Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29168

Cross-Modality Attention with TransUNet for Robust Multimodal Brain Tumor Segmentation

Ruocun Wang

Dept of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China
1155189441@link.cuhk.edu.hk

Abstract. Segmentation of brain tumors is an important but difficult part of medical image processing. Automatic segmentation serves as an effective aid for physicians to achieve precise diagnoses and determine optimal treatment strategies for their patients. This article presents an innovative framework that integrates a cross-modality attention mechanism with TransUNet to tackle the complexities of multi-modal segmentation in brain MRI. The suggested method uses information from different MRI sequences to make segmentation more accurate. A cross-modality attention block is trained to capture how different modalities interact with one another, which makes modal-wise fusion work better. The approach shows the full structure of the tumor by using CNN-based multi-channel local feature extraction and transformer-based multi-head self-attention modules. The network achieved an average Dice score of 0.873, tested on the BraTS2021 dataset. The suggested method demonstrates superior efficacy in characterizing tumor morphology compared to existing models while maintaining computational efficiency. The findings suggest that this technique could be advantageous in clinical environments for accurate and reliable brain tumor segmentation.

Keywords: Brain Tumor Segmentation, Multimodal MRI, TransUnet, Image Segmentation.

1. Introduction

Brain tumors pose a significant and fatal threat to patients. Accurate evaluations and treatment strategies are essential for enhancing patients' well-being and extending their lifespan. Medical imaging is a crucial and essential method for diagnosing and evaluating brain cancers. Magnetic resonance imaging (MRI) is often considered the first reference for tumors because of its ability to capture particular anatomical structures. Radiologists are prone to encounter many problems when they manually segment brain tumors, such as subjectivity, unpredictability, and the fact that it takes experience.

Because of these limits, tools based on artificial intelligence have been built that work better, more consistently, and on a larger scale. AI-based segmentation systems that use deep learning techniques have shown that they can be more accurate and efficient than traditional methods, which makes them an essential tool in current medical imaging. Over the years, many AI frameworks have been created and evolved for medical picture segmentation, such as brain MRI processing. Each framework backbone has its strengths and weaknesses: Convolutional Neural Networks (CNNs) are

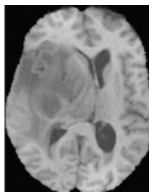
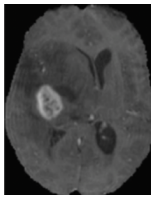
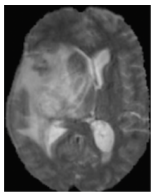
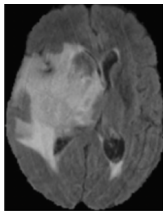
© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ29168

the basic structures that deep learning uses to analyze images. CNNs excel in learning spatial features and pattern arrangements, but they may have drawbacks when dealing with the whole picture. U-Net is a well-known architecture made particularly for jobs that need segmenting numerous objects. The utilization of skip connections between the CNN-based encoder and decoder is quite beneficial for keeping spatial information, which makes it very efficient at breaking down complex structures like brain tissues. Vision transformer models utilize a self-attention technique to allow for global feature modeling. They outperform CNNs when monitoring vast amounts of data or complex scenes, like long-range dependencies. These frameworks can do well at specific segmentation tasks, and multi-modal MRI images have the potential to further enhance their performance by providing extra dependent information.

MRI scans are typically obtained in various sequences or modalities, each delineating different tissue characteristics and exhibiting unique efficacy in brain tumors, as illustrated in Table 1. Single-modality segmentation models might work well in some situations, but they usually have trouble getting all the details about a tumor because a single sequence doesn't furnish them enough information or follow the rules of physical imaging. This problem becomes worse when dealing with rare cases or complicated diseases. On the other hand, multimodal MRI combines information from several sequences to give a better picture of the tumor's composition and edges. However, properly integrating multimodal information continues to pose a substantial difficulty because of variations in spatial resolution, intensity distributions, and modality-specific noise. Direct fusion may also result in heightened noise and a deficiency of significant inter-modality complementary information.

Table 1. Modality features of brain MRI

Sequence	T1 Weighted Imaging	T1-Weighted Contrast-Enhanced Imaging (T1CE)	T2-Weighted Imaging	Fluid-Attenuated Inversion Recovery (FLAIR)f
Image				
Focus	Anatomy and structure	Tumor enhancement and vascularity	Edema and fluid content	Lesions near cerebrospinal fluid
Significance in Brain Tumor Segmentation	Identifies structural boundaries	Highlights active tumor regions	Detects tumor-associated swelling	Differentiates lesions from surrounding fluid

In order to overcome these obstacles and capitalize on the advantages of multimodal MRI, this paper suggests a novel dual-stage brain tumor segmentation framework made up of TransUNet and Cross-Modality Attention. The following are this work's primary contributions:

• A cross-modality attention mechanism is introduced to dynamically model interactions between different MRI modalities. This module learns modality-specific importance weights, enabling effective multimodal fusion.

• Constructed a hybrid segmentation model that integrates transformer-based architectures and U-Net. This model processes multimodal 3D MRI data, leveraging global and local features for tumor segmentation.

DOI: 10.54254/2755-2721/2026.TJ29168

• The method is tested on the BraTS2021 dataset [1], which is a standard for separating brain tumors. The method works better than single-modality networks and other segmentation frameworks that have already been tested.

In summary, the proposed framework demonstrates the significant potential of combining cross-modality attention and a transformer-based U-Net for robust and accurate brain tumor segmentation.

2. Related work

Medical image analysis has recently emerged as a significant research domain. Using multimodal brain MRI to separate brain cancers is becoming more and more important. As deep learning-based methods have quickly progressed, researchers have worked hard to make better models that use multimodal MRI data to make classification tasks more accurate. This part discusses both the benefits and drawbacks of the several ways to segment brain tumors that have already been constructed and evaluated.

2.1. CNN- and Unet-based methods

Balamurugan et al. put forward a hybrid DCNN model with a modified LuNet classifier for brain tumor segmentation and classification [2]. This framework uses GLCM and VGG-16 along with Fuzzy C-Means for segmentation. The technique gets 99.7% accuracy on the SRM and BraTS datasets, which is better than other frameworks that are less complicated. Abdul Haseeb Nizamani et al. came up with FE-HU-NET, a new way to segregate brain tumors in MRI scans with proprietary UNet designs and CNN-based post-processing to make segmentation more accurate [3]. The model is tested on the BraTS2020 and LGG-MRI datasets and gets more than 99% accuracy. Ahmed M. Gab Allah et al. created Edge U-Nets [4], which use an Edge Guidance Block (EGB) to include boundary information and bespoke loss algorithms that combine boundary loss with cross-entropy loss. CNN and U-Net have an inevitable problem with this task because they both utilize local convolutions, which makes it hard for them to get the big picture and separate dependencies. This can degrade their performance on tumors with complicated forms or many locations with a risk of overfitting as well.

2.2. Transformer-based methods

Fethi Ghazouani et al. devised a method to segment brain tumors using Swin Transformer and Enhanced Local Self-Attention (ELSA) [5]. They built a structure that encodes and decodes by putting Transformer and CNN modules together. The BraTS 2021 dataset gave it an average Dice score of 89.77%, which is better than other frameworks with similar backbones. Yun Jiang et al. came up with a new 3D multimodal brain tumor segmentation method called SwinBTS [6]. This method combines the Swin Transformer with a CNN. The Enhanced Transformer (ETrans) module significantly improves feature extraction. At BraTS 2019, SwinBTS achieved a Dice score of 81.15%, indicating that it outperformed other methods in terms of segmentation. The use of transformers for this task is challenging due to their substantial memory and processing requirements, particularly when dealing with high-resolution medical images or small datasets. When the training data volume is relatively low, it struggles to converge to a reliable segmentator.

DOI: 10.54254/2755-2721/2026.TJ29168

3. Methods

3.1. Framework architecture

The suggested system, seen in Figure 1, utilizes an adapted multi-modal 3D TransUNet architecture to attain precise brain tumor segmentation. This architecture is made to manage the complicated nature of multi-modal MRI data, which includes inputs from T1, T1CE, T2, and FLAIR modalities. The first step is a patch extraction module, which breaks the 3D multi-modal MRI inputs into smaller, machine-friendly patches. This splitting makes the calculations easier and helps with image transformations.

After being transformed and augmented, patches are processed by a modality attention module. This module is essential for getting first cross-modality information because it dynamically models how different modalities interact with each other. Unique characteristics of each modality are recorded, and at the same time, complimentary information that improves segmentation accuracy is noted. The cross-modality information that has been extracted and normalized is then projected back into the 4-channel space, where it is mixed with the original patch data. This way, the network may use both the raw input and the multi-modal feature representations.

Three CNN-based multi-modal encoders are put together, each with more channels than the last. The job of each encoder is to pull out features from each modality, which lets the network learn details that are distinct to each modality. To improve feature integration, a concatenator combines the outputs of these encoders into a single representation. A set of transformer blocks is used to determine the relationships between the representations. There are eight transformer blocks in the framework, and they are all connected in series. Correlations between spatial and modality-specific data are learned by the transformer blocks.

In the decoder block, the recovered features are first up-sampled to a higher resolution. This makes sure that the spatial details with varied scales are combined together. The decoder also has skip connections from the encoder layers, which serve as residual information. This procedure enables the network to get back fine-grained details and keep the input and output consistent. The combined feature map is then processed through two convolution layers to improve and integrate the features. This stage makes sure that the features from different scales will be mixed together and be optimized for segmentation.

Finally, the decoder module puts the segmentation mask back together by combining patch-wise features and gradually increasing their resolution to match the original input resolution. Sliding window recovery is adopted to recover the whole image from the cropped patches.

DOI: 10.54254/2755-2721/2026.TJ29168

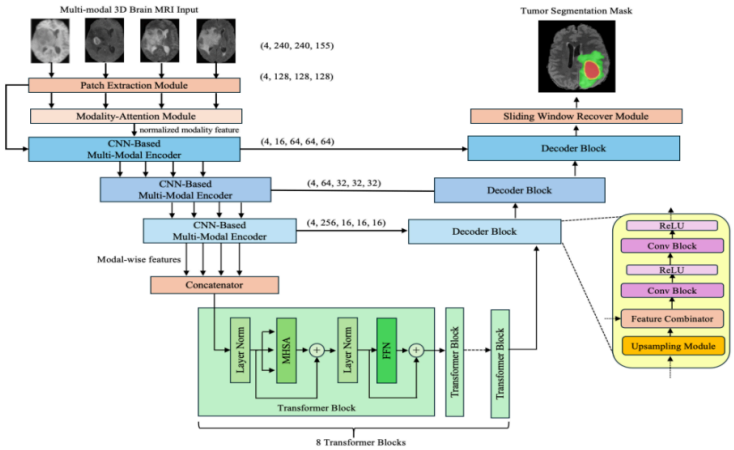


Figure 1. The proposed Trans-Unet framework for multi-modal brain tumor segmentation

3.2. Data preprocessing

The data preprocessing pipeline comprises three primary parts: normalizing, transforming the data, and cropping images. This module is used to ensure that the input data for training is of satisfactory quality and follows a consistent format. First, the multi-modal MRI data will be standardized to make voxel intensities the same for all patients and modalities. In particular, voxel intensity I is normalized by

$$I' = \frac{I-\mu}{\sigma} \tag{1}$$

where μ and σ stand for the mean and standard deviation of the intensities in the brain region. This makes sure that the intensity values are centered and scaled so that the data is always shown the same way.

Next, transform functions from the MONAI library are used to change the data. These functions are useful for improving model generalizability. The BraTS2021 dataset only has a limited number of 3D images. Therefore, the model needs to be strong enough to handle data modification and augmentation. Transformations include rotating the object in space, changing the intensity, and adding Gaussian blur. Gaussian blurring is used to smooth out changes in intensity.

Finally, data cropping is done to address the problem of MRI data having an uneven distribution of foreground (tumor) and background areas. The cropping approach makes sure that each patch has an equal amount of tumor label and backdrop, unlike random cropping. The crop operation picks a region C from the whole volume V, where the proportion P of foreground voxels meets the following condition:

$$\alpha \leq \frac{|C_{foreground}|}{|C|} \leq \beta \tag{2}$$

In this case, α and β are preset limits that keep the ratio of tumor to background voxels from being too high or too low. This makes sure that the clipped patches are useful and representational, which makes the segmentation model work better.

DOI: 10.54254/2755-2721/2026.TJ29168

3.3. Training strategy

There are three parts to the dataset: validation, testing, and training. The normal mix is 70%, 15%, and 15%. The training takes place in a GPU environment using the PyTorch framework to speed up calculations and make it easier to work with medical images.

A hybrid loss function that combines Cross Entropy Loss and Dice Loss is used by the training system. Dice Loss alone converges slowly since the tumor masks are small, and Cross-Entropy Loss alone typically makes predictions that are skewed toward the background. The combined loss function can be expressed as

$$L = \lambda_1 L_{Dice} + \lambda_2 L_{CE} \tag{3}$$

where λ_1 and λ_2 are weighting factors.

4. Experiment & results

4.1. Experiment setup

The BraTS 2021 dataset, a benchmark for brain tumor segmentation tasks targeting glioblastomas (GBM) and lower-grade gliomas (LGG), was utilized to train and assess the model. All images are pre-processed to make sure that the voxel resolution is the same (1×1×1 mm³, skull-stripped, and co-registered). The tests utilized an NVIDIA RTX A6000 GPU equipped with 48 GB of RAM, with PyTorch as the principal deep learning framework.

4.2. Results

Figure 2 shows that the training loss curves change at the beginning of the training, but after around 35 epochs, both loss functions start to go down steadily. The framework shows that optimization is stable across 100 epochs.

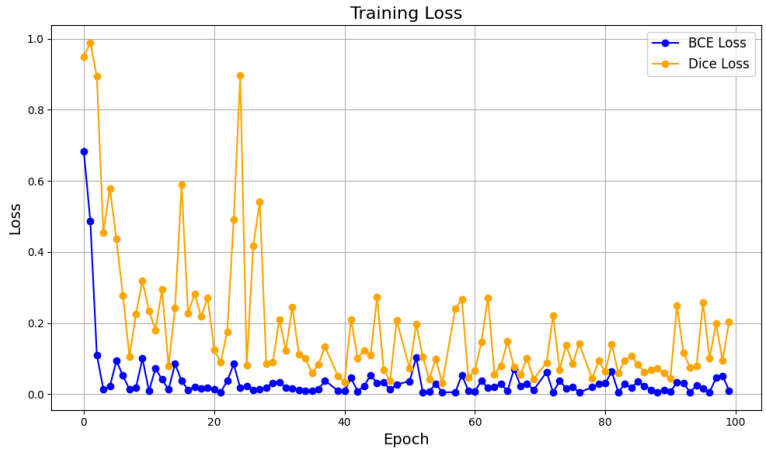


Figure 2. Training loss. The training process consists of two types of loss: Binary Cross Entropy (BCE) loss and dice loss

DOI: 10.54254/2755-2721/2026.TJ29168

The suggested model is compared with four other typical techniques. Table 2 shows a quantitative comparison that shows this framework gets state-of-the-art Dice scores. In particular, the method works better than older methods like MMCFormer and TransBTSV2, showing that it is better at handling multimodal MRI data and accurately capturing tumor features.

Table 2. Dice score comparison

Models	Dataset	WT	Dice Score TC	ET	Reference
Missing modality compensation transformer (MMC Former)	BraTS 2018	0.897	0.872	0.801	Karimijafarbigloo et al. [7]
Improved DNN with fast Fuzzy C-Means (FFCM)	BraTS 2020	0.901	0.849	0.846	Sahoo et al. [8]
Generative adversarial network (GAN) based ontransformers	BraTS 2020	0.905	0.834	0.762	Huang et al. [9]
TransBTSV2	BraTS 2020	0.892	0.849	0.803	Li et al. [10]
Cross-Modality Attention with TransUNet (This Paper)	BraTS 2021	0.904	0.872	0.842	This Paper

5. Conclusion

This research suggested a new way to use cross-modality attention and TransUNet together to segment brain tumors in multiple ways. The method efficiently encompasses the strengths of U-Net for representing local features and Transformers for modeling global context. The cross-modality attention mechanism facilitates the integration of multi-modal MRI data. Experiments using the BraTS 2021 dataset showed that the framework works at the highest level, indicating its potential for use in medicine. Future endeavors will concentrate on generalizing across various datasets, enhancing computational efficiency, addressing absent modalities, and improving result comprehensibility towards application in clinical environments.

References

[1] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, et al. (2019) Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge

[2] T. Balamurugan and E. Gnanamanoharan. (2022) Brain Tumor Segmentation and Classification using Hybrid Deep CNN with LuNet Classifier, ResearchkArticle,

[3] Jiang, Y.; Zhang, Y.; Lin, X.; Dong, J.; Cheng, T.; Liang, J. (2022) SwinBTS: A Method for 3D Multimodal Brain Tumor Segmentation Using Swin Transformer. Brain Sci.k2022, 12, 797.

[4] A.H. Nizamani, Z. Chen, A.A. Nizamani, and U.A. Bhatti. (2023) Advance brain tumor segmentation using feature fusion methods with deep U-Net model with CNN for MRI data," Journal of King Saud University – Computer and Information Sciences, vol. 35.

[5] F. Ghazouani, P. Vera, and S. Ruan. (2024) Efficient brain tumor segmentation using Swin transformer and enhanced local self-attention. International Journal of Computer Assisted Radiology and Surgery, vol. 19, pp. 273–

DOI: 10.54254/2755-2721/2026.TJ29168

281.
[6] Yousef, R.; Khan, S.; Gupta, G.; Siddiqui, T. (2023) Albahlal, B.M.; Alajlan, S.A.; Haq, M.A. U-Net-Based Models towards Optimal MR Brain Image Segmentation. Diagnostics 2023, 13, 1624.
[7] Karimijafarbigloo, S., Azad, R., Kazerouni, A., Ebadollahi, S., and Merhof, D. (2023). Mmcformer: missing modality compensation transformer for brain tumor segmentation, in Medical imaging with deep learning.
[8] Sahoo, A. K., Parida, P., Muralibabu, K., and Dash, S. (2023). An improved dnn with ffcm method for multimodal brain tumor segmentation, in Intelligent systemskwith applications.
[9] Huang, L., Zhu, E., Chen, L., Wang, Z., Chai, S., and Zhang, B. (2022b). A transformer-based generative adversarial network for brain tumor segmentation. Front. Neurosci. 16, 1054948.
[10] Li, J., Wang, W., Chen, C., Zhang, T., Zha,kS., Wang, J., et al. (2022). Transbtsv2: towards better and more efficient volumetric segmentation of medical images.