MTSegFormer: A Multitask Learning Approach for Brain Tumor Segmentation using Transformer

Meng Zhou[1] **Xudong Liu**[1] **Reyna Wu**[1]
[1]Department of Computer Science, University of Toronto, Canada
{simonzhou,xudong,reynawu}@cs.toronto.edu

Abstract

Accurate brain tumor segmentation in MRI images is crucial for effective diagnosis and treatment planning. However, traditional U-Net architecture faces challenges in capturing long-range dependencies and preserving features of small-sized tumors, which limits its performance. Hence, we present MTSegFormer, a novel learning framework for 2D brain tumor segmentation using latent transformer through the Multi-task learning paradigm. We use a UNet-like structure with a latent space transformer, and a self-supervised image decoder to build up the overall framework. We also introduce the Breath-wise Cross Attention module that aims to refine the skip connection features. Experiment shows our proposed framework achieves superior performance compared to other baselines by up to 11% in Dice and 10% in IoU score. The code is available at `https://github.com/simonZhou86/csc2516_proj`

1 Introduction

Brain tumor segmentation is a critical task in medical image processing that plays an important role in the diagnosis and treatment of brain tumors. Automated segmentation of brain tumors can assist physicians and provide an accurate and reproducible solution for tumor monitoring.

Deep learning based segmentation methods have achieved great success in various semantic segmentation tasks. Convolutional neural networks (CNN) are able to learn from the images and realize end-to-end dense semantic segmentation with impressive segmentation accuracy [1]. Though CNN has shown excellent representation capacity, it suffers from the lack of long-distance dependency due to limited receptive fields. Recently, Vision Transformers (ViT) have also reached state-of-the-art performance on various computer vision tasks [2]. By leveraging the transformer structure, ViT is more powerful for learning long-distance dependencies.

Multi-task learning is a technique where the model optimizes an auxiliary task additionally to improve the performance of the main task or accelerate the training process. One approach to multi-task learning is adding a self-supervised auxiliary task in which no manually labeled data is needed. For the semantic segmentation task, a natural choice is to include an additional autoencoder task to help the model preserve fine-grained information [3].

In this work, we propose a semantic segmentation method for 2D brain tumor segmentation from MRI. We adapt a UNet-like structure [4] and append a transformer module preceding the decoder to capture the global long-term dependency in the latent space. Besides, we add a self-supervised auxiliary reconstruction task in parallel with the segmentation task to preserve fine-grained information. The proposed method demonstrates the capacity to improve the accuracy of brain tumor segmentation.

Preprint. Under review.

2 Related Work

Recent advances in deep learning have enabled highly accurate and efficient medical image segmentation. U-Net, a convolutional neural network architecture consisted of a contracting path and a symmetric expanding path, has been adopted for medical image segmentation tasks [4]. However, U-Net has limitations in explicitly modeling long-range dependencies due to the intrinsic locality of convolution operations [5]. To address this issue, transformer-based architectures like TransUNet and TransBTS have been introduced. While TransUNet, a 2D network that adopts ViT with a U-Net-like decoder, focuses solely on the spatial correlation between tokenized image patches [5], Transformer in 3D CNN for 3D MRI Brain Tumor Segmentation (TransBTS) is based on 3D CNN to model the long-range dependencies in both depth and spatial dimensions simultaneously for volumetric segmentation [6].

Multi-task learning has been shown to improve segmentation accuracy by integrating related tasks such as classification or detection. For small brain tumor segmentation from MRI, multi-task learning has been applied with U-module to help retain features of small-sized tumors, which are easily overlooked due to the decreasing feature resolution after each encoder layer of U-Net-based models [3]. Similarly, in 3D automated breast ultrasound imaging, multi-task learning has been applied to jointly segment and classify tumors by training the multi-task network using an iterative training strategy to extract auto-context features for both tasks where predicted segmentation maps are added as a part of the input to guide feature extraction [7]. Additionally, a multi-task learning network has been developed to jointly train Glioma segmentation and IDH genotyping in an end-to-end manner by sharing the spatial and global feature representation extracted from the hybrid CNN transformer encoder [8].

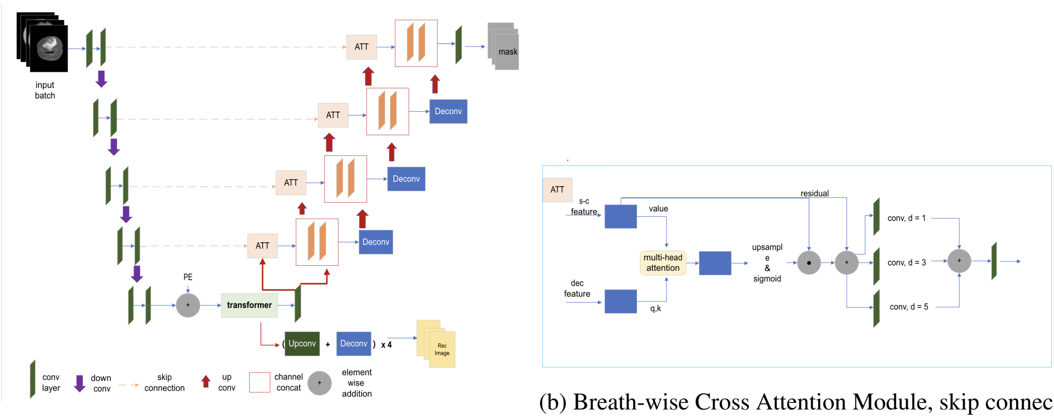
3 Method

3.1 Overall Framework

Figure 1a summarizes the proposed framework for this work, which consists of four core modules for the segmentation task. 1). the image encoder, shown as the left branch in Figure 1a, is utilized to generate compact feature maps that could capture the local spatial information. 2). the transformer module is designed to capture the long-term dependency in a global manner. 3). the segmentor, shown as the right branch in Figure 1a, will gradually up-sample the feature maps with the convolutions to produce a high-resolution segmentation mask. The encoder-segmentor constitutes the UNet-like architecture [4, 9] with skip connection. 4). the image decoder, down right corner in Figure 1a aims to produce the reconstructed image to preserve the relevant and important information in the original image as an auxiliary task. We hypothesize that the image reconstruction task will force the feature extractor (encoder and transformer) retains the important local and global features in the latent space and mitigates the side effects caused by down-sample convolutional layers.

3.2 Breath-wise Cross Attention Module

To ensure the feature maps pass into each convolution in the decoder that contains both the high-resolution features from skip connections and the semantic richness from the deeper layer of the network, we adapt the idea from [10] to design the cross attention module as shown in Figure 1b. The deeper layer features guide the attention module to not focus on the irrelevant or less-informative regions from the skip connection features but highlight the significant regions where the tumor presents. We use the traditional multi-head attention where query and key are the deeper layer features and the value is skip connection features to get the attention output, followed by a residual connection to stabilize the whole training process. Next, considering the medical images require fine-grained information, we use the $\{1, 3, 5\}$-dilated convolution to learn the multi-scale spatial context [11] on the attention output. Finally, we aggregate the features from these three dilated convolutions by element-wise addition.



(a) An overview of our proposed method which contains an encoder, a transformer module, a segmentor and a decoder.

(b) Breath-wise Cross Attention Module, skip connection features and deeper layer features are fed into the multi-head attention module, following by dilated convolutions to extract multi-scale context.

Figure 1: Proposed method and the attention module

3.3 Loss function

Both the main task and auxiliary task are trained simultaneously. The combination of the dice coefficient loss and the binary cross entropy loss is used to supervise the encoder and segmentor, readers can refer to [12] for more details of these two losses.

The loss function for the auxiliary task ($\mathcal{L}_{recon}$) is defined as in Equation (1), where $\hat{I}$ is the reconstructed image and I is the original image. The first loss term measures the pixel differences between $\hat{I}$ and I; the second loss term measures the image gradient differences in x and y-direction; and the third loss term is the perceptual differences [13]. The final loss is the weighted combination of $\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{recon}$, e.g.,$\mathcal{L}_{\text{total}} = c_1\mathcal{L}_{\text{seg}} + c_2\mathcal{L}_{recon}$, where $c_1 + c_2 = 1$.

$$\mathcal{L}_{\text{recon}} = \|\hat{I} - I\|_F^2 + \|\nabla\hat{I} - \nabla I\|_2^2 + \frac{1}{C}\sum_{k=1}^{C}\|f_i^k(\hat{I}) - f_i^k(I)\|_2^2 \tag{1}$$

3.4 Data

Our data cohort contains 259 HGG patients and 76 LGG patients, a total of 335 patients, from the BraTS 2019 dataset [14–16] [1]. Among these patients, 268 patients are used for training and 67 patients are used for testing. The original 3D data are with size $240 \times 240 \times 155$, which we then reshape to $\#S \times 128 \times 128$ where $\#S$ indicates the total single-channel slices we have for all patients. To achieve this, we treat every slice independently to each other, reshape to 128×128, and remove all zero-valued segmentation masks and corresponding brain images afterward, since we are interested in the slices with brain tumor present. Finally, we normalize all images within the range of $[0, 1]$. Due the computational and time limitation, we do not use any data augmentation methods.

4 Experiments and Results

All programs are implemented in Python and PyTorch frameworks. Every experiment runs for 10-12 hours using a single V100 GPU. All models run for 100 epochs with a batch size of 16 and an initial learning rate of 0.001, a decay factor of 0.1 for every 30 epochs, and the optimizer is Adam. We compared to UNet [4], Attention UNet [17], and the 2D version of TransBTS [6], the quantitative segmentation results are reported in Table 1. Our proposed method outperforms all other works in Dice and IoU scores by 11% and 10%, respectively, indicating the superior performance of our auxiliary decoder and the usage of the latent transformer. Below figure shows the original image and segmentation mask, followed by the predicted mask generated from the four models we described

[1] https://www.med.upenn.edu/cbica/brats2019/data.html

above, we use a hard thresholding of value 0.5 to binarize all predicted masks. Next, we explore the importance of the transformer and auxiliary decoder in our proposed framework by conducting ablation studies. Table 2 shows the quantitative results, where *(A)* and *(B)* are the proposed framework without the transformer and decoder module, respectively. *(C)* is changing to a different attention module, and the last one is the proposed method. We can see that, compared to *(B)* and *(A)*, if we do not have either the decoder for the auxiliary task or the transformer, the Dice and IoU score drops significantly, which validate our hypothesis in 3.1. Furthermore, the superior performance of the cross attention over the attention-unet block *(C)* [17] demonstrates the effectiveness of the proposed attention module discussed in 3.2.

	Accuracy % ↑	Dice Score ↑	IoU Score ↑
UNet	0.987	0.632	0.491
Attention UNet	0.988	0.676	0.534
TransBTS	0.985	0.703	0.555
Ours	**0.989**	**0.742**	**0.596**

Table 1: Main results for tumor segmentation. Higher values indicate better performance for all metrics.

	Accuracy % ↑	Dice Score ↑	IoU Score ↑
(A) ours w/o transformer	**0.989**	0.616	0.530
(B) ours w/o recon	0.984	0.640	0.482
(C) ours with diff. att block	0.977	0.651	0.495
(D) ours with cross attention	**0.989**	**0.742**	**0.596**

Table 2: Results for the Ablation study. Higher values indicate better performance for all metrics. w/o and diff. are abbreviations for without and different, respectively.

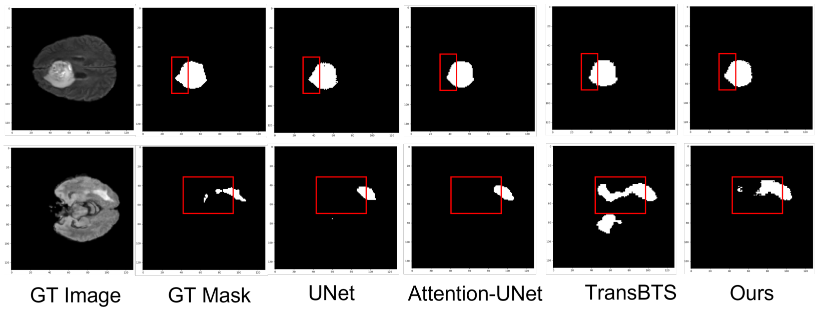


Figure 2: Qualitative Results, "GT" stands for the Ground truth. The red bounding box highlights the difference in the predicted mask between different methods.

5 Conclusion

In this work, we proposed a novel framework for 2D brain tumor segmentation. The proposed method outperforms several segmentation models, such as UNet, Attention UNet, and the 2D version of TransBTS, in terms of Dice and IoU scores. The ablation study we conducted validate the effectiveness of the three core modules: image decoder, latent transformer, and the cross attention. The results have shown the importance of the auxiliary image reconstruction task in preserving fine-grained information and the superiority of the proposed cross-attention module in capturing significant regions where tumor presents. As the future work, we want to investigate the fully transformer architecture, i.e., replace encoder and segmentor with transformer encoder and decoder layers. We also notice that our data cohort is imbalanced, so applying data augmentation techniques to the current dataset may further improved the performance.

6 Acknowledgment

This work is enabled in part by support provided by Compute Canada (https://alliancecan.ca/en)

7 Team Contribution

Meng Z. developed and implemented the proposed method, loss functions, training script, and baseline methods.
Xudong L. implemented the proposed method, training script, evaluation metrics, and baseline methods.
Reyna W. implemented the whole data pre-processing pipeline, training script and necessary helper functions.
All members participated in running the experiments in this work.

References

[1] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[3] Duc-Ky Ngo, Minh-Trieu Tran, Soo-Hyung Kim, Hyung-Jeong Yang, and Guee-Sang Lee. Multi-task learning for small brain tumor segmentation from mri. *Applied Sciences*, 10(21): 7790, 2020.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[5] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *CoRR*, abs/2102.04306, 2021. URL https://arxiv.org/abs/2102.04306.

[6] Wenxuan Wang, Chen Chen, Meng Ding, Jiangyun Li, Hong Yu, and Sen Zha. Transbts: Multimodal brain tumor segmentation using transformer. *CoRR*, abs/2103.04430, 2021. URL https://arxiv.org/abs/2103.04430.

[7] Yue Zhou, Houjin Chen, Yanfeng Li, Qin Liu, Xuanang Xu, Shu Wang, Pew-Thian Yap, and Dinggang Shen. Multi-task learning for segmentation and classification of tumors in 3d automated breast ultrasound images. *Medical Image Analysis*, 70:101918, 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101918. URL https://www.sciencedirect.com/science/article/pii/S1361841520302826.

[8] Jianhong Cheng, Jin Liu, Hulin Kuang, and Jianxin Wang. A fully automated multimodal mri-based multi-task learning for glioma segmentation and idh genotyping. *IEEE Transactions on Medical Imaging*, 41(6):1520–1532, 2022. doi: 10.1109/TMI.2022.3142321.

[9] Mehrdad Noori, Ali Bahri, and Karim Mohammadi. Attention-guided version of 2d unet for automatic brain tumor segmentation. In *2019 9th international conference on computer and knowledge engineering (ICCKE)*, pages 269–275. IEEE, 2019.

[10] Olivier Petit, Nicolas Thome, Clement Rambour, Loic Themyr, Toby Collins, and Luc Soler. U-net transformer: Self and cross attention for medical image segmentation. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction*

with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12, pages 267–276. Springer, 2021.

[11] Meng Zhou, Xiaolan Xu, and Yuxuan Zhang. An attention-based multi-scale feature learning network for multimodal medical image fusion. *arXiv preprint arXiv:2212.04661*, 2022.

[12] Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*, pages 1–7. IEEE, 2020.

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[14] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

[15] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. 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*, 2018.

[16] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[17] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.