Received 28 March 2023, accepted 24 April 2023, date of publication 1 May 2023, date of current version 5 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3272055

RESEARCH ARTICLE

DenseTrans: Multimodal Brain Tumor Segmentation Using Swin Transformer

LI ZONGREN[1], **WUSHOUER SILAMU**[1], **WANG YUZHEN**[2], **AND WEI ZHE**[2]

[1]College of Information Science and Engineering, Xinjiang University, Ürümqi 830047, China

[2]Information Research and Development Center, 940th Hospital of the PLA Joint Logistic Support Force, Lanzhou 730050, China

Corresponding author: Wushouer Silamu (lzr18993156814@163.com)

This work was supported by the National Natural Science Foundation of China under Grant 61433012.

ABSTRACT Aiming at the task of automatic brain tumor segmentation, this paper proposes a new DenseTrans network. In order to alleviate the problem that convolutional neural networks(CNN) cannot establish long-distance dependence and obtain global context information, swin transformer is introduced into UNet++ network, and local feature information is extracted by convolutional layer in UNet++. then, in the high resolution layer, shift window operation of swin transformer is utilized and self-attention learning windows are stacked to obtain global feature information and the capability of long-distance dependency modeling. meanwhile, in order to alleviate the secondary increase of computational complexity caused by full self-attention learning in transformer, deep separable convolution and control of swin transformer layers are adopted to achieve a balance between the increase of accuracy of brain tumor segmentation and the increase of computational complexity. on BraTs2021 data validation set, model performance is as follows: the dice dimilarity score was 93.2%,86.2%,88.3% in the whole tumor,tumor core and enhancing tumor, hausdorff distance(95%) values of 4.58mm,14.8mm and 12.2mm, and a lightweight model with 21.3M parameters and 212G flops was obtained by depth-separable convolution and other operations. in conclusion, the proposed model effectively improves the segmentation accuracy of brain tumors and has high clinical value.

INDEX TERMS Brain tumor segmentation, convolutional neural networks, swin transformer, UNet++.

I. INTRODUCTION

As one of the tumors with high fatality rate, brain tumor has become an important factor endangering human life and health. from the perspective of tracing, brain tumors are usually divided into primary tumors and secondary tumors. primary tumors refer to the tumors initially appearing in the intracranial, which originate from the central nervous system and originate from intracranial neuroepithelial tissue, meningeal tissue cells and pineal cells. secondary brain tumors are relative to primary brain tumors. secondary brain tumors originate from the lungs, digestive tract, mammary gland, uterus and other organs of the body, and metastasize from other organs to the brain, usually with a relatively high degree of malignancy. gliomas are the most common form of brain tumor, caused by cancerous glial cells in the brain and spinal cord. gliomas are classified into low grade (LGG) and high grade (HGG) subtypes [1]. high grade gliomas are aggressive, grow rapidly, have poor survival prognosis, and usually require surgery and radiotherapy. as a reliable diagnostic tool, magnetic resonance imaging (MRI) can realize human examination by radiating energy signals from internal substances to the surrounding environment through high-frequency magnetic field in vitro, which plays an important role in the analysis and detection of brain tumors. there are usually four common 3D modes T1-weighted(T1), T1-weighted with gadolinium contrast enhancement(T1-Gd or T1c), T2-weighted(T2), and Fluid Attenuated Inversion Recovery (FLAIR). different MRI modes can effectively complement each other and fully subdivide the tumor in related areas, thus effectively improving the accuracy of segmentation. as can be seen from Figure 1, MRI data of different morphologies captured different pathological features of tumors.

MRI is the primary method of clinical detection of brain tumors. segmentation of brain tumor regions from

The associate editor coordinating the review of this manuscript and approving it for publication was Liangxiu Han.

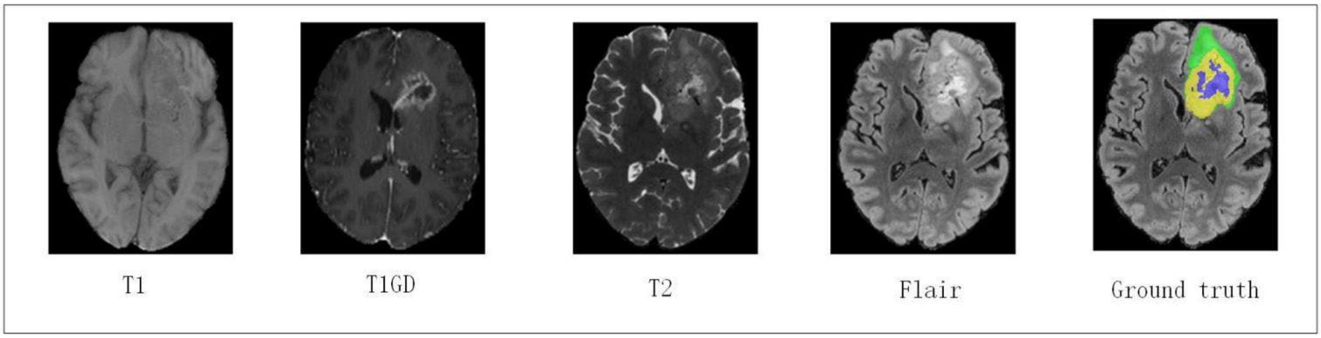

FIGURE 1. An example of multimodal MRI volumes for brain tumor segmentation. The green, blue, and yellow regions in the ground truth indicate edema (ED), non-enhancing tumor and necrosis (NCR/NET), and enhancing tumor (ET), respectively.

multimodal MRI images is helpful for treatment examination, post-diagnosis monitoring and effect evaluation of patients [2]. However, the segmentation of brain tumors in MRI was performed manually by experienced radiographers in the past, which is time-consuming and may lead to inconsistent segmentation results, because artificial segmentation mainly depends on experience. the same medical image is segmented by different technicians with different results. therefore, many researchers try to solve this problem by using the method of computer aided diagnosis to achieve semi-automatic segmentation. with the rapid development of machine learning technology, various automatic segmentation methods for brain tumors emerge in an endless stream, while traditional segmentation methods based on threshold, edge detection, clustering, region and registration have gradually faded out of people's attention due to their high complexity and low segmentation accuracy [3]. machine learning algorithms based on feature selection and classification, such as random forest, adaboost, kmeans clustering, support vector machines, etc., still have limited segmentation performance. moreover, automatic brain tumor segmentation remains a challenge due to extreme intrinsic heterogeneity in appearance, shape, and histology.

In order to solve the above problems, and with the vigorous development of deep learning technology, researchers began to use computer deep learning technology to segment brain tumors, and has achieved obvious advantages in segmentation accuracy. Since the birth of Fully Convolutional Networks(FCN) architecture [4] and UNet architecture [5] in 2015. Due to their excellent encoder-decoder architecture, FCN and UNet have become increasingly popular in the field of brain tumor segmentation [6], [7], [8]. Myronenko et al. [9] proposed a multi-modal semantic segmentation method for 3D brain tumors, which followed the UNet encoder decoder architecture, and added variable autoencoder (VAE) branches to the network to reconstruct input images together with the segmentation, so as to regularise the shared encoder [40]. Jiang et al. [10] designed a new two-level cascade UNet to segment brain tumors. the substructure of brain tumors was trained end-to-end from coarse to fine. after that, the crude segmentation map and the original image are input into the second stage UNet, through which a more accurate segmentation map with more network parameters can be provided. In recent years, Isensee et al. [11] applied nnU-Net network to brain tumor segmentation, and made specific modifications by integrating brain tumor segmentation, including post-processing, region-based training, etc., effectively improving the accuracy of brain tumor segmentation. Luu et al. [12] applied the extended nnU-Net network to brain tumor segmentation, improved on the basis of nnU-Net, replaced batch normalization with group normalization, and improved nnU-Net by using axial attention in the decoder, which further improved the accuracy of brain tumor segmentation.

However, because the current segmentation methods based on brain tumors mostly rely on CNN and its variants, although CNN has achieved excellent performance, it cannot learn global and remote semantic information interaction well due to the locality of convolutional operation [13], [14], [15], lacks the ability to model long-term dependencies explicitly. later, although some researchers have introduced the brilliant transformer architecture in the field of Natural Language Processing(NLP) into the field of image segmentation, transformer and its variants, such as vision transformer, require a large number of data sets for pre-training. However, the lack of current medical image data sets prevents Transformer from further deepening in the field of medical image segmentation.Under the scarcity restriction, many models may under-perform in capturing meaningful patterns in the data [44].

In view of this, this paper proposes a network architecture based on DenseTrans, which can effectively solve the above problems by combining the improvement of UNet++ and swin transformer. the main contributions of this paper are as follows:

(1) Combine swin transformer with the improved UNet++ network innovatively. firstly, extract features through CNN encoder, and then transfer extracted features into swin transformer through patch embedding in the high-resolution layer of UNet++. use the swin transformer layer to learn long-range dependencies and global context information. moreover, due to the introduction of swin transformer,

long-term dependency modeling using transformer no longer requires a large number of medical image data sets for pre-training, which solves the problem that global dependency modeling cannot be carried out when medical image data sets are scarce.

(2) In order to make full use of the information of each mode in MRI, T1-weighted(T1), T1-weighted with gadolinium contrast enhancement(T1-Gd or T1c), T2-weighted(T2), and Fluid Attenuated Inversion Recovery (FLAIR) modes were used for pixel level fusion. the predictive ability of each mode to Enhancing Tumor(ET), Tumor Core(TC) and Whole Tumor is explored to better perform pixel-level classification.

(3) In order to alleviate the increasing complexity of network structure and the increasing number of parameters and computational complexity, deep separable convolution was introduced into DenseTrans network architecture, which effectively improved the efficiency of image segmentation and reduced the number of model parameters and computational complexity.

(4) Further optimize the training model through deep supervision, assist the model to conduct pixel-level classification of brain tumors, and improve the segmentation accuracy of brain tumors. More importantly, this paper uses deep supervision to carry out hierarchical pruning of the model, so that the computational complexity of the deep network with a large number of parameters can be greatly reduced within the acceptable accuracy range. Experimental results on the BraTS 2021 and BraTS 2020 data sets demonstrate the effectiveness of the method.

The specific sections of the article are arranged as follows: the second section introduces the relevant work of the article, and the third section elaborates it in detail The improvement method and the concrete implementation process proposed in this paper, the fourth section through the experiment to verify the proposed method, the fifth section of this paper is summarized.

II. RELATED WORK

A. UNet AND ITS VARIANTS

Since the UNet network was proposed in 2015, it has effectively improved the accuracy of brain tumor segmentation. In addition, the original purpose of UNet network is to solve the problem of medical image segmentation. This network is mainly composed of Encoder layer, decoder layer and skip connection layer. Encoder layer consists of 3 × 3 convolution layer, Rectified Linear Unit(RELU) layer and maxPooling layer. feature extraction was carried out by subsampling. The decoder layer consists of the transposed convolution layer and Rectified Linear Unit(RELU) to form the upper sampling layer. skip connection concat the features of encoder and decoder, providing multi-scale and multi-level information for the later image segmentation, effectively alleviating the problem of space loss caused by downsampling in the structure of single encoder, thus improving the accuracy of image segmentation. traditional UNet designed a four-layer structure, but the researchers speculated that a three-layer or five-layer structure for different data sets could also improve the segmentation accuracy. Zhou et al. [16] proposed the UNet++ network, designed a dense jump connection layer on the basis of UNet, introduced a built-in UNet set of different depths, thus improving the segmentation performance of objects of different sizes, which is equivalent to integrating several UNet networks to train image segmentation at the same time. by pruning the model of UNet ++ through deep supervision, the pruned UNet ++ model achieved significant acceleration, but the performance was only slightly decreased. Milletari et al. [17] proposed a kind of UNet network VNet, which introduced residual connections in the contraction layer and the expansion layer, optimized the convolutional neural network through residual connections, and used the convolution layer to replace the pooling layer during downsampling. VNet replaced the pooling layer with the convolution layer to further optimize the weights. Finally, the network proposed the dice loss function of dynamic adjustment. when the sample categories were unbalanced, the formula was used to adjust the weights dynamically, and there was no need to reweight the samples during training [43]. the sample imbalance is improved. Huang et al. [18] believed that the skip connection layer of UNet++ fused the low-level features of the encoder and the high-level features of the decoder, which would lead to feature loss. therefore, the UNet3+ network with full-scale feature fusion was proposed. the features of the decoder are fused with the lower and sibling features of the encoder and the higher features of the decoder. Qin et al. [19] proposed an improved UNet3+ network. In the encoder stage, phase residual network was adopted to replace the original convolution layer, which improved the performance of network feature extraction to a certain extent and avoided the phenomenon of gradient disappearance. In addition, the batch normalization layer is replaced by the Filter Response Normalization(FRN) layer [44], and the impact of batch size on the network is avoided.

B. TRANSFORMER AND ITS VARIANTS IN MEDICAL IMAGE SEGMENTATION

Although CNN has achieved great success in the field of medical image segmentation, due to the limitation of receptive field in convolution operation, these methods are unable to establish long-range dependence and global context connection. In view of this, some researchers have introduced the transformer architecture in the field of Natural Language Processing(NLP) to medical image segmentation. Vaswani et al. [20] believe that Recurrent Neural Networks(RNN) and other models are difficult to realize parallelization, and Recurrent Neural Networks(RNN) is difficult to remember long distance features. In view of this, Transformer model is proposed, which uses weighted attention to make the model can see all inputs. The multi-head attention mechanism is introduced in order to project onto multiple Spaces to obtain multi-scale features similar to the channels of convolutional neural networks. In the training process, a group of query functions with multiple attention headers are calculated at the same time, and then they are encapsulated in the matrix Q, the functions matched by Q are encapsulated in the matrix K, the corresponding key value is in V, $\mathrm{d_k}$ represents the matrix dimension, and the output calculation matrix is:

$$\mathrm{Attention(Q_i, K_i, V_i) = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i} \tag{1}$$

Long distance dependencies and global context information can be effectively captured through the Transformer model, but the Transformer is primarily used in the NLP domain. Dosovitskiy et al. [21] applies transformer to the field of computer vision and puts forward a Vision transformer model, which divides the input image into patch blocks and uses patch blocks to classify Transformer. patch blocks are taken as 1D input sequence and corresponding position information is added. medical images are classified at pixel level by transformer encoder and MLP layer. however, this model requires a large number of data sets for pre-training, but medical image data sets are scarce. Liu et al. [22] proposed a new architecture based on transformer, swin transformer. the model is a layered transformer, and its representation is shifted by the shifted window. the shifted window scheme improves efficiency by limiting self-focused calculations to non-overlapping local windows while also allowing cross-window connections. Cao et al. [23] proposed a pure swin transformer model of UNet. the model draws on UNet and includes encoder, decoder, jump connection, connection layer, etc. the multi-scale features are extracted by swin transformer from the encoder. the patch expanding module is creatively proposed to increase the image resolution and reduce the feature channel at the decoder layer, thus improving the image segmentation accuracy. Wu et al. [24] proposes a transformer model based on 3DUNet architecture, which proposes a new local attention (LSM) and global attention mechanism (GSM). GS-MSA mimics the vacuous convolution model, and selects a patch every fixed distance to form a global attention unit, and the rest to form a unit, so as to extract global feature information. Hatamizadeh et al. [25] introduced swin transformers into the UNet network, used swin transformers as the encoder to extract features, and then input the extracted features into the convolutional network to restore images through upsampling. The encoder features are passed to the decoder through jump connections during upsampling, and residual blocks are utilized at each layer of the decoder.

III. METHOD

A. OVERALL ARCHITECTURE OF DenseTrans

Although some researchers combine UNet with Transformer in an attempt to solve the problem that CNN cannot establish long-distance dependency and extract global context information, directly combining CNN with Transformer cannot achieve the expected effect. the transformer model and its variants have a weak ability to extract local information and shallow features, and transformer requires a large number of data sets for pre-training.

As shown in Figure 2, DenseTrans is an improved dense hybrid model of UNet++ and swin transformer. DenseTrans contains encoder,decoder dense jump connection layer. first of all, the MRI section $X \in R^{H\times W\times D\times C}$, which H×W represents spatial resolution, D represents dimension, and C represents the number of channels. firstly, 3D CNN is used to extract local shallow features and context information. since transformer cannot directly flatline pixels into 1D for attention calculation, it needs to calculate the attention weight after Patch of the input image, and partial local features will be ignored when patch is used as a unit for calculation. therefore, for shallow features, 3D CNN is still used for feature extraction. after two downsampling with 3D CNN, the obtained features are then introduced into swin transformer for long-distance dependency modeling and global context information acquisition. the features processed by swin transformer encoder are converted into 3D dimensions through the feature mapping layer. finally, the spatial resolution of the features is restored through the decoder layer, and the pixel-level classification is performed by the SoftMax layer to generate segmentation results. at the same time, the improved UNet++ model is adopted in the model, which is different from the traditional UNet, which generally adopts four times of downsampling for spatial resolution restoration. after downsampling twice with 3D CNN, swin transformer blocks are added to each layer. 3D CNN and swin transformer are used to capture features of different layers, connect them by superposition, and integrate more shallow UNet. The scale difference of the feature map during fusion is smaller, and deep supervision is added to each shallow UNet output, so that the complex depth network can greatly reduce the number of parameters within the acceptable accuracy range. More importantly, in order to reduce the computational complexity and the number of parameters, we use depth-separable convolution when using CNN downsampling.

B. NETWORK ENCODER

Different from pure swin transformer [23] to extract features by layered transformer construction of encoder, we took an MRI section of the input, the use of 3D CNN for two down sampling to extract the local characteristics of shallow, alleviate the problem of weak extraction of local feature information in transformer. we adopted depth-separable convolution to carry out the convolution operation. first, the depth information was separated by 3 × 3×1 convolution, and then the channel fusion was carried out by 1 × 1×1 convolution. such operation can reduce the number of model training parameters and reduce the computational complexity. after convolution, the maximum pooling is used for sampling reduction, and the input image is gradually encoded into low-resolution/high-order feature representation $G \in R^{K\times\frac{H}{16}\times\frac{W}{16}\times\frac{D}{16}}$ (the last layer of Encoder K=256). In this case, G is integrated with rich local context information. the integrated information is then passed into swin transformer to learn the long distance dependencies and global context information.

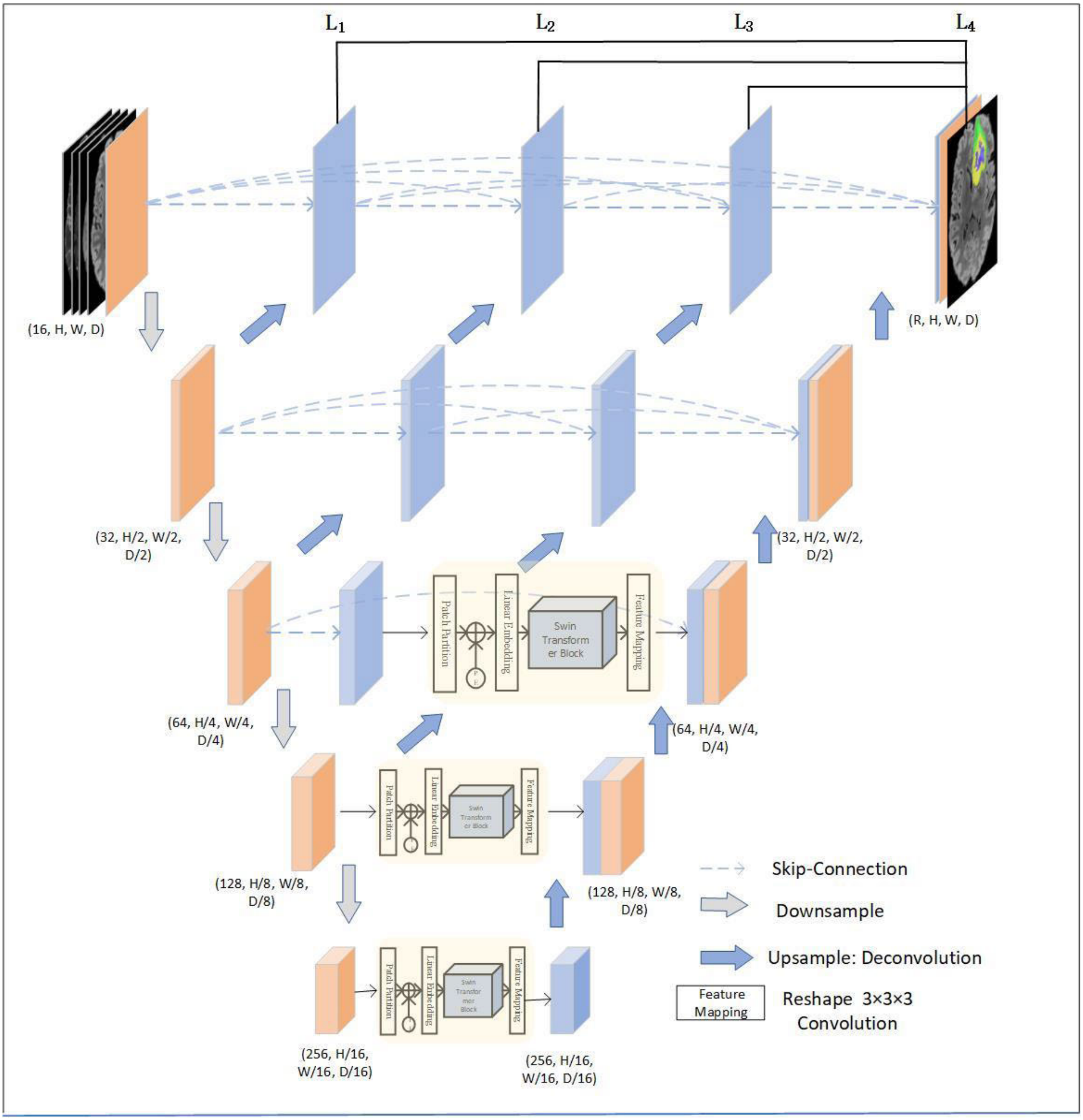

FIGURE 2. Overall architecture of the proposed DenseTrans.

Swin Transformer Block: The model uses swin transformer instead of transformer or vision transformer because the latter requires a large data set for pre-training. some improvements based on traditional techches transformer, first using the patch layer partition the input image into a patch block, to convert the characteristics of the channel, through the linear embedding layer after converted into $X \in R^{H\times W\times D\times K}$, unlike traditional swin transformer, patch merging of image resolution and adding feature channels is not used in the following months. because the model in this paper has already layered the input image, subsampling and layered feature extraction do not need to be done again. after swin transformer core W-MSA, its essence is to carry out transformer in a fixed window, and its computational complexity is as follows:

$$\Omega(\mathrm{W-MSA}) = 4hwD^2 + 2hwM^2D \tag{2}$$

W-MSA makes transformer for the local information within the small window, but it still needs to obtain the global information. SW-MSA is used to obtain the information between the windows, and then the windows are moved in

SW-MSA. cyclic shift and mask are used to make transformer for the pixels within the moving window. In this way, the pixel information between the surrounding windows is obtained, and the global context information is obtained along with the movement of the window, and the long-distance dependence relationship is established. as the window moves, swin transformer is calculated as:

$$\overset{\wedge}{\mathrm{z}}l = \mathrm{W} - \mathrm{MSA}(\mathrm{LN}(z^{l-1})) + z^{l-1} \tag{3}$$

$$z^{l} = \mathrm{MLP}(LN(\overset{\wedge}{\mathrm{z}}l)) + \overset{\wedge}{\mathrm{z}}l \tag{4}$$

$$\overset{\wedge}{\mathrm{z}}l + 1 = \mathrm{SW} - \mathrm{MSA}(\mathrm{LN}(z^{l})) + z^{l} \tag{5}$$

$$z^{l+1} = \mathrm{MLP}(\mathrm{LN}(\overset{\wedge}{\mathrm{z}}l + 1)) + \overset{\wedge}{\mathrm{z}}l + 1 \tag{6}$$

LN represents the layer normalization, and z^{l} represents the swin transformer output of Layer L. the algorithm flow is shown in Figure 3.

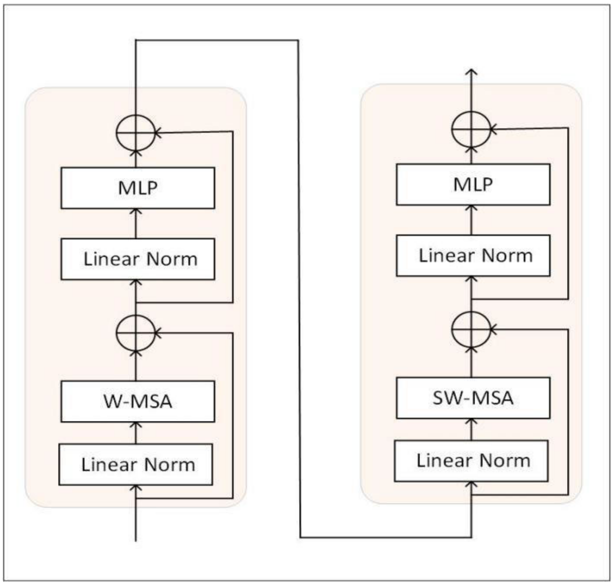

FIGURE 3. Swin transformer block.

Relative Position Bias: swin transformer differs from vision transformer in the use of non-overlapping patches, which reduces the computational load and complexity. however, when W-MSA fixes the local window, it cannot learn the global features. The information between SW-MSA learning windows was introduced, which led to difficulties in position learning. then, relative position coding was used to learn relative position information, and the following was followed when calculating similarity and self-attention:

$$\mathrm{Attention(Q, K, V)} = \mathrm{SoftMax}(QK^{T}\sqrt{d} + T)V \tag{7}$$

where, Q, K and V respectively represent query coefficients, corresponding key and value pairs queried, $Q, K, V \in R^{M^2 \times M^2}$ and M^2 represent the number of patches in the input image, and represent relative position offset, $T \in R^{M^2 \times M^2}$, which is used to represent the relative position information between patches.

C. NETWORK DECODER

Corresponding to the encoder is the symmetric decoder of the DenseTrans block, in order to get the segmentation result of the original input image (H×W×D). We use 3D CNN transpose convolution in the decoder for upsampling and pixel-level segmentation of the extracted depth features. Secondly, the model class in this paper integrates multiple UNet, so depth supervision is added to each layer of UNet, so that the depth network with a huge number of parameters can reduce the number of parameters significantly within the acceptable accuracy range.

Feature Mapping. encoder extracts features, CNN is used for subsampling, and then swin transformer is used for global feature learning. before learning the features extracted by CNN, swin transformer will first input the feature graph PatchEmbedding, so the features processed by swin transformer cannot be directly used for upsampling of 3D CNN. a feature mapping module is designed in this paper. $Z^{L} \in R^{d \times N}$ of swin transformer is mapped to $X_{int} \in R^{d \times \frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}}$, and then the features processed by feature mapping are used for up-sampling.

1) COMBINED FEATURE UPSAMPLING

Each level of UNet in the model performs up-sampling after feature mapping is completed. during the up-sampling process, concatenate the features after transposed convolution and the features transmitted by jump connections, so as to further improve the segmentation accuracy and obtain more abundant semantic information of global context. finally, pixel-level segmentation is performed by SoftMax.

2) DEEP SUPERVISION

Deep Supervision is added into the model, aiming to make the complex and redundant deep network significantly reduce the number of parameters within the acceptable accuracy range.1 × 1 convolution kernel is added to each level of supervision after L_1,L_2,L_3,L_4, which is equivalent to supervising the output of each level or each UNet and analyzing the output results, so as to reduce the number of model parameters and computational complexity within an acceptable range.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. DATASET AND PRE-PROCESSING

The model uses BraTS 2021 as the benchmark data set to demonstrate the proposed method. the BraTS 2021 dataset, provided by the Brain Tumor Segmentation (BraTS) challeng, contained MRI scans from a total of 2000 patients. among them, there were 1251 cases of training set, 219 cases of verification set and 530 cases of test set. The training set contained the original image and the corresponding annotations. the verification set was used to adjust the model but did not provide corresponding annotations. users need to upload data segment to https://ipp.cbica.upenn.edu/ for assessment. each training sample contains data of four modes: they are native T1-weighted(T1), T1-weighted with gadolinium contrast enhancement(T1-Gd or T1c), T2-weighted(T2), and Fluid Attenuated Inversion Recovery (FLAIR). All data sets were manually segmented by one to four raters following the same protocol, and their markings were approved by experienced board-certified neuroradiologists. Annotations included four categories: GD enhances tumor (ET-label 4), peritumoral edema/infiltrating tissue (ED-label 2), and necrotic tumor core (NCR-label 1), background (label 0). since the annotations of verification set and test set are not published, the training set (1251) of BraTS 2021 data set is divided into training set and test set in the training stage. the dimension of each sample is 240 × 240×155, in which there are most background voxels. we cut the sample into 128 × 128×128, and use contrast processing, noise reduction processing and other pretreatment methods. the second 3D data set for experimental verification is the BraTS 2020 data set, which is similar to the BraTS 2021 data set, both of which are data sets of MICCAI brain tumor segmentation competition. the training set of this data set contains 369 cases, and the validation set contains 125 cases. the validation set and the test set are used for online evaluation without publishing corresponding annotations.

B. IMPLEMENTATION AND EVALUATION METRICS

The method adopted in our experiment is consistent with most previous experiments in the field. In the training stage, the training set is divided according to 8:2 to conduct model training and adjustment, and the verification set is used to evaluate the model performance in the Inference stage.the proposed model is run in PyTorch framework with 8 NVIDIA RTX A5000 graphics cards (each with 24G of memory) for 7000 epochs using a batch size of 12. For the optimizer, we set the adam optimizer with an initial learning rate of 0.0004. during optimization, the initial rate decays by a power of 0.8 in each iteration. in the processing process, images of the four modes of MRI were combined into a 4D image (C×H×W×D), where C=4. In addition, the following data enhancement techniques are applied in the processing process: (1) the original image (240 × 240×155) is randomly cropped to (128 × 128×128); (2) the image is simply rotated at the Angle {90,180,270}; (3) Contrast processing and Gaussian denoising.

1) IMPLEMENTATION DETAILS

Our model is trained from scratch in PyTorch, using DenseTrans Net as a split network and adding a swin transform to the encoder's skip connections at a high level stage. In DenseTrans Net, the contraction path has five layers, including bottleneck, each Layer is composed of 3 × 3 × 1 deep convolution and 1 × 1 × 1 channel convolution as well as Layer Normalization and reLu activation. the number of feature channels set in the the first encoder is 16. then, the maximum pooling layer of 2 × 2×2 is used for downsampling, and the stride is 2.the number of channels is set to 32 in the second encoder, and the number of channels at the third and fourth layers as well as bottleneck layers is doubled in turn. in the skip connections, features are flattened through the patch embedding layer, and then swin transform is used to learn long-term dependency and global context information. at the decoder stage, feature mapping layer is used to convert the features processed by swin transform into $X_{int} \in R^{d\times\frac{H}{16}\times\frac{W}{16}\times\frac{D}{16}}$. after that, trilineal interpolation followed by 3 × 3×1 deep convolution and 1 × 1×1 channel convolution are used for up-sampling. the last layer is composed of 1 × 1×1 convolution, and the number of output channels is 3. The segmentation result is then generated.

2) EVALUATION METRICS

Previously, the commonly used evaluation metrics for brain tumor segmentation include (1) the Dice similarity coefficient (DSC): DSC is used to measure the overlap between the segmentation contour obtained by the proposed tumor segmentation method and the manual contour described by experienced doctors.

$$DSC = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{8}$$

(2) the Hausdorff distance (HD): The surface area difference between the segmented profile and the manual profile was measured. HD is more sensitive to boundary segmentation and is used to assess the maximum difference between the surface area of the segmented contour S and the corresponding manual contour M.

$$\begin{aligned} &Hausdorff\,95distance \\ &\quad = P_{95}\left\{Sup_{x\in X}\ d(x, Y), Sup_{y\in Y} d(X, y)\right\} \end{aligned} \tag{9}$$

In this paper, Dice and Hausdorff distance are used to evaluate the segmentation results of the proposed model. to some extent, Hausdorff distance is the complement of Dice coefficient, which can measure the maximum distance between two contour edges. raise the weight of outliers and punishes outliers.therefore, the combination of Dice coefficient and Hausdorff distance is noisier and has stronger generalization and robustness than pure Dice metrics.

3) OPTIMIZER

In order to obtain the global minimum in the model training stage, we conducted various experiments in the optimization of the back propagation loss function, including optimization algorithms such as stochastic gradient descent (SGD), Adam and Momentum. Initially, we tried the SGD optimizer, but because it was sensitive to the learning rate of hyperparameters, If the learning rate is too small, the convergence speed will be too slow, and if the learning rate is too large, the extreme point will be crossed, and the algorithm will be easily stuck at the saddle point in the iterative process. therefore, in the end, we choose Adam optimizer, which is an optimizer algorithm with momentum term, and uses gradient first-order matrix estimation and second-order matrix estimation to dynamically adjust the learning rate of parameters. since the learning rate of each iteration of Adam has a certain

TABLE 1. Comparison of the proposed model with the classical brain tumor segmentation method on the BraTS 2021 data validation set.

References	Whole Tumor(WT)		Tumor Core(TC)		Enhancing Tumor(ET)		Average	
	Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff
Adversarial learning[26]	0.908	5.37	0.854	8.56	0.814	21.83	0.859	11.92
Extending nn-UNet[12]	0.928	3.47	0.88	7.62	0.845	20.73	0.884	10.61
Swin Unter[25]	0.926	5.83	0.885	3.77	0.858	6.02	0.889	5.21
Bitr-unet[27]	0.910	4.51	0.843	16.69	0.819	17.85	0.857	13.02
Atrous Convolutions[28]	0.891	11.78	0.807	21.17	0.780	25.82	0.826	19.59
Attention Mechanism[29]	0.902	6.16	0.816	16.65	0.769	30.21	0.829	17.67
MS UNet[30]	0.919	-	0.863	-	0.824	-	0.869	-
Orthogonal-Nets[31]	0.914	5.43	0.850	9.81	0.832	20.97	0.865	12.07
HNF-Netv2[27]	0.925	3.46	0.880	5.86	0.848	14.18	0.884	7.83
Multi-plane UNet++[32]	0.906	4.54	0.835	10.11	0.792	16.61	0.844	10.42
Our(DenseTrans)	0.932	4.58	0.862	14.8	0.883	12.2	0.89	10.53

range, the parameters are relatively stable, and the step size annealing process can be naturally realized, so it is more suitable for large-scale data scenarios.

$$g_t \leftarrow \nabla_w f_t(w_{t-1}) \tag{10}$$

we first calculate the gradient, and the formula is shown in (10), where g represents the gradient and f represents the noisy objective function. our goal is to calculate the expected value of the function f(w). after that, we update biased first moment estimate.

$$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t \tag{11}$$

β_1 coefficient is exponential decay rate and controls weight distribution.After that, we update biased second raw moment estimate, and calculate the first order moment estimate and second order matrix estimate of deviation correction respectively, and finally update the parameters.

$$w_t \leftarrow w_{t-1} - \alpha \cdot \hat{m} t(\sqrt{\hat{v}} t + \varepsilon) \tag{12}$$

where, $\hat{m} t$ represents the first moment estimate after calculating the deviation, and $\hat{v} t$ represents the second moment estimate after calculating the deviation.we set the initial learning rate to 0.0004, and the initial rate decays 0.8 powers in each iteration.

C. RESULT AND COMPARISONS

1) BraTS 2021

The model proposed by us is different from the ordinary UNet model and also fundamentally different from the pure transform model. In Table 1, the DenseTrans model proposed by us is compared with excellent models in recent years. the segmentation accuracy of our model is 93.23%(DSC) and 4.58(HD) on WT, 86.2%(DSC) and 14.8(HD) on TC, and 88.3%(DSC) and 12.2(HD) on ET. comparing our proposed method with the Brats2021 Challenge champion model method (Extending nn-UNet), our segmentation accuracy on WT is improved by 0.4%, that on TC is reduced by 1.8% and that on ET is improved by 3.8%. on another evaluation metrics, Hausdorff distance metric, our model is better

TABLE 2. Comparison of the proposed model with the classical brain tumor segmentation method on the BraTS 2020 data validation set.

References	Whole Tumor(WT)		Tumor Core(TC)		Enhancing Tumor(ET)		Average	
	Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff
TransBTS[33]	0.901	4.96	0.817	9.77	0.787	17.95	0.835	10.89
3DU-net[34]	0.886	6.67	0.843	19.55	0.785	20.36	0.838	15.53
Dual-Path UNet[35]	0.879	7.45	0.779	28.78	0.752	16.95	0.803	17.73
Attention-Based Fusion[36]	0.876	6.45	0.824	20.23	0.773	27.17	0.826	17.95
Modality-Pairing [33]	0.891	6.24	0.842	19.54	0.816	17.79	0.850	14.52
Vox2Vox[37]	0.872	6.44	0.811	24.36	0.787	18.95	0.823	16.58
Scale Attention[38]	0.904	5.49	0.842	8.34	0.785	20.35	0.844	11.39
Our(DenseTrans)	0.914	6.32	0.853	16.9	0.823	15.2	0.86	12.81

than Extending nn-UNet model. Extending nn-UNet model integrates automatic segmentation of a variety of models, and axial attention mechanism is added to the Decoder process, which makes the model have an obvious effect on the accuracy of TC segmentation. our model obviously outperforms the Extending nn-UNet model on both WT and ET, demonstrating the benefits of applying the swin transform of the attention mechanism to establish global dependencies. compared with the classical Swin Unter model, this model used Swin Transform to construct the Unet-like model for brain tumor segmentation, which played a leading role in the combination of UNet and swin transform. we improved the segmentation accuracy by 0.4% on WT, 2.3% on TC, and 2.5% on ET. on The whole, the DenseTrans model proposed by us has achieved good results in terms of the Dice scores, and improved the accuracy on both WT and ET. compared with the Extending nn-UNet model and Swin Unter model, On TC, our segmentation accuracy decreased slightly, mainly because the model set a lower weight for T1GD mode when the initial input original image was used for multi-mode fusion, and T1GD mode was more suitable for detecting TC region. In terms of Hausdorff distance metric, the DenseTrans also demonstrated better performance. the experimental results show that combining swin transform with improved UNet++ is helpful for long-term dependency modeling and global context information acquisition.

2) QUANTITATIVE ANALYSIS

The segmentation results of our model are compared with the latest segmentation results, and the quantitative comparison results are shown in table 5. since our model combines CNN and Transformer for brain tumor segmentation, we will compare them with models using CNN and Transformer alone. using the same data set and the same input modes, all methods were compared and performance was quantitatively assessed. compared with the classical 3DUNet in convolutional neural networks, our original model has a great improvement in performance, but also a significant increase in computational complexity. however, when the number of layers was reduced from L4 to L3 by combining deep separable convolution and pruning assisted by deep supervision, our computational complexity was significantly reduced. the pruning strategy we adopted is shown in Figure 2. in the training stage of our model, because there are both forward and back propagation, each layer from L1 to L4 is used for weight updating. while in the Inference phase, the input image will only propagate forward, reducing the level of the model will not hinder the output of the model. due to the use of deep supervision, each layer from L1 to L4 produces corresponding segmentation results. The experimental results show that the number of parameters in L3 layer is reduced more than that in L4 layer, and the performance is only slightly decreased. Compared with TransBTS, a classical model using transformer, our

TABLE 3. Fold cross validation results on Brats2021 validation set.

Model	Dice(%)			Hausdorff95		
	Whole Tumor(WT)	Tumor Core(TC)	Enhancing Tumor(ET)	Whole Tumor(WT)	Tumor Core(TC)	Enhancing Tumor(ET)
FOLD1	91.3	82.3	86.8	6.89	13.3	11.8
FOLD2	90.6	81.8	87.2	4.26	15.6	12.6
FOLD3	91.6	83.4	84.6	8.23	14.6	15.4
FOLD4	92.8	82.9	86.7	5.64	16.9	13.2
FOLD5	91.4	84.6	88.1	6.28	17.3	14.9
ENSEMBLE	93.2	86.2	88.3	4.58	14.8	12.2

TABLE 4. Fold cross validation results on Brats2020 validation set.

Model	Dice(%)			Hausdorff95		
	Whole Tumor(WT)	Tumor Core(TC)	Enhancing Tumor(ET)	Whole Tumor(WT)	Tumor Core(TC)	Enhancing Tumor(ET)
FOLD1	89.7	84.8	78.9	10.9	15.4	14.6
FOLD2	88.6	82.5	80.6	11.4	14.3	15.7
FOLD3	90.6	82.7	81.8	9.8	18.9	18.5
FOLD4	91.2	84.3	80.4	7.4	17.6	18.1
FOLD5	89.4	84.9	79.8	10.3	19.2	17.6
ENSEMBLE	91.4	85.3	82.3	6.32	16.9	15.2

model not only improves segmentation accuracy but also reduces computational complexity.

3) QUALITATIVE ANALYSIS

In FIG 4, we visualized the segmentation results of the proposed model on the Brats2021 data set, and conducted a 5 fold cross validation evaluation of the proposed method on the validation set. by comparing other experimental methods in table 1 and table 2, our model has achieved great advantages in the Brats2021 data set, especially in WT and ET, and Dice and Hausdorff metrics are significantly better than other similar methods. moreover, it can be seen from Figure 4 that our model more accurately describes brain tumors, and has obvious advantages for edge segmentation. by modeling the long-term dependence between each volume through Swin Transform, a better segmentation mask can be generated. table 3 and table 4 respectively show the cross-validation results from the first to the fifth fold of our model on the Brats2021 and Brats2020 data sets and the average results of the integration of five models. It can be seen from table 3 that on the Brats2021 data set, the 5-fold cross-validation model integrated by us has the best performance. the segmentation accuracy of the fourth-fold and fifth-fold is close to that of the integrated model, but fluctuates greatly, while the performance of Hausdorff95 is poor. similarly, the stability of the third and fifth fold in Table 4 is poor, and there are peaks. however, the segmentation accuracy of the model in our average integration 5 is higher, and the Hausdorff95 index also has the best performance.

4) BraTS 2020

We also conducted experimental verification on the BraTS 2020 data set. Since the modes and corresponding annotations of the BraTS 2020 data set are consistent with those of BraTS 2021, we directly applied the hyperparameters of the BraTS 2021 data set to the BraTS 2020 data set for verification. as the number of instances is about 1/5 of the BraTS 2021 data set, the segmentation accuracy is

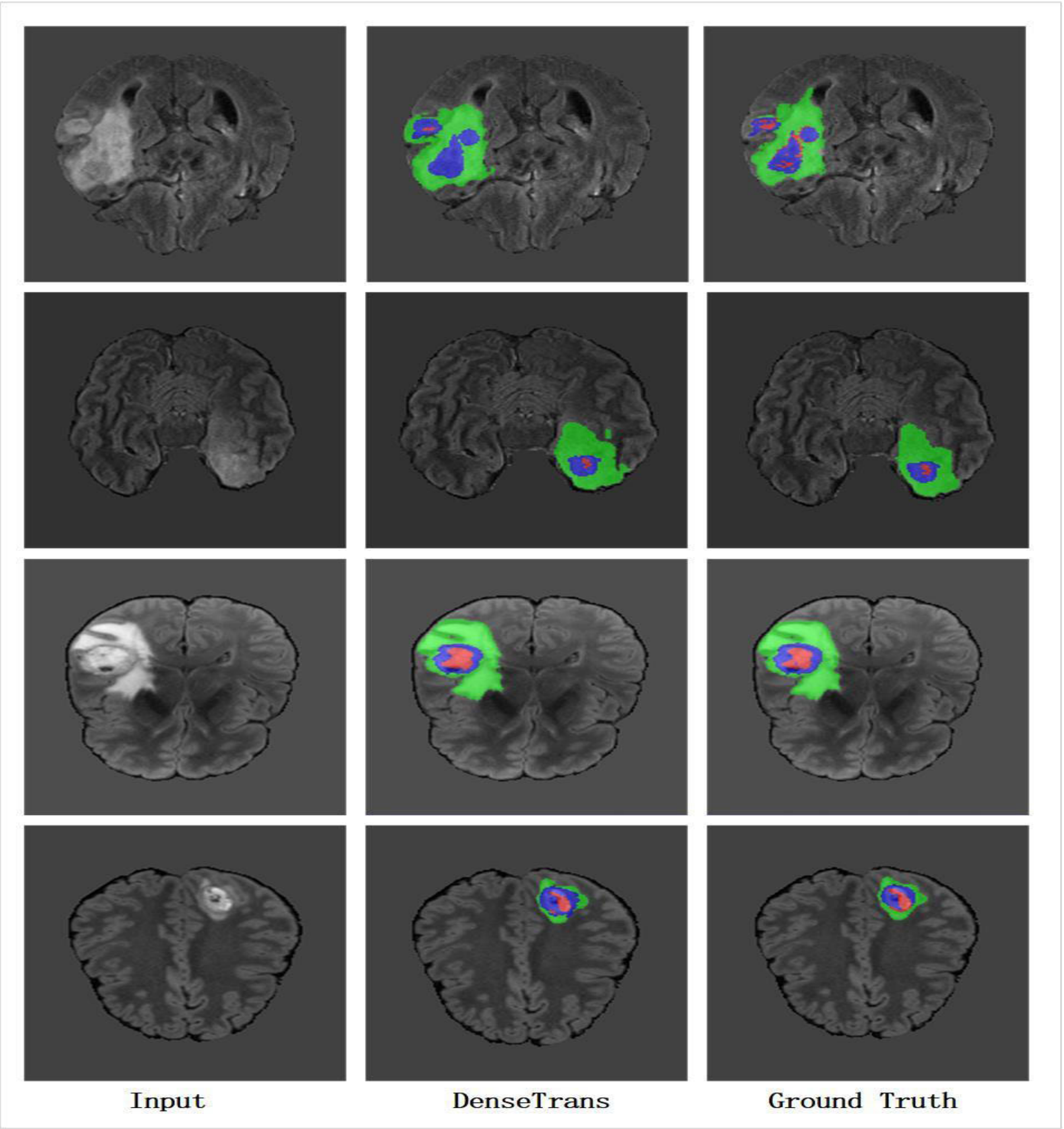

FIGURE 4. Visualized predicted images of different models. The green, red, and blue regions indicate edema (ED), non-enhancing tumor and necrosis (NCR/NET), and enhancing tumor (ET), respectively.

slightly reduced compared with the BraTS 2021 data set. the segmentation accuracy was 91.4%(DSC) and 6.32(HD) on WT, 85.3%(DSC) and 16.9(HD) on TC, 82.3%(DSC) and 15.2(HD) on ET. compared with classical networks such as TransBTS Net, 3DU-net, Dual-Path UNet and Scale Attention Unet, our model has achieved significant advantages in two evaluation metrics, and the segmentation accuracy has been significantly improved. This indicates that the deep fusion method of CNN and swin transfrom adopted by us has a remarkable effect. compared with the traditional 3DU-net, it can be obviously seen that the segmentation accuracy of our model is increased by 2.8%(DSC) on WT, 1.0%(DSC) on TC and 3.8%(DSC) on ET. meanwhile, the Hausdorff distance is also lower than that of 3DU-net. this demonstrates the importance of our improved swin transfrom learning long-term dependencies and global context information.

D. MODEL COMPLEXITY ANALYSIS

Since the model proposed in this paper combines Swin Transform with the improved UNet ++, both the number of

TABLE 5. Performance comparison with other state-of-art models.

Method	Params(M)	FLOPs(G)	Dice(%)			Hausdorff95		
			WT	TC	ET	WT	TC	ET
3D UNet	16.21	1670	88.9	84.8	79.1	6.54	18.68	19.27
TransBTS	26.82	346	89.6	80.8	79.5	5.64	12.48	18.95
Swin UNETR	61.98	394.84	92.6	88.5	85.8	5.83	3.77	6.02
Our(DenseTrans-L4)	51.82	384	93.2	86.2	88.3	4.58	14.8	12.2
Our(DenseTrans-L3depthwise separable convolution)	21.3	212	92.8	85.8	87.2	5.32	15.8	13.6

training parameters and computational complexity are improved, with 51.82M parameters and 384GFlops, which is a medium-scale model. later, we used deep separable convolution and deep supervision to reduce the number of layers in swin transformer during the model downsampling process to alleviate this problem. by reducing swin transformer from L4 to L3, we get a relatively lightweight DenseTrans model with 21.3M parameters and 212GFlops, with only a slight performance drop, due to the depth separable convolution and the reduction in the number of swin transformer layers. compared with TransBTS net model, the number of parameters is reduced by 21%, and the performance is greatly improved. compared with 16.21M parameter and 1670GFlops of the 3D UNet model, the number of parameters is not much different, while the performance improvement is more obvious. our lightweight DenseTrans model has significant advantages in terms of complexity and segmentation accuracy compared to brain tumor segmentation models using Transformer.

In general, the DenseTrans model proposed in this paper, after the introduction of swin transformer layer reduction and depth-separable convolution, obtains a lightweight network. the number of network parameters and computational complexity are shown in table5. compared with 3D UNet, TransBTS and Swin UNETR models, this lightweight model has fewer parameters and computational complexity.the Dice Similarity Score was 92.8%, 85.8%, 87.2% in the whole tumor,tumor core and enhancing tumor, Hausdorff Distance(95%) values of 5.32mm,15.8mm and 13.6mm. the performance is obviously better than previous similar excellent models.

E. DISCUSSION

The experimental results from Brats2021 and Brats2020 data sets show that the segmentation accuracy of the proposed model is significantly improved on both WT and ET by comparing with the traditional excellent models Extending nn-UNet, Swin Unter, TransBTS Net and 3DU-net. the results show the effectiveness and feasibility of the model, and further prove that the combination of swin transform and improved UNet++ is helpful for long-term dependency modeling and global context information acquisition. as for the slight decline in the segmentation accuracy on TC, this is due to the low weight setting of the T1GD mode in the initial input multi-mode fusion, and the T1GD mode is more suitable for the detection of TC region. according to the experimental verification, if the weight of T1GD mode is adjusted during the multi-mode fusion, the segmentation accuracy of WT and ET regions will be affected. therefore, the model is set according to the current superparameter to ensure the optimization of segmentation accuracy. In terms of details, compared with Swin Unter and Extending nn-UNet, although the performance of our model is obviously improved in WT and extending nn-unet, the Dice coefficient in TC is reduced by 2.3% and 1.8% respectively. the more direct reason is that we combine the features extracted by CNN with the feature layer of swin transformer in the high-resolution filter layer of input features, and swin transformer adopts (4 × 4)patch as the self-attention weight of Token in a unit learning window. then, the information between local windows is exchanged by shift window and stack operation to obtain global feature information, but this operation limits the ability to obtain local semantic feature information. therefore, compared with Swin Unter and Extending nn-UNet, the necrotic and non-enhancing tumor parts in the TC region are limited in the ability of our model to obtain local information, thus reducing the segmentation accuracy in this region. In general, through comparative experimental analysis, our model has the following advantages over the best brain tumor segmentation models: (a) It greatly improves the segmentation accuracy of the whole tumor (WT) and the enhancing tumor (ET) regions. (b) compared with other brain tumor transformer segmentation models, our lightweight model has fewer parameters.

F. LIMITATIONS AND FUTURE WORK

The advantages of our proposed model are that it improves the segmentation accuracy of multimodal brain tumors and effectively alleviates the problem of secondary increase of computational complexity caused by the introduction of transformer in the field of medical images.

1) LIMITATIONS

(a) Problems of generality and uncertainty. our model is designed based on a specific brain tumor model, so it does not have a good universality for other tumors. Secondly, multimodal fusion was carried out in model setting and experimental configuration, but it was only carried out for different sequences in MRI, without attempting to fuse CT, MRI and image text.finally, because our experiment was conducted in MRI of brain tumors, there is uncertainty in our model for other tumors or other medical imaging methods such as CT.

(b) Computational complexity. we use deep separable convolution and other operations to reduce the complexity of the model. compared with similar brain tumor segmentation models in transformer, the complexity of our model is much lower. however, due to the introduction of full self-attention mechanism in transformer, the complexity of our model is still higher than that of pure CNN.

2) FUTURE WORK

(a) Universality.we have collected CT, MRI images and image reports of 45 cancer patients, and plan to establish a general model to accurately segment tumors in different sites in the future.

(b) Computational complexity.we will try to take measures including concurrent multi-head self-attention learning mechanism to balance the increase of receptive field and the secondary increase of computational complexity in transformer self-attention learning window.

(c) Clinical practicability. Inspired by the work of [39], [41], and [42] and in order to improve the generality and robustness of the model, we applied the feature extraction framework of the proposed model to the detection and classification of brain tumors to build an excellent computer-aided diagnosis system and provide accurate and reliable reference basis for clinical practice.

V. CONCLUSION

In this paper, we introduce the DenseTrans model, which is a new multimodal brain tumor segmentation model combined with improved UNet++ and Swin Transformer. the model uses the improved Encoder of UNet++ to extract local features, and then each layer in nested UNet transfers the extracted features to swin transformer for learning the long-distance dependency and obtaining the global context information. combining advantages of CNN and transformer, the experimental results on the Brats2021 and Brats2020 dataset show that the model can effectively improve the accuracy of brain tumor segmentation. In future work, we will continue to study the lightweight aspect of DenseTrans model to achieve an efficient semantic segmentation model.

REFERENCES

[1] D. N. Louis, H. Ohgaki, O. D. Wiestler, W. K. Cavenee, P. C. Burger, A. Jouvet, B. W. Scheithauer, and P. Kleihues, “The 2007 WHO classification of tumours of the central nervous system,” *Acta Neuropathologica*, vol. 114, no. 2, pp. 97–109, Aug. 2007.

[2] X. Guan, G. Yang, J. Ye, W. Yang, X. Xu, W. Jiang, and X. Lai, “3D AGSE-VNet: An automatic brain tumor MRI data segmentation framework,” *BMC Med. Imag.*, vol. 22, no. 1, pp. 1–18, Dec. 2022.

[3] J. Liu, M. Li, J. Wang, F. Wu, T. Liu, and Y. Pan, “A survey of MRI-based brain tumor segmentation methods,” *Tsinghua Sci. Technol.*, vol. 19, no. 6, pp. 578–595, Dec. 2014.

[4] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[5] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[7] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[8] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, “Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks,” in *Proc. Int. MICCAI Brainlesion Workshop*, Sep. 2017, pp. 178–190.

[9] A. Myronenko, “3D MRI brain tumor segmentation using autoencoder regularization,” in *Proc. Int. MICCAI Brainlesion Workshop*, Sep. 2018, pp. 311–320.

[10] Z. Jiang, C. Ding, M. Liu, and D. Tao, “Two-stage cascaded U-Net: 1st place solution to brats challenge 2019 segmentation task,” in *Proc. Int. MICCAI Brainlesion Workshop*, Oct. 2019, pp. 231–241.

[11] F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, “nnU-Net for brain tumor segmentation,” in *Proc. Int. MICCAI Brainlesion Workshop*, Oct. 2020, pp. 118–132.

[12] H. M. Luu and S. H. Park, “Extending nn-UNet for brain tumor segmentation,” in *Proc. Int. MICCAI Brainlesion Workshop*, 2022, pp. 173–186.

[13] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “TransUNet: Transformers make strong encoders for medical image segmentation,” 2021, *arXiv:2102.04306*.

[14] T. Magadza and S. Viriri, “Brain tumor segmentation using partial depthwise separable convolutions,” *IEEE Access*, vol. 10, pp. 124206–124216, 2022.

[15] S. Liang, Z. Hua, and J. Li, “Transformer-based multi-scale feature fusion network for remote sensing change detection,” *J. Appl. Remote Sens.*, vol. 16, no. 4, Nov. 2022, Art. no. 046509.

[16] Z. Zhou et al., “UNet++: A nested U-Net architecture for medical image segmentation,” in *Proc. 4th Int. Workshop, DLMIA, 8th Int. Workshop, ML-CDS, Conjunct. MICCAI.* Granada, Spain: Springer, Sep. 2018, pp. 3–11.

[17] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[18] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, “UNet 3+: A full-scale connected UNet for medical image segmentation,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1055–1059.

[19] C. Qin, Y. Wu, W. Liao, J. Zeng, S. Liang, and X. Zhang, “Improved U-Net3+ with stage residual for brain tumor segmentation,” *BMC Med. Imag.*, vol. 22, no. 1, pp. 1–15, Dec. 2022.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ''An image is worth 16×16 words: Transformers for image recognition at scale,'' 2020, *arXiv:2010.11929*.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, ''Swin transformer: Hierarchical vision transformer using shifted windows,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[23] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, ''Swin-UNet: UNet-like pure transformer for medical image segmentation,'' 2021, *arXiv:2105.05537*.

[24] Y. Wu, K. Liao, J. Chen, J. Wang, D. Z. Chen, H. Gao, and J. Wu, ''D-former: A U-shaped dilated transformer for 3D medical image segmentation,'' *Neural Comput. Appl.*, vol. 35, pp. 1931–1944, Oct. 2022.

[25] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, ''Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images,'' in *Proc. Int. MICCAI Brainlesion Workshop*, 2022, pp. 272–284.

[26] H. Peiris, Z. Chen, G. Egan, and M. Harandi, ''Reciprocal adversarial learning for brain tumor segmentation: A solution to BraTS challenge 2021 segmentation task,'' 2022, *arXiv:2201.03777*.

[27] H. Jia, C. Bai, W. Cai, H. Huang, and Y. Xia, ''HNF-Netv2 for brain tumor segmentation using multi-modal MR imaging,'' 2022, *arXiv:2202.05268*.

[28] A. S. Akbar, C. Fatichah, and N. Suciati, ''UNet3D with multiple atrous convolutions attention block for brain tumor segmentation,'' in *Proc. Int. MICCAI Brainlesion Workshop*, 2022, pp. 182–193.

[29] Z. Li, Z. Shen, J. Wen, T. He, and L. Pan, ''Automatic brain tumor segmentation using multi-scale features and attention mechanism,'' in *Proc. Int. MICCAI Brainlesion Workshop*, 2022, pp. 216–226.

[30] P. Ahmad, S. Qamar, L. Shen, S. Q. A. Rizvi, A. Ali, and G. Chetty, ''MS UNet: Multi-scale 3D UNet for brain tumor segmentation,'' in *Proc. Int. MICCAI Brainlesion Workshop*, 2022, pp. 30–41.

[31] K. Pawar, S. Zhong, D. S. Goonatillake, G. Egan, and Z. Chen, ''Orthogonal-nets: A large ensemble of 2D neural networks for 3D brain tumor segmentation,'' in *Proc. Int. MICCAI Brainlesion Workshop*, 2022, pp. 54–67.

[32] J. Roth, J. Keller, S. Franke, T. Neumuth, and D. Schneider, ''Multi-plane UNet++ ensemble for glioblastoma segmentation,'' in *Proc. Int. MICCAI Brainlesion Workshop*, 2022, pp. 285–294.

[33] Y. Wang, Y. Zhang, F. Hou, Y. Liu, J. Tian, and C. Zhong, and Z. He, ''Modality-pairing learning for brain tumor segmentation,'' in *Proc. Int. MICCAI Brainlesion Workshop*, 2021, pp. 230–240.

[34] T. Henry, A. Carré, M. Lerousseau, T. Estienne, C. Robert, N. Paragios, and E. Deutsch, ''Brain tumor segmentation with self-ensembled, deeply-supervised 3D U-Net neural networks: A BraTS 2020 challenge solution,'' in *Proc. Int. MICCAI Brainlesion Workshop*, 2020, pp. 327–339.

[35] W. Jun, X. Haoxiang, and Z. Wang, ''Brain tumor segmentation using dual-path attention U-Net in 3D MRI images,'' in *Proc. Int. MICCAI Brainlesion Workshop*, 2021, pp. 183–193.

[36] C. Liu, W. Ding, L. Li, Z. Zhang, C. Pei, L. Huang, and X. Zhuang, ''Brain tumor segmentation network using attention-based fusion and spatial relationship constraint,'' in *Proc. Int. MICCAI Brainlesion Workshop*, 2021, pp. 219–229.

[37] M. D. Cirillo, D. Abramian, and A. Eklund, ''Vox2Vox: 3D-GAN for brain Tumour segmentation,'' in *Proc. Int. MICCAI Brainlesion Workshop*, 2021, pp. 274–284.

[38] Y. Yuan, ''Automatic brain tumor segmentation with scale attention network,'' in *Proc. Int. MICCAI Brainlesion Workshop*, 2021, pp. 285–294.

[39] V. Rajinikanth, S. Kadry, and Y. Nam, ''Convolutional-neural-network assisted segmentation and SVM classification of brain tumor in clinical MRI slices,'' *Inf. Technol. Control*, vol. 50, no. 2, pp. 342–356, Jun. 2021.

[40] S. Kadry, R. Damasevicius, D. Taniar, V. Rajinikanth, and I. A. Lawal, ''U-Net supported segmentation of ischemic-stroke-lesion from brain MRI slices,'' in *Proc. 7th Int. Conf. Bio Signals, Images, Instrum. (ICBSII)*, Mar. 2021, pp. 1–5.

[41] S. Maqsood, R. Damaševičius, and R. Maskeliūnas, ''Multi-modal brain tumor detection using deep neural network and multiclass SVM,'' *Medicina*, vol. 58, no. 8, p. 1090, Aug. 2022.

[42] S. Maqsood, R. Damasevicius, and F. M. Shah, ''An efficient approach for the detection of brain tumor using fuzzy logic and U-NET CNN classification,'' in *Proc. Int. Conf. Comput. Sci. Appl.* Cham, Switzerland: Springer, Sep. 2021, pp. 105–118.

[43] M. Frank, D. Drikakis, and V. Charissis, ''Machine-learning methods for computational science and engineering,'' *Computation*, vol. 8, no. 1, p. 15, Mar. 2020.

[44] K. Poulinakis, D. Drikakis, I. W. Kokkinakis, and S. M. Spottswood, ''Machine-learning methods on noisy and sparse data,'' *Mathematics*, vol. 11, no. 1, p. 236, Jan. 2023.

LI ZONGREN received the Ph.D. degree and the master's degree in software engineering from Lanzhou Jiaotong University, Lanzhou, China, in 2013 and 2019, respectively. He is currently pursuing the Ph.D. degree in computer science and technology with Xinjiang University, Ürümqi, China, in 2020. His current research interests include medical image analysis and medical data security.

WUSHOUER SILAMU was born in Greenwich Wushuer Lamu, Yili, Xinjiang, in October 1941. He is currently a Multilingual Information Processing Expert, an Academician of the Chinese Academy of Engineering, and a Professor and a Doctoral Supervisor with Xinjiang University.

WANG YUZHEN received the master's degree. She is currently a Senior Engineer, the Deputy Director of the Information Department, 940th Hospital of the PLA Joint Logistic Support Force, and a Medical Information Field Expert, long-term engaged in medical information management, technology research and development, and scientific research.

WEI ZHE was born in Baoji, Shanxi, in October 1963. He is currently pursuing the Ph.D. degree.

He is also a master's supervisor with 30 years of experience in biomedical engineering and information engineering.