STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, December 2026, pp 5392–5412.
Published online in International Academic Press (www.IAPress.org)

Optimized Attention-Based U-Net with Adaptive Hybrid Loss for Accurate Brain Tumor Segmentation in MRI

Younes Kabbadj, Noura Bentaher, Mohamed Bensalah*

LabSIV, Department of Computer Science, Faculty of Science, Ibn Zohr University, Agadir, BP 8106, 80000, Morocco

Abstract
Brain tumors such as gliomas are a major concern in health care. In order to detect them effectively, practitioners inspect medical images, and MRI is one of the main tools used to analyze brain structures and detect any abnormalities. In this work, we introduce an optimized attention-based deep learning approach for glioma segmentation. Our contributions include : (1) optimized attention gates that use Hard Sigmoid activation and 1×1 convolutions to reduce computational complexity from $O(C^2)$ to $O(C)$ while preserving the discriminative capacity of the original gates. (2) An adaptive loss weighting mechanism that adjusts the balance between Dice Loss and Binary Cross-Entropy during training in order to improve learning stability, and (3) a joint segmentation-classification framework that leverages global constraints to reduce False Positives. We have validated this approach on BRATS 2018 and Lower Grade Glioma (LGG) datasets. On BRATS, we achieve a 94.84% Dice coefficient, 90.16% IoU and 96.01% precision; on LGG, 92.85% Dice, 87.15% IoU and 89.76% precision. This corresponds to the best precision on both datasets among all compared methods. We have also noted that the optimal value of the adaptive weight of the loss function, $\alpha_{\max} = 0.7$, maximizes the Intersection over Union on BRATS, and that the classification branch under this configuration produces substantially fewer false positives compared to the ablation where classification is not used.

Keywords Deep Learning, Attention Mechanisms, Brain Tumor Segmentation, Adaptive Loss Weighting, Hybrid Loss, MRI Imaging, U-Net Architecture

AMS 2010 subject classifications 68T45, 92C55

DOI: 10.19139/soic-2310-5070-4300

1. Introduction

Brain tumors such as gliomas are a major concern in health care. In order to detect them effectively, radiologists inspect medical images. Magnetic resonance Imaging is one of the main tools to analyze brain structures and detect any abnormalities including gliomas. This modality is usually used to manually segment images into homogeneous regions. Radiologists tend to use a manual delineation process in order to perform this task but this time-consuming process often induces a high variability in precision and speed, which could hinder its usefulness especially in large-scale studies. These facts are the main reasons for the development of fully automatic, fast and accurate segmentations algorithms as a computer aided detection and diagnosis tools (CAD, and CADx) [29]. Convolutional Neural Networks (CNNs) are a cornerstone in this application thanks to their ability to learn hierarchical feature representations from image data. UNet [1] is a fully convolutional deep neural network model that is intended for semantic segmentation. It features a symmetric lookalike encoder-decoder architecture with skip connections. These skip connections are the key for preserving and reinjecting the spatial details that are lost during downsampling, which allows pixel-level localization. The encoder and the decoder in the U-Net architecture capture, respectively, objects and their location within the image. However, U-Net capacity is limited by its skip

*Correspondence to: Younes Kabbadj (Email: y.kabbadj@uiz.ac.ma). Département d'informatique, Faculté des sciences Agadir,Université Ibn Zohr, B.P 8106 Cité Dakhla Agadir, Morocco

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 2026 International Academic Press

connection and the concatenation of features from encoder to decoder, as the encoder part represents fine details while the decoder represents more abstract features which creates a semantic gap between the encoder and the decoder [36]. To deal with this limitation, UNet++ [2] introduced a densely connected, nested architecture. This redesign of the skip pathways and their diagonal design in U-Net++ progressively creates a refined feature map, effectively mitigating the semantic gap between the encoder and decoder while highlighting fine structural details. An alternative strategy to resolve this limitation is the integration of attention mechanisms. Attention UNet [3] relies on two attention gates, a spatial and a channel attention gate. While the spatial attention gate focuses on spatial details, the channel attention gate computes for each channel a corresponding weight that is proportional to its importance in the context of segmentation. Those attention gates enable the model for a selective focus on feature maps. In another dimension, the optimization objective, defined by the loss function can also be critical to accuracy. Dice loss [4] is widely used for the cases of class imbalance as for brain tumors vs healthy brain tissue. It aims to maximize the overlap fraction between ground truth and prediction. Nevertheless, it might be very instable in the training process, which can induce a poor performance [5, 6, 7]. In addition, due to the generality of its equation it could also be particularly weak for detecting small details. The aim of this work is to introduce a framework that deals with the limitations of U-Net. Based on our previous work, we introduce (1) two new and optimized attention gates to preserve details and improve the network performances based on our previous U-Net based network [26]. (2) We incorporate two outputs into the network: a segmentation and a classification so that each output assists the other in the learning process. (3) We add an adaptive loss function that takes into account binary cross-entropy and dice loss for an improved performance and stability. We have validated this approach on BRATS and LGG datasets demonstrating similar to superior performances over the original and the state of art methods. The remainder of this paper is structured as follows: Section 2 reviews related work, Section 3 details the proposed methodology, Section 4 describes the experimental design, Section 5 reports the results, Section 6 discusses them, and Section 7 concludes the work.

2. Related Work

Automatic detection and segmentation of brain tumors in MRI images has become one of most important topics in biomedical image processing thanks to deep convolutional architecture revolution. U-Net architecture [1] is one of the most used architectures in this field and is particularly adapted to medical images purposes because of its ability to propagate both spatial and semantic information. However U-Net may suffer from its inability to grasp contextual information, particularly for small or irregular tumor regions [8, 9] and may need additional preprocessing or augmentation to achieve this task optimally.

2.1. Variants of U-Net: Dense and Nested Connections

Many models have tried to solve some limitations of U-Net. U-Net++[2] is one of the earliest methods to address the semantic gap between encoder and decoder by inserting diagonal skip connections into the original architecture, which enables a progressive feature maps building and permits a better localization of fine structures. Another influential extension of U-Net is V-Net [4], which adapts the U-Net encoder-decoder architecture. It incorporates residual connections within each stage to improve gradient flow and feature propagation. Other studies like Dense U-Net [23] and UNet 3+ [24] adopted ways to enhance the aggregation of multiscale features. They reported improved Dice Score and sensitivity compared to U-Net. Other authors targeted a different usage of skip connections including multi-path blocks [18, 19]. They have also reported that these enhancements can potentially preserve fine details while maintaining a high level semantic context to guide segmentation. More recently, Echine and Darouichi [38] extended this line of work to the volumetric domain with a 3D ASPP-ResUNet evaluated on BRATS 2020, which confirms that residual and multi-scale designs remain competitive for multimodal brain tumor segmentation.

2.2. Loss Functions and Optimization Challenges

In order for the U-Net architecture to train, one has to choose many hyper-parameters including a loss function. Its choice has a significant impact on the network's convergence and performance. The loss function defines the way to focus on the desired outcome of the network i.e., how to make sure that the segmentation corresponds to the ground truth during training while reaching a good generalization capacity. Dice loss [4] is often used due to its ability to maximize the overlap between predicted and ground-truth masks, which is particularly effective for discriminating small tumor regions from healthy tissue. However, the use of Dice Loss can be tricky because of its instability during training, particularly when predicted masks contain very few positive pixels and in early learning stages. Thus, Hybrid approaches that combine Dice with cross-entropy, focal, or Tversky losses have been proposed to improve both stability and sensitivity [20, 21, 22]. Beyond the choice of the overlap term itself, the objective can also be enriched with additional constraints: Aitmaten et al. [40] add a KL-divergence term on the intensity distribution of the region of interest to a U-Net and report a consistent gain over cross-entropy alone, particularly on degraded images. This supports the more general observation that the formulation of the objective function, and not only the architecture, is a lever for segmentation quality. A separate line of work focuses on how to weight the loss terms during training rather than which terms to combine. Golnari and Diba [46] track the recent history of each loss term and adjust multi-loss weights in real time for breast-cancer segmentation, instead of following a fixed schedule. Outside medical imaging, Kendall et al. [47] weight jointly trained tasks by their homoscedastic uncertainty, learning the balance between objectives end-to-end. Our adaptive weighting mechanism (Section 3.4) uses a simpler, predefined exponential schedule $\alpha(t)$ instead: it is less flexible than these learned approaches, but relies on a single, interpretable hyperparameter $\alpha_{\max}$, tuned once via grid search (Section 5.6).

2.3. Attention Mechanisms

Since their seminal work [3], attention mechanisms were extensively used to improve U-Net's ability to locate relevant features and suppress background regions. Attention U-Net [3] developed spatial and channel attentions to emphasize important features in skip connections. Other variants such as CBAM [12], SE-Net [13], or self-attention [14] have further enhanced the performance of attention mechanisms. These modules permit a particular focus on tumor regions especially in heterogeneous MRI scans. Attention-inspired models like TransUNet[15], Swin-UNet [16], and MedT[17] provide the ability to integrate global context attentions with strong performances. The caveat is that they suffer from high computational cost [10] and from their need for more data compared to U-Net [11]. This cost is precisely what motivates lightweight designs: Hmidi et al. [41] show that a carefully regularized compact CNN reaches 94.08% accuracy on multiclass brain tumor CE-MRI while running at roughly 60 ms per image on a Raspberry Pi 4, which illustrates the practical value of reducing the number of operations rather than increasing model capacity.

3. The Proposed Approach

3.1. Overview of the proposed approach

In our previous studies [26, 25], we explored two directions in medical image segmentation. The first study is a U-Net inspired network that gathers detection (i.e. the existence or not of a tumor) and segmentation (what is the shape of the tumor) in the same place. The network performs classification by using the features of the last decoder. In this approach, both processes have their own loss function and process features in their own way by adapting them to their respective purposes. The second study presents an enhanced U-Net framework that introduces two attention blocks to reduce the semantic gap between the encoder and the decoder in U-Net and to capture fine-grained boundaries. This work also uses a residual block to boost features propagation through the network by solving the gradient vanishing problem [27]. However, we have noted since that R2A-UNet's weakness lies in the way we implemented attention gates. While they are effective in highlighting relevant features, they tend to replicate many unnecessary operations across each scale leading to redundant computations and a higher training cost. To address this limitation, we propose a method that optimizes the number of operations without compromising precision.

The resulting model has the potential to preserve R2A-UNet's ability to detect important features by leveraging the power of attention gates while reducing its computational cost. The network also inherits the design of our joint classification-segmentation framework. While we introduced the latter as two independent schemes that are performed within the same network, we thought about the possibility of integrating classification and segmentation aspects in a single loss function. Indeed, classification can guide the segmentation process and act as a global constraint that can improve results in two ways. By according importance to classification through the addition of dedicated weight in the loss function, it can improve segmentation quality and reduce false positives especially in the cases where training data is insufficient. It is obvious that when an image contains no abnormality it is useless and unproductive to look for positive regions. Another potential gain is the fact that some segmentation loss functions like Dice Loss often lack stability in the early stages, which is the case when overlap between predicted and ground truth pixel is low. With a global classification loss as binary cross entropy, we can mitigate this effect and achieve a faster convergence to an optimal solution as well as consistent results with different hyper-parameters initialization. The benefit of coupling the two tasks has also been observed outside brain imaging: Abdulqader and Abdulazeez [39] combine attention-enhanced segmentation with a lightweight classification stream for lung cancer and report that the two objectives reinforce each other on a multi-dataset benchmark. In summary, such a model with optimized attention gates and the integration of classification to guide the segmentation can achieve strong performances, an improved stability during early learning steps and reduced false positives. Figure 1 is a visual illustration of the R2A-UNet with a global classification output. The model consists of a five-level encoder-decoder network for joint segmentation and classification. The encoder processes the input image through two series of residual convolutional blocks, each followed by Batch Normalization and ReLU. After the first residual convolutional block of each encoder level, an attention mechanism is applied. This mechanism produces two outputs: a channel attention map and a spatial attention map. The channel attention is built by combining global average pooling and global max pooling, passing the result through a dense layer and a sigmoid activation to obtain a set of channel weights. These weights modulate the original feature map channel by channel. Spatial attention applies a spatial mask that highlights the biggest coefficients, which usually correspond to important details. In the encoder, the spatial attention output is passed after max pooling to the next level, while the channel-filtered features are used later during decoding as Normalized Spatial Attention (NSA) in the attention gate. The encoder increases the number of filters as depth increases (64, 128, 256, 512, and 1024), and the bottleneck applies two residual blocks with 1024 filters and dropout. The decoder reconstructs the spatial resolution using transposed convolutions. At each decoding step, an attention gate receives the upsampled decoder features and the NSA and produces a filtered skip connection. This ensures that only the most relevant encoder activations are passed to the decoder. The attention gate (AG) consists of NSA features that are concatenated with the decoder output. This pattern is repeated across all four decoding levels (512, 256, 128, 64 filters). After the final decoding stage, a 1×1 convolution followed by a sigmoid activation produces the segmentation mask. The classification branch is attached to the end of the segmentation path. It receives the final segmentation features, applies Batch Normalization and a 3×3 convolution with 32 filters, and compresses the output into a single vector using Global Average Pooling. A Dense layer with 16 units and a LeakyReLU activation extracts global information, and a final sigmoid neuron predicts whether the input image contains a tumor.

3.2. Channel attention optimization

3.2.1. Original formulation In previous work [26], we formulated a normalized channel attention (NCA) module that weights and adapts each feature channel's discriminative ability by learning inter-channel dependencies. Given that $X \in \mathbb{R}^{H\times W\times C}$ is the input feature map where H, W, and C denote height, width, and number of channels respectively, NCA relies on computing two descriptors:

$$f_{\text{avg}}(c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{i,j,c}, \quad f_{\max}(c) = \max_{i,j}(X_{i,j,c}) \tag{1}$$

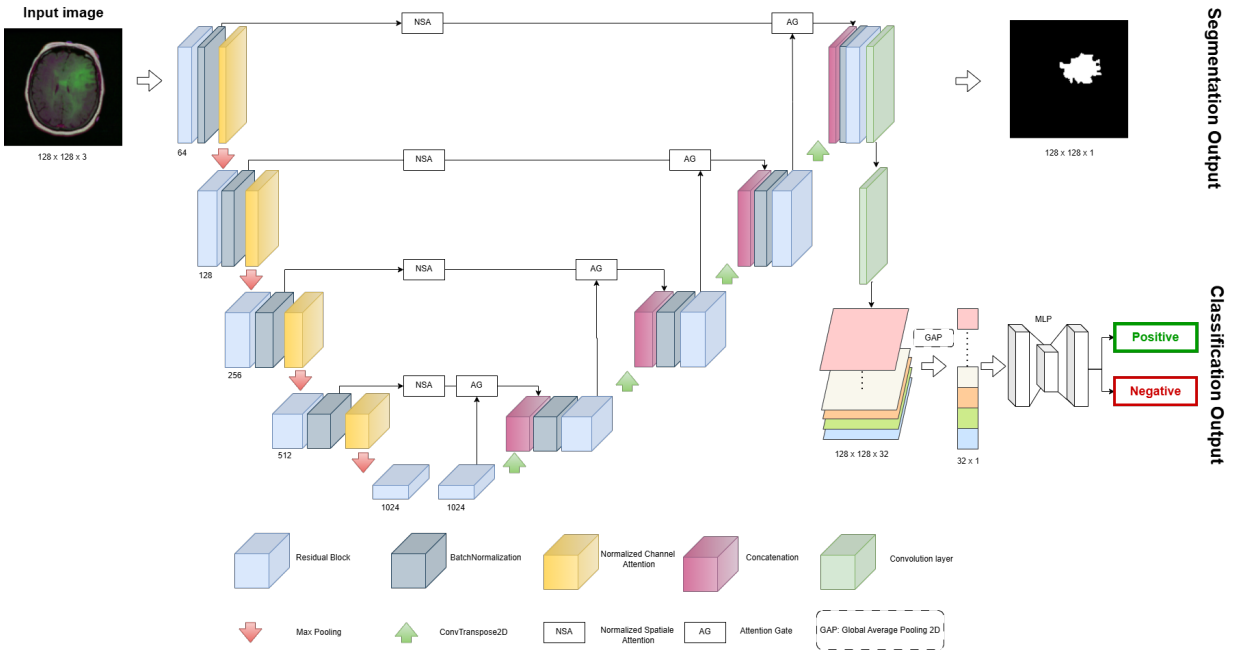


Figure 1. Diagram of the proposed method showing the encoder-decoder architecture with optimized attention mechanisms and joint classification-segmentation outputs.

Thus $f \in \mathbb{R}^{C\times 1\times 1}$. The two descriptors are summed into a single channel descriptor, passed through a fully connected layer, and normalized by sigmoid activation:

$$\beta = \sigma(W f_{\text{sum}} + b) \tag{2}$$

where $W \in \mathbb{R}^{C\times C}$. The resulting weights in β rescale the input feature map:

$$X' = \beta \odot X \tag{3}$$

Computational cost: The dense layer projection involves C^2 multiplications, resulting in $O(C^2)$ complexity.

3.2.2. The optimized formulation of channel attention Standard sigmoid function is extensively used in deep learning especially in classification layers. While it is derivable and straightforward, it is computationally costly as it involves an exponential term. This term needs iterative algorithms as polynomial approximations, which in turn need more operations than simple additions or multiplications. Hard sigmoid is an alternative that allows the same scaling functionality while reducing the number of operations. Hard sigmoid is defined as:

$$\text{HardSigmoid}(x) = \max\big(0,\ \min(1,\ 0.2x + 0.5)\big) \tag{4}$$

In our experiments using the Numpy package, we have noted that hard sigmoid is at least 1.5 times faster than the conventional sigmoid, while preserving a very similar scaling profile (Figure 2). Additionally, we propose to replace the dense layer with 1×1 convolutions, which involve a single parameter per vector element rather than C^2, so the relative computational cost is reduced from $O(C^2)$ to $O(C)$. The optimized reformulation becomes:

$$\tilde{\beta} = \text{HardSigmoid}(\text{Conv}_{1\times 1}(f_{\text{sum}}))\,, \quad \tilde{X} = \tilde{\beta} \odot X \tag{5}$$

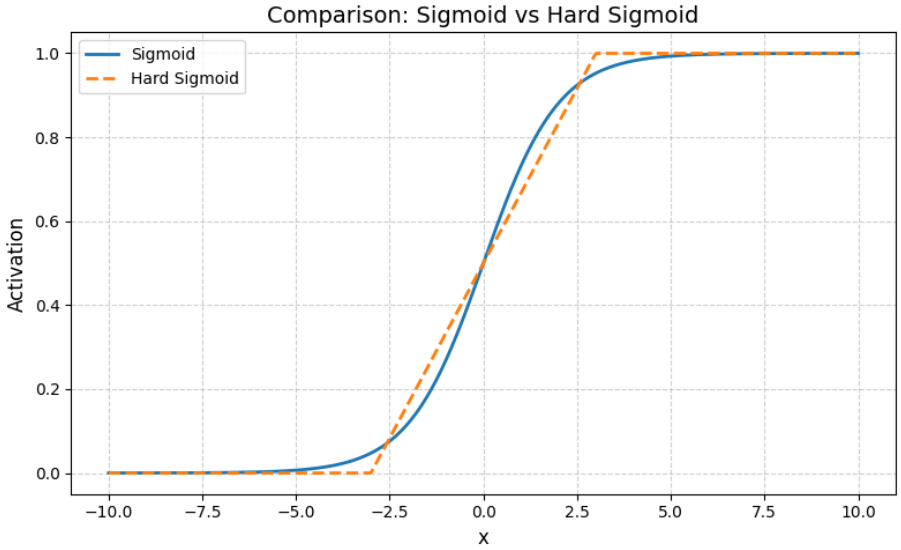


Figure 2. Comparison of sigmoid and hard sigmoid activation functions showing computational efficiency gain while maintaining similar scaling behavior.

3.3. Spatial attention optimization

3.3.1. Original formulation of spatial attention The spatial attention mechanism aims to enhance and highlight spatial regions with high coefficients corresponding to strong responses from previous convolutional layers. These regions typically contain important details like tumors. Given input feature map $X' \in \mathbb{R}^{H\times W\times C}$, the spatial mechanism enhances informative regions by computing weights across spatial dimensions (height and width). Two spatial descriptors are extracted:

$$M_{\text{avg}}(i,j) = \frac{1}{C}\sum_{k=1}^{C} X'_{i,j,k}, \quad M_{\max}(i,j) = \max_{k=1,\ldots,C} X'_{i,j,k} \tag{6}$$

These descriptors are concatenated into a single descriptor:

$$M = [M_{\text{avg}}; M_{\max}] \in \mathbb{R}^{H\times W\times 2} \tag{7}$$

A convolution with kernel size $k \times k$ followed by sigmoid activation produces the spatial attention map:

$$\gamma = \sigma(\text{Conv}_{k\times k}(M)) \tag{8}$$

This highlights meaningful regions within the feature map. Finally, features are modulated:

$$X'' = \gamma \odot X' \tag{9}$$

The computational complexity is approximately $O(HWk^2)$.

3.3.2. Optimized formulation To simplify spatial attention while maintaining discriminative capability, a single map replaces concatenated pooled feature maps:

$$M = \alpha M_{\text{avg}} + (1-\alpha)M_{\max} \tag{10}$$

where $\alpha \in [0,1]$ is a trainable parameter balancing max-pooling versus average-pooling importance. The resulting feature map is processed by 1×1 convolution followed by Hard Sigmoid activation. Single weighted maps

halve intermediate operations, and hard sigmoid reduces computing time by eliminating exponential function approximations. The final transformed feature map equation becomes:

$$\tilde{\gamma} = \text{HardSigmoid}(\text{Conv}_{1\times1}(M)), \quad \tilde{X}'' = \tilde{\gamma} \odot X' \tag{11}$$

3.4. Adaptive Loss Weighting for Stability and False Positive Reduction

The number of false positives as well as false negatives remains one of the main challenges in image segmentation, especially when only a limited number of annotated images is available. When the number of positive pixels (e.g. tumors) is significantly lower than the background, the model tends to overfit in favor of the background creating false negatives [5, 37]. Additionally, in most cases, most positive cases consist of a single tumor. However the network, can produce a set of diffuse positive regions. Those detections could be problematic to the diagnosis, as they can lead to unnecessary biopsies. Therefore, our aim is to use a suitable loss function that combines efficiency with stability, which is crucial for a reliable segmentation.

Dice loss is particularly adapted for the cases of class imbalances [30]. It has become one of the most popular loss functions in medical image segmentation thanks to its ability to measure the overlap accuracy between the predicted regions and the ground truth. It is defined as:

$$L_{\text{Dice}} = 1 - \frac{2\sum_i p_i g_i + \epsilon}{\sum_i p_i^2 + \sum_i g_i^2 + \epsilon} \tag{12}$$

where p_i and g_i are predicted and ground-truth pixel values, respectively, and ϵ is a positive constant. Dice loss is particularly effective for imbalanced problems since it focuses on region overlap, measuring predicted pixels coinciding with ground truth, without being influenced by dominant classes. However, Dice loss suffers from instability during early training stages [31, 32]. In such cases, the denominator is still very small and can cause very large variations of the gradient and a slow convergence. It is important to clarify at this point how the two terms of the hybrid objective are actually computed in our architecture: the Dice term above is produced by the segmentation decoder and evaluated pixel-by-pixel between predicted and ground-truth masks, whereas the Binary Cross-Entropy term introduced below is produced by the classification branch and evaluated at the image level. Consequently, in Eq. (13) the index i ranges over images in the batch and N denotes the number of images, not the number of pixels. As opposed to Dice Loss, **Binary Cross-Entropy (BCE)** provides more stable gradients precisely because it does not depend on pixel-wise overlap: it is computed once per image by the classification branch:

$$L_{\text{BCE}} = -\frac{1}{N}\sum_i [g_i \log(p_i) + (1-g_i)\log(1-p_i)] \tag{13}$$

This architectural coupling changes how the hybrid loss should be interpreted: the Dice term comes from the segmentation decoder and the Binary Cross-Entropy term from the classification branch, so the adaptive weighting described below does not just balance two views of the same pixel map, it also controls the relative influence of the two branches. Section 5.7 quantifies this effect through the resulting reduction in false positives.

We propose a combination of Dice Loss and BCE through a hybrid objective:

$$L_{\text{total}} = \alpha L_{\text{Dice}} + \beta L_{\text{BCE}} \tag{14}$$

where α and β are positive coefficients that determine the relative contribution of each component. While this approach integrates the potential stability of BCE and spatial precision of Dice, using fixed weights during training may lead to some limitations. Dice loss is known to be unstable during early training phases, therefore assigning it a large weight can amplify gradient effects and increase false positive rate. Conversely, assigning a high weight to BCE for the entire training will very likely lead to smoothing and loss of boundaries. In such a way, balance between the two terms can evolve over time as the network learns the basic representations of the tumors and background and refines them.

To overcome this limitation, we propose a dynamic weighting mechanism that adjusts the importance of each loss during training. The idea is to start with a small weight for Dice Loss which allows BCE to dominate and

stabilize the model in early iterations. Then, the Dice Loss contribution will increase gradually as the network learns more. Here is the formulation of the Dice-associated weight:

$$\alpha(t) = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min})\big(1 - e^{-\gamma t}\big), \quad \beta(t) = 1 - \alpha(t) \tag{15}$$

where α and β are Dice and BCE weights, t is current training epoch, $\alpha_{\min}$ and $\alpha_{\max}$ are initial and final Dice weights, and γ controls evolution speed. In short, we propose exponentially evolving weights. The objective function becomes:

$$L_{\text{total}}(t) = \alpha(t)L_{\text{Dice}} + \beta(t)L_{\text{BCE}} \tag{16}$$

In summary, our hybrid loss function balances Dice and BCE dynamically. By leveraging each approach's strengths throughout training, we stabilize the objective function early while controlling false positive rates and providing reliable segmentation.

4. Experimental Design

In this section, we describe the experimental design that we have used in order to test and evaluate performances and robustness of the proposed segmentation model. We have designed the experiments to assess how the model performs across different datasets. For this purpose, we have employed two different datasets: BRATS 2018, a dataset that includes multimodal MRI scans with pixel level annotations of brain tumors, and the Lower Grade Glioma (LGG) dataset, which contains heterogeneous and scattered lesions compared to BRATS. We have normalized to [0,1] and augmented both datasets to have more homogeneous scans and sufficient data for training. In the following sections, we describe the datasets, training environment and evaluation metrics that we have chosen for performance evaluation.

4.1. BRATS 2018 Dataset

Brain Tumor Segmentation (**BRATS 2018**) dataset is a standard for medical image segmentation and evaluation algorithms [33, 34, 35]. It consists of multimodal MRI scans from patients with low grade or high-grade gliomas. Every scan includes four MRI sequences: T1-weighted, T1 contrast-enhanced (T1ce), T2-weighted, and FLAIR. These sequences provide structural and pathological information. While T1ce is likely to be efficient with tumor regions, T2 and FLAIR can accentuate edema and infiltrative areas. Finally, T1 provides anatomical details. The BRATS 2018 dataset contains 285 patient cases (210 high grade and 75 lower grade gliomas). It consists of 7,125 MRI slices, including 5,042 tumor-positive and 2,083 tumor-negative images. Following the patient-level protocol described above, it is split into 80% for training, 10% for validation and 10% for testing.

4.2. LGG Dataset

We have also used the Lower Grade Glioma (LGG) dataset [48]. This dataset is more challenging than BRATS since tumors are in general smaller, less contrasted and more scattered than those in BRATS. Boundaries between tumors and healthy tissue are also not as well-defined as in BRATS, which can make segmentation more difficult [48]. The LGG (TCGA-LGG) dataset contains 110 patient cases of lower-grade glioma that were collected from five medical institutions. We processed the MRI scans to 128×128 pixels consisting of a total of 3,929 slices. This dataset contains 2,556 tumor-positive and 1,373 tumor-negative images. As for BRATS, this dataset is divided into 80% training, 10% validation and 10% testing at the patient level.

4.3. Training Parameters and Setup

All experiments were implemented using the Keras deep learning library with TensorFlow backend, and conducted on the Kaggle GPU environment (Tesla P100, 16GB dedicated memory). We used the Adam optimizer with an initial learning rate of 0.001. Training was run for up to 150 epochs with a batch size of 16, using early stopping (Keras EarlyStopping callback, monitor = val_loss, patience = 10). The adaptive loss schedule used

$\alpha_{\min} = 0.1, \alpha_{\max} = 0.7$ (selected via grid search, Section 5.6), and $\gamma = 0.05$. Consistent with our prior work [26], data augmentation was applied to both datasets, comprising random rotations up to 20%, width and height shifts up to 5%, shear transformations up to 5%, zoom adjustments up to 5%, horizontal flips, and nearest-fill mode for newly introduced pixels. All input images were normalized to [0, 1]. Train/validation/test splits were performed at the patient level (seed = 42) in order to prevent any data leakage between slices of the same patient. For LGG, this corresponds to 88/11/11 patients (80/10/10), i.e. 3,199/357/373 slices; for BRATS 2018, to 228/28/29 patients (80/10/10), combining HGG and LGG cases from the BRATS 2018 training set. Because the number of slices varies from one patient to another, the resulting slice-level ratios differ slightly from an exact 80/10/10 partition; in all cases, no slice from a given patient appears in more than one split.

4.4. Segmentation Evaluation Metrics

To quantify segmentation quality, beyond Dice Coefficient, we chose the **Intersection over Union (IoU)** metric. Given that P is the segmentation mask and G the ground truth, IoU is defined as:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} \tag{17}$$

Dice Coefficient is suited for learning while IoU is more accurate for segmentation quality evaluation. Indeed, IoU could produce zero gradients which are inconvenient for learning [28]. Conversely, Dice Coefficient typically returns higher values when segmenting while weakly penalizing false positives and false negatives because it associates a higher weight to the intersection. In the reformulation in terms of TP, FP and FN:

$$\text{Dice} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{18}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{19}$$

TP is therefore more dominant in the Dice formula which results in a bigger overlap value for Dice, while IoU penalizes errors more strictly.

5. Results

This section presents a comprehensive quantitative and qualitative evaluation of the proposed Optimized R2A-UNet on the BRATS 2018 and LGG datasets. We report segmentation performance against state-of-the-art methods, analyze the influence of the adaptive loss hyperparameter $\alpha_{\max}$ on segmentation quality, evaluate the classification branch via confusion matrices, and provide visual comparisons and statistical robustness analyses.

5.1. Computational Efficiency and Training Time

The computational improvements result primarily from the optimized attention mechanisms. Channel attention complexity dropped from $O(C^2)$ to $O(C)$ through 1×1 convolutions instead of dense layers, and spatial attention was simplified by replacing concatenated pooling with a trainable parameter. We also replaced the classical sigmoid by hard sigmoid. As a result, the number of parameters is reduced from 68.0 million to 66.7 million. This gain translates into a 10% reduction of the time per iteration, from 2.92 s to 2.63 s per step. The total training time decreases from 6 h 00 min to 5 h 13 min, i.e. a 13.1% reduction, which stems both from the shorter time per epoch and from the smaller number of epochs required to converge (110 instead of 120). Table 1 summarizes these measurements.

Reviewer A additionally requested a broader efficiency comparison that includes FLOPs, per-image inference time, and a lighter architectural baseline. Table 2 reports these three metrics for Attention U-Net [3], the original R2A-UNet, and the proposed Optimized R2A-UNet, measured with a dedicated profiling script (parameter count, analytic FLOPs for one forward pass, and wall-clock inference time averaged over 100 runs after GPU warm-up)

Metric	Original R2A-UNet	Optimized R2A-UNet	Reduction
Number of parameters	68.0 M	66.7 M	1.9%
Time per iteration (s)	2.92	2.63	10.0%
Total training time	6 h 00 min	5 h 13 min	13.1%
Epochs to convergence	120	110	8.3%

Table 1. Computational efficiency and training time comparison on the Kaggle GPU environment (Tesla P100, 16 GB). Both models were trained under identical settings.

on the same Kaggle Tesla P100 GPU used for training. On all three metrics the Optimized R2A-UNet improves over the original R2A-UNet, while remaining heavier than the much shallower Attention U-Net, due to its residual and recurrent blocks.

Metric	Attention U-Net	Original R2A-UNet	Optimized R2A-UNet
Number of parameters	8.14 M	68.0 M	66.7 M
FLOPs (1 forward pass)	22.81 G	60.03 G	54.03 G
Inference time per image (ms)	59.60 ± 3.29	82.54 ± 5.47	74.67 ± 3.32

Table 2. Extended computational cost comparison (parameters, FLOPs for one forward pass, and per-image inference time). Inference time measured on batch size 1, Kaggle Tesla P100 (16 GB), averaged over 100 runs after warm-up.

5.2. Ablation Study

To isolate the contribution of each proposed component, we retrained the network by adding one modification at a time to the U-Net baseline: the Hard Sigmoid attention gates alone, the optimized spatial attention alone, the data augmentation pipeline alone, and the combination of the classification branch with the adaptive hybrid loss $\alpha(t)$. Table 3 reports Dice and IoU for each configuration on both the LGG and BRATS datasets; the full Optimized R2A-UNet, which combines all components, is reported for reference on the last row.

Configuration	LGG Dice	LGG IoU	BRATS Dice	BRATS IoU
U-Net (baseline)	90.02%	82.04%	78.22%	67.80%
Hard Sigmoid alone	90.07%	81.93%	83.43%	80.60%
Spatial attention alone	90.58%	83.04%	86.75%	83.20%
Data augmentation alone	89.09%	83.53%	91.25%	87.46%
Classification branch + adaptive loss $\alpha(t)$	91.86%	84.50%	91.26%	88.05%
Optimized R2A-UNet (full model)	**92.85%**	**87.15%**	**94.84%**	**90.16%**

Table 3. Ablation study: contribution of each component, evaluated independently on top of the U-Net baseline.

Every individual component improves over the baseline on BRATS. On LGG, however, data augmentation alone lowers Dice below the U-Net baseline (89.09% vs. 90.02%) while still improving IoU (83.53% vs. 82.04%). We attribute this apparent anomaly to boundary noise: geometric augmentations (rotation, shift, shear, zoom) perturb lesion contours in the smaller and more heterogeneous LGG dataset, which penalizes the boundary-sensitive Dice coefficient more than the more tolerant IoU, and this effect is only corrected once the attention mechanisms are present to compensate for it. This observation motivates combining augmentation with attention rather than relying on either mechanism alone, as the full model does.

5.3. Comparison with Additional State-of-the-Art Architectures

Beyond the U-Net family already reported in Tables 5 and 6, we further position the proposed model against nnU-Net [42] and two transformer-based architectures, TransUNet [15] and UNETR [44]. For nnU-Net, the

BRATS 2018 figures come directly from the original challenge submission of Isensee et al. [42]; the LGG figure is taken from an independent 2D nnU-Net baseline trained by Mostafa et al. [43] on the same LGG-MRI-Segmentation dataset under a patient-level split. For TransUNet and UNETR, no result on either of our two datasets under our own protocol was available; we therefore report the third-party BRATS 2018 benchmark of Sun et al. [45], who re-implemented and evaluated both architectures. Table 4 summarizes these figures alongside their comparability status.

Model	**Dataset**	**Dice (%)**	**Comparability**
nnU-Net [42]	BRATS 2018 (WT / TC / ET)	87.81 / 80.62 / 77.88	Direct
nnU-Net [43]	LGG	84.50 (mean)	Direct
TransUNet [15]	BRATS 2018 (WT / TC / ET)	74.32 / 66.81 / 65.32	Indirect†
UNETR [44]	BRATS 2018 (WT / TC / ET)	73.44 / 66.48 / 66.48	Indirect†

Table 4. Comparison with nnU-Net, TransUNet, and UNETR. †TransUNet and UNETR figures are taken from a third-party re-implementation [45] and evaluated under a different protocol; see caveat below.

We stress a methodological caveat on the two indirect comparisons: in the same benchmark [45], the reported U-Net baseline reaches only 73.75% WT Dice on BRATS 2018, markedly lower than the 78.22% Dice obtained by our own U-Net baseline (Table 6) under identical preprocessing and training settings to the rest of this paper. This gap indicates a different evaluation protocol (e.g. preprocessing, split, or sub-region definition), so the TransUNet and UNETR figures should not be read as a strictly controlled, head-to-head comparison with our model. No TransUNet or UNETR result is available on the LGG dataset. We report these numbers because they position our model, directly comparable to nnU-Net, against transformer-based architectures as well, with the caveat above stated explicitly.

5.4. Quantitative Segmentation Performance

Tables 5 and 6 report Dice coefficient, IoU, Precision, and Recall for all compared methods on the LGG and BRATS datasets, respectively.

Model	Dice	IoU	Precision	Recall
U-Net [1]	90.02	82.04	85.00	86.03
V-Net [4]	91.55	84.56	85.11	81.89
U-Net++ [2]	90.43	82.7	82.63	82.91
Attention U-Net [3]	89.20	81.11	85.11	85.07
R2A-UNet [26]	92.43	86.22	88.32	82.68
Optimized R2A-UNet	**92.85**	**87.15**	**89.76**	82.53

Table 5. Quantitative results for the LGG Dataset. Best results are highlighted in bold.

Model	Dice	IoU	Precision	Recall
U-Net [1]	78.22	67.80	82.03	78.26
V-Net [4]	86.21	77.88	90.21	83.46
U-Net++ [2]	82.69	72.7	82.98	82.45
Attention U-Net [3]	87.10	79.38	91.85	83.64
R2A-UNet [26]	**94.9**	**90.32**	95.66	**94.13**
Optimized R2A-UNet	94.84	90.16	**96.01**	93.07

Table 6. Quantitative results for the BRATS Dataset. Best results are highlighted in bold.

On the LGG dataset (Table 5), the Optimized R2A-UNet achieves the best results across all metrics except Recall: 92.85% Dice, 87.15% IoU, and 89.76% Precision. Compared to the previous R2A-UNet, the proposed

optimizations yield gains of $+0.42\%$ in Dice, $+0.93\%$ in IoU, and $+1.44\%$ in Precision, while Recall remains nearly identical (-0.15%). However, when compared to U-Net, improvements are more pronounced: $+2.83\%$ Dice, $+5.11\%$ IoU, and $+4.76\%$ Precision, which reflects the benefits of attention and loss function improvements.

On the BRATS dataset (Table 6), we achieved 94.84% Dice, 90.16% IoU, and **96.01% Precision**, the highest precision across all methods. While Dice and IoU are slightly below R2A-UNet (-0.06% and -0.16%, respectively), Precision improves by $+0.35\%$. By using classification-segmentation design and adaptive loss weighting, the model sacrifices a negligible amount of true-positive overlap in favour of a substantially cleaner prediction map. Compared to U-Net, the gains are even more pronounced: $+16.62\%$ Dice, $+22.36\%$ IoU, and $+13.98\%$ Precision.

5.5. Robustness to Training Stochasticity (Multi-Seed Analysis)

To assess whether the improvements reported above are robust to training stochasticity rather than an artifact of a single run, we retrained R2A-UNet and the proposed Optimized R2A-UNet five times each ($N = 5$), varying only the training seed (weight initialization, batch shuffling order, and data augmentation draws) while keeping the patient-level train/validation/test split fixed (seed $= 42$, Section 4) across all runs. Table 7 reports the mean $\pm$ standard deviation over the five runs for both models on both datasets. These mean values differ slightly from the single-run figures reported in Tables 5 and 6 due to normal training variability; Table 7 is the more statistically reliable estimate for these two models. This multi-seed analysis was conducted for R2A-UNet and Optimized R2A-UNet only; the remaining baselines in Tables 5 and 6 still reflect single training runs.

Dataset	Model	Dice (%)	IoU (%)	Precision (%)	Recall (%)	N
LGG	R2A-UNet	91.99 ± 0.63	87.42 ± 0.71	90.32 ± 0.20	82.65 ± 0.25	5
LGG	Optimized R2A-UNet	92.56 ± 0.31	87.63 ± 0.09	91.03 ± 0.07	83.01 ± 0.11	5
BRATS	R2A-UNet	94.27 ± 0.26	90.25 ± 0.09	95.72 ± 0.20	92.98 ± 0.60	5
BRATS	Optimized R2A-UNet	95.02 ± 0.08	89.99 ± 0.25	96.53 ± 0.13	92.09 ± 0.32	5

Table 7. Multi-seed robustness: mean $\pm$ standard deviation over $N = 5$ independent training runs (fixed data split, varying training seed).

We tested whether the differences between R2A-UNet and Optimized R2A-UNet are statistically significant using Welch's unpaired t-test (unequal variances, $n_1 = n_2 = 5$). On LGG, the Dice ($p = 0.121$) and IoU ($p = 0.547$) differences are not statistically significant, while the Precision ($p < 0.001$) and Recall ($p = 0.029$) differences are both significant in favor of Optimized R2A-UNet. On BRATS, the Dice ($p = 0.002$) and Precision ($p < 0.001$) differences are significant in favor of Optimized R2A-UNet, the IoU difference is not significant ($p = 0.080$), and the Recall difference is significant in favor of R2A-UNet ($p = 0.026$). Overall, the proposed optimizations improve Dice on BRATS, Precision on both datasets, and Recall on LGG, at a significant cost only in Recall on BRATS, while the IoU differences remain within the noise of training stochasticity on both datasets.

5.6. Influence of $\alpha_{\max}$ on Segmentation Quality

Figure 3 shows the IoU as a function of $\alpha_{\max} \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ on the BRATS dataset. The IoU curve peaks at $\alpha_{\max} = 0.7$, yielding 90.16%. Two opposite regimes explain this behaviour. For low values ($\alpha_{\max} < 0.5$), the Dice term keeps a weak weight throughout training, so the model relies almost exclusively on Binary Cross-Entropy, which does not explicitly optimize spatial overlap; the IoU therefore remains below its optimum. Conversely, for high values ($\alpha_{\max} > 0.7$), the Dice term dominates from the early epochs onward, when the predicted overlap is still very small; the resulting gradient instability increases the number of errors and again lowers the IoU. The value $\alpha_{\max} = 0.7$ therefore provides the best compromise between early BCE stability and late Dice spatial refinement.

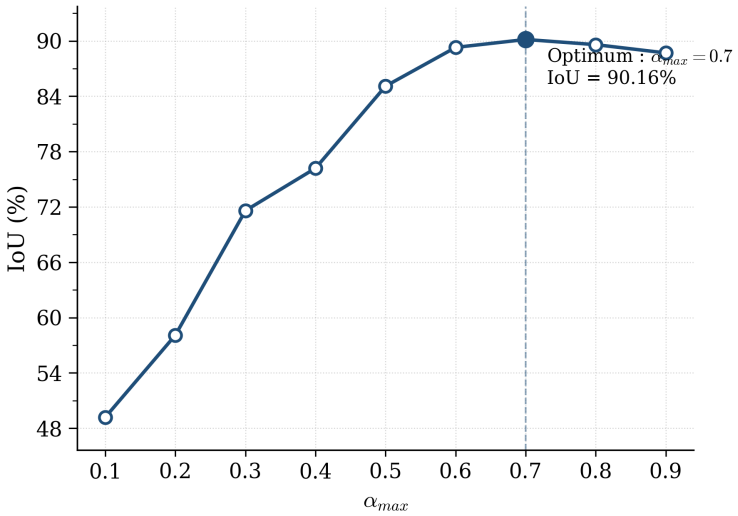


Figure 3. Influence of the adaptive loss parameter $\alpha_{\max}$ on IoU for BRATS dataset.

5.7. Classification branch contribution to the model's accuracy

As explained in Section 3.4, the Binary Cross-Entropy term of the adaptive hybrid loss is produced by this classification branch and is therefore evaluated at the image level rather than pixel-wise; the confusion matrices below quantify the resulting reduction in false positives.

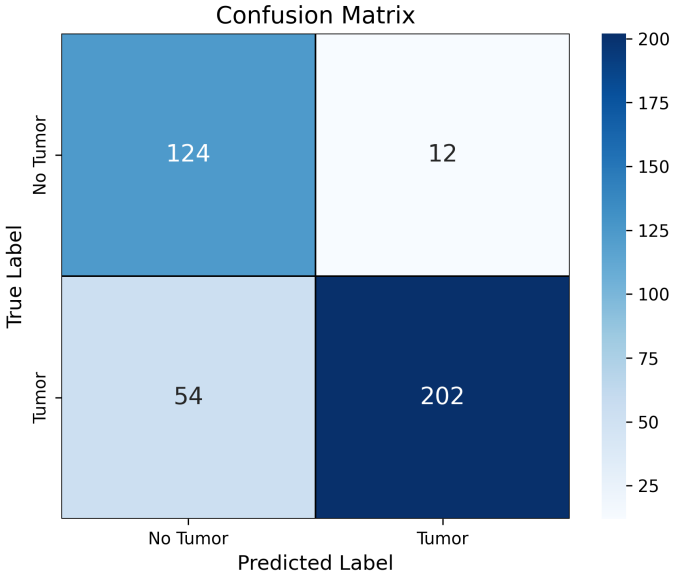


Figure 4. Image-level confusion matrix of the proposed joint model on the LGG test set ($\alpha_{\max} = 0.7$).

In order to quantify the contribution of the classification branch, we do not report a classification confusion matrix in the usual sense: the four cells summarize *detection outcomes at the image level*, i.e. whether the model produces a tumor region in a slice that does or does not contain a glioma. Test images are therefore grouped as follows:

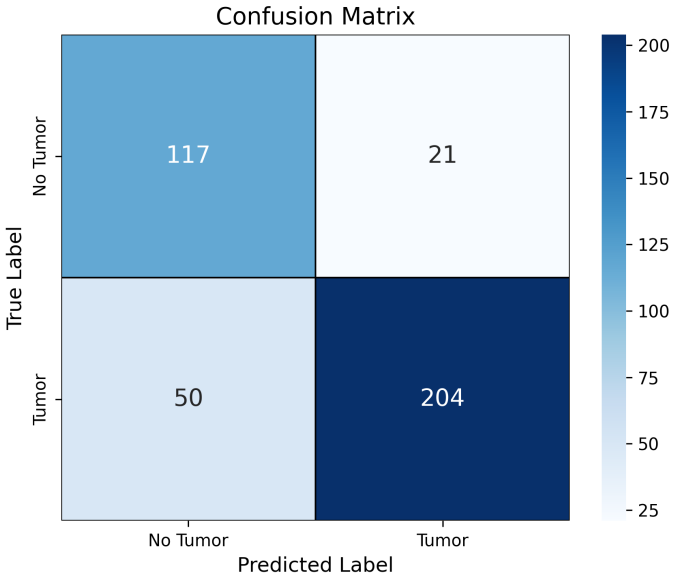


Figure 5. Image-level confusion matrix of the ablation without the classification branch on the LGG test set ($\alpha_{\max} = 1$, i.e. pure Dice loss).

- **TP**: images containing a glioma for which the model produces a segmentation;
- **FP**: images without any glioma for which the model nevertheless segments a region;
- **TN**: images without any glioma for which the model produces no segmentation;
- **FN**: images containing a glioma that the model fails to segment.

Figures 4 and 5 present these matrices under two configurations: the full proposed model with $\alpha_{\max} = 0.7$, and an ablation in which the classification branch is disabled ($\alpha_{\max} = 1$, equivalent to training with the Dice loss only).

With $\alpha_{\max} = 0.7$ (Figure 4), the joint model reaches TP $= 202$, TN $= 124$, FP $= 12$ and FN $= 54$, whereas the ablation without classification (Figure 5) yields TP $= 204$, TN $= 117$, FP $= 21$ and FN $= 50$. True positives and true negatives dominate both matrices, but the number of false positives is reduced by 9 (a 43% relative decrease) when the classification branch and the adaptive Dice–BCE weighting are used, at the cost of 4 additional false negatives. The total number of detection errors therefore drops from 71 to 66. This supports interpreting the classification branch as a global constraint that suppresses spurious detections in tumor-free slices.

5.8. Qualitative Segmentation Results

Qualitative inspection of BRATS results (Figure 6) confirms the quantitative gains. The Optimized R2A-UNet produces segmentation masks with better defined tumor boundaries. Compared to U-Net, which generates coarse segmentations, our method achieves regular shapes closer to ground truth. The results are comparable to those of V-Net and R2A-UNet.

On LGG (Figure 7), the challenge is different. LGG tumors exhibit sparse and infiltrative patterns with complex boundaries that blend strongly into healthy tissue. Despite this, our method recovers tumor shapes with fidelity.

5.9. Statistical Robustness

Box plots in Figures 8 and 9 display the proposed model's statistical robustness against the other tested methods. For U-Net, the Dice distribution spans a wide range with numerous low-score outliers. As for the other methods: V-Net, U-Net++, Attention U-Net, and R2A-UNet, the interquartile range (IQR) is tighter. The Optimized R2A-UNet shows one of the most compact Dice distributions: the IQR is narrow, the median sits in the range of the best competing methods, and the number of low-score outliers is two. Even in worst-case scenarios, the minimal Dice

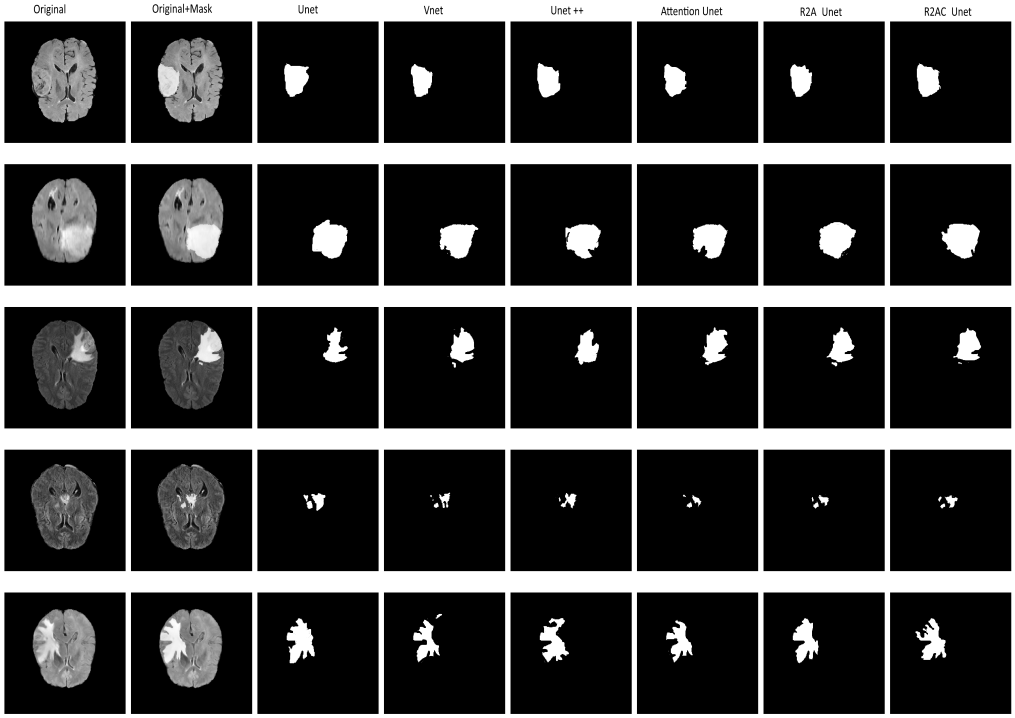


Figure 6. Visual comparison of segmentation results on representative BRATS dataset cases. From left to right: input MRI slice, ground truth mask, U-Net prediction, R2A-UNet prediction, and Optimized R2A-UNet prediction. The proposed method produces clean, well-defined masks with minimal false positives, closely matching ground truth boundaries.

value of the proposed model remains around 80%, a threshold below which all other methods except R2A-UNet fall. This statistical robustness is a direct consequence of the adaptive loss mechanism combined with spatial and channel attention.

5.10. Training Dynamics

Figure 10 shows training and validation learning curves on LGG, illustrating the effect of the adaptive loss schedule. In the first 20 to 40 epochs, with $\alpha(t) \approx \alpha_{\min} = 0.1$, Binary Cross-Entropy dominates giving the model a stable but still under average Dice score. The training loss decreases smoothly and monotonically, and the validation loss follows closely—a hallmark of stable, well-regularized optimization. Around epoch 40, as $\alpha(t)$ begins to rise, the Dice coefficient improves rapidly from approximately 0.65 to 0.8 which reflects that the model is now able to transition from a coarse to a more precise segmentation thanks to the Dice term. Both training and validation metrics increase in tandem during this phase, which indicate a good generalization rather than memorization. By epoch 80–100, the model reaches a plateau near 0.90 Dice on validation, with a small persistent gap between training and validation, which confirms good generalization on unseen cases. As for the accuracy, represented here by the binary cross-entropy term, the model reaches a maximum of about 95% around the 40th epoch. The gap in BCE between training and validation widens as training progresses, which is expected since the objective is increasingly geared toward segmentation as the Dice weight grows from the 40th epoch onward.

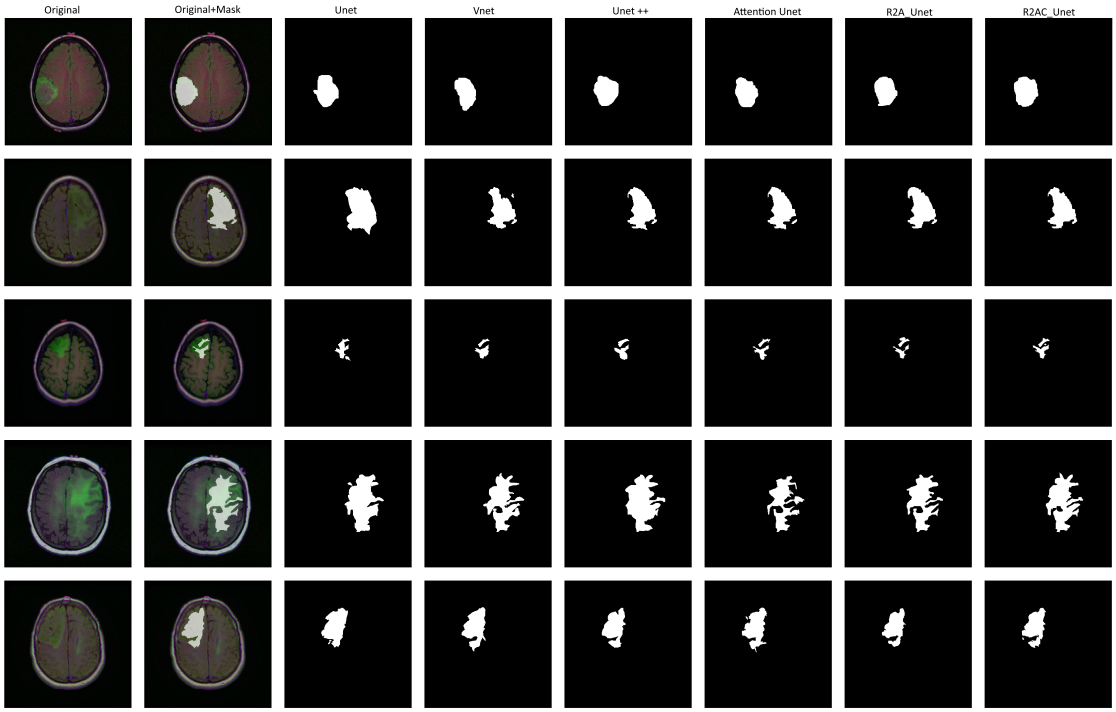


Figure 7. Visual comparison of segmentation results on representative LGG dataset cases showing challenging diffuse tumor patterns. The Optimized R2A-UNet successfully recovers faint boundaries and infiltrative regions that baseline models miss or over-segment.

6. Discussion

The present work addresses three important challenges frequently encountered in medical image segmentation: preserving segmentation accuracy while reducing computational complexity, improving training stability throughout optimization, and limiting false positive detections through a hybrid segmentation and classification loss function. Experimental results show that this strategy maintains state-of-the-art segmentation performance while considerably simplifying the attention modules. This approach may represent a serious alternative to continuously increasing model complexity. One of the most important contributions of this work concerns the redesigned attention mechanisms. Most recent segmentation architectures attempt to improve feature extraction by introducing increasingly sophisticated attention modules, nested skip connections, or transformer encoders capable of modeling long-range dependencies. Although these methods often achieve excellent segmentation accuracy, they usually do so at the expense of computational efficiency and memory consumption. Such architectures generally require larger datasets, longer training times, and more powerful hardware, which limits their practical applicability in many clinical environments. In contrast, the proposed attention modules simplify the computation by replacing dense channel transformations with 1×1 convolutions and by approximating the conventional sigmoid activation with Hard Sigmoid. At first, one could expect these simplifications to reduce the discriminative capability of the attention mechanism. However, the experimental results demonstrate that this is not the case. A possible explanation is that the primary objective of attention is not to generate complex representations but rather to assign relative importance to already extracted features. In other words, attention acts mainly as a feature weighting mechanism. As long as the relative ordering between informative and non-informative channels is preserved, an activation function remains sufficient to emphasize the most relevant structures. Similarly, replacing fully connected transformations with lightweight pointwise convolutions still allows the network to learn channel-specific importance while significantly reducing computational complexity. This interpretation explains why the optimized attention gates

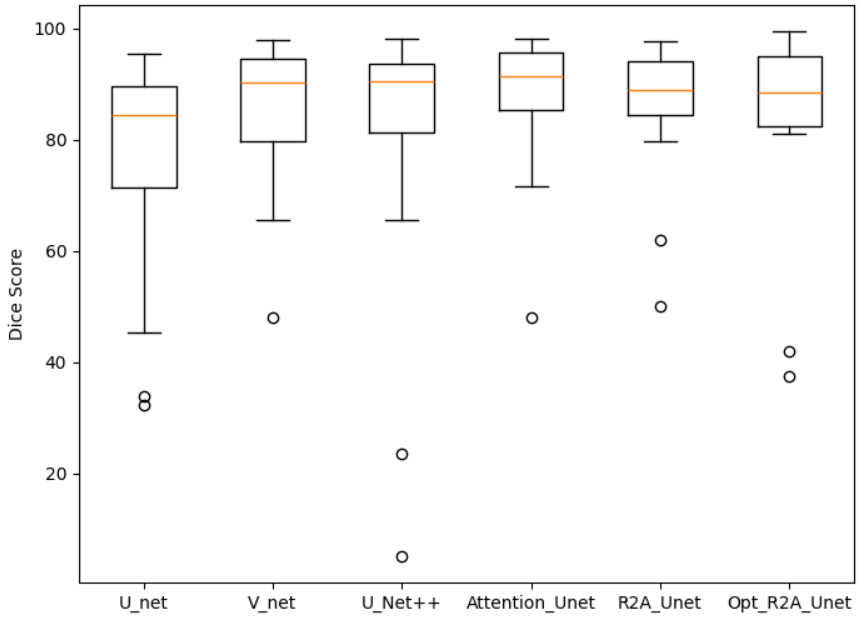


Figure 8. Box plot of Dice scores across all test cases and methods. The Optimized R2A-UNet achieves the highest median and the tightest interquartile range, indicating both superior accuracy and consistency.

preserve segmentation accuracy despite requiring fewer operations. Another major contribution of the proposed framework lies in the joint optimization of segmentation and classification. U-Net based networks generally rely on pixel-level information, which forces the model to decide independently for each pixel whether it belongs to the tumor or to the surrounding healthy tissue. This strategy generally produces accurate segmentations thanks to the encoding and decoding parts, which are enforced with skip connections. However, it does not incorporate global contextual information regarding the presence or absence of a lesion within the entire image. The additional classification branch introduced in this work provides exactly this missing information by learning an image-level representation that predicts whether a tumor exists in the MRI slice. This global supervision naturally complements pixel-level segmentation. When the classification branch predicts that no tumor is present, the segmentation network is implicitly discouraged from producing isolated positive regions. Conversely, when a tumor is detected globally, the segmentation branch can focus on accurately delineating its spatial extent. The confusion matrix analysis supports this interpretation, showing a substantial reduction in false positive detections compared with the ablation experiment. This improvement is achieved with only a negligible decrease in recall, indicating that the global constraint suppresses erroneous detections without significantly compromising tumor sensitivity. From a clinical perspective, this behavior is particularly needed in computer-aided detection systems. False positives often represent a major limitation because they increase the workload of radiologists and may require unnecessary examinations or invasive procedures. Although maximizing sensitivity remains important, a high false positive rate may reduce clinicians' confidence in automated systems. The proposed joint learning strategy therefore provides a compromise by maintaining a high detection rate while substantially improving prediction precision. The fact that our framework achieves the highest precision on both the BRATS and LGG datasets indicates that integrating image-level semantic information into the segmentation process represents a promising research direction for future medical image analysis systems. The adaptive weighting strategy between Dice Loss and Binary Cross-Entropy also contributes significantly to the observed performance improvements. Dice Loss has become one of

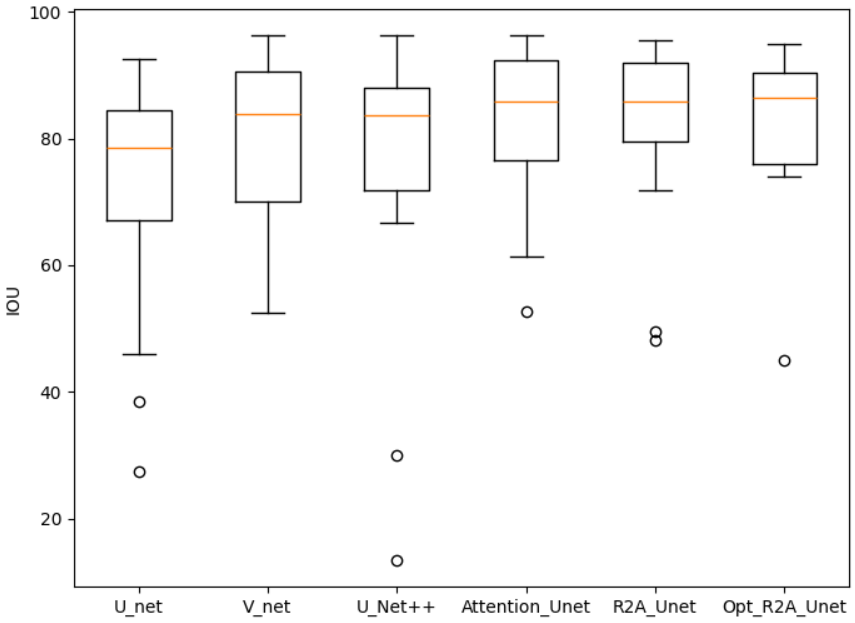


Figure 9. Box plot of IoU scores. The Optimized R2A-UNet achieves superior median IoU with the smallest spread, confirming robustness across diverse and challenging cases.

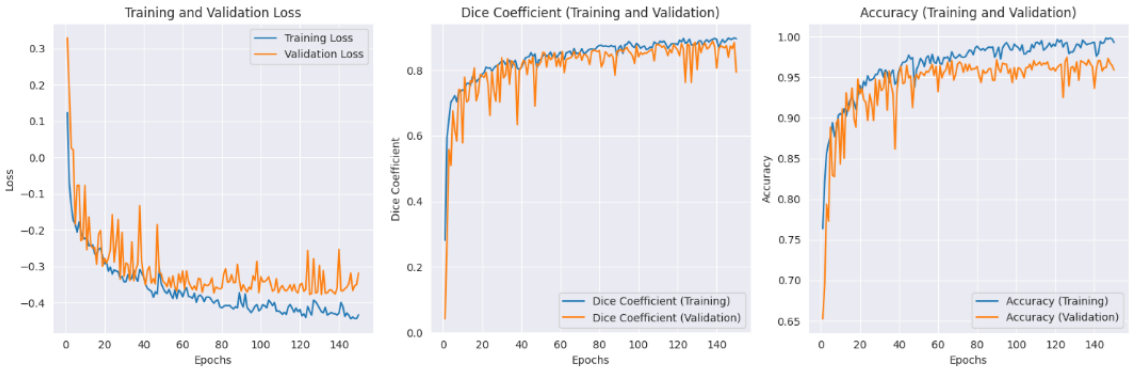


Figure 10. Training and validation learning curves on the LGG dataset.

the most widely adopted objective functions for medical image segmentation because of its robustness to class imbalance. However, several previous studies have reported that Dice Loss generates unstable gradients during the early stages of optimization, particularly when the predicted overlap between the segmentation mask and the ground truth remains very small which is always the case in the early learning phases. Binary Cross-Entropy, on the other hand, provides more stable gradients but does not explicitly optimize spatial overlap. Instead of choosing one objective over the other, the proposed framework progressively shifts the importance process from Binary Cross-Entropy toward Dice Loss. During the first training epochs, Binary Cross-Entropy dominates the optimization and allows the network to learn progressively low-level representations that are proper to each class. Once these

representations become sufficiently stable, the contribution of Dice Loss gradually increases. The learning curves observed throughout training provide experimental evidence that supports this interpretation. We chose BRATS and LGG as benchmark datasets, which provide a quantifiable evaluation of the presented model. Although the framework achieves good performances on both datasets compared to the state of art methods, the characteristics of these databases differ a lot. BRATS tumors generally exhibit relatively well-defined boundaries and stronger contrast against surrounding tissues. In contrast, LGG lesions often present infiltrative growth patterns, irregular shapes, and weak intensity differences with healthy brain tissue. These characteristics make the segmentation task considerably more challenging because the transition between tumor and normal tissue becomes progressively blurred. The proposed attention mechanisms appear sufficiently effective under these difficult conditions. The proposed approach emphasizes informative spatial regions, characterized by relatively high activation coefficients, while suppressing irrelevant background responses. Indeed, the optimized attention gates enable the network to highlight subtle structural information that may otherwise fade because of successive pooling operations. More generally, these results suggest that lightweight attention mechanisms remain relevant for medical images where pathological structures occupy a limited portion of the image and exhibit also an important appearance variability. The present study demonstrates also that optimizing existing architectural components may produce comparable improvements without increasing network complexity. The reduction of computational complexity from quadratic to linear within the attention modules, combined with the reduction in training time, illustrates that computational efficiency and segmentation accuracy are not conflicting goals.

6.1. Limitations

Despite these encouraging results, several limitations should be discussed. First, the proposed framework was evaluated on two publicly available datasets only. Although BRATS and LGG represent widely accepted benchmarks, additional validation on datasets acquired with different scanners and imaging protocols would certainly provide evidence regarding the generalization capability of the proposed architecture. In addition, whereas publicly available datasets provide a benchmark for segmentation approach evaluation, raw images from hospitals with the supervision of radiologists remain the best validation for every computer-aided detection method. Second, the current implementation performs two-dimensional slice-by-slice segmentation and therefore does not explicitly exploit the three-dimensional anatomical continuity naturally present in volumetric MRI examinations. Fully volumetric approaches such as the 3D ASPP-ResUNet of Echine and Darouichi [38] address this continuity directly, at the cost of a substantially higher memory footprint; combining our optimized attention gates with such a volumetric backbone is a natural extension of the present work. Third, following standard BRATS practice, our evaluation reports whole-tumor Dice, IoU, precision, and recall but does not break results down by tumor sub-region (enhancing tumor, tumor core, whole tumor); given the review timeline we were unable to recompute and validate sub-region-specific metrics for this revision, and we identify this as a priority for a future extended evaluation. Fourth, to address this we retrained R2A-UNet and the proposed Optimized R2A-UNet five times each with independent training seeds and report mean $\pm$ standard deviation together with Welch's t-test results in Section 5.5 (Table 7); the results are favorable on most metrics, with a significant loss only in Recall on BRATS. This multi-seed analysis was limited to the two R2A-UNet variants given the review timeline; the remaining baselines in Tables 5 and 6 (U-Net, V-Net, U-Net++, Attention U-Net) still reflect single training runs, and extending the multi-seed protocol to all baselines remains a priority for future work. Fifth, the clinical significance of the reported gains is not directly quantified: patient-level sensitivity, false-positive rate per patient, and tumor-volume measurement error were not computed in this revision, although they would more directly support the clinical relevance argued for the classification branch in Section 5.7. These three points remain priorities for future work.

7. Conclusion

This work proposes an optimized attention-based U-Net architecture for reliable brain tumor segmentation in MRI scans. Our main contributions include: (1) simplified and computationally efficient channel and spatial attention

mechanisms using Hard Sigmoid activation and 1×1 convolutions, reducing computational complexity from $O(C^2)$ to $O(C)$; (2) a novel adaptive loss weighting scheme dynamically adjusting Dice Loss and Binary Cross-Entropy balance; and (3) a joint segmentation-classification framework reducing false positives.

Extensive validation on BRATS 2018 and LGG datasets confirms effectiveness, achieving 94.84% Dice and 96.01% Precision on BRATS, and 92.85% Dice with 89.76% Precision on LGG—the highest precision values among all compared methods. Sensitivity analysis of the adaptive loss parameter $\alpha_{\max}$ identifies 0.7 as the optimal value for BRATS, maximizing IoU by achieving the best balance between early BCE stability and late Dice spatial refinement. Confusion matrix evaluation of the joint classification branch confirms high tumor detection accuracy under this configuration, with substantially fewer false positives and false negatives compared to the ablation without classification, validating the benefit of the joint segmentation-classification design.

Visual analysis demonstrates our method captures both well-defined BRATS tumors and challenging diffuse LGG lesions effectively. The box plot analysis confirms a high consistency across test cases, with few low-score outliers.

Future work will therefore focus on extending the proposed architecture toward fully volumetric three-dimensional segmentation. In addition, validation across additional datasets and real clinical conditions should be considered to evaluate the model's generalization capabilities.

REFERENCES

1. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, pp. 234–241, Springer, 2015.
2. Z. Zhou, M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *DLMIA*, pp. 3–11, Springer, 2018.
3. O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.
4. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *3DV*, pp. 565–571, IEEE, 2016.
5. C. H. Sudre et al., "Generalised Dice overlap as a deep learning loss function," in *DLMIA*, pp. 240–248, Springer, 2017.
6. J. Ma et al., "Loss odyssey in medical image segmentation," *Medical Image Analysis*, vol. 71, p. 102035, 2021.
7. K. P. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "3D segmentation with weighted dice loss," *arXiv preprint arXiv:1802.04574*, 2018.
8. C. Shi et al., "Real-Time ConvNext-Based U-Net with Feature Infusion," *MDPI Agriculture*, vol. 14, no. 9, p. 1655, 2024.
9. Q. Zhang et al., "Improved U-Net network asphalt pavement crack detection," *PLOS ONE*, vol. 19, no. 5, p. e0300679, 2024.
10. M. R. Islam et al., "CoST-UNet: Convolution and Swin Transformer-based Deep U-Net," *Computers in Biology and Medicine*, vol. 157, p. 106667, 2024.
11. J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, et al., "TransUNet: Rethinking the U-Net Architecture Design for Medical Image Segmentation Through the Lens of Transformers," *Medical Image Analysis*, vol. 97, art. 103280, 2024.
12. S. Woo et al., "CBAM: Convolutional block attention module," in *ECCV*, 2018.
13. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.
14. A. Vaswani et al., "Attention is all you need," in *NeurIPS*, 2017.
15. J. Chen et al., "TransUNet: Transformers make strong encoders," *arXiv preprint arXiv:2102.04306*, 2021.
16. H. Cao et al., "Swin-UNet: U-Net-like pure transformer," *Neurocomputing*, 2022.
17. J. Valanarasu et al., "MedT: A lightweight transformer," *arXiv preprint arXiv:2104.06481*, 2021.
18. T. Tong et al., "AC-ResUNet++: Residual and attention modules," *Applied Journal of Science and Technology*, 2024.
19. Y. Wang et al., "MPB-UNet: Multi-path parallel blocks," *Electronics*, vol. 14, no. 1, p. 40, 2024.
20. S. Yeung et al., "Modifying Dice loss for better training stability," *Medical Image Analysis*, 2021.
21. S. Salehi et al., "Tversky loss function for image segmentation," in *MICCAI*, 2017.
22. T.-Y. Lin et al., "Focal loss for dense object detection," in *ICCV*, 2017.
23. X. Li et al., "Dense U-Net for automatic liver and tumor segmentation," in *MICCAI*, pp. 415–423, 2019.
24. Z. Huang et al., "UNet 3+: A full-scale connected UNet," *arXiv preprint arXiv:2004.08790*, 2020.
25. N. Bentaher, Y. Kabbadj, and M. Ben Salah, "Enhancing breast masses detection and segmentation," in *WINCOM*, pp. 1–6, IEEE, 2023.
26. N. Bentaher et al., "R2A-UNet: Double attention mechanisms with residual blocks," *Multimedia Tools and Applications*, pp. 1–31, Springer, 2025.
27. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, pp. 770–778, 2016.
28. M. A. Rahman and Y. Wang, "Optimizing Intersection-Over-Union in Deep Neural Networks," in *ACIVS*, 2016.
29. J. Yanase and E. Triantaphyllou. A systematic survey of computer-aided diagnosis in medicine: Past and present developments. Expert Systems with Applications, 138:112821, 2019. doi: 10.1016/j.eswa.2019.112821.
30. M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Computerized Medical Imaging and Graphics*, vol. 95, p. 102026, 2022.

31. K. C. L. Wong, M. Moradi, H. Tang, and T. Syeda-Mahmood, "3D segmentation with exponential logarithmic loss for highly unbalanced object sizes," in Medical Image Computing and Computer Assisted Intervention (MICCAI), ser. Lecture Notes in Computer Science, vol. 11072, pp. 612–619, Springer, 2018.
32. J. Bertels, T. Eelbode, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimizing the Dice score and Jaccard index for medical image segmentation: Theory and practice," in Medical Image Computing and Computer Assisted Intervention (MICCAI), ser. Lecture Notes in Computer Science, vol. 11765, pp. 92–100, Springer, 2019.
33. B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," IEEE Transactions on Medical Imaging, vol. 34, no. 10, pp. 1993–2024, 2015.
34. S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, et al., "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," Nature Scientific Data, vol. 4, p. 170117, 2017.
35. S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, 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.
36. H. Wang, P. Cao, J. Yang, and O. Zaiane, "Narrowing the Semantic Gaps in U-Net with Learnable Skip Connections: The Case of Medical Image Segmentation," *Neural Networks*, vol. 178, p. 106546, 2024.
37. S. Asgari Taghanaki, Y. Zheng, S. K. Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, and G. Hamarneh, "Combo Loss: Handling Input and Output Imbalance in Multi-Organ Segmentation," *Computerized Medical Imaging and Graphics*, vol. 75, pp. 24–33, Jul. 2019, doi: 10.1016/j.compmedimag.2019.04.005.
38. K. Echine and A. Darouichi, "An Efficient Deep Learning Approach based on 3D Res-UNet for Multimodal Brain Tumor Segmentation," *Statistics, Optimization & Information Computing*, vol. 15, no. 2, pp. 1236–1247, 2026. doi: 10.19139/soic-2310-5070-2786.
39. H. Abdulqader and A. Abdulazeez, "Lung Cancer Segmentation and Classification with Multi-Dataset Integration," *Statistics, Optimization & Information Computing*, vol. 15, no. 3, pp. 1868–1888, 2025. doi: 10.19139/soic-2310-5070-2814.
40. Z. Aitmaten, S. Aloui, and A. Belaid, "Enhancing Echocardiographic Segmentation through KL-Divergence-Based Intensity Distribution Constraints in U-Net Models," *Statistics, Optimization & Information Computing*, vol. 15, no. 5, pp. 3945–3963, 2026. doi: 10.19139/soic-2310-5070-2869.
41. A. Hmidi, N. Kouka, and L. Tekari, "A Fine-Tuned CNN for Multiclass Classification of Brain Tumors on Figshare CE-MRI and its Raspberry Pi Deployment," *Statistics, Optimization & Information Computing*, vol. 14, no. 6, pp. 3199–3213, 2025. doi: 10.19139/soic-2310-5070-2635.
42. F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "No New-Net," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2018)*, Lecture Notes in Computer Science, vol. 11384, Springer, Cham, 2019, pp. 234–244.
43. R. M. Mostafa, E. Mabrouk, A. Ayman, and H. Z. Zidan, "Brain Tumor Segmentation using a Combination of Graph Cut and Optimized U-Net with Genetic Algorithm: An Efficient Hybrid Framework for Enhanced Medical Image Analysis," medRxiv preprint, 2025. doi: 10.64898/2025.11.30.25341319.
44. A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, D. Xu, et al., "UNETR: Transformers for 3D Medical Image Segmentation," in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 1748–1758.
45. Y. Sun, M. Liu, and C. Lian, "AEPL: Automated and Editable Prompt Learning for Brain Tumor Segmentation," arXiv preprint arXiv:2410.19847, 2024.
46. A. Golnari and M. Diba, "Adaptive Real-Time Multi-Loss Function Optimization Using Dynamic Memory Fusion Framework: A Case Study on Breast Cancer Segmentation," arXiv preprint arXiv:2410.19745, 2024.
47. A. Kendall, Y. Gal, and R. Cipolla, "Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7482–7491.
48. M. Buda, A. Saha, and M. A. Mazurowski, "Association of Genomic Subtypes of Lower-Grade Gliomas with Shape Features Automatically Extracted by a Deep Learning Algorithm," *Computers in Biology and Medicine*, vol. 109, pp. 218–225, 2019. doi: 10.1016/j.compbiomed.2019.05.002.