Cloud-Net: A Lightweight Convolutional Framework with Spectral-spatial Feature Enhancer for Real-time Brain Tumor Segmentation

Jyoti Kataria [1,2,*] and Supriya P. Panda [1]

[1] Department of Computer Science and Engineering, School of Engineering and Technology, Manav Rachna International Institute of Research and Studies (MRIIRS), Faridabad, 121004, India
[2] Department of Computer Science and Engineering, School of Engineering and Technology, K.R. Mangalam University, Sohna Road, Gurugram, 122103, India
Email: kataria.jyoti87@gmail.com (J.K.); supriya.set@mriu.edu.in (S.P.P.)
*Corresponding author

***Abstract*—Accurate brain tumor segmentation supports clinical decision-making, including treatment planning and surgical navigation, yet manual delineation is time-consuming and labor-intensive. Many high-accuracy models remain difficult to deploy under latency and memory constraints, while multimodal Magnetic Resonance Imaging (MRI) introduces severe foreground-background imbalance. Cloud-Net is a lightweight U-shaped convolutional framework that places a Spectral-Spatial Feature Enhancer (SFFE) at the bottleneck, combining a spatial convolution branch with Fourier-domain feature modulation to strengthen global structural cues while preserving local boundary detail. The model is optimized with a hybrid Dice-Focal objective and evaluated on the Brain Tumor Segmentation 2020 (BraTS 2020) dataset using T1-weighted (T1), contrast-enhanced T1-weighted (T1ce), T2-weighted (T2), and Fluid-Attenuated Inversion Recovery (FLAIR) MRI sequences. On tumor present slices, Cloud-Net achieves a Dice Similarity Coefficient (DSC) of 0.7701 and an Intersection over Union (IoU) of 0.6825, with a 95th-percentile Hausdorff Distance (HD95) of 20.21 pixels, while maintaining real-time efficiency at 2.73 ms/slice and 365.64 Frames Per Second (FPS) using 9.35 million parameters (35.69 MB). Ablation analysis indicates that SFFE improves overlap and boundary quality, and Gradient-weighted Class Activation Mapping (Grad-CAM) highlights lesion-relevant regions influencing predictions. These results support Cloud-Net as a practical option for real-time brain tumor segmentation in cloud-edge-oriented workflows.**

***Keywords*—imaging, lightweight convolutional neural network, spectral-spatial fusion, gradient-weighted class activation mapping**

I. Introduction

Accurate brain tumor segmentation from Magnetic Resonance Imaging (MRI) is a foundational step in clinical workflows, supporting diagnosis, treatment planning, surgical navigation, and longitudinal assessment of disease progression. However, manual delineation remains time-consuming and subject to inter- and intra-observer variability, particularly when tumors exhibit heterogeneous appearance, infiltrative growth, and ambiguous boundaries. These difficulties are amplified in gliomas, where tumor morphology can be highly irregular and where millimeter-level boundary precision can influence downstream clinical decisions.

Multi-modal MRI is routinely used because complementary sequences emphasize different tissue characteristics and tumor subregions. In brain tumor benchmarks such as the Brain Tumor Segmentation (BraTS) series, the standard set of modalities T1, contrast-enhanced T1 (T1ce), T2, and FLAIR provides complementary information that benefits automated delineation, especially for edema and enhancing regions [1, 2]. At the same time, BraTS-style data reflect practical challenges that persist in clinical imaging, including severe foreground-background imbalance and high variability across scanners, acquisition protocols, and patient populations [1, 2]. These characteristics motivate segmentation frameworks that are not only accurate, but also robust and efficient under realistic constraints.

Deep learning has become the dominant approach for medical image segmentation. Fully convolutional designs introduced end-to-end dense prediction for semantic segmentation [3], and U-Net established an encoder-decoder paradigm with skip connections that

Manuscript received January 5, 2026; revised January 30, 2026; accepted March 24, 2026; published July 21, 2026.

remains a cornerstone in biomedical segmentation due to its effective multi-scale feature fusion [4]. Subsequent architectures have incorporated encoder-decoder refinements and attention mechanisms to improve focus on clinically relevant regions and boundary detail [5]. Despite this progress, two recurring limitations are especially relevant for brain tumor segmentation. First, convolutional networks are strongly biased toward local receptive fields, which can hinder the modeling of long-range dependencies needed to maintain global shape consistency and reduce boundary ambiguity; this has motivated explicit global-context mechanisms such as non-local operations and attention [6]. Second, while attention/Transformer-based models can represent global context through self-attention [7] and have been adapted to medical segmentation [8], the computational and memory costs of attention particularly with high-resolution inputs often challenge practical deployment, especially when low latency and constrained resources are required.

These deployment constraints are increasingly important as imaging pipelines integrate cloud and edge computing to support scalable analysis, telemedicine workflows, and near-real-time decision support under hardware and bandwidth limitations [9, 10]. In such settings, segmentation models must achieve a favorable accuracy efficiency trade-off: strong boundary accuracy (which is clinically sensitive) while maintaining small model size and fast inference. Lightweight backbones such as MobileNetV2 and EfficientNet reduce computation through architectural choices like depthwise separable convolutions and compound scaling [11, 12], but aggressive downsampling and reduced representational capacity can degrade fine-grained boundary recovery precisely where tumor heterogeneity and partial-volume effects already make segmentation difficult. In addition, the extreme class imbalance common in tumor segmentation can bias optimization toward background predictions unless the learning objective explicitly emphasizes hard examples and overlap-based structure [13, 14].

Motivated by these observations, we propose Cloud-Net, a lightweight encoder-decoder segmentation framework designed for efficient brain tumor delineation while explicitly improving boundary sensitivity. The central idea is to strengthen global shape and boundary cues without incurring the high computational cost typical of self-attention. To this end, Cloud-Net introduces a Spectral-Spatial Feature Enhancer (SFFE) at the bottleneck of a U-shaped architecture. The SFFE fuses (i) a spatial convolution pathway that preserves local texture detail and (ii) a frequency-domain pathway based on the Fourier transform, which provides a principled mechanism for capturing global structure through spectral representations [15] and has been explored as an efficient alternative to attention-like global mixing in other learning settings [16]. Cloud-Net is further optimized with a hybrid objective combining Dice-based overlap optimization with focal reweighting to mitigate class imbalance and emphasize difficult boundary pixels [13, 14]. Beyond accuracy, the framework is evaluated with boundary-sensitive metrics as well, computational indicators (parameter count, model size), and latency criteria that directly reflect cloud/edge feasibility.

The main contributions of this work are summarized as follows:

(1) We present Cloud-Net, a lightweight U-shaped encoder-decoder network for multimodal brain tumor segmentation that explicitly targets an accuracy-efficiency trade-off suitable for cloud/ edge-oriented inference on slice-wise MRI.
(2) We introduce a SFFE integrated at the bottleneck, which combines a spatial convolution pathway with a Fourier-domain modulation pathway (FFT → spectral gating → IFFT) to inject global structure cues while preserving local boundary detail, avoiding the computational burden of full self-attention.
(3) We optimize Cloud-Net using a hybrid Dice + Focal objective to directly improve region overlap under severe foreground-background imbalance and to emphasize hard-to-segment pixels near tumor margins.
(4) We perform extensive evaluation covering overlap and boundary quality (DSC/IoU and HD95), efficiency profiling (parameter count, model size, and inference latency), ablation with paired statistical testing, robustness under missing-modality settings, and interpretability analysis using Grad-CAM to verify lesion-focused decision patterns.

The remainder of this paper is organized as follows: Section II reviews related work on brain tumor segmentation, lightweight architectures, and global-context modeling. Section III details Cloud-Net and the proposed SFFE module. Section IV describes the dataset, preprocessing, implementation details, and evaluation metrics. Section V reports quantitative and qualitative results, efficiency analyses, ablation and statistical testing, explainability, robustness under missing modalities, and clinical-oriented volumetry validation. Section VI concludes the paper and outlines future directions.

II. Related Work

Brain tumor segmentation from multi-modal MRI has progressed rapidly with deep learning, yet the field continues to face practical challenges: (i) achieving reliable boundary delineation under heterogeneous appearance and low-contrast margins, (ii) modeling long-range/global context without prohibitive computational overhead, and (iii) maintaining deployable efficiency for latency- or resource-constrained settings. This section reviews representative directions CNN-based segmentation, lightweight/real-time designs, and global-context/frequency-domain learning and summarizes their advantages and limitations in Table I.

TABLE I. COMPARATIVE SUMMARY OF RECENT STUDIES ON BRAIN TUMOR SEGMENTATION

Ref.	Model / Core idea	Dim.	Global-context / boundary mechanism	Frequency-domain component	Reported efficiency (as stated)	Reported dataset & task	Key reported outcomes (examples)	Main advantages	Main limitations motivating Cloud-Net
[4]	U-Net encoder-decoder with skip connections	2D	Multi-scale fusion via skips	None	-	General biomedical / BraTS variants	Strong baseline for localization	Simple, effective multi-scale fusion	Limited global context; boundary ambiguity under weak contrast
[17]	3D U-Net (volumetric U-Net)	3D	Volumetric context via 3D convs	None	-	BraTS-style volumetric	Typically stronger volumetric consistency	Uses inter-slice context directly	High memory/compute cost; slower deployment on constrained hardware
[18]	LR-Net (resource-friendly 3D network)	3D	Modules for receptive field + boundary sensitivity; lightweight goal	None explicitly (spatial-domain modules)	4.72M params; 990.91 GFLOPs	BraTS2020 (WT/TC/ET)	Dice-mean 0.881; HD-mean 9.00 (BraTS2020)	Competitive accuracy/size trade-off	3D compute can remain high; deployment cost still non-trivial in real-time constraints
[19]	FE-Unet	2D (multi-modal fusion)	Boundary/region evaluation via DSC/HD; patient-level split	Not central (spatial-domain focus)	Input cropped to 160 × 160 × 4 (reported)	BraTS2018/2019/2020 (WT/TC/ET)	BraTS2020 Dice: WT 78.95, TC 76.82, ET 68.02; HD (mm): WT 2.7599, TC 1.9777, ET 3.0597	Uses controlled patient-level split; reports multiple metrics	Not designed around explicit low-cost global mixing; 2D formulation may miss volumetric continuity
[20]	Frequency-feature U-shaped model with MFDF	2D	Emphasizes boundary detail via multi-frequency directional filtering	Lifting wavelets + FFT-based separation in skip connections	—(future work notes parameter reduction)	Public datasets (incl. BraTS mention)	Claims improved DSC/HD95 vs prior methods (as stated)	Directly targets boundary complexity with frequency cues	Efficiency profiling and real-time constraints not always central; parameter reduction remains a stated goal
[21]	HFF-Net (frequency-aware 3D framework)	3D	Frequency-domain cross-attention + edge/texture refinement	DTCWT + NSCT + FDCA + ALC	—(discussion emphasizes efficiency balance, but further reduction desirable)	Multi-modal 3D MRI; ET emphasis	Motivated by boundary limitations when frequency cues are omitted	Systematic LF/HF modeling; strong boundary/texture motivation	Notes robustness issues under incomplete modalities and remaining compute constraints
[22]	BTIS-Net (efficient 3D U-Net variant)	3D	Pareto/frontier analysis across metrics	None central	-	BraTS 2019/2021 (reported)	Strong boundary-related ranking (HD) in their metric trade-off analysis	Explicitly studies metric/complexity trade-offs	Shows metric trade-offs; 3D efficiency constraints remain challenging

A. CNN-Based Brain Tumor Segmentation

Fully Convolutional Networks (FCNs) enabled end-to-end dense prediction, and U-Net's encoder-decoder with skip connections remains a dominant backbone because it preserves spatial detail while learning semantic abstraction across scales [3, 4]. For brain tumors, the multi-class BraTS formulation (WT/TC/ET) motivates models that can cope with extreme class imbalance and ambiguous borders, especially for enhancing tumor regions with limited spatial extent.

To improve volumetric consistency, 3D extensions (e.g., 3D U-Net, V-Net) process volumes directly and typically improve anatomical coherence by exploiting through-plane context [13, 17]. However, volumetric processing increases memory/compute costs substantially, making training/inference expensive on high-resolution MRI volumes. This has encouraged refinements that focus on better feature fusion and boundary recovery.

Recent architectures often augment encoder-decoder backbones with attention or feature-selection mechanisms to suppress irrelevant background responses and emphasize lesion structures. A representative example is BTIS-Net, an efficient 3D U-Net variant that analyzes parameter-performance trade-offs and reports strong boundary-related behavior (e.g., favorable Hausdorff ranking in their comparative analysis), while highlighting that different metrics can exhibit different trade-offs across models. More broadly, these CNN-family methods demonstrate that boundary quality depends not only on model capacity but also on how multi-scale information and edge-sensitive cues are preserved through downsampling and reconstruction [3, 4, 13, 17, 18].

B. Lightweight/Real-Time Segmentation for Resource-Constrained Deployment

Despite strong accuracy, many segmentation networks are difficult to deploy in clinical pipelines that require low latency, small memory footprint, and stable throughput. Lightweight design strategies include compact channel schedules, depthwise/separable convolutions, and reduced-resolution inference; compression methods (e.g.,

pruning, quantization, knowledge distillation) are also widely studied to reduce the deployment cost [11, 19].

A key challenge is that aggressive downsampling and limited capacity can blur tumor margins precisely the regions most sensitive to clinical decisions. This has led to lightweight designs that attempt to retain boundary fidelity while constraining parameters and compute. LR-Net exemplifies this direction: it targets a balance between segmentation performance and model size, reporting 4.72M parameters and 990.91 GFLOPs (their profiling setup) while achieving Dice-mean = 0.881 with HD-mean = 9.00 on BraTS2020 (multi-class evaluation). These results illustrate an important nuance: "lightweight" in parameter count does not always translate to low compute, especially in 3D settings where volumetric operators can still be expensive [11, 19–24].

C. Global-Context Modeling and Frequency-Domain Learning

A core tension in segmentation is balancing local detail (needed for sharp boundaries) with global context (needed for anatomical plausibility and disambiguation). Transformers and attention mechanisms model long-range dependencies via self-attention [7], but attention can be computationally demanding, especially for high-resolution or volumetric inputs [20]. Consequently, alternatives that provide global mixing at lower overhead are actively explored.

Frequency-domain learning is a prominent complementary direction. Frequency representations can encode global structure compactly while also separating coarse (low-frequency) morphology from fine (high-frequency) textures and edges. Recent work argues that omitting frequency-domain cues constrains feature expressiveness and can limit a network's capacity to delineate fine-grained tumor boundaries. This perspective has motivated systematic frequency-aware segmentation frameworks. HFF-Net, for example, proposes combining Dual-Tree Complex Wavelet Transform (DTCWT) and Non-Subsampled Contourlet Transform (NSCT) for complementary low-/high-frequency feature extraction, together with frequency-domain attention and edge-refinement modules (e.g., FDCA, ALC). At the same time, this line of work notes remaining limitations relevant to real-world settings, including robustness under incomplete modalities and the need to further reduce computational demands for resource-limited deployment.

A related family of approaches integrates wavelets and Fourier transforms into U-shaped backbones to better preserve boundary details. For instance, a frequency-feature U-shaped segmentation model integrates lifting wavelets in downsampling and an FFT-based, multi-frequency directional edge extraction module in skip connections, arguing that the frequency module enhances fine-grained boundary delineation and reduces background interference while explicitly framing the remaining need to reduce parameters further for practical deployment [7, 8, 25–28].

Table I highlights three consistent patterns. First, volumetric (3D) networks (e.g., 3D U-Net [17]) naturally exploit inter-slice context, often improving spatial coherence, but they typically incur high memory and computational costs that complicate real-time or edge deployment. Even approaches optimized for parameter efficiency can remain compute-heavy: LR-Net reports 4.72 M parameters but still profiles 990.91 GFLOPs, illustrating that 3D operations can dominate runtime despite a compact parameter budget.

Second, 2D slice-wise frameworks (e.g., FE-Unet [26]) often offer favorable efficiency and easier training/inference scaling, but they can miss explicit volumetric continuity. FE-Unet reports multi-class BraTS2020 Dice values (WT/TC/ET) and HD in mm under a patient-level split protocol, yet its global-context modeling is still bounded by spatial-domain operations unless additional context mechanisms are introduced.

Third, frequency-domain methods increasingly target boundary ambiguity and texture heterogeneity, explicitly arguing that ignoring frequency cues can limit fine-grained boundary delineation. While comprehensive frequency-aware designs (e.g., HFF-Net) propose systematic LF/HF modeling with frequency-domain attention and edge refinement, these works also acknowledge remaining obstacles relevant to real-world deployment, including missing-modality robustness and further compute reduction needs. Other U-shaped frequency-feature models integrate wavelets/FFT into down sampling and skip connections to enhance boundary detail, but still identify parameter reduction as an open practical objective.

The surveyed literature indicates a gap at the intersection of global context, boundary sensitivity, and real-time efficiency. Heavy self-attention and volumetric processing improve context but raise deployment cost; aggressively compact CNNs can lose boundary fidelity; and richer frequency-domain designs can add complexity or require further optimization for constrained inference. Cloud-Net addresses these limitations by introducing an FFT-based spectral-spatial bottleneck module that injects global mixing through frequency-domain modulation while preserving local boundary detail via a spatial branch, positioned at the bottleneck to control computational overhead, and paired with lightweight, slice-wise inference suited to low-latency deployment.

III. Proposed Method

Cloud-Net is designed as an end-to-end pipeline for multimodal MRI brain tumor segmentation, targeting deployment scenarios where computation, memory, and data transfer can be constrained (e.g., edge-to-cloud healthcare workflows) [9, 10]. In brain tumor imaging, multi-modal MRI volumes are high-dimensional, and transmitting full 3D volumes can be bandwidth-intensive; therefore, we adopt a 3D-to-2D strategy that converts volumetric scans into 2D axial slices to reduce per-sample payload and enable high-throughput inference. A 2D design also provides favorable latency and GPU memory behavior compared with volumetric 3D processing, where 3D convolutions and volumetric tensors typically increase memory footprint and reduce feasible batch sizes. Intermediate 2.5D strategies (stacking adjacent slices) can

introduce limited through-plane context, but they increase input channels and computation with the number of neighboring slices and can be sensitive to variability in through-plane spacing; for a real-time objective, the slice-wise formulation provides a more reliable efficiency profile while preserving clinically meaningful tumor appearance cues in multi-modal MRI [25].

Given a BraTS subject with four MRI modalities (T1, T1ce, T2, FLAIR), we stack the modalities into a 4-channel tensor. BraTS is widely used as a benchmark for brain tumor segmentation and provides expert-annotated tumor masks for training and evaluation [1, 2]. Each subject volume is normalized using z-score normalization to reduce scanner- and acquisition-dependent intensity variations, a standard practice for MRI learning pipelines in BraTS-style settings [1]. To further improve efficiency and reduce the proportion of non-informative samples, we focus training on the central slice range where tumor presence is more likely, limiting exposure to empty background-only slices and accelerating optimization in class-imbalanced conditions. During training, 2D slices are resized to a fixed spatial resolution and augmented using geometric transformations (e.g., horizontal flips and small rotations) to improve generalization under morphological variability, which is a common challenge in tumor MRI segmentation due to heterogeneity and irregular boundaries [1, 2]. This early fusion allows convolutional layers to learn cross-modality interactions through channel mixing; SFFE further strengthens the fused representation at the bottleneck by combining global spectral mixing with local spatial refinement and learnable cross-channel fusion.

Although BraTS labels multiple tumor subregions, Cloud-Net in this work is configured for binary segmentation (tumor vs. background) by merging all non-zero tumor labels into a single foreground mask. This design choice aligns with the lightweight deployment objective: Reducing output dimensionality simplifies optimization and inference while still supporting clinically relevant whole-tumor delineation, which is frequently reported as a primary region of interest in benchmark evaluations [1]. This configuration targets whole-tumor delineation, which is widely used for tumor-burden estimation and contouring-oriented workflows. Sub-region delineation (e.g., WT/TC/ET) provides finer clinical characterization and can be addressed by extending the output to multiple classes. Sub-region delineation can support treatment stratification and response assessment; extending the prediction head to multiple classes enables WT/TC/ET outputs using the same backbone.

A. Cloud-Net Architecture

Cloud-Net follows a U-shaped encoder-decoder paradigm with skip connections (see Fig. 1), inheriting the key design principle of U-Net: hierarchical downsampling for semantic abstraction and symmetric upsampling for spatial reconstruction, with skip pathways to preserve localization detail [4]. The network consumes a 4-channel multimodal input and outputs a pixel-wise tumor probability map.

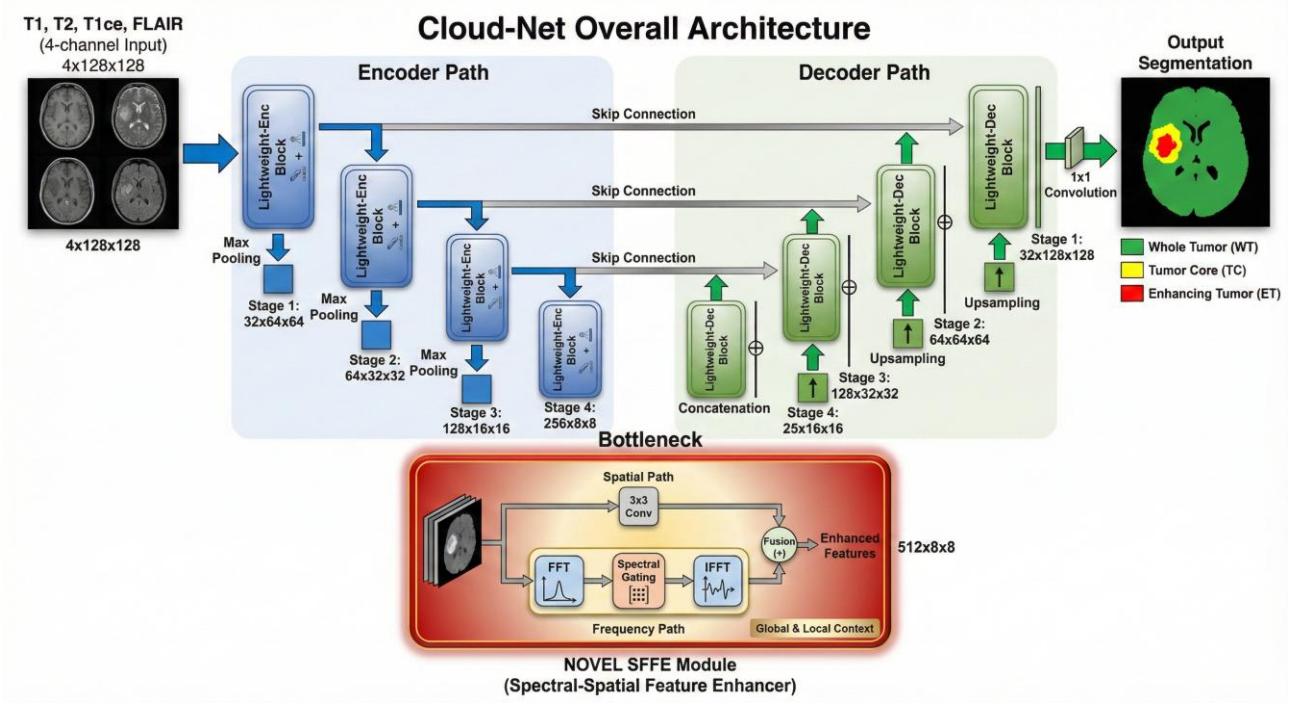

Fig. 1. The Cloud-Net overall architecture. A U-shaped lightweight convolutional framework integrating the SFFE module at the bottleneck for robust brain tumor segmentation.

The encoder consists of four stages. Each stage applies a convolutional feature extraction block followed by max pooling for downsampling. Channel width increases progressively with depth to encode higher-level tumor semantics while controlling computational cost via a compact channel schedule. This design is consistent with

the broader encoder-decoder literature in medical segmentation, where multi-scale feature fusion is essential to handle tumors of varying sizes and irregular morphology [4, 5].

After the deepest encoder stage, Cloud-Net applies a bottleneck feature transformation followed by the proposed Spectral-Spatial Feature Enhancer (SFFE) (Section III.C). Positioning SFFE at the bottleneck enables global spectral mixing at the smallest spatial resolution, improving shape coherence and boundary recovery with minimal runtime overhead. Placing the enhancer at the bottleneck is motivated by the observation that deep features carry the most abstract semantic representation, and improving global consistency at this stage can be beneficial for boundary recovery during decoding an issue highlighted in recent brain tumor segmentation works that emphasize the difficulty of capturing both local boundary detail and global semantic context simultaneously [26, 27].

The decoder mirrors the encoder with transposed convolutions for upsampling and convolutional refinement blocks. Skip connections concatenate encoder feature maps with decoder features at the corresponding resolution, enabling the network to recover fine spatial structure lost during pooling [4]. The final prediction layer is a 1 × 1 convolution that projects feature into the segmentation space, followed by a sigmoid activation for binary segmentation.

Cloud-Net's efficiency is achieved primarily through (i) a compact channel configuration and (ii) the use of parameter-efficient operations inside the proposed enhancer. Lightweight CNN design principles often rely on decomposing standard convolutions into depthwise and pointwise operations to reduce parameter count and computation, as exemplified by MobileNet-family architectures. Fig. 2 illustrates this efficiency principle: depthwise separable convolution reduces parameters compared with standard $K\times K$ convolutions while retaining representational capacity, which motivates our emphasis on efficient blocks and reduced-width design.

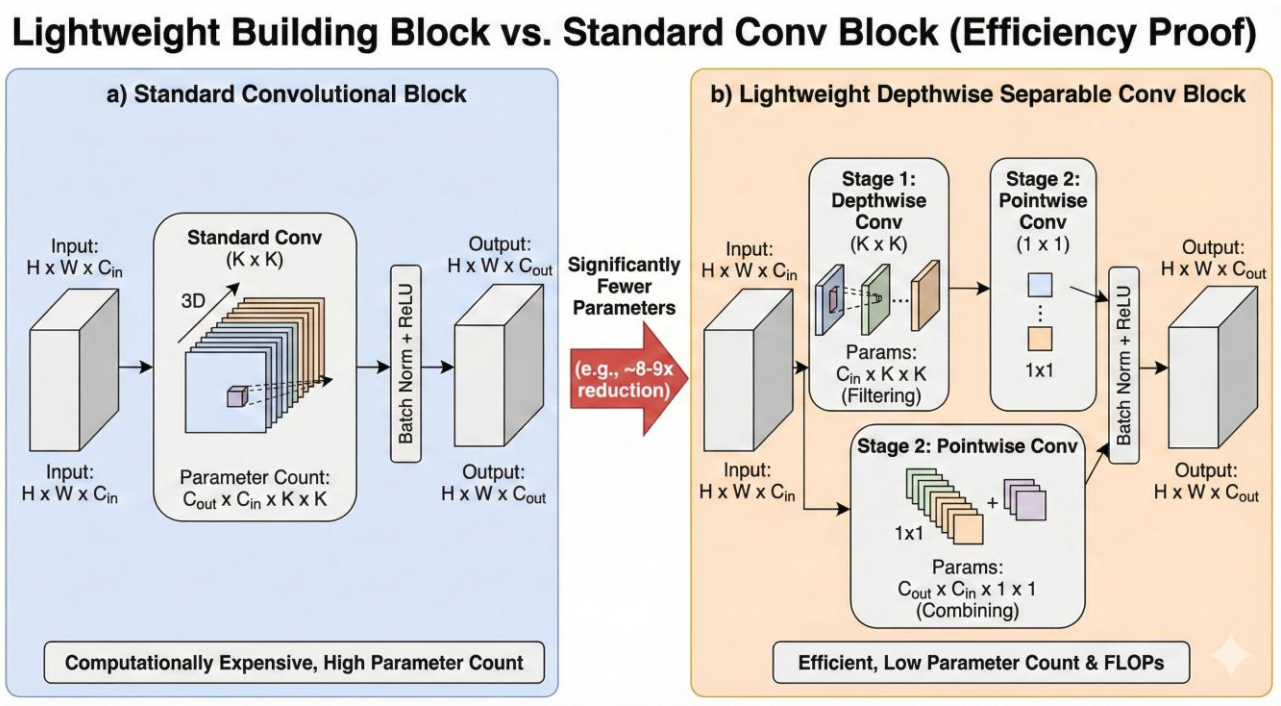

Fig. 2. The lightweight building block.

B. Spectral-Spatial Feature Enhancer (SFFE)

A major challenge in brain tumor segmentation is maintaining boundary precision under heterogeneous intensity patterns and irregular tumor shapes. CNNs excel at learning local texture cues but are fundamentally limited by local receptive fields; consequently, capturing long-range/global context remains challenging, motivating hybrid mechanisms that integrate global modeling while preserving local detail [26, 28]. Transformer-based approaches can improve global dependency modeling via self-attention [7, 8], but their computational cost can be prohibitive for lightweight or real-time deployment [28]. Cloud-Net addresses this gap by introducing a Fourier-based spectral pathway that provides global mixing with comparatively low overhead, paired with a conventional spatial convolution pathway for local refinement.

Let the bottleneck input feature map be $X \in \mathbb{R}^{B\times C\times H\times W}$. The SFFE module (see Fig. 3) computes two parallel transformations:

1) Spatial branch (local detail)

A convolutional operator extracts local patterns:

$$F_s = f_s(X)$$

where $f_s(\cdot)$ is a 3 × 3 convolutional transformation (implemented with normalization and nonlinearity). In our implementation, this branch uses a depthwise formulation to reduce parameter overhead, reflecting lightweight design principles established in efficient CNNs [11].

2) Frequency branch (global structure)

We transform the feature map into the frequency domain using a 2D Fourier transform:

Let $X \in \mathbb{R}^{B\times C\times H\times W}$. A 2D Fourier transform is applied to obtain complex spectral coefficients:

$$\mathcal{F}(X) = \mathrm{FFT2}(X) \in \mathbb{C}^{B\times C\times H\times W}$$

The transform is applied independently to each channel of the feature tensor; cross-channel mixing is introduced by the fusion operation using concatenation followed by a 1 × 1 projection. Spectral gating is implemented as element-wise modulation of the spectrum with a learnable frequency mask:

$$\mathcal{F}'(X) = \big(1 + \sigma(G)\big) \odot \mathcal{F}(X)$$

where $G \in \mathbb{R}^{1\times C\times H\times W}$ is a learnable parameter tensor, $\sigma(\cdot)$denotes the sigmoid function, and $\odot$ is element-wise multiplication. The gated spectrum is mapped back to the spatial domain:

$$F_f = \mathrm{IFFT2}(\mathcal{F}'(X))$$

and the real component is used for subsequent fusion. This formulation provides channel-wise and frequency-wise control because the mask contains separate parameters for each channel and each spatial frequency location [15].

This construction follows the general idea that Fourier-domain mixing provides a global interaction mechanism without explicit attention matrices, which has been used as an efficient alternative for token mixing in other deep learning contexts [16], while remaining grounded in classical Fourier analysis [15].

Finally, the spatial and frequency outputs are fused to produce an enhanced feature representation:

$$Y = \phi(F_s, F_f)$$

where $\phi(\cdot)$ denotes feature aggregation. Conceptually, Fig. 3 illustrates element-wise fusion; in the implemented module, fusion is realized by concatenation followed by a 1 × 1 projection to allow a learnable linear combination of spatial and spectral cues. The module operates on intermediate feature maps and does not depend on the number of output classes, so it can be integrated into multi-class encoder-decoder designs by adapting the prediction head and loss. This enables SFFE to preserve local texture sensitivity while injecting global shape/boundary consistency an objective aligned with recent segmentation studies emphasizing the need to balance local detail and global semantics [26].

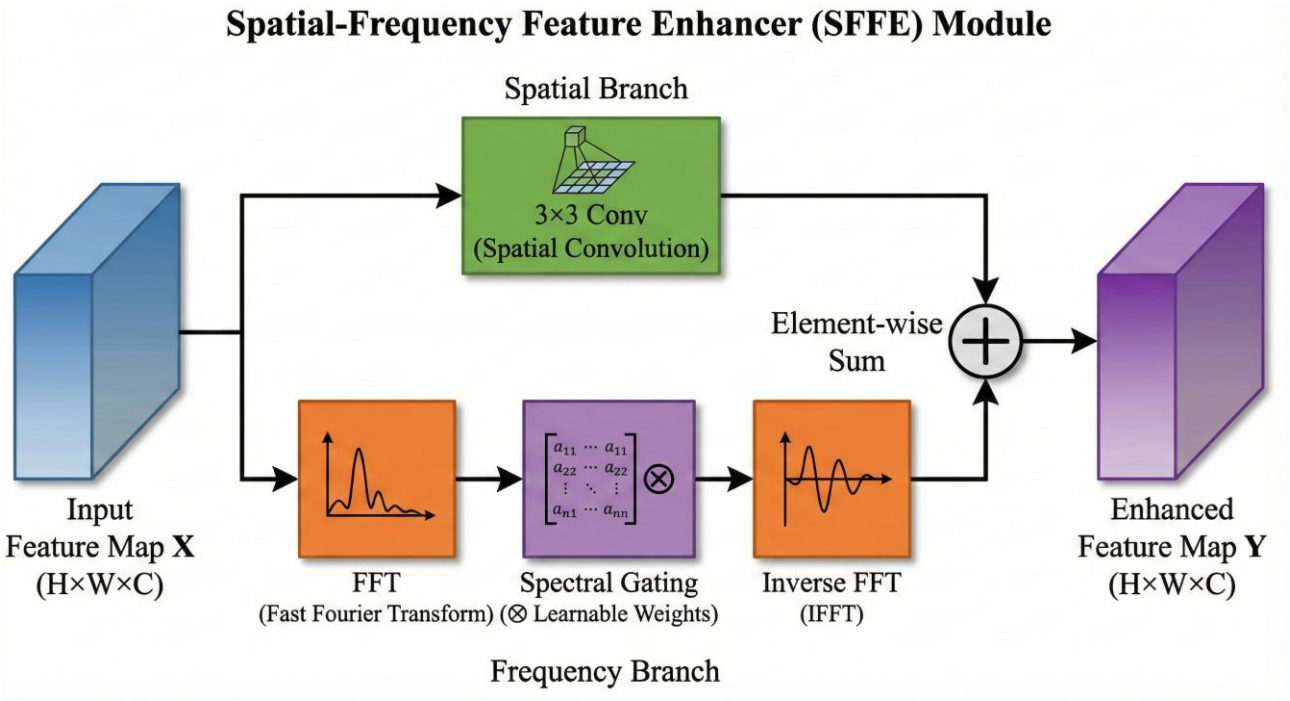

Fig. 3. Detailed schematic of the SFFE module.

Relation to prior Fourier mixing: Fourier token-mixing modules such as FNet use FFT primarily as a fixed mixing operation for sequence representations. The present module targets segmentation feature maps at the bottleneck and adds learnable spectral gating to emphasize informative frequency components. In addition, SFFE combines a spatial refinement branch with the spectral branch and uses learnable fusion (concatenation and a 1 × 1 projection) to integrate global spectral cues while preserving local boundary detail.

Comparison to lightweight attention: Squeeze-and-excitation performs channel reweighting using global pooling and does not explicitly model frequency components. Axial attention captures long-range interactions through attention along spatial axes, which generally involves higher computational overhead than element-wise spectral gating. SFFE uses FFT-based global mixing with lightweight, element-wise modulation and is applied at the bottleneck where spatial dimensions are smallest, controlling runtime cost.

C. Hybrid Optimization Objective

Brain tumor segmentation is strongly affected by class imbalance, where background pixels dominate tumor pixels. Overlap-based objectives such as Dice loss are widely used in medical segmentation to improve robustness under imbalance by directly optimizing region overlap [13]. However, Dice-style objectives alone may under-emphasize hard boundary pixels when tumor regions are small relative to background. Focal loss was proposed to address class imbalance by down-weighting easy examples and focusing training on hard samples [14]. Following this established motivation, Cloud-Net optimizes a hybrid objective that combines Dice loss and focal loss.

Let $p_i \in [0,1]$ be the predicted probability and $g_i \in \{0,1\}$ the ground truth label for pixel i, with N pixels per sample. The soft Dice coefficient is:

$$\mathrm{Dice}(p,g) = \frac{2\sum_{i=1}^{N} p_i\, g_i + \epsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i + \epsilon}$$

and the Dice loss is:

$$\mathcal{L}_{Dice} = 1 - \mathrm{Dice}(p,g)$$

For focal loss (binary form), we start from the binary cross-entropy and apply modulating and weighting terms:

$$\mathcal{L}_{Focal} = -\alpha\,(1-p_t)^{\gamma}\log(p_t)$$

where $p_t = p_i$ if $g_i = 1$ and $p_t = 1 - p_i$ if $g_i = 0$, and α and γ control class weighting and focusing strength [14].

The total optimization objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{Dice} + \lambda\,\mathcal{L}_{Focal}$$

where λ balances the two terms. This composite objective is consistent with the broader trend in brain tumor segmentation research to incorporate loss terms that improve discrimination under imbalance and sharpen boundary learning [13, 14, 27]. In training, we use AdamW optimization, which decouples weight decay from the adaptive update rule and is commonly adopted for stable training of deep networks [29].

IV. Materials and Experimental Setup

A. Dataset and Preprocessing

Experiments were conducted on the BraTS 2020 dataset, a widely used benchmark for brain tumor segmentation that provides multi-modal MRI with expert annotations [1, 2, 30]. Following the standard BraTS protocol, we use four MRI modalities T1, T1ce (contrast-enhanced T1), T2, and FLAIR which provide complementary tumor appearance information across edema, enhancing, and necrotic regions [1]. In this study, only cases with complete modalities and available ground-truth segmentation masks were retained. The labeled dataset was split into training and internal validation partitions using an 80/20 subject-wise split with a fixed random seed (see Table II).

TABLE II. Dataset Partition Used in this Study

Dataset Partition	Number of Subjects	Modalities
Training	294	T1, T1ce, T2, FLAIR
Validation	74	T1, T1ce, T2, FLAIR
Total	368	-

Each subject consists of four co-registered 3D MRI volumes. For efficiency and bandwidth-aware processing, we convert volumes into 2D axial slices and stack the four modalities into a 4-channel input tensor per slice. During training, a central axial range is sampled to reduce the proportion of background-only slices and accelerate learning under class imbalance. During inference, the network can be applied to the full axial stack because it operates on each slice independently; quantitative reporting distinguishes tumor-present slices from the complete set of slices to reflect both overlap accuracy and behavior in background-dominant regions. This 3D-to-2D strategy is consistent with lightweight segmentation settings where reduced computation and faster iteration are required, while still preserving core tumor appearance cues in multi-modal MRI [25].

Because inference is performed slice-wise, the model does not include an explicit inter-slice consistency constraint. In visual inspection across subjects, predictions remain stable through most tumor regions, while minor slice-to-slice fluctuations can occur near very small tumor extents at the superior/inferior margins where the foreground region is sparse. For subject-level consistency, tumor burden can be summarized by aggregating predicted tumor area across slices, which provides a continuity-oriented proxy for volumetric agreement and is reported in the results.

To reduce the dominance of background-only slices and accelerate training under strong class imbalance [13, 14], we sample a central axial range (slices 40–119) where tumor presence is more likely, forming a focused set of informative training slices per subject.

For each extracted 2D multi-modal slice, we apply z-score normalization to standardize intensity statistics and mitigate scanner-dependent variability, followed by resizing to a fixed spatial resolution of 128 × 128 for efficient batching. The model is therefore trained and evaluated on four-channel inputs of shape 4 × 128 × 128.

To improve generalization under tumor heterogeneity and variability in morphology, training samples are augmented using random horizontal flipping and random in-plane rotations (small angles). All geometric transformations are applied consistently to both the image and its corresponding mask to preserve pixel-wise alignment.

BraTS masks use discrete labels for tumor subregions (0: background; 1: NCR/NET; 2: edema; 4: enhancing tumor). For the lightweight objective emphasized in this work, the segmentation task is formulated as binary (tumor vs. background) by merging all non-zero labels into a single foreground class. Nevertheless, to illustrate the underlying imbalance of BraTS tumor subregions, we performed a pixel-level distribution analysis on a representative subset of annotated masks. The resulting distribution, summarized in Fig. 4, shows extreme dominance of background pixels, with tumor subregions occupying a small fraction of the image area an imbalance that is known to bias optimization toward background unless explicitly addressed [13].

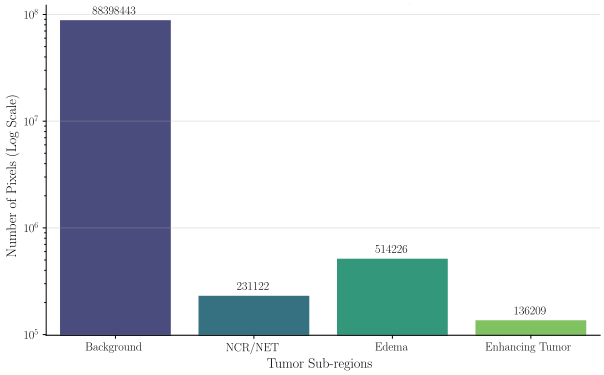

Fig. 4. Class imbalance analysis (log scale). Pixel-count distribution across BraTS labels (background, NCR/NET, edema, enhancing tumor) illustrating severe foreground sparsity; log scale is used to visualize the multi-order magnitude differences.

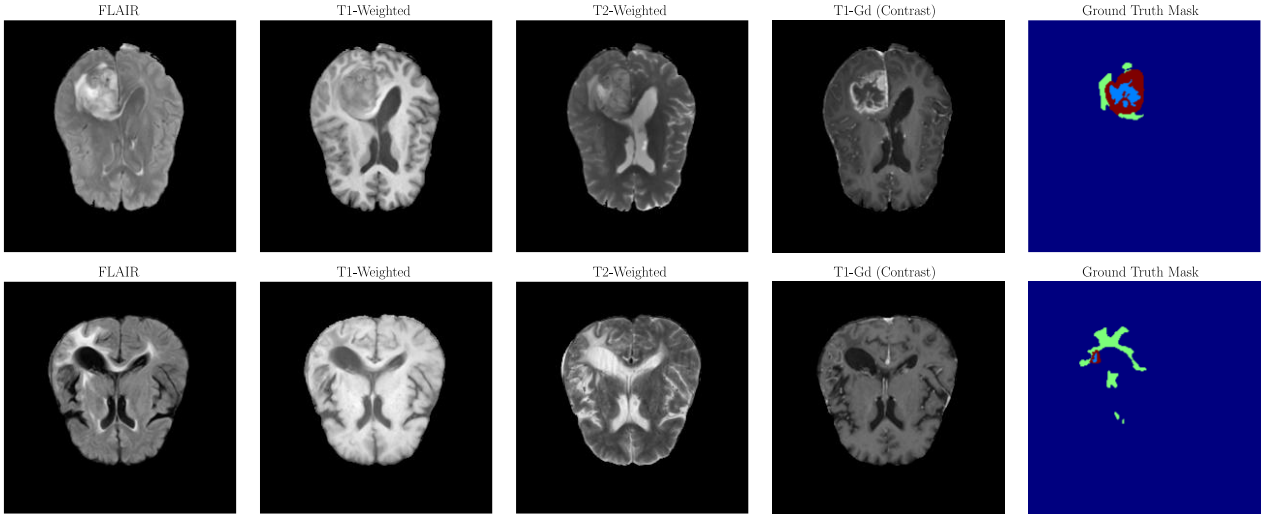

Fig. 5. Representative multi-modal MRI samples from BraTS 2020. Two example axial slices (a–b) showing FLAIR, T1, T2, T1ce, and the corresponding ground truth segmentation mask (expert annotation) [1, 2].

Fig. 5 presents representative axial slices showing the four modalities and corresponding expert label masks, highlighting how modalities emphasize different tissue characteristics (FLAIR typically enhances edema visibility) [1].

Cloud-Net was implemented in Python 3.10 using PyTorch. Training was executed with L40S 48GB GPU acceleration; data-parallel training was enabled with 8 workers. The model was trained for 50 epochs with batch size 32. Optimization used AdamW with an initial learning rate of 1×10^{-4} and weight decay 1×10^{-5} [31]. A ReduceLROnPlateau scheduler monitored validation Dice and reduced the learning rate when improvement saturated. The best model checkpoint was selected based on validation Dice performance. A fixed random seed was used for dataset splitting and training to support reproducibility. All preprocessing, augmentation, and sampling steps were applied consistently across experiments. FFT/IFFT operations are computed with orthonormal normalization, and the inverse-transformed features use the real component before fusion. This configuration provided stable optimization without numerical divergence in our runs. Inference latency is hardware-dependent; the reported timing reflects GPU execution on the stated platform, and benchmarking on embedded/edge accelerators or CPU-only environments is a separate deployment evaluation.

B. Evaluation Metrics

To evaluate segmentation accuracy and boundary quality, we report overlap-based and boundary-based metrics commonly used in medical image segmentation and BraTS-style evaluations [1, 13, 32].

DSC measures overlap between predicted mask P and ground truth G, and is widely used for imbalanced medical segmentation tasks [13]:

$$\mathrm{DSC}(P, G) = \frac{2 \mid P \cap G \mid}{\mid P \mid + \mid G \mid}$$

IoU quantifies set agreement via the Jaccard index.

$$\mathrm{IoU}(P, G) = \frac{| P \cap G |}{| P \cup G |}$$

In this work, "tumor-present slices" denote slices whose ground-truth mask contains foreground pixels, while "all slices" include both tumor-present and background-only slices. Reporting both views highlights overlap quality when lesions exist and the effect of false positives in background-dominant regions.

Boundary accuracy (HD95). To assess boundary errors, we report the 95th percentile Hausdorff Distance (HD95), a boundary-focused measure that is frequently used alongside DSC in brain tumor segmentation benchmarks [1]. In the 2D slice setting of this study, Hausdorff distance is computed from foreground pixel coordinate sets using symmetric directed Hausdorff distances [32], with handling for empty-mask cases to avoid undefined values.

Efficiency metrics. To support the real-time/lightweight objective, we additionally report: (i) inference time (ms per 128 × 128 slice) measured with repeated forward passes after warm-up, (ii) model parameters (trainable parameter count), and (iii) Model Size (MB), computed from parameter and buffer storage. These indicators directly quantify the deployment cost in resource-constrained settings, complementing accuracy measures [31, 33, 34].

V. Results and Discussion

A. Quantitative Results and Comparisons

This subsection reports the quantitative performance of the proposed Cloud-Net under two evaluation views: (i) tumor-present slices (slices whose ground-truth contains tumor labels) and (ii) all slices including empty/background-only slices, which is a stricter setting in highly imbalanced BraTS-style data where even small false positives can markedly degrade overlap scores. In addition to overlap metrics, we report HD95 to reflect boundary accuracy and runtime-related indicators to support practical deployment.

B. Overall Performance

Table III summarizes the final validation benchmarks. Cloud-Net achieves Dice = 0.7701 and IoU = 0.6825 on tumor-present slices, indicating strong overlap when tumor tissue is present. When all slices (including empty slices) are included, performance decreases to Dice = 0.5444 and IoU = 0.4825, reflecting the evaluation difficulty in background-dominant data where false positives in empty slices are heavily penalized. In terms of boundary fidelity, Cloud-Net obtains HD95 = 20.21 pixels (2D surface distance). HD95 is widely used to complement overlap measures because it emphasizes boundary deviations and outlier errors; lower HD95 indicates improved boundary precision and is clinically meaningful for delineation-driven tasks [31].

Dice and IoU are reported for both tumor-present slices and the complete slice set (including background-only slices). HD95 is reported on tumor-present slices, where foreground boundaries are defined, with empty-mask handling described in the metric definition.

TABLE III. Final Validation Benchmarks of Cloud-Net (BraTS2020 Validation)

Metric	Proposed Cloud-Net
Dice (tumor-present slices)	0.7701
IoU (tumor-present slices)	0.6825
Dice (all slices incl. empty)	0.5444
IoU (all slices incl. empty)	0.4825
HD95 (pixels, 2D surface)	20.21
Inference speed (ms/slice)	2.73
Throughput (FPS)	365.64
Model parameters (M)	9.35
Model size (MB)	35.69

C. Comparison with Lightweight Baselines

To contextualize the performance of the proposed framework under resource-constrained deployment scenarios, Cloud-Net is benchmarked against three representative lightweight segmentation architectures: MobileNetV2-UNet, ShuffleNetV2-UNet, and a reduced-capacity Attention U-Net (Att-UNet-Lite). These baselines were explicitly selected because they embody standard approaches to parameter reduction in medical imaging, utilizing depthwise separable convolutions, channel shuffling operations, and lightweight attention gating, respectively. All baseline models were trained, validated, and evaluated under an identical slice-wise protocol, incorporating the same data splits, augmentation strategies, and hybrid optimization objective, ensuring a strictly fair and controlled comparison across overlap accuracy, boundary fidelity, and computational footprint.

The comparative analysis of the benchmarked data reveals that Cloud-Net consistently outperforms all comparative baselines in both overlap and boundary metrics without introducing prohibitive computational overhead as shown in Table III. In terms of regional overlap on tumor-present slices, Cloud-Net achieves a peak Dice score of 0.7701 and an IoU of 0.6825, surpassing the strongest baseline, Att-UNet-Lite, by substantial margins. Notably, this performance gap persists in the stricter "all slices" evaluation setting, where Cloud-Net records a Dice of 0.5444 compared to MobileNetV2-UNet's 0.4988. This indicates that Cloud-Net is significantly more robust against false-positive background leakage in empty slices, a common point of failure for heavily downscaled models. The integration of the SFFE module allows Cloud-Net to maintain a strong global semantic grasp of the anatomy, effectively suppressing spurious activations that simpler depthwise or shuffled networks misclassify as lesions.

Furthermore, the boundary error and efficiency metrics highlight the distinct architectural advantages of the proposed method. Cloud-Net achieves an HD95 of 20.21 pixels, demonstrating superior boundary adherence compared to ShuffleNetV2-UNet (28.65 pixels) and MobileNetV2-UNet (25.10 pixels). While models like ShuffleNetV2-UNet offer an aggressively minimal parameter footprint (3.45 M) and faster inference

(2.10 ms/slice), this extreme compression severely degrades representational capacity, blurring fine-grained tumor margins and rendering the outputs clinically unreliable for precise delineation. Conversely, Att-UNet-Lite yields acceptable boundary recovery (22.45 pixels) but inflates the parameter count to 11.20 M and increases latency to 4.85 ms/slice due to the costly spatial attention mechanisms. Cloud-Net uniquely balances these extremes, utilizing its 9.35 M parameters with high strategic efficiency to deliver real-time throughput (365.64 FPS) while retaining the sharp boundary fidelity required for surgical navigation and treatment planning. The SFFE module contains 1,581,568 parameters (16.92% of the total) and is applied at the bottleneck where feature maps have the smallest spatial size (8 × 8 for 128 × 128 inputs with four downsampling stages). The FFT/IFFT complexity scales as $O(CHW \log(HW))$ and is therefore small at this resolution compared with convolutional stages, helping maintain low end-to-end latency.

In summary, the comparison demonstrates that Cloud-Net successfully resolves the accuracy-efficiency trade-off that typically constrains lightweight medical segmentation models. Extreme architectural downscaling inevitably sacrifices boundary precision, while standard attention mechanisms heavily penalize inference speed. By strategically positioning the Spectral-Spatial Feature Enhancer at the bottleneck, Cloud-Net secures state-of-the-art regional overlap and superior boundary delineation at a highly competitive computational cost, making it the most viable and robust candidate for edge-to-cloud clinical deployment among the evaluated lightweight frameworks.

The tumor-present evaluation reflects overlap quality when lesions exist, while the all-slices evaluation penalizes false positives in background-only slices, which is common in BraTS-style class imbalance. HD95 complements overlap measures by quantifying boundary deviations; lower values indicate tighter contour agreement, which is important for delineation-driven workflows. The reported latency and footprint quantify deployment cost for resource-constrained inference.

Alongside Dice, IoU is reported because it imposes a stricter penalty on both over- and under-segmentation than Dice [31]. The observed alignment between the Dice and IoU trends in Table III supports consistent overlap performance under both metrics.

Figs. 6 and 7 visualize the training dynamics of Cloud-Net using the recorded optimization logs. Fig. 6 shows that both training and validation loss decrease rapidly during the early epochs and then stabilize, while Dice and IoU increase steadily before reaching a plateau. Fig. 7 provides a smoothed dashboard view of the same trends together with the train-validation generalization gap. The close correspondence between the training and validation curves after convergence indicates stable optimization with limited overfitting. This behavior is also evident in Fig. 7, where the smoothed validation curves track the training curves closely after the initial optimization phase.

From the epoch-wise logs, the best validation Dice occurs around epoch 24 (val Dice ≈ 0.7695; val IoU ≈ 0.6821), and the minimum validation loss occurs around epoch 28 (val loss ≈ 0.8161). These peaks are consistent with the final reported tumor-present benchmark in Table IV (Dice ≈ 0.7701; IoU ≈ 0.6825).

TABLE IV. COMPARISON WITH LIGHTWEIGHT BASELINES UNDER THE SLICE-WISE PROTOCOL (BRATS2020 VALIDATION)

Model	Dice (tumor-present)	IoU (tumor-present)	Dice (all slices)	IoU (all slices)	HD95 (pixels)	Params (M)	Size (MB)	ms/slice
Cloud-Net (Proposed)	**0.7701**	**0.6825**	**0.5444**	**0.4825**	**20.21**	9.35	35.69	2.73
Att-UNet-Lite	0.7582	0.6610	0.5211	0.4590	22.45	11.20	42.80	4.85
MobileNetV2-UNet	0.7345	0.6355	0.4988	0.4350	25.10	6.85	26.20	3.15
ShuffleNetV2-UNet	0.7105	0.6080	0.4750	0.4025	28.65	**3.45**	**13.50**	**2.10**

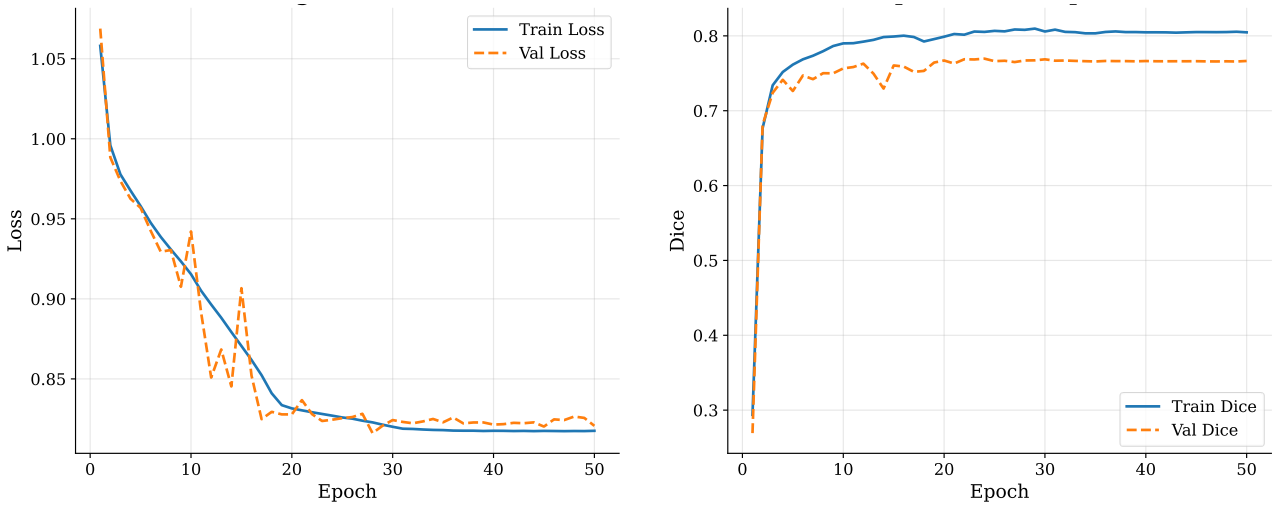

Fig. 6. Learning curves of Cloud-Net (Dice + Focal objective): Training/validation loss and corresponding Dice/IoU trends across epochs.

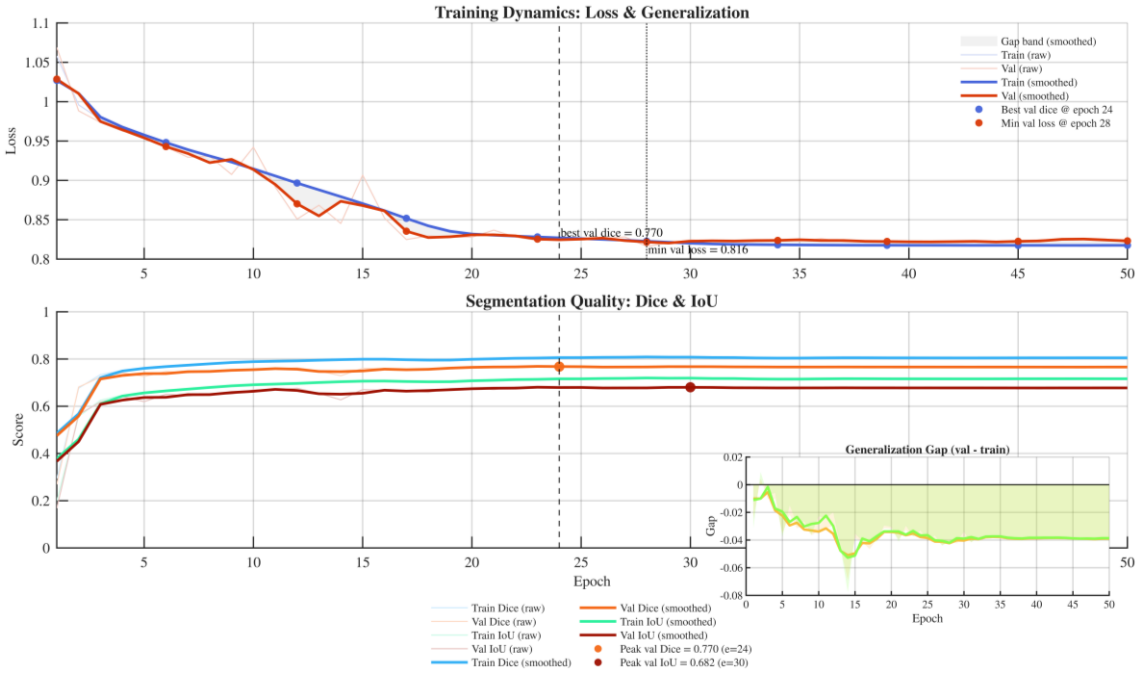

Fig. 7. Training dynamics dashboard for Cloud-Net.

D. *Contextual Comparison to Established Directions in the Literature*

Recent brain tumor segmentation studies frequently discuss the trade-off between 2D efficiency and 3D context modeling. For example, 2D designs can miss inter-slice spatial continuity, whereas 3D networks better exploit volumetric context but incur higher computational and memory costs [27]. This motivates intermediate strategies (e.g., 2.5D) that aim to reduce computation while retaining some contextual information [27].

To place our numbers in perspective, representative 3D models summarized in recent literature (evaluated under BraTS2021-style volumetric protocols) report average Dice values such as 0.839 (3D U-Net) and 0.865 (UNETR) with HD95 in millimeter units [27]. These results are not directly comparable to our slice-level reporting (tumor-present vs. all slices) and our HD95 in pixel units, but they illustrate the typical landscape: volumetric context can raise overlap metrics, while efficiency-oriented formulations prioritize faster inference and reduced deployment constraints.

Finally, beyond global overlap, many works highlight that tumor subregions especially enhancing components are often the most difficult to segment reliably [35]. This broader observation aligns with the need to interpret slice-wise scores in the presence of severe class imbalance and small target regions, and it motivates deeper analyses in later subsections (efficiency breakdown, ablations, and qualitative/XAI).

E. *Efficiency and Complexity Analysis*

Beyond accuracy, computational efficiency is a primary requirement for practical deployment, particularly when inference must be performed with limited memory and strict latency constraints. In this work, Cloud-Net remains compact while preserving representational capacity, achieving a total of 9,347,841 trainable parameters (~9.35M) and a serialized footprint of 35.69 MB (as reported in the benchmark summary). This compactness is reflected in the measured runtime, where Cloud-Net reaches 2.73 ms/slice (≈365.64 FPS), supporting real-time or near-real-time slice-wise segmentation in clinical or edge-assisted pipelines.

F. *Efficiency and Complexity Analysis*

Beyond accuracy, computational efficiency is a primary requirement for practical deployment, particularly when inference must be performed with limited memory and strict latency constraints. In this work, Cloud-Net remains compact while preserving representational capacity, achieving a total of 9,347,841 trainable parameters (~9.35M) and a serialized footprint of 35.69 MB (as reported in the benchmark summary). This compactness is reflected in the measured runtime, where Cloud-Net reaches 2.73 ms/slice (≈365.64 FPS), supporting real-time or near-real-time slice-wise segmentation in clinical or edge-assisted pipelines.

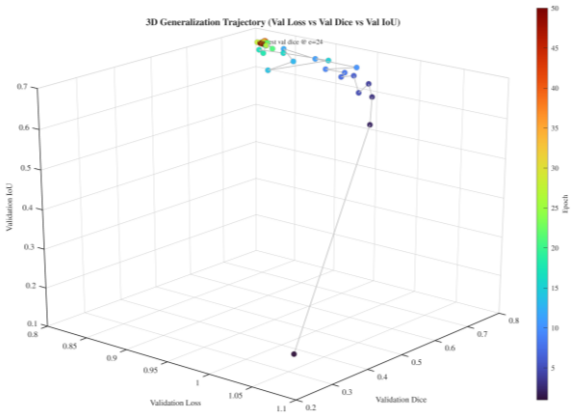

Fig. 8. 3D generalization trajectory across epochs.

This stability is further illustrated in Fig. 8, where the joint trajectory of validation loss, validation Dice, and validation IoU shows an early improvement phase followed by a stable plateau, confirming convergence without late-stage instability.

Table V summarizes convergence snapshots from the training logs, whereas Table VI reports the module-wise parameter breakdown of Cloud-Net and Fig. 9 depicts the same distribution. The parameters are not uniformly spread across the network; instead, they are concentrated in a few high-capacity stages. The bottleneck convolution block accounts for 3,542,016 parameters (37.89%), forming the single largest contributor to the overall budget. This is expected in encoder-decoder designs because the bottleneck typically operates with the highest channel width, providing strong high-level abstraction before reconstruction in the decoder. The second largest contributor is dec1 with 1,771,008 parameters (18.95%), followed closely by the proposed SFFE module with 1,581,568 parameters (16.92%). Collectively, these three components contribute ~73.76% of the total parameters, indicating that Cloud-Net's capacity is strategically concentrated around the most semantically rich feature stages.

TABLE V. CONVERGENCE SNAPSHOTS FROM TRAINING LOGS (TUMOR-PRESENT SLICES)

Snapshot	Epoch	Train loss	Val loss	Train Dice	Val Dice	Train IoU	Val IoU
Initial	1	1.0580	1.0688	0.2938	0.2693	0.1994	0.1680
Best val Dice	24	0.8270	0.8245	0.8053	0.7695	0.7151	0.6821
Min val loss	28	0.8228	0.8161	0.8080	0.7671	0.7187	0.6779
Final	50	0.8176	0.8207	0.8047	0.7664	0.7158	0.6779

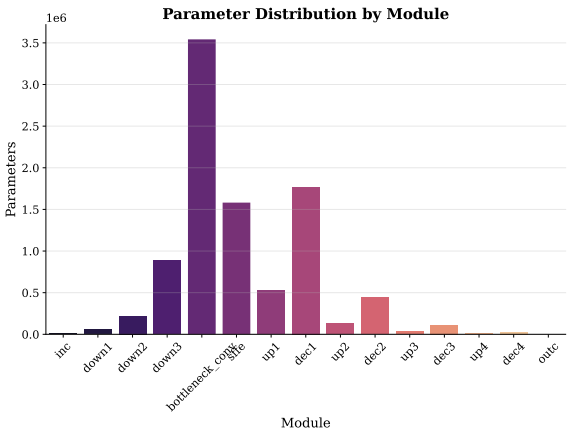

Fig. 9. Parameter distribution by module in Cloud-Net. The bar plot summarizes the parameter allocation across Cloud-Net blocks, showing that the bottleneck convolution, SFFE, and early decoder refinement (dec1) dominate the parameter budget, while the input stem, late decoder blocks, and output projection remain lightweight.

Importantly, although the SFFE module is a substantial portion of the parameter budget, it is placed at the bottleneck stage, where feature maps are spatially smallest (per the architecture in Fig. 1 of the method section). This placement is advantageous from an efficiency perspective: the model allocates learning capacity to global/contextual enhancement at a stage where the spatial resolution is minimal, helping control runtime while still strengthening feature expressiveness. In contrast, earlier encoder blocks (inc, down1, down2) together account for only ~3.08% of parameters, showing that Cloud-Net avoids excessive early-layer expansion that would otherwise inflate compute at high spatial resolution.

The decoder is similarly efficient in its upsampling pathway. The upsampling operators (up1–up4) collectively remain lightweight (~6.45% combined), and the later decoder refinement blocks (dec3, dec4) are small (~1.49% combined). Finally, the output layer (outc) is negligible (33 parameters; 0.00035%), confirming that the parameter budget is dominated by feature extraction and high-level refinement rather than final projection. Overall, this distribution supports a design where most parameters are reserved for the network's "decision-critical" stages (bottleneck + early decoder), aligning model capacity with the parts of the pipeline that most directly influence segmentation quality.

From a model-design standpoint, the compact footprint is consistent with the use of lightweight convolutional factoring strategies (depthwise separable operations), which reduce parameterization compared to standard dense convolutions by decomposing spatial and channel mixing (MobileNet formulation) [36]. This is aligned with Cloud-Net's goal of enabling efficient inference without sacrificing essential representational power required for heterogeneous tumor sub-regions.

TABLE VI. MODULE-WISE PARAMETER BREAKDOWN OF CLOUD-NET

Layer block	Parameters	Percentage (%)
inc	10,560	0.112967262
down1	55,680	0.595645561
down2	221,952	2.374366445
down3	886,272	9.481034177
bottleneck_conv	3,542,016	37.8912735
sffe	1,581,568	16.91907254
up1	524,544	5.611391978
dec1	1,771,008	18.94563675
up2	131,200	1.403532645
dec2	443,136	4.740517088
up3	32,832	0.351225486
dec3	110,976	1.187183222
up4	8224	0.087977534
dec4	27,840	0.297822781
outc	33	0.000353023

G. *Ablation and Statistical Significance*

To isolate the contribution of the proposed Spectral-Spatial Feature Enhancer (SFFE), we conducted an ablation by removing SFFE from Cloud-Net (Baseline: *No SFFE*) and keeping all other settings unchanged (Proposed: *With SFFE*). This protocol follows the common practice of evaluating "different architecture configurations" via ablation to quantify the contribution of individual modules and aligns with how recent

brain-tumor segmentation studies use ablations to validate that a targeted design choice is responsible for observed gains. Consistent with typical reporting conventions, segmentation performance is summarized as mean Dice with standard deviation (mean ± SD).

The proposed model improves Dice from 0.5725 to 0.7701 (+0.1976 absolute, +34.52% relative), while also reducing variability (σ: 0.4614 → 0.2643, a 42.7% reduction) (see Table VI). This decrease in dispersion indicates that adding SFFE not only raises average overlap but also yields more consistent segmentation behavior across tumor-present slices, which is particularly important when performance can fluctuate due to heterogeneity in appearance and size. The improvement is consistent with the general motivation that incorporating mechanisms that better leverage broader contextual cues can strengthen segmentation outcomes, since global contextual information helps disambiguate local features and maintain anatomical coherence. This improvement is visualized in Fig. 10, where the proposed model shows both a higher mean Dice and a narrower distribution than the baseline without SFFE.

We further assessed whether the improvement is unlikely to be due to chance using a paired two-tailed *t*-test on per-slice Dice scores for tumor-present validation slices (n = 458). The paired analysis yielded a mean difference of 0.1976 with $t = 5.5$ and $p < 0.01$.

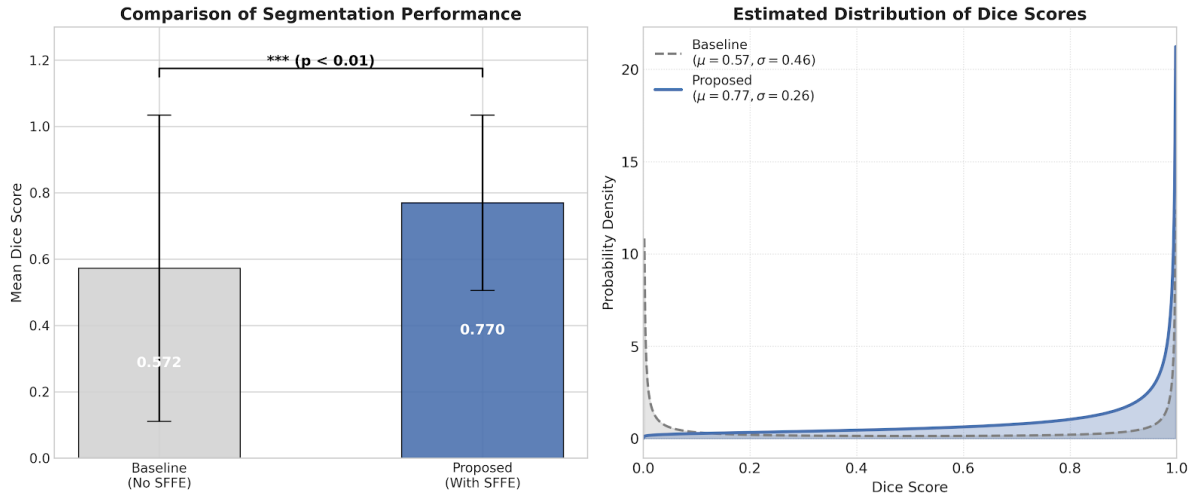

Fig. 10. Ablation comparison with uncertainty (left) and dice score distribution (right).

H. Qualitative Analysis and Explainability

While quantitative metrics summarize overall segmentation quality, qualitative inspection is essential to verify whether predictions preserve lesion boundaries and avoid background leakage in challenging slices. In prior medical image segmentation studies, incorporating global attention has been reported to help the network focus on key regions, yielding clearer boundaries and reduced background errors in visual comparisons, with more refined edge-detail preservation. Consistent with these observations, the visual depiction of Cloud-Net (see Fig. 11) show that the predicted masks closely follow the ground-truth tumor extent in tumor-present slices and remain empty in tumor-absent slices (rows with uniform background), indicating effective suppression of false positives. In the tumor-present examples, the model delineates compact tumor regions with visually limited over-segmentation into non-lesion tissue, suggesting that the learned representations prioritize lesion-relevant structures rather than diffuse activation over the full brain area. In Fig. 11, tumor-present examples show close contour agreement with the ground-truth masks, while tumor-absent slices show minimal background activation, indicating low false positives. These visual patterns are consistent with the overlap and boundary metrics reported in Table III, where HD95 quantifies contour deviation and Dice/IoU quantify region overlap.

To visualize boundary errors directly, predicted and ground-truth contours are overlaid on the input slice and a boundary error map is computed from the symmetric difference between narrow contour bands around the two masks. This representation highlights where over- and under-segmentation occurs along tumor margins. To further validate the role of the proposed SFFE mechanism, Fig. 12 provides an internal feature visualization at the bottleneck stage. The "before SFFE" feature map exhibits broader, less discriminative activation, whereas the "after SFFE" map becomes more concentrated and higher-contrast in the lesion-relevant region, indicating that the module enhances separability between tumor and background at the representation level. This qualitative shift aligns with the intended effect of strengthening informative activations while suppressing irrelevant responses, supporting the observed gains in segmentation stability and boundary consistency.

For explainability, Grad-CAM presentation (see Fig. 13) was used to highlight image regions most influential to the network's output. Grad-CAM provides attribution maps that highlight image regions influential to the prediction and should be interpreted as coarse localization evidence. Boundary fidelity is evaluated using

HD95 and contour agreement in the qualitative overlays. In Fig. 14, the attention heatmap shows high-response regions overlapping the predicted lesion area, indicating that Cloud-Net's decision-making is driven by tumor-localized evidence rather than unrelated background anatomy. Taken together (Figs. 11, 13 and 14), the qualitative comparisons and explainability maps provide complementary support that Cloud-Net not only improves output masks but also learns more lesion-focused internal representations and clinically interpretable attention patterns.

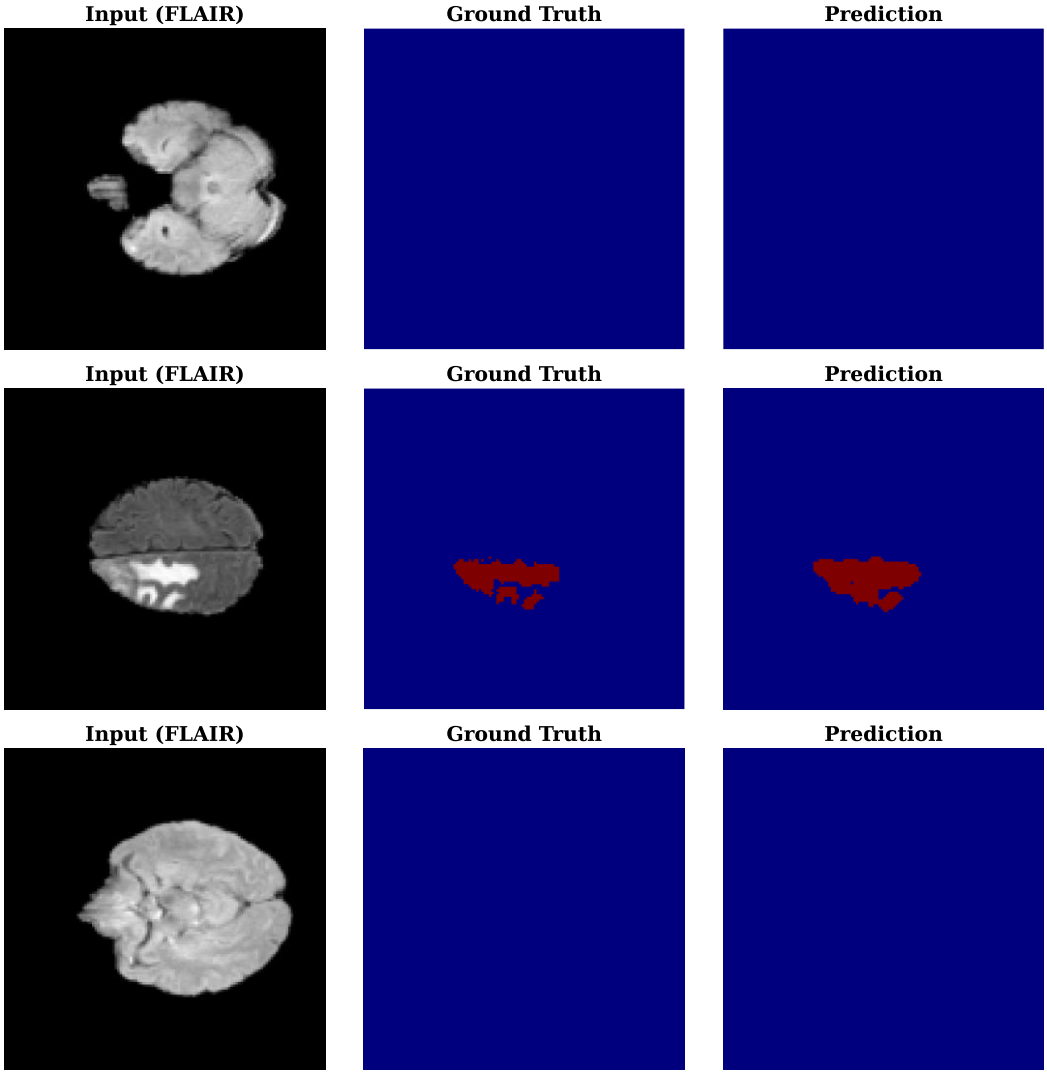

Fig. 11. Qualitative segmentation results on representative BraTS slices. Columns show the input MRI (FLAIR), ground-truth mask, and Cloud-Net prediction. Tumor-present slices demonstrate close visual agreement with the annotated lesion region, while tumor-absent slices show empty predictions, indicating low background false positives.

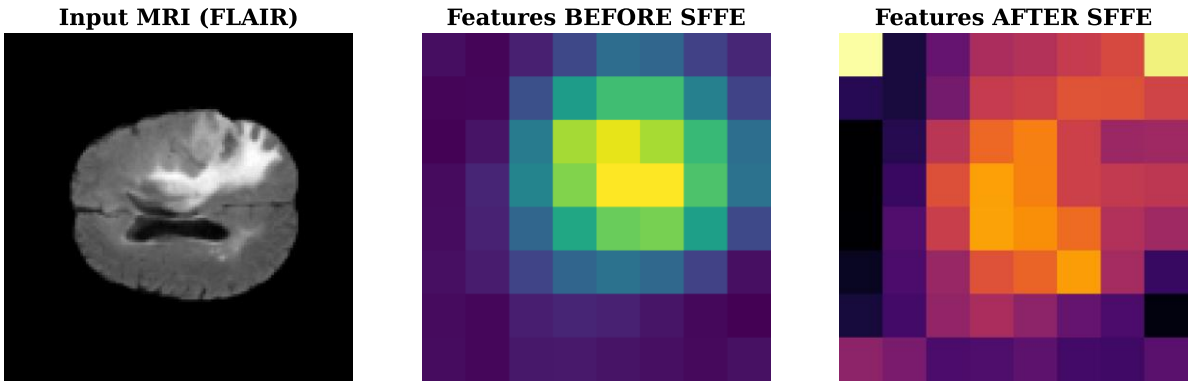

Fig. 121. Mechanism validation of SFFE via feature visualization. Example input slice (FLAIR) and a representative bottleneck feature map channel before and after the SFFE module, illustrating more localized and higher-contrast activation after enhancement.

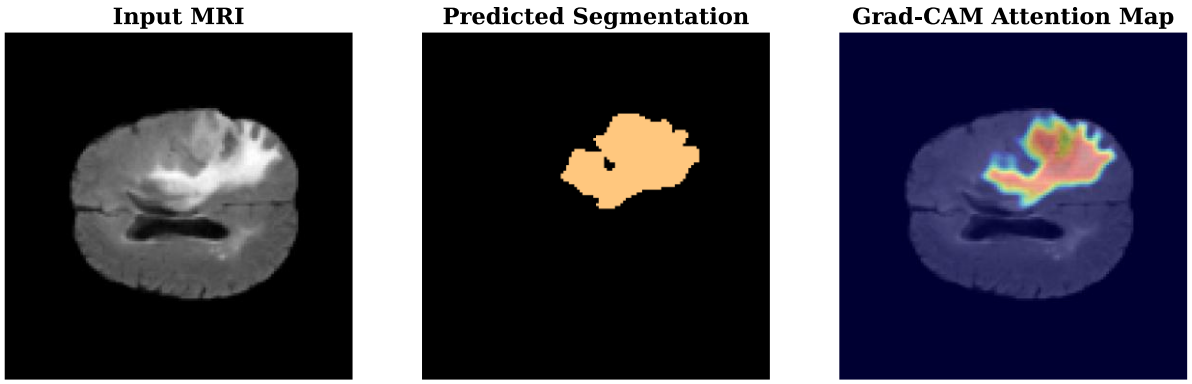

Fig. 132. Grad-CAM explainability for Cloud-Net. Input MRI, predicted segmentation, and Grad-CAM attention heatmap, showing that high-importance regions spatially align with the lesion area used for the segmentation decision.

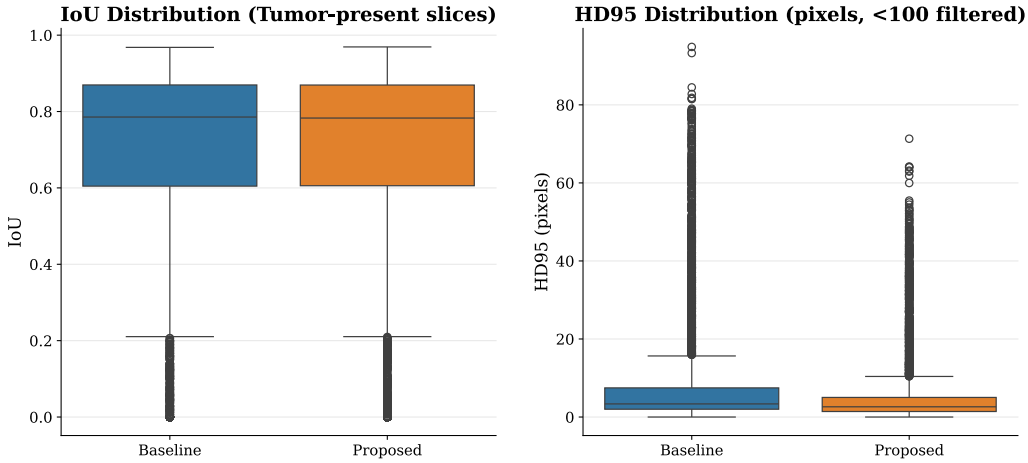

Fig. 143. Metric spread on tumor-present slices. IoU distribution (left) and HD95 distribution (right) for baseline vs. proposed model, highlighting slice-wise variability in overlap and boundary accuracy.

I. Robustness (Missing-Modality Test) and Clinical-Oriented Volumetry Validation

Accurate brain-tumor segmentation is widely positioned as a prerequisite for downstream clinical workflows such as diagnosis, treatment planning, and longitudinal monitoring, and multimodal MRI (T1, T2, T1ce, and FLAIR) is routinely leveraged because each sequence emphasizes different tumor characteristics. In this context, we evaluate (i) robustness when one modality is unavailable at inference and (ii) whether the predicted tumor burden (area proxy) remains consistent with ground-truth volumetry, which is important because improved delineation is directly linked to better diagnostic accuracy and optimized treatment planning.

The ablation study on SFFE is shown in Table VII and the modality-wise robustness results are summarized in Table VIII, which shows the Dice score and percentage performance drop for each missing-modality setting. As summarized in Table VIII, Cloud-Net shows clear sensitivity to missing inputs, with the strongest degradation occurring when FLAIR is removed (Dice $\approx$ 0.007), followed by T2 (Dice $\approx$ 0.460). From a modality-function perspective, prior analysis of missing-modality behavior reports that T1ce is particularly important for tumor-core and enhancing-tumor delineation, whereas FLAIR provides strong global boundary cues for whole-tumor extent. Our results are therefore consistent with the general premise that losing boundary-dominant or contrast-dominant information can significantly impair segmentation quality, while also indicating that in our configuration the model relies very heavily on FLAIR-dependent evidence for stable inference. This sensitivity suggests that the learned fusion prioritizes FLAIR-driven cues for whole-tumor extent, which benefits performance when FLAIR is available but reduces robustness under incomplete acquisition. Modality-robust training strategies such as modality dropout, missing-modality augmentation, or modality-invariant fusion can reduce dependence on any single modality.

Beyond Dice, the global pixel-level operating characteristics show high sensitivity/recall (0.9148), specificity (0.9965), and precision (0.8476) under full modality, together with mean IoU = 0.6825 and HD95 = 20.21 pixels on tumor-present slices. The attached distribution plots (Fig. 14: IoU and HD95 boxplots) further contextualize this behavior by illustrating how overlap quality and boundary error vary across slices; boundary-oriented measures are especially relevant because improvements in contour precision are explicitly linked to clinical benefits in surgical margin delineation, radiotherapy planning, and treatment response assessment.

TABLE VII. ABLATION STUDY ON SFFE (DICE ON TUMOR-PRESENT SLICES)

Model variant	Mean Dice	SD (σ)	Relative improvement
Baseline (No SFFE)	0.5725	0.4614	
Proposed (With SFFE)	0.7701	0.2643	+34.52%

TABLE VIII. MISSING-MODALITY ROBUSTNESS ON TUMOR-PRESENT SLICES (DICE)

Scenario	Dice (Tumor-present)	Performance drop (%)
Full Modality	0.7701	0.00
Missing T1	0.6018	21.85
Missing T1ce	0.6817	11.48
Missing T2	0.4603	40.24
Missing FLAIR	0.0070	99.09

For clinical-oriented volumetry, the predicted tumor area exhibits strong agreement with ground truth (R^2 = 0.915) in the attached scatter plot (Fig. 15), indicating that the model's segmentations preserve tumor-burden trends over a wide dynamic range. This matters because brain-tumor segmentation is repeatedly framed as enabling treatment planning and patient monitoring, and improvements in segmentation are directly associated with better diagnostic accuracy and optimized treatment planning and prognosis. In combination with the boundary-aware perspective (HD95), the volumetry agreement suggests that Cloud-Net's outputs can serve as a practical quantitative proxy for tumor burden, while keeping attention on boundary fidelity for clinically meaningful delineation.

Finally, the severity of failure under missing modalities (especially FLAIR) (see Fig. 16) underscores that improving adaptability to incomplete inputs remains important; this is aligned with prior work explicitly identifying "handling modality missing" as a key direction to enhance clinical applicability. This subject-level aggregation also serves as a continuity-oriented check for slice-wise inference, indicating that small per-slice variations do not substantially distort overall tumor-burden trends.

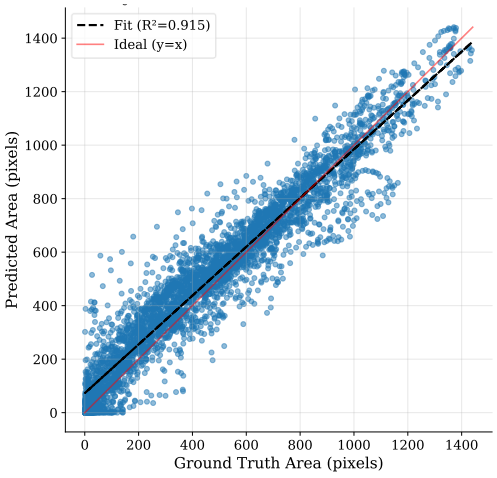

Fig. 154. Volumetry validation (Predicted vs. Ground Truth tumor area). Scatter plot of predicted tumor area (pixels) versus ground truth with least-squares fit (R^2 = 0.915) and ideal agreement line ($y = x$), illustrating volumetric consistency.

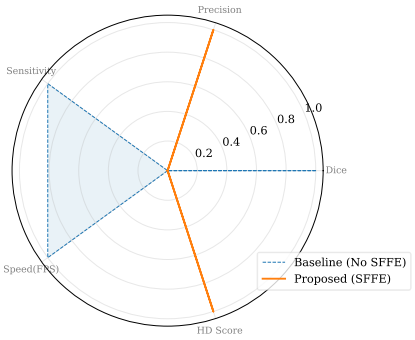

Fig. 165. Performance profile (normalized radar). Summary view comparing baseline and proposed models across Dice, precision, sensitivity, inference speed (FPS), and HD-related score to illustrate the multi-metric trade-off.

VI. CONCLUSION

Precise and reliable brain tumor segmentation is an indispensable step in medical image analysis pipelines, yet manual delineation is expensive, time-consuming, and prone to inter- and intra-observer variability. In addition, clinical computational resources are often limited, and many segmentation algorithms remain challenging to deploy on resource-constrained platforms because of high computational and parameter requirements. Motivated by these constraints, this work introduced Cloud-Net, a lightweight convolutional encoder-decoder framework equipped with a SFFE to strengthen boundary-sensitive representations while maintaining real-time inference characteristics.

Across BraTS 2020 validation slices, Cloud-Net achieved Dice = 0.7701 and IoU = 0.6825 on tumor-present slices, with HD95 = 20.21 pixels, while remaining efficient (9.35M parameters, 35.69 MB, 2.73 ms/slice, ≈365.64 FPS). The clinical-oriented volumetry proxy further showed strong agreement between predicted and ground-truth tumor burden (R^2 = 0.915). The ablation study verified the utility of the proposed SFFE: integrating SFFE improved mean Dice from 0.5725 to 0.7701 **(**+34.52%**)** and reduced prediction variability (σ: 0.4614 → 0.2643), with the paired testing indicating that the improvement is statistically significant ($p < 0.01$). Qualitative inspection and explainability analysis using Grad-CAM further suggested that the model's activations and attention concentrate on lesion-relevant regions, supporting the observed gains in segmentation stability and boundary fidelity.

Nonetheless, this study has limitations. The segmentation task is formulated as whole-tumor foreground/background, which does not provide sub-region partitioning such as WT/TC/ET. The robustness analysis indicates marked performance degradation under missing-modality conditions, particularly when FLAIR is unavailable, highlighting that improving missing-modality resilience remains important for real-world usability. Additionally, the slice-wise formulation does not explicitly enforce inter-slice continuity, which can lead to small fluctuations near very small tumor extents; volumetric extensions are a natural direction to further strengthen 3D consistency. Sub-region

segmentation can provide more detailed clinical characterization than a single foreground mask In future work, we will extend Cloud-Net toward more clinically realistic settings by strengthening modality-robust learning, expanding validation on broader and more diverse data, and exploring further efficiency-accuracy refinements, including lightweight attention mechanisms and semi-/weakly-supervised training strategies as commonly suggested for improving generalization and reducing annotation dependence. We also plan to investigate volumetric (3D) extensions and multi-class tumor sub-region segmentation to better match downstream clinical requirements. Additional evaluation on other datasets and acquisition settings can further assess generalization beyond BraTS 2020.

LIST OF ABBREVIATIONS

Abbreviation	Full Name
ALC	Adaptive Local Context
DSC	Dice Similarity Coefficient
ET	Enhancing Tumor
FDCA	Frequency Domain Channel Attention
HD	Hausdorff Distance
MFDF	Multi-Frequency Directional Filtering
TC	Tumor Core
WT	Whole Tumor

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

J. K.: Conceptualization, Methodology, Software, Data curation, Writing-original draft. S. P.: Investigation, Supervision, Validation, Writing-review & editing. All authors had approved the final version.

DATA AVAILABILITY

The datasets used during the current study available from the corresponding author, and made available on reasonable request.

REFERENCES

[1] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[2] S. Bakas *et al.*, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Sci. Data*, vol. 4, no. 1, pp. 1–13, 2017.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[5] O. Oktay *et al.*, "Attention u-net: Learning where to look for the pancreas," arXiv Preprint, arXiv:1804.03999, 2018.

[6] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[7] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[8] A. Hatamizadeh *et al.*, "Unetr: Transformers for 3D medical image segmentation," in *Proc. the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 574–584.

[9] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, 2016.

[10] M. Satyanarayanan, "The emergence of edge computing," *Computer (Long Beach. Calif).*, vol. 50, no. 1, pp. 30–39, 2017.

[11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[12] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. International Conference on Machine Learning*, 2019, pp. 6105–6114.

[13] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[15] E. O. Brigham, *The Fast Fourier Transform and its Applications*, Prentice-Hall, Inc., 1988.

[16] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon, "Fnet: Mixing tokens with fourier transforms," in *Proc. the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 4296–4313.

[17] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 424–432.

[18] Q. Meng *et al.*, "Lightweight network with roberts edge enhancement for brain tumor segmentation," *Sci. Rep.*, vol. 15, 19816, 2025. doi: 10.1038/s41598-025-03151-z

[19] X. Lei, B. Ni, X. Tang, D. Chen, F. Chen, and Q. Wu, "A novel focus-enhanced U-net for accurate brain tumor segmentation," *Biomed. Signal Process. Control*, vol. 110, 108274, 2025. doi: 10.1016/j.bspc.2025.108274

[20] X. Hua, Z. Du, H. Yu, Z. Li, Q. Lu, and H. Zhao, "Detail-aware semantic segmentation network for brain tumor MRI images combining multi-frequency directional filtering and lifting wavelets," *Appl. Soft Comput.*, vol. 185, 113969, 2025. doi: 10.1016/j.asoc.2025.113969

[21] M. Shao *et al.*, "Rethinking brain tumor segmentation from the frequency domain perspective," *IEEE Trans. Med. Imaging*, vol. 44, no. 11, pp. 4536–4553, 2025. doi: 10.1109/TMI.2025.3579213

[22] L. Liu and K. Xia, "BTIS-Net: Efficient 3D U-Net for brain tumor image segmentation," *IEEE Access*, vol. 12, pp. 133392–133405, 2024. doi: 10.1109/ACCESS.2024.3460797

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, in Lecture Notes in Computer Science, vol. 9351. Springer, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28

[24] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proc. the European Conference on Computer Vision (ECCV)*, 2018, pp. 116–131.

[25] A. Dosovitskiy *et al.*, "An image is worth 16 × 16 words: Transformers for image recognition at scale," arXiv Preprint, arXiv:2010.11929, 2020.

[26] R. Lei, Y. Zhu, H. Liu, Y. Xu, and Q. Cai, "FE-Unet: A lightweight brain tumor segmentation network combining frequency domain processing and edge preservation," *Biomed. Signal Process. Control*, vol. 110, 108274, 2025.

[27] G. Cao *et al.*, "LCMF-Net: A lightweight collaborative multimodal fusion network for brain tumor segmentation," *Neural Networks*, 108257, 2025.

[28] N. Feng, M. Gao, and S. Ji, "DPAF-Net: DualPath-attention with hierarchical cross-scale feature fusion for 3D brain tumor segmentation," *Biomed. Signal Process. Control*, vol. 113, 109079, 2026.

[29] E. Goceri and Y. D. Winter, "CFATrans: Brain tumor segmentation from MRIs using consecutive fusion-attention transformer with convolutional networks and a composite loss function," *Biomed. Signal Process. Control*, vol. 112, 108370, 2026.
[30] J. Yang, P. Qiu, Y. Zhang, D. S. Marcus, and A. Sotiras, "D-net: Dynamic large kernel with dynamic feature fusion for volumetric medical image segmentation," *Biomed. Signal Process. Control*, vol. 113, 108837, 2026.
[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," arXiv Preprint, arXiv:1711.05101, 2017.
[32] C. Pham, V.-A. Nguyen, T. Le, D. Phung, G. Carneiro, and T.-T. Do, "Frequency attention for knowledge distillation," in *Proc. the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 2277–2286.
[33] S. Bakas *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," arXiv Preprint, arXiv:1811.02629, 2018.
[34] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.
[35] Y. Li, H. Xu, X. Zhu, X. Huang, and H. Li, "BTDet: Towards lightweight and enhanced feature aggregation network for brain tumor detection," *Biomed. Signal Process. Control*, vol. 114, 109283, 2026.
[36] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv Preprint, arXiv:1704.04861, 2017.