Attention-Guided Vector Quantized Variational Autoencoder for Brain Tumor Segmentation

Danish Ali[1], Ajmal Mian[1], Naveed Akhtar[1,2], and Ghulam Mubashar Hassan[1]

[1] Computer Science and Software Engineering, The University of Western Australia
danish.ali@research.uwa.edu.au
[2] School of Computing and Information Systems, The University of Melbourne

Abstract. Precise brain tumor segmentation is critical for effective treatment planning and radiotherapy. Existing methods rely on voxel-level supervision and often struggle to accurately delineate tumor boundaries, increasing potential surgical risks. We propose an Attention-Guided Vector Quantized Variational Autoencoder (AG-VQ-VAE) — a two-stage network specifically designed for boundary-focused tumor segmentation. Stage 1 comprises a VQ-VAE which learns a compact, discrete latent representation of segmentation masks. In stage 2, a conditional network extracts contextual features from MRI scans and aligns them with discrete mask embeddings to facilitate precise structural correspondence and improved segmentation fidelity. Additionally, we propose an attention scaling module to reinforce discriminative feature learning and a soft masking module to refine attention in uncertain tumor regions. Comprehensive evaluations on BraTS 2021 demonstrate that our AG-VQ-VAE sets a new benchmark, improving the HD95 metric by 4.83 mm (Whole Tumor), 2.14 mm (Tumor Core), and 2.39 mm (Enhancing Tumor), compared to state-of-the-art methods, while achieving a 0.23% improvement in Dice score for whole tumor. Furthermore, our qualitative results and ablation study demonstrate that feature-level supervision significantly enhances boundary delineation compared to voxel-level approaches. The code is available at https://github.com/danishali6421/AG-VQVAE-MICCAI.

Keywords: Tumor boundary · Vector quantization · Self attention.

1 Introduction

Brain tumor, one of the most common and life-threatening cancers, is caused by the abnormal growth of glial cells. Approximately 300,000 cases are reported annually, with 75% being malignant [24]. Magnetic resonance imaging (MRI) is frequently used to capture the detailed structure of brain tissues. Four major MRI modalities / sequences (T1, $T1_{CE}$, T2, and FLAIR) are used to capture heterogeneous and complementary information from the brain tissues. Expert radiologists segment tumor regions by carefully analyzing each MRI slice across these modalities. However, manual segmentation is a labor-intensive process due to the substantial heterogeneity in tumor morphology, including size, shape, and spatial location within the brain. Furthermore, low contrast between the healthy

and tumorous regions in 3D multimodal MRI scans pose additional challenges in precise boundary delineation.

Advancement in deep learning frameworks has driven researchers to explore their potential for automatic tumor segmentation. Preliminary research [14, 18] adopted a 2D U-Net architecture where each MRI volume was split into 2D slices, and each slice was segmented independently. Several CNN-based 2D U-Net models such as Znet [19], LeU-Net [21], and BU-Net [23], have demonstrated promising results. While 2D U-Net models made considerable progress in advancing brain tumor segmentation, their slice-wise processing of MRI data inherently disrupts spatial continuity across adjacent slices, leading to potential performance degradation [9]. To avoid this, various techniques such as 3DFCNN [1], ERV-Net [34], VAT [20], and MBANet [5] have been proposed that exploit full 3D MRI volumes. In addition, Zhang et al. [32] introduced a hierarchical multi-scale network (HMNET) to capture the tumor structures at various scales. Similarly, in the multi-scale residual U-Net (mResU-Net) [15], dilated convolutions of varying scales are employed to expand the receptive field, leading to improved segmentation accuracy for targets of varying sizes. Although these methods have excellent representation ability, they struggle to establish explicit long-range dependencies due to the limited local receptive fields of convolution kernels.

Recently, attention-based models, particularly Transformers [27], have gained significant recognition for their effectiveness in modeling global contextual relationships. Numerous Transformer models [4, 7, 26] have been proposed for medical image segmentation, achieving favorable results. The TransBTS [28] marks the pioneering attempt to integrate Transformer model into 3D CNN for brain tumor segmentation. This architecture extends the encoder-decoder framework by incorporating a Transformer in the bottleneck layer, leading to significantly improved segmentation. In contrast to TransBTS, UNETR [12] used a Transformer encoder to derive sequence representations of the input volume, effectively capturing global multiscale information. Jia et al. proposed a refined version of TransBTS called BiTr-Unet [13] which employs two sets of ViT layers to extract global information at two different scales. This solution surpasses the bottleneck approach (TransBTS) and offers computational efficiency compared to UNETR.

While transformer models have shown remarkable success in brain tumor segmentation, their attention mechanism involves computationally expensive operations with quadratic cost. To enchance computational efficiency, several window-based transformer models [3, 8, 11, 35] have been proposed, with attention computed within localized windows. Dual-branch vision transformer (DB-Trans) [31] models relationship between non-adjacent windows, significantly improving the segmentation accuracy. Despite considerable progress made by these methods, the challenge of accurately delineating tumor boundaries remains unsolved. Most existing methods emphasize maximizing the overlap between predicted and ground truth masks using voxel-level supervision, often relying on low-level spatial features passed through skip connections. However, they struggle with precise tumor boundary delineation [29], which is crucial for effective

treatment planning and tumor resection while preserving surrounding healthy tissues [25].

To address the above limitations, we propose a novel two-stage brain tumor segmentation network that not only enhances tumor boundary delineation but also eliminates the reliance on low-level spatial features for segmentation tasks, thereby improving computational efficiency. This two-stage design is loosely inspired by recent advances in generative modeling, particularly DALL-E [22] and VQ-Diffusion [10], which leverage discrete latent representations for more structured generation. The first stage of our proposed model employs a Vector Quantized Variational Autoencoder (VQ-VAE) trained on segmentation masks to learn discrete latent representation, effectively capturing fine-grained boundary details. The second stage introduces a cross-latent alignment network, integrating convolutional layers with a Transformer-based bottleneck to effectively map the continuous latent representations of MRI modalities to the discrete representation of the segmentation mask. Our contributions are summarized below:

1. We propose a two-stage Attention-Guided Vector Quantized Variational Autoencoder (AG-VQ-VAE) to focus on structural coherence and accurate boundary delineation across all tumor regions using feature-level supervision, without relying on low-level spatial features from skip connections.
2. We propose a novel attention scaling mechanism that dynamically modulates the contribution of each attention head during training, prioritizing those that are most relevant for feature learning, resulting in improved generalization.
3. We propose a soft masking module to adaptively assign weights to spatial features based on their uncertainty following the attention calculation. This enforces the model to focus more on learning features associated with tumor boundaries that are inherently uncertain [17].

Extensive evaluation on the BraTS 2021 dataset demonstrates that our method significantly improves tumor boundary delineation compared to existing state-of-the-art models [16, 31]. Our method improves the HD95 metric by 4.83mm, 2.14mm and 2.39mm for whole, core and enchancing tumors, respectively.

2 Methodology

We develop a two stage AG-VQ-VAE model for brain tumor segmentation, see Fig. 1. The first stage employs a VQ-VAE to learn a discrete latent representation of the ground truth segmentation masks. In the second stage, MRI conditional network processes the four stacked MRI modalities, mapping their continuous high-dimensional latent features to the discrete indices of the pre-trained VQ-VAE codebook. Constituent components of our model are explained below.

2.1 Vector Quantized Variational Autoencoder

We cast brain tumor segmentation as a MRI-conditioned mask generation problem. To achieve this, three separate VQ-VAE networks are trained, each dedicated to a specific tumor sub-region: Whole Tumor (WT), Tumor Core (TC), and

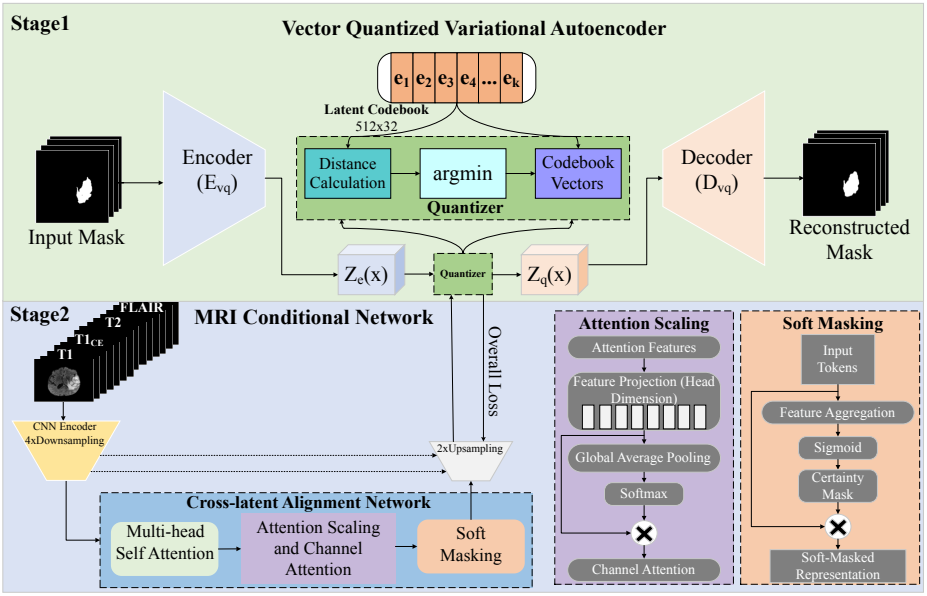

Fig. 1. The overall architecture of the proposed AG-VQ-VAE method. Stage 1 consists of VQ-VAE, while Stage 2 incorporates an attention-based MRI conditional network.

Enhancing Tumor (ET), using their respective 3D ground truth segmentation masks. Figure 1 shows that Stage 1 comprises a hierarchical VQ-VAE network, where the encoder (E_{vq}) processes the ground truth input masks and maps them to continuous structural features ($Z_e(x)$) that capture essential details of the tumor shape, size, and boundaries. These continuous features ($Z_e(x)$) are then passed through the vector quantization layer, which maps them to a finite set of discrete latent features. The vector quantizer consists of $K = 512$ embedding vectors, each of dimensionality $D = 32$ in its codebook E. The quantizer assigns each continuous feature to the nearest embedding vector in the codebook based on Euclidean distance, defined as:

$$Z_q(x) = \arg\min_{e_k \in E} \|Z_e(x) - e_k\|_2 , \tag{1}$$

where $Z_e(x)$ and $Z_q(x)$ are the latent feature maps before and after quantization respectively, and e_k denotes the k-th embedding vector in the codebook E. Finally, the decoder (D_{vq}) reconstructs the mask from these quantized latent features. Dice loss $\mathcal{L}_{\text{Dice}}$ is employed as the reconstruction loss for VQ-VAE training. Straight-through estimator is applied to allow gradients to propagate through the vector quantization step. Codebook vectors are updated using the exponential moving average of the encoder output. Additionally, a commitment loss

$$\mathcal{L}_{\text{ct}} = \|Z_e(x) - s_g[Z_q(x)]\|_2^2 , \tag{2}$$

where s_g denotes the stop-gradient operator, is used to align the encoder's output with the codebook embeddings, to ensure stable training. The overall loss is:

$$\mathcal{L}_{\text{overall}} = \alpha_d \mathcal{L}_{\text{Dice}} + \alpha_c \mathcal{L}_{\text{ct}}, \tag{3}$$

where α_d and α_c are weights to balance Dice loss $\mathcal{L}_{\text{Dice}}$ and commitment loss $\mathcal{L}_{\text{ct}}$, respectively. By learning a discrete latent space, VQ-VAE ensures a structured representation of segmentation masks, where the quantized latent embeddings effectively capture essential tumor characteristics. The organization of similar tumor structural patterns into distinct codebook embeddings allows for a robust encoding of tumor morphology, which can help in precise boundary delineation during reconstruction. Moreover, the adopted discrete encoding discourages the model from memorizing specific instances, thereby enhancing its ability to handle diverse anatomical variations.

2.2 MRI Conditional Network

The MRI conditional network (Fig. 1 stage 2) comprises a CNN encoder, cross-latent alignment network and upsampling block. The input MRI of shape (4, H, W, D), where four modalities are stacked along the channel dimension, is passed to a CNN encoder which employs four convolutional layers to progressively extract local features while downsampling the input to generate a lower-dimensional embedding $\mathbf{e}_i \in \mathbb{R}^{C_i \times H_i \times W_i \times D_i}$ in the i_{th} encoder layer ($i \in [1, 4]$).

This image embedding is then fed into the cross-latent alignment network, where it is processed by a series of transformer layers. We utilize eight transformer layers, each employing multi-head self attention to capture global contextual features and long-range dependencies within the brain's anatomical structures and tumor morphology. The input embeddings, representing both healthy brain tissues and tumorous regions, are tokenized and projected into queries (Q), keys (K), and values (V) through learnable linear transformations, where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N \times d_k}$, with N being the number of tokens in the MRI embedding and d_k denoting the dimension of each key and query vector. The query, key, and value matrices are split into eight attention heads to effectively model the structural variations in the brain, including the boundaries between healthy tissues and tumors. Each attention head independently computes its own set of attention scores using the scaled dot-product attention mechanism:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}. \tag{4}$$

Following attention computation across all heads, the resulting features are concatenated and passed through residual connections and layer normalization, to ensure stable gradient propagation. A multi-layer perceptron (MLP) with two fully connected layers and GELU2 activation is then applied to further refine the learned representations. The refined tokenized embeddings undergo attention scaling and soft masking before being transformed into spatial embeddings. Finally, these embeddings are upsampled using two convolution layers to match

their feature dimensions with those of the pre-trained VQ-VAE codebook vectors, with each feature vector assigned to the closest codebook entry. To align the MRI latent embeddings with the corresponding segmentation mask, two loss functions are employed: A cross-entropy loss that ensures each MRI embedding vector is assigned to the codebook vector index that corresponds to its ground truth (GT) mask feature vector, and a focal loss to prioritize the ambiguous features. The overall loss for training the MRI conditional network is

$$\mathcal{L}_{\text{overall}} = \left[(-\sum_{i} y_i \log(\hat{y}_i)) + (-\alpha(1-\hat{y}_i)^{\gamma} \log(\hat{y}_i)) \right], \tag{5}$$

where, y_i is the target codebook index, $\hat{y}_i$ is the predicted probability, α is a balancing factor, and γ controls the focus on hard examples.

Attention Scaling Module: The multi-head attention mechanism captures diverse features across multiple attention heads, but not all heads contribute equally to the task of discriminative feature extraction. To optimize the learning process, we introduce an attention scaling module that dynamically modulates the contribution of each head based on its relevance in extracting discriminative features. The module works by first projecting the attention features in the head dimension, followed by global average pooling (GAP) within each head to derive a compact descriptor that captures its overall significance. A softmax function is applied across the attention heads to compute a set of normalized weights, allowing the model to prioritize more relevant heads while suppressing less informative ones. The attention scaling process can be formulated as:

$$\begin{aligned} Y &= \text{Proj}(X) \quad \text{where} \quad Y \in \mathbb{R}^{B \times L \times N_{\text{head}} \times D_{\text{head}}}, \\ S &= \text{Softmax}\left(\text{GAP}_{D_{\text{head}}}(Y)\right), \\ X_{\text{scaled}} &= X_{resh} \odot S \quad \text{where} \quad X_{resh} \in R^{B \times L \times N_{\text{head}} \times D_{\text{head}}}, \end{aligned} \tag{6}$$

where "$\odot$" represents element wise rescaling and X_{resh} denotes the reshaped attention features in the head dimension. After attention scaling, the model incorporates squeeze and excitation to refine channel-wise attention, enhancing feature representation and improving model generalization by emphasizing key activations while attenuating redundant information.

Soft Masking: The Soft Masking Module, introduced after the first five transformer layers, aggregates the attention features into a single channel using a learnable transformation, followed by a sigmoid to compute a certainty map. This map generates a mask to prioritize uncertain features in the attention computation of subsequent transformer layers. This results in a more precise and reliable representation of complex tumor structures.

3 Experiments

Setup: We evaluate our method on the BraTS 2021 dataset [2], comprising MRI data of four modalities (T_1, T_2, $T1_{CE}$, FLAIR) from 1,251 patients. Following

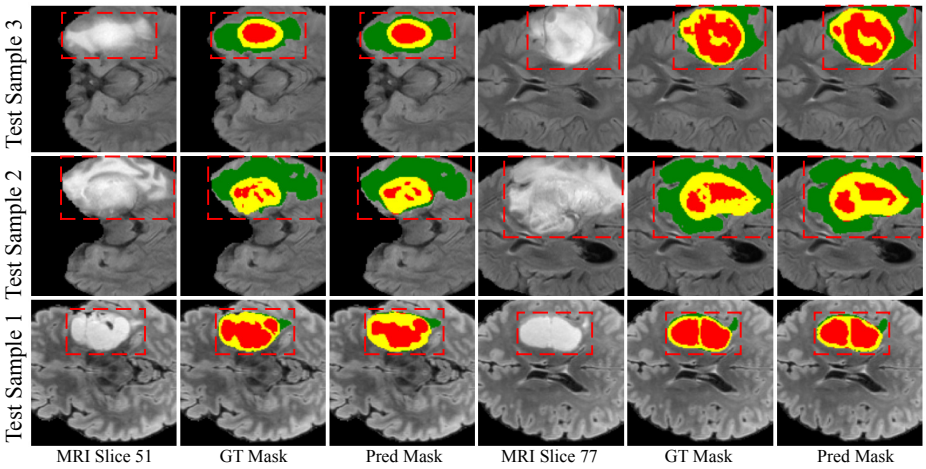

Fig. 2. Qualitative results of our method on three test samples. Red box represents the tumor region, with green, yellow, and red colors representing peritumoral edema (ED), enhancing tumor (ET), and necrotic tumor (NCR) respectively. Test Sample 1 has Avg HD95 (1.24mm) and Avg Dice (91%), Sample 2 has Avg HD95 (1.41mm) and Avg Dice (94%), and Sample 3 has Avg HD95 (1.27mm) and Avg Dice (95%).

recent studies [31, 16], we split the dataset into 834, 208, and 209 for training, validation and test sets respectively. Ground truth is annotated by radiologists into four regions: background, peritumoral edema (ED), necrosis (NCR), and enhancing tumor (ET). Segmentation results are evaluated across three regions: whole tumor (WT), Tumor Core (TC), and enhancing tumor (ET) using Hausdorff distance (HD95) and Dice similarity score. Our model is implemented in PyTorch and trained on each MRI conditional network for 150 epochs on NVIDIA RTX 3090 (24GB) with a batch size of 3. We use GELU2 activation, the Adam optimizer with learning rate of $1e^{-4}$, and a weight decay of $1e^{-4}$. We apply data augmentations, including random flipping across three planes, rotation, intensity shifting, and scaling, with probabilities of 0.5, 0.5, 0.1, and 0.1, respectively.

Results and Analysis: Table 1 compares the performance of our model with state-of-the-art methods including 3D U-Net [6], TransBTS [28], UNETR [12], DBTrans [31], NestedFormer [30], and Cascaded Causal Intervention [16] on the HD95 and Dice score metrics for three tumor regions: WT, TC, and ET.

Compared to nearest competitor DBTrans [31], our boundary-focused feature-level supervision approach improves HD95 by 4.83mm (WT), 2.14mm (TC), and 2.39mm (ET), and the Dice score of WT by 0.23%. Furthermore, our proposed AG-VQ-VAE architecture, comprising three dedicated networks (one per tumor region), results in a total of 441G FLOPs and 40M parameters. In contrast, DBTrans [31] require 146G FLOPs and 25M parameters. Although our model requires approximately 780ms to process a full 3D MRI volume (240×240×155) on NVIDIA RTX 3090 (24GB) compared to 254ms for DBTrans [31], it still produces segmentation results in less than one second, with memory usage be-

Table 1. Comparison on BraTS 2021 dataset across three tumor regions: Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET). Best results are in bold.

Methods	HD95 (mm) ↓				Dice Score (%) ↑			
	WT	TC	ET	AVG	WT	TC	ET	AVG
3D U-Net [6]	11.49	6.18	6.15	7.94	89.59	86.28	83.39	86.42
TransBTS [28]	15.12	8.21	7.83	10.38	89.25	85.35	80.35	84.99
UNETR [12]	15.99	10.01	9.72	11.90	90.10	83.66	79.78	84.51
NestedFormer [30]	10.23	6.43	6.08	7.58	90.12	88.18	85.62	87.97
DBTrans [31]	9.84	6.24	6.13	7.40	92.41	90.26	86.70	89.79
Causal Intervention [16]	13.92	5.85	6.43	8.73	92.32	**91.19**	**87.21**	**90.24**
Ours	**5.01**	**4.10**	**3.74**	**4.28**	**92.64**	89.05	82.25	87.98

low 8GB, and achieves superior HD95 scores, highlighting a favorable trade-off between accuracy and computational cost.

Note that accurate tumor boundary delineation is crucial for surgical decision-making, as minor inconsistencies in inner voxel predictions of tumor regions are more interpretable for clinicians than errors in boundary predictions [33]. Qualitative results in Fig. 2 show that, despite some inconsistencies in the prediction of necrotic regions (red segmentation mask), the boundaries of tumor sub-regions ($WT = ED + ET + NCR$, $TC = ET + NCR$, and ET) remain accurately delineated. This makes boundary-focused approaches, such as our AG-VQ-VAE, more clinically relevant than the models that prioritize inner voxel overlap but struggle with boundary precision.

Ablation Study: We first evaluate the segmentation performance of the single-stage AG-UNet with voxel-level supervision, incorporating both Attention Scaling (AS) and Soft Masking (SM), as shown in Table 2. While this model achieves better Dice scores for WT (93.19%) and ET (85.09%), its higher HD95 values indicate less precise boundary delineation. In contrast, our two-stage AG-VQ-VAE achieves significantly lower HD95 scores, demonstrating superior tumor boundary delineation, even when using only one of the two modules. The best HD95 scores are achieved when using both AS and SM and excluding either one leads to performance degradation, highlighting the importance of both modules for accurate segmentation. These results validate that our two-stage AG-VQ-VAE offers superior tumor boundary delineation compared to single-stage AG-UNet.

Table 2. Ablation study on Attention Scaling (AS) and Soft Masking (SM) to evaluate the performance of the single-stage AG-UNet and two-stage AG-VQ-VAE.

Models	AS	SM	HD95 (mm) ↓			Dice Score (%) ↑		
			WT	TC	ET	WT	TC	ET
AG-UNet (Single-stage)	✓	✓	6.15	5.39	4.37	**93.19**	88.26	**85.09**
AG-VQ-VAE (Two-stage)	✓	✓	**5.01**	**4.10**	**3.74**	92.64	**89.05**	82.25
AG-VQ-VAE (Two-stage)	✓	×	5.87	4.64	4.20	91.35	88.51	80.27
AG-VQ-VAE (Two-stage)	×	✓	5.57	4.99	4.43	91.74	87.88	80.41

4 Conclusion

We proposed a novel two-stage network for brain tumor segmentation with a strong emphasis on precise boundary delineation. By integrating attention scaling and soft masking, our approach enhances feature representation while effectively handling uncertainty in tumor regions. Evaluations on BraTS 2021 demonstrate the superiority of our method. We showed that our feature-level supervision enables superior tumor boundary delineation compared to voxel-level supervision commonly used in other methods. These advancements have significant clinical implications for surgical planning.

Acknowledgments. This research was supported by the Australian Government. Professor Ajmal Mian is the recipient of an ARC Future Fellowship Award (project #FT210100268), funded by the Australian Government. Dr. Naveed Akhtar is a recipient of the ARC Discovery Early Career Researcher Award (project #DE230101058), funded by the Australian Government.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Anand, V.K., Grampurohit, S., Aurangabadkar, P., Kori, A., Khened, M., Bhat, R.S., Krishnamurthi, G.: Brain tumor segmentation and survival prediction using automatic hard mining in 3d cnn architecture. In: International MICCAI Brainles Workshop. pp. 310–319. Springer (2021)
2. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
3. Cai, Y., Long, Y., Han, Z., Liu, M., Zheng, Y., Yang, W., Chen, L.: Swin unet3d: a 3-dimensional medical image segmentation network combining vision transformer and convolution. BMC medical informatics and decision making **23**(1) (2023)
4. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: ECCV. pp. 205–218. Springer (2022)
5. Cao, Y., Zhou, W., Zang, M., An, D., Feng, Y., Yu, B.: Mbanet: A 3d convolutional neural network with multi-branch attention for brain tumor segmentation from mri images. Biomedical Signal Processing and Control **80**, 104296 (2023)
6. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: MICCAI: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. pp. 424–432. Springer (2016)
7. Gao, Y., Zhou, M., Metaxas, D.N.: Utnet: a hybrid transformer architecture for medical image segmentation. In: MICCAI: 24th International Conference, 2021, Proceedings, Part III 24. pp. 61–71. Springer (2021)
8. Ghazouani, F., Vera, P., Ruan, S.: Efficient brain tumor segmentation using swin transformer and enhanced local self-attention. International Journal of Computer Assisted Radiology and Surgery **19**(2), 273–281 (2024)

9. Gholami, A., Subramanian, S., Shenoy, V., Himthani, N., Yue, X., Zhao, S., Jin, P., Biros, G., Keutzer, K.: A novel domain adaptation framework for medical image segmentation. In: International MICCAI Brainles Workshop. pp. 289–298. Springer (2019)
10. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: CVPR. pp. 10696–10706 (2022)
11. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
12. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF WACV. pp. 574–584 (2022)
13. Jia, Q., Shu, H.: Bitr-unet: a cnn-transformer combined network for mri brain tumor segmentation. In: International MICCAI Brainlesion Workshop. Springer (2021)
14. Kermi, A., Mahmoudi, I., Khadir, M.T.: Deep convolutional neural networks using u-net for automatic brain tumor segmentation in multimodal mri volumes. In: International MICCAI Brainles Workshop. pp. 37–48. Springer (2019)
15. Li, P., Li, Z., Wang, Z., Li, C., Wang, M.: mresu-net: multi-scale residual u-net-based brain tumor segmentation from multimodal mri. Medical & Biological Engineering & Computing **62**(3), 641–651 (2024)
16. Liu, H., Li, Q., Nie, W., Xu, Z., Liu, A.: Causal intervention for brain tumor segmentation. In: MICCAI. pp. 160–170. Springer (2024)
17. McKinley, R., Rebsamen, M., Dätwyler, K., Meier, R., Radojewski, P., Wiest, R.: Uncertainty-driven refinement of tumor-core segmentation using 3d-to-2d networks with label uncertainty. In: International MICCAI Brainles Workshop. pp. 401–411. Springer (2021)
18. Munir, K., Frezza, F., Rizzi, A.: Brain tumor segmentation using 2d-unet convolutional neural network. Deep Learning for Cancer Diagnosis pp. 239–248 (2021)
19. Ottom, M.A., Rahman, H.A., Dinov, I.D.: Znet: deep learning approach for 2d mri brain tumor segmentation. IEEE Translational Engg. in Health & Medicine (2022)
20. Peiris, H., Chen, Z., Egan, G., Harandi, M.: Reciprocal adversarial learning for brain tumor segmentation: a solution to brats challenge 2021 segmentation task. In: International MICCAI Brainlesion Workshop. pp. 171–181. Springer (2021)
21. Rai, H.M., Chatterjee, K.: 2d mri image analysis and brain tumor detection using deep learning cnn model leu-net. Multimedia Tools & Applications **80**(28) (2021)
22. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML. pp. 8821–8831 (2021)
23. Rehman, M.U., Cho, S., Kim, J.H., Chong, K.T.: Bu-net: Brain tumor segmentation using modified u-net architecture. Electronics **9**(12), 2203 (2020)
24. Reynoso-Noverón, N., Mohar-Betancourt, A., Ortiz-Rafael, J.: Epidemiology of brain tumors. Principles of neuro-oncology: brain & skull base pp. 15–25 (2021)
25. Scoccianti, S., Detti, B., Gadda, D., Greto, D., Furfaro, I., Meacci, F., Simontacchi, G., Di Brina, L., Bonomo, P., Giacomelli, I., et al.: Organs at risk in the brain and their dose-constraints in adults and in children: a radiation oncologist's guide for delineation in everyday practice. Radiotherapy & Oncology **114**(2) (2015)
26. Tragakis, A., Kaul, C., Murray-Smith, R., Husmeier, D.: The fully convolutional transformer for medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3660–3669 (2023)

27. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
28. Wang, W., Chen, C., Ding, M., Li Jiangyun, Yu Hong, e.a.: Zha sen.(2021a). transbts: Multimodal brain tumor segmentation using transformer. MICCAI (2021)
29. Xiao, Q., Nie, D.: Blurry boundary segmentation with semantic-aware feature learning. In: MIUA. pp. 101–111. Springer (2024)
30. Xing, Z., Yu, L., Wan, L., Han, T., Zhu, L.: Nestedformer: Nested modality-aware transformer for brain tumor segmentation. In: MICCAI. Springer (2022)
31. Zeng, X., Zeng, P., Tang, C., Wang, P., Yan, B., Wang, Y.: Dbtrans: A dual-branch vision transformer for multi-modal brain tumor segmentation. In: MICCAI. pp. 502–512. Springer (2023)
32. Zhang, R., Jia, S., Adamu, M.J., Nie, W., Li, Q., Wu, T.: Hmnet: Hierarchical multi-scale brain tumor segmentation network. Clinical Medicine **12**(2) (2023)
33. Zhang, Z.Z., Shields, L.B., Sun, D.A., Zhang, Y.P., Hunt, M.A., Shields, C.B.: The art of intraoperative glioma identification. Frontiers in oncology **5**, 175 (2015)
34. Zhou, X., Li, X., Hu, K., Zhang, Y., Chen, Z., Gao, X.: Erv-net: An efficient 3d residual neural network for brain tumor segmentation. Expert Systems with Applications **170**, 114566 (2021)
35. Zhu, Z., He, X., Qi, G., Li, Y., Cong, B., Liu, Y.: Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal mri. Information Fusion **91**, 376–387 (2023)