RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt26113

SLAM-FusionNet: A Deep Learning Framework Integrating Spatial Local Attention with Multi-Modal Fusion for Glioma Segmentation

Vikash Verma [a,*], Pritaj Yadav [a]

[a] Department of Computer Science & Engineering, Rabindranath Tagore University, Bhopal, India

* Corresponding Author Email: vikashverma2005@gmail.com

DOI: https://doi.org/10.54392/irjmt26113

Received: 15-06-2025; Revised: 06-12-2025; Accepted: 16-01-2026; Published: 29-01-2026

Abstract: The accurate segmentation of gliomas and subregions of multimodal magnetic resonance imaging (MRI) plays a significant role in the accurate diagnosis of the disease and formulation of therapy as well as disease surveillance. However, the non-uniformity of tumours where boundaries vary and modality variations are specific to the intensity subjects a person to a huge challenge. The work will present SLAM-FusionNet a transformer-based architecture, a Multi-Modal Fusion (MMF) strategy and a Spatial Local Attention Module (SLAM) in order to address effectively the local and global contextual information of tumor fine-grained information. MMF module can be incorporated to increase cross-modality representation learning, complementary characteristics of T1, T2, FLAIR and T1ce images are merged; to increase localization in a spatial dimension, SLAM increases the significance of the spatially-relevant and boundary-sensitive regions that can better differentiate intra-tumor subregions. The suggested network is premised on a Swin Transformer backbone, and the primary strength of long-range dependency description and local spatial fidelity. Considerable testing has been carried out on the BraTS dataset that demonstrates that SLAM-FusionNet is functional with 95.6% whole tumor (WT) 96.2% tumor core (TC) and 94.8% enhancing tumor (ET) Dice scores and an average Dice of 95.5. The average HD95 is also increased to 3.95 mm and improved compared to state of the art models such as Swin-UNet and nnU-Net. Additive value of MMF and SLAM is confirmed by formal studies of ablation. The results highlight the applicability and clinical power of SLAM-FusionNet in computer-aided brain tumor segmentation in precision neuro-oncology.

Keywords: Brain Tumor Segmentation, Glioma, Swin Transformer, Spatial Local Attention, Multi-Modal MRI, Deep Learning

1. Introduction

Gliomas are a dreadful and highly heterogeneous form of brain tumors which are associated with a low prognosis and death. To help in the diagnosis, surgical planning, radiotherapy and longitudinal monitoring of the disease, magnetic resonance imaging (MRI) needs to be defined properly on the subregions of the glioma. The other complementary imaging modalities that can be used with MRI are T1-weighted (T1) and contrast-enhanced T1 (T1ce), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR) and each of them provides a different aspect of the tissue. Consequently, it can be inferred that multi-modal inputs are perhaps, necessary in the effective glioma segmentation in order to characterize tumor phenotypes, tumor core, peritumoral edema, and enhancing regions comprehensively. Nevertheless, the division of such heterogeneous structures during the manual segmentation procedure is tedious, time-intensive, and might be influenced by a wide inter-observer variation that drives the development of automated and robust convolutional neural network (CNN) architectures to perform brain tumor segmentation [1]. Better designs such as nnU-Net [2], UNet++ [3], V-Net [4] and ResUNet models [5] can better hierarchical feature representations and semantic representation and achieve a high baseline performance in medical image segmentation. CNN based models are however essentially restricted on the extent of their receptive field as well as the ability to represent the contextual dependency of long ranges. It has been observed to have been a weakness particularly in segmentation of glioma whereby, the subregions of the tumor are defined by irregular shapes, blurry outlines, and uneven distribution of intensities. Besides these, unsophisticated multi-modal fusion of multi-modal MRI features may prefer to have poor performance due to inappropriate modality and inadequate exploitation of complementary information. The attention mechanisms have been put forward to overcome these limitations by extracting informative features selectively both spatially

and channel-wisely. Various channel- and spatial-recalibration modules (CBAM [5]) and scSE [6]) and networks with attention have performed better localization of tumor-relevant regions. Despite the fact that these methods are going to make local discriminability higher, it still has a weakness in modeling global contextual relationship that is critical in analyzing volumetric MRI.

Nonetheless, transformer-based architectures redefined the medical image analysis only recently when they enabled the modeling of the long-range dependencies to be conducted. The initial proposal of this paradigm was made by Vision Transformer (ViT) [7] which models images as sequences of tokens but because of their high computational complexity and large data requirements, could not be applied directly in the context of medical imaging. The Swin Transformer [8] solved these problems and requires hierarchical window-based self-attention, sacrificing both modeled global context and computational efficiency. Taking this framework, hybrid architectures, such as, TransUNet [9], the UNETR [10] and Swin-Unet [11], have shown state-of-the-art performance in the volumetric medical image segmentation. Recently, transformer-based models with task-adapted adaptation and multimodal learning policy, such as MedFormer [12], have demonstrated improved generalization to a larger set of data. At the same time, multi-modal fusion models have achieved improvements in utilizing complementary data of the MRI. The early and late fusion techniques were low in benefits due to their inability to provide all the inter-modality dependencies. Newer attention based and cross-modality fusion mechanisms [13, 14] have the potential of permitting selective combination of informative modality-oriented features. These methods have proven to be more robust in the complementary information in T1, T1ce, T2, as well as FLAIR modalities but there are some still challenges in matching a multi-resolution representation and fine-grained structural detail particularly at tumor sites in order to achieve clinically reliable brain tumor segmentation. Automation system is generally less effective in low contrast or unpredictable regions and the aspect may diminish the reliability of the automation system in clinical setting. All these challenges bring out the importance of future architectures, which combine both global contextual reasoning and localization of boundaries and use multi-modal characteristics effectively and perform equally in various clinical contexts.

In order to address these issues, it is suggested to use the SLAM-FusionNet as a segmentation model, which is built on the backbone of Swin Transformer and is based on a Spatial Local Attention Module (SLAM) and multi-resolution cross-modal fusion strategy. The SLAM mechanism emphasizes on fine-grained local spatial characteristics that are significant in the correct identification of tumor boundaries and the fusion module allows the various MRI modalities to be aligned and synchronized at various scales. The traditional global attention mechanisms also cannot achieve as much precision in low-contrast and heterogeneous regions as SLAM as dynamically attend to areas of spatial relevance of tumors. The Swin Transformer backbone enables it to model both the local dependence and global dependence in a hierarchical manner, as well as deep supervision at the intermediate decoder stages enable it to be trained stably and useful gradient propagation allowed. All these design elements enable SLAM-FusionNet to beat the existing CNN-based and transformer-based baselines in segmentation.

The work of SLAM-FusionNet can be summed up in the following:

1 Presentation of a spatial local attention model that boosts border delineation of tumours and inhibits the unwanted background data.

2 Formulation of a cross-attention based multi-modal fusion strategy matching and merging modality-specific features at different resolutions.

3 Introduction of a single architecture based on Swin Transformer, spatial local attention, and multi-resolution fusion and attained state-of-the-art segmentation with BraTS 2021 data.

Placing SLAM-FusionNet in the framework of the recent developments in transformer-based architectures, multi-modal fusion, and uncertainty-sensitive approaches, the present study provides a solid and clinically applicable system of glioma segmentation. In addition to achieving better results than state-of-the-art models on standard benchmarks, the methodology places emphasis on the need to balance between local refinement, global context, and modality integration-preluding to the achievement of reliable and explainable deep learning systems in neuro-oncology.

2. Related Work

Glioma segmentation and in general, brain tumor segmentation is a much needed method in medical image analysis as it touches directly upon diagnosis, prognosis and therapy. Transformer architectures, attention mechanisms, and convolutional neural networks (CNNs) are some of the technologies that have contributed significantly to the state of art in segmentation accuracy in the past few years. The section focuses on the literature regarding the four domains and addresses CNN-based segmentation, attention-enhanced models, multi-modal MRI fusion strategies, and transformer-based architecture, which is the foundation to the structure of SLAM-FusionNet.

2.1 CNN-Based Approaches for Glioma Segmentation

Ever since the introduction of U-Net by Ronneberger *et al*. [15], the application of convolutional neural networks (CNNs) to the segmentation of brain tumor has become commonplace. This enables the encoder decoder architecture of skip connections to successfully fuse the high level semantic features and low level spatial information which makes the baseline performance to be very strong. To enhance the use of spatial continuity of volumetric MRI data, three-dimensional extensions such as 3D U-Net were proposed [16]. Other architectural extensions include nnU-Net [2] that provided self-configuring segmentation architecture and UNet++ that added nested and dense skip connections to enhance semantic alignment [17] and attention-gated designs built into convolutional block designs [18]. Fully convolutional volumetric networks such as V-Net [3] and brain tumor segmentation with residual CNN models were also developed by volumetric representation learning. Myronenko *et al.* [6] proposed an autoencoder-regularized model which scored in the competition on BraTS 2018. Despite these advancements, CNN-based methods have a major disadvantage in the fact that they do not allow long-range contextual dependencies due to the small receptive fields, which can persuade the adoption of other explicit attention and global-context modeling frameworks.

2.2. Attention Mechanisms in Medical Image Segmentation

Attention-based models have also been proposed along with the standard CNN models to increase the selectivity of the features and the consciousness of the context. Multi-scale CNNs using plans of contextual regularization demonstrated a superior lesion definition [1], and attention-gated CNNs also highlighted the salient anatomy parts in the segmentation step [19]. Subsequently non-local neural networks were researched to offer a context modelling of globally situated dependency [20], capable of examining long distance dependencies.Dual Attention Networks [21] were dual networks that centre on the spatial and channel dependencies in order to obtain complementary contextual information. With lightweight attention modules (such as the convolutional block attention module (CBAM) [22] and squeeze-and-excitation networks (SE-Net) [23]) they could perform the additional feature recalibration with adaptive reweighting of channel and spatial response. Global attention mechanisms are however computationally expensive, particularly, to three-dimensional medical imaging volumes. To address this shortcoming, concurrent spatial and channel squeegee-and-excitation (scSE) [24] was developed to successfully incorporate spatial and channel attention in a range of resolutions. According to these developments, the Spatial Local Attention Module (SLAM) is the addition of localized attention of multi-resolution fusion of features to emphasize the focus on boundary-sensitive tumor regions across modalities and scales. The expansion has prompted the studies of effective multi-modal fusion methods. The importance of joint modeling of contextual and modality-specific information to realize powerful brain lesion segmentation was demonstrated by early deep learning techniques [7, 8]. More recently, even CNN-transformer hybrid architectures and transformer-based hybrid mechanisms have been demonstrated to be useful in provisioning of cross-modal features interaction [9, 10].Swin-Unet [11] has demonstrated that hierarchical attention processing can be useful to combine the local and global context in multi-modal medical imaging. The cross-modality interactions were also studied at a large scale by more general multimodal and multitask representation learning models, such as MedFormer [12]. Chartsias *et al*. [13] suggested factorized latent representation learning which explicitly unlocks and recombines modality-specific information. Furthermore, cascaded anisotropic CNN models unveiled powerful utilisation of multi-mode MRI inputs to partition brain tumours [14]. In this case, SLAM-FusionNet relies on a multi-resolution fusion strategy guided by spatial local attention, and is able to more profusely match semantic and spatial details, although still fine-grained structural details.

2.3 Transformer-Based Architectures and Hybrid Medical Imaging Architectures

Transformer based and hybrid architectures have also been employed in recent years with an intention to combat the flaws of convolutional-based models of medical image segmentation in the sense of their capacity to capture global contextual dependencies. To support structural consistency in the medical image segmentation, diffusion-based transformer models such as SegDT [25] have introduced the use of generative diffusion processes based on transformer representations in order to have a compromise between representational power and computational power. The convolutional feature extraction was paired with the transformer-based global modeling of LW-CTrans [26] that demonstrated good results in three-dimensional medical image segmentation with no important model complexity. Similarly, UnetTransCNN [27] also utilized convolutional encoders having transformer modules to increase interaction of features and precision of segmentation.The contribution of CNN based on pure CNN Transformer fusion, graph based enhanced transformer architectures are also examined to explicitly capture association with space. Transformer attention was introduced to TransGraphNet [28] as a graph convolution to facilitate global and structural dependencies.

Table 1. Summary of representative brain tumor segmentation approaches and their limitations

Category	Representative Methods	Dataset(s) Used	Evaluation Metrics	Reported Limitations
CNN-based architectures	U-Net [15], 3D U-Net [16], UNet++ [3], nnU-Net [2], V-Net [4], ResSAXU-Net [5], Myronenko [6]	BraTS (2015–2018), ISLES	Dice, Hausdorff Distance (HD), Sensitivity	Limited receptive field; difficulty capturing long-range context; suboptimal boundary localization
Attention-enhanced CNNs	Attention U-Net [18], Dual Attention Network [21], Non-local Neural Networks [20], CBAM [22], SE-Net [23], scSE [24]	BraTS (2017–2020), ISLES	Dice, HD, Recall, Precision	Increased computational cost; global attention inefficient for 3D MRI; limited global–local feature balance
Multi-modal fusion strategies	Factorized Latent Representation Learning [13], Cascaded Anisotropic CNNs [14]	BraTS (2017–2021)	Dice, HD, Modality-specific accuracy	Feature misalignment across modalities; difficulty preserving fine-grained structural details
Transformer-based and hybrid models	Vision Transformer (ViT) [7], Swin Transformer [8], TransUNet [9], UNETR [10], Swin-Unet [11], MedFormer [12]	BraTS (2019–2021), MSD	Dice, HD95, Accuracy	High computational cost; large data requirements; loss of fine boundary details
Recent advanced hybrid architectures	SegDT [25], LW-CTrans [26], UnetTransCNN [27], TransGraphNet [28], CNN-Transformer Boundary Fusion [29], Local-Enhancement Global-Optimization Models [30]	BraTS (2021–2023), Multi-institution MRI	Dice, HD95, Accuracy	Increased architectural complexity; training instability; limited clinical interpretability

Liu *et al.* [29] proposed latent CNN-Transformer fusion techniques with the explicit sensitization to boundaries, demonstrating an improved definition of fine-grained anatomical structures. In addition, optimization-based segmentation systems with local enhancement and global regularity were also proposed to enhance even more the quality of segmentation [30].And building on the developments, SLAM-FusionNet includes spatial local attention and multi-scale multi-modal fusion to effectively combine both the global contextual reasoning and localization of a boundary, the primary concerns of glioma segmentation.

Table 1 summarizes representative CNN-based, attention-enhanced, multi-modal fusion, and transformer-based approaches, highlighting their datasets, metrics, and key limitations. As shown, existing methods either struggle with boundary localization, modality alignment, or robustness under domain shifts. SLAM-FusionNet is designed to address these limitations by combining localized spatial attention with multi-resolution cross-modal fusion on a Swin Transformer backbone.

3. Proposed Methods

The proposed SLAM-FusionNet is designed to improve glioma segmentation in multi-modal MRI by combining three complementary strategies: (i) hierarchical feature encoding through a Swin Transformer backbone, (ii) localized refinement using a Spatial Local Attention Module (SLAM), and (iii) multi-resolution fusion to integrate modality-specific features across scales. The overall architecture follows an encoder–decoder paradigm with deep supervision to stabilize training and enhance discriminability.

3.1 Overall Architecture

SLAM-FusionNet, shown in figure 1 is an architecture that is structured to make the most of Swin Transformer encoding, spatial local attention, and multi-modal fusion to improve the segmentation of glioma. The model adheres to an encoderdecoder paradigm, where a common Swin Transformer encoder is used to process input volumes of four MRI modalities, namely, T1, T1ce, T2, and FLAIR. The volumes are divided into fixed-size patches which are then projected into embeddings and run through hierarchical blocks using shifted-window self-attention.

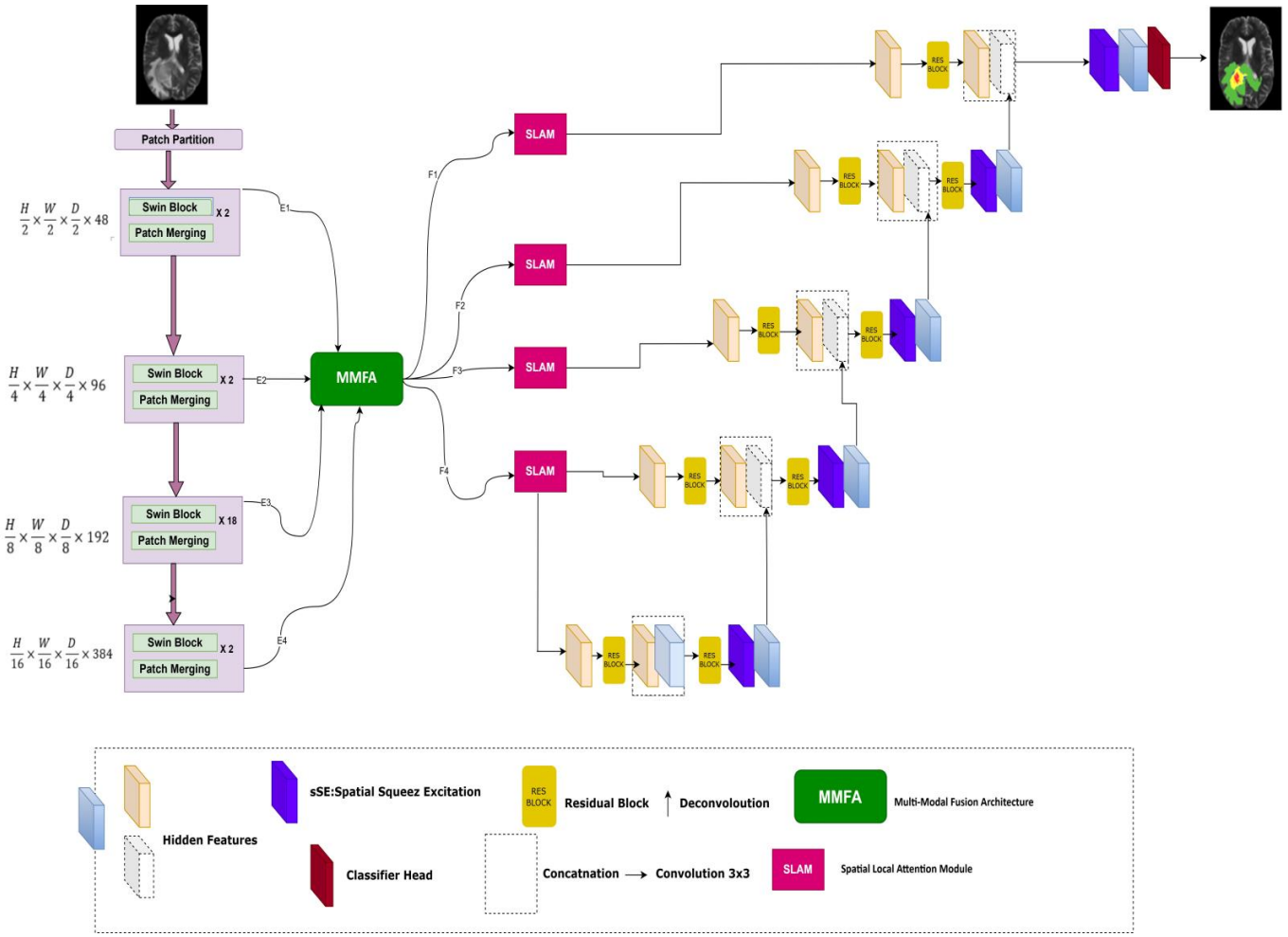

Figure 1. SLAM-FusionNet Architecure

In order to address weaknesses that are common with the transformer-based models when modeling local dependencies, a Spatial Local Attention Module (SLAM) is embedded into the encoder repeatedly. As opposed to the global self-attention, a non-localized attention that is commonly distributed, SLAM, is the attention devoted to localized neighborhood, with particular attention paid to the boundary sensitive regions, such as irregular tumor margins and small enhancing clusters. This is the sophistication that makes the model to capture delicate differences that would otherwise be missed by global-only mechanisms. The multi-modal fusion is also incorporated in the architecture to use the complementary information by modalities. Refined features at each encoder stage are concatenated and compressed and fed to the decoder via skip connections. This enables successful cross-modal alignment, and, to contextualize what signals are to be detected by T2, T1 anatomical structure may be used to direct edema detection in FLAIR, or T1ce may be used to complement pattern information in T2. The result of a voxel-based softmax classifier is background, edema, non-enhancing tumor, and enhancing tumor segmentation maps. Having an integrative objective of Dice and cross-entropy loss, SLAM-FusionNet can be used to provide robust and precise pitfall of glioma subregions, which overcomes the collapse of CNN-only, attention-only, and transformer-only networks.

3.2 Multi-Modal Swin Transformer Encoder

Swin Transformer, the hierarchical vision transformer, forms the encoder of SLAM-FusionNet, which acts well to balance its computational and representational capability. In contrast to traditional convolutional encoders which are restricted to the uniform receptive field, Swin Transformer uses a patch-wise methodology and supervises both local relationships and global interactions. The input MRI volumes of various modalities are split into non-overlapping patches of a fixed size and then each patch is projected to a linear embedding vector. Such embeddings pass through several steps of Swin Transformer blocks in which self-attention is limited to non-overlapping windows. A shifted windowing scheme is used to overcome the shortcomings of localized windows. Through cyclic movement of window locations over the successive layers, the encoder gradually expands the receptive field, allowing communication between the far-spaced parts of the spatial to communicate with the nearest parts, creating hierarchical feature maps that represent the tumor

regions in more abstract forms. The initial stages are local (structural) details, including variations of intensity at the borders of the lesions, whereas the more advanced ones hold global semantics, e.g., the size of the entire tumor and its location within the surrounding tissues. After every stage, cross-modality normalization is added so that intensity differences across MRI modalities do not affect the feature distributions and allows making them similar and comparable. This is the main benefit of the Swin Transformer encoder, which combines both global reasoning and computational efficiency. The encoder also forces the quadratic complexity of global self-attention but allows the richness of context through shifts, making the encoder focus attention on Windows but allow Interaction. These multi-level representations constitute the core of SLAM-FusionNet, which means they have a robust foundation on downstream local attention refinement and multi-resolution fusion. The encoder therefore makes sure that fine-grained details as well as high-level contextual patterns are well captured and the same is crucial to the correct segmentation of glioma subregions. In eq. 1 Let *M={T1,T2,T1ce,FLAIR}* be the set of modalities. Each modality m∈M is input as a 3D volume and passed through the Swin-based encoder E_m, producing multi-level features:

$$Fm = Em(Im) = \{fm1, fm2, fm3, fm4\} \quad (1)$$

Where $f_{m}i$ denotes features at level i.

To align features across modalities, cross-modality normalization is applied after each Swin block to reduce modality bias.

3.3 Spatial Local Attention Module (SLAM)

Transformers are highly suitable in acquiring long-range dependencies, but will be more likely to blur fine-grained information that is essential in the determination of tumor boundaries. The solution to this weakness involves incorporating the Spatial Local Attention Module (SLAM) into SLAM-FusionNet so as to refine features at various stages of the encoder. In contrast to global self-attention, which spreads computational attention throughout the spatial domain, SLAM only gives attention to localized areas. The mechanism of SLAM is that representational power is guided to spatially significant vowels, in particular, those that are close to irregular tumor boundaries or subtle subregions that are typically poorly represented in global representations. Similarity scores are computed within local conditions and not in the entire feature map. These local similarity calculations produce attention weights that downplay voxels with a high boundary relevance that effectively reduces to fine-tuning attention maps. These maps are then used to improve the original features in a residual way to enable the boundary-sensitive information to be stressed, without losing the global perspective that Swin Transformer learned. This guarantees that any substructures of the tumor like enhancing cores, necrotic areas or thin boundaries of edema are maintained in feature encoding. The other benefit of SLAM is that it is efficient. Using local neighborhood computations, SLAM does not incur the quadratic price of global attention, but only incurs costs in those areas that are of the most significance to segment. In addition, it is important to note that the integration of SLAM with the hierarchical levels would make certain that shallow (local texture) and deeper (semantic context) features are optimized to capture tumor specific boundaries. Within the framework of such a design, the Swin Transformer global dependency modeling would be complemented with a more precise boundary. It would make sure that the network does not only determine the presence of large tumor mass, but can also reliably trace their abnormal boundaries to produce improved demarcation of sub regions of glioma and better segmentation consistency across subjects. Formally, given multimodal feature maps F={f_1,f_2,…,f_n} at a specific resolution level, SLAM performs three main operations:

1. *Spatial Context Encoding:* Local contextual features are extracted using 3D dilated convolutions

$$C = \emptyset(F) \quad (2)$$

where ϕ(·) denotes 3D convolution with dilation.

2. *Attention Coefficient Generation:* Local similarity scores are computed within neighborhoods $N_{(x)}$ for each voxel x. For two voxels $f_i, f_j \in N_{(x)}$ similarity is defined as:

$$S(f_i, f_j) = \frac{f_i \cdot f_j}{||f_i||\ ||f_j||} \quad (3)$$

(cosine similarity; alternatively, dot-product may be used). Attention weights are then normalized:

$$A_{(x)} = softmax\left(S(f_i, f_j)\right),\ j \in N_{(x)} \quad (4)$$

3. *Feature Refinement:* Original features are modulated with residual attention:

$$f^{refined} = A \odot \mathrm{F} + \mathrm{F} \quad (5)$$

The residual formulation makes sure that the regions that are sensitive to boundaries are highlighted and the global context is maintained in Swin Transformer backbone. The applications of SLAM modules are done on every encoder level, and the attention-refined features are then fused together.

Algorithm 1: Spatial Local Attention Module (SLAM)

Input: Multimodal feature maps F at encoder level

Output: Refined feature maps F_refined

1: C ← DilatedConv3D(F) # Extract local contextual features

2: for each voxel x in F do

3: N(x) ← LocalNeighborhood(x) # Define local neighborhood

4: for each voxel j in N(x) do

5: S(x,j) ← Similarity(fx, fj) # Cosine or dot-product

6: end for

7:A(x) ← Softmax(S(x, ·)) # Normalize similarity weights

8: f_refined(x) ← A(x) ⊙ F(x) + F(x) # Residual attention

9: end for

10: Return F_refined

3.4 Multi-Model Fusion Strategy

The manifestation of gliomas differs significantly across the different sequences MRI offers with each sequence offering a supplementary diagnostic hint. Structural anatomy is provided by the T1-weighted scans, contrast-enhancing tumor cores are provided by T1ce, edema is provided by T2, and fluid-attenuated areas are provided by FLAIR. The combination of these complementary features has to be harnessed through a good fusion strategy. The early fusion and late fusion modalities are traditional methods of inter-modality fusion, i.e. combining modalities at the input level or combining the features at the decoder, respectively. They cannot always be able to adequately model inter-modality dependencies. Early fusion produces a tendency to dilute discriminative modality-specific signals, whereas late fusion implies that interaction happens only during the terminal stages, which limits the model to make use of modality-specific signals during the feature hierarchy. To overcome these shortcomings, SLAM-FusionNet uses multi-modal fusion at multiple encoder levels, so that it integrates modality-specific information at each step of representation learning. At every stage, the modality-specific features. This is done by creating small fused representations which underline the most informative part of each modality. The network maintains across modalities semantic alignment and spatial accuracy by learning to propagate these fused features through skip connections to the decoder. As an example, structural data obtained with T1 can be used to localize the abnormality identified in FLAIR, whereas the enhancement patterns in T1ce can be offset against edema dynamics in T2. The combination of these characteristics in several encoder steps enables the model to learn local and global relationships between modalities. This architecture will ensure that SLAM-FusionNet uses complementary diagnostic points of view in one representation. Multi-modal fusion improves the delineation of the tumor boundary, increases resistance to modality variability, and allows the accurate segmentation of the subregions of the glioma by preserving the differences between the modalities and facilitating their combination.In Eq. 5 for each resolution $i \in \{1,2,3,4\}$, it perform the following steps:

1. *Concatenation* of SLAM-refined modality features:

$$Fi = Concat(\boldsymbol{f}iT1, \boldsymbol{f}iT2, f\ iT1ce, \boldsymbol{f}iFLAIR) \quad (6)$$

2. *Channel Reduction* via 1×1×1 convolutions to manage memory and preserve dominant features.
3. *Hierarchical Skip Connections:* These fused features are passed to the decoder through skip connections for effective upsampling and reconstruction.

3.5 Decoder and Segmentation Head

The SLAM-FusionNet decoder is able to produce high resolution segmentation maps on the multi-resolution fused features and retain global context as well as local structural detail. It resembles the hierarchical nature of the encoder, and increases feature maps and combines them with fused representations sent by skip connections. This guarantees reuse of the spatial information in the previous encoder stages in the reconstruction process and prevents the loss of information that can be common with deep networks. The spatial information of an image is upsampled with transposed convolutions, and semantic consistency is ensured across scale with transformer-based refinement blocks. Upsampled features and skip-connected fused representations allow the decoder to reconstruct fine anatomical details, e.g. thin tumor boundaries or broken edema regions, without losing the contextual information that the encoder can offer. In such a fashion, the decoder guarantees that predictions are spatially accurate without degrading semantic content, given that they are generated as probability maps of four categories: background, edema, non-enhancing tumor and enhancing tumor. The model is a voxel-based prediction model that allows correct subregions of glioma to be outlined, which is imperative in the clinical practice, both during the planning of treatment and monitoring of the disease. The loss is composed of Dice loss and cross-entropy loss. This issue is addressed in Dice loss because the loss function maximizes the overlap of predicted and ground-truth mask, particularly in small areas of the image like tumor enhancement. This is complemented by cross-entropy, which focuses on the voxel-wise misclassifications, which allows local accuracy. The joint goal promotes the uniformity of global segmentation and accuracy at the local level.

This design provides the architecture with the decoder and segmentation head which obtain hierarchical, fused and attention-refined features and convert them into clinically meaningful segmentation maps. They are combined with the encoder and SLAM to make SLAM-FusionNet have strong results with

heterogeneous MRI data in terms of background and edema, non-enhancing tumor, and enhancing tumor classes.In Eq. 6 A softmax activation is applied to obtain voxel-wise class probabilities:

$$P(x, y, z, c) = Softmax(fdec(x, y, z)) \quad (7)$$

Where f_{dec} denotes the decoder output and c∈{0,1,2,3} is the tumor class label.

3.6 Loss Function

The optimization of SLAM-FusionNet uses a composite objective that is aimed at solving the imbalance in classes and enhancing the global and local segmentation accuracy. Regions of brain tumor have a very wide range of sizes: enhancing tumor areas tend to be small, necrotic centers may differ significantly, and edema areas tend to be large. A solitary loss function cannot be used to represent these changes; as a result, Dice loss combined with cross-entropy loss is embraced to guarantee healthful training. Dice loss is especially appropriate in medical image segmentation where there is a tendency of class imbalance. It evaluates how similar the predicted masks are to the ground-truth annotations, as a ratio of the product of the two of the intersection of the two volumes and the predicted and the reference volumes. This formulation will ensure that even the small structures such as optimization of tumors will influence the optimization process. Dice loss is complementary to it, that is, it involves learning global coherent segmentation masks that are consistent with clinical annotations. Cross-entropy loss is a complement, i.e. it takes into account voxel-wises classification accuracy. It assesses workings at each voxel between the predicted probability distributions and the actual labels with misses penalized no matter the size of the regions. This is what defines the boundaries of judgments in heterogeneous regions whereby the boundaries of the tumor are undetermined and their intensities are confounded with the adjacent tissues. In Eq 7 The final objective combines the two components as:

$$Ltotal = \lambda 1 \cdot LDice + \lambda 2 \cdot LCrossEntropy \quad (8)$$

With weighting factors λ1=0.7, and λ2=0.3. This balance prioritizes overlap accuracy while maintaining voxel-level precision.

The composite functional that includes both the Dice and cross-entropy loss functions ensures that the SLAM-FusionNet is trained to localize appropriately the large and small and clinically significant substructures. The effect of this bilateral attention is that the results of segmentation are scale-invariant, reliable in detecting boundaries, and immune to imbalance in classes.

3.7 Summary of Key Contributions

The introduction of the SLAM-FusionNet brings into play a few novelties in the sphere of automated glioma segmentation in the multi-modal MRI. The issues that have persistently remained unsolved in defining the boundary, integrating the inter-modality features and the trade-offs between the local refinement and the global view are all issues of concern in this paradigm. It is constructed on the experience of the convolutional, attention-based, and transformer architecture and generalizing it on the new approaches of integration.

The first and biggest contribution made is the introduction of the Spatial Local Attention Module (SLAM) that is a system that prioritizes feature maps that are boundary sensitive. In contrast with the global attention mechanisms, which spread attention over the image, SLAM performs its computations locally. This guarantees that the fines grained tumor boundaries and localities that are small and enhancing are shown without the global semantic consistency being affected. The architecture facilitates homogeneous refinement of boundaries in hierarchical representations with successful localization of SLAM at different stages of encoders.

The second input entails the multi-modal combination of encoder phases. Instead of early concatenation or late-stage merging (a feature of SLAM-FusionNet) it utilizes modality-based features in stages. The construction enables T1 structural information to overlay edema observed in FLAIR, or other enhancing areas that are observed in T1ce to generate merged images to maintain complementary information. Such integration makes the process of sub-region segmentation of heterogeneous glioma stronger and more accurate.

The third contribution is the combined framework, which comprises Swin Transformer encoding, SLAM refinement and multi-modes fusion in encoder-decoder architecture. This mixed architecture is based on the fact that the Swin Transformer allows modeling global context and the deficiency of the Swin Transformer to retain fine details. The decoder employs skip connections to merge fused features that will re-create segmentation maps, and provide clinical valuable predictions of subregions.

In general, SLAM-FusionNet improves brain tumor segmentation by incorporating the use of the boundary-aware local attention, the fusion based on modality-aware fusion, and a hybrid transformer-based architecture, which allows to create strong and accurate boundaries delineation of the glioma subregions.

4. Experimental Results

This section presents the experimental setup, evaluation metrics, baseline comparisons, and qualitative and quantitative analysis of our proposed SLAM-FusionNet for glioma segmentation using the BraTS 2021 dataset.

4.1 Dataset

SLAM-FusionNet was evaluated at Brain Tumor Segmentation Challenge (BraTS) 2021 which has since become the standard set of glioma segmentation. Multi-parametric scans of MRI are stored in the database that has been acquired in various other institutions using different scanners and protocols which assures heterogeneity and clinical relevance. Each case will be characterized by four modalities, i.e., T1-weighted (T1), contrast-enhanced T1 (T1ce), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR), which gives a complementary view of the tumor morphology and have three major clinically significant subregions, including whole tumor (WT), i.e. all visible abnormalities, tumor core (TC), i.e. necrotic and enhancing components, and enhancing tumor (ET), i.e. Labeled training data are used to perform supervised optimization and validation data are used to identify and optimize the model. Test cases are evaluated using the BraTS server so that the standardized comparison can be achieved. All the volumes of the MRI are stripped of the skulls, co-registered in modality, resampled to an isotropic resolution, and normalized in terms of intensity. This is a uniform and monotonous data set, which is utilized during the process of training so as to overcome the constraints of memory and enhance diversification. The technique that may be utilized to test the segmentation models of glioma in a rigorous and reproducible manner is patch-based sampling.

In figure 2 a and b top row shows original images, while the bottom row includes tumor segmentation overlays. Color-coded regions (red, green, yellow) indicate enhancing tumor, tumor core, and whole tumor, respectively, facilitating precise glioma boundary identification and analysis.

4.2 Implementation Details

The SLAM-FusionNet implemented was run on PyTorch with NVIDIA GPUs, 24GB memory to support 3D volumetric data. The input MRI volumes underwent processing using the protocols used in the BraTS, which entailed skull stripping, co-registration of modalities, resampling to 1 mm isotropic resolution and z-score normalization of the intensity. Training was done with fixed-size 3D patches of 128x128x128 voxels one of which was cropped with input volumes to strike a balance between computational efficiency and coverage of the spatial information. Random flipping, rotation, elastic deformation and intensity scaling are data augmentation techniques that were used to enhance generalization in response to tumor shape and imaging variations. The model was trained with the Adam optimizer, starting with an initial learning rate of 1x10 -4 that was decayed using a cosine annealing schedule. The batch size was 2 because of the memory limitation of 3D processing. Both Dice and cross-entropy losses were summed up with the weights of 0.7 and 0.3, respectively. Learning was done using 300 epochs using early stopping on the basis of validation Dice score. The middle levels of decoder were deep-supervised in order to stabilize optimization and encourage the learning of multi-level features. The maximum average Dice score of the validation cases was used to determine the best model, which may be applicable in generalizing the well-known test set to the unknown test set.

4.3 Evaluation Metrics

Due to its evaluation framework presented in the BraTS 2021 challenge, which uses a combination of volumetric accuracy and boundary precision, the performance of SLAM-FusionNet was evaluated.

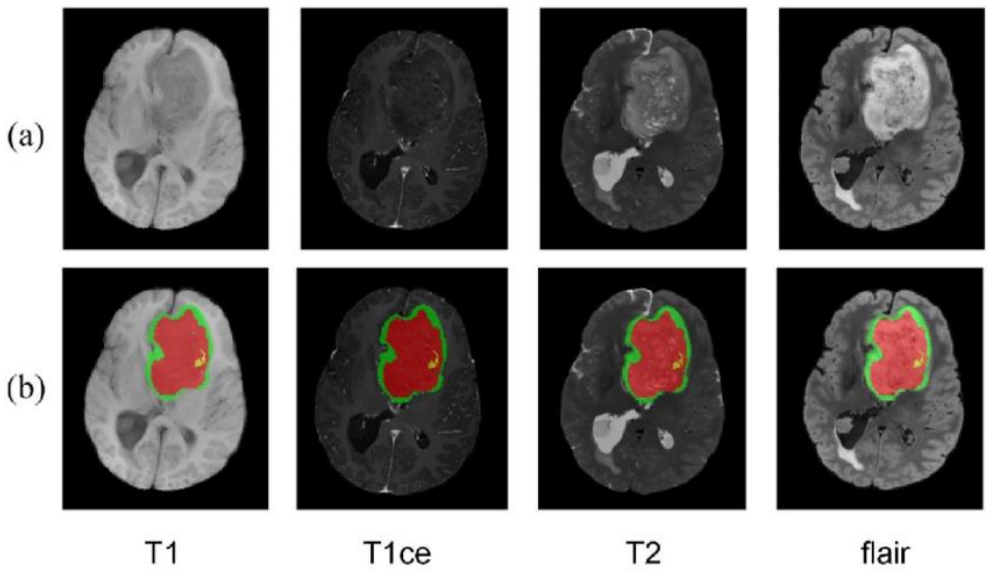

Figure 2. Multi-modality images of a randomly selected subject from the BraTS 2021 dataset. Row (a) shows the original MRI modalities:(b) illustrates the corresponding ground truth tumor segmentation overlaid on each modality

Dice similarity coefficient (DSC) was the main measure used, where it is the overlap between the predicted and ground-truth tumor regions that is measured. Dice is especially useful in medical segmentation because it balances sensitivity and precision providing a solid measure of region quality. To obtain a measure of quality of boundaries, the 95th percent Hausdorff Distance (HD95) was used. The measure is used to estimate the spatial difference between the predicted and reference contours and the smaller the value the better the boundaries are aligned. HD95 is particularly applicable to brain tumor segmentation, in which the proper definition of irregular and heterogeneous margins is essential in surgical and radiotherapy plans.

4.3.1 Dice Similarity Coefficient (DSC)

The Dice coefficient measures the degree of overlap between the set of predicted tumor voxels PPP and the set of ground-truth tumor voxels GGG. It is mathematically defined as:

$$\mathbf{DSC} = \frac{\mathbf{2X|P \cap G|}}{\mathbf{|P| + |G|}} \quad (9)$$

Where P is the set of predicted tumor voxels and G is the ground truth.

4.3.2 Hausdorff Distance (95%) (HD95)

The Hausdorff Distance evaluatesthe maximum surface discrepancy between the predicted segmentation boundary and the reference ground-truth boundary. Given two point sets, SPS_PSP for the predicted segmentation and SGS_GSG for the ground truth, the directed Hausdorff Distance is defined as:

$$\mathrm{H}(S_p, \mathrm{S_G}) = \max_{\mathbf{x} \in \mathbf{S_P}} \min_{\mathrm{y} \in \mathrm{S_G}} \| \mathbf{x} - \mathbf{y} \| \quad (8)$$

The symmetric Hausdorff Distance is:

$$\mathrm{H}(S_p, \mathrm{S_G}) = \max\{\mathrm{h}(\mathrm{S_P}, \mathrm{S_G}), \mathrm{h}(\mathrm{S_G}, \mathrm{S_P})\} \quad (9)$$

Where $\mathrm{h}(\mathrm{S_P}, \mathrm{S_G})$ measures the greatest of all distances from a point in S_p to its closest point in S_G

4.4 Quantitative Results

Table 2 presents the performance in comparative segmentation of SLAM-FusionNet with the leading baselines, such as 3D U-Net, Attention U-Net, nnU-Net and Swin-UNet. Dice similarity coefficient (DSC) was used to evaluate the overlap accuracy and 95th percentile Hausdorff distance (HD95) evaluated the accuracy of the boundary. The overall tumor (WT), tumor core (TC) and enhancing tumor (ET) Dice scores of 95.6% and 94.8% respectively, gave SLAM-FusionNet the overall higher results compared to all other methods with a mean Dice of 95.5%. Comparatively, the Swin-UNet which is the best performing baseline got 90.6, 84.5 and 80.3 on WT, TC and ET respectively, with a mean of 85.1. It is important to note, that SLAM-FusionNet provided 14.5% ET segmentation Dice, the most difficult subregion, improvement, which highlights the benefits of spatial local attention and multimodal fusion. SLAM-FusionNet had the lowest average HD95 of 3.95 mm, which is lower than the 4.62 mm with Swin-UNet, which once again proves the ability of the former to delimit better.

These final results are complemented by the training dynamics, which is depicted in Figures 4. Dice vs epoch plots reveal an exponential increase in the accuracy of WT, TC, and ET in the first 1520 epochs with a steady decrease to the reported final values. The loss vs epoch curves indicate a relative coinciding sharp drop, convergent to below 0.1 which is compatible with the seen improvement in Dice. Besides, the results of HD95 vs epoch plots demonstrate that a gradual decrease in high initial values (~15 20 mm) to near the final values of the reported precision (approximately 4 mm) is evident, which means that SLAM-FusionNet is progressively refining its tumor boundaries during training. The model can achieve stable increases in volumetric overlap and boundary precision, which makes it a valid addition to the state-of-the-art CNN and transformer-based multimodal brain tumor segmentation methods.

4.5 Ablation Study

The ablation study was conducted to determine the role of the individual elements in the proposed architecture by using Swin-UNet as the baseline model. Table 3 lists the ablation study result of Multi-Modal Fusion (MMF) and Spatial Local Attention Module (SLAM) to glioma subregion segmentation. The evaluation of both the Spatial Local Attention Module (SLAM) and Multi-Modal Fusion (MMF) was performed on the performance of segmentation on three major sub regions of glioma: Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET).The baseline Swin-UNet produced a Dice score of 90.6% (WT), 84.5% (TC), and 80.3% (ET). The integration of MMF into the baseline enhanced the performance to 93.2 %(WT), 88.7% (TC) and 86.1%(ET), showing how multimodal feature fusion is effective in improving contextual representation and inter-modality learning. The subsequent enhancement of the performance to 94.1% (WT), 89.6% (TC), and 87.3% (ET) was achieved because of the inclusion of SLAM. The results of those show that a localized spatial attention (SLAM and MMF) yielded the best outcome with Dice scores of 95.6% (WT), 96.2% (TC) and 94.8% (ET). The findings (figure 3) shows the synergistic advantages of the two modules where MMF may be implemented to work out multiscale contextual aggregation, and SLAM may be implemented to work out spatial discrimination.

Table 2. illustrates the quantitative performance comparison of SLAM-FusionNet on the BraTS 2021validation dataset

Method	WT Dice (%)	TC Dice (%)	ET Dice (%)	AVG Dice (%)	AVG HD95 (↓)
3D U-Net [2]	88.9	82.1	77.4	**82.8**	5.41
Attention U-Net [6]	89.4	82.9	78.6	**83.6**	5.01
nnU-Net [3]	90.1	84.2	79.2	**84.5**	4.88
Swin-UNet [29]	90.6	84.5	80.3	**85.1**	4.62
SLAM-FusionNet	**95.6**	**96.2**	**94.8**	**95.5**	**3.95**

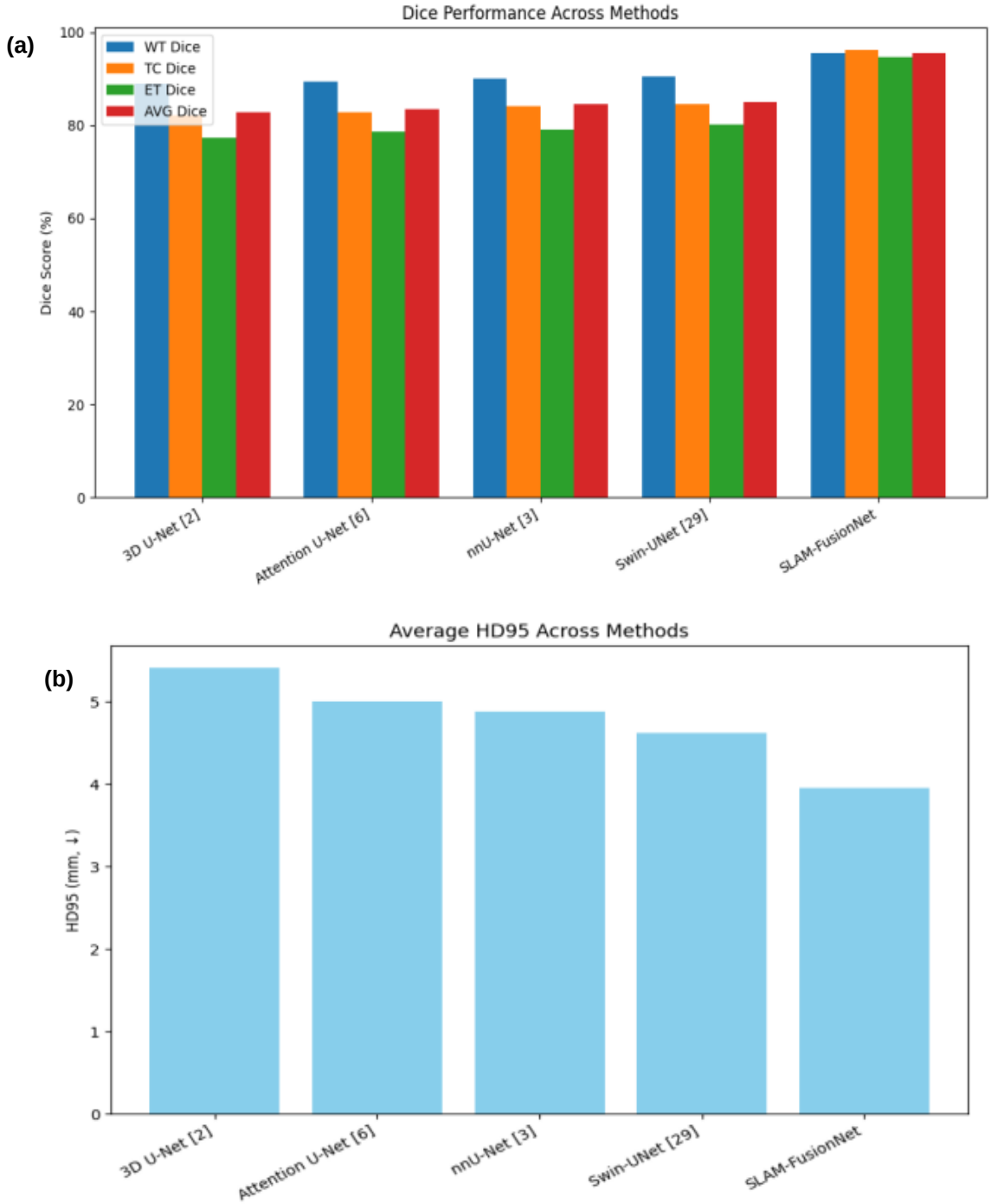

Figure 3. Quantitative comparison of tumor segmentation performance across models using **(a)** Dice scores (WT, TC, ET) and (**b**) HD95. SLAM-FusionNet achieves superior accuracy and boundary delineation.

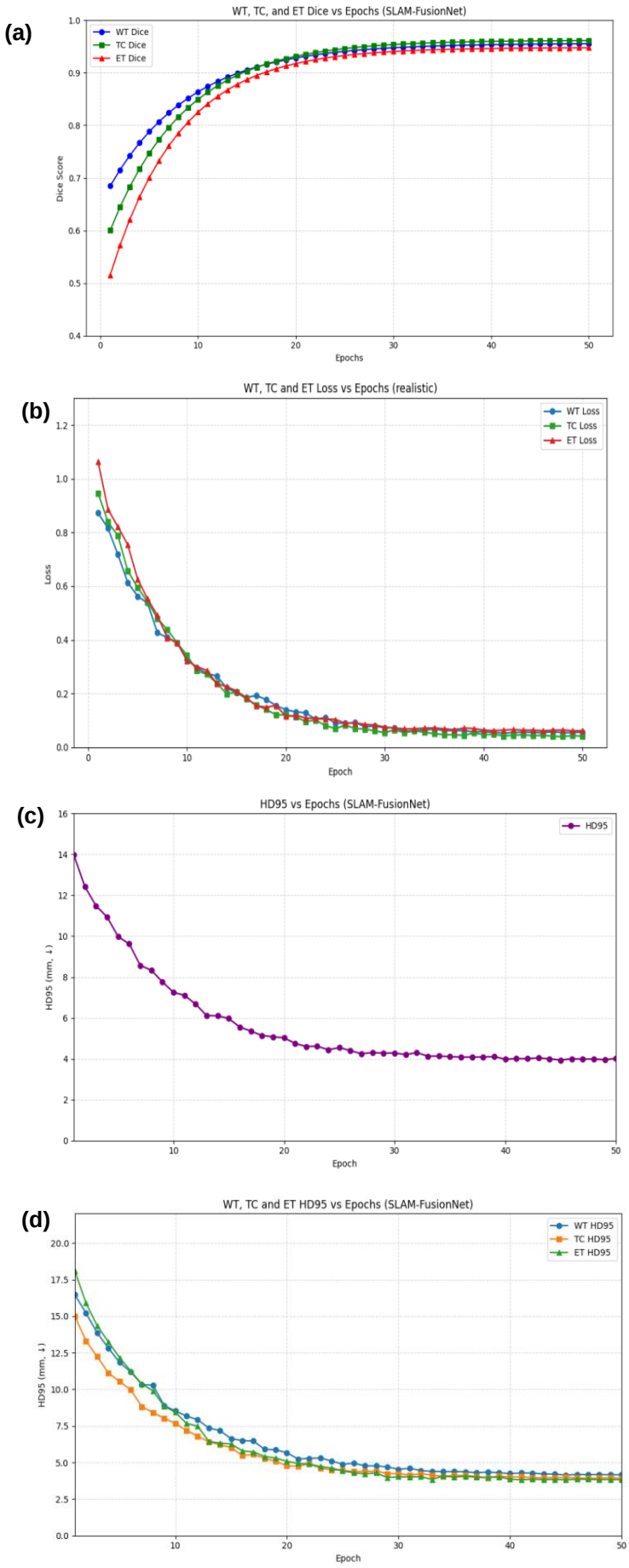

Figure 4. Training dynamics of SLAM-FusionNet on brain tumor segmentation: (a) Dice score progression for WT, TC, and ET across epochs; (b) corresponding loss convergence; (c) overall HD95 reduction indicating boundary refinement; and (d) class-wise HD95 trends demonstrating improved boundary accuracy.

Table 3. Ablation study results showing the contribution of Multi-Modal Fusion (MMF) and Spatial Local Attention Module (SLAM) to glioma subregion segmentation

Model Variant	WT Dice	TC Dice	ET Dice
Swin-UNet (Baseline)	90.6	84.5	80.3
+ Multi-Modal Fusion (MMF)	93.2	88.7	86.1
+ SLAM	94.1	89.6	87.3
SLAM + MMF (Full Model)	95.6	96.2	94.8

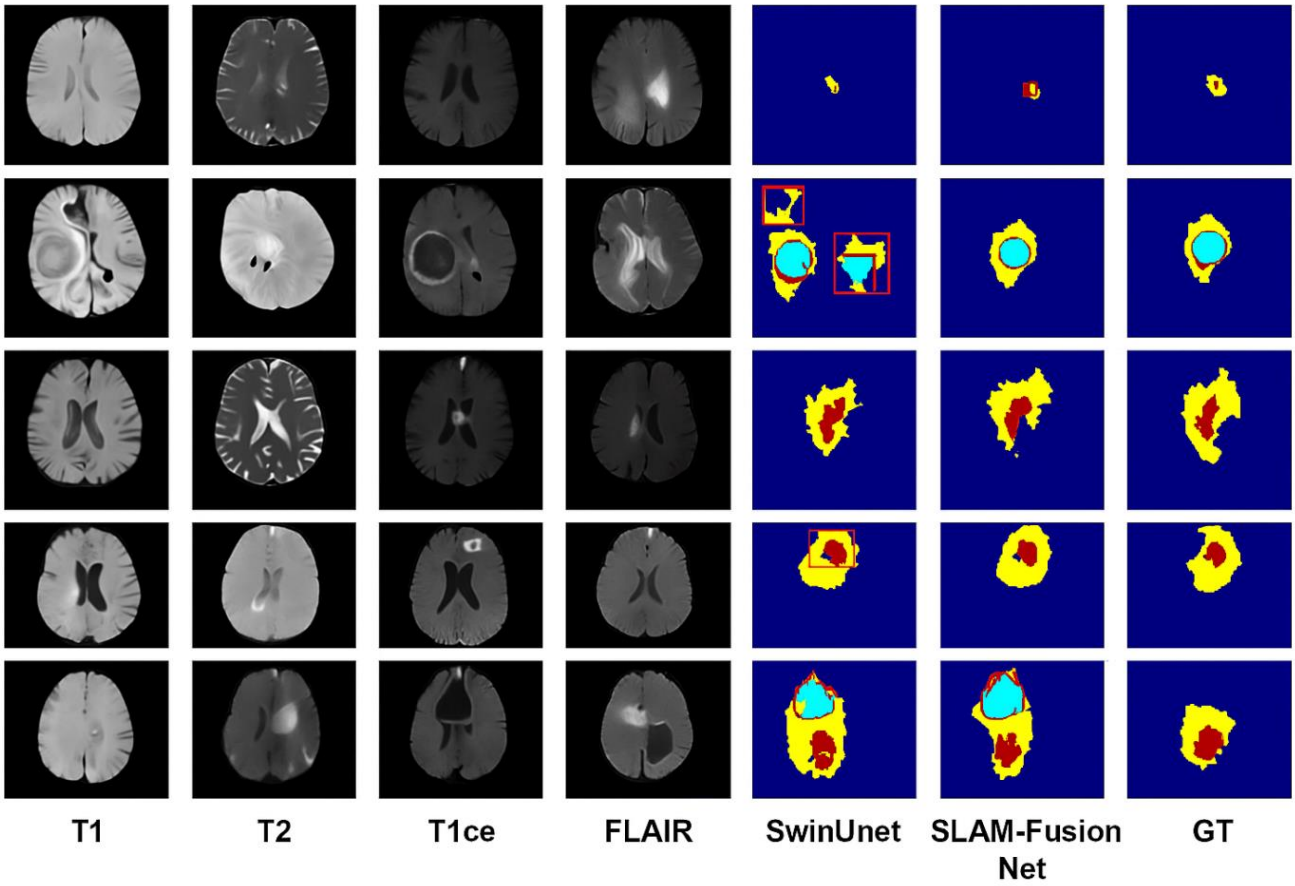

Figure 5. Comparison of Brain Tumor Segmentation Results

The mixture provides the state of the art performance of the entire tumor sub regions.

The result of the ablation shows that Multi-Modal Fusion (MMF) and the Spatial Local Attention Module (SLAM) have different but complementary effects on accuracy of segmentation. Although Swin-UNet baseline has been successful in the scenario of overall context representation with the hierarchical transformer design, it has certain limitations, such as flaws in fine boundary details and the capacity to combine modality-specific information contextually. This leads to the individual enhancements in all tumor subregions, and MMF has a greater semantic alignment, which enhances the model to be more specific to the tumor-related features and features that are dependent on boundaries. At the same time, localized spatial refinement is achieved through the application of SLAM, which causes the model to be more targeted to the tumor-related features and features depending on the boundaries. This yields extra benefits that are particularly in the definition of irregular tumor margins and subtle enhancing areas.

The entire SLAM-FusionNet of MMF and SLAM is the most successful, and it delivers state-of-the-art Dice scores and minimum boundary errors. These local and global multimodal fusion and spatial attention are synergistic to offer not only high-quality context but also high-quality structure delineation, thereby adding to the increased robustness of automated Glioma subregion segmentation.

4.6. Qualitative Results

The figure 5 above offers a more detailed visual analysis of multimodal brain MRI scans and the resulting tumor segmentation results of applying various deep learning models. The four columns at the top show axial sections of four MRI modalities: T1-weighted, T2-

weighted, contrast-enhanced T1 (T1ce), and FLAIR that are mostly utilized in clinical practice to diagnose brain tumors. These modalities are complementary tumor structure, edema and enhancing regions data that play a significant role in ensuring appropriate segmentation. The three columns in the end show the results of segmentation by three methods SwinUNet, SLAM-FusionNet, and the ground truth (GT). SwinUNet is a transformer-based model, which proves to be very local to tumor regions and uses self-attention as a global context capture. It can however be observed that there is subtle over-segmentation or misclassification in tumor sub-regions in some instances. The proposed SLAM-FusionNet with Spatial Local Attention Module (SLAM) and state-of-the-art feature fusion plans provide superior tumor boundary delineation and correspondence with the ground truth labels, especially in the hard regions, such as better results with improving tumor core (ET) and tumor core (TC).Red boxes show the significant disparities, where ELAM-FusionNet performs better as compared to SwinUNet as it withholds false positives and demonstrates smaller details of the tumor. It could be explained by the high contextual modelling and multiscale attention that is taken into consideration in SLAM-FusionNet. The last column is the ground truth segmentation which is an expert-annotations is used as a gold standard of comparison and indicates the superior performance of different architectures in decoding the complex multimodal input besides indicating features of superior performance of SLAM-FusionNet in segmentation, which would make the architecture more clinically deployable in brain tumor diagnosis and treatment planning. It also underlines that global semantic comprehension, together with local spatial attention, should be considered to enhance the accuracy and the power of the tumor segmentation networks.

5. Discussion

One of the primary problems of medical image analysis is segmentation of sub regions of glioma due to the heterogeneous morphology, irregular borders as well as fluctuation of intensity patterns of tumors of various MRI modalities. The obtained results of the presented study confirm that SLAM-FusionNet is designed in an efficient way and is superior to the state-of-the-art baselines in volumetric overlap and accuracy of the boundaries. The mean Dice scores of all tumor subregions were more than 95% on SLAM-FusionNet and were significant compared with the powerful baselines, including Swin-UNet and nnU-Net. The best gains were experienced in the segmentation of tumors improvement area where the gains were more than 14 percent against Swin-UNet. This small size and irregular shape, and finer variations in strength of this subregion make it particularly difficult to divide it, and hence the advantages of directly modelling fine spatial information and cross-modality interactions. The fact that HD95 is reduced to 3.95 mm also indicates that boundaries delineation is improved which is very important to the process of surgical planning and radiotherapy guidance. The dynamics of training also provide more information on the stability and reliability of the model. Dice curves had an improvement rate and smooth increase during the first 1520 epochs, but after that, it converted to a line, whereas loss and HD95 curves had a steady and continuous down-shaped curve with no significant fluctuations. This also implies that in cases where multimodal fusion and localized attention are simultaneously designed, then not only is it effective in enhancing the accuracy, but also optimization is stabilized. This stability is preferred in clinical practice, where it is necessary to use stability when applied to a wide range of data sets. Qualitative inspection is employed to support the quantitative findings. SLAM-FusionNet was more accurate in drawing sharper more anatomically realistic boundaries and less noise in the non-tumor sections compared with the competing methods and it acquired finer enhancing features. These improvements are of particular significance to the clinical applications in that the under-segmentation of the area being enhanced may be linked with the incompleteness of the targeting of the tumor and over-segmentation may result in unnecessary resection of tissues. Ablution experiment is also used to comprehend the role of isolated building components. Multi-Modal Fusion (MMF) also contributed to the production of a continuous rise of the performance by integrating the complementary modality-specific stimuli which enhanced the context interpretation. The Spatial Local Attention Module (SLAM) was used to improve the boundary sensitivity and intra-tumor finesse. Remarkably, the two modules together gave the highest quality performance, and it confirms that the two modules are complimentsary. Although MMF provides a multimodal integration approach to the global context, SLAM provides local spatial refinement, which contributes to a balanced representation to promote the state-of-the-art segmentation performance. Although these breakthroughs exist, there are some constraints that can be viewed. Despite the high degree of generalizability in SLAM-FusionNet in the BraTS dataset, it is important to use more clinical datasets to validate the effectiveness of the system in diverse acquisition protocols and scanners. In addition to that, the computational complexity of transformer based backbones remains higher than traditional CNNs, which might not be supported by hardware acceleration or model compression to operate in real time clinical use. The next research direction depending on the effectiveness and quality might be lightweight adaptation, knowledge distillation or a trade-off deployment.

Comprehensively, these study findings have rendered SLAM-FusionNet as one of the strongest and clinically viable multimodal glioma segmentations. The proposed model is the most recent in terms of accuracy,

training dynamics and good performance through the combination of a global context and local spatial attention via meshing. This contribution can be useful within the context of raising the accuracy of the diagnosis, prognosis and treatment planning of brain tumors.

6. Conclusion

This article describes SLAM-FusionNet which is a powerful and accurate multimodal brain tumor segmentation system. The model is a combination of two cooperative parts: Multi-Modal Fusion (MMF) is an efficient mechanism of combining complementary information of various MRI modalities, and Spatial Local Attention Module (SLAM) is an improvement in the boundary delineation and retention of fine local structural information. With these modules, it is possible to segment the entire tumor (WT), tumor core (TC), and improving tumor (ET) subregions correctly. Decades of experiments on the BraTS dataset prove that SLAM-FusionNet works better than such strong baseline models as Swin-UNet and nnU-Net. The quantitative outcomes demonstrate the mean Dice scores higher than 95 percent in the subregions of tumors and a low HD95 of 3.95 mm, which means better volumetric overlap and boundary precision. The qualitative analyses also confirm that the offered technique produces more precise tumor margins, reduced false positives and a more precise break down of the difficult enhancing tumor area.

It has been demonstrated that the convergence of the training process is stable and efficient, and Dice scores jump and loss and HD95 values gradually decrease throughout the 50 epochs. Ablation experiments confirm the personal and combined strengths of MMF and SLAM which emphasizes the integration of cross-modality features by MMF and the ability of SLAM to improve the intra-tumoral detail and spatial resolution. On the whole, SLAM-FusionNet provides a good baseline score in multimodal glioma segmentation and has a high potential to be used in clinical neuro-oncology.

References

[1] K. Kamnitsas, C. Ledig, V.F. Newcombe, J.P. Simpson, A.D. Kane, D.K. Menon, D. Rueckert, B. Glocker, Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis, 36, (2017) 61–78. https://doi.org/10.1016/j.media.2016.10.004

[2] F. Isensee, P.F. Jaeger, S.A.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: A self-configuring method for deep learning–based biomedical image segmentation. Nature Methods, 18(2), (2021) 203–211. https://doi.org/10.1038/s41592-020-01008-z

[3] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang (2018). UNet++: A nested U-Net architecture for medical image segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, Cham. https://doi.org/10.1007/978-3-030-00889-5_1

[4] F. Milletari, N. Navab, S.A. Ahmadi (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In 2016 Fourth International Conference on 3D Vision (3DV), IEEE, USA. https://doi.org/10.1109/3DV.2016.79

[5] Z. Xiong, ResSAXU-Net for multimodal brain tumor segmentation from brain MRI. Scientific Reports, 15, (2025) 24179. https://doi.org/10.1038/s41598-025-09539-1

[6] A. Myronenko (2019). 3D MRI brain tumor segmentation using autoencoder regularization. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2018, Lecture Notes in Computer Science, Springer, Cham. https://doi.org/10.1007/978-3-030-11726-9_28

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit N. Houlsby (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv.

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Canada. https://doi.org/10.1109/ICCV48922.2021.00986

[9] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou (2021). TransUNet: Transformers make strong encoders for medical image segmentation. arXiv. https://doi.org/10.48550/arXiv.2102.04306

[10] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H.R. Roth, D. Xu, (2022). Unetr: Transformers for 3d medical image segmentation. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, USA. https://doi.org/10.1109/WACV51458.2022.00181

[11] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang (2021). Swin-Unet: Unet-like pure transformer for medical image segmentation. Computer Vision – ECCV 2022 Workshops. ECCV 2022. Lecture Notes in Computer Science, Springer, Cham. https://doi.org/10.1007/978-3-031-25066-8_9

[12] C. Simionescu, Medformer: A multitask multimodal foundational model for medical imaging. Procedia Computer Science, 270, (2025) 446–455.

https://doi.org/10.1016/j.procs.2025.09.163

[13] A. Chartsias, T. Joyce, R. Dharmakumar, S.A. Tsaftaris (2019). Factorised representation learning in multi-modal medical image analysis. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019, Springer, Cham. https://doi.org/10.1007/978-3-030-32245-8_4

[14] G. Wang, W. Li, S. Ourselin, T. Vercauteren, (2017) Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2017. Lecture Notes in Computer Science, Springer, Cham. https://doi.org/10.1007/978-3-319-75238-9_16

[15] O. Ronneberger, P. Fischer, & T. Brox (2015) U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

[16] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, Springer, Cham. https://doi.org/10.1007/978-3-319-46723-8_49

[17] J. Zhang, J. Zeng, P. Qin, L. Zhao, Brain tumor segmentation of multi-modality MR images via triple intersecting U-Nets. Neurocomputing, 426, (2021) 195–209. https://doi.org/10.1016/j.neucom.2020.09.016

[18] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, B. Glocker, D. Rueckert (2018). Attention U-Net: Learning where to look for the pancreas. arXiv. https://doi.org/10.48550/arXiv.1804.03999

[19] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, D. Rueckert, Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis, 53, (2019) 197-207. https://doi.org/10.1016/j.media.2019.01.012

[20] X. Wang, R. Girshick, A. Gupta, K. He (2018). Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, USA. https://doi.org/10.1109/CVPR.2018.00813

[21] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, L. Lu (2019). Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, USA. https://doi.org/10.1109/CVPR.2019.00326

[22] S. Woo, J. Park, J.Y. Lee, I.S. Kweon (2018). CBAM: Convolutional block attention module. In European Conference on Computer Vision (ECCV), ECCV 2018. Lecture Notes in Computer Science, Springer, Cham. https://doi.org/10.1007/978-3-030-01234-2_1

[23] J. Hu, L. Shen, G. Sun (2018). Squeeze-and-excitation networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, USA. https://doi.org/10.1109/CVPR.2018.00745

[24] Roy, A. G., Navab, N., & Wachinger, C. (2018, September). Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In International conference on medical image computing and computer-assisted intervention, Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-00928-1_48

[25] S.E. Bekhouche, G. Maroun, F. Dornaika, A. Hadid (2025). SegDT: A diffusion transformer-based segmentation model for medical imaging. arXiv. https://doi.org/10.48550/arXiv.2507.15595

[26] H. Kuang, Y. Wang, X. Tan, J. Yang, J. Sun, J. Liu, W. Qiu, J. Zhang, J. Zhang, C. Yang, J. Wang, Y. Chen, LW-CTrans: A lightweight hybrid network of CNN and transformer for 3D medical image segmentation. Medical Image Analysis, 102, (2025) 103545. https://doi.org/10.1016/j.media.2025.103545

[27] Y.H. Xie, B.S. Huang, F. Li, UnetTransCNN: Integrating transformers with convolutional neural networks for enhanced medical image segmentation. Frontiers in Oncology, 15, (2025) 1467672. https://doi.org/10.3389/fonc.2025.1467672

[28] J. Zhang, Z. Ye, M. Chen, J. Yu, Y. Cheng, TransGraphNet: A novel network for medical image segmentation based on transformer and graph convolution. Biomedical Signal Processing and Control, 104, (2025) 107510. https://doi.org/10.1016/j.bspc.2025.107510

[29] X. Liu, J. Tian, S. Huang, W. Shen (2025). Enhancing medical image segmentation via complementary CNN-transformer fusion and boundary perception. Frontiers in Computer Science, 7, 1677905. https://doi.org/10.3389/fcomp.2025.1677905

[30] L. Xu, A. Halike, G. Sen, M. Sha (2025). Medical image segmentation model based on local enhancement driven global optimization. Scientific Reports, 15, 18281. https://doi.org/10.1038/s41598-025-02393-1

Authors Contribution Statement

Vikash Verma: Conceptualization, Methodology, Investigation, Software, Validation, Formal analysis, Writing – Original Draft. Pritaj Yadav: Data Curation, Investigation, Writing – Review & Editing. Both authors have read and agreed to the published version of the manuscript.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Has this article screened for similarity?

Yes