electronics

MDPI

Article

CSU-Net: A CNN-Transformer Parallel Network for Multimodal Brain Tumour Segmentation

Yu Chen [1,†], Ming Yin [2,†], Yu Li [2,*] and Qian Cai [2,*]

[1] College of Computer, Hubei University of Education, Wuhan 430205, China; chenyuhue@163.com
[2] School of Electronic and Electrical Engineering, Wuhan Textile University, Wuhan 430200, China; 2015053003@mail.wtu.edu.cn
* Correspondence: leewoo@wtu.edu.cn (Y.L.); 2001001@wtu.edu.cn (Q.C.)
† These authors contributed equally to this work.

Abstract: Medical image segmentation techniques are vital to medical image processing and analysis. Considering the significant clinical applications of brain tumour image segmentation, it represents a focal point of medical image segmentation research. Most of the work in recent times has been centred on Convolutional Neural Networks (CNN) and Transformers. However, CNN has some deficiencies in modelling long-distance information transfer and contextual processing information, while Transformer is relatively weak in acquiring local information. To overcome the above defects, we propose a novel segmentation network with an "encoder–decoder" architecture, namely CSU-Net. The encoder consists of two parallel feature extraction branches based on CNN and Transformer, respectively, in which the features of the same size are fused. The decoder has a dual Swin Transformer decoder block with two learnable parameters for feature upsampling. The features from multiple resolutions in the encoder and decoder are merged via skip connections. On the BraTS 2020, our model achieves 0.8927, 0.8857, and 0.8188 for the Whole Tumour (WT), Tumour Core (TC), and Enhancing Tumour (ET), respectively, in terms of Dice scores.

Keywords: brain tumour segmentation; multimodal MRI; CNN; volumetric transformer

check for updates

Citation: Chen, Y.; Yin, M.; Li, Y.; Cai, Q. CSU-Net: A CNN-Transformer Parallel Network for Multimodal Brain Tumour Segmentation. *Electronics* **2022**, *11*, 2226. https://doi.org/10.3390/electronics11142226

Academic Editor: Abdeldjalil Ouahabi

Received: 31 May 2022
Accepted: 13 July 2022
Published: 16 July 2022

1. Introduction

The glioma represents the most common primary brain tumour [1]. The histological subareas of glioma include Oedema/Invasion, Necrosis, Enhancing, and Non-Enhancing. In routine diagnosis, different tissue features can be highlighted by certain sequences (T1, T1CE, T2, Flair, etc.). For example, low-grade gliomas often show low T1 signals and high T2 signals on MRI. It is difficult to locate and segment the uneven shape and obscure scope of gliomas in MRI. Currently, clinical diagnosis mainly relies on the subjective judgments of medical experts, a process requiring their professional assessment and rich experience. Occasionally, it is difficult for the medical experts to reach a consensus on brain tumour image segmentation of the same patient. The traditional segmentation process is based on manual labelling and is a complex process with poor repeatability. To address the above problems, it is critical to develop an effective algorithm to segment tumour subregions, which can provide the basis for quantitative image analysis, assistant diagnoses and surgical planning, and even patient survival prediction.

High quality MR images are the first step to developing an effective algorithm. However, most MR images have problems such as inconsistent imaging protocols and image noise, which negatively impact data analysis and model inference. Thus, we need to preprocess the original MR images to obtain high quality ones. For imaging protocols, the same set of anatomical templates can be applied to achieve consistency. For image noise, wavelet denoising [2,3] and compressed sensing [4,5] are common solutions: the former approach can preserve edge information of the image during denoising; the latter can restore the image to the high dimension after denoising in the low-dimension space.

Currently, many studies focus on deep learning models to segment tumour subregions based on high quality MR images. The Convolutional Neural Network (CNN) has demonstrated good performance on image segmentation. Chen et al. [6] propose a new semantic segmentation method with combined DCNN and CRF, which obtains relatively accurate results. Aiming at sparse feature maps, they also propose the DeepLab model [7] to avoid insensitivity of the network to targets with various scales. Wang et al. [8] propose pixel contrast learning, a fully supervised semantic segmentation training approach, to learn a structured feature space based on pixel–pixel correspondences across images in training, which outperforms traditional image-based training paradigms.

Many works have extended CNN's application to transfer natural semantic segmentation to medical image segmentation. For example, Huang et al. [9] propose UNet 3+ with improved skip connections and multiscale depth supervision to combine low-level detail with high-level semantics. Zeng et al. [10] present RIC-Unet for nuclei segmentation, using residual blocks and channel attention mechanisms. They have made much progress on medical image segmentation. However, the accuracy remains insufficient to assist medical treatment due to two factors: first, some features of medical images may affect segmentation accuracy, such as blank background information, the small amount of training medical images but with large volume, and multimodality of the same lesion; second, pure CNN cannot effectively be applied to medical image segmentation because of limitations in the size of the perceptual field, slow processing of large data volumes, and insufficiency of long-range dependencies.

Transformer initially applied in NLP (Natural Language Processing) was introduced into computer vision by Dosovitskiy et al. [11] for the first time. In the following research regarding medical image segmentation, Zhang et al. [12] propose a novel method with the integration of multiscale attention and CNN to comprehend relations of different ranges without changing the overall complexity. Zhou et al. [13] propose a simple yet powerful hybrid Transformer network for multi-label cardiac MR image segmentation. However, there are several challenges to medical image segmentation, such as larger volumes of 3D medical and deep layers of Transformer blocks with significant parameters, displaying two main reasons to increase GPU memory. To solve the problem, we usually slice 3D medical images into 2D, which reduces the amount of computation and the demand for GPU memory, but at the loss of 3D context information.

Aiming at solving the above defects relatively, we propose a novel segmentation network with an "encoder–decoder" architecture, namely CSU-Net. The encoder consists of two parallel feature extraction channels based on CNN and Transformer, respectively, in which the features of the same size are fused. The decoder has a dual Swin Transformer decoder block with two learnable parameters for feature upsampling. The features from multiple resolutions in the encoder and decoder are merged via skip connections.

Our main contributions to this work are as follows:

- CSU-Net can directly process 3D dataset without slicing each voxel;
- CSU-Net proposes a novel architecture in which (1) The encoder structure is based on improved CNN and Swin Transformer in parallel, which enables establishment of remote dependencies at a high level and retains the ability of local feature extraction; (2) The information extracted from the CNN branch is applied in guidance from the following information extraction in the Transformer branch, which can accelerate model convergence and reduce training time; and (3) The decoder structure is based on a dual Swin Transformer parallel structure and introduces two learnable parameters to enhance the capability of restoring information;
- We validate the effectiveness of our method on 3D MRI dataset (BraTS). It exceeds the current advanced schemes on WT, ET and TC segmentation regions, to achieve a Dice score of 0.8927, 0.8188 and 0.8857, respectively.

2. Related Work

CNN-based Segmentation Networks: The introduction of CNN has a significant impact on medical segmentation tasks. For 2D segmentation, CE-Net [14] captures high-level information and preserves spatial information by building a network of contextual encoders. For 3D segmentation, 3D U-Net [15] is a simple extension of U-Net applied to 3D image segmentation. These networks aim at extracting different dimensions of information during the downsampling process. A deep encoding layer with a smaller receptive field can obtain more accurate edge information. However, these kinds of serial patterns increase the receptive field of the last network layers, leading to incomplete learning of high-level features and inaccurate segmentation of details.

Some networks use parallel structures in the encoder section to increase the information sources or break through the perceptual domain's limitations. KiU-Net [16] improves performance by fusing Ki-Net and U-Net to detect smaller perceptual regions. A two-stage multiscale framework proposed by Roth et al. [17] achieves superior performance in the image segmentation of the pancreas. However, these may lead to inaccuracy of the segmentation due to their poor learning ability in establishing long-distance information transfer dependencies and handling contextual information.

Transformer-based Networks: Vision Transformer has dramatically advanced the performance of machine vision tasks. In ViT [11], a long-range information model is constructed by transforming images into fixed-size patches and adding positional patch embedding before the sub-attentive mechanism module for more effective global contextual connectivity. Swin-Unet [18] contains a symmetric encoder–decoder structure using a skip-connection of swin transformer modules, which implements local to global self-attention in the encoder. In its decoder, multiscale features are fused with the encoder using skip-connections. In contrast to Swin-Unet, nnFormer [19] is computed mainly using a cross-backbone network and a local 3D image block-based self-attention mechanism. They are far less capable of information extraction compared to CNN.

Lin et al. [20] propose a dual-scale semantic segmentation model based on Swin Transformer to construct long-distance feature relationships between different scales using the self-attention mechanism. It validates the model on several medical datasets, which gains better results. However, adding multiple parallel Swin Transformer channels may increase the amount of model data and computation.

CNN and Transformer Fusion Network: Aiming at the weaknesses of pure CNN networks and pure Transformer networks, the fusion of these two structures can compensate for each other. TransUnet [21] is the first application of Transformer to CNN, while BiTr-Unet [22] is also associated with fusing 3D CNN with Transformer. The Transformer block of the above work is performed after the CNN. The features are then recovered by upsampling layer by layer. To achieve accurate image segmentation, image processing requires a large stack of arithmetic power. The overall data volume and computational complexity rise dramatically when dealing with 3D data.

To solve the above problems, we propose a novel parallel network of CNN and Transformer to establish complete global dependencies and preserve the network's feature extraction ability. The feature extracted by the improved CNN branch is fused with the features of the Swin Transformer and then fed into the next layer of the Swin Transformer module. Throughout the encoding process, the CNN branch uses its powerful feature extraction capabilities to establish guidance for the feature transfer of the Swin Transformer with the different dimensional features.

The key differences between our model and those of other works include the following. (1) CSU-Net is applied to 3D medical images where the network can directly process the volumetric data without needing low-dimensional transformations; (2) CSU-Net uses the improved CNN and Swin Transformer as encoders in parallel, rather than using these encoders in tandem. This parallel structure allows global–local information to be obtained and remains efficient in information processing rates compared to a single deep network.

3. Method

In this section, the overall framework structure of the network is shown first. Then each component, such as DConv, Swin Transformer Encoder and Swin Transformer Decoder, is described in detail.

3.1. Overall Architecture of CSU-Net

This section presents an overview of the proposed CSU-Net model, as shown in Figure 1.

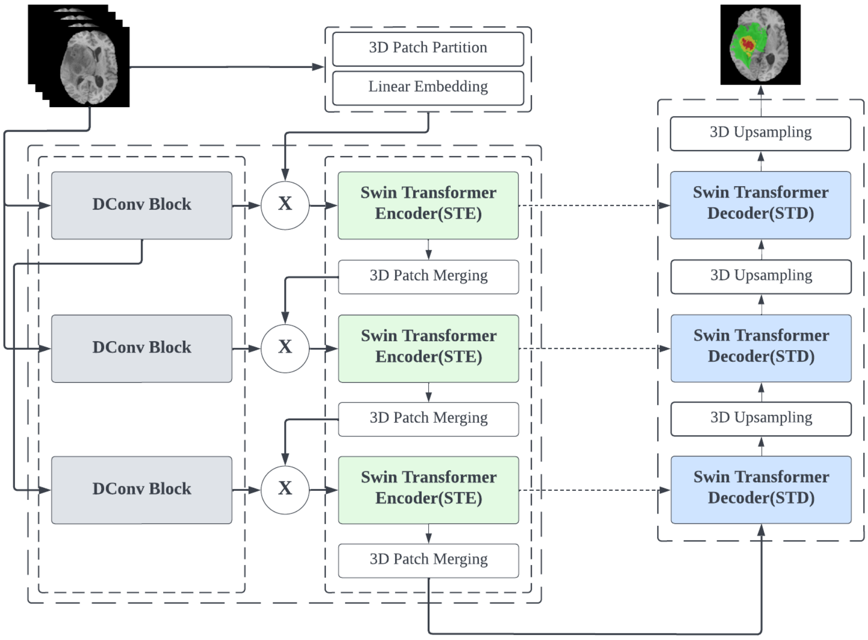

Figure 1. Overall network framework diagram.

CSU-Net utilises a parallel architecture which combines CNN and Transformer as an encoder, interacting with the decoder via skip connections. In the parallel architecture, the CNN downsampling channel consists mainly of DConv modules containing large convolutional kernels and bottleneck structures. The Transformer downsampling channel is primarily constructed using Swin Transformer. Considering the possibility of effective parameter loss and inadequate restoration due to dropout in the decoding, we propose a decoder using a dual Swin Transformer block with learnable parameters. In addition, a classification layer is constructed at the end of the overall network to predict the segmentation results. For the BraTS 2020 dataset, three types of segmentation results will be predicted.

3.2. Encoder

3.2.1. Swin Transformer Encoder (STE) Block

The Embedding Block is responsible for converting each input image $C \times D \times H \times W$ into non-overlapping patches and subsequently mapping to a tensor of a set dimension. In this design, the patch size is $4 \times 4 \times 4$, the sequence length C is set as 96, and then transformed into a high-dimensional tensor $X_{embedding} \in R^{\frac{D}{4} \times \frac{H}{4} \times \frac{W}{4} \times 96}$.

Swin Transformer has creatively designed a shifted window operation. Figure 2 shows the Swin Transformer block internal connections and component block.

Each STE (Swin Transformer Encoder) Block consists of a LayerNorm (LN) layer, a multi-head self-attention (MSA) module, a residual connection, and two layers of MLPs containing GELU activation functions. MSA has two kinds of modules: W-MSA (window-based multi-head self-attention) and SW-MSA (shifted window-based multi-head self-attention). In W-MSA, the volume is cut into non-overlapping blocks of a specified size.

SW-MSA uses the shifted window mechanism to link unassociated adjacent blocks in W-MSA. The STE Block implements the functions as Equation (1):

$$\begin{aligned} \hat{z}^{l} &= W-MSA\left(LN\left(z^{l-1}\right)\right)+z^{l-1}, \\ z^{l} &= MLP\left(\mathrm{LN}\left(\hat{z}^{l}\right)\right)+\hat{z}^{l}, \\ \hat{z}^{l+1} &= SW-MSA\left(LN\left(z^{l}\right)\right)+z^{l}, \\ z^{l+1} &= MLP\left(LN\left(\hat{z}^{l+1}\right)\right)+\hat{z}^{l+1} \end{aligned} \quad (1)$$

where l denotes the block layer; z^l and $\hat{z}^l$ denote the output features of the W-MSA module and the MLP module, respectively.

The W-MSA module and SW-MSA module are mainly composed of self-attention mechanisms and trainable relative position encoding. The overall computation is shown in Equation (2):

$$\text{Attention }(Q,K,V)=\text{soft}\max\left(\frac{QK^{T}}{\sqrt{d_k}}+B\right)V, \quad (2)$$

where Q, K, V denote the query, key and value, respectively; d_k denotes the size of the query and key; B denotes the relative position information deviation value.

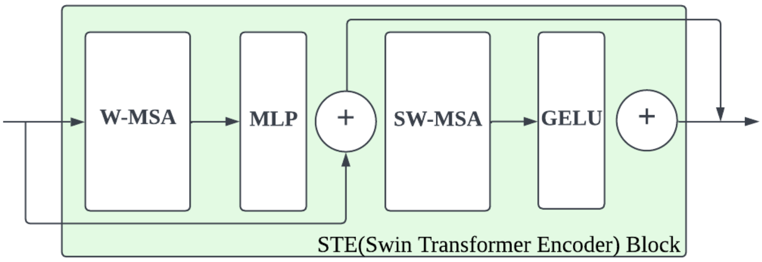

Figure 2. Swin Transformer Module.

To reduce the computational complexity of the attention in 3D images, we build the Sign Transformer Encoder layer by layer using the W-MSA and SW-MSA module and MLP while receiving feature information from the CNN to complement the local attention. The detailed parameters of the STE blocks are shown in Table 1.

Table 1. The parameters of STE blocks.

Conv Name	Dim Length	Depths
STE_0	96	2
STE_1	192	2
STE_2	384	2

3.2.2. DConv Block

To supplement the information lost in the STE's downsampling and compensate for the lack of attention to local features, we add a pure convolutional downsampling block, called DConv, to the existing backbone network.

We choose convolutional layers in the parallel network because they encode spatial information at the pixel level, which is more accurate than the patch-level position encoding used in Transformer. Moreover, the convolutional architecture is comparatively lighter while remaining efficient when the computation complexity is caused by shift window operations in existing backbone networks. The original image used three different sizes of 3D convolution kernels. The parameters are shown in Table 2.

Table 2. The parameters of 3D Conv.

Conv Name	Input Size	Output Size	Kernel Size	Patch
Conv3d_0	4	96	(4, 4, 4)	(4, 4, 4)
Conv3d_1	4	192	(4, 8, 8)	(4, 8, 8)
Conv3d_2	96	384	(1, 4, 4)	(1, 4, 4)

It is worth mentioning that for the third convolution kernel (1, 4, 4), the input does not directly use the original image but the output of the Conv3d_0. This is because the small kernel size helps to reduce the computation complexity compared to the larger kernels while providing the same size perceptual field.

The DConv module consists of the Depthwise Conv, Layer Norm, and GELU. The specific module connections can be found in Figure 3.

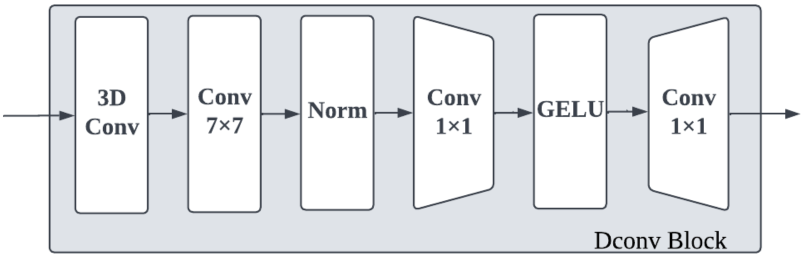

Figure 3. Patch merging process diagram.

3.2.3. Patch Merging

The primary function of this module is to downsample before moving on to the next STE module. This step can be used to reduce the resolution and, at the same time, adjust the number of channels between each layer, saving some arithmetic. To simplify this process, the patch merging process is conducted using a linear transformation. Elements are selected at regular intervals in multiple directions and expanded after the elements are stitched together to form a complete tensor.

3.3. Decoder

We add a skip connection between the encoder and decoder to transmit the feature representations, which can compensate for the loss of partial information, which is a common defect in a U-shaped network.

Referring to the idea of weighted feature fusion in BiFPN proposed by Tan et al. [23], we propose the Swin Transformer Decoder(STD) Block. We design the STD block with a learnable parallel structure formed by the multi-head self-attention (MSA) model. The construction is shown in Figure 4.

Each STD block consists of two identical parallel modules containing W-MSA and SW-MSA. The feature maps from the upper STD block layers are fed into the same two SA modules and subsequently output. When two identical modules fuse, the process is computed as shown in Equation (3):

$$F^l = W_1 \times MSA_1^l(X) + W_2 \times MSA_2^l(X) \tag{3}$$

where W_1, W_2 is the learnable covariate; l denotes the block layer; F denotes the output of the STD block; X denotes the input of the STD block. In contrast to the single MSA architecture used in Swin-Unet [18], the present parallel MSA architecture allows increasing self-supervision in the upsampling process. The source of features feeding into the decoder is increased. The possible bias caused by the random dropout in the upsampling process is corrected relatively by the learnable weight variates.

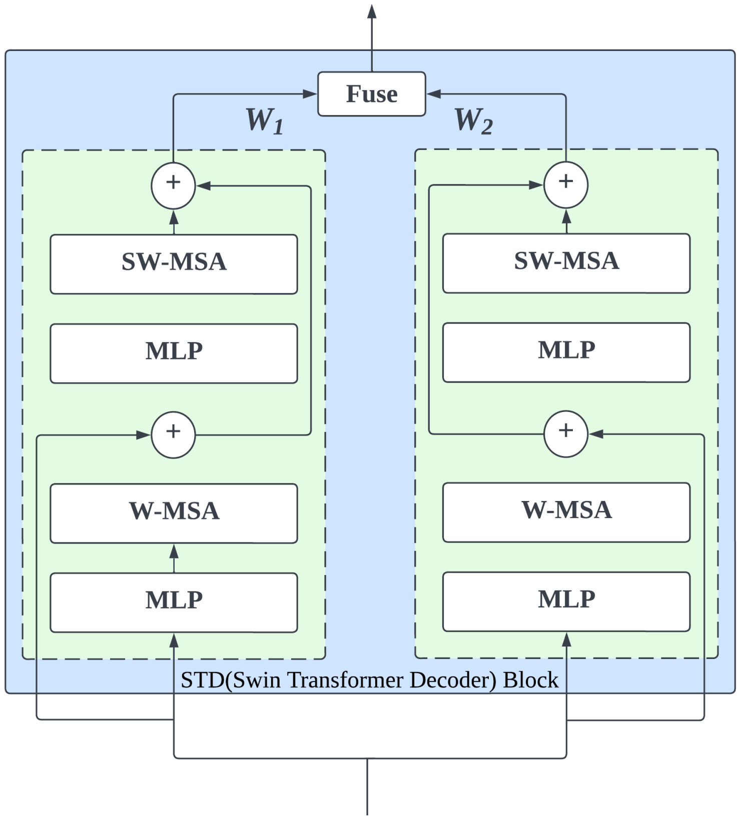

Figure 4. Swin Transformer Decoder (STD) Block.

3.4. Classifier Layer

After the decoding session is completed, the classification layer is introduced, and the feature map of depth C is mapped into N categories using a 3D convolutional layer. Thus, the scale of the predicted output is $(N \times D \times H \times W)$.

3.5. Loss Function

The loss function is a combination of the Dice Loss [24] and the Binary Cross Entropy Loss. Dice Loss is defined as Equation (4):

$$L_{\text{Dice}}(G, Y) = 1 - \frac{2}{N}\sum_{n=1}^{N} \frac{\sum_{m=1}^{M} G_{m,n} Y_{m,n}}{\sum_{m=1}^{M} G_{m,n}^2 + \sum_{m=1}^{M} Y_{m,n}^2} \tag{4}$$

Binary Cross Entropy Loss is defined as Equation (5):

$$L_{CE}(G, Y) = -\sum_{n=1}^{N}\left[\sum_{m=1}^{M}(G_{m,n} \cdot \ln Y_{m,n}) + \sum_{m=1}^{M}(1 - G_{m,n})(1 - \ln Y_{m,n})\right] \tag{5}$$

In Equations (4) and (5), M denotes the number of voxels; N is the number of classes; $Y_{m,n}$ and $G_{m,n}$ denote the probability of output and one-hot-encoded ground truth for class n at voxel m, respectively.

The overall split loss can therefore be defined as Equation (6):

$$L = \alpha L_{\text{Dice}} + (1 - \alpha) L_{CE} \tag{6}$$

In Equation (6), set to a practical value of 0.7.

4. Experiments

In this section, the dataset, the evaluation metrics and other implementation details are described.

4.1. Data and Evaluation Metric

The 3D MRI dataset used in the experiments is provided by the Brain Tumour Segmentation (BraTS) 2020 Challenge [25–27]. All samples contained gliomas. Different sequences (T1, T1CE, T2, and FLAIR) were typically generated by various influences on the MR signal, as shown in Figure 5. These influences can highlight features at different tumour regions to localize the nidus and determine the size of the tumour. The size of each pattern is 240 × 240 × 155. To reduce the effect of unnecessary background on experimental results, all data are cropped to a fixed size of 128 × 128 × 128. Our data preprocessing involves cropping, random rotation and random flipping, which is consistent with baselines. In addition, this dataset consists of relatively high quality MR images. For these two reasons, we do not apply denoising to this dataset.

All four modalities for a sample share the same segmentation file. We aim to output three types of segmentation regions: the Enhancing Tumour (ET): label 1; the Tumour Core (TC): labels 1 and 4; the Whole Tumour (WT): labels 1, 2 and 4.

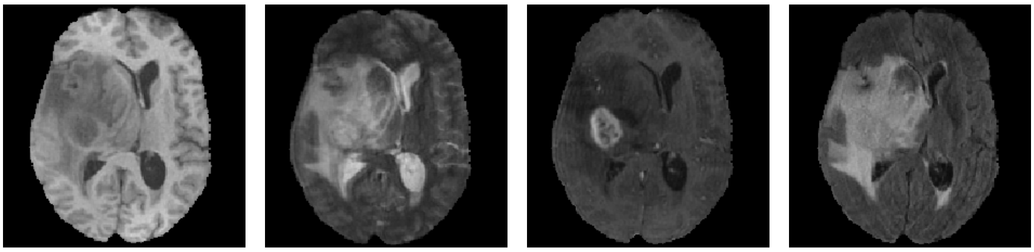

Figure 5. T1, T2, T1CE and FLAIR MR image.

4.2. Implementation Details

The proposed network implements in PyTorch and trained using an NVIDIA RTX 3090 GPU (24 GB of video memory), using a batch size of 1 and executing 300 epochs from zero. We use the Adam optimizer, with the initial learning rate set to 0.0001. In preliminary tests, training the model directly requires a lot of time, so a pre-trained model is used in this experiment. Swin-T, a model pre-trained in ImageNet-1K by Liu et al. [28], was used to accelerate the convergence rate of the experiments.

4.3. Evaluation Metrics

We use Dice score and sensitivity to measure the performance of our model.

The Dice score is calculated as Equation (7):

$$Dice = \frac{2TP}{FN+2TP+FP} \tag{7}$$

Sensitivity is an important index in the medical imaging field that can be used to predict the true positive rate of sample images, which can assess the validity and stability of the algorithm model. It is defined as Equation (8):

$$Sensitivity = \frac{TP}{FN+TP} \tag{8}$$

In Equations (7) and (8), TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively.

5. The Results

In this section, we show and discuss the results of the main experiment and ablation experiment.

5.1. Main Results

We evaluate CSU-Net on BraTS 2020, and the results are reported in Table 3. We compare the performance of CSU-Net against CNN and CNN-transformer. CSU-Net achieves an overall average Dice score of 0.8644 and outperforms the second, third and fourth top-ranked methodologies by 1.813%, 3.496% and 3.732%, respectively. Specifically, on ET and

TC, our method outperforms the second-best baselines by 2.826% and 4.817% in terms of Dice score.

Table 3. Quantitative comparisons of segmentation performance in brain tumour segmentation tasks of the BraTS dataset. WT, ET and TC denote Whole Tumour, Enhancing Tumour and Tumour Core subregions, respectively.

Method	**Dice Score**			
	ET	**TC**	**WT**	**AVG**
3D U-Net	0.6876	0.7906	0.8411	0.7731
ME-Net [29]	0.7020	0.7390	0.8830	0.7747
Liu et al. [30]	0.7637	0.8012	0.8823	0.8157
Ghaffari et al. [31]	0.7800	0.8200	0.9000	0.8333
TransBTS	0.7873	0.8173	0.9009	0.8352
TransBTS V2	0.7963	0.8450	**0.9056**	0.8490
Ours	**0.8188**	**0.8857**	0.8927	**0.8644**

This shows that CSU-Net can compensate for the information loss during downsampling and outperform other CNN-Transformer fusion methods in image restoration.

In Table 4, we compare the performance against CNN-based and CNN-Transformer-based baselines in average Dice, parameters and epochs. CSU-Net outperforms the third top-ranked methodologies by 3.50% and 38.46% in average Dice and parameters. Compared with the closest baseline, although given the slightly increasing parameters, CSU-Net outperforms by 1.5% on average Dice and decreases the training epochs dramatically from 6000 to 300.

Table 4. Comparison of parametric numbers in the BraTS 2020 dataset. AVG represents the average of WT, ET and TC, and Parameters represent the parametric size of the model.

Method	**Dice Score (AVG)**	**Parameters**	**Epochs**
3D U-Net	0.7731	16.13 M	-
TransBTS	0.8352	32.33 M	6000
TransBTS V2	0.8490	**15.03 M**	6000
Ours	**0.8644**	23.35 M	300

To analyse the quality of segmentation results, we show a visual comparison of brain tumour segmentation results of four methods in Figure 6.

As shown in Figure 6, compared with other methods, our model is more accurate in segmenting each subregion of brain tumour, with smoother segmentation edges and clearer boundaries between multiple regions. In the actual clinical detection process, the size of each tumour area can be displayed to help specialists determine the pathological stage of the tumour and facilitate timely planning of treatment.

In Table 5, sensitivity, a commonly used index in medical imaging, is quantified for the visual images of brain tumour segmentation in Figure 6. It is shown that our model matches the ground truth well with the smallest area difference. This is consistent with the high value of sensitivity shown in Table 5, which demonstrates high segmentation accuracy, smooth segmentation edges and complete segmentation area. Therefore, our model has a wide range of clinical applications.

Table 5. Quantitative sensitivity analysis of the segmentation performance in brain tumour segmentation tasks of the BraTS dataset.

Method	**Senstivity**			
	ET	**TC**	**WT**	**AVG**
Ours	0.8077	0.8665	0.9072	0.8604

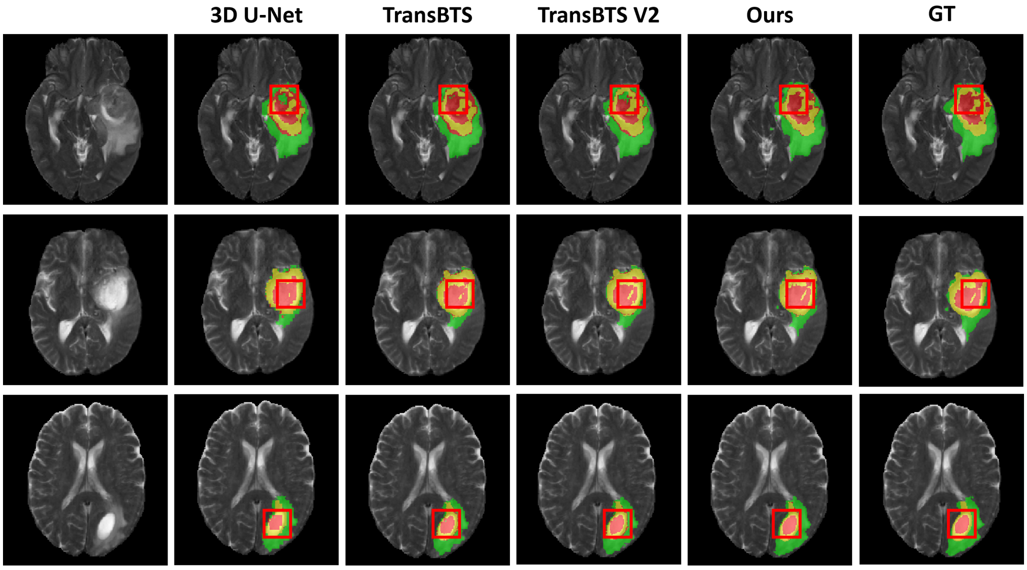

Figure 6. A Visual Comparison of MRI Brain Tumour Segmentation Results. COLUMN-1: the original image. COLUMN-2 to COLUMN-5: the results predicted by different methods. COLUMN-6: the ground truth.

5.2. Ablation Experiments

In this section, we will discuss how the essential components of the overall network model species contribute to improving the overall performance.

Table 6 below confirms that the DConv network's introduction effectively enhances the network information extraction, and the introduction of the parallel Swin Transformer model helps in the information reduction process.

Table 6. Ablation Study on Individual Components.

Method	Dice Score			
	ET	TC	WT	AVG
CSU-Net	**0.8188**	**0.8857**	0.8927	**0.8644**
CSU-Net w/o STD	0.8132	0.8715	**0.8998**	0.8615
CSU-Net w/o STD and DConv	0.8084	0.8613	0.8987	0.8561

Comparison of three models' segmentation images in Figure 7.

Sequence Length C. The sequence length C in Swin Transformer is discussed in detail. Its length is related to the model complexity and the overall training convergence speed. The three lengths of 48, 72, and 96 are discussed in Table 7.

In Table 7, we test parameters and Dice Loss (ET, TC, WT and average) at 48, 72, and 96, respectively. Accuracy and parameters are in direct proportion to length. We also test the model parameters with a length of 120, which amounts to over 36 M. The larger amount of data in the model leads to less computing speed and more convergence time. Therefore, a length of 96 is chosen for our model.

Table 7. Ablation Study on Sequence Length.

Dim	Dice Score				Parameters
	ET	TC	WT	AVG	
48	0.7678	0.7969	0.8658	0.8102	6,150,794
72	0.7857	0.8240	0.8814	0.8304	13,341,074
96	0.8188	0.8857	0.8927	0.8644	23,353,754

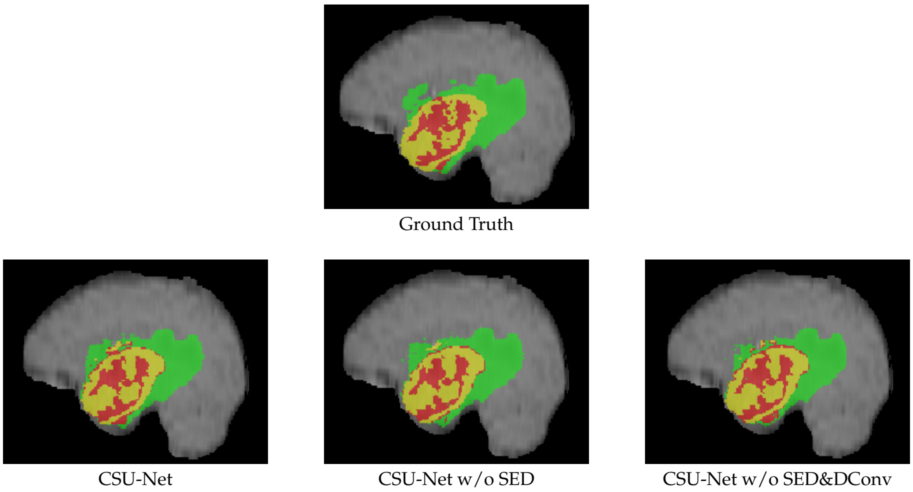

Figure 7. The visible MRI brain tumour segmentation results. Green, yellow and red represent the WT, ET and TC. ROW-1: Ground Truth(GT). ROW-2: Prediction result.

6. Conclusions

In this paper, we introduce a novel CNN-Transformer-based architecture, dubbed as CSU-Net. It presents a parallel hybrid segmentation framework that effectively fuses 3D Swin Transformer and CNN into the network framework for 3D multimodal brain tumour segmentation in MRI. We proposed to fuse the CNN and Transformer to increase the capability for capturing the global–local information and learning long-range spatial dependencies.

We validated the effectiveness of CSU-Net on the BraTS 2020 dataset. CSU-Net achieves 0.8927, 0.8857, and 0.8188 for the WT, TC, and ET, respectively, outperforming competing methodologies. Our method provides clinical assistance in brain tumour location and diagnosis. However, there exist some shortcomings in our proposed method. Firstly, our test was only performed on BraTS 2020, which lacks promotion in diverse medical images. Secondly, there is still further room to reduce our model's parameters. We hope to propose a more lightweight Transformer network and work towards developing more efficient medical image segmentation models in future.

Author Contributions: Conceptualization, Y.C. and M.Y.; Data curation, Y.C.; Formal analysis, M.Y.; Methodology, M.Y.; Project administration, Y.L.; Writing—review & editing, Q.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Bingtuan Science and Technology Program: No. 2019BC008, Bingtuan Science and Technology Program: No. 2022DB005.

Data Availability Statement: Not applicable.

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

References

1. Wesseling, P.; Capper, D. WHO 2016 Classification of gliomas. *Neuropathol. Appl. Neurobiol.* **2018**, *44*, 139–150. [CrossRef] [PubMed]
2. Ouahabi, A. A review of wavelet denoising in medical imaging. In Proceedings of the International Workshop on Systems, Signal Processing and Their Applications, Algiers, Algeria, 12–15 May 2013.
3. Ouahabi, A. *Signal and Image Multiresolution Analysis*; ISTE-Wiley: London, UK; Hoboken, NJ, USA, 2013.
4. Haneche, H.; Ouahabi, A.; Boudraa, B. New mobile communication system design for Rayleigh environments based on compressed sensing-source coding. *IET Commun.* **2019**, *13*, 2375–2385. [CrossRef]
5. Mahdaoui, A.E.; Ouahabi, A.; Moulay, M.S. Image Denoising Using a Compressive Sensing Approach Based on Regularization Constraints. *Sensors* **2022**, *22*, 2199. [CrossRef] [PubMed]

6. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
7. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *40*, 834–848. [CrossRef] [PubMed]
8. Wang, W.; Zhou, T.; Yu, F.; Dai, J.; Konukoglu, E.; Gool, L.V. Exploring Cross-Image Pixel Contrast for Semantic Segmentation. In Proceedings of the International Conference on Computer Vision, Online, 22–24 September 2021.
9. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.H.; Chen, Y.W.; Wu, J. UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. In Proceedings of the International Conference on Acoustics Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020.
10. Zeng, Z.; Xie, W.; Zhang, Y.; Lu, Y. RIC-Unet: An Improved Neural Network Based on Unet for Nuclei Segmentation in Histology Images. *IEEE Access* **2019**, *7*, 21420–21428. [CrossRef]
11. 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**. [CrossRef]
12. Zhang, Z.; Zhang, W.; Sun, B. Pyramid Medical Transformer for Medical Image Segmentation. *arXiv* **2021**. [CrossRef]
13. Zhou, M.; Gao, Y.; Metaxas, D.N. UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation. *arXiv* **2021**. [CrossRef]
14. Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; Liu, J. CE-Net: Context Encoder Network for 2D Medical Image Segmentation. *IEEE Trans. Med. Imaging* **2019**, *38*, 2281–2292. [CrossRef] [PubMed]
15. 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, Athens, Greece, 17–21 October 2016.
16. Valanarasu, J.M.J.; Sindagi, V.A.; Hacihaliloglu, I.; Patel, V.M. KiU-Net: Overcomplete Convolutional Architectures for Biomedical Image and Volumetric Segmentation. *arXiv* **2020**. [CrossRef] [PubMed]
17. Roth, H.R.; Oda, H.; Hayashi, Y.; Oda, M.; Shimizu, N.; Fujiwara, M.; Misawa, K.; Mori, K. Hierarchical 3D fully convolutional networks for multi-organ segmentation. *arXiv* **2017**, arXiv:1704.06382. [CrossRef]
18. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. *arXiv* **2021**, arXiv:2105.05537. [CrossRef]
19. Zhou, H.Y.; Guo, J.; Zhang, Y.; Yu, L.; Wang, L.; Yu, Y. nnFormer: Interleaved Transformer for Volumetric Segmentation. *arXiv* **2021**, arXiv:2109.03201. [CrossRef]
20. Lin, A.; Chen, B.; Xu, J.; Zhang, Z.; Lu, G. DS-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation. *arXiv* **2021**. [CrossRef]
21. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306. [CrossRef]
22. Jia, Q.; Shu, H. BiTr-Unet: A CNN-Transformer Combined Network for MRI Brain Tumor Segmentation. *arXiv* **2021**, arXiv:2109.12271. [CrossRef]
23. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
24. Milletari, F.; Navab, N.; Ahmadi, S.A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the International Conference on 3D Vision, Stanford, CA, USA, 25–28 October 2016.
25. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging* **2015**, *34*, 1993–2024. [CrossRef] [PubMed]
26. Bakas, S.; Akbari, H.; Sotiras, A.; Bilello, M.; Rozycki, M.; Kirby, J.; Freymann, J.; Farahani, K.; Davatzikos, C. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* **2017**, *4*, 170117. [CrossRef] [PubMed]
27. Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R.T.; Berger, C.; Ha, S.M.; Rozycki, M.; et al. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. *arXiv* **2018**, arXiv:1811.02629. [CrossRef]
28. 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 International Conference on Computer Vision, Online, 22–24 September 2021.
29. Zhang, W.; Yang, G.; Huang, H.; Yang, W.; Xu, X.; Liu, Y.; Lai, X. ME-Net: Multi-encoder net framework for brain tumor segmentation. *Int. J. Imaging Syst. Technol.* **2021**, *31*, 1834–1848. [CrossRef]
30. Liu, C.; Ding, W.; Li, L.; Zhang, Z.; Pei, C.; Huang, L.; Zhuang, X. Brain Tumor Segmentation Network Using Attention-Based Fusion and Spatial Relationship Constraint. In Proceedings of the International MICCAI Brainlesion Workshop, Lima, Peru, 4–8 October 2020.
31. Ghaffari, M.; Sowmya, A.; Oliver, R. Brain tumour segmentation using cascaded 3D densely-connected U-net. In Proceedings of the International MICCAI Brainlesion Workshop, Lima, Peru, 4–8 October 2020.