

Article

SSGNet: Selective Multi-Scale Receptive Field and Kernel Self-Attention Based on Group-Wise Modality for Brain Tumor Segmentation

Bin Guo ^{1,2} , Ning Cao ^{1,*}, Peng Yang ² and Ruihao Zhang ²¹ College of Information Science and Engineering, Hohai University, Nanjing 210098, China; drbinguo@163.com² College of Computer and Information Engineering, Xinjiang Agricultural University, Urumqi 830052, China; yangpeng990625@163.com (P.Y.); dlstudyzrh@163.com (R.Z.)

* Correspondence: caoning@vip.163.com

Abstract: Medical image processing has been used in medical image analysis for many years and has achieved great success. However, one challenge is that medical image processing algorithms ineffectively utilize multi-modality characteristics to further extract features. To address this issue, we propose SSGNet based on UNet, which comprises a selective multi-scale receptive field (SMRF) module, a selective kernel self-attention (SKSA) module, and a skip connection attention module (SCAM). The SMRF and SKSA modules have the same function but work in different modality groups. SMRF functions in the T1 and T1ce modality groups, while SKSA is implemented in the T2 and FLAIR modality groups. Their main tasks are to reduce the image size by half, further extract fused features within the groups, and prevent information loss during downsampling. The SCAM uses high-level features to guide the selection of low-level features in skip connections. To improve performance, SSGNet also utilizes deep supervision. Multiple experiments were conducted to evaluate the effectiveness of our model on the BraTS2018 dataset. SSGNet achieved Dice coefficient scores for the whole tumor (WT), tumor core (TC), and enhancing tumor (ET) of 91.04, 86.64, and 81.11, respectively. The results show that the proposed model achieved state-of-the-art performance compared with more than twelve benchmarks.

Keywords: brain tumor segmentation; MRI; medical image; deep learning; multi-modality fusion



Citation: Guo, B.; Cao, N.; Yang, P.; Zhang, R. SSGNet: Selective Multi-Scale Receptive Field and Kernel Self-Attention Based on Group-Wise Modality for Brain Tumor Segmentation. *Electronics* **2024**, *13*, 1915. <https://doi.org/10.3390/electronics13101915>

Academic Editors: Luca Mesin, Antonio Lanata, Irena Galić and Marija Habijan

Received: 22 April 2024

Revised: 11 May 2024

Accepted: 12 May 2024

Published: 14 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Brain tumors, also known as intracranial tumors, are the most common disease in the field of neurosurgery [1,2]. Generally, if a tumor is found in the brain, it is relatively malignant and difficult to cure [3]. Additionally, symptoms such as increased intracranial pressure, headache, vomiting, blurred vision, and coma may occur [4–6]. If brain tumors are not treated in a timely manner, they can grow larger and cause more severe clinical symptoms over time and as the condition progresses [7]. The most common type of malignant brain tumor is glioma [8,9], which is classified in ascending order of malignancy into grade 1, grade 2, grade 3, and grade 4 gliomas [10]. Grade 1 and grade 2 gliomas are relatively benign, and patients can survive for a long time after surgery. Grade 3 and grade 4 gliomas are malignant gliomas and have blurred boundaries with the surrounding brain tissue. There are many diagnosis methods for brain tumors, and a plan should be formulated based on the patient's physical condition [11]. Typically, diagnostic methods include skull plain films [12], cerebral angiography [13], computed tomography (CT) [14], magnetic resonance imaging (MRI) [15], and neuronuclear medicine. MRI is suitable for observing lesions in the sellar region and skull base [16], especially in the posterior cranial fossa, due to its many advantages, e.g., no skeletal artifacts, multi-dimensional scanning cross-sections, and multiple parameter imaging. The application of the vascular flow effect can display the relationship and blood supply between the tumor and surrounding blood

vessels without the use of contrast agents, making MRI an important tool for the diagnosis of brain tumors [17]. Most brain tumors have an invasive growth pattern [18], where the boundary between the tumor and brain tissue is unclear. This makes it difficult to completely segment tumors through human diagnosis [19].

Unlike the traditional human segmentation of brain tumors, deep-learning-based methods can automatically and accurately perform segmentation, to avoid unnecessary human errors [20,21]. At present, although CNN-based methods have achieved excellent performance in the field of medical image segmentation, they still cannot fully satisfy the strict requirements for segmentation accuracy in practical medical applications. This remains a challenging task in medical image segmentation [22]. One of the challenges is the effective utilization of multi-modality data. It is well known that there are four modalities included in the BraTS2018 dataset, namely T1-Weighted (T1), T1-enhanced contrast (T1ce), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR) [23–26]. T1 imaging can be used to display normal brain tissue structures and help doctors detect pathological changes in the brain. T1 imaging is more sensitive to the diagnosis of intracranial hemorrhage, sclerotic lesions, tumors, and other lesions [27]. T1ce imaging involves T1 imaging performed on the basis of injecting contrast agents, which can help doctors detect tumors, inflammation, infections, and other lesions in the brain. T1ce imaging has high diagnostic sensitivity for brain tumors [28]. T2 weighted imaging is sensitive to tissue water content and can effectively distinguish different structures, such as tumors, blood vessels, cerebrospinal fluid, and white matter [29]; thus, it achieves high contrast. FLAIR imaging is a modality that removes cerebrospinal fluid signals and that is sensitive to gray matter lesions, especially in the early stages of the disease. FLAIR imaging can be used to detect various brain diseases, such as meningioma, multiple sclerosis, and so on [30]. However, FLAIR imaging is not sensitive enough to diagnose lesions such as calcium deposition and bleeding [31]. UNet, which integrates the characteristics of the four modalities, has achieved great success in brain tumor image segmentation. It extracts the features of the four modalities through the encoding block in the left branch and restores the image in the decoding layer in the right branch, thereby achieving segmentation of brain tumor images. Over time, many variants of UNet have emerged, most of them simultaneously integrating the four modalities to achieve segmentation. Although the attention mechanism has played a significant role in improving performance, the utilization of the characteristics of each modality is still insufficient [32]. Transformer, which has achieved great success in natural language processing [33], also performs exceptionally well in medical image segmentation, utilizing the four modalities for simultaneous and identical processing. Recently, certain researchers have specialized in the performance of each modality in segmentation, but further improvement in the results is needed. Owing to the advantages and disadvantages of the different modalities, combining them with different attention mechanisms would definitely improve the results; that said, there has been less research in this area.

The main contributions of this study are as follows:

- (1) We proposed SSGNet, which can fully utilize the special relationship combinations between modalities to improve segmentation performance.
- (2) We designed two different attention mechanisms, SMRF and SKSA, for use in the downsampling process of the encoder and applied them to groups T1 and T1ce, and groups T2 and FLAIR, respectively, to reduce information loss and capture more detailed features.
- (3) We developed a SCAM to be used for high-level semantic features, to guide the selection of low-level semantic features in skip connection layers.
- (4) Depth supervision was introduced to further improve the segmentation effect. The proposed model achieved state-of-the-art performance compared with more than twelve benchmarks.

The rest of this paper is structured as follows. Related work is described in Section 2. The methods and details of SSGNet are presented in Section 3. The datasets, architecture parameters, evaluation metrics, and experimental configurations are presented in Section 4

in detail. Finally, Section 5 comprises the comparison of results and ablation experiments. The discussion and conclusions are presented in Sections 6 and 7, respectively.

2. Related Work

2.1. Deep-Learning-Based Methods for Medical Image Segmentation

Deep learning has achieved good performance in brain tumor segmentation tasks, effectively solving manual errors and improving segmentation efficiency. This has inspired a large number of researchers to explore its application and further advancements [34]. In the early days, convolutional neural networks (CNNs) demonstrated the potential of deep learning in image processing and strengthened the confidence of many researchers. Jonathan Long et al. [35] proposed fully convolutional networks (FCNs), to lay the foundation for the application of deep learning in the field of semantic segmentation, solving the problem caused by a CNN that only inputs fixed image sizes. UNet, which was proposed by Olaf Ronneberger et al. [36], has attracted the attention of medical image researchers due to its unique structure. It has a symmetrical network structure, resulting in relatively fewer network parameters. This is important for tasks such as medical image segmentation with small amounts of data, and it can reduce the risk of overfitting and improve the generalization ability of the model [37]. UNet introduced a skip connection structure to retain more spatial and contextual information, helping to improve the accuracy of the segmentation results. For brain tumor segmentation tasks, it can directly process multi-channel inputs, allowing the model to fully utilize the information between different channels and improving the segmentation performance. In recent years, a large number of variants of UNet have emerged, which have achieved excellent performance in brain tumor segmentation tasks. Özgün Çiçek et al. [38] designed a 3D-UNet, which captures more contextual information, especially depth direction. A network architecture named UNet++ with dense connections was proposed by Zhou et al. [39], which makes full use of the characteristics of different levels to improve accuracy. In UNet++, the network parameters are reduced to an acceptable range using a deep supervision mechanism. The V-Net architecture based on UNet, which uses a convolution layer to replace upsampling and downsampling layers and performs well, was proposed by Fausto Milletari et al. [40]. Chen et al. [41] proposed a 3D dilated multifiber network (DMF) architecture with 3D multi-fibers to solve the problem of the model being unsuitable for practical large-scale applications. To encode the multi-scale multi-view context, Luo et al. [42] developed HDC-Net using group decoupled convolutions. Guan et al. [43] designed MVKSNet to better represent the complex boundaries and improve segmentation accuracy by adopting a multi-branch structure with different sized convolution kernels. Liu et al. [44] proposed SPA-Net based on 3D-UNet, which focuses on multi-scale spatial details and contextual information. In summary, the UNet architecture and multi-scale structure can effectively improve the effectiveness of brain tumor segmentation, so we adopted a U-shaped structure as our architecture.

2.2. The Attention-Based Module for Medical Image Segmentation

The attention mechanism pays more attention to the important parts of the network and ignores the unimportant ones [45], and it has recently been applied in many networks that have effectively improved the segmentation performance. In networks based on UNet, it is generally applied in the upsampling layer, skip connection layer, or feature extraction blocks of each layer. Liu et al. [46] proposed SGEResU-Net with residual blocks and spatial group-wise enhance (SGE) attention blocks. To enhance the feature learning ability and reduce noise, SGE attention was introduced in the skip connection layers. To improve brain tumor segmentation, Tian et al. [47] introduced axial attention to AABTS-Net. The axial attention mechanism was applied between the upsampling layer and the decoder block to capture local-global contextual information. Yang et al. [48] designed CFHA-Net, which combines a cross-scale feature, hybrid attention mechanism, hybrid pooling module, and hybrid upsampling module. The cross-scale fusion module is applied to the encoder part,

triple hybrid attention is exploited for the skip connection layers, the hybrid pooling is used during downsampling, and the hybrid upsampling module is utilized during upsampling.

Recent studies have revealed that Transformers have shown good performance in medical image segmentation. Chen et al. [49] proposed TransUNet, which combines both a Transformer and UNet, to enhance finer details by recovering localized spatial information from medical image segmentation. Lu et al. [50] combined the local modeling of CNN and the long-range representation of a Transformer in an auxiliary MetaFormer. Zhang et al. [51] incorporated an efficient spatial-channel attention layer into the bottleneck layer for global interaction, to further capture high-level semantic information and highlight important semantic features from the encoder path output. Undoubtedly, using the attention mechanism in various parts of UNet has indeed improved segmentation performance, but it is still less commonly used in the downsampling process.

2.3. Modality-Based Methods for Medical Image Segmentation

Recently, many researchers have shown great interest in utilizing segmentation results from different modalities, such as cross-modalities and group-wise modalities. Many studies have shown that fully utilizing the interaction and fusion between different modalities can effectively improve segmentation performance. Wang et al. [52] proposed a novel end-to-end modality-pairing learning method for brain tumor segmentation, which exploits different modality features to solve the problem of ignoring the latent relationship among different modalities. The method achieved second place in the BraTS2020 Challenge in the tumor segmentation tasks. To demonstrate the important role of multi-modality in brain tumor segmentation, a multi-modality and single-modality feature recalibration network (MSFR-Net) was proposed for brain tumor segmentation by Li et al. [53]. Zhou et al. [54] designed a multi-modality segmentation network guided by a novel tri-attention fusion, which is used in the encoder layer; in their work, the four modalities could capture modality-specific features independently. Dual attention can also be used to emphasize the useful parts of each modality at different positions. Lin et al. [55] developed CKD-TransBTS to divide modalities into two groups, according to the imaging principles of MRI. The model achieved state-of-the-art segmentation performance compared with all the competitors. To fully exploit the brain tumor features of different modalities, Jiao et al. [56] proposed RFTNet, which combines a dual-branch vision Transformer to effectively fuse the images of the different modalities. Zhuang et al. designed ACMinet [57] to refine multi-modality features. The cross-modality feature interaction module is made up of three stages: grouping, interaction, and fusion. Previous studies have shown that the study of group-wise modality is valuable for improving segmentation performance and fully utilizing modal information.

3. Methodology

3.1. Network Architecture

The overall architecture of our method is depicted in Figure 1. The proposed SSGNet is an improvement on the UNet structure, so our network is similar to UNet. In the left half of the network is the encoder branch, and in the right half of the network is the decoder section. The skip connections in the middle connect the encoder module and decoder module of the corresponding layer. The encoders mainly perform feature extraction of images, and the decoders enable image restoration. The middle skip connections allow communication between the encoders and decoders, using features with original information to help the decoders more effectively restore images. In each stage of the encoder, each modality group contains two convolutions with $3 \times 3 \times 3$ kernels and a stride of 1. The group normalization (GN) and Gaussian error linear unit (GeLU) are employed after each convolution. In order to prevent overfitting and improve the generalization ability of the model, a dropout regularization method was added to Stage 1, which has a value of 0.3. There are also two convolutions with two $3 \times 3 \times 3$ kernels and a stride of 1 in each decoder layer. In the upsampling stage, deconvolution is used to restore the resolution of the image. Empirically,

the residual block is used between the convolutions, then a squeeze-and-excitation (SE) module is added to the residual block. To make the low-level semantic feature more effective, a SCAM is located in the third and fourth layers of our SSGNet. To provide more detailed information on the features that are transmitted to the decoder, the features from the T1 and T1ce groups, as well as the information from the T2 and FLAIR groups, are combined and convolved through skip connection layers before being sent to the corresponding layer of the decoder. Two convolutions with kernel sizes of $3 \times 3 \times 3$ and a stride of 1 are used in each layer of the first and second layers, and the GN and GeLU are utilized after each convolution of the skip connections. In particular, to leverage the advantages between MRI modalities, the selective multi-scale receptive field (SMRF) plays an important role in each stage in the T1 and T1ce modalities, and the selective kernel self-attention (SKSA) enhances its value in each stage in the T2 and FLAIR modalities during downsampling. The input dimensions are $2 \times 128 \times 128 \times 128$. As the number of layers increases, the image shape becomes smaller, with a feature size of $256 \times 16 \times 16 \times 16$ in the bottleneck layer. Sequentially, after the last layer of the decoder, the shape is restored to $32 \times 128 \times 128 \times 128$. Finally, after the classifier, the dimensions are $4 \times 128 \times 128 \times 128$.

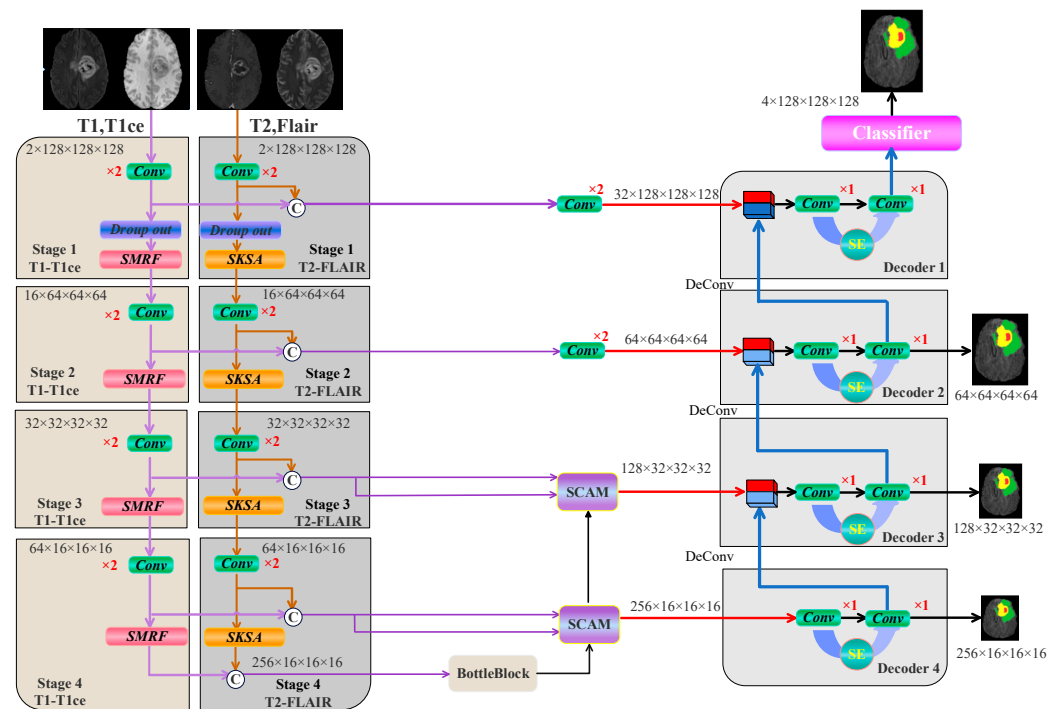


Figure 1. Illustration of the proposed SSGNet for brain tumor image segmentation.

3.2. Selective Multi-Scale Receptive Field Module (SMRF)

The receptive field is one of the challenges of medical image segmentation, which plays an important role in further feature extraction and information retention during the downsampling stages. Zhang et al. [58] proposed receptive-field attention (RFAConv) to determine the significance of each feature in the receptive field, which further improves the efficiency of the extraction features and the performance of the model. Inspired by RFAConv, we proposed a selective multi-scale receptive field module (SMRF), which is depicted in Figure 2. The T1 modality provides information about the anatomical situation. In T1, the white matter is represented as white, the gray matter is displayed as gray, and the cerebrospinal fluid is indicated as black. The T1ce sequence is used to create a contrast agent in the blood for MR imaging. The bright areas have an abundant blood supply, and the enhanced display indicates rich blood flow. The tumor site is the area with fast blood flow, and the T1ce sequence can further display the situation inside the tumor and distinguish between tumor and non-tumor lesions. The SMRF is an attention mechanism

that can focus on important information in key areas. Combining T1 and T1ce modalities allows us to more accurately learn the feature expression of brain tumors.

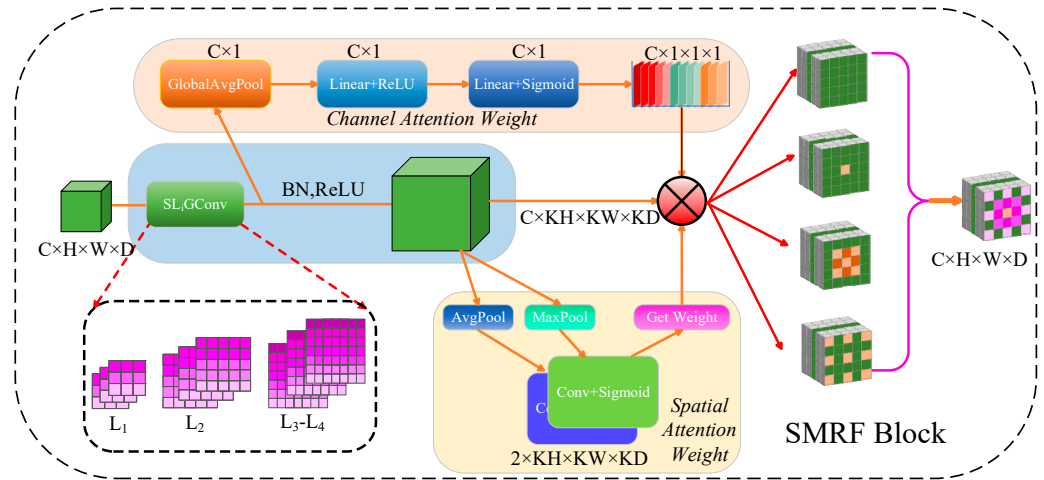


Figure 2. The structure of the proposed SMRF module.

The main purpose of this SMRF module is to maintain the receptive field, extract important features, and prevent information loss. Generally, in traditional UNet, the downsampling part uses max-pooling or convolution with a stride of 2 to change the image size. In this case, it not only lacks the ability to further extract features but can also easily lose information. Our model uses SL_iGConv , which is a group convolution with different kernel sizes, to perform downsampling operations, gradually reducing the image size. However, considering the image size of different layers, and to prevent information loss, different convolution kernels are used on different layers. Empirically, the number of groups is the same as the number of input channels. After the SL_iGConv operation, the image size becomes half of its original size. SL_iGConv selects different convolution kernels according to the position of the layer, which can be represented as follows:

$$U_i = f_{sli}(x) \quad (1)$$

where f_{sli} denotes the group convolution with different kernels of the SL_iGConv module, whose kernel sizes are $\{3 \times 3 \times 3, 5 \times 5 \times 5, 7 \times 7 \times 7, 7 \times 7 \times 7\}$. U_i is the result calculated by the SL_iGConv module. Index i represents the i th layer. We define the convolution kernels according to the different positions of the layer, as follows:

- (1) Layer 1: the kernel size is $3 \times 3 \times 3 \rightarrow U1 \rightarrow f_{sl1} \in R^{C \times H \times W \times D}$;
- (2) Layer 2: the kernel size is $5 \times 5 \times 5 \rightarrow U2 \rightarrow f_{sl2} \in R^{2C \times \frac{1}{2}H \times \frac{1}{2}W \times \frac{1}{2}D}$;
- (3) Layer 3: the kernel size is $7 \times 7 \times 7 \rightarrow U3 \rightarrow f_{sl3} \in R^{4C \times \frac{1}{4}H \times \frac{1}{4}W \times \frac{1}{4}D}$;
- (4) Layer 4: the kernel size is $7 \times 7 \times 7 \rightarrow U4 \rightarrow f_{sl4} \in R^{8C \times \frac{1}{8}H \times \frac{1}{8}W \times \frac{1}{8}D}$.

Channel attention aims to learn the correlation between different channels and automatically extract the importance of each feature channel. Finally, different weight coefficients are assigned to each channel to enhance important features and suppress unimportant ones. In the upper part of Figure 2, we adopt the channel attention weight, which can enhance the interaction between features from a channel perspective. The specific expression is as follows:

$$CA_i = LNS(LNR(GP(x))) \quad (2)$$

where x signifies the input, CA_i indicates the channel attention weight, LNS represents layer normalization and sigmoid, LNR indicates layer normalization and ReLU, and GP is the global average pool.

The spatial attention weight is presented in the bottom branch. The specific description is as follows:

$$SA_i = GW(CS(concat(AP(U_i), MP(U_i)))) \quad (3)$$

where U_i denotes the result of SL_iGConv , SA_i indicates the spatial attention weight, MP is max-pooling, AP means average pooling, CS is defined as convolution and sigmoid, and GW indicates the get weight, which can be written as follows:

$$GW_i(\bullet) = AP(GC(\bullet)) \quad (4)$$

where GC represents group convolution, whose convolution kernel is defined as 1, and the number of groups is identical to the number of input channels. To accurately extract image features with larger receptive fields, the three results above were multiplied. The specific description is as follows:

$$T_i = U_i \times SA_i \times CA_i \quad (5)$$

where T_i denotes the feature values of the large receptive field with spatial and channel weights, SA_i indicates the spatial attention weight, and CA_i indicates the channel attention weight.

At the end of the SL_iGConv block, we process the feature values T_i with multi-scale receptive fields, which can be represented as follows:

$$Z_{SMRF} = conv_{1 \times 1 \times 1}(cat(atrous(AP(T_i), \theta))) \quad (6)$$

where T_i is the input feature, and Z_{SMRF} denotes the output result. The *atrous* function is an atrous convolution operation with a rate of θ , $\theta \in \{1, 2, 3, 4\}$. AP represents the average pool. $Conv 1 \times 1 \times 1$ denotes convolution with a kernel of $1 \times 1 \times 1$.

3.3. Selective Kernel Self-Attention Module (SKSA)

The T2 signal is related to the water volume, and the T2 signal of many lesions is stronger than that of surrounding normal tissues, often showing a bright state. Therefore, the size of the lesions can be clearly seen from the T2 sequence. The MRI liquid attenuation reversal sequence, also known as FLAIR, can darken the cerebrospinal fluid in T2, allowing for clear edges of the brain tumor. We place T2 and FLAIR into a group to retrieve information on the size and boundaries of brain tumors. With similar operations to the SMRF module, the SKSA module, which is depicted in Figure 3, focuses more on the relationships between global features. The feature values of spatial and channel fusion with receptive fields from Formula (6) can then be obtained. The acquisition of the relationship between global features can be represented as follows:

$$Z_{SKSA} = \alpha \times torch.bmm(Conv(T_i), sigmoid(torch.bmm(Conv(T_i), Conv(T_i)^T))) + T_i \quad (7)$$

where Z_{SKSA} denotes the output result of the SKSA module, T_i denotes the input feature, and $Conv$ represents convolution with a kernel of $1 \times 1 \times 1$. α indicates learning parameters, which can learn suitable parameter values through model training. The *torch.bmm* operation is defined as a matrix product.

3.4. Skip Connection Attention Module (SCAM)

A skip connection is commonly used to transfer information from the encoder layer to the corresponding decoder layer. On the one hand, this can solve the problem of vanishing or exploding gradients, and on the other hand, it is concatenated to help the network better obtain features according to the channel dimension. One challenge is that the skip connection does not fully utilize its role in fusing deep information. Hang et al. [59] proposed a CAM, which uses high-level features to guide the selection of low-level features, establishing the relationship between the two. Inspired by this CAM, we designed a SCAM to capture feature representations effectively in the third and fourth skip connection layers.

A diagram of the SCAM is shown in Figure 4. Unlike the CAM, our model not only utilizes rough processed high-level semantic information but also incorporates high-level semantic information generated by two convolution operations. This shows that the selection of low-level features is more effective with high-level semantic information. As can be seen, the sizes of inputs A and B are twice as large as that of input C. To fuse the two features accurately, the feature size of input C is adjusted by deconvolution. The operation for input C is divided into two parts: one is used to calculate weights and fuse them with the features of input B, and the other is used to integrate them into the features after the size changes. Finally, the finely processed features for input A are integrated to enhance the expression of low-level semantic information. The SCAM can be represented as follows:

$$Z = \text{concat}(\text{conv}_{3 \times 3 \times 3}(b) \times \text{conv}_{1 \times 1 \times 1}(\text{GP}(c)) + \text{DC}_{3 \times 3 \times 3}(c), \text{conv}_{3 \times 3 \times 3}(\text{conv}_{3 \times 3 \times 3}(a))) \quad (8)$$

where $\text{conv}_{1 \times 1 \times 1}$ and $\text{conv}_{3 \times 3 \times 3}$ denote $1 \times 1 \times 1$ and $3 \times 3 \times 3$ convolutions, respectively. GP represents the global average pool. DC indicates deconvolution. a denotes input A, b is defined as input B, and c is defined as input C.

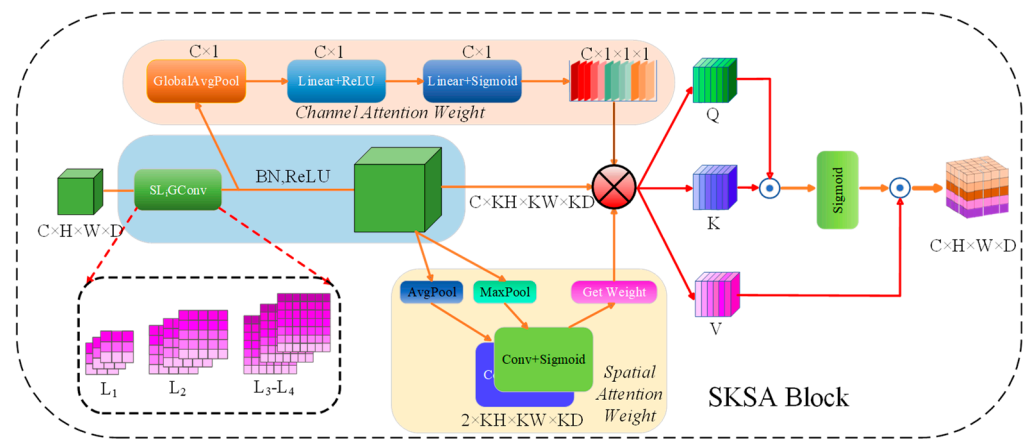


Figure 3. The structure of the proposed SKSA module.

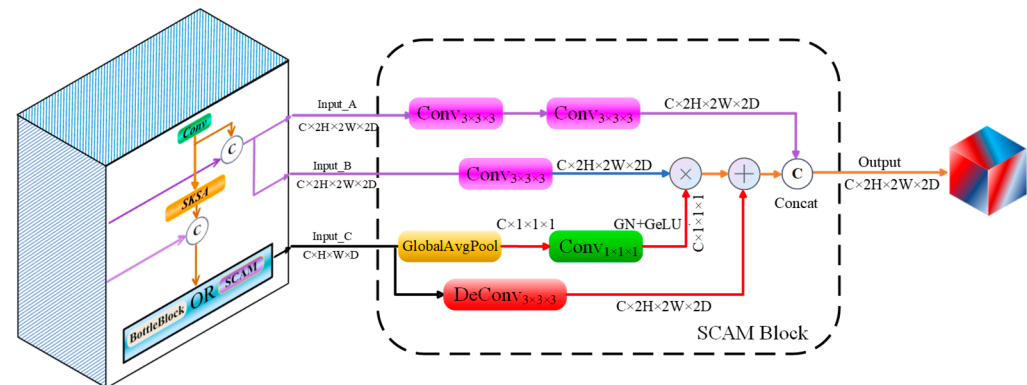


Figure 4. The structure of the SCAM. Input A and input B are at the same level, but they differ by one level from input C.

4. Experiments

4.1. Datasets and Preprocessing

The BraTS dataset refers to the brain tumor segmentation challenge dataset, which is a public medical image dataset used to research and develop brain tumor segmentation algorithms. The dataset is provided by multiple medical centers and contains image data from multiple patients who underwent brain tumor MRI. A total of 285 cases were used for training, and 66 cases were used for online validation in BraTS2018 [24–26], which is a popular database. Generally, the 285 cases all contain a ground truth labeled by board-certified

neuroradiologists. In comparison, the ground truths of 66 cases are hidden to the public, and the results can only be obtained through online validation. Our strategy involved using all 285 cases for our training model. Our prediction results were evaluated on the official BraTS platform (<https://www.med.upenn.edu/sbia/brats2018.html>) (accessed on 11 May 2024).

To enable our network to segment brain tumor images normally, we first read the BraTS2018 dataset into our program in the preprocessing stage. After processing with SimpleITK and MONAI, we used the Z-score method to standardize each image. Sequentially, we reduced the background as much as possible while ensuring that all of the brain was included, and then randomly re-cropped the fixed patch size of the image to $128 \times 128 \times 128$. All intensity values were clipped to the 1st and 99th percentiles of the non-zero voxel distribution of the volume. In this research, we used rotating, adding noise, blurring, and adding gamma as data augmentation techniques. After optimizing the model, we adjusted the image size back to the original image size. Finally, we submitted our results to the official platform for evaluation.

4.2. Implementation Details

Our network was constructed in Python 3.8.10 and PyTorch 1.10.0. A single NVIDIA GeForce RTX 3090 with 24 GB of memory and an Intel(R) Xeon(R) Platinum 8255C were used during training. As shown in Table 1, the initial learning rate was 1×10^4 , and the batch size was 1. Ranger was used to optimize our network. As an important part of the algorithm, the objective loss function played an important role in helping the model converge quickly. Unlike the hybrid loss, only the ordinary soft Dice loss [40] was trained in our network. The dimensions of the input in the first stage and the output of the last layer were $128 \times 128 \times 128$.

Table 1. Model parameter configuration.

Basic Configuration	Value
PyTorch Version	1.10.0
Python	3.8.10
GPU	NVIDIA GeForce RTX 3090(24G)
Cuda	11.3
Learning Rate	1×10^4
Optimizer	ranger
Batch Size	1
Input Size	$128 \times 128 \times 128$
Output Size	$128 \times 128 \times 128$

4.3. Evaluation Metrics

Quantitative and qualitative analyses were carried out using evaluation metrics, including the Dice similarity coefficient (Dice) score and the Hausdorff distance (HD).

Dice is a measure of the similarity between two effects. It is used to measure the similarity between network segmentation predicted results and manual masking in the field of image segmentation, which can be represented as follows:

$$Dice = \frac{2TP}{2TP + FP + FN} \quad (9)$$

where TP , FP , and FN represent true-positive cases, false-positive cases, and false-negative cases, respectively.

HD represents the maximum distance between the predicted segmentation region boundary and the real region boundary. The smaller the value, the smaller the prediction boundary segmentation error and the better the quality. The HD can be represented as follows:

$$HD(P, T) = \max\{sup_{t \in T} in f_{p \in P} d(t, p), sup_{p \in P} in f_{t \in T} d(t, p)\} \quad (10)$$

where t and p represent the real region boundary and predicted segmentation region boundary, respectively. $d(\bullet)$ represents the distance between t and p . $\sup$ denotes the supremum.

The sensitivity is referred to as the true-positive rate. It quantifies the accurate probability of complete positive detection. The sensitivity can be represented as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (11)$$

where TP and FN represent true-positive cases and false-negative cases, respectively. A higher sensitivity corresponds to a smaller discrepancy between glioma segmentation and the ground truth.

The specificity represents the true-negative rate, which reflects the probability of complete negative detection. The specificity can be represented as follows:

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (12)$$

where TN and FP represent true-negative cases and false-positive cases, respectively. The higher the specificity, the smaller the difference between the segmentation and ground truth for the normal tissue.

5. Results

5.1. Comparison with Other Methods

We compared the proposed model with twelve advanced models to evaluate its advantages. The number of compared networks were 3, 3, 2, 1, 1, and 2 for 2024, 2023, 2022, 2020, 2019, and 2016, respectively. We conducted comparative experiments on the models from the perspectives of UNet, Transformer, and fusion of modality. The 3D UNet, V-Net, DMFNet, HDCNet, MVKS-Net, SPANet, ACMinet, and MSFRNet are architecture variants based on the basic UNet, while TransUNet, RFTNet, ETUNet, and mmformer are structures based on Transformer. Among them, ACMinet, MSFR-Net, RFTNet, and ETUNet represent the fusion of modality. By comparing with these common 3D UNet, network variants based on UNet, Transformer, Transformer + CNN, and fusion networks with multi-modality, we could prove the segmentation performance of our network. Among the above networks, variants based on UNet and CNN + Transformer networks based on Transformer are currently the mainstream research directions, and research on multi-modality utilization is relatively scarce. We used whole tumor (WT), tumor core (TC), enhancing tumor (ET), and average Dice as objective indicators to evaluate the performance of the proposed network, the results of which are presented in the form of tables and graphs.

Table 2, Figures 5 and 6 show that the Dice coefficients of SSGNet were 91.04, 86.64, and 81.11 with HDs of 4.62, 5.75, and 2.68 for the three tumor subregions (WT, TC, and ET), respectively. In comparison to the traditional 3D UNet, our model showed improvements of 2.51, 14.87, and 5.15 in terms of WT, TC, and ET metrics, respectively. Compared with the advanced MSFR-Net, which is the best UNet-based variant in Table 2, our model had improved Dice values of 0.14, 0.84, and 0.41 in WT, TC, and ET, respectively. We also compared the best results of the Transformer-based models (ETUNet) and found improvements in WT, TC, and ET by 1.04, 1.44, and 0.11, respectively, using our model. Finally, we compared our results with four multi-modality fusion models, with our results being the highest.

Table 2. Comparison of different methods on BraTS2018 validation dataset (best indicated in **bold**). Legend: whole tumor (WT), tumor core (TC), enhancing tumor (ER), average Dice (AVG), and Hausdorff95 (HD95).

Method	WT		TC		ET		Avg	
	Dice	HD ₉₅	Dice	HD ₉₅	Dice	HD ₉₅	Dice	HD ₉₅
3D U-Net [38] (2016)	88.53	17.10	71.77	11.62	75.96	6.04	78.75	11.59
V-Net [40] (2016)	89.60	6.54	81.00	7.82	76.60	7.21	82.40	7.19
DMFNet [41] (2019)	89.90	4.86	83.50	7.74	78.10	3.38	83.83	5.33
HDCNet [42] (2020)	88.50	7.89	84.80	7.09	76.60	7.21	83.30	7.40
TransUNet [49] (2022)	89.95	7.11	82.04	7.67	78.38	4.28	83.46	6.35
Mmformer [60] (2022)	89.56	4.43	83.33	8.04	78.75	3.27	83.88	5.25
MVKS-Net [43] (2023)	90.00	3.95	83.39	7.63	79.88	2.31	84.42	4.63
ACMINet [57] (2023)	90.41	-	84.96	-	81.52	-	85.63	-
MSFR-Net [53] (2023)	90.90	4.24	85.80	6.72	80.70	2.73	85.80	4.82
RFTNet [56] (2024)	90.30	5.97	82.15	6.41	80.24	3.16	84.23	5.18
ETUNet [51] (2024)	90.00	6.67	85.20	7.40	81.00	6.01	85.40	6.69
SPA-Net [44] (2024)	89.63	4.79	85.89	5.40	79.90	2.77	85.14	4.32
SSGNET (Ours)	91.04	4.62	86.64	5.75	81.11	2.68	86.26	4.35

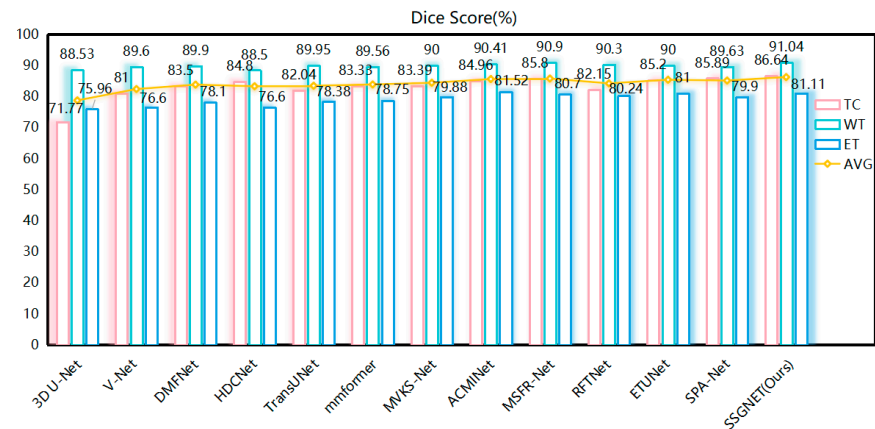


Figure 5. Comparison of the Dice results of the different segmentation methods.

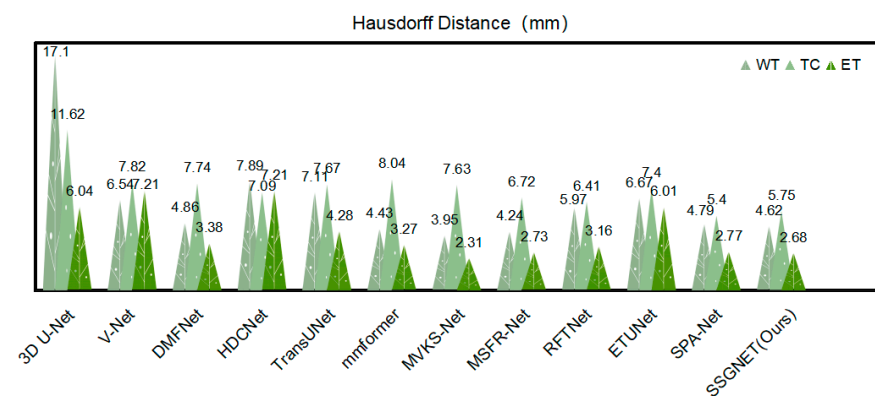


Figure 6. Comparison of HD results of different segmentation methods.

Regarding the results as a whole, the WT, TC, and Dice values were higher than the other models, indicating that the proposed model and modules provided significant improvements in segmentation tasks. Figure 7 shows the visualization results of the SSGNet model on the BraTS2021 dataset [61], in which five cases were randomly chosen. The results from left to right are T1, T1ce, T2, FLAIR, segmented results by SSGNet, and ground truth. Letters a-e present different visualized cases in Figure 7, which are results segmented by SSGNet. Green, yellow, and red represent the WT, TC, and ET, respectively. The results of

SSGNet are close to the labeled ground truth. The results suggest that our architecture and modules can provide a good basis for subsequent research.

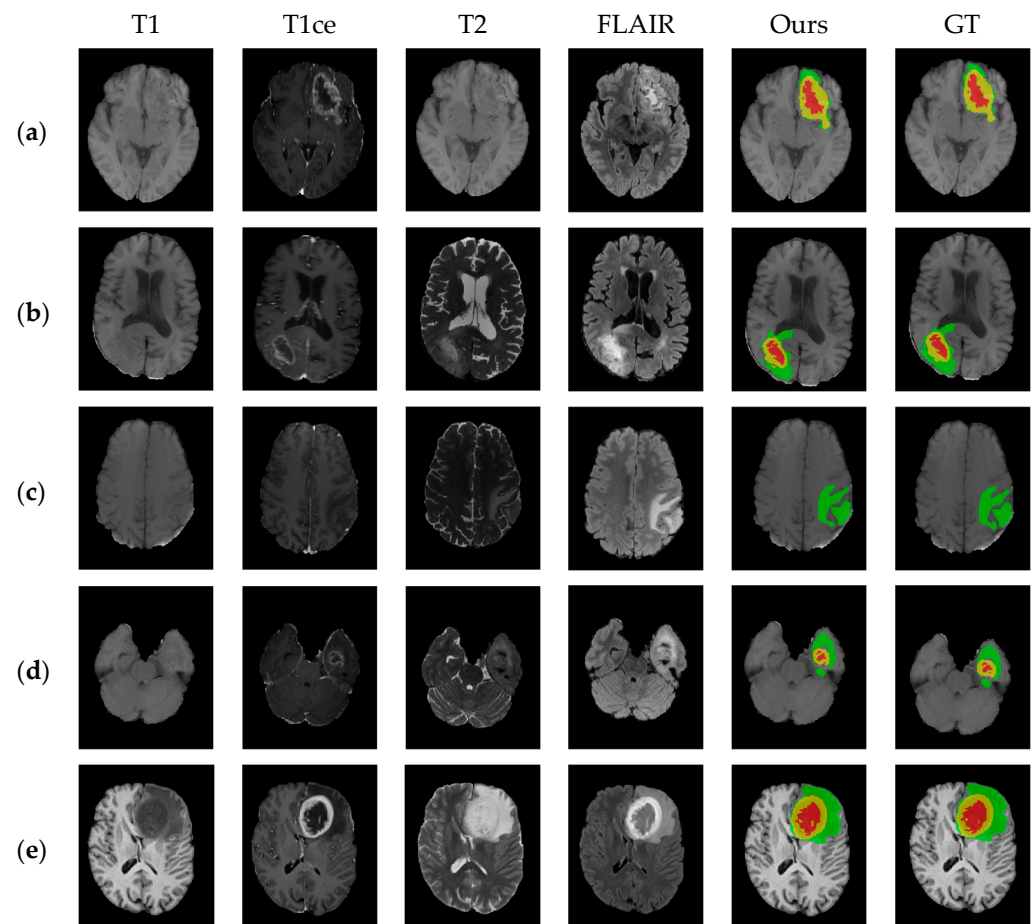


Figure 7. Visualization results of medical cases. (a–e) present different visualized cases, which are results segmented by SSGNet. Green, yellow, and red represent the WT, TC, and ET, respectively.

5.2. Ablation Experiments

5.2.1. Ablation Study of Each Module in SSGNet

We conducted ablation experiments to verify the effects of SMRF, SKSA, SCAM, and deep supervision (DP) in this architecture. Table 3 and Figure 8 show that the best results were obtained with the WT, TC, ET, and average Dice (Avg) values being 91.04, 86.64, 81.11, and 86.26, respectively, when all modules were used on the model. When the SMRF and SKSA were not all used in the networks (Experiment A), the WT, TC, ET, and average Dice values decreased by 1.75, 4.28, 2.14, and 2.72, respectively. When we removed the SCAM without deep supervision (Experiment B), the WT, TC, ET, and average Dice values decreased by 0.77, 2.58, 2.38, and 1.9, respectively. Experiment D with deep supervision garnered WT, TC, ET, and average Dice value decreases of 1.41, 1.28, 1.96, and 1.55, respectively. It can be seen that our developed module indeed played an important role in our structure. When we removed the deep supervision (Experiment C), the WT, TC, ET, and average Dice values became 90.11, 85.55, 80.37, and 0.92, respectively. By comparing Experiment B, Experiment D, and Experiment E, it can be seen that the proposed SCAM was of great help in improving the network results. Comparing the results of Experiment A and Experiment B, it can be seen that the developed SMRF and SKSA modules played important roles in the network. It can thus be stated that the module effectively improved the brain tumor segmentation performance.

Table 3. The results of the ablation study for each module in SSGNet, with the best performance highlighted in **bold** (WT: whole tumor, TC: tumor core, ET: enhancing tumor, Avg: average Dice score).

Experiment	SMRF	SKSA	SCAM	DP	Dice (%)			Avg
					WT	TC	ET	
A					89.29	82.36	78.97	83.54
B	✓	✓			90.27	84.06	78.73	84.36
C	✓	✓	✓		90.11	85.55	80.37	85.34
D	✓	✓		✓	89.63	85.36	79.15	84.71
E (SSGNet)	✓	✓	✓	✓	91.04	86.64	81.11	86.26

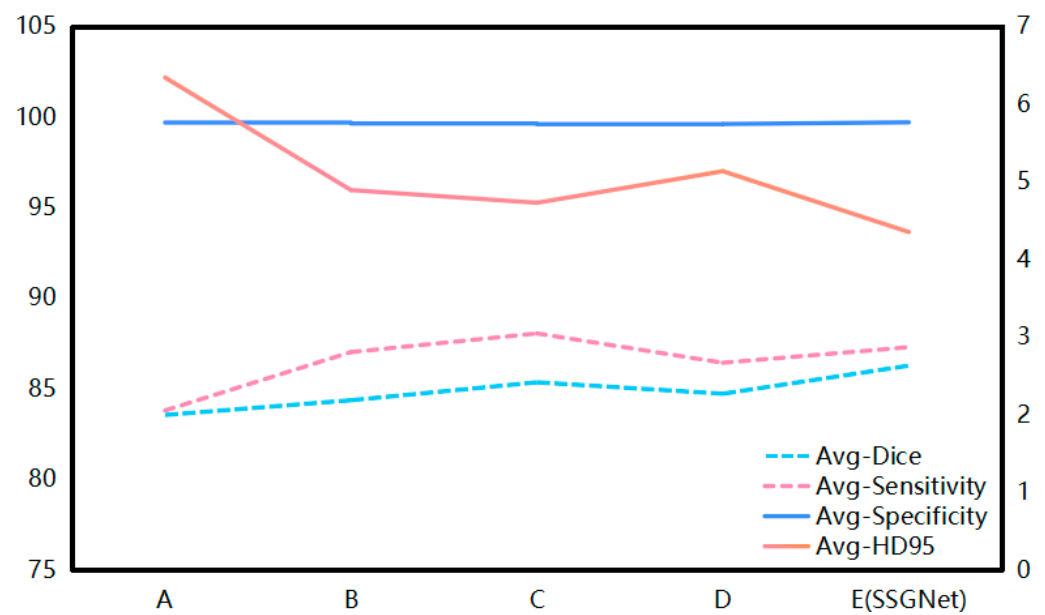


Figure 8. The results of the ablation study for each module in SSGNet.

5.2.2. Ablation Experiments for SMRF and SKSA Modules

Three studies were performed to evaluate the effect of SMRF and SKSA in different modality groups, consisting of SMRF and SKSA position exchange experiments and the same attention mechanism being applied in different modalities.

From the experimental results, which are shown in Table 4, Figures 9 and 10, it can be seen that, with the WT, TC, and ET being 91.04, 86.64, and 81.11, respectively, the result was the best when the SMRF module was applied to the T1 and T1ce modal groups and the SKSA module was applied to the T2 and FLAIR modal groups. Properly expanding the receptive field was more effective for retaining the expression of important information in the brain tumor regions in the T1 and T1ce modalities and capturing more helpful global feature relationships in the T2 and FLAIR modalities.

Table 4. The results of ablation experiments for the SMRF and SKSA modules, with the best performance highlighted in **bold** (WT: whole tumor, TC: tumor core, ET: enhancing tumor, Avg: average Dice score).

Experiment	Dice (%)			Hausdorff 95 (mm)		
	WT	TC	ET	WT	TC	ET
a (SSGNet)	91.04	86.64	81.11	4.62	5.75	2.68
b	90.02	85.12	80.67	5.37	7.51	2.80
c	91.20	85.19	79.37	4.50	7.50	2.66
d	90.32	85.06	79.82	5.21	7.54	2.61

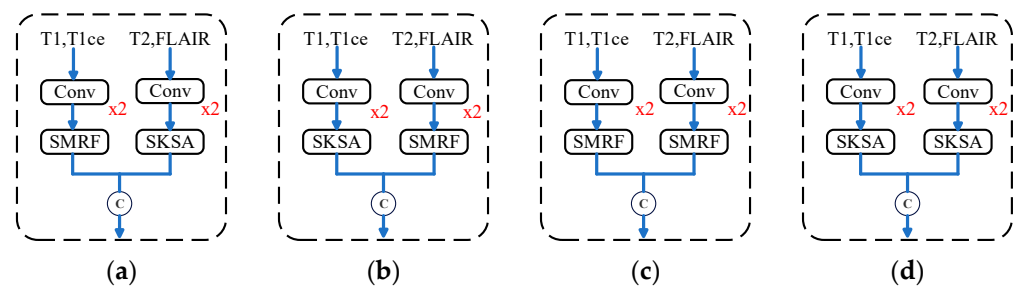


Figure 9. Diagram showing the SMRF and SKSA modules. (a) Original position; (b) SKSA module in T1, T1ce modality group and SMRF module in T2, FLAIR modality group; (c) SMRF module in T1, T1ce group and T2, FLAIR group, simultaneously; (d) SKSA module in T1, T1ce group and T2, FLAIR group, simultaneously.

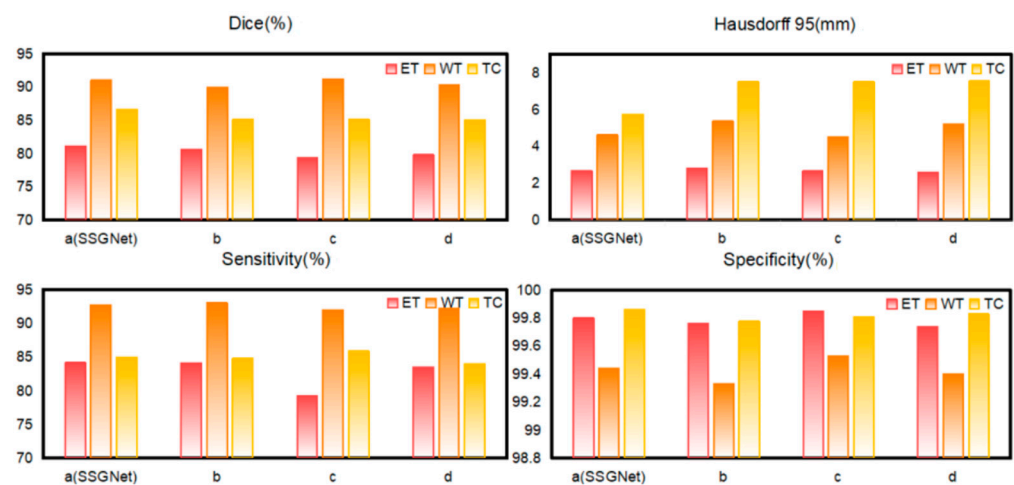


Figure 10. The results of ablation experiments for the SMRF and SKSA modules.

Further experiments were carried out with group-wise modalities. To verify the necessity of the grouping modality, we conducted experiments by directly connecting SMRF and SKSA in series or in parallel to the structure without the grouping modality.

The results are shown in Table 5, Figures 11 and 12. When SMRF and SKSA were connected in parallel (Experiment f), the Dice values of WT, TC, and ET were 88.89, 84.62, and 79.86, respectively. Similarly, when SMRF and SKSA were connected in series (Experiment g), the values of WT, TC, and ET were 88.20, 80.17, and 77.99, respectively. We conducted a total of five experiments to verify the effective positions and combinations of SMRF and SKSA. From the results, it can be concluded that the fixed current positions and combinations of SMRF and SKSA placed in both groups were the best in our network. The results indicate that effectively grouping the four modalities in the dataset was indeed helpful in improving the segmentation effect of brain tumors in our structure.

Table 5. The results of the ablation experiments for SMRF and SKSA in parallel or in series to the structure without the grouping modality (best indicated in **bold**). Legend: whole tumor (WT), tumor core (TC), enhancing tumor (ER), average Dice (AVG), and Hausdorff95 (HD95).

Experiment	Dice (%)			Hausdorff 95 (mm)		
	WT	TC	ET	WT	TC	ET
e (SSGNet)	91.04	86.64	81.11	4.62	5.75	2.68
f	88.89	84.62	79.86	14.91	10.10	3.48
g	88.20	80.17	77.99	8.34	10.79	3.77

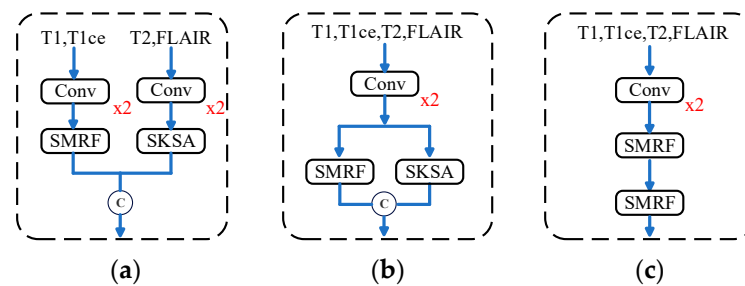


Figure 11. Diagram of SMRF and SKSA modules in series or parallel. (a) Original position. (b) SMRF and SKSA sequentially in parallel to four modalities. (c) SMRF in series to four modalities.

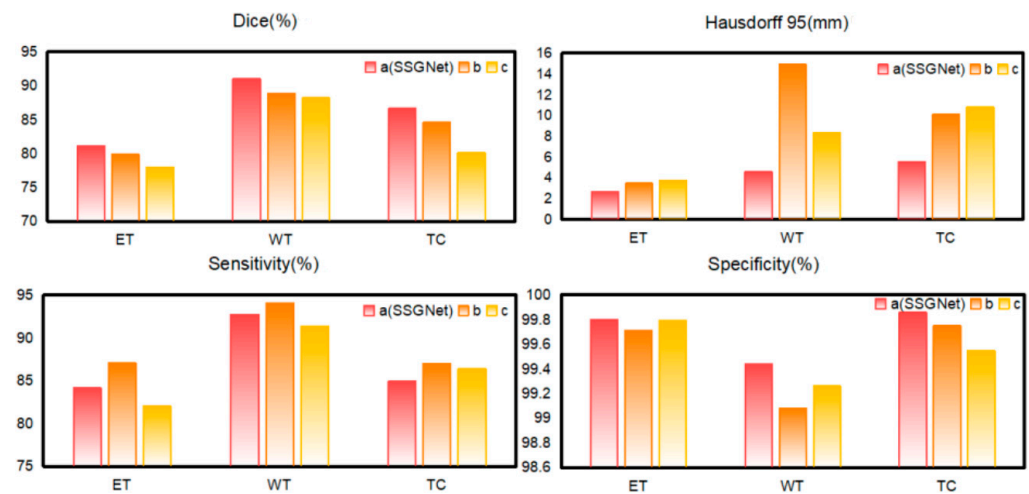


Figure 12. The results of the ablation experiments for SMRF and SKSA in series or in parallel to the structure without the grouping modality.

5.2.3. Ablation Experiments for SCAM Module

The significance of this research lies in the relationships between low-level semantic features and high-level semantic features. The SCAM is a module used in skip connection layers, but due to the different information provided by each layer, it varies depending on the depth of the layer. That is to say, the importance level of information transmitted by different encoder layers to their corresponding decoder layers is also different. To verify the above theory, we conducted ablation experiments using the SCAM. In this experiment, we applied the SCAM to the skip connection layers. Table 6 and Figure 13 show that when the third and fourth layers of SCAM were applied on the skip connection layers, the WT, TC, and ET were 91.04, 86.64, and 81.11, respectively, and the effect was the best. This suggests that the deeper the layers, the more high-level semantic information is needed to guide selection of the low-level features in our structure.

Table 6. Dice results of ablation experiments for the SCAM module, with the best performance highlighted in **bold** (WT: whole tumor, TC: tumor core, ET: enhancing tumor, Avg: average Dice score).

Experiment	SCAM				Dice (%)		
	Layer 1	Layer 2	Layer 3	Layer 4	WT	TC	ET
A					89.63	85.36	79.15
B	✓				90.86	85.49	79.33
C (SSGNet)	✓	✓			91.04	86.64	81.11
D	✓	✓	✓		90.89	86.10	80.41
E	✓	✓	✓	✓	90.88	85.00	80.50

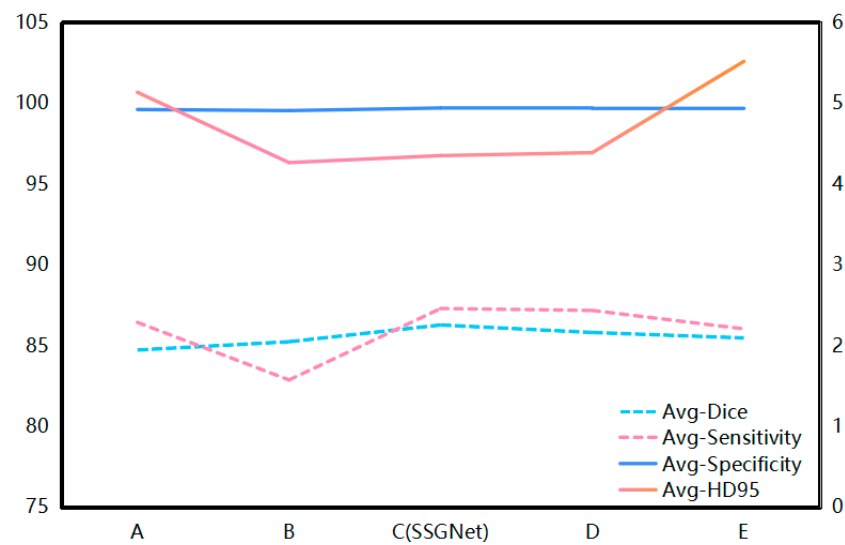


Figure 13. The impact results for the SCAM module.

5.3. Comparative Experiment SMRF, SKSA, and SCAM Module with RFACnv and CAM Module

To evaluate the effectiveness of our proposed SMRF, SKSA, and SCAM modules compared to existing modules, we replaced the SMRF and SKSA modules with RFACnv and the SCAM with CAM in the SSGNet network, respectively. The specific details of the experiment are shown in Table 7 and Figure 14. After both the SMRF and SKSA modules had been replaced with RFA, the WT, TC, ET, and average Dice values in SSGNet were 86.45, 84.57, 79.46, and 83.49, respectively. When only replacing SCAM with CAM, the WT, TC, ET, and average Dice values were 90.86, 86.44, 77.67, and 84.99, respectively. The above results were all lower than the results of our proposed network, which also demonstrates that the SMRF, SKSA, and SCAM modules are effective in our proposed network.

Table 7. Dice results of comparative experiments on the BraTS2018 dataset, with the best performance highlighted in **bold** (WT: whole tumor, TC: tumor core, ET: enhancing tumor, Avg: average Dice score).

Method	Dice(%)			Avg	Dataset
	WT	TC	ET		
(a): Replace SCAM with CAM in the proposed model	90.86	86.44	77.67	84.99	BraTS2018
(b): Replace SMRF and SKSA with RFACnv in the proposed model	86.45	84.57	79.46	83.49	
(c): Proposed model (SSGNet)	91.04	86.64	81.11	86.26	

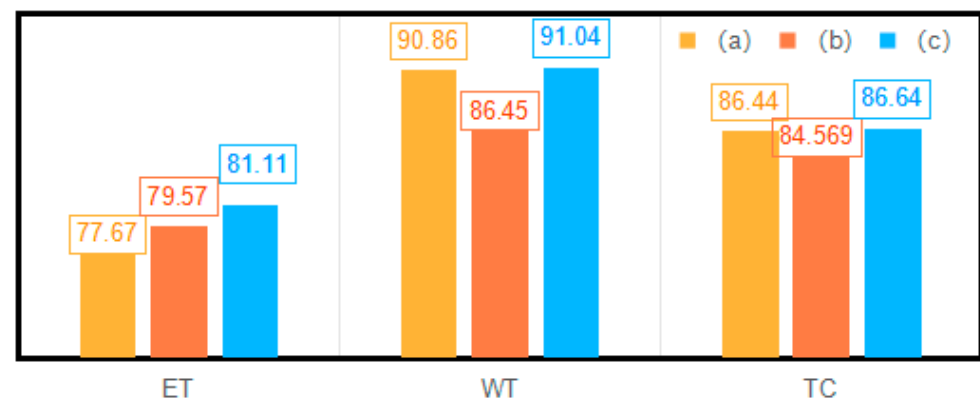


Figure 14. Dice results of comparative experiments on the BraTS2018 dataset.

6. Discussion

The multi-modality imaging of brain tumors provides structured information that can help segment brain tumors and achieve a reliable diagnosis. Tables 4 and 5 illustrate that, due to the differences between the different modalities, applying the same attention mechanism to all four modalities simultaneously means that their capacities cannot be fully exploited. However, adding different attention mechanisms to all four modalities will inevitably lead to a sharp increase in the number of parameters. Therefore, grouping different modalities according to the principles of medical imaging and adding different attention mechanisms within different groups would fully capitalize on the potential of the model; this would significantly improve the segmentation results. It is clear from experiments A and B, which are shown in Table 3, that the Dice results of the WT, TC, ET, and average Dice with and without an attention mechanism after grouping changed from 89.29, 82.36, 78.98, and 83.54 to 90.27, 84.06, 78.73, and 84.36, respectively.

In this study, to integrate the characteristics of the different modalities and improve the performance of the network, we proposed the SSGNet, which is based on group-wise modalities; these include an SMRF module, an SKSA attention mechanism, a SCAM, and deep supervision. We designed SMRF and SKSA to reduce the image size by half, further extract fused features within the modality group, and prevent the loss of information. The SMRF was used in the T1 and T1ce modality groups because the T1 sequence delivers an anatomical overview of the brain; meanwhile, the T1CE sequence enhances the highly vascularized and viable parts of the tumor. T1ce is the contrast-enhanced T1 and appears to be similar to T1 with regard to tumor regions, which are sensitive to necrosis and enhancing tumors. Owing to T2 and FLAIR sequences, which facilitate the evaluation of peritumoral edema and the non-contrast-enhancing parts of the tumor in gliomas, as well as the extent of the main tumor mass in non-contrast-enhancing low-grade gliomas [24–26,61], the SKSA module was leveraged in T2 and the FLAIR modalities group. The SMRF module involves group convolution with different kernel sizes, channel and spatial attention, and atrous convolution operation; meanwhile, the SKSA module contains group convolution with different kernel sizes, channel and spatial attention, and self-attention. Similarly, they all use group convolution with different kernel sizes and spatial and channel attention mechanisms, performing downsampling operations with a receptive field, and spatial and channel weights. One of the differences between these two modules is the atrous convolution operation and self-attention mechanism. The T1 and T1ce groups focus on necrosis and enhancing tumors, and do not require a global dependency relationship; only an appropriate receptive field is needed to achieve better performance. Meanwhile, due to their characteristics, T2 and FLAIR place greater emphasis on long-range dependencies between features. In Table 7, it can be seen that the SMRF and SKSA modules are superior to RFACnv, indicating that their use of multi-scale receptive fields and global receptive fields plays an important role. Tables 4 and 5 show that when the WT, TC, and ET were 91.04, 86.64, and 81.11, respectively, the optimum result was obtained when the SMRF module was applied to the T1 and T1ce modal groups and the SKSA module was applied to the T2 and FLAIR modal groups. This indicates that different attention mechanisms have an impact on different modal groups.

The fusion of the four modalities is also necessary in segmentation tasks for brain tumor images. This stage of the process is primarily completed by the skip connection layer, which is mainly responsible for transmitting information from the encoder layer to the corresponding decoder layer. We continuously used two convolutions on the skip connection layer to extract four modal features, but the outcome was not good, due to the influence of deep information. From Table 7, it can be seen that the CAM module lacked the fusion of information obtained from the four modalities in the group-wise network; therefore, we added an information fusion function on the basis of the CAM, thus improving the effect. Table 6 shows that when the third and fourth layers of the SCAM were applied to the skip connection layers, the WT, TC, and ET were 91.04, 86.64, and 81.11, respectively, with the best effect obtained. This suggests that the deeper the layers, the

more high-level semantic information is needed to guide the selection of the low-level features in our structure.

We also compared our model with UNet and its variants, Transformer, and Transformer-based networks in 2024, 2023, 2022, 2020, 2019, and 2016. We obtained Dice values of 91.04, 86.64, and 81.11, respectively, for the three tumor subregions of WT, TC, and ET; this is shown in Table 2. The results show that the proposed model achieved state-of-the-art performance compared with the twelve benchmarks. Compared with variant networks based on UNet and Transformer, the advantage of our network lies in its full utilization of the multi-modality characteristics of brain tumor MRI data. Moreover, unlike other multi-modality fusion models, our network could be built using the different attention mechanisms in different modality groups; this would give more relevance to the features, allow the network to obtain additional important information, and thus increase its performance.

However, our research also has some limitations. The extensive use of attention mechanisms will inevitably increase the complexity of the model, while improving its effectiveness. Therefore, we aim to perform lightweight research on models in the future.

Although our network exhibited a better performance when tested on the BraTS2018 dataset for brain tumor segmentation tasks, we did not attempt to evaluate this network using other datasets, especially those that are not relevant to brain tumors. In the field of medical imaging, some images, such as X-ray, do not contain four modalities; this would lead to the inability to use our model. Additionally, the medical image model used for real clinical diagnosis requires a larger number of experimentations, but our current experiments were far from sufficient. In the next stage, we aim to enhance the practical applicability of our model by testing it on more medical image datasets and conducting extensive experiments with real medical images, in order to obtain the best results.

7. Conclusions

In this paper, we proposed an improved model of SSGNet combined with SMRF, SKSA, and SCAM modules. Our results demonstrated that a group-wise modality and our proposed modules played an important role in SSGNet compared with more than twelve benchmarks. We also conducted ablation experiments on the SMRF, SKSA, and SCAM modules, which demonstrated the effectiveness of our modules. We believe that the encouraging results obtained with SSGNet will inspire further research into brain tumor segmentation.

Author Contributions: Conceptualization, N.C. and B.G.; Methodology, B.G.; Software, P.Y.; Data Curation, R.Z.; Writing—Original Draft, B.G.; Writing—Review and Editing, N.C. and B.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 41830110.

Data Availability Statement: Publicly available datasets were analyzed in this study. The dataset can be found in the BraTS 2018 dataset: <https://www.med.upenn.edu/sbia/brats2018/data.html> (accessed on 11 May 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. McFaline-Figueroa, J.R.; Lee, E.Q. Brain tumors. *Am. J. Med.* **2018**, *131*, 874–882. [CrossRef]
2. DeAngelis, L.M. Brain tumors. *N. Engl. J. Med.* **2001**, *344*, 114–123. [CrossRef] [PubMed]
3. Castro, M.G.; Cowen, R.; Williamson, I.K.; David, A.; Jimenez-Dalmaroni, M.J.; Yuan, X.; Bigliari, A.; Williams, J.C.; Hu, J.; Lowenstein, P.R. Current and future strategies for the treatment of malignant brain tumors. *Pharmacol. Ther.* **2003**, *98*, 71–108. [CrossRef] [PubMed]
4. Dandy, W.E. Intracranial pressure without brain tumor: Diagnosis and treatment. *Ann. Surg.* **1937**, *106*, 492–513. [CrossRef] [PubMed]

5. Alther, B.; Mylius, V.; Weller, M.; Gantenbein, A. From first symptoms to diagnosis: Initial clinical presentation of primary brain tumors. *Clin. Transl. Neurosci.* **2020**, *4*, 2514183X20968368. [\[CrossRef\]](#)
6. Alentorn, A.; Hoang-Xuan, K.; Mikkelsen, T. Presenting signs and symptoms in brain tumors. *Handb. Clin. Neurol.* **2016**, *134*, 19–26.
7. Brandsma, D.; Stalpers, L.; Taal, W.; Sminia, P.; van den Bent, M.J. Clinical features, mechanisms, and management of pseudoprogression in malignant gliomas. *Lancet Oncol.* **2008**, *9*, 453–461. [\[CrossRef\]](#)
8. Omuro, A.; DeAngelis, L.M. Glioblastoma and other malignant gliomas: A clinical review. *JAMA* **2013**, *310*, 1842–1850. [\[CrossRef\]](#)
9. Barnholtz-Sloan, J.S.; Ostrom, Q.T.; Cote, D. Epidemiology of brain tumors. *Neurol. Clin.* **2018**, *36*, 395–419. [\[CrossRef\]](#)
10. Inoue, T.; Ogasawara, K.; Beppu, T.; Ogawa, A.; Kabasawa, H. Diffusion tensor imaging for preoperative evaluation of tumor grade in gliomas. *Clin. Neurol. Neurosurg.* **2005**, *107*, 174–180. [\[CrossRef\]](#)
11. Bauer, S.; Wiest, R.; Nolte, L.-P.; Reyes, M. A survey of MRI-based medical image analysis for brain tumor studies. *Phys. Med. Biol.* **2013**, *58*, R97–R129. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Thoman, W.J.; Ammirati, M.; Caragine, L.P., Jr.; McGregor, J.M.; Sarkar, A.; Chiocca, E.A. Brain tumor imaging and surgical management: The neurosurgeon's perspective. *Top. Magn. Reson. Imaging* **2006**, *17*, 121–126. [\[CrossRef\]](#) [\[PubMed\]](#)
13. McAfee, J.G.; Taxdal, D.R. Comparison of radioisotope scanning with cerebral angiography and air studies in brain tumor localization. *Radiology* **1961**, *77*, 207–222. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Schillaci, O.; Filippi, L.; Manni, C.; Santoni, R. Single-photon emission computed tomography/computed tomography in brain tumors. *Semin. Nucl. Med.* **2007**, *37*, 34–47. [\[CrossRef\]](#)
15. Tonarelli, L. Magnetic Resonance Imaging of Brain Tumor. 2013. Available online: <https://cewebsource.com/lander> (accessed on 11 May 2024).
16. Auer, L.M.; Van Velthoven, V. Practical Handling of the US Probe During Investigation. In *Intraoperative Ultrasound Imaging in Neurosurgery: Comparison with CT and MRI*; Springer: Berlin/Heidelberg, Germany, 1990; pp. 10–21.
17. Abd-Ellah, M.K.; Awad, A.I.; Khalaf, A.A.; Hamed, H.F. A review on brain tumor diagnosis from MRI images: Practical implications, key achievements, and lessons learned. *Magn. Reson. Imaging* **2019**, *61*, 300–318. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Sander, L.M.; Deisboeck, T.S. Growth patterns of microscopic brain tumors. *Phys. Rev. E* **2002**, *66*, 66–73. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Pereira, S.; Pinto, A.; Alves, V.; Silva, C.A. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans. Med. Imaging* **2016**, *35*, 1240–1251. [\[CrossRef\]](#)
20. Alam, M.T.; Nawal, N.; Nishi, N.J.; Sahan, M.; Islam, M.T. Automatic Brain Tumor Segmentation Using U-ResUNet Chain Model Approach. Ph.D. Thesis, Brac University, Dhaka, Bangladesh, 2021.
21. Blanc, D. Artificial Intelligence Methods for Object Recognition: Applications in Biomedical Imaging. Ph.D. Thesis, Université de Montpellier, Montpellier, France, 2022.
22. Xue, Y. Deep Generative Models for Medical Images and Beyond. Ph.D. Thesis, Pennsylvania State University, Pennsylvania, PA, USA, 2021.
23. Weninger, L.; Rippel, O.; Koppers, S.; Merhof, D. Segmentation of brain tumors and patient survival prediction: Methods for the brats 2018 challenge. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Proceedings of the 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Spain, Granada, 16 September 2019*; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 3–12.
24. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* **2014**, *34*, 1993–2024. [\[CrossRef\]](#)
25. Bakas, S.; Akbari, H.; Sotiras, A.; Bilello, M.; Rozycki, M.; Kirby, J.S.; Freymann, J.B.; Farahani, K.; Davatzikos, C. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* **2017**, *4*, 170117. [\[CrossRef\]](#)
26. Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R.T.; Berger, C.; Ha, S.M.; Rozycki, M. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv* **2018**, arXiv:1811.02629.
27. Ginat, D.T.; Meyers, S.P. Intracranial lesions with high signal intensity on T1-weighted MR images: Differential diagnosis. *Radiographics* **2012**, *32*, 499–516. [\[CrossRef\]](#)
28. Zacharaki, E.I.; Wang, S.; Chawla, S.; Soo Yoo, D.; Wolf, R.; Melhem, E.R.; Davatzikos, C. Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. *Magn. Reson. Med. Off. J. Int. Soc. Magn. Reson. Med.* **2009**, *62*, 1609–1618. [\[CrossRef\]](#)
29. Whittall, K.P.; Mackay, A.L.; Graeb, D.A.; Nugent, R.A.; Li, D.K.; Paty, D.W. In vivo measurement of T2 distributions and water contents in normal human brain. *Magn. Reson. Med.* **1997**, *37*, 34–43. [\[CrossRef\]](#)
30. Roozpeykar, S.; Azizian, M.; Zamani, Z.; Farzan, M.R.; Veshnavei, H.A.; Tavoosi, N.; Toghyani, A.; Sadeghian, A.; Afzali, M. Contrast-enhanced weighted-T1 and FLAIR sequences in MRI of meningeal lesions. *Am. J. Nucl. Med. Mol. Imaging* **2022**, *12*, 63–70.
31. Liu, C.; Li, W.; Tong, K.A.; Yeom, K.W.; Kuzminski, S. Susceptibility-weighted imaging and quantitative susceptibility mapping in the brain. *J. Magn. Reson. Imaging* **2015**, *42*, 23–41. [\[CrossRef\]](#)
32. Shi, T.; Jiang, H.; Zheng, B. C 2 MA-Net: Cross-modal cross-attention network for acute ischemic stroke lesion segmentation based on CT perfusion scans. *IEEE Trans. Biomed. Eng.* **2021**, *69*, 108–118. [\[CrossRef\]](#)

33. Kalyan, K.S.; Rajasekharan, A.; Sangeetha, S. Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv* **2021**, arXiv:2108.05542.
34. Iqbal, S.; Ghani Khan, M.U.; Saba, T.; Mehmood, Z.; Javaid, N.; Rehman, A.; Abbasi, R. Deep learning model integrating features and novel classifiers fusion for brain tumor segmentation. *Microsc. Res. Tech.* **2019**, *82*, 1302–1315. [[CrossRef](#)]
35. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
36. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
37. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1055–1059.
38. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, 17–21 October 2016; pp. 424–432.
39. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the 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, 20 September 2018; pp. 3–11.
40. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
41. Chen, C.; Liu, X.; Ding, M.; Zheng, J.; Li, J. 3D dilated multi-fiber network for real-time brain tumor segmentation in MRI. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, 13–17 October 2019; pp. 184–192.
42. Luo, Z.; Jia, Z.; Yuan, Z.; Peng, J. HDC-Net: Hierarchical decoupled convolution network for brain tumor segmentation. *IEEE J. Biomed. Health Inform.* **2020**, *25*, 737–745. [[CrossRef](#)] [[PubMed](#)]
43. Guan, X.; Zhao, Y.; Nyatega, C.O.; Li, Q. Brain tumor segmentation network with multi-view ensemble discrimination and kernel-sharing dilated convolution. *Brain Sci.* **2023**, *13*, 650. [[CrossRef](#)] [[PubMed](#)]
44. Liu, H.; Huang, J.; Li, Q.; Guan, X.; Tseng, M. A deep convolutional neural network for the automatic segmentation of glioblastoma brain tumor: Joint spatial pyramid module and attention mechanism network. *Artif. Intell. Med.* **2024**, *148*, 102776. [[CrossRef](#)]
45. Xu, W.; Wang, J.; Wang, Y.; Xu, G.; Lin, D.; Dai, W.; Wu, Y. Where is the model looking at?–Concentrate and explain the network attention. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 506–516. [[CrossRef](#)]
46. Liu, D.; Sheng, N.; He, T.; Wang, W.; Zhang, J.; Zhang, J. SGEResU-Net for brain tumor segmentation. *Math. Biosci. Eng.* **2022**, *19*, 5576–5590. [[CrossRef](#)]
47. Tian, W.; Li, D.; Lv, M.; Huang, P. Axial attention convolutional neural network for brain tumor segmentation with multi-modality MRI scans. *Brain Sci.* **2022**, *13*, 12. [[CrossRef](#)]
48. Yang, L.; Zhai, C.; Liu, Y.; Yu, H. CFHA-Net: A polyp segmentation method with cross-scale fusion strategy and hybrid attention. *Comput. Biol. Med.* **2023**, *164*, 107301. [[CrossRef](#)]
49. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
50. Lu, Y.; Chang, Y.; Zheng, Z.; Sun, Y.; Zhao, M.; Yu, B.; Tian, C.; Zhang, Y. GMetaNet: Multi-scale ghost convolutional neural network with auxiliary MetaFormer decoding path for brain tumor segmentation. *Biomed. Signal Process. Control* **2023**, *83*, 104694. [[CrossRef](#)]
51. Zhang, W.; Chen, S.; Ma, Y.; Liu, Y.; Cao, X. ETUNet: Exploring efficient transformer enhanced UNet for 3D brain tumor segmentation. *Comput. Biol. Med.* **2024**, *171*, 108005. [[CrossRef](#)]
52. Wang, Y.; Zhang, Y.; Hou, F.; Liu, Y.; Tian, J.; Zhong, C.; Zhang, Y.; He, Z. Modality-pairing learning for brain tumor segmentation. In Proceedings of the Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, 4 October 2021; pp. 230–240.
53. Li, X.; Jiang, Y.; Li, M.; Zhang, J.; Yin, S.; Luo, H. MSFR-Net: Multi-modality and single-modality feature recalibration network for brain tumor segmentation. *Med. Phys.* **2023**, *50*, 2249–2262. [[CrossRef](#)]
54. Zhou, T.; Ruan, S.; Vera, P.; Canu, S. A Tri-Attention fusion guided multi-modal segmentation network. *Pattern Recognit.* **2022**, *124*, 108417. [[CrossRef](#)]
55. Lin, J.; Lin, J.; Lu, C.; Chen, H.; Lin, H.; Zhao, B.; Shi, Z.; Qiu, B.; Pan, X.; Xu, Z. CKD-TransBTS: Clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation. *IEEE Trans. Med. Imaging* **2023**, *42*, 2451–2461. [[CrossRef](#)] [[PubMed](#)]
56. Jiao, C.; Yang, T.; Yan, Y.; Yang, A. RFTNet: Region–Attention Fusion Network Combined with Dual-Branch Vision Transformer for Multimodal Brain Tumor Image Segmentation. *Electronics* **2023**, *13*, 77. [[CrossRef](#)]
57. Zhuang, Y.; Liu, H.; Song, E.; Hung, C.C. A 3D cross-modality feature interaction network with volumetric feature alignment for brain tumor and tissue segmentation. *IEEE J. Biomed. Health Inform.* **2022**, *27*, 75–86. [[CrossRef](#)] [[PubMed](#)]

58. Zhang, X.; Liu, C.; Yang, D.; Song, T.; Ye, Y.; Li, K.; Song, Y. Rfaconv: Innovating spatital attention and standard convolutional operation. *arXiv* **2023**, arXiv:2304.03198.
59. Huang, G.; Zhu, J.; Li, J.; Wang, Z.; Cheng, L.; Liu, L.; Li, H.; Zhou, J. Channel-attention U-Net: Channel attention mechanism for semantic segmentation of esophagus and esophageal cancer. *IEEE Access*. **2020**, *8*, 122798–122810. [[CrossRef](#)]
60. Zhang, Y.; He, N.; Yang, J.; Li, Y.; Wei, D.; Huang, Y.; Zhang, Y.; He, Z.; Zheng, Y. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Singapore, 18–22 September 2022; Springer Nature: Cham, Switzerland, 2022; pp. 107–117.
61. Baid, U.; Ghodasara, S.; Mohan, S.; Bilello, M.; Calabrese, E.; Colak, E.; Farahani, K.; Kalpathy-Cramer, J.; Kitamura, F.C.; Pati, S. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv* **2021**, arXiv:2107.02314.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.