New Approach for Brain Tumor Segmentation Based on Gabor Convolution and Attention Mechanism

Yuan Cao and Yinglei Song *

School of Automation, Jiangsu University of Science and Technology, Zhenjiang 212100, China; 221110303105@stu.just.edu.cn
* Correspondence: syinglei2013@163.com

Abstract: In the treatment process of brain tumors, it is of great importance to develop a set of MRI image segmentation methods with high accuracy and low cost. In order to extract the feature information for each region of the brain tumor more effectively, this paper proposes a new model Ga-U-Net based on Gabor convolution and an attention mechanism. Based on 3D U-Net, Gabor convolution is added at the shallow layer of the encoder, which is able to learn the local structure and texture information of the tumor better. After that, the CBAM attention mechanism is added after the output of each layer of the encoder, which not only enhances the network’s ability to perceive the brain tumor boundary information but also reduces some redundant information by allocating the attention to the two dimensions of space and channel. Experimental results show that the model performs well for multiple tumor regions (WT, TC, ET) on the brain tumor dataset BraTS 2021, with Dice coefficients of 0.910, 0.897, and 0.856, respectively, which are improved by 0.3%, 2%, and 1.7% compared to the base network, the U-Net network, with an average Dice of 0.887 and an average Hausdorff distance of 9.12, all of which are better than a few other state-of-the-art deep models for biomedical image segmentation.

Keywords: brain tumor; Gabor convolution; CBAM attention mechanism; medical image segmentation

1. Introduction

With the rapid development of medical imaging technology, brain tumor image segmentation has made remarkable progress as an important medical image processing task. So far, two major types of brain tumors have been identified, including the primary and metastatic tumors based on whether the brain tumor cells are generated autonomously or migrate to the brain for generation [1]. Primary brain tumors originate in the brain tissue, while metastatic brain tumors are metastases of tumors from other sites. By histological type, they can be classified as glioma (G), meningioma (M), and pituitary tumor (P) [2], of which glioma of the brain is the most commonly seen type of primary tumor with a high mortality rate. Therefore, in order to treat these patients and extend their life expectancy, doctors usually need to analyze and develop accurate and reasonable treatment plans based on their brain images, and the most commonly used technique is magnetic resonance imaging (MRI) [3]. MRI can provide four different modalities, which are complementary to each other, and the combination of which can generate more detailed and comprehensive information about the structure and function of the brain. These four modalities complement each other and can be used in combination to provide more detailed and comprehensive information about brain structure and lesions, and such information can help doctors make more accurate diagnoses and develop more effective treatment plans [4].

MRI images are of significant value in diagnosing early brain tumor conditions. However, at present, in clinical practice, the process mainly relies on the experience of radiologists to determine the category of the tumor and manually mark the location of the brain tumor. This method not only wastes a large amount of human and material resources but
also may lead to misjudgments and omissions due to the subjective judgment of experts. Therefore, there is an urgent need for the availability of computer programs that can assist doctors in the classification and segmentation labeling of brain tumors. Such computer programs will help doctors focus more on developing individualized treatment plans and reduce their work pressure.

When facing these tumors with irregular shapes and complex boundaries, manual segmentation is both a waste of time and prone to errors; scientists have thus been developing methods that can automatically segment brain tumors with high accuracy in recent years. Traditional methods include segmentation methods based on thresholds, boundaries, and regions [5], but the accuracy of these methods is not very high. In recent years, with the rapid development in the field of artificial intelligence, the integration of artificial intelligence into applications in the various fields has become increasingly important [6–8]. The application of automated computer-aided technology to the detection of brain tumor patients not only improves the efficiency of diagnosis but also reduces the work pressure of doctors. In brain tumor segmentation, compared with the inefficiency and high error of traditional methods, deep learning-based techniques have significantly improved the recognition rates of brain tumors and the segmentation accuracy of their tissues by automatically extracting features from MRI images. However, the development of highly accurate segmentation algorithms remains an important challenge to further improve the accuracy and robustness of brain tumor diagnosis.

Long et al. [8] proposed the FCN architecture in 2015; different from the traditional CNN architecture, it replaces the fully connected structure with convolutional layers and adopts an up-sampling technology, which can obtain a segmented image with the same size as the original image. Compared with the traditional convolutional network, FCN effectively reduces the computation time and improves the segmentation accuracy. Therefore, FCN has become a pioneering work for the field of semantic segmentation in deep learning. Many researchers have used it as the basic framework for deeper research. For example, Shen et al. [9] designed a symmetric differential image-based FCN network for improved segmentation accuracy, which is characterized by three up-sampling structures for feature extraction. Subsequently, Ronneberger et al. [10] proposed the U-Net architecture based on the FCN network; the U-Net utilizes a training strategy that combines the features of the encoding path and those of the decoding path through long connections to reduce information loss. In addition, the U-Net network uses methods such as data augmentation and discard regularization during training to improve the generalization ability and robustness of the network. It has been verified that U-Net has achieved excellent success in the field of medical images and also has a great impact on brain tumor medical segmentation. For example, Liu et al. [11] introduced the cavity convolution structure into the U-Net architecture for improved segmentation accuracy; the cavity convolution can control the receptive field without changing the size of the feature map, and the multi-scale information can thus be accurately extracted.

Ai et al. [12] added the attention mechanism into the U-Net architecture for brain tumor segmentation, the improved performance of this algorithm can be ascribed to a dense residual block utilized to replace the double convolution structure and the attention mechanism incorporated into decoding. Wen et al. [13] also added the attention mechanism into the U-Net model; the difference is that it incorporates the attention module into the jump connectivity path of U-Net, which consists of a spatial module and an attention module. The attention model is capable of modeling both spatial and channel attention. It leads to higher segmentation accuracy with reduced memory cost. Maji D et al. [14], on the other hand, incorporated the attention mechanism into the structure of the Res-U-Net [15] architecture and added a bootstrap decoder. As a result, more accurate features can be obtained and the segmentation accuracy is thus improved. The 3D Convolutional Block Attention Module (3D CBAM) is an advanced attention mechanism designed to enhance the performance of 3D convolutional neural networks. Wang et al. [16] illustrated the effectiveness of 3D CBAM in brain tumor segmentation by incorporating it into a
U-Net framework, achieving significant improvements in accuracy due to enhanced feature extraction from multimodal MRI images. Zhou [17] et al. proposed a new model for brain tumor segmentation. One of the most important features of the model is that the segmentation is performed in multiple stages. Firstly, an initial network is used to generate a region that serves to constrain the tumor site that is about to be segmented. Immediately after that, a 3D U-Net network that also contains an attention mechanism is used to perform a multiregional segmentation of the constrained region. Finally, in order to solve the multi-segmentation problem in a better way, a new loss function is developed to refine the network. The work by Jiang [18] et al. also incorporates multiple stages into the network for segmentation. A two-stage cascaded U-Net is designed for brain tumor segmentation, where rough results obtained in the first stage are refined in the second stage to obtain segmentation results with improved accuracy.

Although many of the above methods have achieved excellent results for brain tumor segmentation, their performance on some tumor images remains unsatisfactory. In practice, accurate segmentation of some tumor images is difficult due to the various distributions of gray values and complex local textures in these images. In this paper, a new method is proposed for brain tumor segmentation. The proposed method performs segmentation of multimodal brain tumor MRI images with a new deep learning model. Specifically, based on the 3D U-Net architecture, the Gabor convolution is incorporated into the encoder. The input image is processed with Gabor filtering and the obtained information is combined with the original convolution result to extract more accurate features for brain tumors, as well as edema regions and other related details, especially fine tumors. In addition, the loss of fine local features in an image due to multiple convolutions can be substantially reduced. The 3D CBAM attention mechanism and the subsequent down-sampling operations are finally utilized to generate segmentation results. Experimental results show that the proposed method can achieve improved segmentation accuracy and outperform a few state-of-the-art methods for the segmentation of tumor images. The major contributions of the paper can be summarized as follows.

1. A new deep learning model that incorporates both the Gabor convolution and 3D CBAM attention mechanism into the 3D U-Net architecture is proposed for the segmentation of MRI images of brain tumors.
2. Experiments have been performed to evaluate the performance of the proposed approach. Its segmentation accuracy is compared with that of several other state-of-the-art methods for brain tumor segmentation.

2. Materials and Methods

When performing the task of brain tumor segmentation on 3D MRI images, it is usually necessary to slice the images, which may lead to the waste of some useful slice information. Although existing segmentation algorithms for multi-feature fusion are able to utilize the 3D data of medical images to a certain extent, these algorithms usually compensate for the lost information by fusing the results of the segmented results instead of actively utilizing the slicing information of the 3D data of the medical images in the segmentation process. Therefore, in order to utilize the contextual information more effectively, a 3D multimodal MRI brain tumor image segmentation network called Ga-U-Net is proposed in this study. Its network structure is shown in Figure 1. The network as a whole adopts the 3D U-Net encoder-decoder architecture.

Firstly, in order to improve the accuracy of the features obtained from brain tumors and edema regions, especially the information regarding fine tumors, a Gabor convolution kernel is added to the original encoder part to form the GA module of the proposed architecture, which firstly performs a convolution operation on the input image, and since the Gabor convolution kernel, as a kind of visual sensory field, possesses a superior ability of spatial and frequency feature extraction, it is expected to effectively identify and segment regions such as brain tumors and edema through its sensitivity analysis of texture so as to enhance the feature extraction effect. After that, its convolved result is added with
of spatial and frequency feature extraction, it is expected to ... Gabor convolution is shown in green color and the other two ordinary convolutional components are shown in blue color.

while the deep encoder can deal with the higher-level and abstract information of the brain tumor and its semantic features. Therefore, in order to improve the effectiveness of filtering.

The overall structure of the proposed Ga-U-Net network, where the meanings of modules and arrows in different colors are shown in the bottom of the figure.

2.1. The GA Module

Gabor filters can effectively extract local features from images. They are based on Gabor functions and are a multi-scale, multi-directional filter that can capture local features such as textures, edges, and details in images. During the training process of the network, the shallow encoder in the coding part usually deals with the lower level information and mainly focuses on the low-level features of the image, such as edges, colors, textures, etc., while the deep encoder can deal with the higher-level and abstract information of the brain tumor image, which pays more attention to the overall structure of a brain tumor and its semantic features. Therefore, in order to improve the effectiveness of the extracted low-level features for segmentation, additional feature information obtained with Gabor filtering is incorporated into the original structure of the 3D U-Net model. The GA module shown in Figure 2 is utilized for the operations associated with the Gabor filtering.

The GA module is specifically designed as follows. The input original image is subjected to the Gabor convolution and two ordinary convolutions, and then their final structures are summed up. Since the sensory field of the ordinary convolution is limited by the size of the convolution kernel, the ability of extracting the features is limited, while the Gabor
convolution kernel consists of sinusoidal functions, with the characteristics of multi-scale and multi-direction, which makes it able to capture the image features more effectively in different frequencies and directions, and the ability to capture features is good both in the low-level encoder and high-level encoder parts; especially in the low-level encoder, the Gabor convolution decomposes the image in the frequency domain, which can better capture the local structure and texture information of the image. The feature summation of these two convolutions can enable the network to learn more comprehensive information about the tumor image. The encoder is responsible for extracting high-level abstract features from the input image, while the decoder maps these features back to the original image space to generate semantic segmentation results. The combination of encoder and decoder enables the network to simultaneously consider both global and detailed information, thus achieving good performance in image segmentation tasks. Compared with feature concatenation, the feature summation network has fewer parameters and consumes less computational resources.

2.2. Gabor Filters

As early as 1946, D. Gabor [19] proposed a one-dimensional Gabor function in his published paper. In 1980, Daugman [20] proposed a two-dimensional form of the Gabor function. The design of this filter is inspired by the Gabor wavelet in biology, which is based on the working principle of the human visual system and simulates the response properties of simple cells in the visual cortex. The results of researchers such as Daugman [21] and Porat [22] showed that biological cells have different perceptions of stimuli of different frequencies during visual processing, and their response patterns can be described by a two-dimensional Gabor filter. This filter presents excellent selectivity in the frequency domain and is able to capture features of different frequencies in the image. Gabor convolution is a technique that integrates the interpretability of Gabor filters with the powerful learning capabilities of convolutional neural networks (CNNs) [23].

The Gabor filter is a linear filter mainly used for image edge detection. Its design is inspired by the visual response of simple cells, and it can effectively mimic the properties of the human visual system and can perform feature extraction from different scales and directions. As a result, Gabor filters excel in capturing texture features, the Gabor filter can recognize useful features in the sample image to texture features, the Gabor filter can also recognize useful features in the sample image and is sensitive to changes in image brightness, contrast, etc., thus giving the model greater robustness. The Gabor function consists of a Gaussian function and a cosine function, whose expression is in complex form and consists of a real part and an imaginary part, thus giving the model greater robustness. The Gabor function consists of a Gaussian function and a cosine function, whose expression is in complex form and consists of a real part and an imaginary part, which are orthogonal to each other. Among them, the filtering of the real part can be used to smooth the image, while the filtering of the imaginary part is used for edge detection. Therefore, based on their excellent performance, Gabor filters are widely used in many fields such as texture analysis, image classification, and action recognition. The 3D Gabor filter can be described by Equations (1)–(3), where \( x', y', z' \) are coordinates adjusted relative to the center position in space, calculated with Equations (4)–(6), where \( \theta, \phi, \sigma, \gamma, \lambda \) are the parameters of the filter, and the specific explanation of each parameter is shown in Table 1. The relationship between \( b \) and \( \sigma / \lambda \) is given by Equations (7) and (8).

\[
G(x, y, z; \theta, \phi, \sigma, \gamma, \lambda) = G_{re}(x, y, z; \theta, \phi, \sigma, \gamma, \lambda) + jG_{im}(x, y, z; \theta, \phi, \sigma, \gamma, \lambda)
\]  
\[
G_{re}(x, y, z; \theta, \phi, \sigma, \gamma, \lambda) = \exp\left(-x'^2 + \gamma^2 y'^2 + \gamma^2 z'^2\right) \cos\left(2\pi \frac{x'}{\lambda} + \phi\right)
\]  
\[
G_{im}(x, y, z; \theta, \phi, \sigma, \gamma, \lambda) = \exp\left(-x'^2 + \gamma^2 y'^2 + \gamma^2 z'^2\right) \sin\left(2\pi \frac{x'}{\lambda} + \phi\right)
\]

\[
x' = x\cos(\theta) + y\sin(\theta) + z\phi
\]
\[
y' = -x\sin(\theta) + y\cos(\theta)
\]
\[
z' = z\cos(\theta) - y\phi
\]

\[
x' = \frac{x}{\sigma} + \gamma \frac{y}{\sigma} + \gamma \frac{z}{\sigma}
\]

\[
y' = -x' + \gamma y' + \gamma z'
\]

\[
z' = x' + y' + z'
\]
\[ b = \log_2 \left( \frac{\sigma \lambda}{\pi} + \sqrt{\frac{\ln 2}{2}} \right) \] 

\[ \frac{\sigma}{\lambda} = \frac{1}{\pi} \sqrt{\frac{\ln 2}{2} 2^b + 1} \] 

\[ \frac{\sigma}{\lambda} = \frac{1}{\pi} \sqrt{\frac{\ln 2}{2} 2^b - 1} \] 

Table 1. Descriptions of all parameters of the filter.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta )</td>
<td>Parallel stripe direction of the filter, value range ([0\text{~to} 360^\circ])</td>
</tr>
<tr>
<td>( \phi )</td>
<td>Offset phase of the filter, value range ([-180\text{~to} 180^\circ])</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>The standard deviation of the Gaussian function that determines the size of the filter space. This value is related only to the semi-response spatial frequency bandwidth, when ( b = 1 ), ( \sigma \approx 0.56 \lambda ).</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>The spatial aspect ratio, used to adjust the shape of the Gabor filter in the ( x ) and ( y ) directions. When the value is 1, it is circular; when the value is less than 1, the filter shape is elongated with the direction of the parallel stripes.</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>Spatial frequency-dependent wavelengths associated with sinusoidal components</td>
</tr>
<tr>
<td>( b )</td>
<td>Half-response spatial frequency bandwidth, related to ( \sigma / \lambda ), is a positive real number, generally taken as 1.</td>
</tr>
</tbody>
</table>

The filtering performance of a Gabor filter is mainly determined by the size of the convolution kernel. Different convolution kernel sizes may have different effects on the Gabor filter. If the edge length of the convolution kernel is larger than the wavelength, it has no effect on the result of filtering. If the edge length of the convolution kernel is smaller than the wavelength, then the entire waveform is not fully included in the convolution calculation, resulting in poor filtering of the waveform edges. It can be concluded that the edge length of the convolution kernel of the Gabor filter must be greater than the wavelength to ensure that the entire waveform is completely contained in the convolution kernel to maximize the filtering effect of the Gabor filter. Additionally, a change in phase may also have effects on the entire Gabor filter. The change in phase brings about a change in the waveform at the center of the convolution kernel of the Gabor filter. If the center point of the filter kernel is directly in front of the wave peak (phase 0), the filtering effect of the whole image will be enhanced. On the contrary, if the center point of the filter kernel is directly opposite to the trough (phase 180), the filtering effect will be weakened. Therefore, it is necessary to avoid coinciding the filter center point with the zero-crossing of the waveform; otherwise, the effect of the filter may not be seen. Therefore, when using the Gabor convolution kernel, it is important to adjust each parameter to the appropriate range so that the desired effect may be seen.

2.3. The CBAM Module

CBAM [24] is a model of attention mechanism that can be used to enhance the performance of convolutional neural networks. Specifically, the attention mechanism provides guidance to the network such that it can automatically learn what to pay attention to in a sequence of pictures or text, and its purpose is to enhance the perception of a deep convolutional neural network and the utilization of image features and to improve accuracy. The CBAM starts from two key domains of action, namely, the channel and the space. Two analytical dimensions, spatial attention and channel attention, are introduced. Such a design implements a sequential attention structure from channel to space, where spatial attention allows a neural network to focus more on the pixel regions of the image that are important for classification, thus ignoring irrelevant regions. Meanwhile, channel attention is used to deal with the allocation relationship of the feature map channels. In other words, CBAM makes the neural network more focused on processing the image information by considering both spatial and channel attention, selectively emphasizing important channels and spatial locations in the image, thus improving the network performance. The so-called channel attention mechanism informs the network what to pay attention to, and the spatial
attention mechanism tells the network where to pay attention to, and its overall network structure diagram is shown in Figure 3.

![Figure 3](image-url)

**Figure 3.** The general structure of a CBAM module, where the input and output feature maps are shown in blue color, the channel attention module and the spatial attention module are shown in green and purple colors respectively.

Figure 3 shows the overall structure of CBAM; the green module represents a channel attention module and the purple color represents a spatial attention module. The input feature map $F$ passes through the channel attention module and the spatial attention module in turn. The overall process can be expressed as follows: for the feature map generated by the network, $F \in R^{C \times H \times W}$, CBAM generates channel attention feature maps separately, $M_c \in R^{C \times 1 \times 1}$, and spatial attention feature maps, $M_s \in R^{1 \times H \times W}$. It can be described by Equations (9) and (10), where the operator $\otimes$ denotes element-level multiplication, with a broadcast mechanism for dimensional transformation and matching in between.

$$F' = M_c(F) \otimes F \quad (9)$$
$$F'' = M_s(F') \otimes F' \quad (10)$$

### 2.3.1. Channel Attention Module

Channel attention focuses on the information useful for the desired task in the image; the specific structure is shown in Figure 4, which consists of 3D maximum pooling, 3D average pooling, and a multi-layer perceptron. It mainly focuses on the relationship between different channels in the feature map and emphasizes the channel information that is more important to the current task by learning the channel weights. Specifically, 3D global average pooling and 3D maximum pooling are first performed on the input feature graph in the channel dimension to downscale the input feature graph $F$ from the data dimension direction to one dimension, and then the feature is fed into a shared multilayer perceptron (MLP) network in order to extract channel-related features and compute their importance scores. This process generates a channel attention graph, $M_c \in R^{C \times 1 \times 1}$, which is used to adjust the importance of each channel. In order to reduce the number of parameters, a reduced dimensionality coefficient $r$ is used in the MLP, $M_c \in R^{\frac{C}{r} \times 1 \times 1}$. Finally, it is normalized by a sigmoid function so that the weight of each channel is between 0 and 1, and the sum is multiplied with each channel in the original feature map to generate a new feature map as defined by Equation (11). Based on Equation (9), the output feature map $M_c(F)$ can be obtained with Equation (12).

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \quad (11)$$
$$M_c(F) = \sigma(W_1(W_0(F_{avg}^{c}) + W_0(F_{max}^{c}))) \quad (12)$$

where $F$ denotes the input feature map; $AvgPool$ denotes the three-dimensional average pooling operation; $MaxPool$ denotes the three-dimensional maximum pooling operation; $F_{avg}^{c}$ and $F_{max}^{c}$ denote the features from the average pooling and maximum pooling oper-
Finally, it is normalized by a sigmoid function so that the weight of each element in the feature map can be obtained. The output can be obtained with $F' = M_c(F) \odot F$.

**Figure 4.** The structure of the attention channel module. The MaxPool layer is shown in dark blue color and the output of the shared MLP network on its output is shown in light blue color; The AvgPool layer is shown in dark orange color and the output of the shared MLP network on its output is shown in light orange color; input feature maps $F$ and $M_c$ are shown in light and dark green colors respectively.

### 2.3.2. Spatial Attention Module

Different from the channel attention, spatial attention focuses on the locations of the effective information in the feature map, and its structure is shown in Figure 5, which mainly focuses on the relationship between different spatial locations in the feature map, and guides the network to focus on more important image regions by learning spatial weights. Specifically, firstly, two new feature maps are generated by downscaling the input feature map $F'$ in spatial dimensions with 3D average pooling and 3D maximum pooling. Their channel numbers are both 1. These two feature maps are spliced together. Secondly, the spliced feature maps are subjected to convolution operation with a $7 \times 7$ convolution kernel and then activated by a sigmoid function to generate a new feature map, and then finally apply it to the original feature map. In summary, the new feature map $M_s(F)$ can be defined with Equation (13):

$$M_s(F) = \sigma\left(f^{7 \times 7}\left([\text{AvgPool}(F); \text{MaxPool}(F)]\right)\right) = \sigma\left(f^{7 \times 7}\left(F^{s_{\text{avg}}}; F^{s_{\text{max}}}\right)\right)$$

(13)

where $M_s(F)$ denotes the output feature map; $F$ denotes the input feature map; AvgPool denotes the three-dimensional average pooling operation; MaxPool denotes the three-dimensional maximum pooling operation; $F^{s_{\text{avg}}}$ and $F^{s_{\text{max}}}$ denote the features obtained with the average pooling and maximum pooling operations respectively; $\sigma$ is the sigmoid activation function. Finally the output can be obtained with $F'' = M_s(F') \odot F'$.

**Figure 5.** The structure of the spatial channel module, where the layer for MaxPool is shown in blue color and the layer for AvgPool is shown in orange color. In addition, the input feature and output feature maps are shown in light and dark green colors respectively.

In convolutional neural networks, the convolution operation is a key step used to extract features from the input data. However, the convolution operation may introduce some redundant information or irrelevant features. The proposed approach adds the Gabor convolution operation to the 3D model, which may introduce redundant information to the model; in order to solve this problem, the CBAM module is added after the convolution to weight the feature map after the convolution operation, which enhances the expressive ability and feature selectivity of the network, and strengthens the network’s ability to perceive the boundary information of the brain tumor. A more accurate location of the tumor region can thus be obtained. As a result, the segmentation accuracy can be improved.
3. Results

3.1. Experimental Setup and Dataset

The experimental configuration is as follows: a Windows 11 system with a 32-core Intel(R) Xeon(R) Platinum 8350C processor made by the Intel Corporation, Santa Clara, CA, USA and an RTX A5000 (24 GB) GPU graphics card made by the Nvidia, Santa Clara, CA, USA; the program runs on a 64-bit ubuntu 20.04 environment with Cuda 11.3, Python 3.8, and Pytorch 1.11.0.

The public dataset of brain tumor segmentation challenge (BraTS Challenge) 2021 [25] is utilized to evaluate the segmentation accuracy of the proposed approach. It has a total of 2040 cases; 1251 of them are included in the training set cases, 219 of them are included in the validation set, and the remaining 570 cases are in the test set. Each case contains four modalities of MRI, as shown in Figure 6, where, T1 (T1-weighted images) modality can be used to show the anatomical structure of the tumor; T2 (T2-weighted images) modality can be used to detect edema, inflammation, etc.; the T1CE (Contrast-Enhanced Imaging) modality can show the internal condition of tumors and hemangiomas, and FLAIR (Fluid Attenuated Inversion Recovery) modality can better show the condition around the tumor as well as the edema area. The dimensionality of each modality is $240 \times 240 \times 155$. Since the BraTS 2021 training set has labels, while the validation set and test set have no labels, the dataset is divided into a training set, a validation set, and a test set in a ratio of 8:1:1 to evaluate the segmentation accuracy of the proposed approach.

**Figure 6.** Examples of the four modalities of a case: (a-d) are the FLAIR, T1, T1CE, and T2 modalities for the case.
In this experiment, the network is trained for 120 iterations due to the large number of 3D image parameters, as well as the amount of computation needed for training, and due to the limited hardware available for the experiment, the batch size of this experiment is set to be 1. The weights of the network are updated using the SGD optimizer, while the warm-up learning strategy is used to warm up the network by firstly preheating for 10 epochs, with a minimum learning rate of 0.002 and a maximum learning rate of 0.004.

3.2. Data Preprocessing

Generally, when images are input into the network for training, a series of preprocessing work needs to be performed on the input images. The two-dimensional data and three-dimensional data processing strategy are slightly different, and since the dataset contains three-dimensional data and the size of each image is \(240\times240\times155\), the slice processing is not necessary. The main steps include the following. (1) Data normalization: in order to achieve faster convergence of the model and reduce the training time, an input MRI image is processed for Z-score normalization so that the mean of the image is 0, and the variance of the image is 1. (2) Image cropping: crop the image size from \(240\times240\times155\) to \(160\times160\times128\) to remove the black border area around the tumor, which not only improves the accuracy but also reduces the amount of computation. (3) Image enhancement: data enhancement operations, such as random flipping, random rotating, Gaussian noise changing, contrast transforming, brightness transforming, and so on are performed on the image during the training process, which improves the model’s generalization ability. The above steps help to remove irrelevant information from the image and enrich the features of the image, which helps to improve the training effect of the model.

3.3. Evaluation Metrics

Evaluation metrics are metrics used to measure the difference between the final segmentation results and the true labeling and are important for outcome diagnosis, as well as subsequent treatment options. The experiment finally needs to segment three regions: the enhanced tumor region (ET), tumor core region (TC), and whole tumor region (WT). Therefore, in order to evaluate the network model proposed in this paper, the Dice coefficient (Dice coefficient) and the Hausdorff distance (HD) are used to assess the accuracy of the final segmentation results. The two measures are the two most important metrics for most of the current brain tumor segmentation methods.

The Dice coefficient measures the degree of overlap between two sets, and the range for its value is \([0,1]\); the larger the Dice coefficient is, the more accurate the segmentation result is. In the field of medical images, the Dice coefficient is widely used as the main index to judge the degree of conformity between a predicted segmented region and the ground truth, which is calculated based on Equation (14), where \(X\) denotes the set of values in the ground truth and \(Y\) denotes the set of predicted values; \(X \cap Y\) denotes the intersection of the predicted result with the ground truth.

\[
D(X, Y) = \frac{2 \cdot |X \cap Y|}{|X| + |Y|} \tag{14}
\]

The Hausdorff distance (HD) is used to measure the distance between two point sets and evaluate the distance between a predicted edge and a real edge. A smaller value for HD usually suggests higher segmentation accuracy. In addition, the result of the 95th quantile, HD95, is used as the value of HD. Equation (15) is used to calculate HD, where \(d(x, y)\) denotes the Euclidean distance between \(x, y\). As shown in Figure 7, \(x\) and \(y\) are two points from the predicted edge and the corresponding real edge, respectively.

\[
HD(X, Y) = \max\left\{\sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y)\right\} \tag{15}
\]
3.3. Evaluation Metrics
Evaluation metrics are metrics used to measure the difference between the final segmentation results. The two measures are the two most important metrics for its value is $[0, 1]$; the larger the Dice coefficient is, the more accurate the segmentation results are used to obtain $L$. Equations (16)–(18) are used to obtain them, respectively.

\[
L_{CE} = -\sum_{i=1}^{K} \omega_i f_i \log p_i \tag{16}
\]

\[
L_{Dice} = 1 - \frac{2\sum_{i=1}^{K} \omega_i f_i p_i + \epsilon}{\sum_{i=1}^{K} \omega_i (f_i + p_i) + \epsilon} \tag{17}
\]

\[
L = L_{CE} + L_{Dice} \tag{18}
\]

3.4. Loss Function

The loss function describes the deviation between the predicted value of a sample and its true value and is an important means of determining the accuracy of a prediction model. A smaller value often suggests that a predicted value is closer to the true value. Meanwhile, since the problem of class imbalance also exists in the brain tumor segmentation task, the weighted cross-entropy loss function $L_{CE}$ and the sum of the Dice loss function with smoothing coefficients $L_{Dice}$ are used to obtain $L$. Equations (16)–(18) are used to obtain them, respectively.

\[
L_{CE} = -\sum_{i=1}^{K} \omega_i f_i \log p_i \tag{16}
\]

\[
L_{Dice} = 1 - \frac{2\sum_{i=1}^{K} \omega_i f_i p_i + \epsilon}{\sum_{i=1}^{K} \omega_i (f_i + p_i) + \epsilon} \tag{17}
\]

\[
L = L_{CE} + L_{Dice} \tag{18}
\]

3.5. Analysis of Results

In this paper, the proposed network Ga-U-Net is analyzed and compared with the two basic networks U-Net and Attention U-Net. In addition, five other models are also tested with the dataset BraTS 2021, and the two metrics of the Dice coefficient and Hausdorff distance (HD95) proposed above are used to evaluate the segmentation results obtained for three regions, including the enhanced tumor region (ET), tumor core region (TC), and whole tumor region (WT). The comparison results for the different models are shown in Tables 2 and 3.

As can be seen from Table 2, the overall performance of the network Ga-U-Net proposed in this paper is higher than that of the other models. It can achieve the highest average value for the Dice coefficients. Clearly, the Dice coefficients of the proposed approach on the enhanced tumor region (ET), the tumor core region (TC), and the whole tumor region (WT) are 0.856, 0.897, and 0.910, respectively. An improvement of 0.017, 0.02, and 0.003 is thus achieved by the proposed approach on ET, TC, and WT, respectively. It is also clear from Table 2 that the improvements in the segmentation accuracy for WT are not as significant as those for ET and TC. This is probably due to the fact that the Gabor convolution in the GA module is able to capture the local structural and textural information of the image more accurately. In addition, the results in Table 2 suggest that the CBAM attention mechanism improves the weights of the features of the tumor core region, suppresses the surrounding irrelevant regions as much as possible, and thus enables the network to locate the region of interest more accurately. Therefore, the segmentation accuracy on the tumor site is improved.
Table 2. Comparison results of Dice coefficients for different models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dice Coefficient (%)</th>
<th>WT</th>
<th>TC</th>
<th>ET</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Net</td>
<td>90.70</td>
<td>91.50</td>
<td>91.40</td>
<td>91.80</td>
<td>90.60</td>
</tr>
<tr>
<td>AttentionU-Net</td>
<td></td>
<td>87.70</td>
<td>85.40</td>
<td>83.50</td>
<td>86.07</td>
</tr>
<tr>
<td>Peiris et al. [26]</td>
<td></td>
<td>83.90</td>
<td>81.70</td>
<td>83.20</td>
<td>84.43</td>
</tr>
<tr>
<td>Jiang et al. [27]</td>
<td></td>
<td>87.43</td>
<td>86.17</td>
<td>86.60</td>
<td>88.77</td>
</tr>
<tr>
<td>Roth et al. [28]</td>
<td></td>
<td>88.07</td>
<td>86.07</td>
<td>84.43</td>
<td>88.77</td>
</tr>
<tr>
<td>Andrade-Miranda et al. [29]</td>
<td></td>
<td>86.17</td>
<td>86.60</td>
<td>84.43</td>
<td>88.77</td>
</tr>
<tr>
<td>Shi et al. [30]</td>
<td></td>
<td>86.07</td>
<td>84.43</td>
<td>88.77</td>
<td>88.77</td>
</tr>
</tbody>
</table>

Table 3. Comparison results for Hausdorff distances for different models.

<table>
<thead>
<tr>
<th>Model</th>
<th>HD95</th>
<th>WT</th>
<th>TC</th>
<th>ET</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jiang et al. [27]</td>
<td>10.19</td>
<td>9.97</td>
<td>9.98</td>
<td>11.40</td>
<td>10.243</td>
</tr>
<tr>
<td>Roth et al. [28]</td>
<td>12.65</td>
<td>12.66</td>
<td>14.81</td>
<td>16.03</td>
<td>11.40</td>
</tr>
<tr>
<td>Andrade-Miranda et al. [29]</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Shi et al. [30]</td>
<td>12.03</td>
<td>12.03</td>
<td>14.81</td>
<td>16.03</td>
<td>12.03</td>
</tr>
<tr>
<td>Ga-U-Net</td>
<td>9.12</td>
<td>13.09</td>
<td>14.40</td>
<td>16.03</td>
<td>12.03</td>
</tr>
</tbody>
</table>

Table 3 shows a comparison of the Hausdorff distances achieved by all tested methods. It can be seen from Table 3 that the proposed Ga-U-Net model can also outperform several other network models on the HD95 index in the tumor core region, and its overall average index is better than the other network models. The proposed approach thus can accomplish the task of boundary segmentation with higher accuracy in the above three regions.

Figure 8 shows examples of the segmentation results obtained with the trained model on the BraTS 2021 dataset, with the coronal, sagittal, and horizontal planes in an order from top to bottom. From left to right are the T1 modality, the real label, and the segmentation results. From the segmentation results, it can be seen from Figure 8 that the segmentation results on the tumor core region are highly close to the ground truths with little amounts of error. However, in the enhanced tumor region (i.e., the white region), a certain amount of error exists in the coronal and sagittal planes. In addition, there is also a modest amount of error for the segmentation of the edema region (i.e., the light grey region). In general, the proposed approach is able to separate the tumor from the healthy region, and it retains more complete information from the tumor.
3.6. Ablation Experiments

In order to verify that the Gabor convolution works better in the low-level encoder, this paper uses the 3D U-Net as the base network and adds the Gabor convolution to the low-level encoder (i.e., the network has 1 GA module) and to each layer of the encoder (i.e., the network has 4 GA modules) to test the effect. Six sets of comparison experiments are performed to test the addition of the CBAM attention mechanism after convolution to see if the segmentation accuracy can be improved. The results of the experiments are shown in Table 4, and the loss curves obtained for the training set and the validation set of the experiments are shown in Figures 9 and 10.

Table 4. Experimental comparison results after adding different modules.

<table>
<thead>
<tr>
<th>Model</th>
<th>Evaluation Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average Dice Coefficient (%)</td>
</tr>
<tr>
<td>U-Net</td>
<td>87.43</td>
</tr>
<tr>
<td>with 1 GA</td>
<td>87.67</td>
</tr>
<tr>
<td>with 4 GA</td>
<td>80.33</td>
</tr>
<tr>
<td>with CBAM</td>
<td>82.54</td>
</tr>
<tr>
<td>with 1 GA+CBAM</td>
<td>88.77</td>
</tr>
<tr>
<td>with 4 GA+CBAM</td>
<td>86.10</td>
</tr>
</tbody>
</table>

Figure 8. Examples of segmentation results using the Ga-U-Net model.

Figure 9. Loss curves obtained for the training set.
which is an overall improvement over the underlying network, whereas the effect instead has advantages for tasks that need to focus on local information. The dataset used for experiments is three-dimensional and the tumors in the dataset are generally small and complex. Theoretically, improved segmentation accuracy can be achieved by the introduction of Gabor convolution. Experimental results have shown that this method is indeed effective. In addition, because of the applicability and generality of the CBAM attention mechanism, it is added to the encoder part in the hope that the input features can be better extracted. Finally, experimental results show that the introduction of the attention mechanism can improve the performance of the model.

Figure 10. Loss curves obtained for the validation set.

It is clear from Table 4 that the average Dice coefficient is slightly improved by introducing the GA module in the first layer of the encoder, as can also be seen from HD95, which is an overall improvement over the underlying network, whereas the effect instead slips more when the GA module is introduced in each layer of the encoder, proving that Gabor convolution indeed performs better in the low-level decoder and can achieve better results. In addition, when the CBAM attention mechanism is added to the network, the segmentation performance is also improved for both the 1 GA module and 4 GA modules. It is because the traditional attention mechanism based on convolutional neural network mainly focuses on the analysis of the channel domain and is limited to the descriptions of the interactions between the channels of the feature map. In contrast, the CBAM mechanism introduces two analysis dimensions, spatial attention and channel attention. By allocating attention to these two dimensions, the performance of the model is improved, and the feasibility of this attention mechanism is also demonstrated. It can also be seen from the loss curves for the training and validation sets that the loss of introducing 4 GA modules is slightly larger than those of the others, both in the training and validation sets, and the amount of loss in each round is more than 0.1, while 1 GA module coupled with the CBAM attention mechanism has the best overall results.

4. Conclusions

In this paper, a new Ga-U-Net model is proposed to combine Gabor convolution and an attention mechanism with an overall 3D U-Net encoder-decoder architecture. It contains both GA modules and CBAM modules. Since the Gabor convolution kernel, as a visual sensory field, has certain advantages in multi-scale and multi-directional feature extraction, Gabor convolution is very effective for texture feature analysis and extraction, which can assist on the capturing of local structures and features in the image and often has advantages for tasks that need to focus on local information. The dataset used for experiments is three-dimensional and the in dataset are generally small and complex. Theoretically, improved segmentation accuracy can be achieved by the introduction of Gabor convolution. Experimental results have shown that this method is indeed effective. In addition, because of the applicability and generality of the CBAM attention mechanism, it is added to the encoder part in the hope that the input features can be better extracted. Finally, experimental results show that the introduction of the attention mechanism can improve the performance of the model.
The model proposed in this paper also has some disadvantages. For example, the computation of 3D Gabor convolution requires relatively more time. In addition, some parameters of the filter, such as scale and direction, need to be selected reasonably. Improper selection may lead to performance degradation. The training cost becomes even higher due to the introduction of the attention mechanism. In addition, the advantages of Gabor convolution can only be realized in specific scenarios. In the future, improvements will continue to be made to this network in the hope that it can be adapted to other medical segmentation tasks for improved accuracy.

Author Contributions: Conceptualization, Y.C. and Y.S.; methodology, Y.C. and Y.S.; software, Y.C.; validation, Y.C. and Y.S.; formal analysis, Y.C.; investigation, Y.C.; resources, Y.C.; data curation, Y.C.; writing—original draft preparation, Y.C.; writing—review and editing, Y.S.; visualization, Y.C.; supervision, Y.S.; project administration, Y.S.; funding acquisition, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Jiangsu University of Science and Technology, grant number 1132921208.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data in this research can be obtained at the following link: http://braintumorsegmentation.org/ (accessed on 4 August 2023). The source code of an implementation of the proposed approach is available at https://github.com/cy253/Gabor-u-net.git (accessed on 10 May 2024).

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

References
2. Zhang, T. Research on MRI Brain Tumor Classification and Segmentation Method Based on Deep Learning. Master’s Thesis, Yanshan University, Qinhuangdao, China, 2022. [CrossRef]


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(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.