AUT Journal of Electrical Engineering
10.22060/EEJ.2026.25612.5970

DASA-Net: A Hybrid Dual-Attention Swin-Transformer Architecture for Precise Segmentation of Brain Tumors from MRI

Noor Isam Abdulnabi[1], Mansoor Fateh[2,*], Saideh Ferdowsi[3]

[1]PhD student, Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran
[2]Associate Professor, Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran
[3]Assistant Professor, School of Mathematics, Statistics and Actuarial Science, University of Essex, Colchester CO4 3SQ, UK
* Corresponding author: mansoor_fateh@shahroodut.ac.ir

Uncorrected Proof

Abstract:

Brain tumor segmentation from Magnetic Resonance Imaging is a critical step for diagnosis and treatment planning, yet manual delineation is time-consuming and subjective. While deep learning models like Convolutional Neural Networks and Transformers have advanced automated segmentation, they face inherent limitations. CNNs struggle with long-range dependencies, while pure Transformers are often data-inefficient. To address these challenges, we propose DASA-Net, a novel hybrid Dual-Attention Swin-Transformer Architecture for precise brain tumor segmentation. Our model integrates a pre-trained Swin Transformer as the encoder to capture rich hierarchical and global contextual features. We introduce an Enhanced Spatial Attention module with residual connections to refine spatial details in the encoder's feature maps, and incorporate Squeeze-and-Excitation blocks in the decoder for adaptive channel-wise feature recalibration. Additionally, a Dynamic Hybrid Loss function is employed to balance Binary Cross-Entropy, Dice, and Focal losses, effectively addressing class imbalance and boundary ambiguity. Extensive experiments on the BRISC and Figshare brain tumor datasets demonstrate that DASA-Net achieves state-of-the-art performance, with a Dice score of 90.3% and IoU of 82.3% on BRISC, significantly outperforming existing methods.

Keywords:

Brain Tumor, MRI, Segmentation, Attention Mechanism

1. Introduction

A brain tumor is an abnormal proliferation of cells within the brain. These growths can be non-cancerous (benign) or cancerous (malignant), with malignant tumors like Glioblastoma being notoriously aggressive [1-3]. Globally, brain tumors represent a significant health burden. For instance, in the United Kingdom, over 12,000 individuals are diagnosed with a primary brain tumor annually [4]. A critical challenge in managing this disease is delayed diagnosis; reports indicate that a substantial percentage of brain cancers are identified only during emergency presentations, and many patients require multiple consultations to obtain a definitive diagnosis [5, 6]. Magnetic Resonance Imaging (MRI) serves as the gold standard for non-invasive diagnosis, offering superior soft-tissue contrast across multiple modalities (e.g., T1-weighted, T2-weighted, FLAIR) that reveal detailed pathological and anatomical information [7-9]. Manual delineation of tumor boundaries from these voluminous scans is the conventional approach for treatment planning. However, this process is exceptionally time-consuming, operator-dependent, and subject to significant inter-observer variability [10, 11]. Consequently, the development of robust, automated, and accurate segmentation systems is an urgent clinical need to improve diagnostic efficiency, aid in surgical planning, and effectively monitor therapeutic response [12, 13].

The advent of deep learning has revolutionized the field of medical image analysis, with automated segmentation models demonstrating performance that can meet or exceed clinical standards [14]. Initially, Convolutional Neural Networks (CNNs) became the dominant paradigm. Architectures based on the U-Net model, characterized by an encoder-decoder structure and skip connections, have been widely adopted and proven highly effective for various medical segmentation tasks, including brain tumors [12, 15]. More recently, Transformer models, which leverage self-attention mechanisms to capture global dependencies, have achieved state-of-the-art results in natural language processing and computer vision [16]. This success has catalyzed their adaptation for medical imaging, with models like the Swin Transformer demonstrating

Uncorrected Proof

a powerful capacity for modeling long-range contextual information within hierarchical feature maps [17, 18].

Despite their respective successes, both architectural classes possess inherent limitations. CNNs, by virtue of their convolutional kernels, excel at extracting local spatial features but are structurally limited in their ability to model long-range spatial dependencies and global context [16]. This "locality bias" can be detrimental when segmenting large, heterogeneous tumors with diffuse boundaries [19]. Conversely, pure Transformer models lack the innate inductive biases of CNNs, such as translation invariance and a locally-focused receptive field. As a result, they are often less data-efficient and may require extensive pre-training on large-scale datasets to achieve optimal performance, posing a challenge in the often data-scarce medical domain [16, 20]. Therefore, developing hybrid architectures that synergistically fuse the local feature extraction strength of CNNs with the global context modeling of Transformers remains a critical and highly active area of research [1, 16].

To address these challenges, we propose Dual-Attention Swin-Transformer Architecture (DASA-Net), a novel hybrid encoder-decoder network for precise brain tumor segmentation. Our model is meticulously designed to leverage the complementary strengths of Transformers and attention mechanisms within a unified framework. The architecture employs a pre-trained Swin Transformer as its encoder backbone, which generates rich, hierarchical feature representations. We introduce a novel Enhanced Spatial Attention (SPA) module, applied with a residual connection to each stage of the encoder's feature maps, to adaptively refine spatial information and focus the network on the most salient image regions. Furthermore, the decoder path is augmented with Squeeze-and-Excitation (SE) blocks, which perform channel-wise feature re-calibration during the upsampling process to selectively amplify the most informative features from the skip connections. To optimize the model, we utilize a Dynamic Hybrid Loss function that intelligently balances three distinct loss components to effectively manage severe class imbalance and focus on difficult boundary pixels [21].

Uncorrected Proof

our contribution summarized as follow:

1. The design of DASA-Net, a novel hybrid segmentation architecture, which integrates a fine-tuned Swin Transformer backbone as a powerful feature encoder.
2. The introduction of an Enhanced SPA module with a residual connection, which is applied to the encoder's multi-scale features to enhance spatial feature refinement.
3. The integration of SE blocks within the decoder to perform adaptive channel-wise feature re-calibration, optimizing the fusion of multi-scale features.
4. The implementation of a Dynamic Hybrid Loss function that adaptively weights three complementary loss components to improve segmentation accuracy, particularly for imbalanced tumor sub-regions.

The remainder of this paper is organized as follows: Section 2 reviews the related work in DR detection. Section 3 details the architecture of the proposed method. Section 4 presents the experimental setup and results. Finally, Section 5 concludes the paper and outlines future directions.

2. Related Work

The automated segmentation of brain tumors from MRI scans has been a prominent research area for decades. The methodologies have undergone a significant evolution, progressing from classical image processing techniques to the current state-of-the-art deep learning architectures. This section reviews this progression, categorized into four key areas: traditional and machine learning methods, foundational deep learning models, transfer learning-based approaches, and the emergence of Transformer and attention-based networks.

2-1- Traditional Methods and Machine Learning

Prior to the deep learning era, brain tumor segmentation relied on traditional image processing and classical machine learning models. Early approaches were intensity-based, such as thresholding and region-growing

Uncorrected Proof

[22]. These methods segment images by grouping pixels based on intensity values. However, they are highly sensitive to image noise, intensity inhomogeneity (common in MRI), and the poor contrast often found at tumor boundaries, making them unreliable for complex segmentation tasks [8].

More advanced methods employed unsupervised machine learning algorithms like K-Means clustering and Fuzzy C-Means (FCM). These techniques group pixels into a predefined number of clusters, but they primarily consider pixel intensity and fail to incorporate crucial spatial information, often resulting in noisy and fragmented segmentation masks [8]. A study by Bhimavarapu et al. [23] highlighted these shortcomings but also proposed an improved Fuzzy C-Means algorithm to enhance unsupervised segmentation, indicating that this area, while dated, is still being refined.

Beyond pixel-clustering techniques, boundary-driven classical approaches such as Active Contour Models (Snakes) and Level Set methods were widely investigated for capturing anatomical boundaries [22]. These energy-minimization paradigms deform a dynamic curve toward high-gradient image regions to delineate complex shapes. Despite their mathematical elegance and capability to yield smooth, continuous contours, they are prone to sub-optimal local minima under poor soft-tissue contrast, suffer heavily from manual initialization dependencies, and demand heavy iterative adjustments per MRI volume. This limitation severely hinders their capability to robustly handle highly heterogeneous brain tumors autonomously [23].

To address these limitations, supervised machine learning models were introduced, including SVMs [24] and Random Forests (RFs) [23, 25]. Unlike unsupervised methods, these models are trained on manually annotated data. However, their primary and most significant drawback is their reliance on hand-crafted feature engineering. This process requires domain experts to manually extract a set of discriminative features which are then fed to the classifier [25, 26]. This feature extraction step is not only time-consuming and subjective but also results in models that struggle to generalize to unseen data with different tumor appearances or imaging protocols.

Uncorrected Proof

2-2- Deep Learning Methods

The advent of deep learning, particularly Convolutional Neural Networks (CNNs), marked a paradigm shift. CNNs automatically learn a hierarchical cascade of discriminative features directly from the raw image data, eliminating the need for manual feature engineering [8]. The Fully Convolutional Network (FCN) was a seminal work that first proposed an end-to-end CNN for semantic segmentation, replacing the dense, fully-connected layers of classification networks with convolutions to produce a pixel-wise prediction map [27]. While revolutionary, FCNs often produced coarse segmentation maps due to the significant loss of spatial resolution in the encoding path.

The most influential architecture in medical image segmentation remains the U-Net [28]. The U-Net introduced a symmetric encoder-decoder architecture. The encoder (contracting path) captures context by progressively downsampling feature maps, while the decoder (expanding path) progressively upsamples these features to reconstruct a full-resolution segmentation map. The key innovation of U-Net is its use of skip connections, which concatenate feature maps from the encoder path with the corresponding feature maps in the decoder path [29]. This fusion allows the network to combine deep, semantic information (what) with shallow, high-resolution information (where), resulting in precise localization and excellent boundary definition [28].

Given that MRI data is inherently volumetric, 2D U-Net models that process scans slice-by-slice fail to capture the full 3D spatial context [8]. This led to the development of 3D variants, such as the 3D U-Net [30] and V-Net [31]. These 3D models have become the standard in brain tumor segmentation challenges like BraTS [8]. Numerous U-Net variants have since been proposed, such as U-Net++ for its use of nested and dense skip connections to improve gradient flow and feature fusion at multiple scales [32].

2-3- Transfer Learning in Segmentation

Uncorrected Proof

While foundational models like U-Net are powerful, their encoders are typically trained from scratch on the target medical dataset. Medical imaging datasets, however, are often small due to the high cost and time required for expert annotation. Training very deep networks on limited data can lead to overfitting and poor generalization.

Transfer learning emerged as a dominant strategy to overcome this [33, 34]. This approach leverages a deep CNN backbone that has been pre-trained on a large-scale natural image dataset, such as ImageNet [33, 34]. The pre-trained encoder acts as a highly effective feature extractor, having already learned a rich hierarchy of visual features. This backbone is then integrated into a U-Net-style encoder-decoder architecture, and the entire model is fine-tuned on the specific medical task [33, 34]. A study by Rastogi. D. et al. [34] demonstrated this by fine-tuning several models, finding that an Xception-based architecture achieved 96.11% accuracy. This approach leads to faster convergence, requires less training data, and often results in superior performance by capitalizing on the generalized features from the source domain.

2-4- Transformer-Based Methods and Attention Mechanisms

Despite their success, all CNN-based models, including those using transfer learning, have an inherent inductive bias: locality. The convolutional kernel operates on a small, local neighborhood, and the network must stack many layers to build a large receptive field. This makes it challenging for CNNs to model long-range spatial dependencies and global context, which is vital for understanding the full scope of a large, heterogeneous brain tumor [1, 35].

Transformers, which rely entirely on self-attention mechanisms, were introduced to overcome this limitation [36]. The Vision Transformer (ViT) first adapted this for image classification by splitting an image into patches and using self-attention to weigh the importance of every patch relative to every other patch, thereby capturing a global field of view from the very first layer [1].

Uncorrected Proof

For segmentation, hybrid models were proposed to combine the strengths of both architectures. TransU-Net was a pioneering example, which uses a CNN encoder to extract robust feature maps and then passes them through a Transformer in the bottleneck to model global context before upsampling with a CNN decoder. In parallel, automated self-configuring architectures such as nnU-Net emerged, setting major benchmarks by systematically optimizing preprocessing, patch sizes, and training hyperparameters within standardized standard U-Net frameworks without altering the deep core layers. While highly robust, pure CNN pipelines like nnU-Net fundamentally struggle with long-range relationship modeling, creating an open space for hybrid models like our proposed network [1, 37].

A more integrated approach is the Swin-U-Net [38], which builds a pure Transformer U-Net. It uses the Swin Transformer [39] as its core component. The Swin Transformer introduces a hierarchical design and computes self-attention within non-overlapping, shifted local windows. This makes it computationally efficient like a CNN, while the window-shifting mechanism allows for information to be exchanged across windows, ultimately achieving a global receptive field [38]. Recent studies confirm that Swin-U-Net and its variants are highly effective, merging the hierarchical structure of U-Net with the global context modeling of Transformers [38, 40].

In parallel, attention mechanisms have also been integrated as modules to enhance standard CNNs. The Attention U-Net [41] adds "attention gates" to the skip connections. These gates learn to spatially suppress irrelevant regions from the encoder's feature map before fusion with the decoder, forcing the model to focus on the regions of interest. Recent studies have consistently shown that incorporating attention mechanisms, such as the Convolutional Block Attention Module (CBAM) or other custom spatial and channel modules, significantly improves segmentation accuracy over baseline U-Net models [37, 41, 42]. For instance, Ga-U-Net [42], integrated Gabor convolutions and a CBAM module into a 3D U-Net to enhance texture information and boundary perception. Our proposed model builds directly upon these latter two trends, integrating a Swin Transformer backbone with novel spatial and channel attention modules.

Uncorrected Proof

3. Proposed Method

In this section, we present our novel architecture, DASA-Net, which is meticulously designed for the automated segmentation of brain tumors from MRI. We first formalize the segmentation task, then provide a high-level overview of the network's data flow, and finally detail each of its core components.

3-1- Problem Definition

The task addressed in this work is semantic segmentation, a pixel-level classification problem. Given a input brain MRI, $I \in R^{H\times W\times C}$, where H, W, and C represent the height, width, and input channels which is one or three, respectively, our objective is to train a deep neural network f with parameters θ. This network must produce a corresponding binary segmentation mask, $\hat{M} \in [0,1]^{H\times W}$, which accurately delineates the tumorous region. The model, defined as $\hat{M} = f(I;\theta)$, is optimized by minimizing a composite loss function $\mathcal{L}$ that quantifies the dissimilarity between the predicted mask $\hat{M}$ and the binary ground-truth mask $M_{gt} \in [0,1]^{H\times W}$. The formal optimization objective is to find the optimal parameters θ^* that minimize the loss, as shown in Eq. (1).

$$\theta^* = \arg\min \backslash\text{below}\,\theta \mathcal{L}\left(f(I;\theta), M_{gt}\right) \tag{1}$$

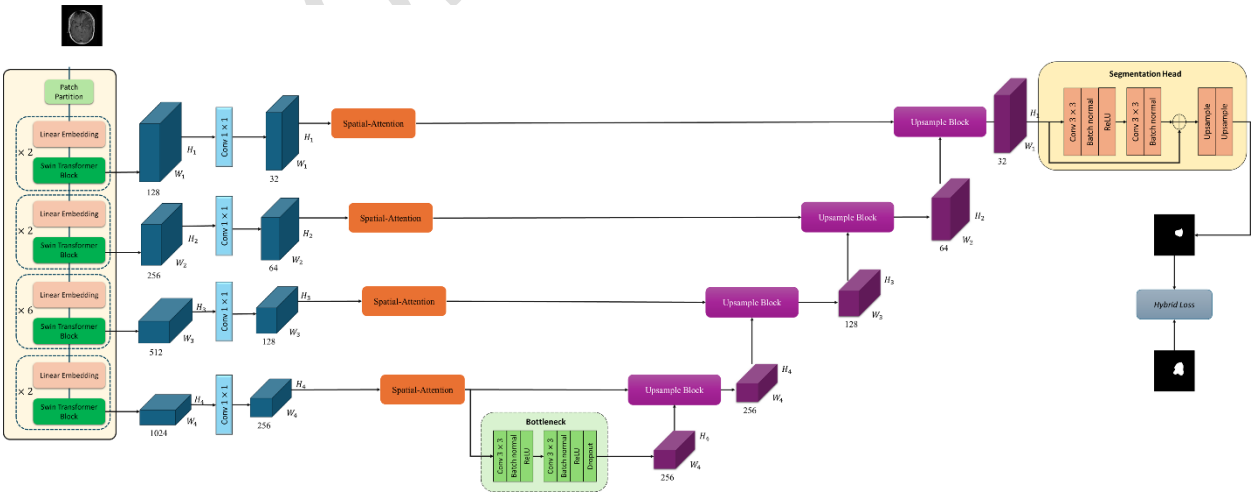

Fig. 1. The overall architecture of the proposed DASA-Net. The model employs a Swin Transformer encoder featuring Enhanced Spatial Attention (SPA) modules on the skip connections, coupled with a CNN-based decoder augmented by Squeeze-and-Excitation (SE) blocks within each decoder stage

3-2- Architectural Overview of DASA-Net

DASA-Net represents an advanced encoder-decoder network architecture, building upon the foundational U-Net topology while introducing significant modifications to enhance feature representation, global context modeling, and effective feature fusion. The designation "Dual-Attention" signifies the synergistic integration of two distinct and complementary attention mechanisms positioned at critical stages within the architecture. Firstly, a novel Enhanced Spatial Attention (SPA) module is embedded within the encoder path. This module is designed to spatially refine the feature maps generated by the encoder backbone before they are transmitted to the decoder via skip connections, thereby emphasizing salient spatial regions relevant to the segmentation task. Secondly, a Squeeze-and-Excitation (SE) Block is incorporated into each constituent block of the decoder. The purpose of the SE block is to perform adaptive channel-wise feature re-calibration during the fusion of upsampled feature maps from the preceding decoder stage and the corresponding feature maps from the encoder's skip connections, allowing the network to dynamically weigh the importance of different feature channels.

The overall architecture, illustrated schematically in Fig. (1), orchestrates a flow of information across four primary stages. The Encoder stage utilizes a pre-trained Swin Transformer to generate a hierarchy of feature maps at multiple scales. These feature maps are immediately processed by our proposed SPA-based refinement block to enhance their spatial fidelity. The refined features are then made available for the decoder path. The deepest feature map produced by the encoder first passes through a convolutional Bottleneck module, which functions as a bridge connecting the contracting (encoder) and expanding (decoder) paths. The Decoder stage is composed of four custom-designed decoder blocks. These blocks progressively upsample the feature maps originating from the bottleneck, integrating the corresponding

Uncorrected Proof

refined skip connection features from the encoder at each resolution level through an SE-augmented fusion strategy. Finally, a dedicated Segmentation Head module processes the high-resolution feature map from the final decoder block to generate the ultimate full-resolution binary segmentation mask. The entire network is trained end-to-end, guided by a novel Dynamic Hybrid Loss function specifically designed to address common challenges in medical image segmentation, such as class imbalance and boundary ambiguity.

To understand the architectural merit of DASA-Net, it is informative to contrast its layout against standard pure or hybrid Swin-Transformer models, such as Swin-U-Net. Conventional Swin-based networks rely heavily on symmetric shifted-window self-attention blocks to capture context but typically transfer encoder features directly to the decoder path through unrefined skip connections. In brain MRI analysis, this direct routing often floods the expanding path with background non-tumorous noise or structural scanning artifacts. DASA-Net departs fundamentally from this passive transmission paradigm. By inserting the Enhanced Spatial Attention (SPA) module directly inside the skip connections, the network enforces an active multi-scale spatial filter that refines target geometry before concatenation. Subsequently, the Squeeze-and-Excitation (SE) blocks within the decoder resolve feature conflicts arising from fusing abstract semantic maps with high-resolution localized features. This two-tiered gating mechanism guarantees that spatial details and channel distributions are meticulously managed at every scale, yielding a far more robust boundary definition than standard window-bound networks.

3-3- Encoder

The encoder, representing the contracting path of the network, is fundamentally responsible for extracting robust, discriminative, and multi-scale feature representations from the input MRI volumes. This hierarchical feature extraction provides the foundation for subsequent segmentation.

3-3-1 Hierarchical Transformer Backbone

Uncorrected Proof

Departing from conventional CNN-based encoders, we employ a hierarchical Vision Transformer, specifically a variant of the Swin Transformer architecture, as the primary feature extractor. The Swin Transformer is particularly well-suited for dense prediction tasks like segmentation. It effectively mitigates the quadratic computational complexity typically associated with standard ViTs applied to high-resolution images, while still retaining the capacity to capture long-range dependencies and global context. This is achieved through its core mechanism of computing self-attention within local, non-overlapping windows. Crucially, to enable information propagation across these localized windows, it utilizes a shifted window (SW-MSA) partitioning approach in successive layers. This strategy allows for cross-window connections, progressively building a hierarchical feature representation that efficiently models global context, analogous to the expanding receptive fields found in deep CNNs.

We initialize the Swin Transformer backbone with parameters pre-trained on a large-scale natural image dataset (ImageNet). Recognizing the domain shift between natural images and medical MRI scans, and aiming to leverage the powerful pre-trained features while adapting the model to the nuances of brain tumor segmentation, we implement a specific transfer learning strategy. We selectively fine-tune only the deeper stages of the transformer network, maintaining the parameters of the initial layers as fixed (frozen). This approach is motivated by the understanding that early layers in deep networks learn general, low-level features (such as edges, corners, and textures) which are broadly applicable across visual domains, while deeper layers capture more abstract, task-specific information. By freezing the initial layers, we retain the robust, generalized feature extraction capabilities learned from the extensive ImageNet dataset. Concurrently, fine-tuning the deeper layers allows the model to specialize its feature representations for the complex and specific morphological characteristics exhibited by various types of brain tumors and surrounding tissues in MRI. This backbone architecture produces four hierarchical feature maps, denoted as f_1, f_2, f_3, f_4, corresponding to different stages of the network, characterized by progressively decreasing spatial resolution and increasing channel depth.

Uncorrected Proof

Immediately following feature extraction by the Swin Transformer backbone, each output feature map f_i undergoes a dimensionality reduction step via a 1×1 convolution. This operation projects the features onto a lower-dimensional embedding space, yielding more compact representations denoted as $f_i^{'}$. This step serves two primary purposes: firstly, it significantly reduces the overall parameter count of the network, leading to a more lightweight model; secondly, it decreases the computational load for the subsequent attention mechanisms, allowing them to operate more efficiently without sacrificing significant representational capacity. These dimensionally-reduced feature maps $f_i^{'}$ are the ones passed forward to the SPA refinement modules and subsequently to the decoder via skip connections.

3-3-2 Enhanced Spatial Attention (SPA) Refinement

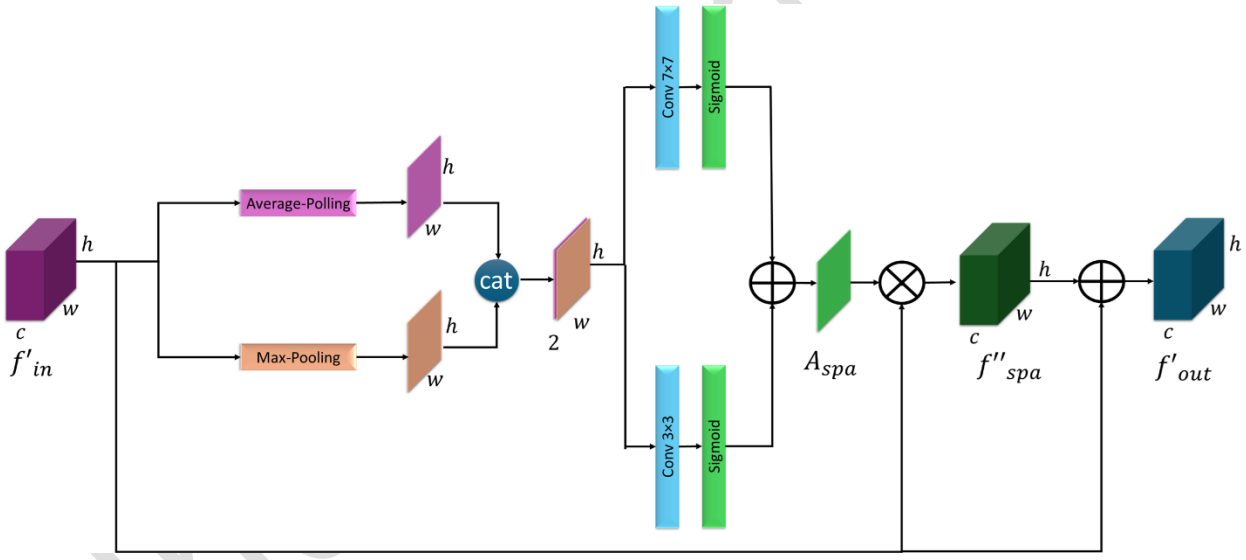

Fig. 2. The architecture of the Enhanced Spatial Attention (SPA) module. It utilizes parallel convolutions with differing kernel sizes applied to channel-pooled features (average and max) to generate a multi-scale spatial attention map

A significant innovation within the DASA-Net architecture is the refinement applied to the dimensionally-reduced feature maps ($f_i^{'}$) before they are propagated to the decoder via skip connections. Traditional spatial attention mechanisms typically employ a single convolutional layer over channel-pooled features,

restricting feature capture to a fixed receptive field. To overcome this limitation, we introduce a novel Enhanced Spatial Attention (SPA) refinement block (Fig. (2)) that constitutes a deliberate architectural variation optimized for capturing multi-scale structures. This module is engineered to simultaneously capture multi-scale spatial context, which is crucial for identifying tumor regions of varying sizes and shapes.

The core SPA module is specifically engineered to capture multi-scale spatial context, which is crucial for identifying tumor regions of varying sizes and shapes. It takes the reduced feature map $f_i^{'}$ as input. It begins by performing both average-pooling and max-pooling operations along the channel dimension. These pooling operations effectively compress the channel-wise information, producing two distinct spatial maps ($H_i \times W_i \times 1$) that summarize the average and maximum feature responses across channels at each spatial location, highlighting regions of general activation and peak activation, respectively. These two spatial summary maps are then concatenated channel-wise and fed into two parallel convolutional paths. A key aspect of our SPA design is the use of differing kernel sizes (e.g., 7×7 and 3×3) in these parallel paths. This multi-kernel approach enables the module to aggregate spatial context from different receptive fields simultaneously, capturing both broader regional information and finer local details. The outputs resulting from these two parallel convolutional paths are subsequently summed element-wise. This summation integrates the multi-scale spatial information. The combined map is then passed through a sigmoid activation function, which normalizes the values to the range [0, 1], producing the final spatial attention map A_{spa}. This attention map represents the learned spatial importance for each pixel. Finally, the attention map A_{spa} is applied to the input feature map $f_i^{'}$ through element-wise multiplication, as shown in Eq. (2).

$$f_{spa}^{''} = f_i^{'} \otimes A_{spa} \tag{2}$$

Uncorrected Proof

This operation selectively amplifies features in salient spatial locations (identified by higher attention weights) and suppresses features in less relevant regions.

To ensure stable training and preserve the original feature information while incorporating the learned spatial refinements, we employ a residual connection. The attention-refined map f''_{spa} is added back element-wise to the original, input map f'_i, according to Eq. (3).

$$f'_{out} = f''_{spa} + f'_i \tag{3}$$

This residual learning formulation allows the network to learn the spatial refinements as an additive modification, ensuring that the module primarily focuses on enhancing the feature representation rather than potentially degrading it. This design choice stabilizes the learning dynamics and guarantees that the rich features extracted by the Swin Transformer backbone (after dimensionality reduction) are effectively retained and utilized in the skip connections forwarded to the decoder.

3-4 Bottleneck Module

Following the final stage of the encoder and its associated SPA refinement, the resulting deepest feature map, f'_{out} (corresponding to the original f_4), is directed into a bottleneck module. This module serves as a crucial intermediary, bridging the contracting path (encoder) and the expanding path (decoder). Architecturally, the bottleneck consists of two sequential convolutional blocks. Each block comprises a 3×3 convolution operation, followed by Batch Normalization for stabilizing activations, and a Rectified Linear Unit (ReLU) activation function for introducing non-linearity. To enhance the model's generalization capability and mitigate potential overfitting, a dropout layer is applied subsequent to these convolutional blocks. Dropout acts as a regularization technique by randomly setting a fraction of neuron activations to zero during training, thus preventing excessive co-adaptation between neurons. This regularization is particularly beneficial at the bottleneck stage, where the feature representations are highly

Uncorrected Proof

abstract and semantically dense, making them potentially more susceptible to overfitting on the training data.

3-5 Decoder

The decoder, constituting the expanding path of the network, is tasked with progressively reconstructing the spatial resolution while systematically integrating the spatially-refined contextual information provided by the skip connections (the f'_{out} maps) from the encoder. This process is orchestrated through a sequence of four custom-designed Decoder Block modules, the internal structure of which is detailed in Fig. (3). Each Decoder Block executes a sophisticated feature fusion and refinement operation.

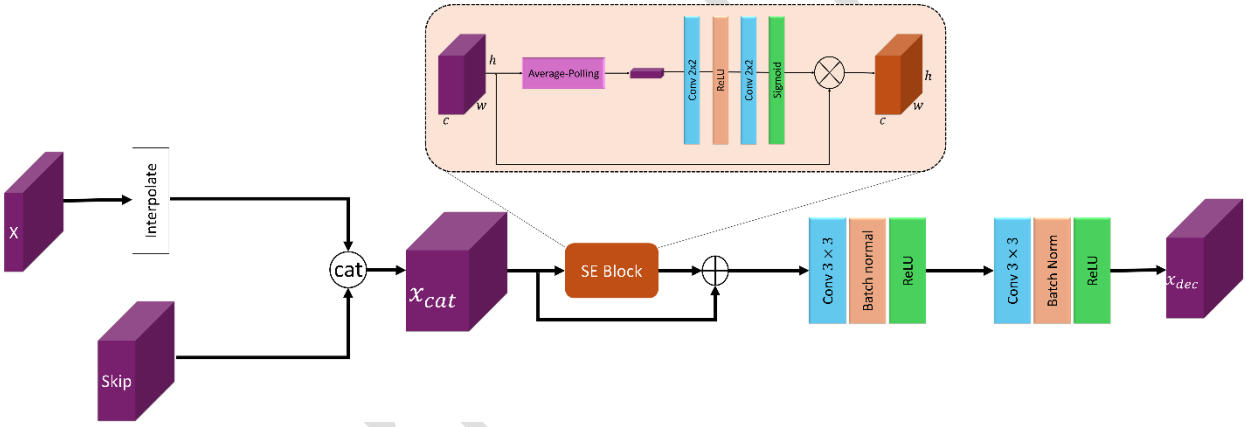

Fig. 3. The architecture of the custom Decoder Block. It illustrates the upsampling operation, concatenation with the corresponding skip connection, application of the Squeeze-and-Excitation (SE) block for channel attention, the subsequent residual connection, and final convolutional refinement layers.

The process within each decoder block begins by receiving input from two sources: the feature map from the preceding decoder stage and the corresponding refined skip-connection feature map (f'_{out}) from the encoder. For the deepest decoder block (directly following the bottleneck), both inputs possess the same spatial resolution, eliminating the need for initial upsampling. However, for all subsequent, shallower decoder blocks, the feature map received from the deeper stage has a lower resolution. Therefore, it is first upsampled by a factor of two, typically using bilinear interpolation, to match the spatial dimensions of the

corresponding skip connection map. This upsampled feature map is then concatenated along the channel dimension with the skip-connection feature map. This concatenation operation yields a composite feature map, x_{cat}, which synergistically combines high-level semantic information propagated through the decoder path with high-resolution spatial details preserved via the skip connection.

Immediately following concatenation, the x_{cat} feature map is processed by the second pillar of our dual-attention strategy: a Squeeze-and-Excitation (SE) block. The SE block implements dynamic, channel-wise attention, enabling the network to adaptively re-calibrate the importance of individual feature channels within the fused representation. The "Squeeze" operation employs global average pooling to aggregate spatial information across the entire feature map ($H \times W$) into a single channel descriptor vector ($1 \times 1 \times C$). This vector encapsulates the global response statistics for each channel. Subsequently, the "Excitation" operation passes this channel descriptor through a small feed-forward network, typically consisting of two 1×1 convolutional layers with an intermediate non-linearity, followed by a sigmoid activation function. This process learns a set of channel-specific modulation weights, ranging between 0 and 1, which signify the relative importance or relevance of each feature channel in the current context. The original concatenated feature map x_{cat} is then element-wise multiplied (scaled) by these learned channel weights. This adaptive re-calibration mechanism effectively amplifies features from channels deemed more informative while attenuating responses from less relevant channels.

Mirroring the design philosophy employed in the encoder's SPA refinement block, we incorporate a residual connection within the decoder block as well. The output of the SE block, x_{se} (which is x_{cat} scaled by the channel attention weights), is added element-wise back to the original concatenated map. This residual link facilitates improved gradient propagation during backpropagation and allows the decoder block to leverage both the raw fused features (x_{cat}) and the channel-refined features (x_{se}), thereby enhancing the overall representational capacity and feature robustness. Finally, the resulting feature map x_{dec} undergoes further processing through two consecutive standard convolutional blocks, each comprising a 3×3 convolution,

Uncorrected Proof

Batch Normalization, and ReLU activation. These blocks serve to further refine the features and prepare the output for the subsequent, shallower decoder stage.

3-6- Segmentation Head

The feature map d_1, produced by the final and shallowest decoder block, possesses the highest spatial resolution within the decoder path and contains a refined synthesis of multi-scale semantic and spatial information. This feature map is directed to the segmentation head module. The primary function of the segmentation head is to further refine these high-resolution features and ultimately project them into the final pixel-wise segmentation prediction map. Architecturally, it consists of two convolutional blocks, incorporating a residual connection to facilitate one last stage of feature refinement while ensuring stable gradient flow. Following these refinement blocks, a final 1×1 convolution layer is applied. This layer acts as a pixel-wise classifier, projecting the multi-channel feature map onto a single channel. The output of this 1×1 convolution represents the raw prediction scores, or logits, for each pixel belonging to the target class (tumor). Since the feature map at this stage typically has a lower spatial resolution than the original input image, a final upsampling step, usually employing bilinear interpolation, is applied to resize the logit map to match the original input dimensions ($H\times W$). This produces the final, full-resolution segmentation prediction.

3-7- Loss Function: Dynamic Hybrid Loss

The effective training of DASA-Net is guided by a novel Dynamic Hybrid Loss function, specifically formulated to address the inherent challenges encountered in medical image segmentation tasks. Standard loss functions often exhibit specific limitations when applied to this domain, characterized by significant class imbalance (e.g., small tumor regions within a large background volume) and the presence of ambiguous or poorly defined boundaries between different tissue types. Our approach combines three well-established and complementary loss functions: Binary Cross-Entropy (BCE), Dice Loss, and Focal Loss.

Uncorrected Proof

BCE provides a stable, pixel-wise measure of classification error but is known to be sensitive to class imbalance. Dice Loss, a region-based metric derived from the Dice Similarity Coefficient, directly optimizes for overlap and is inherently robust to class imbalance, making it suitable for segmenting small structures, although it can exhibit instability during training, particularly in early stages or with noisy predictions. Focal Loss modifies the standard BCE loss to down-weight the contribution of easily classified pixels (e.g., large background regions), thereby focusing the model's learning efforts on more challenging examples, such as pixels near tumor boundaries.

The key innovation of our Dynamic Hybrid Loss lies in its adaptive weighting strategy. Instead of relying on fixed, manually tuned hyperparameters to balance the contributions of L_{bce}, L_{dice}, and L_{focal}, our function calculates these weights dynamically during each training iteration. The process begins by computing the current values for each individual loss component based on the model's predictions and the ground-truth masks. The total loss, L_{total}, is then calculated as the sum of these individual losses (augmented with a small epsilon, $1e-8$, for numerical stability). Dynamic weights (e.g., α_{dice}) are subsequently computed by normalizing each individual loss component with respect to this total loss. For instance, the dynamic weight for the Dice loss component is determined as shown in Eq. (4).

$$\alpha_{dice} = \frac{L_{dice}}{L_{bce} + L_{dice} + L_{focal} + \epsilon} \tag{4}$$

Crucially, these dynamically computed weights are detached from the backpropagation computation graph. This ensures that they function purely as scaling factors for their respective loss terms and are not treated as trainable parameters themselves. The final composite loss, L_{final}, used to update the network parameters, is then calculated as a weighted sum employing these dynamic weights, as defined in Eq. (5).

Uncorrected Proof

$$\mathcal{L}_{final} = \alpha_{bce} \cdot L_{bce} + \alpha_{dice} \cdot L_{dice} + \alpha_{focal} \cdot L_{focal} \quad (5)$$

This dynamic weighting scheme confers a significant advantage by enabling the model to autonomously adapt its learning objective based on the current training state. If, for example, the model is primarily struggling with accurately segmenting the overall tumor region due to class imbalance (reflected by a relatively high L_{dice} value), the corresponding dynamic weight α_{dice} will automatically increase. This implicitly directs the optimization process to prioritize the improvement of overlap metrics. Conversely, if boundary definition becomes the main challenge (potentially indicated by a higher L_{focal}), its influence will grow accordingly. This self-adjusting mechanism fosters a more robust, stable, and ultimately more effective optimization process compared to using fixed loss weights.

4. Experimental Result

4-1- Dataset

To comprehensively evaluate the performance of our proposed DASA-Net, we utilize two distinct, publicly available 2D MRI datasets for brain tumor segmentation. The selection of these datasets allows for a thorough assessment of our model's effectiveness, robustness, and generalization capabilities across different data sources, imaging characteristics, and class distributions.

4-1-1- BRISC 2025 Dataset

The BRISC (BRain tumor Image Segmentation and Classification) dataset is a recent, large-scale collection of MRI scans designed to address limitations in existing datasets, such as class imbalance and annotation inconsistencies [43]. The dataset contains 6,000 T1-weighted contrast-enhanced MRI images, divided into a training set of 5,000 images and a testing set of 1,000 images. A key feature of this dataset is its comprehensive labeling for both segmentation and classification tasks, encompassing the three most

Uncorrected Proof

common brain tumor types, Glioma, Meningioma, and Pituitary as well as a "non-tumorous" category to help models distinguish between healthy and pathological scans.

The dataset was meticulously curated from several sources and underwent a rigorous re-annotation process using advanced tools and validation by certified radiologists and physicians to ensure the accuracy of the tumor masks [43]. Furthermore, the dataset is balanced across imaging planes, containing a near-uniform distribution of axial, sagittal, and coronal views. This multi-planar and multi-class composition, combined with its high-quality annotations, presents a challenging and realistic benchmark for evaluating robust segmentation models. For our experiments, we adhere strictly to the official training and testing splits provided with the dataset.

4-1-2- Figshare Brain Tumor Dataset

The second dataset utilized in our study is a widely-cited 2D brain tumor collection containing 3,064 T1-weighted contrast-enhanced MRI images, aggregated from 233 patients [44, 45]. This dataset includes three types of brain tumors: Meningioma (708 slices), Glioma (1426 slices), and Pituitary tumor (930 slices). The data was acquired at Nanfang Hospital and General Hospital, Tianjin Medical University, China, with an in-plane resolution of 512×512 pixels.

4-2- Experimental Setting

All experiments were conducted using the PyTorch deep learning framework. The training process was executed on the Kaggle platform, leveraging the computational power of a Tesla T4 GPU. For a consistent and fair evaluation across all models, a fixed set of hyperparameters was used for both the BRISC 2025 [43] and Figshare [44, 45] datasets, as summarized in Table (1). To ensure a completely rigorous and unbiased comparative evaluation, every network configuration and baseline model was trained independently 3 times using different random initializations. The metrics reported in the subsequent sections represent the mathematical average of these 3 runs. Notably, the standard deviation across all trials

Uncorrected Proof

remained consistently below 0.01%, demonstrating exceptional statistical stability and confirming that the performance gains are highly reproducible rather than artifacts of a specific seed selection.

Table 1. Experimental setting and hyperparameters.

Parameter	Value
Deep Learning Framework	PyTorch
Training Environment	Kaggle
GPU	Tesla T4
Input Image Size	512×512
Batch Size	8
Epochs	60
Learning Rate	1×10^{-4}

4-3- Evaluation Metrics

To quantitatively assess and compare the segmentation performance of our proposed DASA-Net architecture, we employ two standard and widely-used metrics in the field of medical image segmentation: the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU). Both metrics measure the spatial overlap between the predicted segmentation mask ($\hat{M}$) and the ground-truth mask (M_{gt}).

Dice Similarity Coefficient (DSC): This metric is a measure of overlap between two sets and is particularly effective for evaluating segmentation performance, especially in cases with significant class imbalance. As shown in Eq. (6), it is defined as twice the area of the intersection divided by the sum of the areas of the two masks. Values range from 0 (no overlap) to 1 (perfect overlap).

$$DSC\left(\hat{M},M_{gt}\right)=\frac{2\times\left|\hat{M}\ \backslash\mathrm{capM}_{gt}\right|}{\left|\hat{M}\right|+\left|M_{gt}\right|} \tag{6}$$

Intersection over Union (IoU): Also known as the Jaccard Index, IoU quantifies the overlap between the predicted mask and the ground-truth mask. As shown in Eq. (7), it is defined as the area of the intersection divided by the area of the union. Like DSC, its value ranges from 0 to 1.

Uncorrected Proof

$$IoU\left(\hat{M}, M_{gt}\right) = \frac{\left|\hat{M} \backslash \text{capM}_{gt}\right|}{\left|\hat{M} \backslash \text{cupM}_{gt}\right|} \tag{7}$$

For both metrics, a higher score indicates a better segmentation performance, with a score of 1.0 signifying a perfect match with the ground-truth segmentation.

4-4- Comparison with State-of-the-Art (SOTA)

The comparative results presented in Table (2) clearly demonstrate the superior performance of our proposed DASA-Net on the BRISC 2025 dataset.

Our model achieves the highest scores in the primary overall metrics, with a Weighted IoU of 82.3% and a DSC of 90.3%. This marks a significant improvement over all other evaluated architectures. The next-best performing model, SaberNet [46], achieved a Weighted IoU of 80.6% and a DSC of 88.6%. Our DASA-Net surpasses this by a considerable margin of 1.7% in both IoU and DSC, a substantial gain in the context of medical image segmentation.

As shown in Table (2), a key advantage of the BRISC 2025 dataset is its multi-class nature, providing distinct labels for Glioma, Meningioma, and Pituitary tumors. This allows for a more granular, per-class analysis of segmentation performance.

Table 2. Quantitative comparison with SOTA models on the BRISC 2025 dataset. The best results are highlighted in bold

Model	Glioma IoU (%)	Meningioma IoU (%)	Pituitary IoU (%)	Weighted IoU (%)	Dice (DSC) (%)
U-Net [28]	69.7	77.1	79.3	75.7	86.2
U-Net++ [47]	71.7	74.2	79.7	75.3	85.9
MANet [48]	72.4	77.5	78.0	76.2	86.5
LinkNet [49]	71.7	74.8	79.0	75.3	85.9
DeepLabV3+ [50]	72.0	77.5	78.7	76.3	86.5
PAN [51]	72.0	74.5	80.7	75.9	86.3
EINet [52]	73.6	78.4	80.3	77.7	87.4
EU-Net [53]	71.7	76.1	78.3	75.6	86.1
TransU-Net [37]	73.1	81.4	82.0	78.6	87.9

Uncorrected Proof

nnU-Net [2]	74.8	82.0	83.5	79.9	88.3
DAD [54]	**75.2**	80.4	82.3	79.5	88.1
BASNet [55]	74.0	77.5	81.7	77.9	87.5
SaberNet [46]	74.0	82.4	84.3	80.6	88.6
ABANet [56]	72.4	80.4	**84.7**	79.5	88.1
DASA-Net (Ours)	72.0	**91.4**	81.7	**82.3**	**90.3**

This detailed analysis suggests that while DASA-Net is robust across all tumor types, its novel dual-attention mechanism is particularly adept at identifying and delineating the features of meningiomas, which subsequently drives its superior weighted average score.

Furthermore, when compared to the foundational U-Net model (75.7% Weighted IoU, 86.2% DSC), DASA-Net shows a remarkable improvement of 6.6% in IoU and 4.1% in DSC. This highlights the limitations of previous SOTA models. The superior performance can be attributed to the powerful global context modeling of the Swin Transformer backbone, combined with the precision of our dual-attention mechanisms and the robust optimization provided by the Dynamic Hybrid Loss.

To further validate the robustness and generalization capability of our proposed DASA-Net, we conducted a comparative analysis on the widely-used Figshare Brain Tumor dataset [44, 45]. For this comparison, we evaluated DASA-Net against a variety of methods. These include established deep learning baselines and recent models.

Table 3. Quantitative comparison on the Figshare Brain Tumor dataset. The best results among the compared methods are in bold.

Model	IoU (%)	DSC (%)
U-Net (baseline)	53.46	69.68
U-Net (ResNet-34 backbone)	56.43	72.15
U-Net (ResNet-50 backbone)	48.62	65.43
U-Net + Attention Gate	61.79	76.39
U-NET-AG-MHA	69.60	82.08
MedCLIP-SAM	50.06	66.72
DASA-Net (Ours)	**78.30**	**87.83**

Uncorrected Proof

The results, summarized in Table (3), confirm that DASA-Net also achieves state-of-the-art performance on this second, challenging dataset. Our model's IOU of 78.30% and DSC of 87.83% represent a substantial lead over the next-best performing model, U-NET-AG-MHA, which scored 69.60% IOU and 82.08% DSC. This is a significant performance gap, with DASA-Net improving upon the next-best architecture by 8.7% in IOU and 5.75% in DSC.

This improvement is even more pronounced when compared to the foundational U-Net baseline. It is also noteworthy that while adding attention mechanisms to U-Net progressively improves performance, our model's scores indicate that its architectural design is far more effective. Furthermore, recent models like MedCLIP-SAM [57] struggle with this specific task, performing even below the U-Net baseline.

To thoroughly assess the real-world practicality and deployment feasibility of DASA-Net, we conducted an empirical complexity profiling of our network. All operational metrics were recorded under identical conditions utilizing a standard Tesla T4 GPU training environment with a uniform slice resolution of 512×512 pixels. Our proposed architecture possesses a parameter footprint of 119.02M, which is predominantly attributed to the deeper unfreezed layers (L_3 and L_4) of the hierarchical Swin Transformer backbone. During training, this layout achieves an efficient processing speed, requiring an average of 94.6 seconds per epoch. Crucially, during the testing phase, DASA-Net demonstrates high clinical viability by executing inference at an exceptional rate of 14.2 milliseconds per individual MRI slice. This rapid sub-20ms execution latency confirms that our dual-attention framework can handle voluminous clinical testing datasets practically instantaneously, making it highly suitable for active software deployment in radiology workflows.

4-5- Ablation Study

To validate the effectiveness of our proposed DASA-Net and analyze the contribution of each individual component, we conducted a comprehensive set of ablation studies on the BRISC 2025 dataset. We analyze

Uncorrected Proof

the choice of encoder backbone, the impact of trainable encoder layers, the additive contribution of our novel modules, the internal design of the SPA module, and the effectiveness of our hybrid loss function.

We first compared the performance of several popular pre-trained backbones. For this test, all backbone layers were kept frozen to evaluate only the feature extraction capability. As shown in Table (4), the Swin Transformer baseline achieved a DSC of 89.40%, significantly outperforming ResNet, EfficientNet, and MobileNet. This result confirms its powerful hierarchical feature representation for this segmentation task and justifies its choice as our encoder.

Table 4. Comparison of different encoder backbones (all layers frozen)

Backbone	IoU (%)	DSC (%)
ResNet	72.45	84.02
EfficientNet	75.65	86.14
MobileNet	75.75	86.20
Swin Transformer	**80.83**	**89.40**

After selecting the Swin Transformer, we explored the impact of unfreezing its layers during training to fine-tune the model, starting from the final stage. As detailed in Table (5), unfreezing only the final stage (L4) provided no benefit over a fully frozen model. However, unfreezing the last two stages (L3+L4) yielded the peak performance of 90.28% DSC and 82.29% IOU. Interestingly, unfreezing more layers (L2-L4 or L1-L4) led to a slight performance degradation, likely due to overfitting on the smaller dataset. Therefore, we adopted the setting of unfreezing L3 and L4 for our final DASA-Net configuration.

Table 5. Effect of unfreezing Swin Transformer layers during training

Unfrozen Layers	IoU (%)	DSC (%)	Params (M)
None (Frozen)	80.83	89.40	57.98
L4	80.83	89.40	97.95
L3, L4 (Ours)	**82.29**	**90.28**	119.02
L2, L3, L4	81.72	89.94	137.98
L1, L2, L3, L4	80.93	89.46	144.86

Uncorrected Proof

We conducted an exhaustive ablation study to validate the contribution of each proposed module, testing all combinations of components built upon the Swin Transformer encoder. The results are presented in Table (6).

The baseline model, consisting of only the encoder, achieved a DSC of 75.72%. Individual components provided substantial improvements: adding the SPA Decoder increased DSC to 76.25%, while incorporating the SE Block yielded a more significant gain to 80.20%. Most notably, the combination of both modules demonstrated strong synergistic effects, achieving the peak performance of 88.65% DSC and 82.29% IOU.

Table 6. Ablation study on the core components of DASA-Net. All models use the Swin Transformer encoder

SPA Decode	SE Block	IoU (%)	DSC (%)
		65.03	75.72
✓		65.68	76.25
	✓	71.26	80.20
✓	✓	**82.29**	**88.65**

The notable performance gains observed when combining the SPA and SE modules stem from their complementary optimization paths across the spatial and channel dimensions. As demonstrated by the quantitative configurations in Table 6, utilizing the hierarchical transformer backbone on its own provides a strong semantic baseline but lacks optimal fine-grained reconstruction capabilities. When the Enhanced SPA module is introduced, its parallel multi-scale convolutions function as an active spatial filter on the skip links. This configuration preserves sharp boundary configurations and ensures that localized positional information is not degraded by background scanning artifacts. Conversely, the SE block handles channel-wise feature re-calibration within the decoder path. By dynamically scaling the most discriminative feature channels during the fusion of upsampled maps and skip features, the SE block eliminates representational conflicts. The synergy of these two components establishes a balanced network behavior: the SPA module dictates precise target alignment along complex tumor borders, while the SE block refines the relevant

Uncorrected Proof

semantic content passed through the expanding path, explaining the distinct jump in overall Dice and IoU metrics.

To justify the design of our Spatial Pyramid Attention (SPA) module, we tested its internal components. The SPA module uses parallel 7×7 and 3×3 convolutions to capture multi-scale spatial features. As shown in Table (7), using only the 7×7 conv or only the 3×3 conv resulted in lower performance. The proposed parallel combination of both 7×7 and 3×3 convolutions achieved the best result (90.28% DSC), validating our multi-scale design.

Table 7. Ablation study on the components of the SPA module

SPA Configuration	IoU (%)	DSC (%)
Conv 7×7 only	80.29	89.07
Conv 3×3 only	79.91	88.83
Parallel $7\times7+3\times3$ (Ours)	**82.29**	**90.28**

Finally, we conducted a rigorous ablation study to validate the design choices of our optimization objective by evaluating individual loss components against both static and dynamic fusion strategies. As summarized in Table 8, training the architecture using individual components yields suboptimal performance due to their inherent structural vulnerabilities to severe class imbalance or boundary ambiguity. Combining these three distinct terms into a standard Static Hybrid Loss with uniform, fixed weights provide a baseline improvement to 89.78% DSC. However, our proposed Dynamic Hybrid Loss achieves the peak performance of 90.28% DSC and 82.29% IoU. This noticeable improvement demonstrates that omitting manual, heuristic tuning in favor of an online, data-driven adaptive normalization scheme effectively focuses the network on the most demanding learning challenges-balancing pixel-level accuracy, regional overlap, and hard boundary elements dynamically throughout the training process.

Table 8. Effect of the Hybrid Loss function

Loss Function	IoU (%)	DSC (%)
Cross Entropy only	81.05	89.53
Dice Loss Only	79.42	88.35

Uncorrected Proof

Focal Loss only	80.14	88.91
Static Hybrid Loss (Fixed Weights)	81.60	89.78
Dynamic Hybrid Loss (Ours)	**82.29**	**90.28**

4-6 Limitations and Disadvantages

While DASA-Net achieves state-of-the-art segmentation accuracy across multiple benchmarks, it exhibits certain inherent architectural and practical limitations that must be acknowledged:

- Computational Footprint: The integration of a hierarchical Swin Transformer backbone alongside multi-scale spatial pyramid attention (SPA) and channel re-calibration (SE) blocks naturally increases the network's parameter volume and computational complexity. This results in longer training times and higher GPU memory allocation demands compared to traditional, lightweight CNN baselines.
- Dependency on Structured Pre-training: Although the transfer learning strategy mitigates the data-inefficiency of pure transformers, the deeper unfrozen layers remain highly sensitive to domain shifts and require standardized, high-quality expert annotations to prevent sub-optimal convergence.
- Lack of Volumetric Modeling: The proposed framework is strictly designed as a 2D architecture that processes MRI volumes on a slice-by-slice basis. Consequently, it cannot inherently capture inter-slice spatial dependencies along the orthogonal axes, which can be critical for delineating highly irregular, deep-seated isotropic tumor volumes.

4-7 Qualitative Evaluation

To supplement our quantitative findings, we introduce a qualitative visual comparison in Fig. (4), illustrating the segmentation masks generated by DASA-Net alongside representative high-performing baselines, namely ABANet, SaberNet, and BASNet, across challenging MRI slices.

Uncorrected Proof

As visually apparent in Fig. (4), baseline models face noticeable difficulties when handling tumors with highly irregular geometries or diffuse boundaries. Specifically, BASNet exhibits a tendency to under-segment subtle tumor boundaries, while SaberNet and ABANet occasionally introduce false positive regions within non-tumorous background tissue due to complex imaging artifacts. Conversely, DASA-Net demonstrates exceptional boundary fidelity, capturing intricate morphological variations precisely and aligning almost perfectly with the expert ground-truth annotations. This visual precision confirms that combining multi-scale SPA with channel-wise decoder re-calibration successfully filters out background noise while preserving highly localization-sensitive boundary details.

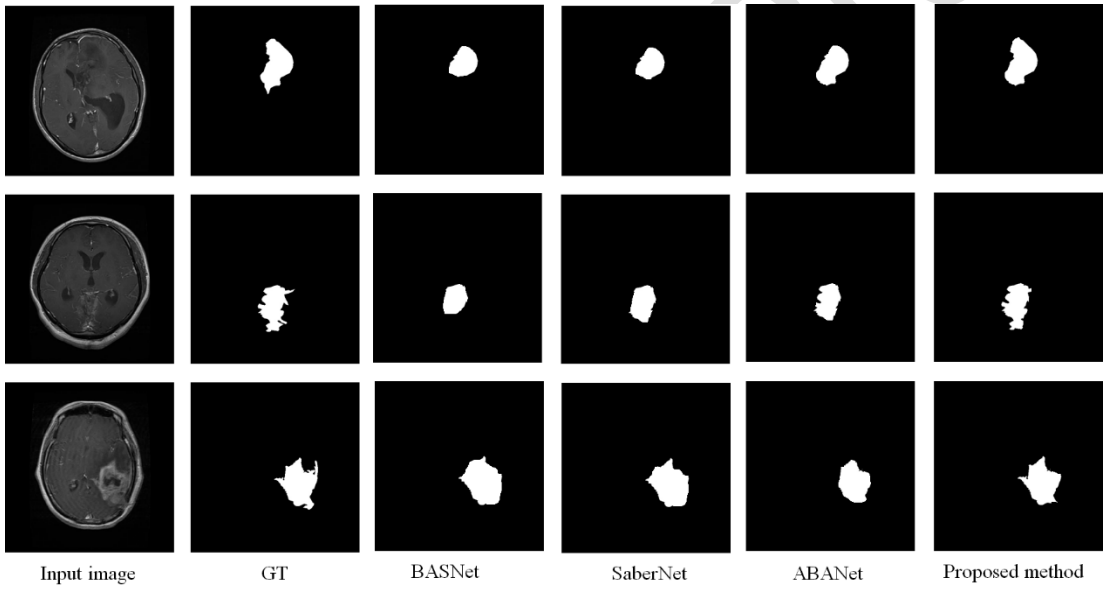

Fig. 4. Qualitative segmentation comparison on representative brain tumor MRI slices. Columns from left to right display the original input MRI scan, the ground-truth annotation, and the prediction masks generated by BASNet, SaberNet, ABANet, and our proposed DASA-Net, respectively.

5. Conclusion

This paper presented DASA-Net, a novel hybrid dual-attention architecture for precise brain tumor segmentation in MRI. By integrating a Swin Transformer encoder with our proposed Enhanced Spatial Attention (SPA) modules and a decoder augmented with Squeeze-and-Excitation (SE) blocks, DASA-Net

effectively captures both global context and local details. The model is optimized by a Dynamic Hybrid Loss that adaptively handles class imbalance. Extensive experiments on the BRISC 2025 and Figshare datasets demonstrate that DASA-Net achieves state-of-the-art performance, significantly outperforming existing methods. Ablation studies confirm the critical contribution of each component.

From a practical standpoint, DASA-Net offers high potential across several clinical applications. First, its sub-20ms inference latency can serve as an active automated screening tool in high-volume radiology centers, reducing diagnostic bottlenecks by immediately highlighting regions of interest. Second, the exceptional boundary fidelity demonstrated by our dual-attention framework is uniquely suited for pre-operative surgical planning, offering neurosurgeons a precise, non-invasive digital template to maximize tumor resection boundaries while preserving healthy brain parenchyma. Finally, the deterministic, automated nature of the model eliminates inter-observer variability, providing a standardized baseline to accurately quantify tumor regression or progression during longitudinal follow-ups in chemotherapy and radiotherapy regimes.

Future work will focus on expanding the architecture along several distinct technical paths. We plan to extend the multi-scale attention mechanisms natively into 3D space to process full, volumetric isotropic MRI grids directly. Additionally, we intend to investigate post-training network quantization and knowledge distillation to develop a compressed, highly efficient edge deployment module capable of running inside low-compute institutional servers. Lastly, we will explore multi-modal inputs to integrate multi-sequence protocols (such as FLAIR, T1ce, and T2) alongside functional imaging sequences, further bolstering the framework's robustness against complex, diffuse tumor phenotypes.

Uncorrected Proof

References

[1] J. Greg, Bridged CNN--Transformer Architectures for Improved Brain Tumor Segmentation in MRI Scans, (2025).

[2] M. Kharaji, H. Abbasi, Y. Orouskhani, M. Shomalzadeh, F. Kazemi, M. Orouskhani, Brain tumor segmentation with advanced nnU-Net: pediatrics and adults tumors, Neuroscience Informatics, 4(2) (2024) 100156.
[3] S.K. Agrawal, I.P. Dubey, A.K. Nair, A. Jain, A. Mahato, R. Kumar, Neuroimaging informatics framework for analysing rare brain metastasis patterns in pleural mesothelioma using hybrid PET CT, Neuroscience Informatics, (2025) 100207.
[4] T.B.T. Charity, Source K: Brain tumour statistics, in, 2023.
[5] M. Soltani-Gol, A. Asgharzadeh-Bonab, H. Soltanian-Zadeh, J. Mazlum, RDCU-Net: A Multi-Scale Residual Dilated Convolution U-Net with Spatial Pyramid Pooling for Brain Tumor Segmentation, AUT Journal of Electrical Engineering, 56(2) (2024) 203-212.
[6] M. Ahangar, E. Najafiaghdam, A. Gharibi, Diagnosing Cancerous Tumor Using Photoacoustic Imaging and Thermoacoustic Tomography with Electric Excitation Techniques—A Numerical Approach for Comparison, AUT Journal of Electrical Engineering, 54(2 (Special Issue)) (2022) 409-420.
[7] A. Kotte, V.K. Prasad, Hybrid 3D U-Net and Attention Mechanisms for Whole Heart Segmentation from CT Images, Engineering, Technology \& Applied Science Research, 15(2) (2025) 21822--21828.
[8] M.M. Saleh, M.E. Salih, M.A.A. Ahmed, A.M. Hussein, From traditional methods to 3d u-net: A comprehensive review of brain tumour segmentation techniques, Journal of Biomedical Science and Engineering, 18(1) (2025) 1--32.
[9] A. Deshpande, V.V. Estrela, A. Jude, J. Hemanth, Special issue on computational intelligence in neuroinformatics: technologies and data analytics, in: Neuroscience Informatics, 2025, pp. 100187.
[10] R. Li, Y. Liao, Y. Huang, X. Ma, G. Zhao, Y. Wang, C. Song, DeepGlioSeg: advanced glioma MRI data segmentation with integrated local-global representation architecture, Frontiers in Oncology, 15 (2025) 1449911.
[11] Mohana, S. Sowmiya, Vinieth, S. Savitha, S. Mohanapriya, K. Dinesh, Optimizing Brain Tumor Segmentation Using Attention U-NET and ASPP U-NET, in: Proceedings of the 3rd International Conference on Futuristic Technology - Volume 2: INCOFT, SciTePress, 2025, pp. 223-228.
[12] V. Swathi, K. Sinduja, V.R. Kumar, A. Mahendar, G.V. Prasad, B. Samya, Deep learning-based brain tumor detection: An MRI segmentation approach, in, 2024, pp. 01157.
[13] B. Gupta, G.K. Jegannathan, M.S. Alam, K.S. Yogi, J.V.N. Ramesh, V.J. Sowmya, I. Bayhan, Multimodal Lightweight Neural Network for Alzheimer's Disease Diagnosis Integrating Neuroimaging and Cognitive Scores, Neuroscience Informatics, (2025) 100218.
[14] C. Diana-Albelda, l. Garca-Martn, J. Bescos, A Review on Deep Learning Methods for Glioma Segmentation, Limitations, and Future Perspectives, Journal of Imaging, 11(8) (2025) 269.
[15] A.S. Pillai, Utilizing Deep Learning in Medical Image Analysis for Enhanced Diagnostic Accuracy and Patient Care: Challenges, Opportunities, and Ethical Implications, Journal of Deep Learning in Genomic Data Analysis, (2021).
[16] X. Liu, J. Tian, S. Huang, W. Shen, Enhancing medical image segmentation via complementary CNN-transformer fusion and boundary perception, Frontiers in Computer Science, 7 (2025) 1677905.
[17] A. Nur, K. Nurhanafi, E.R. Putri, Brain Tumor Segmentation in MR Images Using Swin Transformer, Atom Indonesia, 51(2) (2025).
[18] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in, 2022, pp. 205--218.
[19] T. Griebel, A. Archit, C. Pape, Segment anything for histopathology, arXiv preprint arXiv:2502.00408, (2025).
[20] R.A. Zeineldin, M.E. Karar, Z. Elshaer, J. Coburger, C.R. Wirtz, O. Burgert, F. Mathis-Ullrich, Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI, Scientific reports, 14(1) (2024) 3713.
[21] W. Yang, Z. Li, C. Du, S.K.K. Chow, HLNet: high-level attention mechanism U-Net++ for brain tumor segmentation in MRI, Applied Intelligence, 55(7) (2025) 692.
[22] R.C. Gonzalez, Digital image processing, 2009.

Uncorrected Proof

[23] U. Bhimavarapu, N. Chintalapudi, G. Battineni, Brain tumor detection and categorization with segmentation of improved unsupervised clustering approach and machine learning classifier, Bioengineering, 11(3) (2024) 266.
[24] S. Bauer, L.-P. Nolte, M. Reyes, Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization, in, 2011, pp. 354--361.
[25] T.A. Soomro, L. Zheng, A.J. Afifi, A. Ali, S. Soomro, M. Yin, J. Gao, Image segmentation for MR brain tumor detection using machine learning: a review, IEEE Reviews in Biomedical Engineering, 16 (2022) 70--90.
[26] A. Ellwaa, A. Hussein, E. AlNaggar, M. Zidan, M. Zaki, M.A. Ismail, N.M. Ghanem, Brain tumor segmantation using random forest trained on iteratively selected patients, in, 2016, pp. 129--137.
[27] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in, 2015, pp. 3431--3440.
[28] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in, 2015, pp. 234--241.
[29] M.A.-A. Al-Murshidawy, O. Al-Shamma, A review of deep learning models (U-Net architectures) for segmenting brain tumors, Bulletin of Electrical Engineering and Informatics, 13(2) (2024) 1015--1030.
[30] z. iek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: learning dense volumetric segmentation from sparse annotation, in, 2016, pp. 424--432.
[31] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in, 2016, pp. 565--571.
[32] C. Makwana, N. Nayak, M. Sagar, Brain tumor segmentation using U-Net++ with dense connection, Multidisciplinary Science Journal, 6(12) (2024) 2024264--2024264.
[33] S.M. Rasa, M.M. Islam, M.A. Talukder, M.A. Uddin, M. Khalid, M. Kazi, M.Z. Kazi, Brain tumor classification using fine-tuned transfer learning models on magnetic resonance imaging (MRI) images, Digital health, 10 (2024) 20552076241286140.
[34] D. Rastogi, P. Johri, M. Donelli, L. Kumar, S. Bindewari, A. Raghav, S.K. Khatri, Brain tumor detection and prediction in MRI images utilizing a Fine-Tuned transfer learning model integrated within deep learning frameworks, Life, 15(3) (2025) 327.
[35] B. Hou, S. Guan, Brain tumor segmentation using deep learning: A Review, Journal of Computing and Electronic Information Management, 16(1) (2025) 26--32.
[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, u. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems, 30 (2017).
[37] X. Chen, L. Yang, Brain tumor segmentation based on CBAM-TransUNet, in, 2022, pp. 33--38.
[38] M.a.M. DIVYA, MOHAMED ANAS and N, MOHAMED ARSATH, Tumor Spotting Using Swin UNET, International Journal of Creative Research Thoughts (IJCRT), (2025).
[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in, 2021, pp. 10012--10022.
[40] Z. Liao, H. Peng, T. Liu, Brain tumor segmentation based on improved swin-UNet, in, 2023, pp. 222--225.
[41] B.M. Mulani, Multi-Modal Brain Tumor Segmentation with Attention Mechanisms, (2025).
[42] Y. Cao, Y. Song, New Approach for Brain Tumor Segmentation Based on Gabor Convolution and Attention Mechanism, Applied Sciences, 14(11) (2024) 4919.
[43] A. Fateh, Y. Rezvani, S. Moayedi, S. Rezvani, F. Fateh, M. Fateh, V. Abolghasemi, Brisc: Annotated dataset for brain tumor segmentation and classification with swin-hafnet, arXiv preprint arXiv:2506.14318, (2025).
[44] J. Cheng, W. Huang, S. Cao, R. Yang, W. Yang, Z. Yun, Z. Wang, Q. Feng, Enhanced performance of brain tumor classification via tumor region augmentation and partition, PloS one, 10(10) (2015) e0140381.

Uncorrected Proof

[45] J. Cheng, W. Yang, M. Huang, W. Huang, J. Jiang, Y. Zhou, R. Yang, J. Zhao, Y. Feng, Q. Feng, et al., Retrieval of brain tumors by adaptive spatial pooling and fisher vector representation, PloS one, 11(6) (2016) e0157112.
[46] A. Saber, P. Parhami, A. Siahkarzadeh, M. Fateh, A. Fateh, Efficient and accurate pneumonia detection using a novel multi-scale transformer approach, arXiv preprint arXiv:2408.04290, (2024).
[47] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in, 2018, pp. 3--11.
[48] T. Fan, G. Wang, Y. Li, H. Wang, Ma-net: A multi-scale attention network for liver and tumor segmentation, IEEE Access, 8 (2020) 179656--179665.
[49] A. Chaurasia, E. Culurciello, Linknet: Exploiting encoder representations for efficient semantic segmentation, in, 2017, pp. 1--4.
[50] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in, 2018, pp. 801--818.
[51] H. Li, P. Xiong, J. An, L. Wang, Pyramid attention network for semantic segmentation, arXiv preprint arXiv:1805.10180, (2018).
[52] C. Li, G. Jiao, EINet: camouflaged object detection with pyramid vision transformer, Journal of Electronic Imaging, 31(5) (2022) 053002--053002.
[53] K. Patel, A.M. Bur, G. Wang, Enhanced u-net: A feature enhancement network for polyp segmentation, in, 2021, pp. 181--188.
[54] J. Li, W. He, H. Zhang, Towards complex backgrounds: A unified difference-aware decoder for binary segmentation, arXiv preprint arXiv:2210.15156, (2022).
[55] X. Qin, D.-P. Fan, C. Huang, C. Diagne, Z. Zhang, A.C. Sant'Anna, A. Suarez, M. Jagersand, L. Shao, Boundary-aware segmentation network for mobile and web applications, arXiv preprint arXiv:2101.04704, (2021).
[56] S. Rezvani, M. Fateh, H. Khosravi, ABANet: Attention boundary-aware network for image segmentation, Expert Systems, 41(9) (2024) e13625.
[57] T. Koleilat, H. Asgariandehkordi, H. Rivaz, Y. Xiao, Medclip-sam: Bridging text and image towards universal medical image segmentation, in, 2024, pp. 643--653.

Uncorrected Proof