nnUnetFormer: An Automatic Method Based on nnUnet and Transformer for Brain Tumor Segmentation with Multimodal MR Images

Shunchao Guo, Qijian Chen, Li Wang, Lihui Wang, Yue-Min Zhu

nnUnetFormer: An Automatic Method Based on nnUnet and Transformer for Brain Tumor Segmentation with Multimodal MR Images

Shunchao Guo [1] [2], Qijian Chen [1], Li Wang [1], Lihui Wang [1], Yuemin Zhu [3]

- [1] Engineering Research Center of Text Computing & Cognitive Intelligence, Ministry of Education, Key Laboratory of Intelligent Medical Image Analysis and Precise Diagnosis of Guizhou Province, State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang, People's Republic of China.
- [2] Key Laboratory of Complex Systems and Intelligent Optimization of Guizhou Province, Institute of Big Data Application and Artificial Intelligence, School of Computer and Information, Qiannan Normal University for Nationalities, Duyun, People's Republic of China.
- [3] CREATIS, CNRS UMR 5220, Inserm U1294, INSA Lyon, University of Lyon, F-69621 Lyon, France.

Abstract：

Objective: Both local and global context information is crucial semantic features for brain tumor segmentation, while almost all the CNN-based methods cannot learn global spatial dependencies very well due to the limitation of convolution operations. The purpose of this paper is to build a new framework to make full use of local and global features from multimodal MR images for improving the performance of brain tumor segmentation.

Approach: A new automated segmentation method named nnUnetFormer was proposed based on nnUnet and transformer. It fused transformer modules into the deeper layers of the nnUnet framework to efficiently obtain both local and global features of lesion regions from multimodal MR images.

Main Results: We evaluated our method on BraTS 2021 dataset by 5-fold cross-validation and achieved excellent performance with Dice similarity coefficient (DSC) 0.936, 0.921 and 0.872, and 95th percentile of Hausdorff distance (HD95) 3.96, 4.57 and 10.45 for the regions of whole tumor (WT), tumor core (TC), and enhancing tumor (ET), respectively, which outperformed recent state-of-the-art methods in terms of both average DSC and average HD95. Besides, ablation experiments showed that fusing transformer into our modified nnUnet framework improves the performance of brain tumor segmentation, especially for the TC region. Moreover, for validating the generalization capacity of our method, we further conducted experiments on FeTS 2021 dataset and achieved satisfactory segmentation performance on 11 unseen institutions with DSC 0.912, 0.872 and 0.759, and HD95 6.16, 8.81 and 38.50 for the regions of WT, TC, and ET, respectively.

Significance: Extensive qualitative and quantitative experimental results demonstrated that the proposed method has competitive performance against the state-of-the-art methods, indicating its interest for clinical applications.

Keywords: brain tumor segmentation, multimodal MR images, nnUnet, transformer

---

1. Introduction

Brain tumor is an abnormal cell population that occurs in the brain (Jayalakshmi *et al.* 2016, Lapointe *et al.* 2018), which has fateful health hazards and high mortality. In clinic, brain tumor segmentation (BTS), which aims at accurately delineating different semantic sub-regions of brain tumor images, can provide clinicians with fundamental guidance and quantitative assessment to assist subsequent clinical diagnosis, treatment planning, progression monitoring (Liu *et al.* 2021a, Ghaffari *et al.* 2022). At present, structural magnetic resonance imaging (MRI) is a commonly used technology for brain tumor clinical routine examination and diagnosis, since it can offer good soft tissue contrast without radiation in a non-invasive way (Moser *et al.* 2009). Typically, the widely used MR sequences are T1-weighted (T1), T2-weighted (T2), contrast-enhanced T1-weighted (T1ce) and fluid attenuated inversion recovery (Flair), which can provide different information about the tumor. In this work, we refer to different imaging sequences as modalities. Compared to single modality, multiple modalities provide complementary information for analyzing different sub-regions of brain tumors (Zhou *et al.* 2021). Even so, accurate brain tumor segmentation in MR images is still a challenging task due to the well-known fact that brain tumors are highly heterogeneous in terms of the location, size, structure, and morphological appearance (Wang *et al.* 2022). Currently, the gold standard method for brain tumor segmentation is manual segmentation (Ghaffari *et al.* 2022, Zeineldin *et al.* 2022) while the latter is quite laborious, time-consuming, subjective, and highly dependent on the experience of clinicians.

Therefore, in clinical settings, a reliable automated brain tumor segmentation method is highly necessary and beneficial.

Recently, with the development of deep learning methods, convolutional neural network (CNN) has been effectively applied to learn high dimensional discriminative features instead of limited hand-crafted features from data and achieved state-of-the-art performance in various medical image analysis tasks (Litjens *et al.* 2017). Among various CNN architectures, the U-Net as well as its variants make the feature extraction more efficient by building a symmetric encoder-decoder structure with skip connections (Jia *et al.* 2022a) and stand out as the most promising approaches for medical image segmentation tasks. In the context of the multimodal Brain tumor Segmentation Challenge (BraTS), which is well known as the benchmark dataset with fully-annotated multi-parametric MR images for evaluating automatic brain tumor segmentation algorithms, U-Net has become the baseline network and has been often more or less modified for better performance. For instance, the recent winning contributions of BraTS 2018 and 2019 extended the U-Net architecture by respectively adding Variational Autoencoder (VAE) (Myronenko *et al.* 2018) and modifying the initial U-Net as a two-stage cascaded network (Lyu *et al.* 2020). Furthermore, a so-called no new U-Net (nnUnet) (Isensee *et al.* 2021), a modified U-Net, made the initial U-Net adaptive and won the first place in both BraTS 2020 (Isensee *et al.* 2020) and BraTS 2021 (Luu *et al.* 2022). Despite the above CNN-based methods having prominent feature representation ability and yielding good performance to some extent, their capability of modeling long-range spatial dependencies is restricted to the limited receptive fields of their convolution kernels, thereby raising challenges to capture explicit global context information that is critical for accurate segmentation of targets with various shapes and sizes. Moreover, introducing fully connected layers to the network can improve the model's global feature expression ability to some extent, but this would greatly increase the number of learning parameters and memory usage for the model. Therefore, segmentation tasks demand a better solution to improve the efficiency of modeling global context information while maintaining the strong power of capturing localized details (Zhang *et al.* 2021a).

Inspired by its success in natural language processing, the attention mechanism was introduced into CNN models to promote the global features extraction ability of the CNN-based methods, leading to channel attention (Hu *et al.* 2018), temporal-spatial attention (Chen *et al.* 2022), axial-attention (Luu *et al.* 2022), and self-attention (Wang *et al.* 2018). Especially, transformer (Vaswani *et al.* 2017), initially proposed to model long-range spatial dependencies based on the self-attention mechanism in sequence-to-sequence prediction tasks, has attracted tremendous interest in the field of natural language processing and computer vision. Compared with previous CNN-based methods that are built solely on convolutional layers, transformers overcome their limitation of locality and therefore makes predictions with more considerations (Jia *et al.* 2022a). In view of different and complementary properties between convolution and transformer, there exists a potential possibility to benefit from both paradigms by integrating them together. Hence, in order to obtain better performance in brain tumor segmentation, numerous efforts (Jia *et al.* 2022a, Wang *et al.* 2021, Li *et al.* 2022) have investigated the effectiveness of integrating the transformer layers into the encoder of CNN-based architectures in a plug-and-play manner. Unfortunately, their segmentation performance for small targets in tumor core region still needs to be further improved, whereas these targets play an important role in estimating the tumor and prognosis for the patients in clinical practice.

In this paper, we propose an innovative end-to-end semantic segmentation method, called nnUnetFormer. The idea is to combine the localization power of nnUnet and the global context sensitivity of transformer for further improving the performance of brain tumor segmentation in multimodal MR images. The proposed method consists of embedding Tranformer modules into the deeper layers of the modified nnUnet to more efficiently extract local and global features, which enables us to segment out small targets more precisely.

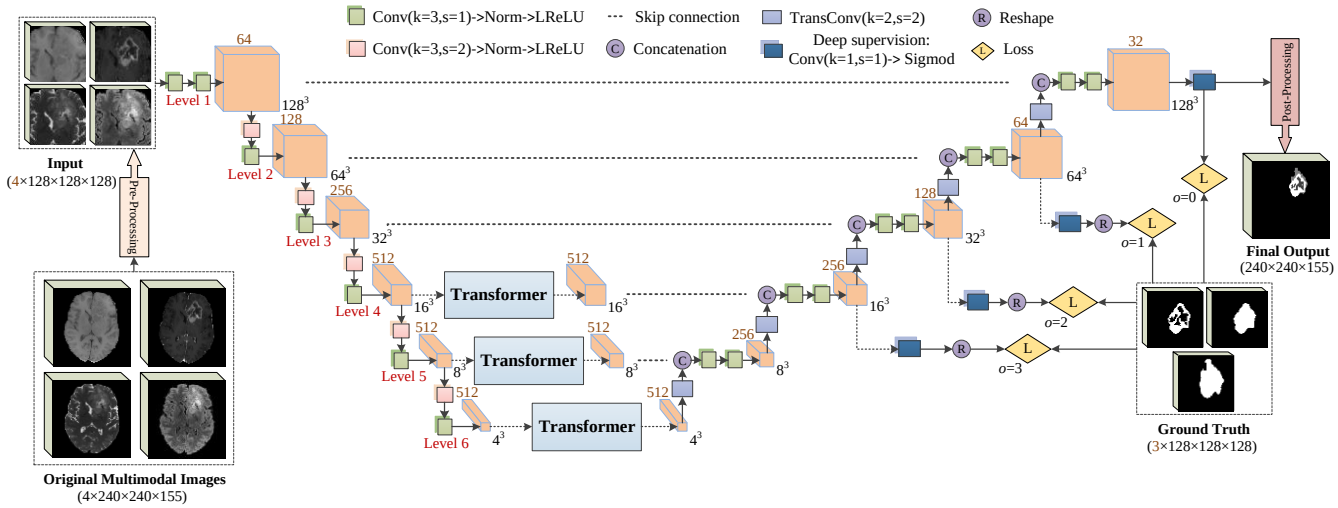

Figure 1. Overview of our proposed nnUnetFormer Framework. Between the input and the post-processing is our designed network architecture.

2. Methodology

Among the deep learning-based methods for brain tumor segmentation, the most common way is to construct a CNN-based network, especially the U-Net, to

encode images into high-level feature representations and then decode them back to the full spatial resolution.

2.1. Architecture overview

Overall, our nnUnetFormer consists of a modified nnUnet as the main network framework and the embedded transformer modules, as illustrated in Figure 1. It takes the multimodal MR images of patch size 128×128×128 as input to obtain the fused local and global features, and finally outputs the predicted segmentation maps. We first utilize the encoder of our nnUnetFormer to efficiently generate a series of feature maps by extracting multi-scale spatial features. In this way, rich local 3D context features from the encoder are effectively embedded in high-level features step by step, which are subsequently fed into the transformer modules to further learn the long-range dependencies and supplement the local context features with a global receptive field. After that, we repeatedly apply the concatenation operation on the corresponding two branches of up-sampling and skip-connection, followed by convolutional layers to gradually produce a high-resolution segmentation map.

Concretely, since the features in shallower layers tend to generate more detailed spatial information and those in deeper layers extract more discriminative semantic information (Jia *et al.* 2022b), we have modified the original nnUnet architecture by using a deeper encoder, which was inspired by Luu et al. (2022). This modification aims to increase the network capacity to better model greater data diversity. Here, we used 6 levels of same-resolution convolutional layers and 5 strided convolution (stride=2) operations for down-sampling in the encoder to extract more discriminative semantic information. The decoder followed the same structure with a concatenation operation and two convolution operations. Note that the concatenation operation is used to concatenate the transpose convolution (TransConv) operations for up-sampling with the skip features from the encoder branch at the same-resolution level. To be more specific, the combination of batch normalization and Leaky ReLU (LReLU) with negative slope of 0.01 was applied after every convolution operation; the number of channels was set from 4 to 64 at the initial level (level 1) and doubled for every down-sampling. Besides, with respect to the initial nnUnet architecture, additional sigmoid outputs were added to every resolution except for the two lowest levels as auxiliary supervision branches according to deep supervision learning technique (Wang *et al.* 2015) for enhancing gradient propagation to the earlier layers as well as improving the speed of deeper network convergence in training phase.

2.2. Transformer module

We found that due to the GPU memory constraint, it was inadvisable to apply the transformer module to the high-resolution features. Hence, we fused up to three transformer modules into the three lowest resolution scales (Level 4, Level 5, and Level 6) of our modified nnUnet while keeping the skip connections unchanged. The structure of the transformer module is illustrated in Figure 2. Specifically, assuming the input of our transformer layer is the feature map $F \in R^{C\times W\times H\times D}$, which firstly goes through a convolution operation to change its channel dimension from C to the preassigned number n, and then reshapes the 3D spatial dimensions of the new feature representations after the convolution operation as 1D dimension of size S (numerically equal to $W\times H\times D$). This process is named linear projection, resulting in feature map $F' \in R^{n\times S}$. After the linear projection, we add a 1D learnable positional embedding (PE) module to F', forming S n-dimensional embedded patches (or tokens). Subsequently, a stack of transformer layers, which mainly consists of multi-head attention (MHA) and multilayer perceptron (MLP) sub-layers, was used to learn long-range dependencies with global receptive field. Let F_l denote the output of the l-th ($l \in [1,2,...,L]$) transformer layer, it can be calculated by:

$$F_{l-1}^{'} = MHA(LN(F_{l-1})) + F_{l-1} \tag{1}$$

$$F_l = MLP(LN(F_{l-1}^{'})) + F_{l-1}^{'} \tag{2}$$

where $LN(*)$ is the layer normalization. In transformer module, the core element is a MHA sub-layer, which was used to update the states of each embedded patch by aggregating information globally. The MHA sub-layer comprises P parallel self-attention (SA) heads and one linear layer. Each head has its own learnable weight matrices. The SA block is a parameterized function that learns the mapping between a query and the corresponding key and value representations in a feature Fa (Hatamizadeh *et al.* 2022). Let Q, K, and V denote respectively query, key and value; the output of i-th SA head is defined as

$$\mathrm{SA}(Fa_i) = \mathrm{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d}}\right) V_i \tag{3}$$

where $i \in [1,2,...,P]$, and d is the dimension of Q and K. The MLP sub-layer is composed of two linear layers linked by one GELU activation function. To fit the input dimension of 3D CNN decoder, we use a feature mapping module to project the patches back to the feature map, which owns the same dimension as the input of the transformer module.

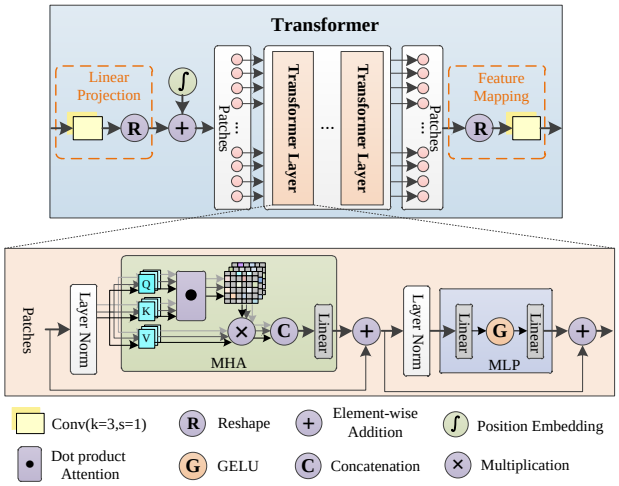

Figure 2. Illustration of the transformer module.

2.3. Loss function

In MR images of brain tumor patients, the volume of WT region is much smaller than that of the whole brain, while either TC or ET region occupies even less, which makes the BTS task suffer from a serious class imbalance problem. To alleviate the class imbalance problem, we adopt a combination of Dice (Milletari *et al.* 2016) and weighted cross-entropy (WCE) (Akil *et al.* 2020) losses as the training loss for our network, each of which is expressed as follows:

$$L_{Dice} = 1 - 2\frac{\sum_{j=0}^{M} g_j p_j}{\sum_{j=0}^{M} (g_j + p_j)} \tag{4}$$

$$L_{WCE} = -\sum_{j=0}^{M} w_j g_j \log(p_j) \tag{5}$$

$$L_{seg} = L_{Dice} + L_{WCE} \tag{6}$$

where M is the number of classes, g_j the ground-truth for class j, p_j the model prediction for class j, and w_j the weight of class j.

In view of sigmoid outputs we added (Figure 1), our network was trained to optimize the combined loss, calculated at the final full resolution output as well as at the auxiliary outputs of lower resolution. The combined loss $L_{combined}$ is defined as

$$L_{combined} = \sum_{o=0}^{O} weight_o \times L_{seg}(o) \tag{7}$$

where O represents the number of the outputs at various resolutions, o the o-th output of the model, $L_{seg}(o)$ the loss of the o-th output. Especially, the weights for each output from full resolution (o=0) to lower resolution are calculated as follows:

$$weight_o = 2^{-o} / \sum_{o=0}^{O} 2^{-o} \tag{8}$$

2.4. Datasets

The multimodal magnetic resonance images used for the model training and evaluation in this study were obtained from the publicly available datasets of BraTS 2021 and Federated Tumor Segmentation challenge (FeTS) 2021 (Isik-Polat *et al.* 2022). In BraTS 2021 and FeTS 2021, their training sets contain 1251 and 341 cases collected from multiple institutions, respectively. In FeTS 2021, the cases in the training set were partitioned into 22 clients according to their original institutions and tumor sizes (Yin *et al.* 2022), and the statistical results of the cases in each client is illustrated in Figure 3. In the training set of both datasets, each case contains four 3D modalities (T1, T2, T1ce and Flair) MR images along with the corresponding segmentation annotations of tumor sub-types. All the MR images and the associated annotations were resized to 240×240×155 and stored as NIfTI files (.nii.gz). Following the two challenges, all the MR images have been preliminarily pre-processed by organizers: co-registered to the same anatomical template, resampled to 1 mm^3 isotropic

resolution, and skull-stripped. The annotations of each case have been achieved manually following the same annotation protocol and comprise the GD-enhancing tumor (label 4), the peritumoral edema (ED, label 2), the necrotic and non-enhancing tumor (NCR/NET, label 1), and the background (label 0). Note that label 3 does not appear here since it was merged into label 1 by organizers. Furthermore, the almost homogeneous annotations have been clustered together to compose three mutually inclusive tumor sub-regions, representing the enhancing tumor (ET: label 4), tumor core (TC: label 1 + label 4), and whole tumor (WT: label 1 + label 2 + label 4), respectively.

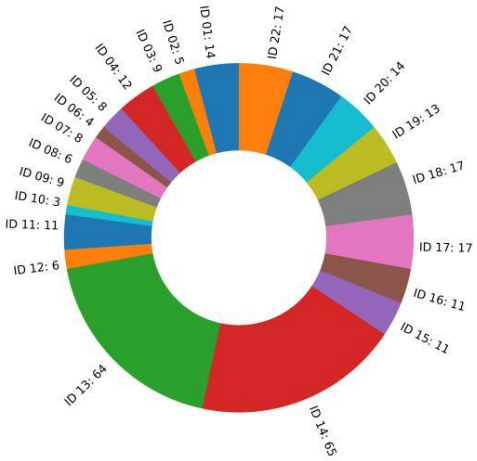

Figure 3. The statistical results of the cases in each client on FeTS 2021 dataset. ID represents the identifier of client.

As for BraTS 2021, to overcome the bias caused by a particular selection of the pair of training and testing sets, each network was trained from scratch and evaluated using 5-fold cross-validation. Hence, the performance evaluated on this dataset was measured using the averaged evaluation metrics from the 5 validation folds. Moreover, since the purpose of task 2 in FeTS 2021 challenge was to evaluate the robustness and generalization of different segmentation algorithms (Yin *et al.* 2022, Nalawade *et al.* 2022), we further used FeTS 2021 to conduct more experiments to evaluate the generalization capability of our method. Here, following the principle of making up the training set with as few the institutions or centers as possible, we used 257 cases from clients 11, 13, 14, 15, 16, 17, 18, 19, 20, 21 and 22 for training and the remaining 84 cases from other unseen 11 clients (independent institutions, to be precise) for testing with an approximate 3:1 ratio.

2.5. Evaluation metrics

To effectively evaluate the performance of the proposed nnUnetFormer method, the official evaluation metrics of Dice similarity coefficient (DSC) and 95th percentile of Hausdorff distance (HD95) were used in this study. DSC was used to measure the spatial overlap between the segmentation predictions and the ground-truth annotations, which is numerically equal to 1-L_{Dice}. Hausdorff distance (HD) originates from set theory and measures the maximum distance of a point set to the nearest point in another set. As opposed to DSC, HD represents a global measure for spatial overlap and is sensitive to local differences (Momin *et al.* 2022). DSC and HD are defined by:

$$DSC = \frac{2|G \cap P|}{|G| + |P|} \tag{9}$$

$$HD(G,P) = \max\{\max_{g \in G} \min_{p \in P} ||g - p||, \max_{p \in P} \min_{g \in G} ||p - g||\} \tag{10}$$

where G denotes the ground-truth annotations, and P the segmentation predictions. For each metric, three tumor regions of ET, TC and WT were evaluated individually, and we took the average DSC of the three regions as the main metric to evaluate the performances of trained models.

2.6. Implementation details

Our proposed nnUnetFormer method is implemented using PyTorch. Keeping consistent with the training methodology of nnUnet, we also adopted the region-based strategy for training instead of predicting the mutually exclusive brain tumor sub-types, and used the cropping and z-score normalization operations for original MR images in the pre-processing step. Considering the over-fitting phenomenon is almost unavoidable in deep learning methods when

training a large neural network containing a huge number of parameters to learn with limited training data (Liu *et al.* 2021a, Zhang *et al.* 2021b), we applied a large variety of spatial data augmentation techniques, including random rotation, random cropping, random scaling, random elastic deformation, and random gamma correction on the fly during network training to boost the generalization capability of our model. In the training phase, the initial learning rate was set to 1×10^{-2} with the updating strategy of scheduler optimization. The optimizer used stochastic gradient descent algorithm with weight decay of 5×10^{-5} and Nesterov momentum of 0.99. The batch size was set to 2. The number of training epoch was set to 200. The number of SA heads in transformer layer was set to 8.

Table 1. Ablation study of the proposed network architecture with different modifications on validation set from BraTS 2021. Bold value indicates best result, underline value runner-up result. AVG means the average score of WT, TC and ET. # indicates p-value < 0.05.

Model	DSC				HD95				FLOPs (G)	Params (M)
	WT	TC	ET	AVG	WT	TC	ET	AVG		
A: nnUnet (baseline)	0.933	0.906	0.858	0.899	4.78	6.81	12.29	7.96	502.50	29.75
B: A+Deep Supervision	0.936#	0.909#	0.857	0.901#	4.21#	5.49#	11.74	7.14#	502.59	29.75
C: B+Deeper U-Net	0.938#	0.911#	0.862#	0.904#	4.76#	4.88#	11.64#	7.09#	1661.11	83.36
D: C+WCE loss	0.936	0.910#	0.868#	0.905#	5.10#	5.52#	10.89#	7.17#	1661.11	83.36
nnUnetFormer (Level 4)	0.936#	0.921#	0.870#	<u>0.909</u>#	4.29#	4.49#	11.38#	6.72#	1685.17	89.38
nnUnetFormer (Level 5)	<u>0.937</u>#	<u>0.920</u>#	0.871#	<u>0.909</u>#	<u>4.02</u>#	<u>4.55</u>#	10.88#	<u>6.48</u>#	1664.12	89.38
nnUnetFormer (Level 6)	0.932#	0.917#	0.874#	0.908#	5.03#	5.43#	10.19#	6.88#	1661.48	89.38
nnUnetFormer (Levels 4-6)	0.936#	0.921#	<u>0.872</u>#	0.910#	3.96#	4.57#	<u>10.45</u>#	6.32#	1688.56	101.41

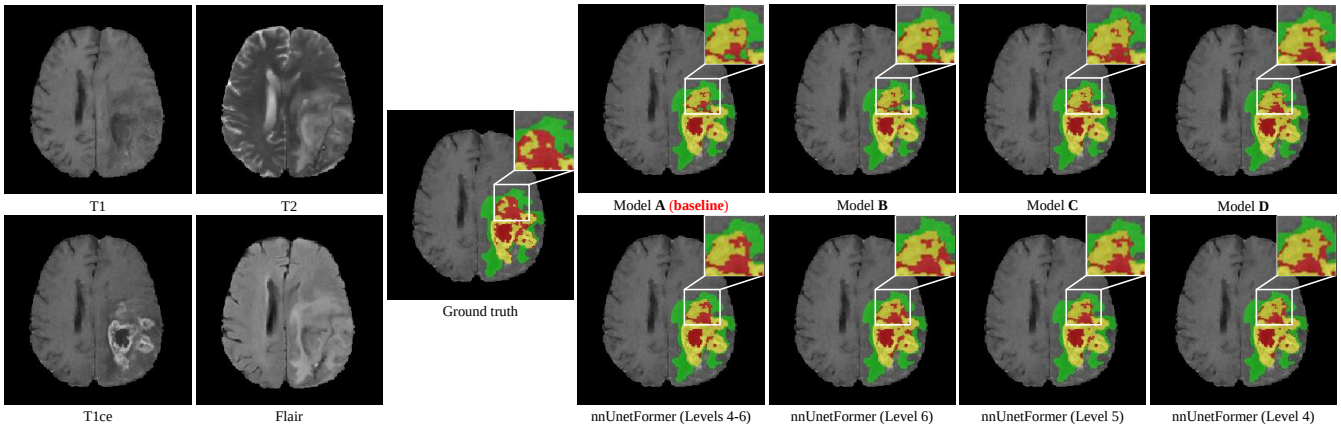

Figure 4. An example of segmentation results using different modifications based on nnUnet baseline. WT: red+yellow+green, TC: red+yellow, ET: yellow. The patient ID is BraTS2021_00105. Model A: nnUnet, model B: A+Deep Supervision, model C: B+Deeper U-Net, model D: C+WCE loss.

3. Experiments and results

3.1. Ablation studies

In this study, we took the original nnUnet as baseline. Only one transformer module was embedded into level 4, level 5, level 6, thus leading to nnUnetFormer (Level 4), nnUnetFormer (Level 5), and nnUnetFormer (Level 6), respectively. When embedding one transformer module in each of levels 4, 5, 6, we have nnUnetFormer (Levels 4-6). To validate the performance under different modifications on baseline, a variety of ablation studies were performed. Paired Wilcoxon signed-rank tests were conducted for the results between the nnUnet baseline and our modified models at the significance level of 0.05. Firstly, the quantitative validation results are given in Table 1. We can observe that our nnUnetFormer (Levels 4-6) model achieves best overall performance both with DSC 0.936, 0.921 and 0.872, and with HD95 3.96, 4.57 and 10.45 for the three regions of WT, TC, and ET, respectively. Especially, the overall performance of our nnUnetFormer (Levels 4-6) model was slightly better than baseline, with the average DSC increased by 1.22% (0.910 vs 0.899). Besides, incorporating the modifications into baseline led to a consistent improvement in terms of average DSC. In particular, integrating the three transformer modules into our modified nnUnet architecture (model D) improved the performance for the task of BTS, with the average DSC increased by 0.55% (0.910 vs 0.905). As for our nnUnetFormer (Level 4), nnUnetFormer (Level 5), and nnUnetFormer (Level 6) models, we found that no matter the transformer module was embedded into Level 4, Level 5 or Level 6, we achieved better segmentation performance in terms of both average DSC and average HD95 compared with other models without using transformer module. By comparing the runtime (FLOPs) and the number of learning parameters of each model, it can be seen that our nnUnetFormer model has no advantages. Secondly, for better understanding our nnUnetFormer method, we also illustrate an example of segmentation results using different modifications based on nnUnet baseline on BraTS 2021, as shown in Figure 4. From Table 1 and Figure 4, we can observe that the segmentation results are gradually improved when the proposed modifications are integrated, especially for the TC and ET regions.

Table 2. Quantitative comparison of different methods on BraTS 2021 dataset. Bold value indicates best result, underline value runner-up result. * means the results were obtained from the corresponding paper directly, and AVG means the average score of WT, TC and ET. # indicates p-value < 0.05.

Method	DSC				HD95			
	WT	TC	ET	AVG	WT	TC	ET	AVG
HNF-Netv2[0]*	0.925	0.880	0.848	0.884	3.46	5.86	14.18	7.83
ResUNet[0]*	0.899	0.850	0.820	0.856	4.30	9.89	17.89	10.69
nnSegNet[0]*	0.928	0.878	0.847	0.884	<u>3.47</u>	9.34	7.56	<u>6.79</u>
BiTr-Unet[0]*	0.910	0.843	0.819	0.857	4.51	16.69	17.85	13.02
E1D3 U-net[0]*	0.924	0.865	0.822	0.870	4.23	9.61	19.73	11.19
SGEResU-Net[0]*	0.916	0.869	0.833	0.873	5.95	7.57	19.28	10.93
MMEF-nnUNet[0]*	0.928	0.901	0.873	0.901	5.10	9.68	<u>10.09</u>	8.29
VT-UNet[0]	0.926#	0.887#	0.862#	0.892#	5.25#	6.27#	11.23#	7.58#
SegResNet[0]	0.928#	0.884#	0.869#	0.894#	4.62#	<u>5.57</u>#	11.78#	6.99#
Extending nnUnet[0]	<u>0.933</u>	<u>0.911</u>#	0.868#	<u>0.904</u>#	4.47	5.62#	10.73#	6.94#
nnUnetFormer (Levels 4-6)	0.936	0.921	<u>0.872</u>	0.910	3.96	4.57	10.45	6.32

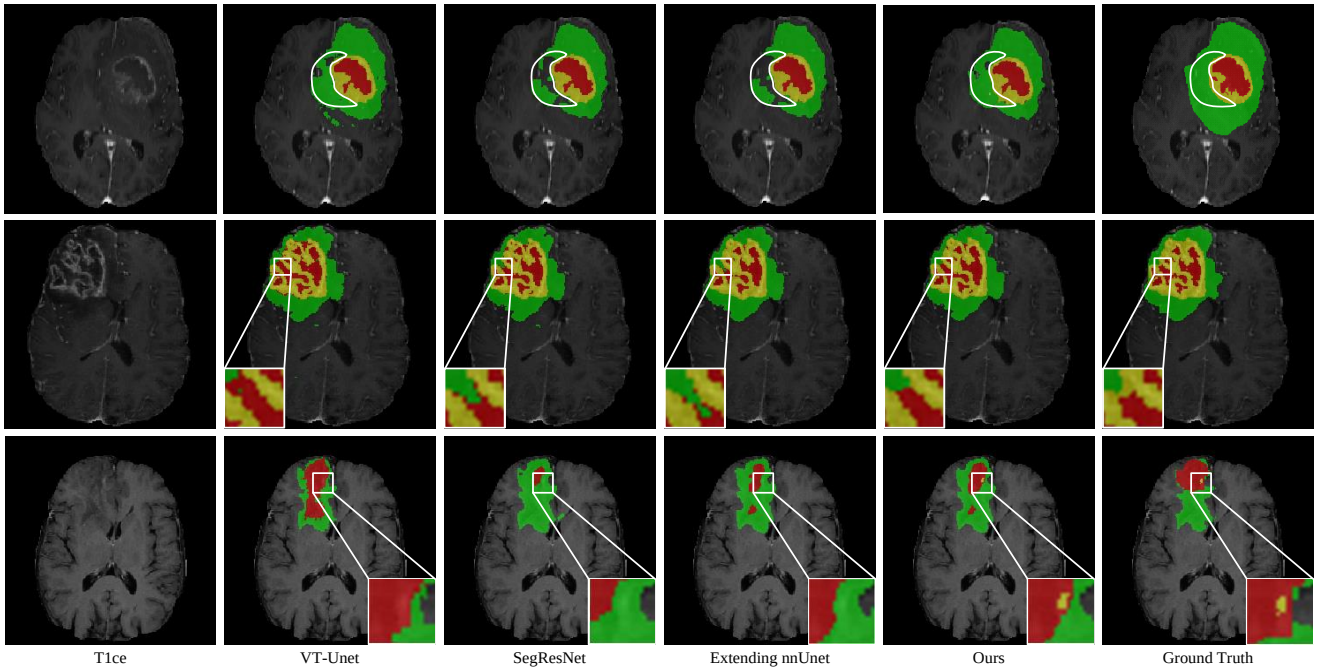

Figure 5. Illustration of qualitative comparison using different methods on BraTS 2021 dataset. WT: red+yellow+green, TC: red+yellow, ET: yellow. From row 1 to 3, the patient ID is BraTS2021_00002, BraTS2021_00074, and BraTS2021_01503, respectively.

3.2. Comparison with state-of-the-art methods

We quantitatively compared our proposed nnUnetFormer method in our internal 5-fold cross-validation split against other recent state-of-the-art methods such as HNF-Netv2 (Jia *et al.* 2022b), ResUNet (Pei *et al.* 2022), nnSegNet (Jabareen *et al.* 2022), BiTr-Unet (Jia *et al.* 2022a), E1D3 Unet (Bukhari *et al.* 2022), SGEResU-Net (Liu *et al.* 2022), MMEF-nnUNet (Huang *et al.* 2022), VT-UNet (Peiris *et al.* 2021), SegResNet (Rahman *et al.* 2022), Extending nnUnet (Luu *et al.* 2022) on BraTS 2021 dataset. Considering we have reproduced three networks VT-UNet, SegResNet, and Extending nnUnet according to the public available codes, paired Wilcoxon rank-sum tests for the results were conducted between them and our nnUnetFormer method. From the quantitative experimental results summarized in Table 2, we found that our nnUnetFormer (Levels 4-6) model shows slightly better in terms of average DSC and average HD95 compared with the other methods and outperforms the second best method by 0.66% (0.910 vs. 0.904) and 6.92% (6.32 vs. 6.79), respectively. Specifically, our model also achieved the best DSC scores in WT and TC regions, although its DSC score of ET region is slightly lower than MMEF-nnUNet method. As for TC region, our model outperformed the second best method by 1.1% (0.921 vs. 0.911) in terms of DSC score.

Furthermore, we also conducted qualitative analysis by visually comparing the segmentation results of different methods including VT-UNet, SegResNet, Extending nnUnet and our nnUnetFormer (Levels 4-6). From Figure 5, we observed that different methods have different segmentation results for the three tumor regions, and our segmentation results are relatively close to the ground-truth compared with the other three methods. Specifically, row 1 indicates that our model can segment the peritumoral edema relatively well while the GD-enhancing tumor and the NCR/NET are accurately segmented. Rows 2 and 3 indicate that our model can capture and identify small targets (NCR/NET and GD-enhancing tumor) better. Note that as for the MR images of patient in row 3, the number of voxels in ET region is only 94, which is much smaller than the number of voxels in TC region.

3.3 Generalization ability validation

Taking the segmentation performance, training/validation time, and parameters number of our four different nnUnetFormer models into consideration, we used the nnUnetFormer (Level 6) model to conduct more experiments on FeTS 2021 for evaluating the generalization capacity of our nnUnetFormer method. The results of different unseen clients in the testing set are summarized in Table 3. Notably, although the segmentation performance of our nnUnetFormer model varies from client to client, it meets the requirements of clinical applications for most clients. What is worth mentioning is that there is a significant difference in image quality between the 3 clients (clients 9, 10, 12), thereby a depressing performance drop when our model shifted from the training set to the unseen cases coming from these 3 independent institutions. Especially, for the cases without GD-enhancing tumor, there are 3.5% in our training set, while there are respectively 44.4%, 66.7%, 66.7% in clients 9, 10, 12 and 0.0% in the other 8 unseen clients. Taking all the testing cases into consideration, our nnUnetFormer model achieves overall segmentation performance with DSC 0.912, 0.872 and 0.759, and HD95 6.16, 8.81 and 38.50 for the regions of WT, TC, and ET, respectively. Besides, the overall DSC performance for all cases in the testing set was also demonstrated with box and whisker plots in

Figure 6 for better illustration. From Figure 6, we can easily observe that our nnUnetFormer method achieves high-performance segmentation predictions for WT, TC and ET regions in almost all patients. Finally, we also compared our method with the methods we reproduced such as VT-UNet (Peiris *et al.* 2021), SegResNet (Rahman *et al.* 2022), Extending nnUnet (Luu *et al.* 2022) on FeTS 2021, as shown in Table 4. From Table 4, we can find our results are slightly better for the segmentation with best average DSC and average HD95. The average DSC score of our method exceeds the methods of VT-UNet, SegResNet, and Extending nnUnet by 1.4% (0.848 vs. 0.836), 2.4% (0.848 vs. 0.828), and 0.6% (0.848 vs. 0.843), respectively. Furthermore, more quantitative performance comparison with five state-of-the-art methods using DSC score collected from corresponding papers is shown in Figure 7. Overall, although our method did not pre-process the images coming from different clients, it is better than the other methods in segmenting both WT and TC regions, except that the DSC score of ET region is slightly lower than that of Yin et al.. by 0.1%, which was implemented by training a nnUnet model with all cases in FeTS 2021 training set and took the first place for task 2. Particularly, as for TC region, the DSC score of our method exceeds the methods of Nalawade *et al.* (2022), Mächler *et al.* (2022), Isik-Polat *et al.* (022), Khan *et al.* (0022), and Yin *et al.* (2022) by 12.8% (0.872 vs. 0.773), 10.0% (0.872 vs. 0.793), 17.7% (0.872 vs. 0.741), 22.1% (0.872 vs. 0.714), and 12.8% (0.872 vs. 0.773), respectively.

Table 3. Segmentation result statistics for institutions were not involved in training on FeTS 2021 dataset. AVG means the average score of WT, TC and ET.

Client ID	DSC				HD95			
	WT	TC	ET	AVG	WT	TC	ET	AVG
1	0.922	0.877	0.804	0.868	3.56	5.20	3.27	4.01
2	0.918	0.897	0.869	0.895	2.33	2.68	1.54	2.18
3	0.958	0.965	0.950	0.958	1.50	1.83	1.09	1.47
4	0.928	0.933	0.887	0.916	5.98	2.18	1.67	3.28
5	0.929	0.893	0.892	0.905	3.88	2.77	2.54	3.06
6	0.937	0.898	0.769	0.868	3.76	3.19	4.72	3.89
7	0.909	0.889	0.733	0.843	5.47	7.86	7.61	6.98
8	0.933	0.921	0.873	0.909	3.01	2.01	1.81	2.28
9	0.786	0.631	0.533	0.650	24.21	50.18	127.58	67.32
10	0.958	0.867	0.455	0.760	2.05	4.40	130.24	45.56
12	0.897	0.822	0.248	0.655	6.53	6.11	250.09	87.58
All	0.912	0.872	0.759	0.848	6.16	8.81	38.50	17.82

Table 4. Quantitative comparison of different methods on FeTS 2021 dataset. Bold value indicates best result, and AVG means the average score of WT, TC and ET. # indicates p-value < 0.05.

Method	DSC	HD95

	WT	TC	ET	AVG	WT	TC	ET	AVG
VT-UNet	0.904[#]	0.848[#]	0.756[#]	0.836[#]	6.32[#]	9.77[#]	42.46[#]	19.52[#]
SegResNet	0.912[#]	0.843[#]	0.730[#]	0.828[#]	6.11[#]	7.46[#]	50.48[#]	21.35[#]
Extending nnUnet	0.917[#]	0.864[#]	0.748[#]	0.843[#]	5.81	8.34[#]	43.45[#]	19.20[#]
Ours	0.912	0.872	0.759	0.848	6.16	8.81	38.50	17.82

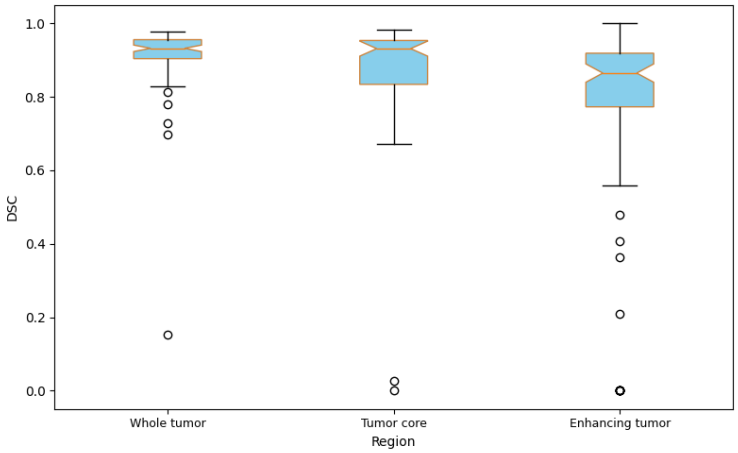

Figure 6. The box and whisker plots of the distribution for DSC scores of testing cases on FeTS 2021 dataset. Each circle means one case.

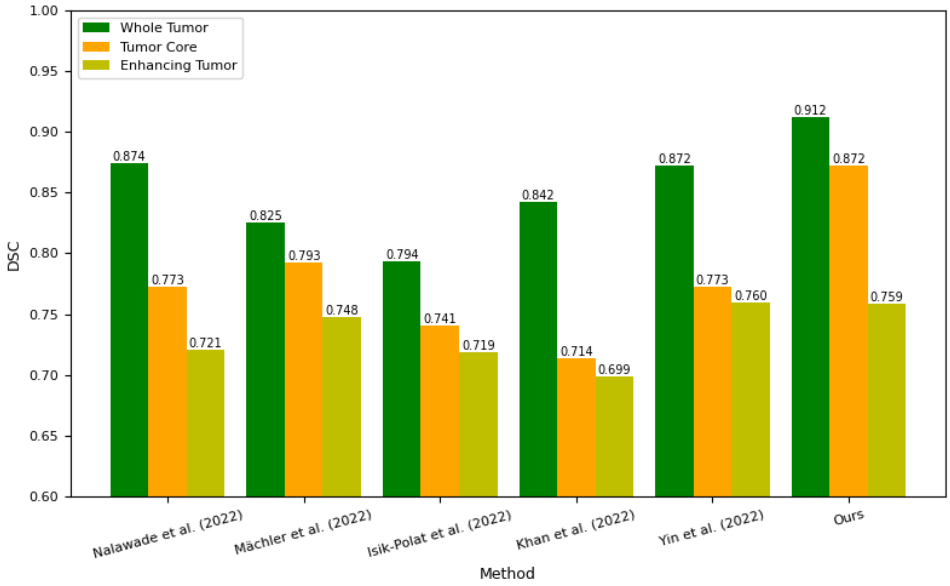

Figure 7. Quantitative comparison of different methods in terms of DSC score on FeTS 2021 dataset.

4. Discussion

We presented a new method, nnUnetFormer, to achieve automatic brain tumor segmentation with multimodal MR images. In our method, based on the nnUnet baseline, we extended it with deeper encoder to model the larger data variety, and added the deep supervision module to the decoder of our network. To alleviate the class imbalance issue caused by the great divergence among the volume of tumor sub-regions, we assigned the class weights for each tumor region to the cross-entropy loss, which is combined with Dice loss to optimize the model training process. More importantly, we embedded the transformer modules into the deeper layers of our network to model long-range dependencies, which further enhances global context features.

A finding of this study is that both local and global context information is all crucial semantic features for brain tumor segmentation; the former is essential to clearly segment boundaries and small targets whereas the latter is effective for modeling the relationship between spatially distant pixels if the targets spread over a large receptive field. Compared to other CNN-based methods, nnUnetFormer not only retains the superiority of convolution in encoding precise local spatial information and producing hierarchical representations, but also inherits the strong power of transformer in modeling long-range dependencies with global receptive field that helps enhance global feature extraction. Therefore, our nnUnetFormer can segment brain tumors more precisely with finer grained details of tumor structures, thus promoting the performance of brain tumor segmentation. The results of our comparison experiments on BraTS 2021 demonstrates the superiority of our nnUnetFormer method over other state-of-the-art methods in BTS task.

As for the multi-class task of brain tumor segmentation, one of the most common challenges is distinguishing small blood vessels in the tumor core region from enhancing tumor region due to the unclear boundaries between adjacent structures as well as the smooth intensity gradients (Isensee *et al*. 2018, Ullah *et al.* 2022). Even worse, the low-grade glioma patients in both BraTS and FeTS datasets may have no GD-enhancing tumor at all. It should be noted that we took the average DSC of the three regions as the main metric to evaluate the performances of trained models in our experiments. In our ablation experiments, we paid more attention on the average DSC of the three regions instead of just one region, although model C achieved the highest DSC for WT. In our method, through adding the modifications into basic nnUnet architecture, the global feature representations were enhanced with a global receptive field, thereby improving the segmentation performance, especially for the TC and ET regions, which have small areas and intricate edges. The quantitative and qualitative comparisons in our ablation studies indicate that the effectiveness of our proposed method.

From the description of the datasets we used in this study, it is easy to conclude that ET can be regarded as a sub-region of TC and TC as a sub-region of WT. Therefore, the whole tumor region is the easiest to be accurately segmented out, while the enhancing tumor region is the most difficult, when we take into account the volume factor of the targets to be segmented. This is consistent with the results of our method as well as the comparison methods listed in Table 2 and Figure 7. As for ET region, we also observed that there is a significant difference in terms of HD95 among different methods. Upon closer inspection, we found that this was caused by several cases (we call them outliers) where ET region was predicted but absent from the ground-truth annotation. According to the BraTS and the FeTS challenges, for the almost correctly predicted cases, their HD95 was scored below 10, whereas for the outliers, their HD95 was scored up to 373.13, thus resulting in a serious skewing. This explanation is consistent with recent studies (Jabareen *et al.* 2022, Kotowski *et al.* 2022). Therefore, the patients without ET region will influence the prediction performance of the ET region and subsequently the overall segmentation performance. That is why many researchers specially re-labeled all enhancing tumor predictions according to the preset threshold based on repeated verification in their post-processing operations (Zeineldin *et al.* 2022, Jia *et al.* 2022a, Luu *et al.* 2022, Jia *et al*. 2022b, Isensee *et al.* 2018).

There is no doubt that enhancing the model's global feature extraction ability is really helpful for improving the performance of brain tumor segmentation. Although using FC layers can improve the model's ability to extract global features to some extent, the number of their parameters is very large due to the fully connected characteristics. Therefore, it was inadvisable to add the FC layers to the high-resolution features directly since the constraint of limited GPU memory. As we know, the transformer network was born to address the inherent locality limitation of CNN-based methods. After careful validation, compared with using FC layers for enhancing global features extraction, the advantage of using transformer is that it can better segment small target regions (TC and ET) with adding fewer model parameters. Besides, simply fusing the vanilla transformer modules into nnUnet in our method inevitably increased the number of parameters and the computational complexity, therefore inference time. Moreover, due to the GPU memory constraint, it was inadvisable to apply the transformer module to the high-resolution features. Currently, swin transformer (Liu *et al*. 2021b), which is based on the hierarchical design and the shifted window scheme, is the most popular way to reduce the computational complexity of the vanilla transformer since its computational complexity is $O(N)$ with N indicating the number of tokens. Hence, as for the higher resolution images or features, SwinT may be a better choice for efficiently extracting the global dependencies.

It is well recognized that most deep learning models exhibit limited generalization ability when applied to the datasets acquired from different imaging devices, parameter settings and patients. For this reason, the generalization problem of trained models caused by different domain distributions between trained samples and unseen samples, should also deserve our special attention, especially in multi-center study. In our method, the discriminative global and local feature representations of brain tumor were learned from images and can provide complementary information to each other, thereby improving the

robustness and generalization of our trained nnUnetFormer models. From the experiments conducted on FeTS 2021, we can see that our proposed method exhibits robust generalization ability on the whole for conducting multi-center studies.

Furthermore, considering that U-Net as well as its variants are the mainstream methods for brain tumor segmentation with multimodal MR images, finding an appropriate U-Net architecture configuration for clinical application is crucial for achieving good segmentation performance. During the implementation of our proposed method, we noticed that our nnUnetFormer model inherits the ease of use and extensibility characteristics of nnUnet, thus can be used as a general and excellent 3D medical image segmentation tool without any complex configurations.

Although encouraging, our method still has several limitations. Firstly, there is no doubt that the cases where ET region was predicted but absent from the ground-truth annotation often have a significant impact on segmentation performance. Considering that most of the strategies for tackling this issue rely on expert-level experiences and require repeated verifications, we didn't address it during the testing phase with additional post-processing operations. Therefore, it would be interesting to develop some strategies to handle this problem for further improving the performance of brain tumor segmentation. Secondly, similarly to most deep learning-based methods, the fusion strategy for multimodal MR images used in the present study might not be optimal because simple channel fusion results as the input to the model may not fully mine the complementary information between modalities. To further improve the performance of brain tumor segmentation with multimodal MR images, we might try to study other image fusion strategies in the future. Finally, the MR images used in this study contained complete modalities for each patient, whereas the situation of missing one or more modalities data is common in clinical practice and this may result in the collapse of the trained models relying on complete modalities data. To handle this problem, more common features of multimodal images might be extracted by learning correlation information among inherent multimodal features.

5. Conclusion

We proposed and evaluated the nnUnetFormer architecture, a combination of nnUnet and transformer specially designed to capture both local and global context information of brain tumors, for the task of segmenting brain tumor into its intrinsic sub-regions automatically with multimodal MR images. Based on the initial nnUnet framework, we first modified it by adding extra outputs, deepening the encoder and adding weights for the cross entropy loss. Then, the efficient transformer was embedded into the deeper layers to model long-range dependencies, thereby enhancing the ability of global features extraction. Extensive experiments on BraTS 2021 and FeTS 2021 datasets showed that our method improved nnUnet and other state-of-the-art methods significantly on average, especially for TC and ET regions, which suggests its potential use in clinical practice for brain tumor segmentation.

References

Jayalakshmi, C., & Sathiyasekar, K. (2016, May). Analysis of brain tumor using intelligent techniques. In 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT) (pp. 48-52). IEEE.

Lapointe, S., Perry, A., & Butowski, N. A. (2018). Primary brain tumours in adults. The Lancet, 392(10145), 432-446.

Liu, Z., Tong, L., Chen, L., Zhou, F., Jiang, Z., Zhang, Q., ... & Zhou, H. (2021a). Canet: Context aware network for brain glioma segmentation. IEEE Transactions on Medical Imaging, 40(7), 1763-1777.

Ghaffari, M., Samarasinghe, G., Jameson, M., Aly, F., Holloway, L., Chlap, P., ... & Oliver, R. (2022). Automated post-operative brain tumour segmentation: A deep learning model based on transfer learning from pre-operative images. Magnetic resonance imaging, 86, 28-36.

Moser, E., Stadlbauer, A., Windischberger, C., Quick, H. H., & Ladd, M. E. (2009). Magnetic resonance imaging methodology. European journal of nuclear medicine and molecular imaging, 36, 30-41.

Zhou, T., Canu, S., Vera, P., & Ruan, S. (2021). Feature-enhanced generation and multi-modality fusion based deep neural network for brain tumor segmentation with missing MR modalities. Neurocomputing, 466, 102-112.

Wang, P., & Chung, A. C. (2022). Relax and focus on brain tumor segmentation. Medical Image Analysis, 75, 102259.

Zeineldin, R. A., Karar, M. E., Mathis-Ullrich, F., & Burgert, O. (2022). Ensemble CNN Networks for GBM Tumors Segmentation Using Multi-parametric MRI. In International MICCAI Brainlesion Workshop (pp. 473-483). Springer, Cham.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical image analysis, 42, 60-88.

Jia, Q., & Shu, H. (2022a). Bitr-unet: a cnn-transformer combined network for mri brain tumor segmentation. In International MICCAI Brainlesion Workshop (pp. 3-14).

Springer, Cham.

Myronenko, A. (2018, September). 3D MRI brain tumor segmentation using autoencoder regularization. In International MICCAI Brainlesion Workshop (pp. 311-320). Springer, Cham.

Lyu, C., & Shu, H. (2020, October). A two-stage cascade model with variational autoencoders and attention gates for MRI brain tumor segmentation. In International MICCAI Brainlesion Workshop (pp. 435-447). Springer, Cham.

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 18(2), 203-211.

Isensee, F., Jäger, P. F., Full, P. M., Vollmuth, P., & Maier-Hein, K. H. (2020, October). nnU-Net for brain tumor segmentation. In International MICCAI Brainlesion Workshop (pp. 118-132). Springer, Cham.

Luu, H. M., & Park, S. H. (2022). Extending nn-UNet for brain tumor segmentation. In International MICCAI Brainlesion Workshop (pp. 173-186). Springer, Cham.

Zhang, Y., Liu, H., & Hu, Q. (2021a). Transfuse: Fusing transformers and cnns for medical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 14-24). Springer, Cham.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

Chen, G., Qu, S., Li, Z., Zhu, H., Dong, J., Liu, M., & Conradt, J. (2022). Neuromorphic vision-based fall localization in event streams with temporal-spatial attention weighted network. IEEE transactions on cybernetics.

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7794-7803).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., & Li, J. (2021, September). Transbts: Multimodal brain tumor segmentation using transformer. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 109-119). Springer, Cham.

Li, J., Wang, W., Chen, C., Zhang, T., Zha, S., Yu, H., & Wang, J. (2022). TransBTSV2: Wider Instead of Deeper Transformer for Medical Image Segmentation. arXiv preprint arXiv:2201.12785.

Jia, H., Bai, C., Cai, W., Huang, H., & Xia, Y. (2022b). HNF-Netv2 for Brain Tumor Segmentation using multi-modal MR Imaging. In International MICCAI Brainlesion Workshop (pp. 106-115). Springer, Cham.

Wang, L., Lee, C. Y., Tu, Z., & Lazebnik, S. (2015). Training deeper convolutional networks with deep supervision. arXiv preprint arXiv:1505.02496.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., ... & Xu, D. (2022). Unetr: Transformers for 3d medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 574-584).

Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In 2016 fourth international conference on 3D vision (3DV) (pp. 565-571). IEEE.

Akil, M., Saouli, R., & Kachouri, R. (2020). Fully automatic brain tumor segmentation with deep learning-based selective attention using overlapping patches and multi-class weighted cross-entropy. Medical image analysis, 63, 101692.

Isik-Polat, E., Polat, G., Kocyigit, A., & Temizel, A. (2022). Evaluation and Analysis of Different Aggregation and Hyperparameter Selection Methods for Federated Brain Tumor Segmentation. arXiv preprint arXiv:2202.08261.

Yin, Y., Yang, H., Liu, Q., Jiang, M., Chen, C., Dou, Q., & Heng, P. A. (2022). Efficient Federated Tumor Segmentation via Normalized Tensor Aggregation and Client Pruning. In International MICCAI Brainlesion Workshop (pp. 433-443). Springer, Cham.

Nalawade, S., Ganesh, C., Wagner, B., Reddy, D., Das, Y., Yu, F. F., ... & Maldjian, J. A. (2022). Federated Learning for Brain Tumor Segmentation Using MRI and Transformers. In International MICCAI Brainlesion Workshop (pp. 444-454). Springer, Cham.

Momin, S., Lei, Y., Tian, Z., Roper, J., Lin, J., Kahn, S., ... & Yang, X. (2022). Cascaded mutual enhancing networks for brain tumor subregion segmentation in multiparametric MRI. Physics in Medicine & Biology, 67(8), 085015.

Zhang, D., Huang, G., Zhang, Q., Han, J., Han, J., & Yu, Y. (2021b). Cross-modality deep feature learning for brain tumor segmentation. Pattern Recognition, 110, 107562.

Pei, L., & Liu, Y. (2022). Multimodal Brain Tumor Segmentation Using a 3D ResUNet in BraTS 2021. In International MICCAI Brainlesion Workshop (pp. 315-323). Springer, Cham.

Jabareen, N., & Lukassen, S. (2022). Segmenting Brain Tumors in Multi-modal MRI Scans Using a 3D SegNet Architecture. In International MICCAI Brainlesion Workshop (pp. 377-388). Springer, Cham.

Bukhari, S. T., & Mohy-ud-Din, H. (2022). E1D3 U-Net for Brain Tumor Segmentation: Submission to the RSNA-ASNR-MICCAI BraTS 2021 challenge. In International MICCAI Brainlesion Workshop (pp. 276-288). Springer, Cham.

Liu, D., Sheng, N., He, T., Wang, W., Zhang, J., & Zhang, J. (2022). SGEResU-Net for brain tumor segmentation. Mathematical Biosciences and Engineering, 19(6), 5576-5590.

Huang, L., Denoeux, T., Vera, P., & Ruan, S. (2022). Evidence fusion with contextual discounting for multi-modality medical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 401-411). Springer, Cham.

Peiris, H., Hayat, M., Chen, Z., Egan, G., & Harandi, M. (2021). A volumetric transformer for accurate 3d tumor segmentation. arXiv preprint arXiv:2111.13300.

Rahman Siddiquee, M. M., & Myronenko, A. (2022). Redundancy Reduction in Semantic Segmentation of 3D Brain Tumor MRIs. In International MICCAI Brainlesion Workshop (pp. 163-172). Springer, Cham.

Khan, M. I., Jafaritadi, M., Alhoniemi, E., Kontio, E., & Khan, S. A. (2022). Adaptive Weight Aggregation in Federated Learning for Brain Tumor Segmentation. In International MICCAI Brainlesion Workshop (pp. 455-469). Springer, Cham.

Mächler, L., Ezhov, I., Kofler, F., Shit, S., Paetzold, J. C., Loehr, T., ... & Menze, B. H. (2022). FedCostWAvg: A new averaging for better Federated Learning. In International MICCAI Brainlesion Workshop (pp. 383-391). Springer, Cham.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., & Maier-Hein, K. H. (2018). No new-net. In International MICCAI Brainlesion Workshop (pp. 234-244). Springer, Cham.

Ullah, Z., Usman, M., Jeon, M., & Gwak, J. (2022). Cascade multiscale residual attention cnns with adaptive roi for automatic brain tumor segmentation. Information Sciences, 608, 1541-1556.

Kotowski, K., Adamski, S., Machura, B., Zarudzki, L., & Nalepa, J. (2022). Coupling nnU-Nets with Expert Knowledge for Accurate Brain Tumor Segmentation from MRI. In International MICCAI Brainlesion Workshop (pp. 197-209). Springer, Cham.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021b). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).