BRAIN TUMOR SEGMENTATION THROUGH SUPERVOXEL TRANSFORMER

Yunfei Xie[1,*], *Ce Zhou*[1,*], *Jieru Mei*[3], *Xianhang Li*[2], *Cihang Xie*[2], *Yuyin Zhou*[2]

[1] Huazhong University of Science and Technology
[2] University of California, Santa Cruz
[3] The Johns Hopkins University

ABSTRACT

Segmenting brain tumors presents a multifaceted challenge due to their varied appearances and scales. This study endeavors to develop segmentation models with broad applicability across brain tumors, highlighting the importance of models capable of accommodating diverse lesion types, institutions, and demographic characteristics. Specifically, we investigate two CNN-Transformer hybrid models within the BraTS-ISBI 2024 challenge, which seeks to advance generalized segmentation models for brain tumors. The first model is the decoder-only variant of the 3D TransUNet, built upon nnUNet, featuring a CNN encoder and a hybrid CNN-Transformer decoder, which has shown superior performance on the brain metastases segmentation. The Transformer decoder treats segmentation as a binary mask classification task, refining queries through cross-attention. Additionally, we introduce a novel supervoxel Transformer to improve interpretability by efficiently clustering voxels with similar characteristics. This model incorporates a supervoxel cross-attention mechanism to iteratively refine assignments and features. Both models undergo training on an expanded augmented BraTS-ISBI 2024 challenge dataset and are combined for validation through ensembling techniques. Our best solution achieved 2nd place on the leaderboard during the validation phase.

***Index Terms*—** Brain Tumor Segmentation, Transformers, nnUNet

1. INTRODUCTION

Tumors, with their subtle intensity variations, often pose difficulties, as evidenced by inconsistencies in manual annotations[1, 2]. The wide variance in tumor appearances and dimensions across patients challenges traditional shape and location models [3]. While segmentation is extensively studied for brain imaging, this study investigates the BraTS-ISBI 2024 challenge which aims to develop generalized segmentation models across brain tumors, emphasizing models capable of spanning lesion types, institutions, and demographics.

Convolutional Neural Networks (CNNs), especially Fully Convolutional Networks (FCNs)[4], have established their prominence. The U-shaped architecture, known as U-Net [5], excels at preserving image intricacies. However, these methods often struggle with modeling long-range dependencies. Researchers have turned to Transformers, which rely on attention mechanisms, showcasing success in capturing global contexts [6]. TransUNet [7], a hybrid CNN-Transformer model, enjoys both convolution's localization with Transformer's long-range dependency.

We employ two CNN-Transformer hybrid models to enhance the generalization of brain tumor segmentation Firstly, we use the 3D version of TransUNet using nnUNet as the backbone [8]. Specifically, we use the decoder-only variant of the publicly available 3D TransUNet, which is comprised of a CNN encoder and a hybrid CNN-Transformer decoder, selected for its superior performance on brain tumors compared to other variants and the original nnUNet backbone. It treats segmentation as a binary mask classification problem, updating queries iteratively through cross-attention in Transformer decoder layers. Additionally, we introduce a novel supervoxel Transformer designed to leverage 3D supervoxel representations, which efficiently group voxels with similar characteristics in space, thereby enhancing interpretability. To enable learning from these representations, we introduce a Supervoxel Cross-Attention mechanism, refining voxel-to-supervoxel assignments and features iteratively through sliding window cross attention. This approach dynamically clusters similar voxels into supervoxels, promoting locally coherent structures in the data. Both models undergo training on an expanded BraTS-ISBI 2024 challenge dataset augmented through image registration. We then ensemble these models by averaging their predictions for validation.

2. METHOD

Given a 3D medical image (*e.g.*, CT/MR scan) $\mathrm{x} \in R^{D \times H \times W \times C}$ with the spatial resolution of $D \times H \times W$ and C number of channels, our goal is to predict the corresponding pixel-wise labelmap with size $D \times H \times W$. In this paper, we train two models for the segmentation of different types of brain tumors to enhance the generalization of brain tumor segmentation.

[*] equal contribution
Yunfei Xie and Ce Zhou's work done as interns at UCSC.

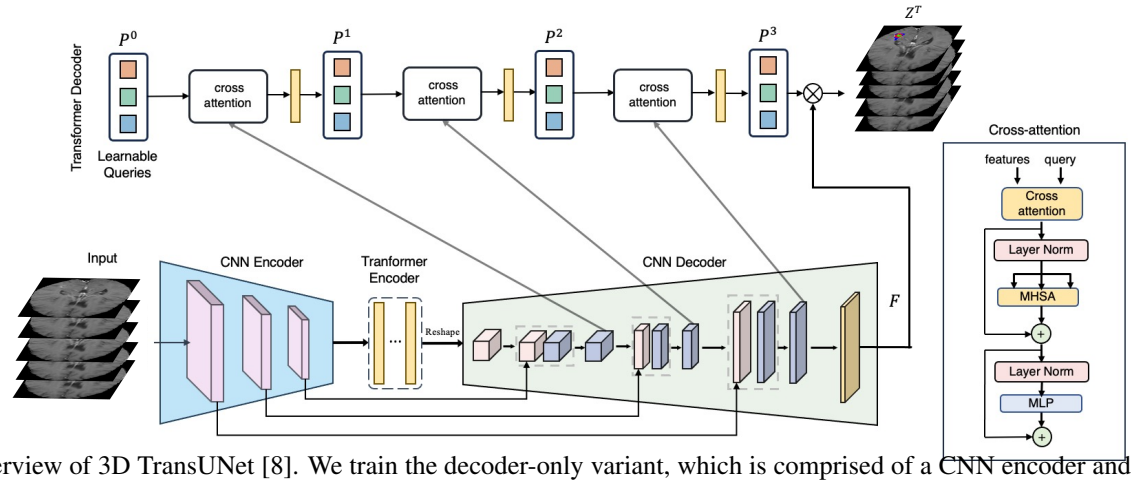

Fig. 1. Overview of 3D TransUNet [8]. We train the decoder-only variant, which is comprised of a CNN encoder and a hybrid CNN-Transformer decoder, for generalized brain tumor segmentation.

One is the decoder-only variant of the publicly available 3D TransUNet, chosen for its superior performance compared to other variants (encoder-only and encoder+decoder) and the original nnUNet backbone. In addition, we also designed a novel supervoxel Transformer, which introduces supervoxel representations in 3D volumetric medical segmentation, enabling more interpretable modeling for various substructures of brain tumors, including enhancing tumor (ET), tumor core (TC), and whole tumor (WT) with structural coherence. Both the 3D TransUNet decoder-only model and the Supervoxel Transformer are trained on an expanded training set of the BraTS-ISBI 2024 challenge using image registration for augmentation. Then we ensemble these two models by utilizing the average prediction for validation.

2.1. 3D TransUNet

3D TransUNet is a 3D hybrid model that combines CNN with Transformer based on the 3D nnUNet backbone [9], where the overall framework is demonstrated in Figure 1. In this solution, we train the decoder-only variant, which is comprised of a CNN encoder and a hybrid CNN-Transformer decoder, as detailed below.

The key idea is to formulate a regular segmentation task into a binary mask classification problem, inspired by the set prediction mechanism proposed in DETR [10]. As shown in Figure 1, we train the CNN decoder and the Transformer decoder simultaneously, allowing for the refinement of queries and feature maps. Specifically, in the t-th layer of the Transformer decoder, the refined queries are denoted by $\mathbf{P}^t \in R^{N \times d_{dec}}$. Alongside this, an intermediate feature from the U-Net is transformed into a d_{dec}-dimensional feature, represented by $\mathcal{F}$. The number of upsampling blocks in the CNN decoder aligns with the Transformer decoder layers, so multi-scale CNN features are effectively projected into the feature space $\mathcal{F} \in R^{(D_t H_t W_t) \times d_{dec}}$, where D_t, H_t, and W_t define the spatial dimensions of the feature map at the t-th upsampling block. Transitioning from the t-th to the $t+1$-th layer, the queries $\mathbf{P}^t$ are updated through the cross-attention mechanism as described by the following formula:

$$\mathbf{P}^{t+1} = \mathbf{P}^t + \text{Softmax}\left((\mathbf{P}^t \mathbf{w}_q)(\mathcal{F}^t \mathbf{w}_k)^\top\right)\mathcal{F}^t \mathbf{w}_v, \quad (1)$$

where $\mathbf{w}_q \in R^{d_{dec} \times d_q}$, $\mathbf{w}_k \in R^{d_{dec} \times d_k}$, and $\mathbf{w}_v \in R^{d_{dec} \times d_v}$ are the weight matrices that linearly project the t-th query features, keys, and values for the subsequent layer. This process is repeated, with a residual connection updating $\mathbf{P}$ after each layer, in line with the previous method ([11]). The final prediction, $\mathbf{Z}^T$, is derived through Equation 2, which details the process of converting $\mathbf{P}^T$ into the binarized segmentation map. It involves a dot product with the U-Net's last block feature, $\mathbf{F}$, resulting in $\mathbf{Z}^T$.

$$\mathbf{Z}^T = g(\mathbf{P} \times \mathbf{F}^\top), \quad (2)$$

where $g(\cdot)$ is the sigmoid activation followed by a hard thresholding operation with a threshold set at 0.5, such that it decodes region-wise binary brain tumor masks.

2.2. Supervoxel Transformer

We introduce a novel supervoxel Transformer tailored for the integration of supervoxel representations into 3D volumetric medical segmentation. Central to our approach is the development of a supervoxel cross-attention mechanism, which iteratively refines voxel-to-supervoxel assignments and features through sliding window cross attention. This iterative process enhances the cohesion of proximate and similar voxels, thereby facilitating more interpretable 3D medical image representation learning.

Supervoxel Representation. We draw inspiration from successful techniques that incorporate supervoxels into neural networks [12, 13, 14, 15, 16]. We transform voxels into a more manageable form called supervoxels—a supervoxel

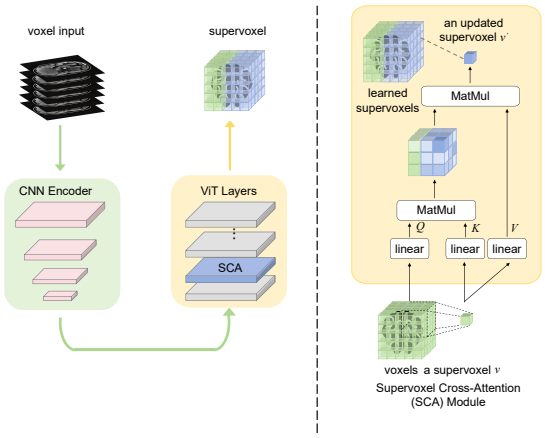

Fig. 2. Supervoxel Cross-Attention module. Illustrates aggregation of a single supervoxel for clarity.

groups together voxels that are close to each other in space and have similar characteristics. This process is noted for its efficiency in $\mathbf{S} \in R^{s_h \times s_w \times s_d \times s_c}$, where $\mathbf{S}$ represents the supervoxel representation with dimensions for height (s_h), width (s_w), depth (s_d), and channels (s_c). Supervoxels naturally follow the contours of organs or structures within the volume, helping to cluster the organ volume into meaningful groups that accurately reflect 3D structures. This not only makes our model more interpretable by closely aligning with how organs actually appear in 3D space but also enhances processing efficiency by reducing the complexity of the data.

Supervoxel Cross-Attention Our method clusters adjacent and similar voxels into a denser supervoxel representation by iteratively refining voxel-to-supervoxel assignments and their respective features using sliding window cross attention [16]. This process involves evaluating the potential assignment of each voxel i to its supervoxel neighbors within a $3 \times 3 \times 3$ vicinity, denoted as $\mathcal{N}_i$. Similarly, each supervoxel v is considered to potentially encompass a $9 \times 9 \times 9$ neighborhood of voxels, represented by $\mathcal{W}_v$. Illustrated in fig. 2, at each iteration t, the feature $\mathbf{S}_v^t$ of a supervoxel v is updated by aggregating information from the voxels assigned to it:

$$\mathbf{S}_v^t = \mathbf{S}_v^{t-1} + \sum_{i \in \mathcal{W}_v} \text{softmax}\left(\mathbf{q}_{\mathbf{S}_v^{t-1}} \cdot \mathbf{k}_{\mathbf{V}_i^{t-1}}\right) \mathbf{v}_{\mathbf{V}_i^{t-1}}, \quad (3)$$

Similarly, voxel features $\mathbf{V}_i^t$ are updated by gathering information from its corresponding supervoxels.

$$\mathbf{V}_i^t = \mathbf{V}_i^{t-1} + \sum_{v \in \mathcal{N}_i} \text{softmax}\left(\mathbf{q}_{\mathbf{V}_i^{t-1}} \cdot \mathbf{k}_{\mathbf{S}_v^{t-1}}\right) \mathbf{v}_{\mathbf{S}_v^{t-1}}, \quad (4)$$

where q, k and v are the query, key and value tensors, generated by an MLP applied to its respective feature of supervoxels ($\mathbf{S}_v^{t-1}$) and voxels ($\mathbf{V}_i^{t-1}$) in the previous iteration $t-1$.

The proposed supervoxel cross-attention mechanism effectively consolidates proximate and similar voxels into cohesive supervoxels that accurately delineate organ boundaries. By redefining supervoxel generation through sliding window cross attention, our approach enables the seamless integration of supervoxel representation into the network, allowing for end-to-end training.

2.3. Data Augmentation via Registration

Following the previous solution[17] for the Brain Tumor Segmentation (BraTS) challenge 2023 a larger dataset was built via registration. Each scan can be registered with any other scan and then warped into the other. Advanced Normalization Tools (ANTs [1]) package is used to perform this registration. It is a software package used for normalizing data to a template, and it provides different scripts that enable the application of different transformations on the images, such as rigid, affine, non-linear and all of them combined. Usually, one (first) image is used as the moving image, and the other (second) one as the fixed image - meaning that the moving image is warped into the fixed image space by applying the computed transformation, and also the inverse transformation in order to warp the fixed image into the moving image space. By applying registration to our training data, we extend the original training set from 2252 scans to 21,542 scans.

3. EXPERIMENTS

3.1. Dataset and Evaluation

We use 21,542 synthetically generated cases using registration from the original 2252 cases of BraTS-ISBI2024 challenge for training, and 360 cases for validation.

We follow the official guidelines and evaluate our model using lesion-wise Dice score(DSC) and Hausdorff distance-95 (HD95). These metrics assess overlap and maximum boundary distance between predicted and ground truth lesions, respectively. Evaluating the model lesion-by-lesion provides insights into its ability to detect and segment abnormalities without bias towards only capturing large lesions.

3.2. Implementation Details

We train both the 3D TransUNet and Supervoxel Transformer with a learning rate of 0.0003 and a batch size of 2 for 1000 epochs and 550 epochs separately. The original volume is cropped into $128 \times 128 \times 128$ patches during training, and we apply the same data augmentation techniques as in the 3D TransUNet.

For the ensemble, we use the most straightforward method of averaging the probabilities from all checkpoints. We en-

[1] http://stnava.github.io/ANTs/

Solutions	HD95				DSC			
	WT	TC	ET	Mean	WT	TC	ET	Mean
S_T	30.444	26.865	30.538	29.27	0.857	0.849	0.845	0.850
S_V	25.943	29.732	25.943	27.04	0.868	0.843	0.841	0.851
$S_{T,V}$	24.833	23.93	27.676	**25.47**	0.869	0.852	0.849	**0.857**

Table 1. Results of the validation set.

semble two checkpoints, one from each of the two models, to obtain the final results.

3.3. Post-processing

The post-processing step plays a crucial role in the overall performance, especially for recent year's BraTs competition, as the organizer decided to use the new evaluation metrics from study-wise to lesion-wise performance, where False positive (FP) and negative (FN) are penalized severely with HD95 and DSC metrics[18]. We also apply the region-base learning to better understand the lesion-wise label. Our experiments show that raw segmentation prediction contains many FPs due to small-size predicted lesions. To alleviate this, we do the some following traditional computer vision methods like thresholding and connected components for each output channel (TC, WT, and ET).

The post-processing step involves three main stages: thresholding, connected component analysis, and filtering.

In the thresholding stage, we define threshold values T_{TC}, T_{WT}, and T_{ET} for each channel (TC, WT, and ET) respectively. We then get the input heatmap x_p from the network and iterate through each voxel in the input heatmap x to apply the thresholding.

Next, we perform connected component analysis on the binary output map x_b to obtain connected components y_{cc} and count the number of connected components N_{cc} in y_{cc}. This step groups the predicted connected tumor voxels into lesions.

Finally, we filter every group based on tumorous voxel count and the mean of tumorous voxel probabilities. We define upper size threshold ($T_{s,u}$), lower size threshold ($T_{s,l}$), upper probability threshold ($T_{p,u}$), and mid probability threshold ($T_{p,m}$) from our prior knowledge and experience, ensuring that $T_{s,u} \geq T_{s,l}$, $T_{s,l} \geq 0$, and $0 \leq T_{p,u}, T_{p,m} < 1$. For each connected component, we count the number of tumorous pixels $size$ and calculate the mean probability $mean$ using the corresponding voxels in the predicted tumor heatmap x_p. If the $size$ and the mean probability $mean$ of the n-th connected component satisfy that $size \geq T_{s,u}, mean \geq T_{p,u}$, or $T_{s,l} \leq size < T_{s,u}, mean \geq T_{p,m}$, we add the connected component to the output tensor y.

3.4. Qualitative Results

We refer to our solutions using the following abbreviations:

- **T**: 3D TransUNet.
- **V**: Supervoxel Transformer.

The results in Table 1 demonstrate the performance differences between the S_V and S_T models for tumor segmentation. The S_V model outperforms the S_T model in terms of average HD95 (27.04 vs. 29.27) and DSC (0.851 vs. 0.850) metrics, suggesting its superiority in overall segmentation accuracy and robustness.

The S_V model's performance is particularly impressive in the ET region, where it achieves the best HD95 (25.943) and DSC (0.841) scores among the three solutions. Although the S_T model obtains the second-best DSC scores in the WT (0.857) and TC (0.849) regions, its overall performance is notably lower compared to the S_V model.

Building upon the strong performance of the S_V model, the $S_{T,V}$ ensemble solution further enhances the segmentation results across all tumor regions. The ensemble model consistently achieves the lowest mean HD95 (25.47) and the highest mean DSC (0.857), demonstrating the benefits of combining the strengths of both S_T and S_V models. The $S_{T,V}$ model obtains the best HD95 scores for WT (24.833) and TC (23.93), and the highest DSC scores for WT (0.869) and TC (0.852).

In conclusion, the S_V model exhibits superior performance compared to the S_T model, especially in the challenging ET region. The $S_{T,V}$ ensemble solution further improves upon the already impressive results of the S_V model, showcasing the value of ensemble techniques in achieving consistent and robust segmentation performance across all tumor regions. These findings underscore the importance of leveraging advanced modeling approaches, such as the Supervoxels and ensemble methods, to obtain optimal results in brain tumor segmentation tasks.

4. CONCLUSIONS

In this study, we examine two CNN-Transformer hybrid models, 3D TransUNet and a novel Supervoxel Transformer, for brain tumor segmentation. The models are trained on an augmented dataset and combined via ensembling. Results show the superiority of the Supervoxel Transformer, especially in the challenging enhancing tumor area. The ensemble model further boosts accuracy and robustness across all tumor regions. These findings stress the significance of advanced modeling methods and ensemble techniques for optimal performance in brain tumor segmentation.

5. ACKNOWLEDGMENTS

We thank the AWS Cloud Credit for Research Program for supporting our computing needs.

6. REFERENCES

[1] Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, and et al., "The brain tumor segmentation (brats) challenge 2023: Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds)," 2024.

[2] Maruf Adewole, Jeffrey D. Rudie, Anu Gbadamosi, and et al., "The brain tumor segmentation (brats) challenge 2023: Glioma segmentation in sub-saharan africa patient population (brats-africa)," 2023.

[3] Félix Renard, Soulaimane Guedria, Noel De Palma, and Nicolas Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Scientific Reports*, vol. 10, no. 1, pp. 13724, 2020.

[4] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[7] 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," *arXiv preprint arXiv:2102.04306*, 2021.

[8] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew Lungren, Lei Xing, Le Lu, Alan L Yuille, and Yuyin Zhou, "3d transunet: Advancing medical image segmentation through vision transformers," *arXiv preprint arXiv:2310.07781*, 2023.

[9] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020.

[11] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[12] Varun Jampani, Deqing Sun, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz, "Superpixel sampling networks," in *ECCV*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, Eds., 2018.

[13] Fengting Yang, Qian Sun, Hailin Jin, and Zihan Zhou, "Superpixel segmentation with fully convolutional networks," in *CVPR*, 2020.

[14] Huaibo Huang, Xiaoqiang Zhou, Jie Cao, Ran He, and Tieniu Tan, "Vision transformer with super token sampling," in *CVPR*, 2023.

[15] Alex Zihao Zhu, Jieru Mei, Siyuan Qiao, Hang Yan, Yukun Zhu, Liang-Chieh Chen, and Henrik Kretzschmar, "Superpixel transformers for efficient semantic segmentation," in *IROS*, 2023.

[16] Jieru Mei, Liang-Chieh Chen, Alan Yuille, and Cihang Xie, "Spformer: Enhancing vision transformer with superpixel representation," 2024.

[17] André Ferreira, Naida Solak, Jianning Li, Philipp Dammann, Jens Kleesiek, Victor Alves, and Jan Egger, "How we won brats 2023 adult glioma challenge? just faking it! enhanced synthetic data augmentation and model ensemble for brain tumour segmentation," *arXiv preprint arXiv:2402.17317*, 2024.

[18] Fadillah Maani, Anees Ur Rehman Hashmi, Mariam Aljuboory, Numan Saeed, Ikboljon Sobirov, and Mohammad Yaqub, "Advanced tumor segmentation in medical imaging: An ensemble approach for brats 2023 adult glioma and pediatric tumor tasks," *arXiv preprint arXiv:2403.09262*, 2024.