The experimental results are presented in Table 5 and Table 6.

By observing Table 5 and Table 6, we found that the proposed method achieved DSC scores of 0.864, 0.884, and 0.801 for WT, TC, and ET, respectively. These scores improved by 1.5 %, 2.4 %, and 1.8 % than the other top scores. The HD scores for WT, TC, and ET were 2.560, 1.537, and 2.713, respectively, which decreased by 0.034, 0.094, and 0.041, respectively, compared to the other top scores. Therefore, MVSI-Net outperformed the state-of-the-art methods in DSC and HD metrics. For the Precision metric, the segmentation results for WT, TC, and ET regions reached 0.880, 0.912, and 0.813, respectively. These scores improved by 0.7 %, 1.0 %, and 0.9 % than the other top scores. Regarding the Sensitivity metric, the segmentation results for WT, TC, and ET regions achieved 0.895, 0.920, and 0.850, respectively. The Sensitivity scores increased by 1.2 % and 0.7 % than the other top scores for the WT and ET regions, respectively, at the same time remaining on par with the other top results for the TC region.

Fig. 10 shows the segmentation results of MVSI-Net with other methods. By comparing with other models, it can be observed that the proposed method performs well in segmenting the target regions and preserves the edge details of different categories effectively. When the target regions are small or have irregular shapes, other models may produce inaccurate segmentation results due to some interfering information. At the same time, our proposed method can preserve the detailed information of the regions of interest effectively.

(2) Experimental results on the BraTS 2020 dataset.

The experiments were conducted on the BraTS 2020 dataset. Similar

Table 3
Detailed ablation studies on Dice Similarity Coefficient and Hausdorff Distance of the proposed method.

Dataset	Method	DSC			HD		
		WT	TC	ET	WT	TC	ET
BraTs19	UNet	0.847	0.855	0.780	2.594	1.631	2.780
	UNet + MVA	0.860	0.867	0.791	2.606	1.593	2.749
	UNet + MSI	0.856	0.855	0.793	2.626	1.607	2.803
	UNet + MVA + MSI	**0.864**	**0.884**	**0.801**	**2.560**	**1.537**	**2.713**
BraTs20	UNet	0.804	0.834	0.723	2.772	1.608	2.931
	UNet + MVA	0.869	0.863	0.798	2.654	1.526	2.720
	UNet + MSI	0.853	0.846	0.776	2.691	1.533	2.843
	UNet + MVA + MSI	**0.876**	**0.875**	**0.817**	**2.501**	**1.426**	**2.602**
BraTs21	UNet	0.819	0.837	0.750	2.751	1.618	2.937
	UNet + MVA	0.837	0.849	0.763	2.749	1.581	2.913
	UNet + MSI	0.827	0.840	0.758	2.728	1.613	2.905
	UNet + MVA + MSI	**0.853**	**0.861**	**0.781**	**2.662**	**1.523**	**2.824**

Note: The bold part is the optimal value.

Table 4
Detailed ablation studies on Precision and Sensitivity of the proposed method.

Dataset	Method	Precision			Sensitivity		
		WT	TC	ET	WT	TC	ET
BraTs19	UNet	0.870	0.889	0.790	0.871	0.907	0.830
	UNet + MVA	0.859	0.899	0.795	0.874	**0.913**	0.827
	UNet + MSI	**0.883**	0.880	0.785	0.879	0.906	0.841
	UNet + MVA + MSI	0.875	**0.912**	**0.796**	**0.885**	0.911	**0.850**
BraTs20	UNet	0.842	0.897	0.761	0.811	0.879	0.775
	UNet + MVA	0.859	0.901	0.768	0.840	0.887	0.808
	UNet + MSI	0.825	0.890	0.775	0.834	0.877	0.788
	UNet + MVA + MSI	**0.878**	**0.934**	**0.814**	**0.866**	**0.910**	**0.825**
BraTs21	UNet	**0.880**	0.907	0.803	0.828	0.871	0.782
	UNet + MVA	0.852	0.909	0.784	0.843	0.883	0.796
	UNet + MSI	0.876	0.908	0.801	0.836	0.873	0.789
	UNet + MVA + MSI	0.870	**0.924**	**0.803**	**0.868**	**0.889**	**0.825**

Note: The bold part is the optimal value.

to the evaluation metrics and comparison methods used on the BraTS 2019 dataset, we evaluated different types of brain tumors (WT, TC, ET) using four evaluation metrics (DSC, HD, Precision, and Sensitivity). Table 7 and Table 8 present the segmentation performance of our proposed method and other state-of-the-art methods, respectively.

Observing Table 7 and Table 8, we can see that our method outperforms other state-of-the-art methods in DSC and HD metrics. Specifically, the DSC values have improved by 3.1 %, 1.0 %, and 3.4 %, respectively. At the same time, the HD values have decreased by 0.049, 0.033, and 0.059, respectively. Similarly, the segmentation results of all categories are superior to other state-of-the-art methods in terms of Precision and Sensitivity metrics. Specifically, Precision has increased by 0.1 % and 1.0 % in the WT and TC regions, respectively, and remains comparable to other optimal results in the ET region. Sensitivity has increased by 0.9 % and 2.6 % in the TC and ET regions, respectively, and remains comparable to other optimal results in the WT region. This further demonstrates the stability and generalization of our model.

(3) Experimental results on the BraTS 2021 dataset.

The experiments were conducted on the BraTS 2021 dataset, using the same evaluation metrics as the previous two datasets (BraTS 2019 and BraTS 2020). Table 9 and Table 10 present the segmentation performance of our proposed method and other state-of-the-art methods, respectively.

Observing Table 9 and Table 10, we can see that our method

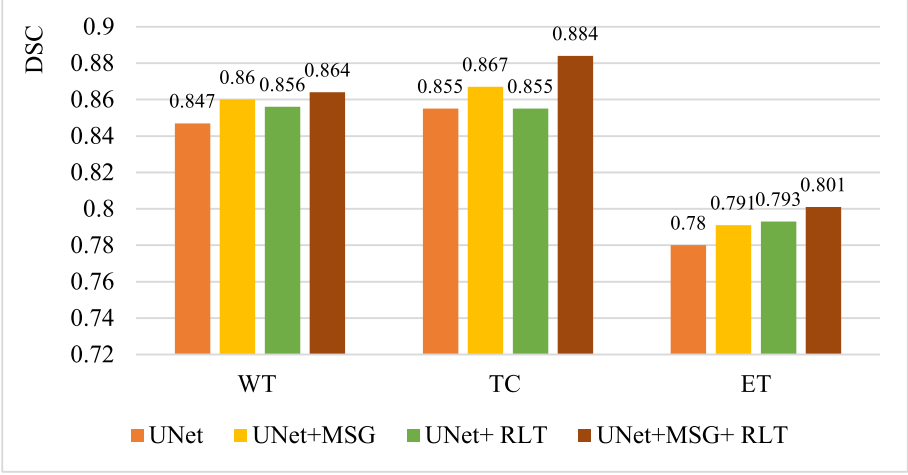

Fig. 7. The values of ablation analysis on Dice Similarity Coefficient (DSC) in BraTS 2019.

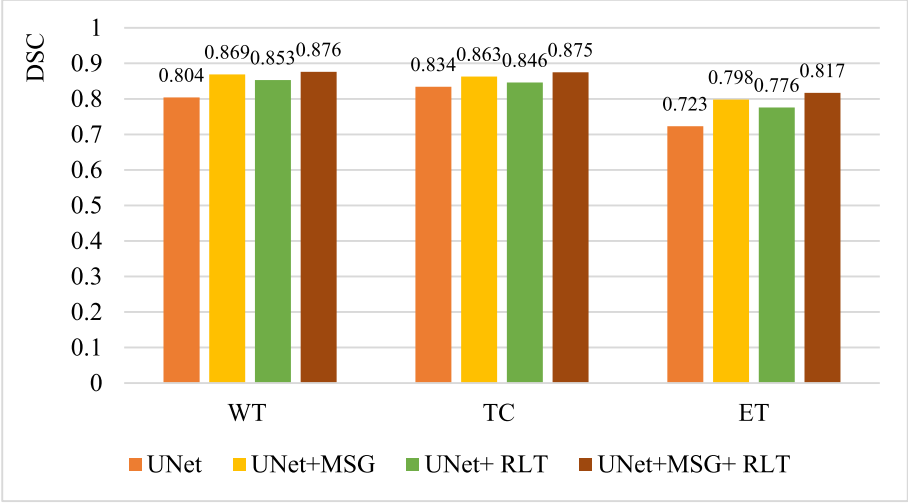

Fig. 8. The values of ablation analysis on Dice Similarity Coefficient (DSC) in BraTS 2020.

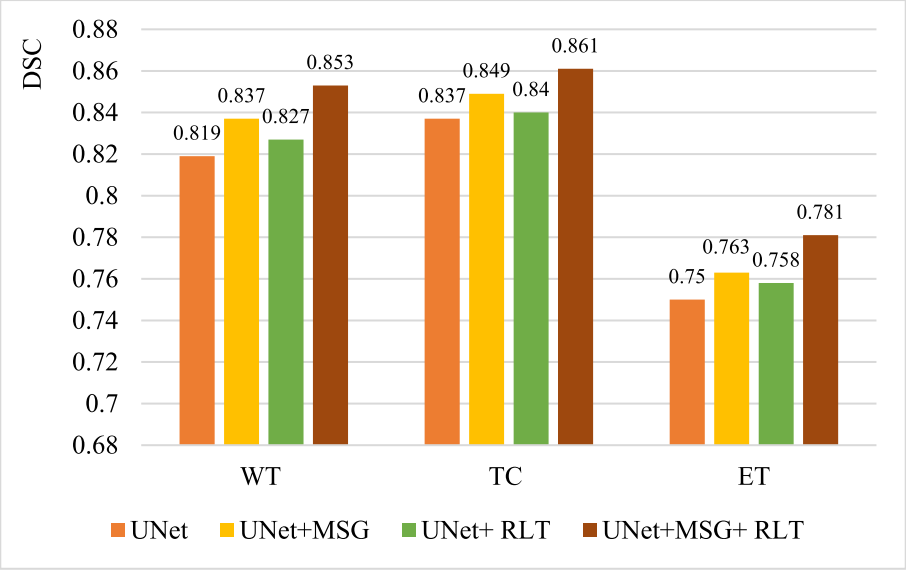

Fig. 9. The values of ablation analysis on Dice Similarity Coefficient (DSC) in BraTS 2021.

Table 5
Performance comparison of different methods on DSC and HD.

Network	DSC			HD		
	WT	TC	ET	WT	TC	ET
UNet [18]	0.847	0.855	0.780	2.594	1.631	2.780
UNet2+ [41]	0.836	0.849	0.764	2.635	1.660	2.833
UNet3+ [42]	0.821	0.815	0.757	2.734	1.651	2.854
DeepResUnet [43]	0.844	0.860	0.780	2.604	1.633	2.883
ResUnet-a [44]	0.819	0.807	0.752	2.794	1.631	2.858
ResUnet++ [45]	0.834	0.840	0.761	2.659	1.694	2.808
DoubleUnet [46]	0.842	0.853	0.779	2.611	1.639	2.769
MultiResUnet [21]	0.849	0.842	0.774	2.617	1.709	2.791
AttentionUnet [47]	0.848	0.851	0.783	2.605	1.653	2.775
DenseUnet [48]	0.847	0.854	0.781	2.601	1.648	2.754
Ours MVSI-Net	**0.864**	**0.884**	**0.801**	**2.560**	**1.537**	**2.713**

Note: The bold part is the optimal value.

Table 6
Performance comparison of different methods on Precision and Sensitivity.

Network	Precision			Sensitivity		
	WT	TC	ET	WT	TC	ET
UNet [18]	0.870	0.889	0.795	0.871	0.907	0.830
UNet2+ [41]	0.867	0.890	0.794	0.864	0.901	0.820
UNet3+ [42]	0.859	0.874	0.781	0.793	0.877	0.742
DeepResUnet [43]	0.865	0.902	0.791	0.883	0.905	0.843
ResUnet-a [44]	0.851	0.891	0.774	0.818	0.886	0.761
ResUnet++ [45]	0.863	0.873	0.789	0.874	0.906	0.828
DoubleUnet [46]	0.860	0.887	0.793	0.878	0.917	0.830
MultiResUnet [21]	0.867	0.860	0.799	0.870	0.920	0.810
AttentionUnet [47]	0.840	0.869	0.770	0.880	0.912	0.834
DenseUnet [48]	0.873	0.884	0.804	0.870	0.915	0.824
Ours MVSI-Net	**0.880**	**0.912**	**0.813**	**0.895**	**0.920**	**0.850**

Note: The bold part is the optimal value.

outperforms other state-of-the-art methods in the DSC metric and, except for the TC region, exceeds the performance of the most advanced methods in the HD metric for all other categories. Specifically, the DSC values have increased by 1.1 %, 0.8 %, and 0.5 %, respectively. At the same time, the HD values have decreased by 0.013 and 0.017 in the WT and ET regions. Furthermore, the segmentation results of all categories are superior to other state-of-the-art methods in Sensitivity metrics, with only the TC region surpassing other leading methods in terms of Precision. Specifically, there have been improvements of 0.4 %, 0.2 %, and 0.3 % in Sensitivity and an increase of 0.5 % in Precision for the TC region. These findings demonstrate the stability and generalization capabilities of our model.

(4) Comparison of computational complexity and number of parameters.

MVSI-Net improves segmentation accuracy and has a limited number of parameters (Params) compared to similar models. Additionally, its computational complexity (Flops) compared to other models is not the highest. Table 11 displays the computational complexity and parameter usage of different models. Specifically, our model has Flops of 7.32 × 1010, which is lower than UNet3+, ResUnet-a, and AttentionUnet. Our model has Params of 4.76 × 107, which is lower than ResUnet-a, MultiResUnet, and AttentionUnet. Through experimental verification, it has been demonstrated that our model effectively improves segmentation accuracy while controlling the number of parameters and computational complexity.

5. Conclusions and future directions

This paper proposes MVSI-Net, an MRI brain tumor segmentation method that integrates a multi-view attention mechanism and multi-scale feature interaction to precisely segment brain tumors. Considering the uncertainty in the location, shape, and size of brain tumors and the unclear segmentation of the edge contours of the three categories:

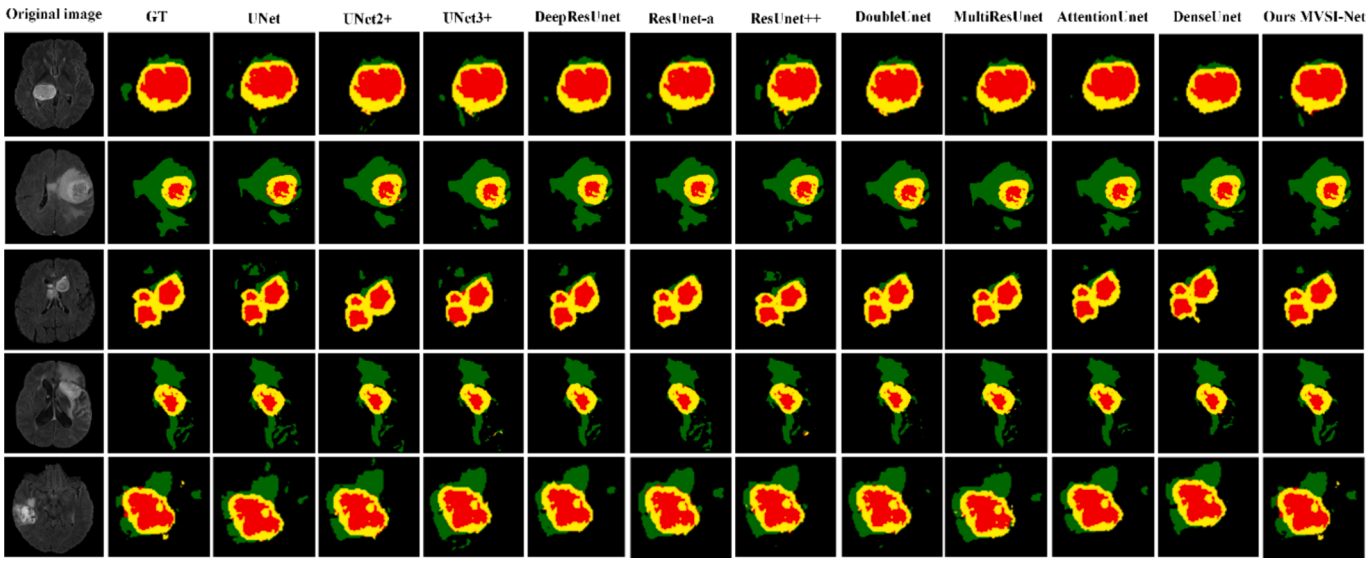

Fig. 10. Segmentation results of different methods on the BraTS 2019 dataset.

Table 7
Performance comparison of different methods on DSC and HD.

Network	DSC			HD		
	WT	TC	ET	WT	TC	ET
UNet [18]	0.804	0.834	0.723	2.772	1.608	2.931
UNet2+ [41]	0.787	0.822	0.711	2.798	1.661	2.963
UNet3+ [42]	0.795	0.842	0.727	2.702	1.658	2.740
DeepResUnet [43]	0.828	0.864	0.770	2.550	1.469	2.661
ResUnet-a [44]	0.788	0.838	0.713	2.815	1.604	2.981
ResUnet++ [45]	0.790	0.823	0.719	2.760	1.570	2.908
DoubleUnet [46]	0.845	0.852	0.754	2.796	1.536	2.823
MultiResUnet [21]	0.840	0.848	0.776	2.657	1.481	2.708
AttentionUnet [47]	0.836	0.859	0.780	2.601	1.502	2.804
DenseUnet [48]	0.828	0.865	0.768	2.578	1.459	2.690
Ours MVSI-Net	**0.876**	**0.875**	**0.817**	**2.501**	**1.426**	**2.602**

Note: The bold part is the optimal value.

Table 8
Performance Comparison of Different Methods on Precision and Sensitivity.

Network	Precision			Sensitivity		
	WT	TC	ET	WT	TC	ET
UNet [18]	0.842	0.897	0.761	0.811	0.879	0.775
UNet2+ [41]	0.848	0.885	0.767	0.805	0.872	0.754
UNet3+ [42]	0.846	0.877	0.762	0.756	0.852	0.670
DeepResUnet [43]	0.890	0.911	0.829	0.842	0.901	0.797
ResUnet-a [44]	0.828	0.901	0.747	0.837	0.880	0.789
ResUnet++ [45]	0.854	0.910	0.787	0.835	0.890	0.790
DoubleUnet [46]	0.805	0.916	0.730	0.839	0.877	0.778
MultiResUnet [21]	0.887	0.904	0.803	0.837	0.886	0.787
AttentionUnet [47]	0.853	0.804	0.824	0.866	0.887	0.798
DenseUnet [48]	0.880	0.924	0.815	0.842	0.890	0.799
Ours MVSI-Net	**0.891**	**0.934**	**0.829**	**0.866**	**0.910**	**0.825**

Note: The bold part is the optimal value.

Table 9
Performance comparison of different methods on DSC and HD.

Network	DSC			HD		
	WT	TC	ET	WT	TC	ET
UNet [18]	0.819	0.837	0.750	2.751	1.618	2.937
UNet2+ [41]	0.816	0.833	0.745	2.763	1.641	2.941
UNet3+ [42]	0.820	0.841	0.741	2.750	1.650	2.900
DeepResUnet [43]	0.818	0.849	0.762	2.740	**1.515**	2.864
ResUnet-a [44]	0.813	0.847	0.741	2.761	1.649	2.909
ResUnet++ [45]	0.816	0.832	0.757	2.756	1.650	2.917
DoubleUnet [46]	0.838	0.841	0.770	2.725	1.661	2.927
MultiResUnet [21]	0.839	0.853	0.767	2.681	1.547	2.856
AttentionUnet [47]	0.842	0.830	0.776	2.746	1.717	2.953
DenseUnet [48]	0.840	0.850	0.775	2.675	1.562	2.841
Ours MVSI-Net	**0.853**	**0.861**	**0.781**	**2.662**	1.523	**2.824**

Note: The bold part is the optimal value.

Table 10
Performance Comparison of Different Methods on Precision and Sensitivity.

Network	Precision			Sensitivity		
	WT	TC	ET	WT	TC	ET
UNet [18]	0.880	0.907	0.803	0.828	0.871	0.782
UNet2+ [41]	0.884	0.904	0.808	0.822	0.868	0.773
UNet3+ [42]	0.885	0.899	0.810	0.816	0.871	0.766
DeepResUnet [43]	0.818	0.904	0.745	0.860	0.887	0.821
ResUnet-a [44]	0.878	0.906	0.804	0.830	0.875	0.777
ResUnet++ [45]	0.883	0.894	0.810	0.824	0.874	0.774
DoubleUnet [46]	**0.891**	0.901	0.818	0.847	0.880	0.797
MultiResUnet [21]	0.889	0.919	**0.828**	0.847	0.880	0.804
AttentionUnet [47]	0.869	0.891	0.796	0.864	0.879	0.810
DenseUnet [48]	0.882	0.918	0.812	0.846	0.876	0.822
Ours MVSI-Net	0.870	**0.924**	0.803	**0.868**	**0.889**	**0.825**

Note: The bold part is the optimal value.

Table 11
Computational complexity and number of parameters for different models.

Model	Flops	Params
UNet [18]	6.29×10^{10}	3.94×10^{7}
UNet2+ [41]	2.73×10^{10}	9.16×10^{6}
UNet3+ [42]	1.57×10^{11}	2.70×10^{7}
DeepResUnet [43]	4.49×10^{10}	3.19×10^{7}
ResUnet-a [44]	9.53×10^{10}	5.28×10^{7}
ResUnet++ [45]	5.55×10^{10}	1.45×10^{7}
DoubleUnet [46]	4.23×10^{10}	2.93×10^{7}
MultiResUnet [21]	5.61×10^{10}	5.17×10^{7}
AttentionUnet [47]	7.80×10^{10}	5.24×10^{7}
DenseUnet [48]	2.97×10^{10}	1.89×10^{6}
Ours MVSI-Net	7.32×10^{10}	4.76×10^{7}

the whole tumor region, the tumor core region, and the tumor enhancement region, we propose the multi-view attention (MVA) mechanism and multi-scale feature interaction (MSI) module, respectively. According to the different levels of attention paid to the focal region by various network locations, MVA was added to the last two layers of the encoder and decoder separately to realize the localization of the focal region and better focus on the areas of interest. To segment different tumor edge contours precisely, we replace the double convolution module at the bottleneck of the UNet network with MSI, which selects useful feature information from multiple receptive fields.

We conducted experiments on the publicly available brain tumor challenge datasets, BraTS 2019, BraTS 2020, and BraTS 2021. The experimental results demonstrate that MVSI-Net achieves superior segmentation outcomes than similar approaches. Furthermore, in the future, we will emphasize collaboration with medical professionals and institutions to ensure the reliability and effectiveness of the proposed model. Before applying the model in the brain tumor department of the hospital, we will engage domain experts to compare the model's segmentation results with manually segmented results by doctors and evaluate the model's performance across different real-life cases. We will also conduct clinical feasibility studies to assess the adaptability and accuracy of the model in various cases, while closely monitoring possible false positive or false negative results. Simultaneously, we plan to enhance the segmentation performance of MRI brain tumors using a 3D network architecture and extend this method to other categories of medical image datasets to improve the model's generalization capability.

6. Institutional review board statement

Not applicable.

7. Informed consent statement

Not applicable.

8. Data availability statement

Not applicable.

Funding

This paper is partially supported by the National Natural Science Foundation of China (62276092); Key Science and Technology Program of Henan Province, CN (212102310084); Key Scientific Research Projects of Colleges and Universities in Henan Province, CN (22A520027); MRC (MC_PCss_17171); Royal Society (RP202G0230); BHF (AA/18/3/34220); Hope Foundation for Cancer Research (RM60G0680); GCRF (P202PF11); Sino-UK Industrial Fund (RP202G0289); LIAS (P202ED10, P202RE969); Data Science Enhancement Fund (P202RE237); Fight for Sight (24NN201); Sino-UK Education Fund (OP202006); BBSRC (RM32G0178B8).

CRediT authorship contribution statement

Junding Sun: Data curation, Methodology, Validation, Writing – original draft. **Ming Hu:** Formal analysis, Methodology, Software, Visualization. **Xiaosheng Wu:** Conceptualization, Data curation, Formal analysis, Software. **Chaosheng Tang:** Investigation, Resources, Visualization, Writing – review & editing. **Husam Lahza:** Methodology, Software, Validation, Writing – review & editing. **Shuihua Wang:** Methodology, Visualization, Writing – original draft, Writing – review & editing. **Yudong Zhang:** Methodology, Resources, Validation, Writing – review & editing.

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.

Data availability

The authors do not have permission to share data.

References

[1] Q. Li, Z. Gao, Q. Wang, J. Xia, H. Zhang, H. Zhang, H. Liu, S. Li, Glioma segmentation with a unified algorithm in multimodal MRI images, IEEE Access 6 (2018) 9543–9553.
[2] H. Li, G. Jiang, J. Zhang, R. Wang, Z. Wang, W.-S. Zheng, B. Menze, Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images, Neuroimage 183 (2018) 650–665.
[3] A. Ghosh, S. Thakur, Review of Brain Tumor MRI Image Segmentation Methods for BraTS Challenge Dataset, in: 2022 12th International Conference on Cloud Computing, Data Science & Engineering (confluence); 2022, IEEE, 2022, pp. 405–410.
[4] P. Jyothi, A.R. Singh, Deep learning models and traditional automated techniques for brain tumor segmentation in MRI: a review, Artif. Intell. Rev. 56 (4) (2023) 2923–2969.
[5] T.A. Soomro, L. Zheng, A.J. Afifi, A. Ali, S. Soomro, M. Yin, J. Gao, Image segmentation for MR brain tumor detection using machine learning: A Review, IEEE Rev. Biomed. Eng. (2022).
[6] A. Zhao, G. Balakrishnan, F. Durand, J.V. Guttag, A.V. Dalca, Data augmentation using learned transformations for one-shot medical image segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019; 2019. p. 8543-8553.
[7] O. Tarkhaneh, H. Shen, An adaptive differential evolution algorithm to optimal multi-level thresholding for MRI brain image segmentation, Expert Syst. Appl. 138 (2019) 112820.
[8] A. Srinivasa Reddy, R.P. Chenna, MRI brain tumor segmentation and prediction using modified region growing and adaptive SVM, Soft. Comput. 25 (2021) 4135–4148.
[9] L. Sun, W. Shao, M. Wang, D. Zhang, M. Liu, High-order feature learning for multi-atlas based label fusion: Application to brain segmentation with MRI, IEEE Trans. Image Process. 29 (2019) 2702–2713.
[10] Y. Liu, Y. Wei, C. Wang, Subcortical brain segmentation based on atlas registration and linearized kernel sparse representative classifier, IEEE Access 7 (2019) 31547–31557.
[11] Y. Jiang, J. Hou, X. Xiao, H. Deng, A brain tumor segmentation new method based on statistical thresholding and multiscale CNN. Intelligent Computing Methodologies: 14th International Conference, ICIC 2018, Wuhan, China, August 15-18, 2018, Proceedings, Part III 14; 2018: Springer; 2018. p. 235-245.
[12] A. Sharma, S. Kumar, S.N. Singh, Brain tumor segmentation using DE embedded OTSU method and neural network, Multidim. Syst. Sign. Process. 30 (2019) 1263–1291.
[13] R. Meier, V. Karamitsou, S. Habegger, R. Wiest, M. Reyes, Parameter learning for CRF-based tissue segmentation of brain tumors. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: First International Workshop, Brainles 2015, Held in Conjunction with MICCAI 2015, Munich, Germany, October 5, 2015, Revised Selected Papers 1; 2016: Springer; 2016. p. 156-167.
[14] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, Y. Fan, A deep learning model integrating FCNNs and CRFs for brain tumor segmentation, Med. Image Anal. 43 (2018) 98–111.
[15] K. Hao, S. Lin, J. Qiao, Y. Tu, A generalized pooling for brain tumor segmentation, IEEE Access 9 (2021) 159283–159290.
[16] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition; 2015; 2015. p. 3431-3440.

[17] H. Shen, R. Wang, J. Zhang, S. McKenna, Multi-task fully convolutional network for brain tumour segmentation. Medical Image Understanding and Analysis: 21st Annual Conference, MIUA 2017, Edinburgh, UK, July 11–13, 2017, Proceedings 21; 2017: Springer; 20p. 239-248.
[18] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18; 2015: Springer; 2015. p. 234-241.
[19] H. Dong, G. Yang, F. Liu, Y. Mo, Y. Guo, Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks. Medical Image Understanding and Analysis: 21st Annual Conference, MIUA 2017, Edinburgh, UK, July 11–13, 2017, Proceedings 21; 2017: Springer; 2017. p. 506-517.
[20] M. Aghalari, A. Aghagolzadeh, M. Ezoji, Brain tumor image segmentation via asymmetric/symmetric UNet based on two-pathway-residual blocks, Biomed. Signal Process. Control 69 (2021) 102841.
[21] N. Ibtehaz, M.S. Rahman, MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation, Neural Netw. 121 (2020) 74–87.
[22] T. Fan, G. Wang, Y. Li, H. Wang, Ma-net: A multi-scale attention network for liver and tumor segmentation, IEEE Access 8 (2020) 179656–179665.
[23] N. Siddique, S. Paheding, C.P. Elkin, V. Devabhaktuni, U-net and its variants for medical image segmentation: A review of theory and applications, IEEE Access 9 (2021) 82031–82057.
[24] D. Maji, P. Sigedar, M. Singh, Attention Res-UNet with Guided Decoder for semantic segmentation of brain tumors, Biomed. Signal Process. Control 71 (2022) 103077.
[25] J. Zhang, Z. Jiang, J. Dong, Y. Hou, B. Liu, Attention gate resU-Net for automatic MRI brain tumor segmentation, IEEE Access 8 (2020) 58533–58545.
[26] T. Zhou, S. Ruan, P. Vera, S. Canu, A Tri-Attention fusion guided multi-modal segmentation network, Pattern Recogn. 124 (2022) 108417.
[27] Z. Huang, Y. Zhao, Y. Liu, G. Song, GCAUNet: A group cross-channel attention residual UNet for slice based brain tumor segmentation, Biomed. Signal Process. Control 70 (2021) 102958.
[28] G. Cheng, H. Ji, Z. Ding, Spatial-channel relation learning for brain tumor segmentation, Med. Phys. 47 (10) (2020) 4885–4894.
[29] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, K. Barnard, Attentional feature fusion. Proceedings of the IEEE/CVF winter conference on applications of computer vision; 2021; 2021. p. 3560-3569.
[30] Z. Shen, I. Bello, R. Vemulapalli, X. Jia, C.-H. Chen, Global self-attention networks for image recognition. arXiv preprint arXiv:201003019 2020.
[31] Z. Xiao, K. He, J. Liu, W. Zhang, Multi-view hierarchical split network for brain tumor segmentation, Biomed. Signal Process. Control 69 (2021) 102897.
[32] M. Mazher, A. Qayyum, D. Puig, M. Abdel-Nasser, Effective Approaches to Fetal Brain Segmentation in MRI and Gestational Age Estimation by Utilizing a Multiview Deep Inception Residual Network and Radiomics, Entropy 24 (12) (2022) 1708.
[33] Y. Lu, Y. Chang, Z. Zheng, Y. Sun, M. Zhao, B. Yu, C. Tian, Y. Zhang, GMetaNet: Multi-scale ghost convolutional neural network with auxiliary MetaFormer decoding path for brain tumor segmentation, Biomed. Signal Process. Control 83 (2023) 104694.
[34] N.S. Syazwany, J.-H. Nam, S.-C. Lee, MM-BiFPN: multi-modality fusion network with Bi-FPN for MRI brain tumor segmentation, IEEE Access 9 (2021) 160708–160720.
[35] M.U. Rehman, S. Cho, J.H. Kim, K.T. Chong, Bu-net: Brain tumor segmentation using modified u-net architecture, Electronics 9 (12) (2020) 2203.
[36] J. Wang, J. Gao, J. Ren, Z. Luan, Z. Yu, Y. Zhao, Y. Zhao, DFP-ResUNet: Convolutional neural network with a dilated convolutional feature pyramid for multimodal brain tumor segmentation, Comput. Methods Programs Biomed. 208 (2021) 106208.
[37] X. Wei, F. Ye, H. Wan, J. Xu, W. Min, TANet: Triple Attention Network for medical image segmentation, Biomed. Signal Process. Control 82 (2023) 104608.
[38] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, S.-M. Hu, Visual attention network, Computational Visual Media (2023) 1–20.
[39] Q. Xu, Z. Ma, H. Na, W. Duan, DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation, Comput. Biol. Med. 154 (2023) 106626.
[40] D. Misra, T. Nalamada, A.U. Arasanipalai, Q. Hou, Rotate to attend: Convolutional triplet attention module. Proceedings of the IEEE/CVF winter conference on applications of computer vision; 2021; 2021. p. 3139-3148.
[41] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation. 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; 2018: Springer; 2018. p. 3-11.
[42] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, J. Wu, in: Unet 3+: A Full-Scale Connected Unet for Medical Image Segmentation, IEEE, 2020, pp. 1055–1059.
[43] Z. Zhang, Q. Liu, Y. Wang, Road extraction by deep residual u-net, IEEE Geosci. Remote Sens. Lett. 15 (5) (2018) 749–753.
[44] F.I. Diakogiannis, F. Waldner, P. Caccetta, C. Wu, ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data, ISPRS J. Photogramm. Remote Sens. 162 (2020) 94–114.
[45] Jha D, Smedsrud PH, Riegler MA, Johansen D, De Lange T, Halvorsen P, Johansen HD. Resunet++: An advanced architecture for medical image segmentation. 2019 IEEE international symposium on multimedia (ISM); 2019: IEEE; 2019. p. 225-2255.
[46] D. Jha, M.A. Riegler, D. Johansen, P. Halvorsen, J.HD. Doubleu-net, in: A Deep Convolutional Neural Network for Medical Image Segmentation, IEEE, 2020, 2020, pp. 558–564.
[47] Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:180403999 2018.
[48] Kaku A, Hegde CV, Huang J, Chung S, Wang X, Young M, Radmanesh A, Lui YW, Razavian N. DARTS: DenseUnet-based automatic rapid tool for brain segmentation. arXiv preprint arXiv:191105567 2019.