

<Conference abbreviation>

<Conference Series name>

<Volume number and Year> <DOI Number>

Explicit Attention with CNNs vs Implicit Attention with Transformers: Insights from Brain Tumor Segmentation

S. Annie Priyadarshini; Rudru Gowtham; Premkumar Borugadda; Dulla Srinivas; K. Martina Rani; Kodali Sai Praveen

S. Annie Priyadarshini and K. Martina Rani: Karunya School of Management, Karunya Institute of Technology and Sciences, Karunya Nagar, Coimbatore - 641114, India; anniepriya@karunya.edu; martinarani@karunya.edu

Rudru Gowtham, Dulla Srinivas, and Kodali Sai Praveen: Department of Computer Science and Engineering, Vignan's Foundation for Science, Technology and Research, Guntur, Andhra Pradesh - 522213, India; rudrugowtham9@gmail.com; dsrinivas2907@gmail.com; kodalisaipraveen7@gmail.com

Premkumar Borugadda: Department of CSE (AI & ML), Bapatla Engineering College, Bapatla, Andhra Pradesh - 522101, India; premkumar.borugadda@becbapatla.ac.in

Abstract. Accurate brain tumor segmentation from magnetic resonance imaging (MRI) is important for diagnosis, treatment planning, and longitudinal assessment. This study presents a controlled comparison of U-Net, Attention U-Net (AUNET), and UNETR to examine explicit attention in convolutional networks against implicit self-attention in transformers. All models use the same Figshare brain tumor dataset, preprocessing pipeline, data split, loss formulation, and evaluation protocol. Performance is assessed using Dice Similarity Coefficient (DSC), 95th-percentile Hausdorff Distance (HD95), precision, sensitivity, and specificity for whole tumor (WT), tumor core (TC), and enhancing tumor (ET). On the test set, UNet achieved the strongest WT performance with Dice 0.8336 and HD95 19.30, followed closely by AUNET with Dice 0.8271 and HD95 20.06. UNETR obtained Dice 0.5962 and HD95 35.99 under the same conditions. TC and ET results show the same trend. The findings indicate that convolutional inductive bias remains effective for fine spatial localization when training data are limited, whereas transformer models may require larger and more diverse datasets. The comparison motivates future hybrid CNNtransformer designs that combine local boundary sensitivity with global contextual modeling.

Keywords. medical image analysis, brain tumor segmentation, deep learning, U-Net, Attention U-Net, UNETR, convolutional neural networks, transformers

1. INTRODUCTION

Brain tumors range from benign lesions to aggressive malignancies and show substantial variation in shape, location, and boundary appearance [1, 2]. MRI is central to diagnosis because of its soft-tissue contrast, but manual segmentation is time-consuming, observer-dependent, and sensitive to tumor heterogeneity and imaging artifacts [3–7].

Deep learning has improved automated segmentation. U-Net uses an encoder-decoder structure with skip connections for local detail [10]; Attention U-Net adds explicit gates that emphasize relevant encoder features [11]. UNETR instead uses transformer self-attention for global interactions and a convolutional decoder for localization [8, 14].

All Open Access articles are peer-reviewed, distributed under the Creative Commons Attribution-Non Commercial 4.0 International

CNNs are strong local feature extractors but have limited global receptive fields, whereas transformers model long-range relationships but are more data- and memory-intensive. This study therefore compares U-Net, Attention U-Net, and UNETR under the same data, preprocessing, training, and metrics, with emphasis on WT, TC, ET and explicit-versus-implicit attention.

2. RELATED WORK

Classical segmentation used thresholding, region growing, watershed methods, and engineered intensity, texture, or shape features. Although computationally simple, these methods were sensitive to low contrast, noise, and heterogeneous tumors and often required post-processing [5, 7].

CNNs replaced handcrafted features with learned hierarchical representations. UNet became a standard baseline through its encoder-decoder and skip connections [10]. Attention U-Net further filters skip features using decoder-guided attention, helping suppress irrelevant anatomy and emphasize small targets [11].

Transformers use self-attention to capture long-range token relationships [14]. UNETR applies a transformer encoder with convolutional decoding [8], while TransUNet follows a related hybrid strategy [9]. Their global modeling is attractive but can demand more data and memory. The research gap addressed here is a controlled comparison of explicit CNN attention and implicit transformer selfattention under identical conditions [7, 12].

3. METHODOLOGY

3.1. Dataset and preprocessing

The Figshare dataset [13] contains 3,064 contrast-enhanced T1-weighted MRI images from glioma, meningioma, and pituitary cases. A 70/15/15 split gives 2,144 training, 460 validation, and 460 test images. Images are resized to 256×256 , normalized to $[0,1]$, and masks are mapped consistently to WT, TC, and ET; imagemask alignment is verified before loading through a custom PyTorch dataset class.

3.2. Compared architectures

U-Net is the convolutional baseline: the encoder extracts hierarchical features, the decoder restores resolution, and skip connections preserve localization. Attention U-Net inserts gates into these skip pathways; encoder features are weighted by decoder-guided attention coefficients before fusion, suppressing irrelevant responses and emphasizing tumor structures [11].

UNETR divides $X \in \mathbb{R}^{(H \times W \times D \times C)}$ into non-overlapping $P_H \times P_W \times P_D$ patches, giving $N = (H/P_H)(W/P_W)(D/P_D)$ tokens. Embedded patches with positional information pass through transformer layers using multi-head self-attention [8, 14]. Intermediate features are reshaped and fused by a convolutional decoder to recover the segmentation map.

3.3. Training and evaluation

Training uses Dice + cross-entropy loss [15], Adam [16] at 1×10^{-4} , ReduceLROnPlateau, 50 epochs, and batch size 8. PyTorch [19] experiments run on an NVIDIA Tesla V100 (32 GB). DSC, HD95, precision, sensitivity, and specificity are evaluated for WT, TC, and ET [17, 18].

4. EXPERIMENTAL RESULTS

4.1. Quantitative findings

On the test set, U-Net achieves Dice-WT 0.8336 and TC/ET 0.8812/0.8695; AUNET obtains 0.8271 and 0.8745/0.8629; UNETR records 0.5962 and 0.6457/0.6321. Training/validation results showed the same CNN advantage, with a larger generalization gap for UNETR.

4.2. Comprehensive test analysis

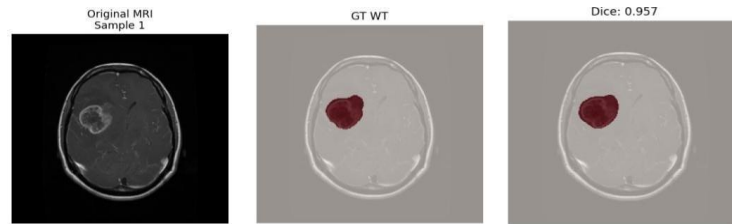
Table 1 shows the full test comparison. U-Net has the best Dice-WT (0.8336) and lowest HD95 (19.30); AUNET is close (0.8271, 20.06). UNETR has lower Dice (0.5962), higher HD95 (35.99), precision 0.5931, and sensitivity 0.5276. High specificity for all models partly reflects background dominance.

Table 1. Test-set whole tumor segmentation performance

Model	Loss	Dice-WT	HD95-WT	Precision	Sensitivity	Specificity
U-Net	0.0063	0.8336	19.30	0.8393	0.8090	0.9981
AUNET	0.0070	0.8271	20.06	0.8359	0.7861	0.9980
UNETR	0.0118	0.5962	35.99	0.5931	0.5276	0.9969

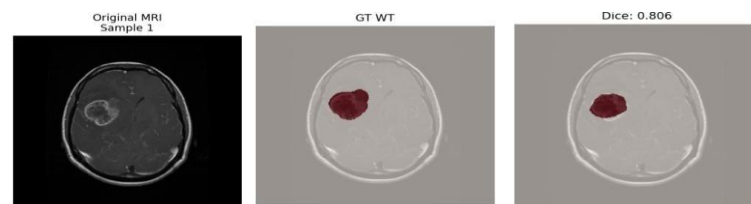
4.3. Qualitative assessment

Representative qualitative predictions yielded Dice 0.957 for U-Net, 0.909 for AUNET, and 0.806 for UNETR. U-Net most closely follows ground-truth boundaries; AUNET preserves fine structures through gated feature selection; UNETR shows more under-segmentation and boundary inconsistency. Thus, qualitative evidence supports the test metrics.

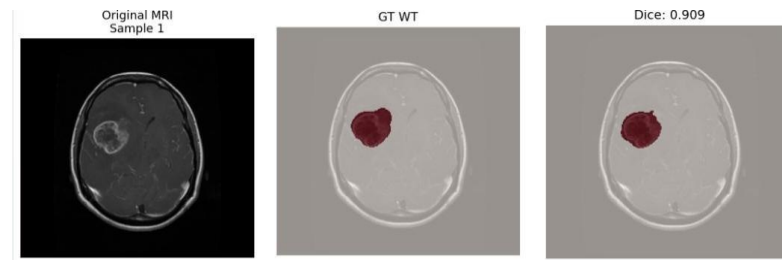


(a) U-Net prediction (Dice = 0.957)

4



(b) UNETR prediction (Dice = 0.806)



(c) AUNET prediction (Dice = 0.909)

Figure 4. Qualitative segmentation results showing original MRI, ground truth, and predicted masks

5. DISCUSSION AND FUTURE DIRECTIONS

U-Net provides the strongest overall segmentation, while AUNET remains close and confirms the value of explicit gating. UNETR is weaker in Dice, sensitivity, and boundary distance. This does not imply that transformers are generally inferior; rather, self-attention appears more sensitive to data volume and training conditions, whereas CNN inductive bias remains effective for local tumor boundaries.

Future work should use larger multi-institutional, multi-modal cohorts, stronger pretraining/augmentation, repeated-split statistics, and external validation. Hybrid CNN-transformer models may combine local boundary sensitivity with global context; uncertainty, calibration, interpretability, robustness, and complete multiclass sub-region evaluation should also be studied.

6. REFERENCES

- [1] D. N. Louis, A. Perry, G. Reifenberger, A. von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues, and D. W. Ellison, "The 2021 WHO classification of tumors of the central nervous system: a summary," *Acta Neuropathologica*, vol. 141, no. 3, pp. 495–508, 2021.
- [2] K. Aldape, K. M. Brindle, L. Chesler, et al., "Challenges to curing primary brain tumours," *Nature Reviews Clinical Oncology*, vol. 16, no. 8, pp. 509–520, 2019.
- [3] F. Bray, J. Ferlay, I. Soerjomataram, et al., "Global cancer statistics 2022," *CA: A Cancer Journal for Clinicians*, vol. 73, no. 6, 2023.
- [4] American Cancer Society, *Cancer Facts & Figures 2025*. Atlanta, GA, USA: American Cancer Society, 2025.
- [5] S. Bauer, R. Wiest, L.-P. Nolte, and M. Reyes, "A survey of MRI-based medical image analysis for brain tumor studies," *Physics in Medicine & Biology*, vol. 58, no. 13, pp. R97–R129, 2013.
- [6] N. Gordillo, E. Montseny, and P. Sobrevilla, "State of the art in MRI brain tumor segmentation," *Magnetic Resonance Imaging*, vol. 31, no. 8, pp. 1426–1438, 2013.
- [7] B. H. Menze, A. Jakab, S. Bauer, et al., "The multimodal brain tumor image segmentation benchmark (BraTS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

- [8] A. Hatamizadeh, Y. Tang, V. Nath, et al., "UNETR: Transformers for 3D medical image segmentation," in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022.
- [9] J. Chen, Y. Lu, Q. Yu, et al., "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.
- [10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and ComputerAssisted Intervention (MICCAI), 2015.
- [11] O. Oktay, J. Schlemper, L. Le Folgoc, et al., "Attention U-Net: Learning where to look for the pancreas," arXiv preprint arXiv:1804.03999, 2018.
- [12] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," Medical Image Analysis, vol. 36, pp. 61–78, 2017.
- [13] J. Cheng, "Brain tumor dataset," Figshare, 2017. [Online]. Available: https://figshare.com/articles/dataset/brain_tumor_dataset/1512427.
- [14] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017.
- [15] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in Proc. Fourth International Conference on 3D Vision (3DV), 2016.
- [16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
- [17] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," BMC Medical Imaging, vol. 15, article 29, 2015.
- [18] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 9, pp. 850–863, 1993.
- [19] A. Paszke, S. Gross, F. Massa, et al., "PyTorch: An imperative style, highperformance deep learning library," in Advances in Neural Information Processing Systems (NeurIPS), 2019.