applied sciences

MDPI

Article

DualTrans: A Novel Glioma Segmentation Framework Based on a Dual-Path Encoder Network and Multi-View Dynamic Fusion Model

Zongren Li, Wushouer Silamu *, Yajing Ma and Yanbing Li

School of Information Science and Engineering, Xinjiang University, Urumqi 830017, China; lizongren@stu.xju.edu.cn (Z.L.); myj18993156814@163.com (Y.M.); lzr549073488@163.com (Y.L.)
* Correspondence: wushouersilamu@126.com

Abstract: Segmentation methods based on convolutional neural networks (CNN) have achieved remarkable results in the field of medical image segmentation due to their powerful representation capabilities. However, for brain-tumor segmentation, owing to the significant variations in shape, texture, and location, traditional convolutional neural networks (CNNs) with limited convolutional kernel-receptive fields struggle to model explicit long-range (global) dependencies, thereby restricting segmentation accuracy and making it difficult to accurately identify tumor boundaries in medical imaging. As a result, researchers have introduced the Swin Transformer, which has the capability to model long-distance dependencies, into the field of brain-tumor segmentation, offering unique advantages in the global modeling and semantic interaction of remote information. However, due to the high computational complexity of the Swin Transformer and its reliance on large-scale pretraining, it faces constraints when processing large-scale medical images. Therefore, this study addresses this issue by proposing a smaller network, consisting of a dual-encoder network, which also resolves the instability issue that arises in the training process of large-scale visual models with the Swin Transformer, where activation values of residual units accumulate layer by layer, leading to a significant increase in differences in activation amplitudes across layers and causing model instability. The results of the experimental validation using real data show that our dual-encoder network has achieved significant performance improvements, and it also demonstrates a strong appeal in reducing computational complexity.

Keywords: brain-tumor segmentation; Swin Transformer; CNN; dual encoder; stability

check for updates

Citation: Li, Z.; Silamu, W.; Ma, Y.; Li, Y. DualTrans: A Novel Glioma Segmentation Framework Based on a Dual-Path Encoder Network and Multi-View Dynamic Fusion Model. *Appl. Sci.* **2024**, *14*, 4834. https://doi.org/10.3390/app14114834

Academic Editor: Claudio Belvedere

Received: 18 April 2024
Revised: 25 May 2024
Accepted: 25 May 2024
Published: 3 June 2024

1. Introduction

Given the rapid advancements in medical imaging technology, magnetic resonance imaging (MRI) has emerged as the preferred method for brain diagnosis and treatment planning. The swift and precise automatic segmentation of MRI images through computer-aided diagnosis technology is crucial for the treatment and prognosis of brain tumors. Researchers are progressively turning to computer-assisted methods for brain-tumor segmentation, leveraging a variety of machine-learning and deep-learning algorithms to achieve optimal results.

In 2012, the authors of [1] designed AlexNet, a deep convolutional neural network, and were the winners of ImageNet LSVRC in the same year. Subsequently, other researchers then introduced AlexNet into the fields of image segmentation and object detection. AlexNet introduced the ReLU activation function and Dropout technique: The ReLU activation function addressed the gradient vanishing problem of sigmoid in deep networks, while the Dropout technique effectively reduced overfitting. However, in AlexNet, information propagation primarily relied on local convolution operations, which meant that information interactions mainly occurred between adjacent pixels or feature maps. This locality constraint limited the network's ability to perceive global information, making it difficult for the network to fully utilize contextual information across the entire image. In 2014, ref. [2] proposed the deep convolutional structure VGGNet, which uses pre-training to train a shallow network and then reuses the weight of the previous layer to train the subsequent layer of the network. Through repeated training, problems, such as the weight initialization of the model, are solved. It also allows the model to converge faster during training. VGGNet consists of five convolutional layers, three fully connected layers, and a softmax output layer. Max-pooling is used to separate layers, and multiple smaller convolutional layers (3 × 3) are used instead of larger convolutional layers to reduce the number of parameters. However, this approach also limits the network's ability to perceive global information. In 2015, ref. [3] proposed the UNet network, which adopts an encoder–decoder architecture, combining both lightweight and high performance and intensively integrating shallow features and deep features, making this model outstanding in the field of medical image segmentation. However, traditional UNet encoders and decoders mainly perform local operations, which limits the model's ability to integrate global information. In biomedical image segmentation, the integration of global information is crucial for improving segmentation accuracy. In 2018, ref. [4] proposed an improved UNet++ network based on the UNET network. The model not only uses the structure of UNet for reference but also adopts the dense connection mode of the DenseNet network and introduces deep supervision. It not only preserves and reconstructs global information and local information but also makes the model more efficient. Due to the introduction of more nested structures and skip connections in UNet++, this results in an increase in the model's parameter count and computational complexity. Later, ref. [5] proposed the vision transformer and applied it in the field of brain-tumor segmentation. Compared with convolutional neural networks, the Vision Transformer enables the model to better obtain global semantic information. However, since ViT mainly captures global information through self-attention mechanisms, it may not be as effective as convolutional neural networks (CNNs) in handling local details and texture features. CNNs can naturally extract local features of images through convolution operations, while ViT requires additional design or integration with other techniques to enhance its ability to capture local context information. Researchers have found that Vision Transformers cannot effectively capture local semantic information, while convolutional neural networks have limitations in capturing global semantic information. The trend in the field of brain-tumor segmentation is to deeply integrate the two, such as the work by [5] proposing the TransBTS network, which combines Transformers for the first time in 3D MRI brain-tumor segmentation, allowing it to capture both global and local features simultaneously. The encoder first utilizes a 3D CNN to extract volumetric spatial feature maps, then models global features using a transformer, enabling the comprehensive capturing of global and local features in the image. However, due to the quadratic computational complexity and sequence length of the Transformer model, TransBTS requires higher computational resources when dealing with large-scale or high-resolution images. The model also has a higher number of parameters, and since the Transformer does not adopt a hierarchical architecture, it has a certain impact on segmentation accuracy. In conclusion, the current brain-tumor segmentation methods that combine CNN and Transformer are able to capture both global and local features simultaneously. However, how to optimally integrate these two types of features to improve algorithm segmentation accuracy and efficiency remains a challenge. In view of this, we proposed a glioma segmentation framework based on a dual-path encoder network and multi-view dynamic fusion model. The main contributions of this model are as follows:

(a) An innovatively proposed dual-path encoder architecture based on CNN and Swin Transformer, combined with a CNN and an improved Swin Transformer, a convolution operation is used to extract local dependencies and rich local features, and an improved Swin Transformer is used to learn global dependencies for global modeling. Then, feature fusion and upsampling are carried out to produce the segmentation results. Deeply integrating CNN and Transformer, leveraging the strengths of both frameworks, effectively enhances the accuracy of brain-tumor boundary recognition.

Finally, considering that as the depth of the Swin Transformer model increases, the differences in the amplitudes of cross-layer activations significantly grow, mainly due to the outputs of residual units directly added to the main branch. This instability issue in large-scale models of the Swin Transformer is addressed by normalizing the activation values of each residual branch and merging them back to the main branch, thereby enhancing the stability of training through structural improvements.

(b) A new location coding module is proposed. By adding a trainable parameter in the local window (M × M × M) and integrating location information in self-attention training, the Swin Transformer encoder structure can obtain rich location information, which helps to improve the segmentation accuracy of the brain-tumor model, especially for the recognition of the brain-tumor boundary region. M represents the local window size during Swin Transformer training, and r represents the relative positional offset.

(c) For validation of the benchmark dataset, this study utilized publicly available datasets named Brats 2021 [6–8] and Brats 2019 [9], which were provided by the organizers of MICCAI (International Conference on Medical Image Computing and Computer-Assisted Intervention) and served as part of the BraTS challenge. The experimental results of the Brats 2021 and Brats 2019 datasets demonstrated the effectiveness of the model, further improving the segmentation accuracy of brain tumors.

2. Related Work

2.1. Swin Transformer

The Transformer [10] was first proposed in the field of natural language processing (NLP) and was published by Google in Computation and Language. The emergence of the Transformer solves problems in the NLP field, such as the inability to be parallelized, limited memory length, etc., previously faced by RNN, LSTM, and other networks. Inspired by the success of the Transformer in the field of NLP, ref. [5] proposed the Vision Transformer, which applies the standard Transformer model to the field of vision with minimal modifications. While keeping the core structure of the Transformer model unchanged, necessary adjustments were made to adapt it to image-processing tasks. The images are divided into patch blocks, which are then used as input token sequences for the model. Through the Vision Transformer's unique image patch processing, token sequence input, self-attention mechanism, and global information modeling capabilities, it effectively addresses the shortcomings of previous CNNs in capturing global semantic information. Therefore, the Vision Transformer has been highly sought after by researchers in the vision field since it came out. For example, in reference [11], researchers utilized the powerful feature-extraction and sequence-modeling capabilities of the Transformer architecture to capture the spatial–temporal relationships between multi-view images, proposing a 3D human pose estimation method based on a Transformer that incorporates multi-view spatial–temporal relationships. However, as research has progressed, researchers have found that, limited by hardware conditions, the Vision Transformer has been shown to have a poor application impact in the field of image segmentation and target detection. Then, in 2021, Microsoft Research published a paper on the Swin Transformer [12] in ICCV, which performed well on multiple visual tasks once published. Compared to the Vision Transformer, the Swin Transformer adopts a hierarchical structure, where the size of feature maps decreases gradually as the network depth increases while increasing the level of feature abstraction. This design is akin to the downsampling operation in convolutional neural networks (CNNs), but the Swin Transformer achieves this process through the self-attention mechanism. Given that computing global self-attention is very computationally intensive, the Swin Transformer uses local windows, and within each window, the Swin Transformer employs the self-attention mechanism to model relationships between pixels. This mechanism allows the model to reference other pixels within the window when processing a pixel, capturing global contextual information. Another key feature of the Swin Transformer is the use of Shifted Windows for computing self-attention, which confines the self-attention computation within non-overlapping local windows, thus improving computational efficiency. Additionally, by moving windows within consecutive blocks, the model can transmit information between adjacent windows, maintaining the communication of global information and effectively alleviating the computational burden and hardware limitations that arise when using the Vision Transformer for image segmentation. However, with the training and application of large-scale visual tasks, the Swin Transformer suffers from problems, such as training instability and a resolution gap between pre-training and fine-tuning. In [13], the Swin Transformer V2 model is proposed, which effectively solves the problem of training instability by using the scaled cosine attention module.

2.2. Position Embedding

In 2017, ref. [10] first proposed position coding, which is called sinusoidal position embedding. This coding allows the Transformer architecture to capture the inherent sequence order and absolute positions of elements. Consequently, the Transformer architecture may not effectively capture the relative position information of elements during the calculation of the self-attention matrix. Currently, researchers generally divide location coding into two categories: absolute location coding and relative location coding. Absolute location coding adds the absolute location information of tokens to the sequence, while relative location coding considers the relative location information between tokens when calculating the self-attention distribution. Absolute location coding had been used earlier, but ref. [14,15] proposed that using relative position coding could significantly improve segmentation accuracy. However, relative location coding generally has the disadvantage of a high overhead and cannot be combined with existing self-attention acceleration algorithms. In view of this, ref. [16] proposed conditional location coding. Different from previous fixed or learned location coding, conditional location coding is dynamically generated and conditioned on the local domain of input tags. The authors of [17] proposed rotating position coding, which realized the effect of relative position coding through the form of absolute position coding, organically unified the two, and reflected the relative position information between tokens in the form of a self-attention matrix bias. Therefore, this study draws on rotational position coding and conditional position coding to enhance location items as context sensitive, making attention location sensitive at a marginal cost. Advantages and limitations of various position coding are shown in Table 1.

Table 1. Advantages and limitations of various position coding in medical image segmentation.

Positional Encoding Type	Advantages	Limitations
Absolute Positional Encoding	Provides precise location information vital for accurate segmentation.	1. Lacks flexibility and struggles with deformations or transformations in the image. 2. Susceptible to changes in image orientation or scale.
Relative Positional Encoding	1. Accommodates image deformations, crucial in tumor-growth monitoring. 2. Enhances the model's ability to handle variations in tumor shape and size.	Demands complex computations to determine relative positions accurately.
Conditional Positional Encoding	Adapts to the context of the image, offering superior segmentation accuracy.	Requires sophisticated architectures and training strategies.
Conditional Positional Encoding	Addresses rotational variations, essential in handling different imaging angles.	Might not be as relevant in standard brain-tumor segmentation, where rotations are minimal.
DualTrans Positional Encoding (Our)	1. During the training process, positional offset information is learned to provide more precise segmentation accuracy. 2. Captures intricate spatial relationships within the image.	Due to the dynamic learning strategy, overfitting is prone to occur.

3. Materials and Methods

3.1. Overall Architecture

The overall network architecture is shown in Figure 1. The dual-encoder network we have proposed is a deep integration of a CNN and an improved Swin Transformer, which solves the limitations of a CNN in terms of global modeling, remote context interaction, and spatial dependency. Furthermore, the prior knowledge of the CNN's hierarchy, locality, and translational invariance is introduced into the transformer to build an improved Swin Transformer model. The improvement of the Swin Transformer model mainly involves performing modifications to the basic blocks or residual blocks, incorporating the DualTrans Positional Encoding. The details of these enhancements to the encoding block are described in detail in Section 3.2. Additionally, to address the stability issues of the large Swin Transformer model, a configuration known as residual post-normalization (res-post-norm) is adopted, moving it from the front of each residual unit to the back. This reduces the differences in activation amplitudes across layers, thereby enhancing the training stability. Meanwhile, Swin Transformer blocks are sliced in axial, coronal, and sagittal dimensions to reduce the computational complexity. After that, the Swin Transformer is used to analyze the 2D slices. Finally, the slices are fused specifically, given a multimodal brain-tumor image input of $X \in R^{H \times W \times D \times C}$, where the image size is $H \times W \times D$, and the number of input channels of the image is represented by C. Owing to its excellent performance, we use an encoder–decoder architecture as the main structure of the network, and we use a dual-encoder network to extract rich spatial information and semantic information. First, based on the locality and translational invariance of convolutional neural networks, we use a CNN to extract local information and features. Secondly, considering the unique advantages of the Swin Transformer in global modeling and remote contextual semantic information interactions, we use the Swin Transformer block to layer and downsample the input images for the second encoder. Then, the features processed by the Swin Transformer block are converted into three dimensions through the feature-mapping layer and concatenated with the features processed by CNN. After that, we restore the spatial resolution through the decoder layer and repeatedly stack the upper sampling and convolutional layers to gradually produce high-resolution segmentation results. The authors of [13] proposed the Swin Transformer V2 model and conducted experiments on four representative visual benchmarks, demonstrating the poor stability of the Swin Transformer when handling large-scale datasets. Given this, we normalize the output-activation values of each residual branch of the Swin Transformer before recombining them back into the main branch. By introducing post-normalization, we aim to enhance the training stability of large visual models. Finally, considering that the Swin Transformer does not save the position information when transforming the image into a patch through patch partition, which gives the sequence arrangement equal edges and limits the expression ability of visual tasks, we innovatively propose to incorporate the initially defined relative position offset into the self-attention learning mechanism. The dynamic priors are generated in the sensitivity field ($M \times M$), and the position offset is used as a trainable parameter to participate in the self-attention calculation. Compared with absolute position coding and relative position coding, this method can significantly improve accuracy.

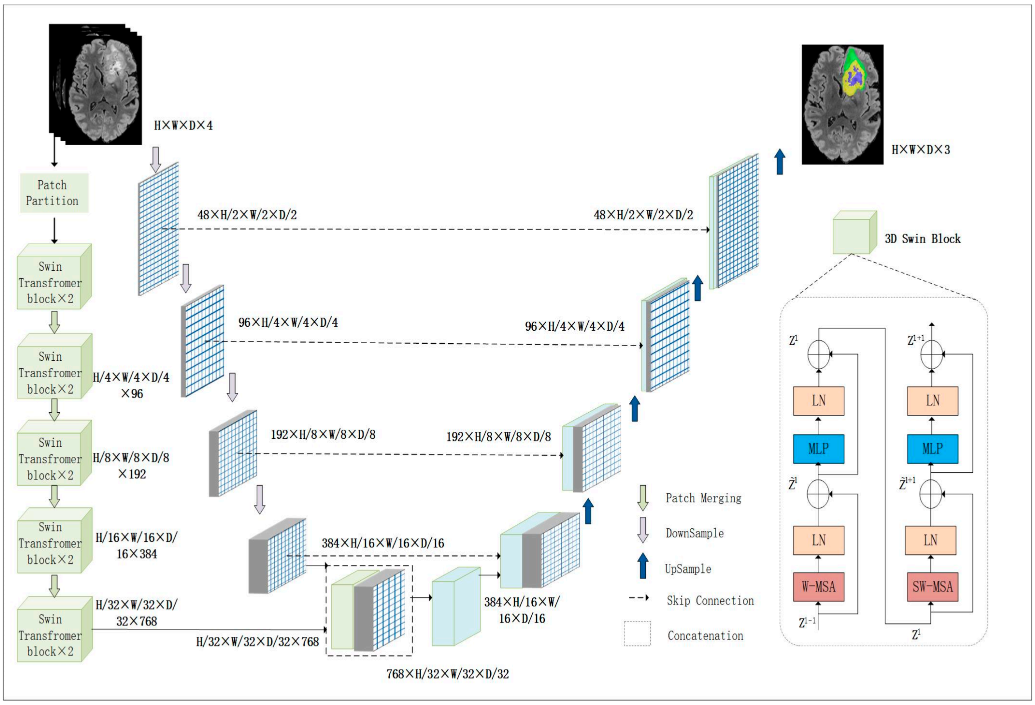

Figure 1. Overview of glioma segmentation framework for dual-path encoder networks and multi-view dynamic fusion models. In the encoder architecture, the subsampling path on the left represents the Swin Transformer encoder structure, and the subsampling path on the right represents the CNN encoder structure.

3.2. Network Encoder

Different from the Vision Transformer, which divides input images into H × W × C-sized patches and converts them into sequences for the self-attention calculation combined with location information, we will first encode the input images in two directions: Swin Transformer and CNN. Figure 2 illustrates the encoder architecture and input-data processing flow of the DualTrans model.

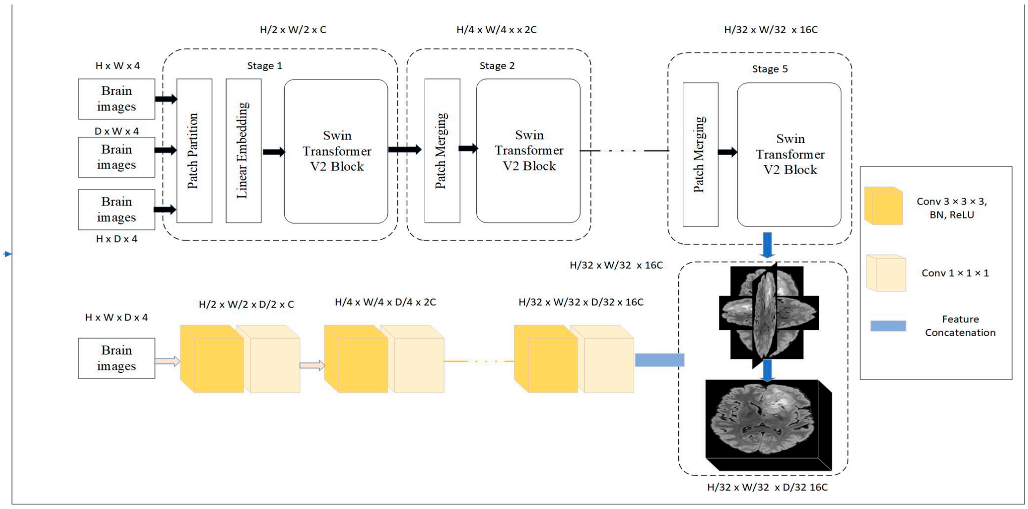

Figure 2. Encoder architecture diagram of the DualTrans model.

For the **Swin Transformer encoder,** compared with the 3D volume data, the Vision Transformer directly divides data into non-overlapping 3D blocks. However, such direct partitioning renders the Transformer unable to model image local context information across spatial and depth dimensions. To solve this problem, we encode the Swin Transformer block, which is composed of the LayerNorm layer, residual connection, multi-head attention module, and nonlinear multilayer perceptron (MLP). Utilizing Shifted Windows multi-head self-attention (W-MSA) versus Shifted Windows multi-head self-attention (SW-MSA) within a Swin Transformer block, the feature map is divided into multiple non-overlapping windows, and multi-head self-attention is only carried out in each window. In this way, the calculation load can be reduced, and the Transformer can better handle visual tasks. However, multi-head self-attention in each independent window will also isolate the information interactions between windows from the space, so SW-MSA is used to strengthen the information transmission between windows. For input $X \in R^{H\times W\times D\times C}$ of the given multimodal brain-tumor feature map, we slice it according to the three dimensions of axial, coronal, and sagittal, and then input $X_{input} \in R^{(H\times W)\times C}$, $X_{input} \in R^{(H\times D)\times C}$, and $X_{input} \in R^{(W\times D)\times C}$ into the Swin Transformer block, respectively, as shown in Figure 2 for the specific structure. Then, we merge them to form a 3D feature map. In Figure 1, we uniformly described the 3D feature map. We first create a sequence of tokens of dimension $\frac{H}{M} \times \frac{W}{M} \times \frac{D}{M}$ using the patch partition and, then, map them to the embedding space of dimension C through the linear embedding layer. The Swin Transformer encoder has four stages, with each stage containing a patch-merging layer and 2^N Swin Transformer blocks. In order to maintain the hierarchical structure of the encoder, in each stage, the patch-merging layer reduces the feature resolution by a factor of two, and then, the grouped patches are merged to obtain 4C feature embedding. In order to maintain the same operation as that in convolution, 4C features are transformed into 2C through the fully connected layer, which is similar to the pooling operation of convolution. Based on this local window self-attention mechanism, the output of layers l and l + 1 in the Swin Transformer encoder layer is calculated as follows:

$$\hat{z}l = W - MSA(LN(\hat{z}l - 1)) + Z^{l-1} \quad (1)$$

$$Z^{l} = MLP(LN(\hat{z}l)) + \hat{z}l \quad (2)$$

$$\hat{z}l + 1 = SW - MSA\left(LN\left(Z^{l}\right)\right) + Z^{l} \quad (3)$$

$$Z^{l+1} = MLP(LN(\hat{z}l + 1)) + \hat{z}l + 1 \quad (4)$$

where (1) represents the output of W-MSA, (2) represents the output of SW-MSA, Z^l represents the output features of the MLP module, W-MSA denotes multi-head attention using a regular window, and SW-MSA denotes multi-head attention using a shifted window. Using only the W-MSA module would prevent adjacent windows from interacting, leading to the model losing its ability for global modeling. Therefore, the W-MSA module and the SW-MSA module appear together in pairs. In order to efficiently calculate the shifted window mechanism, the self-attention calculation formula is as follows:

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

where Q, K, and V represent the query, key, and value matrices, respectively. $B \in R^{M^2\times M^2}$ is the relative position offset, used to represent the relative position information between patches.

DualTrans Positional Encoding. Since position information is not taken into account in the calculation of multi-head attention, the position order of the patch has no influence on the results of self-attention calculation, while position information is crucial for capturing spatial structure or shape in visual tasks, which will greatly reduce the accuracy of brain-tumor segmentation. Therefore, researchers are increasingly adding the location information back; for example, [18] further included the location information into the calculation of the multi-head attention mechanism and used trainable parameters to further optimize the relative position information. First, the relative position distance between each position abc ϵ $\underset{\smile}{\mathrm{H}}_{m\times m\times m}$ and the center point was defined as ijk in the window. Relative distances are decomposed across dimensions, so each element in abc ϵ $\underset{\smile}{\mathrm{H}}_{m\times m\times m}$ accepts distances in three dimensions, namely the X-axis offset r_{a-i}, the Y-axis offset r_{b-j}, and the Z-axis offset r_{c-k}. The X, Y, and Z axis offset is embedded and connects together to form $r_{a-i,b-j,c-k}$, and the relative attention of this spatial structure can be defined as:

$$y_{ijk}=\sum_{a,b,c\epsilon\underset{\smile}{\mathrm{H}}_{o(i,j,k)}}\mathrm{softmax}_{abc}(Q_{ijk}^{\intercal}K_{abc}+Q_{ijk}^{\intercal}r_{a-i,b-j,c-k}^{q})V_{abc} \tag{6}$$

Experiments have proven that this method can effectively improve the classification accuracy of visual tasks. However, we noticed that the position deviation proposed by this method only depends on the query pixel Q_{ijk} rather than key pixel V_{abc}, but the corresponding position information should also be paid attention to in key pixel V_{abc}. Therefore, we propose attention based on three-dimensional relative positional embeddings. First, we define the relative distance from ijk to each element abc ϵ $\underset{\smile}{\mathrm{H}}_{m\times m\times m}$, where the relative distance is decomposed into three dimensions. On this basis, we add a positional bias term $r_{a-i,b-j,c-k}^{v}$ related to the key pixel, with the specific formula being:

$$y_{ijk}=\sum_{a,b,c\epsilon\underset{\smile}{\mathrm{H}}_{o(i,j,k)}}\mathrm{softmax}_{abc}(Q_{ijk}^{\intercal}K_{abc}+Q_{ijk}^{\intercal}r_{a-i,b-j,c-k}^{q})(V_{abc}+r_{a-i,b-j,c-k}^{v}) \tag{7}$$

where $r_{a-i,b-j,c-k}^{q}$ is the position code of the learnable query, and $r_{a-i,b-j,c-k}^{v}$ is the position code of learnable value. Compared with Formula 6, the increase in vector $r_{a-i,b-j,c-k}^{v}$ does not introduce more parameters because they are shared among the attention heads between each layer, while the increase in $r_{a-i,b-j,c-k}^{v}$ can obtain more position information, which helps to improve the segmentation accuracy of brain tumors. Considering that $r_{a-i,b-j,c-k}^{q}$ is the matrix multiplication for all pixels in the local window (M × M × M), and $r_{a-i,b-j,c-k}^{v}$ is the position information of each pixel value in the local window, we did not add the $r_{a-i,b-j,c-k}^{k}$ vector to represent the position coding of learnable keys because it would introduce additional parameters. At the same time, it is not beneficial to obtain more position information. It is easy to have redundancy in location-information training.

Convolutional Neural Network Encoder. In the second encoder layer, we use a common convolutional neural network. First, the 3 × 3 × 3 convolution block is used for subsampling (stride = 2 convolution), and then, each layer also contains a 1 × 1 × 1 convolution block to convert the number of channels, gradually converting the input image into a feature map with low resolution but high feature representation. Rich local information can be effectively extracted through the locality and spatial invariance of convolution, and the $\frac{H}{32}\times\frac{W}{32}\times\frac{D}{32}\times 768$ feature block can be obtained from the last layer of the encoder. The feature block can be spliced with the feature block obtained by the Swin Transformer, and the local context feature information and global context feature information can be fused.

3.3. Network Decoder

As shown in Figure 3, our research adopts the encoder and decoder architecture, and the encoder corresponds to the decoder. Unlike patch merging in the Swin Transformer encoder and step convolution in the convolutional neural network encoder for downsampling, transposed convolution is used for upsampling at the decoder layer. First, in the bottleneck layer, in order to fit the input dimensions of the 3D CNN decoder, the output results of the Swin Transformer encoder are converted through the feature-mapping block designed at the feature-mapping layer, and the sequence data are re-mapped to the feature space. The feature block of $\frac{H}{32}\times\frac{W}{32}\times\frac{D}{32}\times 768$ is obtained, and then, it is spliced with the feature block obtained by the CNN encoder to obtain rich feature information. After the feature map is cascaded, it is upsampled. In order to avoid the loss of partial local information during feature downsampling, we also use skip connections to splice the features of the CNN encoder and decoder, so as to obtain more abundant spatial feature information.

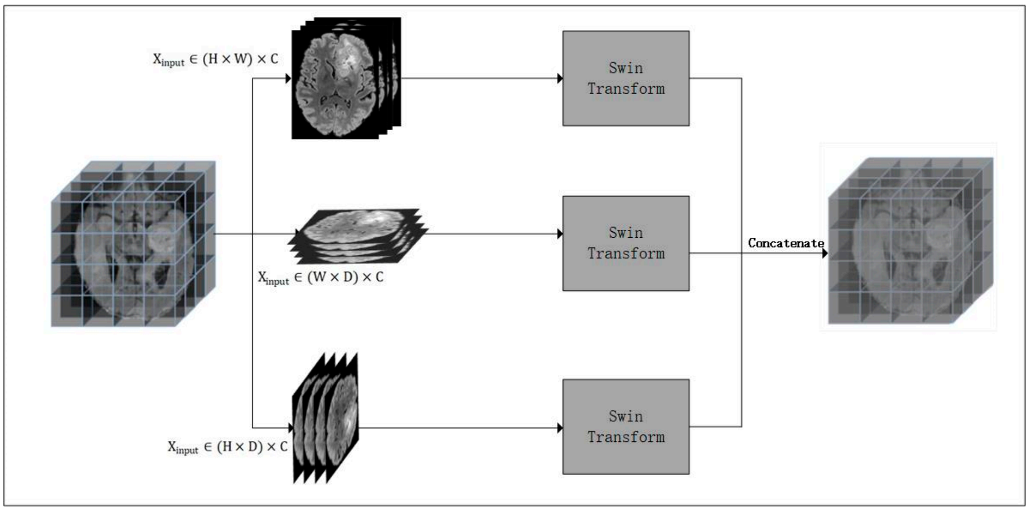
$X_{input} \in (H \times W) \times C$
$X_{input} \in (W \times D) \times C$
$X_{input} \in (H \times D) \times C$
Swin Transform
Swin Transform
Swin Transform
Concatenate

Figure 3. Overall architecture of the Swin Transformer encoder. The 3D feature map was sliced along the axial plane, coronal plane, and sagittal plane and was fused after the training.

4. Experimental Results

In order to further verify the feasibility and effectiveness of the proposed model, two datasets, Brats 2019 and Brats 2021, were used to verify the model, and the innovative sensitive location coding module mentioned in the study was proven through ablation experiments. Both the Brats 2019 and Brats 2021 datasets were provided by previous brain-tumor segmentation competitions. The event is organized by international authorities, such as The Radiological Society of North America (RSNA) and the American Society of Neuroradiology (ASNR). Over the years, the event has become a prestigious competition in the field of brain-tumor segmentation, producing multi-scale 3D CNN, nnU-Net, Extending nn-UNet [19–21], and other advanced brain-tumor segmentation algorithms.

4.1. Data and Evaluation Indicators

For the **dataset**, the first dataset used in this study is the Brats 2019 dataset provided by Brats, which uses MRI scans of patients' pre-operation brain conditions from multiple institutions, focusing on heterogeneous segmentation (appearance, shape, and tissue) of brain tumors, which in this case mainly refer to gliomas. Data are obtained from scanners at different institutions with different treatment regimens, and these MRI scans contain four modes of data: T1-weighted (T1), contrast-enhanced T1-weighted (T1ce), T2-weighted (T2), and T2 fluid-attenuated inversion recovery (T2-FLAIR) volumes. All datasets were manually segmented by 1–4 professional and experienced radiologists according to the same protocol standards. The segmented labels included enhancing tumor (ET-label 4), peritumoral edematous/invaded tissue (ED), and necrotic tumor core (NCR), for a total of three tumor regions. The Brats 2019 dataset contains two preoperative MRI sequence datasets: the training set, including 335 multimodal MRI cases and the corresponding brain-tumor labeling, and the validation set, containing 125 cases without any public labeling. These images have been preprocessed, including stripping the skull, co-registering all MRI volumes to the same anatomic template, and resampling at an isotropic resolution of 1 mm^3, resulting in a 128 × 128 × 128 brain-tumor image. The second dataset adopted in this study is the Brats 2021 dataset, which is consistent with the Brats 2019 dataset except for the number of cases, including the mode, output size, etc. Brats 2021 contains 8160 MRI scans of 2040 patients, among which 1251 cases are used as training sets and publicly labeled. A total of 219 cases of data were classified as validation sets, with the corresponding annotations not disclosed, and another 570 cases were classified as test datasets, with undisclosed data.

For **dataset preprocessing,** the BraTS dataset contains MRI images of multiple modalities (T1, T1Gd, T2, and T2-FLAIR). We integrate the data from these different modalities to effectively utilize multi-modal information for brain-tumor segmentation. However, we standardize these data, as the dataset provided by BraTS is composed of sequences acquired by different institutions using devices from various manufacturers, thus resulting in inconsistent intensities of MRI volumes. After removing background pixels, the images are cropped to a fixed patch size of 128 × 128 × 128. The four sequence images are then placed in the same dimension, resulting in each processed sample image having dimensions of (4, 128, 128, 128). The mask images undergo the same processing during training. To prevent overfitting, this study employs data-augmentation techniques, which are crucial in the training process. These techniques ensure that the model possesses a degree of invariance to specific natural transformations, effectively enhancing the model's generalization capabilities.

(a) The background region was clipped, considering the proportion of tumor region and non-tumor region, and 128 × 128 × 128 feature blocks were extracted.
(b) The scaling coefficient is 0.8–1.2; the probability is 0.15.
(c) Gaussian N(0, 0.01) noise is added.
(d) Gaussian smoothing is performed with $\alpha \in [0.5, 1.15]$.
(e) The probability of random flipping on the axial, coronal, and sagittal planes is 0.5.

For the **training details,** the dual-path encoder brain-tumor segmentation model proposed in this study is implemented in PyTorch. It is trained from scratch using six NVIDIA RTX 3090 GPUs (each with 24 GB of memory) with a batch size of 12 and trained for 7000 iterations. The model is trained using the Adam optimizer with a multi-learning strategy, starting with an initial learning rate of 0.0002 and decaying by 0.9 at each iteration.

For the **evaluation indicators,** the segmentation accuracy of the brain tumor was measured using the dice evaluation function and Hausdorff distance (95%) index, respectively, for the core tumor region (TC, label 1 and label 4), enhanced tumor region (ET, label 1), and entire tumor region (WT, label 1, label 2, and label 4).

(a) The dice evaluation function is a commonly used index to measure the segmentation accuracy of brain tumors. This index measures the accuracy of the model by calculating the overlap between the predicted results of the model and the real label. When the dice coefficient is close to one, the higher the overlap and the better the performance.

$$\mathrm{Dice} = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{8}$$

where true positive (TP) represents the number of pixels that the model correctly predicts to be positive, false positive (FP) indicates the number of pixels that the model incorrectly predicts to be positive, and false negative (FN) indicates the number of pixels that the model incorrectly predicts as a negative class.

(b) The Hausdorff distance (95%) represents the surface distance between the prediction and ground truth. The 95% quantized value of the maximum distance is different from the dice coefficient, which is sensitive to the inner filling of the mask. The Hausdorff distance is sensitive to the segmentation boundary of the brain tumor. It can effectively identify the boundary of the enhanced tumor region and tumor core region. The Hausdorff distance measures the distance between two subsets in the space, where d represents the element-by-element distance closest to the voxels from the first set of voxels to the second set of identical labels, X represents the true value label of the voxel, and Y represents the predicted value label of the voxel.

$$\text{Hausdorff distance (95\%)} \ = 95\%(\mathrm{d}(\mathrm{X}, \mathrm{Y}) || \mathrm{d}(\mathrm{Y}, \mathrm{X})) \tag{9}$$

4.2. Main Result

For **BraTS 2019,** the training set of this dataset contains 335 multi-modal MRI cases. A five-fold cross-validation evaluation was performed on the training set. The dice coefficients of our dual-path encoder model in WT, ET, and TC were, respectively, 93.68%, 90.75%, and 89.2%, and the Hausdorff distance 95% (HD) values were 6.54, 7.02, and 4.32, respectively. At the same time, we also compared our model with other SOTA models using the BraTS 2019 validation set, which contains 125 cases without any annotation. We uploaded all data to https://ipp.cbica.upenn.edu/ (accessed on 18 January 2024) to verify this model and compare it with other models, and the results are detailed in Table 2. We put forward the model of our dual-path encoder, and the dice coefficients in WT, ET, and TC were 91.42%, 85.63%, and 87.2%, respectively. For HD95, the values were 7.21, 8.32, and 5.64, respectively.

Table 2. Comparison between the DualTrans model and other SOTA models based on BraTS 2019 validation set.

Model	Dice Score (%)			Hausdorff 95 (mm)		
	ET	TC	WT	ET	TC	WT
TransBTS [22]	78.36	81.41	88.89	5.91	7.58	7.60
Attention-based [23]	75.9	80.7	89.3	4.19	7.66	6.96
Cross-Sequence [24]	78.09	84.32	90.81	2.88	5.74	5.27
Two-Stage Cascaded [25]	80.21	86.47	90.94	3.16	5.43	4.26
3D-UNet [26]	70.86	72.48	87.38	5.06	8.71	9.43
Swin UNETR [27]	85.2	86.9	90.8	8.78	5.62	6.23
Pei et al. [28]	81.33	84.08	88.62	4.21	8.02	5.46
DualTrans (ours)	85.63	87.2	91.42	8.32	5.64	7.21

We applied the proposed dual-path encoder model (DualTrans) to the BraTS 2019 validation dataset and evaluated the performance online on the Challenge web. Compared with Swin UNETR, our model showed great advantages in both indicators and showed significant improvement. Swin UNETR is a reformulation of the task of 3D semantic segmentation as a sequence-to-sequence prediction problem, and maps multimodal MRI features to one-dimensional sequences as the input to the layered Swin Transformer encoder. The model was verified with BraTS 2021; we obtained the source code of the model and verified it with BraTS 2019 according to the same training details, and the results were obtained as shown in Table 2. Swin UNETR combines the architectures of the Swin Transformer and UNETR (Transformer-based U-Net), utilizing the Swin Transformer to extract global context information. In comparison to Swin UNETR, our proposed DualTrans model achieved improvements in the dice coefficient: by 0.4% in the ET region, 0.3% in the TC region, and 0.42% in the WT region. This indicates that our dual-path encoder model, which leverages CNN to extract local context information, outperforms Swin UNETR, which solely employs the Swin Transformer in the encoder for segmentation tasks. On the other hand, the 3D-UNet model uses convolutional neural networks to extract feature information. Although CNN has unique advantages in extracting local context information, it has limitations in capturing global context information. In comparison to 3D-UNet, our DualTrans model showed significant improvements in the dice coefficient: by 14.77% in the ET region, 14.72% in the TC region, and 4.04% in the WT region. This demonstrates that, by leveraging the Swin Transformer for global context information and a CNN for local context information, and then combining these extracted features for upsampling, our model captures richer semantic features.

At the same time, we also show the comparative analysis results of excellent models, such as TransBTS, Attention-based, Cross-Sequence, Two-Stage Cascaded, 3D-UNet, etc. Compared with the TransBTS model, in brain-tumor areas such as ET, TC, and WT, we have achieved an increase of 7.3%, 6.8%, and 2.5%, respectively, in the dice coefficient. The TransBTS model combines a CNN and the Transformer to first use a 3D CNN brain-tumor-volume spatial feature map and then transforms the feature mapping, thus proving it to be an excellent example of utilizing the Transformer for global modeling. As a classic model combining a CNN and the Transformer in the field of brain-tumor segmentation, this model has great reference significance for our model transformation. It should be noted that the results of the 3D-UNet model in Table 2 on the BraTS 2019 validation set are based on the comparative experimental analysis results of the TransBTS paper. Compared with the Two-Stage Cascaded model, the brain-tumor regions of ET, TC, and WT were increased by 5.4%, 0.7%, and 0.5% respectively. Compared with the model by [28], the brain-tumor areas such as ET, TC, and WT were increased by 4.3%, 3.1%, and 2.8%, respectively. It can be seen that the improvement effect of our model is relatively obvious. Figure 4 shows the visual segmentation results on the Brats 2019 validation set.

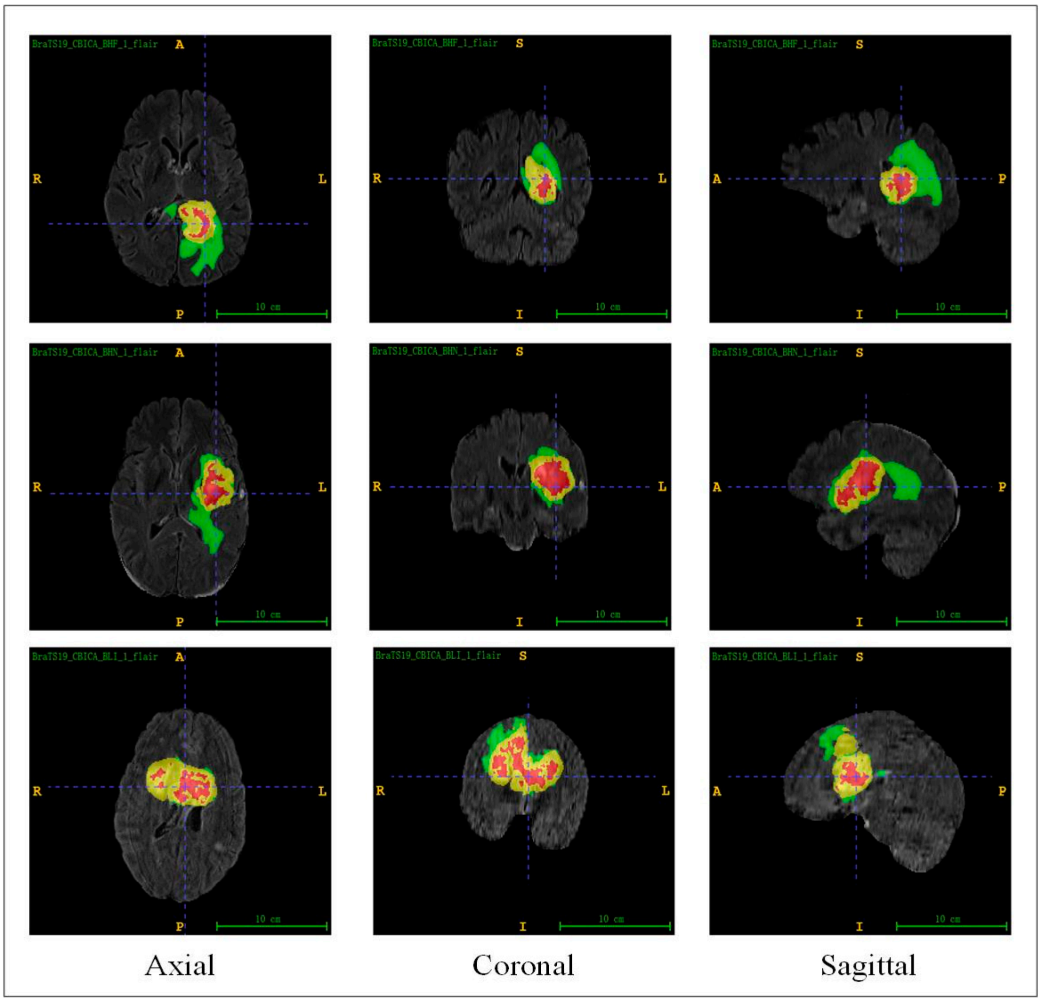
BraTS19_CBICA_BHF_1_flair
A
R
L
P
10 cm
BraTS19_CBICA_BHF_1_flair
S
R
L
I
10 cm
BraTS19_CBICA_BHF_1_flair
S
A
P
I
10 cm
BraTS19_CBICA_BHN_1_flair
A
R
L
P
10 cm
BraTS19_CBICA_BHN_1_flair
S
R
L
I
10 cm
BraTS19_CBICA_BHN_1_flair
S
A
P
I
10 cm
BraTS19_CBICA_BLI_1_flair
A
R
L
P
10 cm
BraTS19_CBICA_BLI_1_flair
S
R
L
I
10 cm
BraTS19_CBICA_BLI_1_flair
S
A
P
I
10 cm
Axial
Coronal
Sagittal

Figure 4. BraTS 2019 validation set on brain-tumor segmentation visual interface. The effect diagrams of axial, coronal, and sagittal plane segmentation using our proposed DualTrans model are shown, respectively.

The training set of the **BraTS 2021** dataset contains 1251 MRI multimodal brain-tumor cases, which is consistent with the operation of the Brats 2019 dataset. First, a five-fold cross-validation evaluation was performed on the training set. Our dual-path encoder model (DualTrans) was evaluated, and the dice coefficients of WT, ET, and TC were 93.89%, 91.23%, and 89.8%, and the Hausdorff distance 95% (HD) values were 4.36, 5.02, and 3.68, respectively. Table 3 demonstrates the comparison of segmentation accuracy of our proposed DualTrans model with other excellent models on the BraTS 2021 validation set.

Table 3. Comparison between the DualTrans model and other SOTA models based on BraTS 2021 validation set.

Model	Dice Score (%)			Hausdorff 95 (mm)		
	ET	TC	WT	ET	TC	WT
Swin UNETR [27]	85.8	88.5	92.6	6.02	3.77	5.83
Reciprocal Adversarial [29]	81.38	85.63	90.77	21.83	8.56	5.37
Qiran Jia et al. [30]	81.87	84.34	90.97	17.85	16.69	4.51
Yuan et al. [31]	84.79	86.55	92.65	12.75	11.19	3.67
Orthogonal-Nets [32]	83.2	84.99	91.38	20.97	9.81	5.43
Attention and Ensemble [33]	83.79	86.47	91.99	6.39	7.81	3.86
Swin–Unet [34]	85.37	87.26	92.08	14.32	9.80	11.28
Extending-nnUNet [21]	84.51	87.81	92.75	20.73	7.62	3.47
DualTrans (ours)	86.23	88.12	92.83	6.37	3.64	4.51

We uploaded 219 unlabeled MRI multimodal brain-tumor cases from Brats 2021 and evaluated the performance online on the Challenge web. The dice coefficients of the proposed dual-path encoder model DualTrans in the ET, TC, and WT brain-tumor regions on the validation set were 86.23%, 88.12%, and 92.83%, with HD95 values of 6.37, 3.64, and 4.51, respectively. Compared with other excellent models, our proposed DualTrans model has obvious advantages. Compared with the classic Swin UNETR model, the dice coefficients in the ET, TC, and WT brain-tumor regions were increased by 0.43%, −0.38%, and 0.23%, respectively. The dice coefficients of Swin UNETR in the ET and WT regions were lower than in our proposed model, while in the TC region, it was mainly the segmentation of the label 4 region that was less accurate. The dice coefficient of this model is higher than ours, mainly because our model is more complex, and the overall training dataset of brain-tumor segmentation is small, which fails to reflect the advantages of our model. Secondly, our model does not obtain the feature information of the Swin Transformer encoder path through skip connection during upsampling. Some global feature information is missing. However, adding this path will further increase the computational complexity of the model, but the improvement in the model accuracy is very limited. Compared with the Swin–Unet model, the dice coefficients are increased by 0.86%, 0.86%, and 0.75% respectively. Swin–Unet is a pure medical image-segmentation model similar to UNet. First, the tagged images are input into the encoder–decoder architecture based on the Transformer. Moreover, skip connections are used to learn local and global semantic features. Swin–Unet has a higher segmentation accuracy and a weaker ability to extract local semantic feature information compared with our model. Moreover, the model slices 3D feature maps into 2D feature maps and inputs the model without a good integration of 3D spatial semantic feature information.

In addition, we compared the proposed model with the classic model nnU-Net in the field of medical image segmentation. The dice coefficients for the ET, TC, and WT brain-tumor regions increased by 1.72%, 0.31%, and 0.08%, respectively. nnU-Net is an ensemble of multiple models that has performed well in various medical image-segmentation challenges, including brain-tumor segmentation. The comparison results with the nnU-Net model also demonstrate the wide applicability and efficiency of our proposed model in the field of medical image segmentation.

4.3. Ablation Study

We designed a large number of ablation experiments to demonstrate the validity of our proposed dual-path encoder model principle. And based on five-fold cross-validation evaluations on the BraTS 2021 training set to verify our design principle, our design mainly includes the following aspects: (a) the impact of the Swin Transformer encoder on segmentation accuracy in the DualTrans model and (b) a focus on the influence of our innovative sensitive position coding on model accuracy when extracting feature information from the Swin Transformer encoder.

As to the **Swin Transformer encoder path,** our model adopts dual-path encoders, the first one being the Swin Transformer encoder, and the second one being a convolutional neural network encoder. As shown in Figure 2, when the Swin Transformer encoder path is removed, the network structure is similar to the UNet network architecture. In terms of specific operations, the input image is gradually transformed into a feature map with a low resolution but a high feature representation by using a 3 × 3 × 3 convolution block for downsampling (stride = 2 convolution), and rich local information is extracted via convolution. In order to prevent the loss of feature information in the downsampling process, the concat feature is carried out by jumping connections. The feature block of $\frac{H}{32} \times \frac{W}{32} \times \frac{D}{32} \times 768$ is obtained in the last layer of the encoder, and then, the upsampling operation is carried out through transposed convolution. The whole process is consistent with the 3D-UNet architecture, and the segmentation accuracy is shown in Table 4.

Table 4. Ablation study on cnn encoder (BraTS 2019).

Model	Dice Score (%)		
	ET	TC	WT
3D-UNet	70.86	72.48	87.38
DualTrans (ours)	86.23	88.12	92.83

As can be seen from Table 4, the segmentation accuracy of our proposed model is greatly improved after adding the Swin Transformer encoder, and the dice coefficients increase by 16.63%, 16.64%, and 5.45% in the ET, TC, and WT brain-tumor regions, respectively. It is strongly proven that combining the Swin Transformer encoder path is of great significance for extracting global semantic feature information. A convolution operation is used to extract local dependencies and rich local features, and the improved Swin Transformer is used to learn global dependencies for global modeling. Then, feature fusion is carried out. The model can, therefore, obtain rich semantic information.

For the **DualTrans position,** the position information is included in the calculation of the multi-head attention mechanism, and the relative position information is further optimized by using trainable parameters. First, the relative position distance between each position abc ϵ $\breve{H}_{m\times m\times m}$ and the center point is defined as ijk in the window, and the relative distance is decomposed across dimensions. When attention is calculated within the range of the M × M × M local window, the $r^{v}_{a-i,b-j,c-k}$ vector is added, through which more position information can be obtained, and since they are shared between the attention heads between each layer, no additional parameters are introduced.

The BraTS 2019 validation set is a subset of the BraTS 2019 dataset that does not come with specific labels. It is used to evaluate the performance of algorithms in brain-tumor segmentation tasks. Segmentation results need to be uploaded to the official website for evaluation. Table 5 shows the segmentation accuracy of the model when position-free coding information, relative position coding, and our proposed DualTrans position coding are adopted into the BraTS 2019 verification set. It can be seen that the segmentation accuracy of the model decreases significantly when there is no position information. When the relative position coding information is used, the segmentation accuracy of the model is greatly improved. When our proposed DualTrans position information is used, the Dice coefficient of the relative position coding model is increased by 0.17%, 0.14%, and 0.18% in the ET, TC, and WT regions, respectively. The results show that adding the DualTrans location information in the M × M × M window is helpful for improving the accuracy of brain-tumor segmentation.

Table 5. Ablation study on DualTrans position self-attention (BraTS 2019).

Model	Dice Score (%)		
	ET	TC	WT
DualTrans (no position)	85.38	87.01	90.46
DualTrans (rel. position)	86.06	87.98	92.65
DualTrans (DualTrans position)	86.23	88.12	92.83

5. Conclusions

In this study, we propose a novel brain-tumor segmentation architecture DualTrans, which deeply integrates a convolutional neural network and the Swin Transformer. The model not only has the advantages of a 3D CNN for obtaining local semantic information but also uses the Swin Transformer to obtain global semantic information. Second, in the Swin Transformer encoder path, an innovative DualTrans position coding structure is proposed to incorporate the position information into the calculation of the multi-head attention mechanism and uses trainable parameters to further optimize the relative position information to further improve the segmentation accuracy of brain tumors. Finally, in order to solve the stability problem of the Swin Transformer, the output-activation values of each residual branch are normalized and merged back into the main branch so that the model has a better generalization performance. The experimental results on Brats 2019 and Brats 2021 have proven the effectiveness of the model. In future work, we will explore the use of "large convolution" to increase the convolutional receptor field and combine it with the Swin Transformer to develop a more efficient brain-tumor segmentation model.

Author Contributions: Author Z.L. wrote the main content of the paper and conducted experimental research. Author W.S. designed and planned the overall structure of the article. Authors Y.M. and Y.L. assisted in the experimental validation of the proposed model in the article. 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 61433012.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository. The original data presented in the study are openly available in [https://www.med.upenn.edu/cbica/brats2021/#Data2] at [https://doi.10.1109/TMI.2014.2377694], (accessed on 17 April 2024).

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

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*. Available online: https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (accessed on 17 April 2024). [CrossRef]
2. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arxiv* **2014**, arXiv:1409.1556.
3. 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; Proceedings, Part III 18; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
4. 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; Proceedings 4; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
5. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

6. Baid, U.; Ghodasara, S.; Mohan, S.; Bilello, M.; Calabrese, E.; Colak, E.; Bakas, S. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. *arXiv* **2021**, arXiv:2107.02314.
7. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Van Leemput, K. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging* **2015**, *34*, 1993–2024. [CrossRef] [PubMed]
8. Bakas, S.; Akbari, H.; Sotiras, A.; Bilello, M.; Rozycki, M.; Kirby, J.S.; Davatzikos, C. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Nat. Sci. Data* **2017**, *4*, 170117. [CrossRef] [PubMed]
9. BraTS Challenge Organizers. BraTS2019 Challenge Dataset [Dataset]. 2019. Available online: https://www.med.upenn.edu/cbica/brats-2019/ (accessed on 17 April 2024).
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (accessed on 17 April 2024).
11. Jiao, J.; Cheng, X.; Chen, W.; Yin, X.; Shi, H.; Yang, K. Towards Precise 3D Human Pose Estimation with Multi-Perspective Spatial-Temporal Relational Transformers. *arxiv* **2024**, arXiv:2401.16700.
12. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
13. Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. Swin transformer v2: Scaling up capacity and resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12009–12019.
14. Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-attention with relative position representations. *arXiv* **2018**, arXiv:1803.02155.
15. Huang, C.Z.A.; Vaswani, A.; Uszkoreit, J.; Shazeer, N.; Hawthorne, C.; Dai, A.M.; Eck, D. Music transformer: Generating music with long-term structure (2018). *arXiv* **2018**, arXiv:1809.04281.
16. Chu, X.; Tian, Z.; Zhang, B.; Wang, X.; Shen, C. Conditional Positional Encodings for Vision Transformers. *arXiv* **2021**, arXiv:2102.10882.
17. Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **2024**, *568*, 127063. [CrossRef]
18. Ramachandran, P.; Parmar, N.; Vaswani, A.; Bello, I.; Levskaya, A.; Shlens, J. Stand-alone self-attention in vision models. *Adv. Neural Inf. Process. Syst.* **2019**, *32*. Available online: https://proceedings.neurips.cc/paper_files/paper/2019/file/3416a75f4cea9109507cacd8e2f2aefc-Paper.pdf (accessed on 17 April 2024).
19. Kamnitsas, K.; Ledig, C.; Newcombe, V.F.; Simpson, J.P.; Kane, A.D.; Menon, D.K.; Glocker, B. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* **2017**, *36*, 61–78. [CrossRef] [PubMed]
20. Isensee, F.; Jäger, P.F.; Full, P.M.; Vollmuth, P.; Maier-Hein, K.H. nnU-Net 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 2020; Revised Selected Papers, Part II 6; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 118–132.
21. Luu, H.M.; Park, S.H. Extending nn-UNet for brain tumor segmentation. In Proceedings of the International MICCAI Brainlesion Workshop, Virtual, 27 September 2021; Springer International Publishing: Cham, Switzerland, 2021; pp. 173–186.
22. Wang, W.; Chen, C.; Ding, M.; Yu, H.; Zha, S.; Li, J. Transbts: Multimodal brain tumor segmentation using transformer. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021; Proceedings, Part I 24; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 109–119.
23. Xu, X.; Zhao, W.; Zhao, J. Brain tumor segmentation using attention-based network in 3D MRI images. In Proceedings of the Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, 17 October 2019; Revised Selected Papers, Part II 5; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 3–13.
24. Zhao, G.; Zhang, J.; Xia, Y. Improving brain tumor segmentation in multi-sequence MR images using cross-sequence MR image generation. In Proceedings of the Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, 17 October 2019; Revised Selected Papers, Part II 5; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 27–36.
25. Jiang, Z.; Ding, C.; Liu, M.; Tao, D. Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In Proceedings of the Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, 17 October 2019; Revised Selected Papers, Part I 5; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 231–241.
26. 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; Proceedings, Part II 19; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 424–432.

27. Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H.R.; Xu, D. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In Proceedings of the International MICCAI Brainlesion Workshop, Virtual, 27 September 2021; Springer International Publishing: Cham, Switzerland, 2021; pp. 272–284.
28. Pei, L.; Vidyaratne, L.; Monibor Rahman, M.; Shboul, Z.A.; Iftekharuddin, K.M. Multimodal brain tumor segmentation and survival prediction using hybrid machine learning. In Proceedings of the Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, 17 October 2019; Revised Selected Papers, Part II 5; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 73–81.
29. Peiris, H.; Chen, Z.; Egan, G.; Harandi, M. Reciprocal adversarial learning for brain tumor segmentation: A solution to BraTS challenge 2021 segmentation task. In Proceedings of the International MICCAI Brainlesion Workshop, Virtual, 27 September 2021; Springer International Publishing: Cham, Switzerland, 2021; pp. 171–181.
30. Jia, Q.; Shu, H. Bitr-unet: A cnn-transformer combined network for mri brain tumor segmentation. In Proceedings of the International MICCAI Brainlesion Workshop, Virtual, 27 September 2021; Springer International Publishing: Cham, Switzerland, 2021; pp. 3–14.
31. Yuan, Y. Evaluating scale attention network for automatic brain tumor segmentation with large multi-parametric MRI database. In Proceedings of the International MICCAI Brainlesion Workshop, Virtual, 27 September 2021; Springer International Publishing: Cham, Switzerland, 2021; pp. 42–53.
32. Pawar, K.; Zhong, S.; Goonatillake, D.S.; Egan, G.; Chen, Z. Orthogonal-Nets: A Large Ensemble of 2D Neural Networks for 3D Brain Tumor Segmentation. In Proceedings of the International MICCAI Brainlesion Workshop, Virtual, 27 September 2021; Springer International Publishing: Cham, Switzerland, 2021; pp. 54–67.
33. Cai, X.; Lou, S.; Shuai, M.; An, Z. Feature learning by attention and ensemble with 3d u-net to glioma tumor segmentation. In Proceedings of the International MICCAI Brainlesion Workshop, Virtual, 27 September 2021; Springer International Publishing: Cham, Switzerland, 2021; pp. 68–79.
34. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer Nature Switzerland: Cham, Switzerland, 2022; pp. 205–218.

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.