frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 30 November 2022
DOI 10.3389/fnins.2022.1054948

OPEN ACCESS

EDITED BY
Shu Zhang,
Northwestern Polytechnical University, China

REVIEWED BY
Yuan Xue,
Johns Hopkins University, United States
Shiqiang Ma,
Tianjin University, China
Lu Zhang,
University of Texas at Arlington, United States

*CORRESPONDENCE
Senchun Chai
chaisc97@163.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 27 September 2022
ACCEPTED 07 November 2022
PUBLISHED 30 November 2022

CITATION
Huang L, Zhu E, Chen L, Wang Z, Chai S and Zhang B (2022) A transformer-based generative adversarial network for brain tumor segmentation.
Front. Neurosci. 16:1054948.
doi: 10.3389/fnins.2022.1054948

A transformer-based generative adversarial network for brain tumor segmentation

Liqun Huang[1†], Enjun Zhu[2†], Long Chen[1], Zhaoyang Wang[1], Senchun Chai[1*] and Baihai Zhang[1]

[1]The School of Automation, Beijing Institute of Technology, Beijing, China, [2]Department of Cardiac Surgery, Beijing Anzhen Hospital, Capital Medical University, Beijing, China

Brain tumor segmentation remains a challenge in medical image segmentation tasks. With the application of transformer in various computer vision tasks, transformer blocks show the capability of learning long-distance dependency in global space, which is complementary to CNNs. In this paper, we proposed a novel transformer-based generative adversarial network to automatically segment brain tumors with multi-modalities MRI. Our architecture consists of a generator and a discriminator, which is trained in min–max game progress. The generator is based on a typical "U-shaped" encoder–decoder architecture, whose bottom layer is composed of transformer blocks with Resnet. Besides, the generator is trained with deep supervision technology. The discriminator we designed is a CNN-based network with multi-scale L_1 loss, which is proved to be effective for medical semantic image segmentation. To validate the effectiveness of our method, we conducted exclusive experiments on BRATS2015 dataset, achieving comparable or better performance than previous state-of-the-art methods. On additional datasets, including BRATS2018 and BRATS2020, experimental results prove that our technique is capable of generalizing successfully.

KEYWORDS

generative adversarial network, transformer, deep learning, automatic segmentation, brain tumor

1. Introduction

Semantic medical image segmentation is an indispensable step in computer-aided diagnosis (Stoitsis et al., 2006; Le, 2017; Razmjooy et al., 2020; Khan et al., 2021). In clinical practice, tumor delineation is usually performed manually or semi-manually, which is time-consuming and labor-intensive. As a result, it is of vital importance to explore automatic volumetric segmentation methods with the help of medical images to accelerate the computer-aided diagnosis. In this paper, we focus on the segmentation of brain tumors with the help of magnetic resonance imaging (MRI) consisting of multi-modality scans. The automatic segmentation of gliomas remains one of the most challenging medical segmentation problems stemming from some aspects, such as arbitrary shape and location, poorly contrasted, and blurred boundary with surrounding issues.

Since the advent of deep learning, Convolutional Neural Networks (CNN) have achieved great success in various computer vision tasks, ranging from classification (LeCun et al., 1998; Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Szegedy et al., 2015; Huang et al., 2017), object detection (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015; Liu et al., 2016; Redmon et al., 2016; He et al., 2017; Redmon and Farhadi, 2017, 2018; Bochkovskiy et al., 2020) to segmentation (Chen et al., 2014, 2017; Long et al., 2015; Ronneberger et al., 2015; Lin et al., 2017). Fully Convolution Networks (FCN Long et al., 2015) and in particular "U-shaped" encoder–decoder architectures have realized state-of-the-art results in medical semantic segmentation tasks. U-Net (Ronneberger et al., 2015), which consists of symmetric encoder and decoder, uses the skip connections to merge the extracted features from encoder with those from decoder at different resolutions, aiming at recovering the lost details during downsampling. Owing to the impressive results in plenty of medical applications, U-Net and its variants have become the mainstream architectures in medical semantic segmentation.

In spite of their prevalence, FCN-based approaches are incapable of modeling long-range dependency because of its intrinsic limited receptive field and the locality of convolution operations. Inspired by the great success of transformer-based models in Natural Language Processing (NLP) (Devlin et al., 2018; Radford et al., 2018; Liu et al., 2019; Yang et al., 2019; Clark et al., 2020), a growing number of researchers propose to apply the self-attention mechanism to medical image segmentation, attempting to overcome the limitations brought by the inductive bias of convolution, so as to extract the long-range dependency and context–dependent features. Especially, unlike prior convolution operations, transformers encode a sequence of patches and leverage the power of self-attention modules to pre-train on a large-scale dataset for downstream tasks, like Vision Transformer (ViT) (Dosovitskiy et al., 2020) and its variants.

Simultaneously, for the Transformers applied in medical image segmentation, Generative Adversarial Networks (GAN) has revealed great performance in semantic segmentation. In a typical GAN architecture used for segmentation, GAN consists of two competing networks, a discriminator and a generator. The generator learns the capability of contexture representations, minimizing the distance between prediction and masks, while the discriminator on the contrary maximizes the distance to distinguish the difference between them. The two networks are trained in an alternating fashion to improve the performance of the other. Furthermore, some GAN-based methods like SegAN (Xue et al., 2018) achieve more effective segmentation performance than FCN-based approaches.

In this paper, we explore the integrated performance of transformer and generative adversarial network in segmentation tasks and propose a novel transformer-based generative adversarial network for brain tumor segmentation. Owing to the attention mechanism, transformer has a global receptive field from the very first layer to the last layer, instead of focusing solely on the local information from convolution kernel in each layer, thus contributing to the pixel-level classification and being more suitable for medical segmentation tasks. Besides, CNN learns representative features at different resolutions through cascading relationships, while the attention mechanism pays more attention to the relationship between features, thus transformer-based methods are easily-generalized and not completely dependent on the data itself, such as experiments with incomplete images input in Naseer et al. (2021). Inspired by some attempts (Wang W. et al., 2021; Hatamizadeh et al., 2022) of fusing transformer with 3D CNNs, we design an encoder–decoder generator with deep supervision, where both encoder and decoder are 3D CNNs but the bridge of them is composed of transformer blocks with Resnet. In the contrast of typical "U-shaped" decoder–encoder network, our transformer block is designed to replace the traditional convolution-based bottleneck, for the reason that the self-attention mechanism inside transformer can learn long-range contextual representations while the finite kernel size limits the CNN's capability of learning global information. For pixel-wise brain tumor segmentation task, replacing CNN with transformer blocks on the bottleneck contributes to capturing more features from encoder. Inspired by SegAN (Xue et al., 2018), we adopt the multi-scale L_1 loss to our method with only one generator and one discriminator, measuring the distance of the hierarchical features between generated segmentation and ground truth. Experimental results on BRATS2015 dataset show that our method achieves comparable or better performance than some previous state-of-the-art methods. Compared to existing methods, the main contributions of our approach are listed as follows:

- A novel transformer-based generative adversarial network is proposed to address the brain tumor segmentation task with multi-modalities MRI. To enhance the efficiency of brain tumor segmentation, our method incorporates the concepts of "Transformer" and "Generative adversarial". The generator makes use of the transformer blocks to facilitate the process of learning global contextual representations. As far as we are aware, our work is among the very first ones to explore the combination of transformer and generative adversarial networks and achieve excellent performance in the brain tumor segmentation task.
- Our generator exploits transformer with Resnet module in 3D CNN for segmenting multi-modalities MRI brain tumors. Building upon the encoder–decoder structure, both encoder and decoder in our proposed generator are mainly composed of traditional 3D convolution layers, while the bottom layer of the "U-shaped" structure is a transformer with Resnet module. With Resnet, the

transformer block captures both global and local spatial dependencies effectively, thus preparing embedded features for progressive upsampling to full resolution predicted maps.

- Our loss functions are suitable and effectively applied in generator and discriminator. Adopting the idea of deep supervision (Zhu Q. et al., 2017), we take the output of the last three decoder layers of generator to calculate weighted loss for better gradient propagation. Besides, we leverage a CNN-based discriminator to compute multi-scale L_1 norm distance of hierarchical features extracted from ground truth and segmentation maps, respectively.
- The exclusive experimental results evaluated on BRATS2015 dataset show the effectiveness of each part of our proposed methods, including transformer with Resnet module and loss functions. Comparing to existing methods, the proposed method can obtain significant improvements in brain tumor segmentation. Moreover, our method successfully generalizes in other brain tumor segmentation datasets: BRATS2018 and BRATS2020.

The following outlines the structure of this paper: Section 2 reviews the related work. Section 3 presents the detail of our proposed architecture. Section 4 describes the experimental setup and evaluates the performance of our method. Section 5 summarizes this work.

2. Related works

2.1. Vision transformer

The Transformers were first proposed by Vaswani et al. (2017) on machine translation tasks and achieved a quantity of state-of-the-art results in NLP tasks (Devlin et al., 2018; Radford et al., 2018). Dosovitskiy et al. (2020) then applied Transformers to image classification tasks by directly training a pure Transformer on sequences of image patches as words in NLP, and achieved state-of-the-art benchmarks on the ImageNet dataset. In object detection, Carion et al. (2020) proposed transformer-based DETR, a transformer encoder–decoder architecture, which demonstrated accuracy and run-time performance on par with the highly optimized Faster R-CNN (Ren et al., 2015) on COCO dataset.

Recently, various approaches were proposed to explore the applications of the transformer-based model for semantic segmentation tasks. Chen et al. (2021) proposed TransUNet, which added transformer layers to the encoder to achieve competitive performance for 2D multi-organ medical image segmentation. As for 3D medical image segmentation, Wang W. et al. (2021) exploited Transformer in 3D CNN for segmenting MRI brain tumors and proposed to use a transformer in the bottleneck of "U-shaped" network on BRATS2019 and BRATS2020 datasets. Similarly, Hatamizadeh et al. (2022) proposed an encoder–decoder network named UNETR, which employed transformer modules as the encoder and CNN modules as the decoder, for the brain tumor and spleen volumetric medical image segmentation.

Compared to these approaches above, our method is tailored for 3D segmentation and is based on generative adversarial network. Our generator produces sequences fed into transformer by utilizing a backbone encoder–decoder CNN, where the transformer with Resnet module is placed in the bottleneck. With Resnet, the encoder captures features not only from CNN-based encoder but also from transformer blocks. Moreover, the last three output layers of the encoder are considered to calculate the loss function for better performance. Networks like UNETR employ transformer layers as encoder in low-dimension semantic level, and taking this network as backbone in our method without pre-training easily leads to model collapse during the adversarial training phase. Therefore, we do not choose these networks as our backbone. We find that taking transformer as encoder in low-dimension semantic level needs quantities of pre-training tasks on other datasets to get good results, like TransUNet and UNETR above. As shown in our experiments Section 4.6, transformer-based encoder in low-dimension semantic level performances inferior to CNN-based one when training from scratch. Therefore, we choose to apply transformer only in bottleneck, and remain the low-dimension encode layers as convolutional layers. In this way, we can train from scratch, meanwhile achieving good performance.

2.2. Generative adversarial networks

The GAN (Goodfellow et al., 2014) is originally introduced for image generation (Mirza and Osindero, 2014; Chen et al., 2016; Odena et al., 2017; Zhu J.-Y. et al., 2017), making the core idea of competing training with a generator and a discriminator, respectively, known outside of fixed circle. However, there exists a problem that it is troublesome for the original GAN to remain in a stable state, hence making us cautious to balance the training level of the generator and the discriminator in practice. Arjovsky et al. proposed Wasserstein GAN (WGAN) as a thorough solution to the instability by replacing the Kullback-Leibler (KL) divergence with the Earth Mover (EM) distance.

Various methods (Isola et al., 2017; Han et al., 2018; Xue et al., 2018; Choi et al., 2019; Dong et al., 2019; Oh et al., 2020; Ding et al., 2021; He et al., 2021; Nishio et al., 2021; Wang T. et al., 2021; Zhan et al., 2021; Asis-Cruz et al., 2022) were proposed to explore the possibility of GAN in medical image segmentation. Xue et al. (2018) used U-Net as the generator and proposed a multi-scale L_1 loss to minimize the distance of the feature maps of predictions and masks for the medical image segmentation of brain tumors. Oh et al. (2020) took residual blocks into account under the framework of pix2pix

(Isola et al., 2017) and segmented the white matter in FDG-PET images. Ding et al. (2021) took an encoder–decoder network as the generator and designed a discriminator based on Condition GAN (CGAN) on BRATS2015 dataset, adopting the image labels as the extra input.

Unlike these approaches, our method incorporates the concepts of "Transformer" and "GAN." Our discriminator is based on CNN instead of transformer. In our opinion, owing to the attention mechanism inside transformer, transformer has a more global receptive field than CNN with limited kernel size, thus contributing to pixel-level classification and being more suitable for medical segmentation tasks. However, for image-level medical classification, transformer-based discriminator seems to be less appropriate for its weakness of requiring huge datasets to support pre-training, while CNN is strong enough for classification tasks without pre-training. Motivated by viewpoints above, in our method, the transformer-based generator and CNN-based discriminator are combined to facilitate the progress of segmentation under the supervision of a multi-scale L_1 loss.

3. Materials and methods

3.1. Overall architecture

The overview of our proposed model is presented in Figure 1. Our framework consists of a generator and discriminator for competing training. The generator G is a transformer-based encoder–decoder architecture. Given a multi modalities (T1, T1c, T2, and FLAIR) MRI scan $X \in R^{C\times H\times W\times D}$ with 3D resolution (H, W, D) and C channels, we utilize 3D CNN-based down-sampling encoder to produce high dimension semantic feature maps, and then these semantic information flow to 3D CNN-based up-sampling decoder through the intermediate Transformer block with Resnet (He et al., 2016). With skip connection, the long-range and short-range spatial relations extracted by encoder from each stage flow to the decoder. For deep supervision (Zhu Q. et al., 2017), the output of decoder consists of three parts: the output of last three convolution layers after sigmoid. Inspired by Xue et al. (2018), the discriminator D we used has the similar structure as encoder in G, extracting hierarchical feature maps from ground truth (GT) and prediction separately to compute multi-scale L_1 loss.

3.2. Generator

Encoder is the contracting path which has seven spatial levels. Patches of size 160 × 192 × 160 with four channels are randomly cropped from brain tumor images as input, followed by six down-sampling layers with 3D 3×3×3 convolution (stride = 2). Each convolution operation is followed by an Instance Normalization (IN) layer and a LeakyReLU activation layer.

At the bottom of the encoder, we leverage the transformer with Resnet module to model the long-distance dependency in a global space. The feature maps produced by the encoder is sequenced first and then create the feature embeddings by simply fusing the learnable position embeddings with sequenced feature map by element-wise addition. After the position embeddings, we introduce L transformer layers to extract the long-range dependency and context dependent features. Each transformer layer consists of a Multi-Head Attention (MHA) block after layer normalization (LN) and a feed forward network (FFN) after layer normalization. In attention block, the input sequence is fed into three convolution layers to produce three metrics: queries Q, keys K and values V. To combine the advantages of both CNN and Transformer, we simply short cut the input and output of Transformer block. Thus, as in Vaswani et al. (2017) and Wang W. et al. (2021), given the input X, the output of the transformer with Resnet module Y can be calculated by:

$$Y = x + y_L \tag{1}$$

$$y_i = FFN\left(LN\left(y_i^{'}\right)\right) + y_i^{'} \tag{2}$$

$$y_i^{'} = MHA\left(LN\left(y_{i-1}\right)\right) + y_{i-1} \tag{3}$$

$$MHA\left(Q, K, V\right) = Concat\left(head_1, ..., head_h\right) W^O \tag{4}$$

$$head_i = Attention\left(Q, K, V\right) = softmax\left(QK^T/\sqrt{d_k}\right) V \tag{5}$$

where y_i denotes the output of ith ($i \in [1, 2, ..., L]$) Transformer layer, y_0 denotes X, W^O are projection metrics, d_k denotes the dimension of K.

Unlike the encoder, the decoder uses 3D 2 × 2 × 2 transpose convolution for up-sampling, followed by skip connection and two 3D 3 × 3 × 3 convolution layers. For a better gradient flow and a better supervision performance, a technology called deep supervision is introduced to utilize the last three decoder levels to calculate loss function. Concretely, we downsampled the GT to the same resolution with these outputs, thus making weighted sum of loss functions in different levels.

The detailed structure of our transformer-based generator is presented in Table 1. In the encoder part, patches of size 160 × 192 × 160 voxels with four channels are randomly cropped from the original brain tumor images as input. At each level, there are two successive 3 × 3 × 3 unbiased convolution layers followed by normalization, activation layers and dropout layers. Beginning from the second level, the resolution of the feature maps is reduced by a factor of 2. These features, e.g., areas of white matter, edges of brain, dots and lines, etc., are extracted by sufficient convolution kernels for next blocks. The

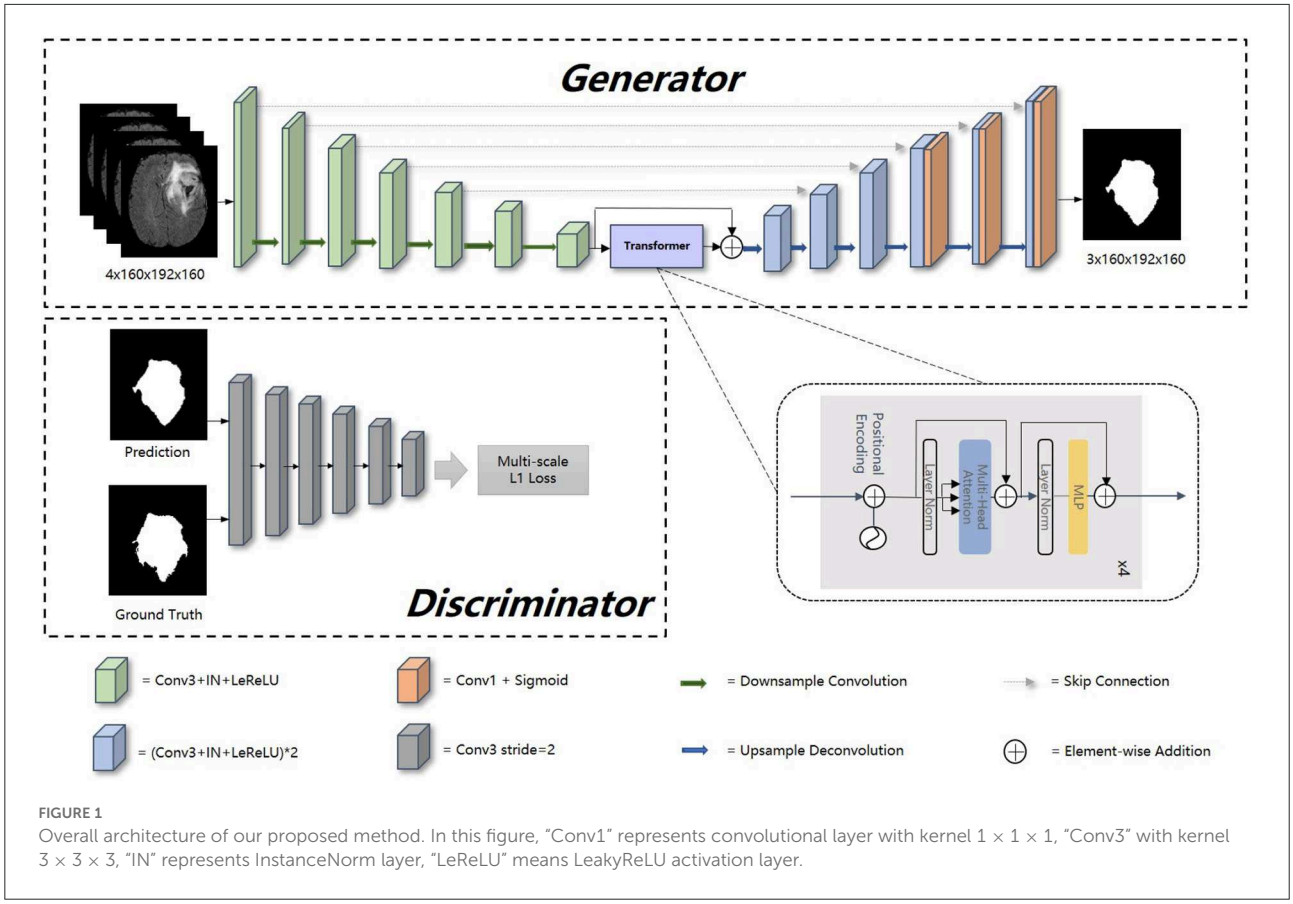

FIGURE 1
Overall architecture of our proposed method. In this figure, "Conv1" represents convolutional layer with kernel $1 \times 1 \times 1$, "Conv3" with kernel $3 \times 3 \times 3$, "IN" represents InstanceNorm layer, "LeReLU" means LeakyReLU activation layer.

transformer block enriches the global contextual representation based on the attention mechanism, forcing features located in the desired regions unchanged while suppressing those in other regions. The shortcut branch crossing the transformer block fusing the features from both encoder part and transformer block by element-wise addition, indicating that our generator is capable of learning short-range and long-range spatial relations with neither extra parameter nor computation complexity. According to the attributes of Resnet (He et al., 2016), $y = f(x) + x$, where $f(x)$ in our method represents transformer blocks, x is the output of CNN-based encoder, whose contexture representations in feature maps are relatively short-range than transformer's. With Resnet, the element-wise addition of $f(x)$ and x can directly fuse the short-range spatial relations from CNN-based encoder and long-range spatial relations from transformer-based bottleneck. Additionally, unlike neural network layers, element-wise addition is a math operation with no more memory cost and negligible computation time cost. The decoder part contains amounts of upsampling layers and skip connection to progressively recover semantic information as well as resolution. The first upsampling layer is implemented by interpolation while the other upsampling layers adapt the form of deconvolution with stride set to 2. At level $i \in [1, 5]$, the encoder block D_i doubles the spatial resolution, followed by skip connection to fuse high-level (from D_i) and low-level (from encoder block E_i) contextual representation so as to segment the desired tumor regions. For a better supervision performance, the outputs of D_i where $i \in [1, 3]$ are fed into $1 \times 1 \times 1$ convolution layer and sigmoid layer to predict segmentation maps with different resolution. Accordingly, the ground truth is downsampled to different shapes such that they match the shapes of those segmentation maps.

Our generator's vital part is the transformer with Resnet module. As shown in Table 1, our transformer with Resnet module consists of transformer block and Resnet, while transformer block is composed of position encodings module, several transformer layers depicted in Figure 2 and features projection module. To make use of the order of the input sequence reshaped from bottom layer feature maps, we introduce a learnable positional encoding vector to represent some information about position of tokens in the sequence, instead of sine and cosine functions. After position encoding and normalization, the input sequence is fed into three different linear layers to create queries, keys, and values. Then, we compute the dot products of keys with queries. To avoid extremely small gradients after softmax function, we scale the dot-products by a factor related to dimensions of queries, as shown in Equation 5. Multiplying scaled weights with

TABLE 1 The detailed structure of proposed generator.

Stage	Name	Details	Output size
Encoder	E1	[Conv3, IN, LeReLU, Dropout]	64*160*192*160
		[Conv3, IN, LeReLU, Dropout]	
	E2	[Conv3(stride2), IN, LeReLU, Dropout]	96*80*96*80
		[Conv3, IN, LeReLU, Dropout]	
	E3	[Conv3(stride2), IN, LeReLU, Dropout]	128*40*48*40
		[Conv3, IN, LeReLU, Dropout]	
	E4	[Conv3(stride2), IN, LeReLU, Dropout]	192*20*24*20
		[Conv3, IN, LeReLU, Dropout]	
	E5	[Conv3(stride2), IN, LeReLU, Dropout]	256*10*12*10
		[Conv3, IN, LeReLU, Dropout]	
	E6	[Conv3(stride2), IN, LeReLU, Dropout]	384*5*6*5
		[Conv3, IN, LeReLU, Dropout]	
	E7	[Conv3(stride2), IN, LeReLU, Dropout]	512*3*3*3
		[Conv3, IN, LeReLU, Dropout]	
Transformer	ResTransBlock	Reshape	512*3*3*3
		PE	
		Transformer Layer*4	
		Reshape	
		Resnet	
Decoder	D6	Upsample	384*5*6*5
		[Conv3, IN, LeReLU, Dropout] x 2	
	D5	Deconv	256*10*12*10
		Concat	
		[Conv3, IN, LeReLU, Dropout]	
		[Conv3, IN, LeReLU, Dropout]	
	D4	Deconv	192*20*24*20
		Concat	
		[Conv3, IN, LeReLU, Dropout]	
		[Conv3, IN, LeReLU, Dropout]	
		Deconv	128*40*48*40
	D3	Concat	
		[Conv3, IN, LeReLU, Dropout]	
		[Conv3, IN, LeReLU, Dropout]	
	Output3	Conv1 + Sigmoid	4*40*48*40
		Deconv	96*80*96*80
	D2	Concat	
		[Conv3, IN, LeReLU, Dropout]	
		[Conv3, IN, LeReLU, Dropout]	
	Output2	Conv1 + Sigmoid	4*80*96*80
		Deconv	64*160*192*160
	D1	Concat	
		[Conv3, IN, LeReLU, Dropout]	
		[Conv3, IN, LeReLU, Dropout]	
	Output1	Conv1 + Sigmoid	3*160*192*160

values, we obtain a single attention output, which is then concatenated with other heads' outputs to produce the multi-head attention outputs. Subsequently, normalization, dropout, and multi-layer perception (MLP) layers are utilized to produce the transformer layer's ultimate output. While convolution layers have local connections, shared weights, and translation equivariance, attention layers are global. We take advantage of both by residual connection to learn both short-range and long-range spatial relations with no more memory cost and negligible computational time cost.

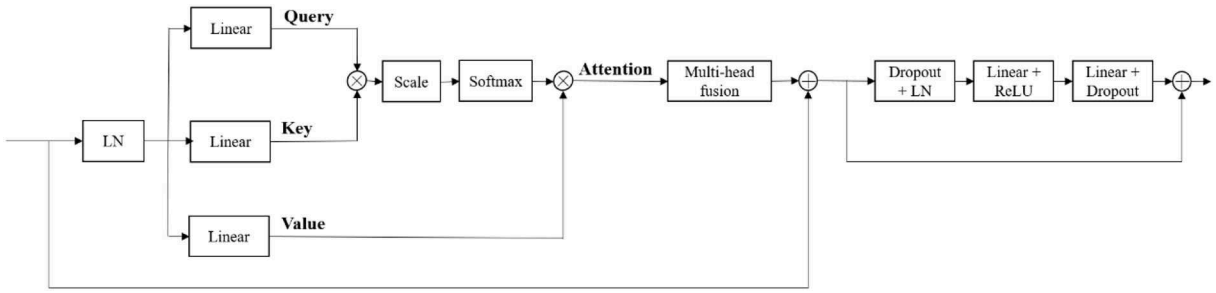

FIGURE 2
The structure of transformer layer.

3.3. Discriminator and loss function

To distinguish the difference between the prediction and GT, the discriminator D extracts features of GT and prediction to calculate L_1 norm distance between them. The discriminator is composed of six similar blocks. Each of these blocks consists of a 3 × 3 × 3 convolution layer with a stride of 2, a batch normalization layer and a LeakyReLU activation layer. Instead of only using the final output of D, we leverage the jth output feature $f_j^i(x)$ extracted by ith ($i \in [1, 2, \ldots, L]$) layers from image x to calculate multi-scale L_1 loss ℓ_D as follows:

$$\ell_D\left(x, x^{'}\right) = \frac{1}{L * M}\sum_{i=1}^{L}\sum_{j=1}^{M}\left\|f_j^i(x) - f_j^i\left(x^{'}\right)\right\|_1 \tag{6}$$

where M denotes the number of extracted features of a layer in D.

Referring to the loss function of GAN (Goodfellow et al., 2014), our loss function of the whole adversarial process is described as follows:

$$\begin{aligned}\min_{\theta_G}\max_{\theta_D}\mathcal{L}\left(\theta_G, \theta_D\right) = \mathbb{E}_{x\sim P_{data}}\left(\ell_D\left(G(x), y\right)\right) \\ +\mathbb{E}_{x\sim P_{data}}\left(\ell_{deep\_bce\_dice}\left(G(x), y\right)\right)\end{aligned} \tag{7}$$

where x, y denote the input image and ground truth, respectively, $\ell_{deep\_bce\_dice}$ denotes that the segmentation maps of generator are used to calculate the BCE loss together with the Dice loss under deep supervision. Concretely, $\ell_{deep\_bce\_dice}$ is a weighted sum of $\ell_{deep\_bce\_dice}\left(p_i, y_i\right), i \in [1, 2, 3]$ for prediction p_i and mask y_i where i denotes the ith level of decoder (D_i).

The detailed training process is presented in Algorithm 1, which interprets the procedure of sampling data and following updating discriminator and generator with corresponding loss function respectively.

```
for number of training epoches do
  for steps of training discriminator do
    Get n input images from p_data {x^1, ..., x^n} and
    corresponding labels {y^1, ..., y^n}.
    Update discriminator by maximizing the loss
  below:
```

$$\frac{1}{n}\sum_{i=1}^{n}\left[\ell_D\left(G\left(x^i\right), y^i\right)\right]$$

```
    Clip the weights of discriminator.
  end for
  Get n input images from p_data {x^1, ..., x^n} and
  corresponding labels {y^1, ..., y^n}.
  Update generator by minimizing the loss
  below:
```

$$\frac{1}{n}\sum_{i=1}^{n}\left[\ell_{deep\_bce\_dice}\left(G\left(x^i\right), y^i\right) + \ell_D\left(G\left(x^i\right), y^i\right)\right]$$

```
end for
```

Algorithm 1. The detailed training process. $\ell_{deep\_bce\_dice}$ represents BCE Dice loss with deep supervision, ℓ_D represents multi-scale L_1 loss.

4. Experimental results

4.1. Dataset

In the experiments, we evaluated our method using the Brain Tumor Image Segmentation Challenge 2015 (BRATS2015) dataset. In BRATS2015, the training dataset contains manual annotation by clinical experts for 220 patient cases with high-grade glioma (HGG) and 55 patient cases with low-grade glioma (LGG), whereas 110 patient cases are supplied in the online testing dataset without annotation. Four 3D MRI modalities—T1, T1c, T2, and FLAIR—are used for all patient cases, as depicted in Figure 3. Each modality has the origin size 240 × 240 × 155 with the same voxel spacing. The ground truth has five classes: background (label 0), necrosis (label 1), edema (label 2), non-enhancing tumor (label 3), and enhancing tumor (label 4).

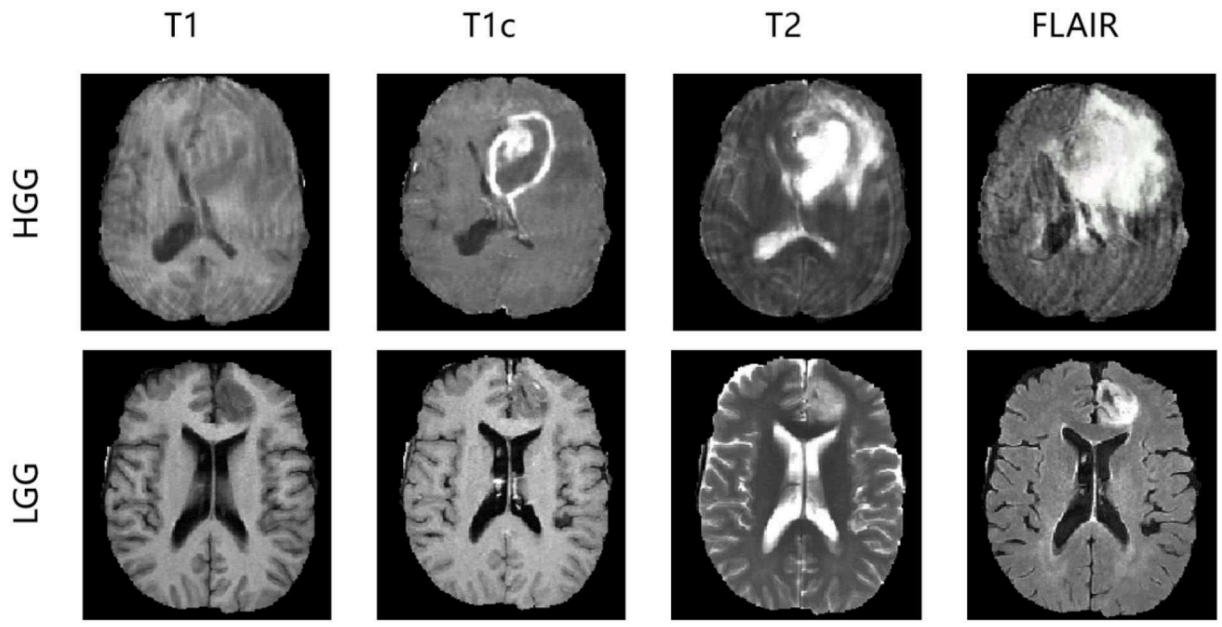

FIGURE 3
HGG and LGG cases with four modalities: T1, T1c, T2, FLAIR on BRATS2015 dataset.

We divided the 275 training cases into a training set and a validation set with the ratio 9:1 both in HGG and LGG. During training and validation, we padded the origin size $240 \times 240 \times 155$ to size $240 \times 240 \times 160$ with zeros and then randomly cropped into size $160 \times 192 \times 160$, which makes sure that the most image content is included.

4.2. Evaluation metric

To evaluate the effectiveness of a segmentation method, the most basic thing is to compare it with the ground truth. In the task of brain tumor segmentation, there are three main evaluation metrics compared with the ground truth: Dice, Positive predictive Value (PPV), and Sensitivity, defined as follows:

$$Dice\,(P,\ T) = \frac{1}{2} \times \frac{\left|P_1 \bigcap T_1\right|}{(|P_1| + |T_1|)} \tag{8}$$

$$PPV\,(P, T) = \frac{\left|P_1 \bigcap T_1\right|}{|P_1|} \tag{9}$$

$$Sensitivity\,(P, T) = \frac{\left|P_0 \bigcap T_0\right|}{|T_0|} \tag{10}$$

where P represents the prediction segmented by our proposed methods, T represents the corresponding ground truth. P_1 and T_1 denote the brain tumor region in P and T, P_0 and T_0 denote the other region except brain tumor in P and T, respectively, $|\cdot|$ calculates the number of voxels inside region, $\cap$ calculates the intersection of two regions. When Dice is larger, PPV and Sensitivity are larger at the same time, the predicted segmentation is considered to be more similar to ground truth, proving that the segmentation method is more effective.

4.3. Implementation details

Experiments were run on NVIDIA A100-PCIE (4×40GB) system for 1,000 epochs (about 3 days) using the Adam optimizer (Kingma and Ba, 2014). The target segmentation maps are reorganized into three tumor subregions: whole tumor (WT), tumor core (TC), and enhancing tumor (ET). The initial learning rate is 0.0001 and batch size is 4. The data augmentation consists of three parts: (1) padding the data from $240 \times 240 \times 155$ to $240 \times 240 \times 160$ with zeros; (2) randomizing the data's cropping from $240 \times 240 \times 160$ to $160 \times 192 \times 160$; (3) random flipping the data across three axes by a probability with 0.5. Impacted by the volumetric input size, the number of parameters of our network is larger than common 2D networks, generator: 58.0127M, transformer blocks inside generator: 11.3977M, discriminator: 75.4524M. Both the Dice loss in deep supervision and multi-scale L_1 loss are employed to train the network in competing progress. In inference, we converted the transformed three subregions (WT, TC, ET) back to the original labels. Specially, we replace the enhancing tumor with necrosis when the possibility

of enhancing tumor in segmentation map is less than the threshold, which is chosen according to the online testing scores.

4.4. Impact of the number of generators and discriminators

As the BRATS2015 is a multi-label segmentation task, our architecture can be implemented with schemes where the number of generators and discriminators are different. Each implementation scheme in Table 2 is specifically described as follows:

- 1G-1D. The network is composed of one generator and one discriminator. The generator outputs three-channel segmentation maps corresponding to three brain tumor subregions, while the discriminator is fed with three-class masked images concatenated in channel dimension.
- 1G-3D. The network is composed of one generator and three discriminators. The generator outputs three-channel segmentation maps while the discriminators output three one-channel maps, each for one class.
- 3G-3D. The network is composed of three generators and three discriminators. Each generator or discriminator is built for one class. There are three pairs of generators and discriminators, indicating that each pair is trained independently for one class.

4.5. Evaluating the transformer with Resnet module

To evaluate the effectiveness of the transformer with Resnet module, we conduct some ablation experiments. We design the bottom layer of our proposed generator with different schemes as follows:

- Transformer with Resnet. The bottom layer is composed of Transformer with Resnet we proposed.
- Transformer w/o Resnet. The bottom layer is composed of Transformer block, ranging from projection, position embedding to transformer layers, without shortcut crossing them.
- CNN with Resnet. The bottom layer is composed of convolutional layers together with a shortcut crossing them.
- Shortcut. The bottom layer is simply a shortcut connection from the encoder part to the decoder part.

TABLE 2 Results of different number of generators and discriminators.

Method	Dice			Positive predictive value			Sensitivity		
	Whole	Core	Enha.	Whole	Core	Enha.	Whole	Core	Enha.
1G-3D	**0.85**	**0.73**	**0.63**	**0.83**	**0.79**	0.59	**0.90**	**0.73**	**0.73**
1G-1D	0.84	0.72	0.62	0.82	0.78	0.58	0.89	0.72	0.71
3G-3D	0.81	0.68	0.60	0.83	0.74	**0.62**	0.84	0.70	0.63

Whole, whole tumor; Core, tumor core; Enha., enhancing tumor.

TABLE 3 Results of different bottom layer in generator.

Method	Dice			Positive predictive value			Sensitivity		
	Whole	Core	Enha.	Whole	Core	Enha.	Whole	Core	Enha.
Transformer with Resnet	**0.85**	**0.73**	**0.63**	**0.83**	**0.79**	0.59	0.90	**0.73**	**0.73**
Transformer w/o Resnet	0.85	0.71	0.61	0.83	0.79	0.60	0.90	0.69	0.68
CNN with Resnet	0.83	0.68	0.58	0.80	0.78	0.58	**0.91**	0.66	0.62
Shortcut	0.82	0.67	0.60	0.82	0.77	**0.63**	0.87	0.67	0.63

Whole, whole tumor; Core, tumor core; Enha., enhancing tumor.

TABLE 4 Results of different discriminators training from scratch.

Method	Dice			Positive predictive value			Sensitivity		
	Whole	Core	Enha.	Whole	Core	Enha.	Whole	Core	Enha.
CNN-based	**0.85**	**0.73**	**0.63**	**0.83**	**0.79**	**0.59**	**0.90**	**0.73**	**0.73**
Transformer-based	0.79	0.66	0.58	0.79	0.77	0.55	0.86	0.64	0.66

Whole, whole tumor; Core, tumor core; Enha., enhancing tumor.

TABLE 5 Results of different loss function.

Method	Dice			Positive predictive value			Sensitivity		
	Whole	Core	Enha.	Whole	Core	Enha.	Whole	Core	Enha.
Our method	**0.85**	**0.73**	**0.63**	**0.83**	**0.79**	**0.59**	**0.90**	**0.73**	**0.73**
w/o deep supervision	0.85	0.72	0.61	0.83	0.78	0.57	0.90	0.73	0.71
Single-scale L_1 loss	0.84	0.72	0.61	0.82	0.78	0.58	0.89	0.72	0.71

Whole, whole tumor; Core, tumor core; Enha., enhancing tumor.

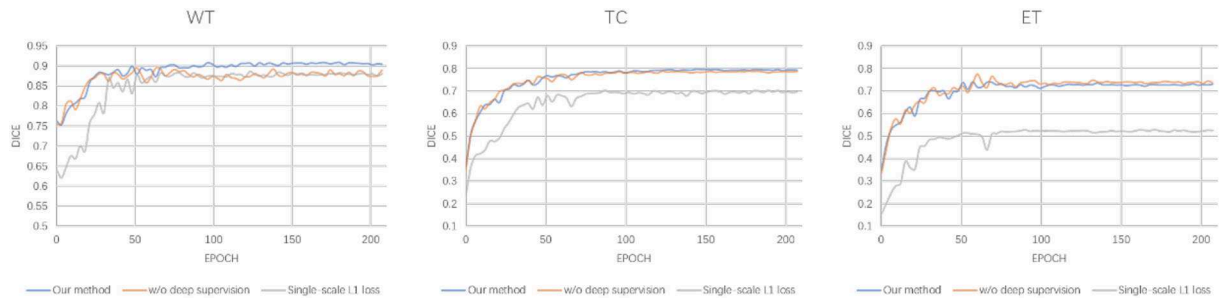

FIGURE 4
Detailed evaluation curves of different loss function.

TABLE 6 Performance of some methods on BRATS2015 testing dataset.

Method	Dice			Positive predictive value			Sensitivity		
	Whole	Core	Enha.	Whole	Core	Enha.	Whole	Core	Enha.
UNET (Ronneberger et al., 2015)	0.80	0.63	0.64	0.83	0.81	**0.78**	0.80	0.58	0.60
ToStaGAN (Ding et al., 2021)	**0.85**	0.71	0.62	0.87	**0.86**	0.63	0.87	0.68	0.69
3D Fusing (Zhao et al., 2018)	0.84	**0.73**	0.62	0.89	0.76	0.63	0.82	**0.76**	0.67
FSENet (Chen et al., 2018)	**0.85**	0.72	0.61	0.86	0.83	0.66	0.85	0.68	0.63
SegAN (Xue et al., 2018)	**0.85**	0.70	**0.66**	**0.92**	0.80	0.69	0.80	0.65	0.62
Our method	**0.85**	**0.73**	0.63	0.83	0.79	0.59	**0.90**	0.73	**0.73**

Whole, whole tumor; Core, tumor core; Enha., enhancing tumor.

The comparation results are shown in Table 3. From the results, we demonstrate the transformer's superiority and irreplaceability, and we can conclude that transformer with Resnet module make the best of features from transformer block and convolutional encoder to improve the segmentation performance.

4.6. Evaluating the CNN-based discriminator

We select the CNN-based discriminator instead of the transformer-based one as our final discriminator in our architecture, due to our opinion that transformer-based multi-layers discriminator requires huge datasets to support pre-training. To prove that, we conduct ablation experiments to compare their performance by training from scratch. The transformer-based discriminator is implemented using the inspiration of Jiang et al. (2021). Table 4 shows the results on BRATS2015 testing dataset using different discriminators, from which our CNN-based discriminator shows its superior capability of classifying the ground truth and segmentation outputs from scratch. Without pre-training, the CNN-based discriminator appears to be better than the transformer-based one.

4.7. Evaluating the loss function

In this section, we evaluate the effectiveness of the loss function in our proposed methods. As shown in Equation 7, our loss function is divided into two parts: the deep supervision

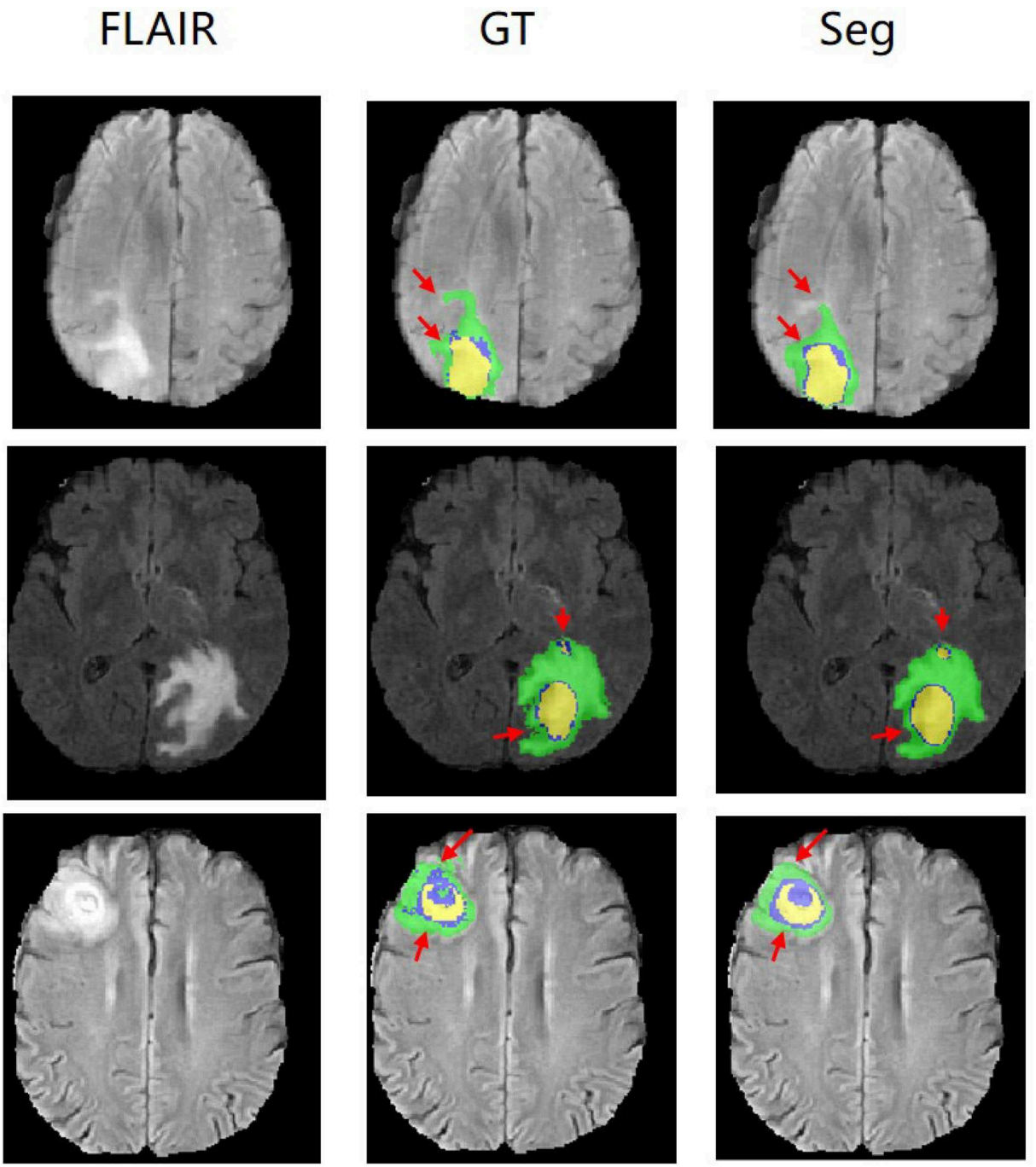

FIGURE 5
Experimental results with corresponding slices on BRATS2015 validation set. The red arrows locate the mainly different regions between GT and segmentation results.

loss and multi-scale L_1 loss. We conduct two ablation experiments: one model with single-scale L_1 loss, the other model without deep supervision loss. It is worth noting that the implementation of these models is the same as 1G-3D where the network consists of one generator and three discriminators and employs the transformer with Resnet module in the bottom layer. From Table 5, we find that our loss function achieves better performance under the same other experimental environment.

The detailed segmentation evaluation scores curves with different loss function are depicted in Figure 4. It is clear that the segmentation performance of all approaches steadily increases

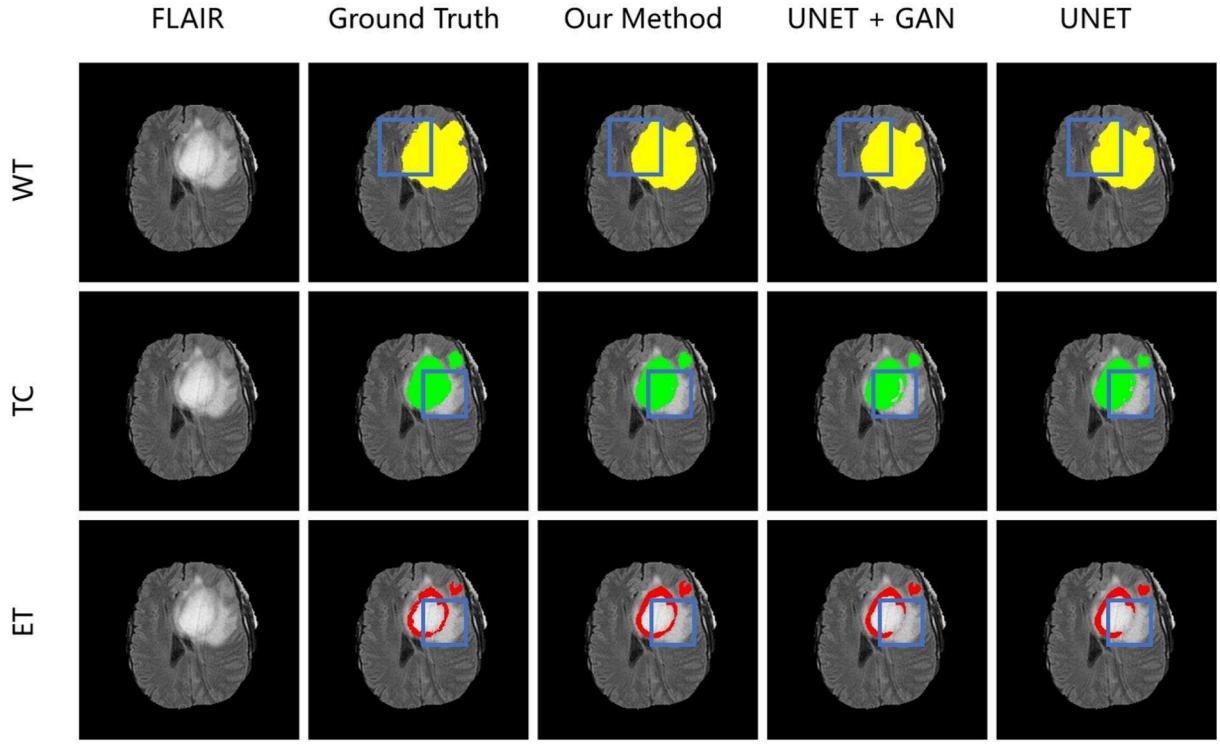

FIGURE 6
From left to right is the visualization results of an original image in FLAIR modality, ground truth, model in our method, model in the form of UNET + GAN, model UNET. From up to down in each column is three segmentation maps predicted with the same method. The blue boxes outline the difference between results from different methods.

as the number of epochs increases until it reaches a steady state. Ranging from WT, TC to ET, our method shows an increasing performance boost over other methods. As a consequence, our method yields the best results in all evaluation metrics listed in Table 5.

4.8. Comparison with other methods

To obtain a more robust prediction, we ensemble 10 models trained with the whole training dataset to average the segmentation probability maps. We upload the results of our methods on the BRATS2015 dataset and get the testing scores computed *via* the online evaluation platform, as listed in Table 6.

Figure 5 shows our qualitative segmentation output on BRATS2015 validation set. This figure illustrates different slices of different patient cases in ground truth and predictions separately.

4.9. Qualitative analysis

To demonstrate the performance of our proposed method, we randomly choose a slice of one patient on BRATS2015 validation set to visualize and compare the result in Figure 6. In Figure 6, images in the same column are produced from the same method, and images in the same row are belonging to the same segmentation label. Concretely, the column FLAIR represents the original image with modality of FLAIR, while other columns are segmentation maps with corresponding categories and colors: WT is yellow, TC is green, and ET is red. The column UNET represents that the corresponding three segmentation maps are inferenced with model UNET. The model of the column UNET plus GAN is built based on UNET, with an addition of GAN, where the generator is UNET with deep supervision and discriminator is a CNN-based network with multi-scale L_1 loss. A deep insight of Figure 6 reveals that with the help of deep supervision and multi-scale L_1 loss, the UNET+GAN method segments fuller edges and richer details than UNET method. When the transformer block is applied, our method produces more smooth borders on the tumor core regions, and more complete contours on enhancing tumor regions. The reason for this improvement seems to be that the transformer with Resnet module can effectively model the short-range and long-range dependency, and collect both local and global contexture representation information. Owing to more complete features, our method achieves the better performance.

TABLE 7 Comparison to other methods on BRATS2018 validation dataset.

Method	Dice(mean)			Hausdorff(mm)		
	Enha.	Whole	Core	Enha.	Whole	Core
Myronenko (2018)	0.7664	0.8836	**0.8154**	3.7731	5.9044	**4.8091**
Hu et al. (2019)	0.7178	0.8824	0.7481	**2.8000**	**4.4800**	7.0700
Chandra et al. (2018)	0.7406	0.8719	0.7990	5.5757	5.0379	9.5884
Liu (2018)	0.7639	0.8958	0.7905	4.0714	4.4924	8.1971
Our method	**0.7686**	**0.9021**	0.8089	5.7116	5.4183	9.4049

Whole, whole tumor; Core, tumor core; Enha., enhancing tumor.

TABLE 8 Comparison to other methods on BRATS2020 validation dataset.

Method	Dice(mean)			Hausdorff(mm)		
	Enha.	Whole	Core	Enha.	Whole	Core
Tang et al. (2020)	0.703	0.893	0.790	**34.306**	**4.629**	10.071
Zhou et al. (2022)	0.647	0.818	0.759	44.400	10.000	14.600
Anand et al. (2020)	**0.710**	0.880	0.740	38.310	6.880	32.000
Zhang et al. (2021)	0.700	0.880	0.740	38.600	7.000	30.200
Our method	0.708	**0.903**	**0.815**	37.579	4.909	**7.494**

Whole, whole tumor; Core, tumor core; Enha., enhancing tumor.

4.10. Generalization on other datasets

To evaluate generalization of our proposed method, we conduct additional experiments on other datasets relative to brain tumor segmentation, BRATS2018 and BRATS2020, which are composed of more practical patient cases. These datasets differ from BRATS2015 dataset in labels, number of cases and difficulty. The detailed inference performance are listed in Tables 7, 8. On BRATS2018 validation dataset, our proposed method achieves Dice score of 0.7686, 0.9021, and 0.8089, and Hausdorff (HD) of 5.7116, 5.4183, and 9.4049 mm on ET, WT, and TC, respectively. On BRATS2020 validation dataset, our method also realizes Dice score of 0.708, 0.903, and 0.815 and HD of 37.579, 4.909, and 7.494 mm on ET, WT, and TC, respectively. These excellent scores reveal the great generalization of our transformer-based generative adversarial network.

5. Discussion and conclusion

In this paper, we explored the application of a transformer-based generative adversarial network for segmenting 3D MRI brain tumors. Unlike many other encoder–decoder architectures, our generator employs a transformer with Resnet module to effectively model the long-distance dependency in a global space, simultaneously inheriting the advantage of CNNs for learning the capability of local contexture representations. Moreover, the application of deep supervision improves the flowability of gradient to some extent. Our discriminator is applied to measure the norm distance of hierarchical features from predictions and masks. Particularly, we calculate multi-scale L_1 loss between the generator segmentation maps and ground truth. Experimental results on BRATS2015, BRATS2018, and BRATS2020 datasets show a better performance of our proposed method in comparison of other state-of-the-art methods, which proves the superior generalization of our method in brain tumor segmentation.

Data availability statement

The dataset BRATS2015 (Menze et al., 2014; Kistler et al., 2013) for this study can be found in the https://www.smir.ch/BRATS/Start2015. The dataset BRATS2018, BRATS2020 (Menze et al., 2014; Bakas et al., 2017, 2018) and online evaluation platform can be found in this https://ipp.cbica.upenn.edu.

Author contributions

LH: conceptualization, methodology, software, project administration, writing—original draft, writing—review, and editing. EZ: medical expert. LC: validation and project administration. ZW and BZ: supervision. SC: supervision, resources, formal analysis, and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This work was funded in part by the National Natural Science Foundation of China (Grant Nos. 82170374 and 82202139), and also supported in part by the Capital Medical Funds for Health Improvement and Research (CHF2020-1-1053).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Anand, V. K., Grampurohit, S., Aurangabadkar, P., Kori, A., Khened, M., Bhat, R. S., et al. (2020). "Brain tumor segmentation and survival prediction using automatic hard mining in 3d CNN architecture," in *International MICCAI Brainlesion Workshop* (Virtual: Springer), 310–319.

Asis-Cruz, J. D., Krishnamurthy, D., Jose, C., Cook, K. M., and Limperopoulos, C. (2022). Fetalgan: automated segmentation of fetal functional brain mri using deep generative adversarial learning and multi-scale 3D u-Net. *Front. Neurosci.* 16, 887634. doi: 10.3389/fnins.2022.887634

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data.* 4, 1–13. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv [Preprint]*. arXiv: 1811.02629. Available online at: https://arxiv.org/pdf/1811.02629.pdf

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*. doi: 10.48550/arXiv.2004.10934

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). "End-to-end object detection with transformers," in *European Conference on Computer Vision* (Glasgow: Springer), 213–229.

Chandra, S., Vakalopoulou, M., Fidon, L., Battistella, E., Estienne, T., Sun, R., et al. (2018). "Context aware 3-D residual networks for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Granada), 74–82.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). Transunet: transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*. doi: 10.48550/arXiv.2102.04306

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*. doi: 10.48550/arXiv.1412.7062

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern. Anal. Mach. Intell.* 40, 834–848. doi: 10.1109/TPAMI.2017.2699184

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). "Infogan: interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems, Vol. 29* (Barcelona).

Chen, X., Liew, J. H., Xiong, W., Chui, C.-K., and Ong, S.-H. (2018). "Focus, segment and erase: an efficient network for multi-label brain tumor segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 654–669.

Choi, J., Kim, T., and Kim, C. (2019). "Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul), 6830–6840.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*. doi: 10.48550/arXiv.2003.10555

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. doi: 10.48550/arXiv.1810.04805

Ding, Y., Zhang, C., Cao, M., Wang, Y., Chen, D., Zhang, N., et al. (2021). Tostagan: an end-to-end two-stage generative adversarial network for brain tumor segmentation. *Neurocomputing* 462, 141–153. doi: 10.1016/j.neucom.2021.07.066

Dong, X., Lei, Y., Wang, T., Thomas, M., Tang, L., Curran, W. J., et al. (2019). Automatic multiorgan segmentation in thorax ct images using u-net-gan. *Med. Phys.* 46, 2157–2168. doi: 10.1002/mp.13458

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. doi: 10.48550/arXiv.2010.11929

Girshick, R. (2015). "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago: IEEE), 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH: IEEE), 580–587.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems, Vol. 27* (Montreal, QC).

Han, Z., Wei, B., Mercado, A., Leung, S., and Li, S. (2018). Spine-gan: Semantic segmentation of multiple spinal structures. *Med. Image Anal.* 50, 23–35. doi: 10.1016/j.media.2018.08.005

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., et al. (2022). "UNETR: transformers for 3D medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Waikoloa, HI: IEEE), 574–584.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 2961–2969.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 770–778.

He, R., Xu, S., Liu, Y., Li, Q., Liu, Y., Zhao, N., et al. (2021). Three-dimensional liver image segmentation using generative adversarial networks based on feature restoration. *Front. Med.* 8, 794969. doi: 10.3389/fmed.2021.794969

Hu, K., Gan, Q., Zhang, Y., Deng, S., Xiao, F., Huang, W., et al. (2019). Brain tumor segmentation using multi-cascaded convolutional neural networks and conditional random field. *IEEE Access* 7, 92615–92629. doi: 10.1109/ACCESS.2019.2927433

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 4700–4708.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 1125–1134.

Jiang, Y., Chang, S., and Wang, Z. (2021). Transgan: two pure transformers can make one strong gan, and that can scale up. *Adv. Neural Inf. Process. Syst.* 34, 14745–14758. doi: 10.48550/arXiv.2102.07074

Khan, M. Z., Gajendran, M. K., Lee, Y., and Khan, M. A. (2021). Deep neural architectures for medical image semantic segmentation. *IEEE Access* 9, 83002–83024. doi: 10.1109/ACCESS.2021.3086530

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. doi: 10.48550/arXiv.1412.6980

Kistler, M., Bonaretti S., Pfahrer, M., Niklaus, R., and Buchler, P. (2013). The virtual skeleton database: An open access repository for biomedical research and collaboration. *J. Med. Internet Res.* 15, e245. doi: 10.2196/jmir.2930

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems, Vol. 25* (Lake Tahoe).

Le L, Yefeng Z, Gustavo C, Lin Y. (2017). "Deep learning and convolutional neural networks for medical image computing - precision medicine, high performance and large-scale datasets," in *Advances in Computer Vision and Pattern Recognition* (Springer).

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Lin, G., Milan, A., Shen, C., and Reid, I. (2017). "Refinenet: multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 1925–1934.

Liu, M. (2018). "Coarse-to-fine deep convolutional neural networks for multi-modality brain tumor semantic segmentation," in *MICCAI BraTs Conference* (Granada).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016). "Ssd: single shot multibox detector," in *European Conference on Computer Vision* (Amsterdam: Springer), 21–37.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). Roberta: a robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. doi: 10.48550/arXiv.1907.11692

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston), 3431–3440.

Menze, B., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2014). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imag.* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Mirza, M., and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*. doi: 10.48550/arXiv.1411.1784

Myronenko, A. (2018). "3D MRI brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop* (Granada: Springer), 311–320.

Naseer, M. M., Ranasinghe, K., Khan, S. H., Hayat, M., Shahbaz Khan, F., and Yang, M.-H. (2021). Intriguing properties of vision transformers. *Adv. Neural Inf. Process. Syst*. 34, 23296–23308. doi: 10.48550/arXiv.2105.10497

Nishio, M., Fujimoto, K., Matsuo, H., Muramatsu, C., Sakamoto, R., and Fujita, H. (2021). Lung cancer segmentation with transfer learning: usefulness of a pretrained model constructed from an artificial dataset generated using a generative adversarial network. *Front. Artif. Intell.* 4, 694815. doi: 10.3389/frai.2021.694815

Odena, A., Olah, C., and Shlens, J. (2017). "Conditional image synthesis with auxiliary classifier gans," in *International Conference on Machine Learning* (Sydney), 2642–2651.

Oh, K. T., Lee, S., Lee, H., Yun, M., and Yoo, S. K. (2020). Semantic segmentation of white matter in fdg-pet using generative adversarial network. *J. Digit. Imaging* 33, 816–825. doi: 10.1007/s10278-020-00321-5

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. (2018). Available online at: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Razmjooy, N., Ashourian, M., Karimifard, M., Estrela, V. V., Loschi, H. J., Do Nascimento, D., et al. (2020). Computer-aided diagnosis of skin cancer: a review. *Curr. Med. Imaging* 16, 781–793. doi: 10.2174/1573405616666200129095242

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 779–788.

Redmon, J., and Farhadi, A. (2017). "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 7263–7271.

Redmon, J., and Farhadi, A. (2018). Yolov3: an incremental improvement. *arXiv preprint arXiv:1804.02767*. doi: 10.48550/arXiv.1804.02767

Ren, S., He, K., Girshick, R., and Sun, J. (2015). "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems, Vol. 28* (Montreal, QC: Sydney).

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Munich: Springer), 234–241.

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. doi: 10.48550/arXiv.1409.1556

Stoitsis, J., Valavanis, I., Mougiakakou, S. G., Golemati, S., Nikita, A., and Nikita, K. S. (2006). Computer aided diagnosis based on medical image processing and artificial intelligence methods. *Nucl. Instrum. Methods Phys. Res. A: Accel. Spectrom. Detect. Assoc. Equip.* 569, 591–595. doi: 10.1016/j.nima.2006.08.134

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA: IEEE), 1–9.

Tang, J., Li, T., Shu, H., and Zhu, H. (2020). "Variational-autoencoder regularized 3D multiresunet for the brats 2020 brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Virtual: Springer), 431–440.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems*. p. 5998–6008.

Wang, T., Wang, M., Zhu, W., Wang, L., Chen, Z., Peng, Y., et al. (2021). Semi-msst-gan: a semi-supervised segmentation method for corneal ulcer segmentation in slit-lamp images. *Front. Neurosci.* 15, 793377. doi: 10.3389/fnins.2021.793377

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., and Li, J. (2021). "Transbts: multimodal brain tumor segmentation using transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Virtual: Springer), 109–119.

Xue, Y., Xu, T., Zhang, H., Long, L. R., and Huang, X. (2018). Segan: adversarial network with multi-scale l1 loss for medical image segmentation. *Neuroinformatics* 16, 383–392. doi: 10.1007/s12021-018-9377-x

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). "Xlnet: generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems, Vol. 32* (Vancouver, BC).

Zhan, Q., Liu, Y., Liu, Y., and Hu, W. (2021). Frontal cortex segmentation of brain pet imaging using deep neural networks. *Front. Neurosci*. 15, 796172. doi: 10.3389/fnins.2021.796172

Zhang, W., Yang, G., Huang, H., Yang, W., Xu, X., Liu, Y., et al. (2021). Me-net: multi-encoder net framework for brain tumor segmentation. *Int. J. Imaging Syst. Technol*. 31, 1834–1848. doi: 10.1002/ima.22571

Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., and Fan, Y. (2018). A deep learning model integrating fcnns and crfs for brain tumor segmentation. *Med. Image Anal.* 43, 98–111. doi: 10.1016/j.media.2017.10.002

Zhou, J., Ye, J., Liang, Y., Zhao, J., Wu, Y., Luo, S., et al. (2022). scse-nl v-net: A brain tumor automatic segmentation method based on spatial and channel "squeeze-and-excitation" network with non-local block. *Front. Neurosci.* 16, 916818. doi: 10.3389/fnins.2022.916818

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 2223–2232.

Zhu, Q., Du, B., Turkbey, B., Choyke, P. L., and Yan, P. (2017). "Deeply-supervised cnn for prostate segmentation," in *2017 International Joint Conference on Neural Networks (IJCNN)* (Anchorage, AK: IEEE), 178–184.