International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(15s) 2026 pp. 982-997
https://doi.org/10.70917/ijcisim-2026-4496
Article

An Efficient Multilevel-Fusion 3-D Framework for Brain Tumor Segmentation in Multimodal MRI

Polagani Rama Devi[1], [2], Srikanth Vemuru[3]

[1]Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Greenfields, Vaddeswaram, Guntur, Andhra Pradesh, India.
[2]Assistant Professor, Department of Information Technology, Siddhartha Academy of Higher Education (Deemed to be University), Vijayawada, Andhra Pradesh, India.
[3]Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Greenfields, Vaddeswaram, Guntur, Andhra Pradesh, India.

Abstract: Accurate volumetric delineation of glioma subregions from multimodal magnetic resonance imaging (MRI) supports quantitative assessment, treatment planning, and longitudinal follow-up. This paper presents a reproducibility-aware rewrite of a multilevel-fusion framework that integrates three established ideas: robust modality-wise intensity conditioning, a 3-D residual inception encoder-decoder with guided attention and multiscale fusion, and fuzzy graph-cut refinement of voxelwise posterior probabilities. The formulation preserves local detail through residual paths, models long-range context at the bottleneck, and imposes spatial coherence during refinement. Three algorithms are specified in standard pseudocode, followed by mathematical results for range preservation, convex-gated fusion, probability normalization, descent of the alternating refinement energy, and the limits of those guarantees. The experimental record provided with the source manuscript identifies BraTS 2020, 128 x 128 x 128 inputs, Adam optimization, 100 epochs, and an NVIDIA A100. It reports aggregate Dice, intersection-over-union, sensitivity, specificity, accuracy, recall, precision, and HD95 values. These values are reproduced as reported and are not presented as independently re-estimated results, since subject-level predictions, split identifiers, repeated runs, and code were not supplied. The paper therefore distinguishes architectural analysis from empirical evidence, documents reporting inconsistencies, and defines the minimum validation package needed for a defensible comparison.
Keywords: Attention, BraTS, brain tumor, fuzzy clustering, graph cut, multimodal MRI, multilevel fusion, 3-D segmentation.

1. INTRODUCTION

Gliomas are infiltrative central-nervous-system tumors whose enhancing tissue, necrotic or nonenhancing core, and surrounding edema can occupy irregular and nested regions. Multimodal MRI offers complementary soft-tissue contrasts without ionizing radiation, but the same lesion may have weak margins, heterogeneous intensities, treatment-related changes, and severe class imbalance. The BraTS benchmark established common data definitions and expert reference annotations for systematic comparison [1]. Curated TCGA-derived glioma volumes and manually corrected labels have extended the available research material and supported quantitative imaging studies [2].

The BraTS 2020 training set contains 369 preoperative cases acquired across institutions, with co-registered T1-weighted, contrast-enhanced T1-weighted (T1ce), T2-weighted, and fluid-attenuated inversion recovery (FLAIR) volumes. The standard evaluation targets are whole tumor (WT), tumor core (TC), and enhancing tumor (ET), which are nested composites of the original labels. Strong challenge systems treat preprocessing, patch sampling, augmentation, loss design, ensembling, and postprocessing as a coupled pipeline rather than attributing performance to a single network layer [3]. This observation is central to the present framework: fusion is performed across modalities and resolution levels, but its role must be evaluated within an explicitly defined pipeline.

U-Net introduced the encoder-decoder and skip-connection pattern that now dominates biomedical segmentation [4]. The 3-D U-Net extended this pattern to volumetric context [5], and V-Net showed that end-to-end 3-D fully convolutional learning can be paired with an overlap-based objective [6]. Volumetric processing reduces the discontinuity that may arise when independent 2-D slices are classified, yet it increases memory use and can amplify foreground-background imbalance. A practical design must preserve small enhancing regions, carry coarse contextual evidence, and remain trainable at feasible patch sizes.

Early deep brain-tumor systems demonstrated that modality-specific normalization, local and global pathways, and cascaded decisions improve the use of heterogeneous MRI evidence [7]-[9]. Autoencoder regularization can stabilize the latent representation [10], and nnU-Net demonstrated that disciplined, data-dependent configuration can outperform manually embellished pipelines [11]. These findings argue against presenting a collection of modules as proof of novelty. The scientific question is whether a defined integration changes segmentation quality under a controlled split and a reproducible evaluation.

Residual learning supplies an identity path through deep transformations and helps optimization of deep feature extractors [12]. Inception-style parallel kernels expand the effective receptive-field mixture [13], and attention gates suppress skip features that are weakly relevant to the current decoding context [14]. A published GAIR-U-Net already combines guided attention, inception, residual connections, and dilated 3-D convolution for multimodal brain-tumor segmentation [15]. The framework in this paper therefore cites those components as prior art. Its proposed element is a pipeline-level integration of localized intensity conditioning, modality- and scale-aware fusion, and an explicit fuzzy graph-cut refinement stage.

Transformers provide a complementary mechanism for long-range dependence. TransBTS embeds transformer processing in a 3-D CNN encoder-decoder for multimodal brain-tumor segmentation [16]. UNETR uses a transformer encoder with multiresolution decoder connections [17], and Swin UNETR reduces attention scope through hierarchical shifted windows [18]. These models inherit the token-interaction principle introduced by self-attention [19]. Their strengths do not remove the need for local edge evidence: tumor boundaries can span only a few voxels, and pure global mixing may blur rare enhancing components if local features and class-aware losses are weak.

The manuscript addresses four connected problems. First, intensities from different scanners and sequences are not directly comparable, so each modality is conditioned inside the nonzero brain region. Second, modality evidence must be combined without forcing every sequence to contribute equally at every scale. Third, deep 3-D encoders require stable gradient paths and a decoder that filters irrelevant skip content. Fourth, isolated posterior errors should be corrected with a spatial objective whose behavior is mathematically interpretable.

These problems are coupled by the nested structure of the BraTS targets. An error in the enhancing component changes ET, TC, and WT differently, and a visually small false-positive island can have a large effect on HD95. A single pooled score therefore cannot establish that a pipeline preserves every clinically relevant subregion. The design developed here keeps the four base-label posteriors until the final mapping to WT, TC, and ET, applies refinement to the posterior field rather than to independently thresholded composite masks, and requires classwise and casewise reporting. This choice preserves label nesting by construction and makes it possible to identify whether a gain comes from improved core delineation, edema coverage, or removal of distant false positives.

Cross-institutional variation creates a second evaluation risk. Scanner vendor, acquisition protocol, voxel spacing, and intensity distribution may correlate with the data partition, so a random patch split can leak subject-specific appearance and produce optimistic estimates. A valid experiment separates complete subjects, derives every learned or tuned quantity from the training and validation partitions, and keeps the test labels inaccessible until the model and postprocessing parameters are frozen. The benchmark definitions in [1] and the self-configuring evidence in [3], [11] support this pipeline view. External-site testing remains necessary before a result can be interpreted as evidence of transportability beyond BraTS. Reader review of difficult cases can then test whether numerical changes correspond to credible anatomical boundaries rather than metric improvements produced by smoothing or component deletion.

The contributions are: 1) a coherent MLF-3D architecture that replaces the draft's inconsistent references to modified V-Net, Vision Transformer processing, and GAIR-U-Net with one defined computation graph; 2) three standard algorithms for preprocessing, network optimization, and fuzzy graph-cut refinement; 3) a mathematical analysis that proves limited structural properties without claiming a proof of diagnostic accuracy or nonconvex-training convergence; 4) monochrome, publication-style reporting of the numerical values contained in the source

record; and 5) a reproducibility discussion that identifies missing split, code, and subject-level evidence before clinical or state-of-the-art claims are made.

The remainder of the paper is organized as follows. Section II reviews volumetric CNNs, attention, transformer fusion, preprocessing, loss functions, and spatial refinement. Section III defines the proposed pipeline. Section IV presents the mathematical results for the three algorithms. Section V states the available experimental protocol and the minimum replication protocol. Sections VI and VII report and interpret the source values. Section VIII summarizes the supported conclusions.

2. RELATED WORK

A. Encoder-Decoder and Volumetric Segmentation

The U-Net family combines low-resolution semantic context with high-resolution localization through skip connections. 3-D variants replace planar operations with volumetric kernels and are better suited to interslice continuity, but their activation tensors impose strict memory limits. V-Net popularized residual volumetric blocks and a Dice objective, and UNet++ redesigned skip pathways with dense intermediate nodes to reduce the semantic gap between encoder and decoder features [20]. Across these models, reported improvements often depend on patch geometry, loss weighting, resampling, and connected-component rules as much as on the nominal backbone.

Brain-tumor pipelines have used dual pathways, cascades, autoencoder regularization, deep supervision, and ensembles. The local pathway helps identify fine edges and small ET foci, whereas coarse features provide anatomical context and reduce confusion with normal high-intensity structures. Deep supervision supplies gradients at more than one decoder level and can regularize coarse-to-fine predictions. Ensembling improves robustness at increased inference cost. No single mechanism resolves cross-site intensity shift, missing sequences, or annotation ambiguity; evaluation must distinguish gains from architecture, training recipe, and postprocessing.

B. Intensity Conditioning and Data Variation

MRI intensity has no universal physical scale across acquisitions. Bias-field correction addresses slow spatial nonuniformity; N4ITK is a widely used improvement of the N3 method [21]. Histogram-based scale standardization maps tissue landmarks to a common intensity scale [22]. In already co-registered and skull-stripped BraTS volumes, many pipelines use foreground-only z-score normalization and percentile clipping. The present method retains a localized CDF transform only as a bounded contrast-conditioning operation. It is not used to create synthetic tumor evidence, and its parameters must be estimated on training data or independently within each volume to avoid leakage.

Spatial and intensity augmentation can improve robustness, but unrealistic deformations or contrast shifts can corrupt subregion semantics. A brain-tumor-specific review found wide variation in augmentation policies and called for controlled evaluation [23]. Appropriate operations include left-right flips when anatomy and labels permit them, modest rotations and scaling, gamma or brightness perturbations, Gaussian noise, and patch sampling biased toward foreground. The same random transform must be applied to all modalities and the label map.

C. Losses and Evaluation

Foreground subregions occupy a small fraction of the volume. Generalized Dice reweights classes to reduce domination by background [24], and the Tversky formulation allows false-positive and false-negative penalties to be adjusted asymmetrically [25]. The proposed training objective combines soft Dice with cross-entropy and an optional boundary term. Each term has a distinct role: overlap optimization addresses imbalance, cross-entropy retains probabilistic supervision, and the boundary term discourages spatially displaced edges. Loss weights are hyperparameters, not mathematical evidence that one error type is clinically preferable.

Dice and intersection-over-union (IoU) measure overlap; sensitivity and precision separate missed and excess foreground; specificity and accuracy can be inflated by the large background; and the 95th-percentile Hausdorff distance (HD95) evaluates boundary discrepancy. Metric definitions and implementation details materially affect rankings [26]. Hausdorff distance was introduced for set comparison in computer vision and remains sensitive to spatial outliers [27]. A credible report should present WT, TC, and ET separately, specify treatment of empty masks, and include distributional summaries or confidence intervals rather than only a pooled percentage.

D. Spatial Refinement and Recent Fusion Models

Graph cuts minimize binary energies with unary evidence and submodular pairwise smoothness terms [28]. Fuzzy c-means (FCM) assigns soft memberships and alternates membership and centroid updates [29]. Classical medical-image segmentation surveys describe the tradeoff between data fidelity, spatial regularity, and initialization [30]. Combining posterior probabilities with a graph model can remove isolated errors and encourage edge-aligned regions, but excessive regularization may erase small ET components. The refinement weight must therefore be selected on a validation set and reported.

Competitive BraTS systems show the value of self-ensembling and deep supervision [31], and the No New-Net result emphasizes that a strong, carefully trained 3-D U-Net can rival more elaborate designs [32]. Recent hybrid models add clinical priors and cross-attention [33], efficient transformer-enhanced U-Net blocks [34], edge-aware fusion for missing modalities [35], or dynamic knowledge distillation for compact models [36]. The research gap is not a shortage of modules. It is the need for a traceable integration in which the contribution of each stage, its computational cost, and its uncertainty are measured under an unchanged data split.

The comparison among these systems must account for the information available to each model. A four-modality network, a missing-modality network, a distilled student, and an ensemble solve related but nonidentical tasks. Parameter count alone does not describe cost: volumetric activation memory, attention-map storage, sliding-window overlap, test-time augmentation, ensemble size, and graph-refinement time can dominate deployment. A fair ablation should begin with one reproducible 3-D U-Net training recipe, then add conditioning, gated multilevel fusion, the bottleneck context block, and refinement one stage at a time. Each row should retain the same subject split, augmentation schedule, epoch budget, and checkpoint rule. Reporting the mean and dispersion across seeds separates a stable contribution from a favorable initialization.

Missing modalities deserve explicit treatment rather than silent zero filling. Edge-aware fusion methods can be trained for incomplete sequences [35], and knowledge distillation can transfer information from a multimodal teacher to a smaller or restricted student [36]. Those strategies require modality-dropout schedules or defined teacher-student pairs and cannot be inferred from a network diagram. The present framework assumes all four registered sequences at both training and inference. Its softmax modality gates express relative use of available sequences; they do not make the system invariant to an absent sequence. Robustness claims would require testing every clinically plausible missing-modality pattern and comparing against retrained or imputation-based baselines.

Spatial refinement has a similarly conditional value. A graph prior can improve connectedness and suppress distant false detections, yet the same prior can merge adjacent structures or erase a small enhancing focus. Evaluation should therefore report results before and after refinement for each target, paired per case, together with the number and physical volume of components removed. Boundary plots and representative failure cases are more informative than a single average overlap. This evidence connects the mathematical bias of the Potts term to the errors observed in the predicted masks.

3. PROPOSED MLF-3D FRAMEWORK

A. Problem Formulation

For subject s, let X = {X_m}_{m=1}^M be M co-registered MRI volumes on voxel domain Omega, with M = 4 for T1, T1ce, T2, and FLAIR. The ground-truth map Y assigns each voxel to background, necrotic or nonenhancing tumor core, edema, or enhancing tumor. The network produces class posteriors P_theta(c|v, X). Composite WT, TC, and ET masks are obtained by the official label unions. The trainable parameters theta are learned only from the training partition; normalization constants, model selection, and threshold tuning may not use test labels.

$$P_{\theta}(c \mid v, X) = \exp(z_c(v)) / \sum_k \exp(z_k(v)) \qquad (1)$$

The architecture has three stages. Algorithm 1 conditions each modality. Algorithm 2 produces multilevel fused probabilities. Algorithm 3 refines the posterior map by coupling fuzzy memberships with a spatial graph. Fig. 1 shows the resulting computation graph.

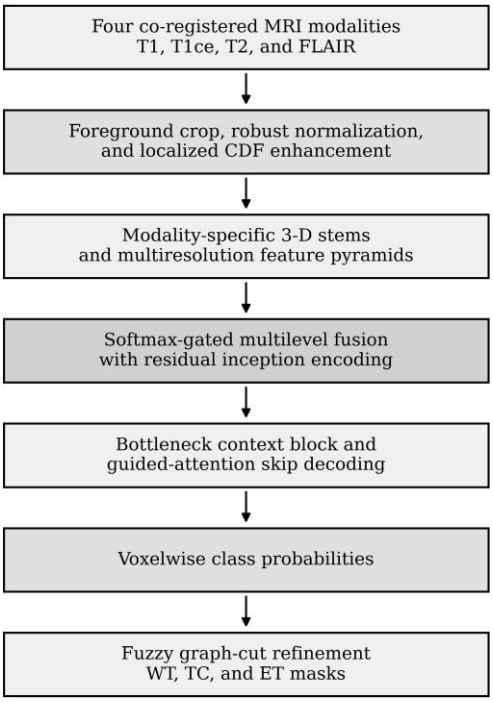

Fig. 1. Proposed MLF-3D pipeline. All blocks are shown in grayscale for IEEE print compatibility.

B. Algorithm 1: Robust Multimodal Conditioning

Foreground-only statistics prevent the large zero background from dominating intensity estimates. For each modality, lower and upper quantiles q_l and q_h are computed within the nonzero brain mask. Intensities are clipped and scaled to [0,1]. A Gaussian-weighted local mean reduces high-frequency perturbations, and a smoothed empirical CDF supplies a monotone contrast coordinate. The final conditioned image is a convex mixture controlled by lambda_e. The same spatial crop is applied to all modalities and labels.

$$\tilde{X}_m(v) = \mathrm{clip}\left(\frac{X_m(v) - q_{l,m}}{q_{h,m} - q_{l,m} + \varepsilon}, 0, 1\right) \quad (2)$$

$$E_m(v) = (1 - \lambda_e)\,\tilde{X}_m(v) + \lambda_e\,F_m(\tilde{X}_m(v)) \quad (3)$$

Algorithm 1 ROI-Aware Multimodal Conditioning
Input: Co-registered volumes {X_m}, quantiles q_l and q_h, Gaussian scale sigma, mixing weight lambda_e.
Output: Conditioned tensor E and common foreground crop.
1: Form the common brain mask Omega_b from nonzero voxels across modalities.
2: Compute a padded bounding box of Omega_b and crop every modality to the same coordinates.
3: For each modality m, estimate q_l,m and q_h,m inside Omega_b; clip and map intensities to [0,1].
4: Compute normalized Gaussian weights and a local mean for noise-resistant neighborhood context.
5: Evaluate a smoothed empirical CDF F_m and form E_m by the convex mixture in (3).
6: Resample or pad E to 128 x 128 x 128; apply the identical transform to the label map during training.

C. Multilevel Fusion and Guided Residual Decoding

Each modality first passes through a shallow 3-D stem. At encoder level l, a residual inception block uses parallel 3 x 3 x 3 and dilated 3 x 3 x 3 paths, followed by channel compression and an identity projection when the channel count changes. The notation IR_l denotes this block. Modality-specific features H_l,m are fused by softmax gates a_l,m computed from global average pooled descriptors. The gate is conditioned on scale, so FLAIR can

dominate edema evidence at one level and T1ce can dominate enhancing-tumor evidence at another without hard-coded modality weights.

$$a_{l,m} = \exp(s_{l,m}) / \sum_r \exp(s_{l,r}), \quad F_l = \sum_m a_{l,m} H_{l,m} \tag{4}$$

The encoder uses channel widths 8, 16, 32, 64, and 128, matching the source record. Downsampling uses stride-two 3-D convolution. A bottleneck context block mixes dilated local features with tokenized coarse features. The decoder upsamples with transpose convolution. Before concatenation, a guided attention gate uses the decoder state g_l to filter F_l. This arrangement retains a residual gradient path, uses global context only at the low-resolution bottleneck, and prevents the transformer block from carrying the full 128-cubed token count.

$$\alpha_l(v) = \sigma(\psi_l^T \mathrm{ReLU}(W_x F_l(v) + W_g g_l(v) + b_l)) \tag{5}$$

$$D_l = IR_l([\mathrm{Up}(D_{l+1}), \alpha_l \odot F_l]) \tag{6}$$

Deep-supervision heads at two intermediate decoder levels are resampled to the output grid during training. The final objective combines class-weighted cross-entropy, soft Dice, and a boundary discrepancy term. For nonnegative weights that sum to one, the total loss is nonnegative. The boundary term should be disabled if a consistent signed-distance implementation is unavailable; an undefined boundary loss is worse than a simpler documented objective.

$$L = \lambda_{ce} L_{CE} + \lambda_d L_{Dice} + \lambda_b L_{boundary} + \lambda_s \sum_l L_{deep,l} \tag{7}$$

Algorithm 2 Training and Inference of MLF-3D

Input: Conditioned tensor E, label Y for training, parameters theta, loss weights, maximum epochs T.

Output: Voxelwise posterior P_theta and discrete mask Y_hat.

1: Encode each modality with a 3-D stem and IR blocks at five spatial resolutions.

2: At every level, compute modality scores, normalize them with softmax, and fuse features by (4).

3: Process the deepest fused feature with the low-resolution context block.

4: Decode coarse-to-fine; gate each skip tensor by (5), concatenate it with the upsampled decoder tensor, and apply an IR block.

5: During training, evaluate (7), update theta, and select the checkpoint using only validation data.

6: During inference, compute P_theta with softmax and assign each voxel to its maximum-posterior class before refinement.

D. Algorithm 3: Fuzzy Graph-Cut Refinement

The neural posterior is converted into unary costs $D_i(c) = -\log(P_i(c) + \varepsilon)$. FCM memberships $u_{i,c}$ and centroids μ_c are initialized from the posterior-weighted feature vectors. Neighboring voxels i and j receive edge weight w_{ij} that decreases with intensity and feature difference. The refinement objective combines fuzzy appearance, posterior fidelity, and a Potts smoothness term. Binary subproblems are solved by s-t cut; multiclass refinement uses alpha-expansion when the pairwise term is metric.

$$E(Y,U,\mu) = \sum_{i,c} u_{i,c}^q \|f_i - \mu_c\|^2 - \eta \sum_i \log(P_{i,Y_i} + \varepsilon) + \beta \sum_{(i,j)} w_{ij} [Y_i \neq Y_j] \tag{8}$$

$$w_{ij} = \exp(-\|E(i)-E(j)\|^2 / (2\sigma_I^2) - \|f_i-f_j\|^2 / (2\sigma_f^2)) \tag{9}$$

Algorithm 3 Posterior-Guided Fuzzy Graph-Cut Refinement

Input: Posterior P, conditioned image E, decoder features f, fuzzifier $q > 1$, weights eta and beta.

Output: Spatially coherent label map Y_star.

1: Initialize memberships $u_{i,c}$ from $P_i(c)$ and compute centroids μ_c.

2: Update fuzzy memberships and centroids using the closed-form FCM block minimizers.

3: Build unary costs from posterior fidelity and fuzzy appearance; build edge weights with (9).

4: Optimize the binary or alpha-expansion graph-cut label subproblem.

5: Repeat membership, centroid, and label updates until the energy change is below tolerance or the iteration limit is reached.

6: Remove only components smaller than a validation-selected physical-volume threshold; map class labels to WT, TC, and ET.

E. Configuration and Computational Scope

TABLE I REPORTED AND DEFINED MODEL CONFIGURATION

Item	**Configuration**
Input	128 x 128 x 128 multimodal volume
Modalities	T1, T1ce, T2, and FLAIR; source tensor elsewhere lists three channels and must be corrected
Encoder widths	8, 16, 32, 64, 128
Downsampling	Stride 2 x 2 x 2
Bottleneck	128 channels; local-dilated and low-resolution context paths
Decoder	Transpose convolution, guided skip gates, residual inception blocks
Output	Four base labels; WT, TC, and ET obtained by label union
Refinement	Posterior-guided fuzzy graph cut with validation-selected beta

4. MATHEMATICAL ANALYSIS OF THE THREE ALGORITHMS

A. Range Preservation and Stability of Algorithm 1

Proposition 1 (range preservation). Assume q_h,m > q_l,m, epsilon > 0, 0 <= lambda_e <= 1, and a valid CDF F_m: [0,1] -> [0,1]. Then every output of (3) lies in [0,1]. If F_m is nondecreasing, E_m is nondecreasing as a function of the normalized intensity X_tilde_m.

Proof. Clipping in (2) gives 0 <= X_tilde_m(v) <= 1. A CDF takes values in [0,1]. Equation (3) is a convex combination of these two quantities, so its minimum cannot be below zero and its maximum cannot exceed one. For x_1 <= x_2, both x and F_m(x) are nondecreasing. Multiplication by nonnegative coefficients and addition preserve the order, giving E_m(x_1) <= E_m(x_2). The transform can alter contrast but cannot invert the ordering of two normalized intensities. This is the precise guarantee; it does not prove that tumor and normal tissue become separable.

The Gaussian local mean has the same boundedness property. Let weights g_vu be nonnegative and normalized so that sum_u g_vu = 1. The local mean mu(v) = sum_u g_vu X_tilde(u) is a convex combination of values in [0,1], hence mu(v) is in [0,1]. Replacing the raw intensity by a convex mixture of X_tilde, mu, and F(X_tilde) preserves the range when all coefficients are nonnegative and sum to one. This is useful for deterministic downstream scaling and prevents a local enhancement stage from generating values outside the network's specified input interval.

Proposition 2 (perturbation bound). Hold q_l,m and q_h,m fixed and let Delta_q = q_h,m - q_l,m + epsilon. If the smoothed CDF is L_F-Lipschitz, then a perturbation delta of the input produces |E_m(x+delta)-E_m(x)| <= [(1-lambda_e)+lambda_e L_F] |delta|/Delta_q, except where clipping makes the change smaller.

Proof. Before clipping, normalization is 1/Delta_q-Lipschitz. Projection onto [0,1] is nonexpansive, so clipping cannot increase the distance. Apply the triangle inequality to (3), then the Lipschitz bound for F_m. The result follows. Empirical step CDFs are not globally Lipschitz, which is why the algorithm specifies a smoothed CDF. Quantile

estimates can change under perturbation; the proposition is conditional on fixed quantiles. Robustness of sample quantiles requires a separate distributional assumption and is not asserted here.

A useful corollary concerns cross-modality comparability. Suppose two modalities are independently mapped by (2) and (3). Their outputs share the same numerical range, but equality of output values does not imply equality of tissue type or physical signal. The transform is order preserving within each modality, not a registration between modality distributions. Multimodal fusion must therefore retain modality identity before the learned gate. Concatenating or averaging the conditioned images as if their intensities were interchangeable would use a property that the range proof does not provide.

The quantile denominator also exposes a stability condition. If q_h,m-q_l,m is very small, the normalization factor becomes large and noise can be amplified despite epsilon. A deterministic implementation should detect near-constant foreground volumes and either omit the modality, substitute a constant zero image with a missingness indicator, or use a declared minimum denominator. This branch must be identical in training and inference. The proposition remains valid under the fallback, but the learned model may behave differently when a modality contains no useful contrast; that behavior requires a modality-ablation experiment.

Crop consistency follows from applying one bounding box to every modality and label. Let T be the crop-resample operator derived from the common brain mask. Using T(X_m) for all m and T(Y) for the label preserves voxel correspondence. Independently estimated modality crops would define different coordinate maps T_m and invalidate voxelwise fusion. The algorithm explicitly forbids that practice. Interpolation must be linear or higher order for images and nearest-neighbor for labels, since interpolating class identifiers creates invalid labels.

B. Bounded Fusion, Probability Validity, and Gradient Paths in Algorithm 2

Proposition 3 (convex-gated fusion). At level l, softmax gates satisfy a_l,m > 0 and sum_m a_l,m = 1. For any norm, ||F_l|| <= sum_m a_l,m ||H_l,m|| <= max_m ||H_l,m|| when the feature tensors share a common vector space.

Proof. Positivity and unit sum follow directly from the softmax definition in (4). The first inequality is the triangle inequality and absolute homogeneity of a norm. The second follows since a weighted average cannot exceed the maximum term. The result means fusion cannot increase the feature norm beyond the largest modality norm solely by convex mixing. It does not bound the subsequent convolution, whose operator norm depends on learned weights. It also does not show that the gate selects the clinically correct modality; that requires ablation and qualitative inspection.

Proposition 4 (valid posterior). For finite logits z_c(v), (1) yields P_theta(c|v,X) > 0 and sum_c P_theta(c|v,X) = 1 for every voxel.

Proof. Exponentials of finite values are positive. Dividing each exponential by their common positive sum gives positive ratios whose sum is exactly one. Numerical implementations subtract max_c z_c before exponentiation for stability; this leaves every ratio unchanged. The maximum-posterior label is therefore defined. Probability validity is weaker than calibration: a normalized posterior may still be overconfident, so reliability analysis is required before probabilities are interpreted as risk.

The attention coefficient in (5) lies strictly between zero and one for finite inputs. Multiplying F_l by alpha_l suppresses or retains a feature; it cannot reverse its sign or increase its magnitude at that voxel. This bounded gate is valuable at skip connections, yet a sigmoid can saturate. Initialization, normalization, and deep supervision are empirical controls for saturation, not formal guarantees.

Proposition 5 (identity contribution to the Jacobian). For a residual block R(x) = x + G(x), the Jacobian is J_R = I + J_G. If a scalar loss has gradient g at the block output, the input gradient is g(I + J_G) = g + gJ_G.

Proof. Differentiate the identity map and G separately. Linearity of differentiation gives I + J_G, and the chain rule gives the gradient expression. The term g is an explicit direct path. This explains why residual blocks can ease optimization compared with a stack whose gradient is only a product of transformation Jacobians. It does not prove that gradients never vanish or explode, since g + gJ_G can cancel or grow depending on J_G and the surrounding network. Any manuscript statement that residual links guarantee convergence would exceed the mathematics.

The self-attention block at the bottleneck receives N tokens, so dense attention requires O(N^2 d + N d^2) arithmetic and O(N^2) attention memory for token dimension d. Applying it after four stride-two reductions changes N by approximately 2^-12 relative to the full voxel grid. This placement makes global interaction feasible but can lose

fine detail; decoder skips provide the complementary high-resolution path. Windowed or axial attention may reduce cost, but their implementation and receptive field differ and should not be conflated.

The loss in (7) is lower bounded by zero when every component is defined as a nonnegative loss and all lambda coefficients are nonnegative. A lower bound prevents loss values from decreasing without limit, but it does not prove that Adam reaches a global or local minimizer of the nonconvex network objective. Training convergence in this paper must therefore be demonstrated with learning curves, repeated seeds, and a fixed stopping rule. The source record supplies none of those artifacts.

Deep supervision changes the gradient received by shared encoder parameters. If L_total = L_final + lambda_s sum_l L_l, differentiation gives grad_theta L_total = grad_theta L_final + lambda_s sum_l grad_theta L_l. The expression supplies shorter paths from intermediate outputs, but those gradients can disagree. For two gradients g_a and g_b, their combined squared norm is ||g_a||^2 + ||g_b||^2 + 2 g_a^T g_b; a negative inner product reveals interference. The weight lambda_s should therefore be selected empirically and the auxiliary heads removed or explicitly retained at inference. The algebra supports the existence of extra gradient signals, not an unconditional improvement in optimization.

The class-loss mixture has a related interpretation. Cross-entropy is a proper scoring rule for the categorical posterior, whereas soft Dice couples voxels through global sums and emphasizes regional overlap. Their convex combination remains nonnegative but is not itself guaranteed to be calibrated. A boundary term adds sensitivity to spatial displacement and may be undefined for an empty class unless the implementation specifies a convention. Stable training requires epsilon constants, empty-mask handling, and class reduction rules to be fixed before evaluation. These details belong in the method rather than being chosen after observing test performance.

C. Descent and Finite Termination of Algorithm 3

For q > 1 and fixed centroids, the FCM membership subproblem with the constraint sum_c u_i,c = 1 has the standard update u_i,c proportional to ||f_i-mu_c||^{-2/(q-1)} when distances are nonzero. For fixed memberships, each centroid is the weighted mean mu_c = sum_i u_i,c^q f_i / sum_i u_i,c^q. These are exact block minimizers of the fuzzy appearance term.

$$u\_i,c = [sum\_k (||f\_i-mu\_c|| / ||f\_i-mu\_k||)^(2/(q-1))]^-1 \quad (10)$$

$$mu\_c = (sum\_i\ u\_i,c^q\ f\_i) / (sum\_i\ u\_i,c^q) \quad (11)$$

Proposition 6 (monotone alternating energy). Suppose each membership-centroid update exactly minimizes its block of (8), each binary graph-cut update exactly minimizes a submodular label block, and all other variables are fixed during that update. Then the sequence of energy values is nonincreasing and converges to a finite limit.

Proof. A block minimization cannot produce an energy larger than the value at the previous block state, so each FCM update and each graph-cut update is nonincreasing. All terms in (8) are nonnegative except the written log term, which is also nonnegative after multiplication by minus eta since 0 < P <= 1. Hence E >= 0. A monotone nonincreasing sequence bounded below converges to a finite limit. The parameter sequence need not converge uniquely, and the limit need not be a global minimum of the joint nonconvex objective.

Proposition 7 (finite label termination). With a finite voxel grid and a deterministic rule that accepts only strictly energy-decreasing label changes, the label-update sequence terminates after finitely many accepted moves.

Proof. For C labels and |Omega| voxels, there are C^|Omega| possible labelings, a finite number. A strict energy decrease prevents revisiting a previous labeling. The accepted sequence can therefore contain at most C^|Omega|-1 changes. In practice, a tolerance and iteration cap are used. For binary submodular energies, s-t cut is globally optimal for the label block. Alpha-expansion for a multiclass metric Potts term gives a strong approximation and a local move optimum, not the global minimizer of the full multiclass energy.

The pairwise weight in (9) is in (0,1] for finite features and positive scale parameters. Similar neighboring voxels receive a larger penalty for label disagreement; sharp intensity or feature changes reduce that penalty. As beta approaches zero, refinement follows the posterior and fuzzy appearance terms. As beta grows, the solution favors piecewise-constant regions and may remove valid small lesions. No universal beta can be proved optimal from (8); it must be selected using validation cases and reported in physical units where component-size thresholds are used.

For a binary label pair, the Potts interaction has costs V(0,0)=V(1,1)=0 and V(0,1)=V(1,0)=beta w_ij. Submodularity requires V(0,0)+V(1,1) <= V(0,1)+V(1,0), which holds whenever beta w_ij is nonnegative. This

condition is why the binary label block can be represented by an s-t cut. If a learned pairwise term becomes negative or asymmetric, the stated cut guarantee no longer applies. Multiclass alpha-expansion further requires a metric or semimetric compatibility cost; the ordinary nonnegative Potts model satisfies that requirement.

Energy descent does not by itself imply better segmentation. The energy contains modeling choices, so minimizing it can move a prediction away from the reference mask when the unary posterior is well localized and the smoothness term is too strong. Let Y_0 be the network labeling and Y_1 the refined labeling. A proper ablation reports Delta Dice = Dice(Y_1,Y_ref)-Dice(Y_0,Y_ref) and the corresponding HD95 change for every case and target. The sign distribution of these paired changes reveals whether refinement consistently helps or merely raises the mean through a subset of large corrections.

D. What the Analysis Does Not Prove

The propositions establish algebraic properties of normalization, fusion, softmax, residual paths, and block-coordinate refinement. They do not prove diagnostic validity, cross-institutional generalization, robustness to missing modalities, or superiority over a baseline. Those are empirical statements. A mathematical section that derives standard equations but labels them a proof of segmentation accuracy is misleading. The paper therefore ends the analysis with explicit boundaries between structural guarantees and evidence that must come from experiments.

5. EXPERIMENTAL PROTOCOL

A. Dataset and Targets

The source manuscript identifies the BraTS 2020 dataset and the four standard MRI sequences. Each original volume has 240 x 240 x 155 voxels and expert-derived tumor labels. WT is the union of edema, necrotic/nonenhancing core, and enhancing tumor; TC is the union of core and enhancing tumor; ET is the enhancing class. Patient-level separation is mandatory. Slices or patches from one subject may not appear in more than one partition.

B. Source-Reported Implementation

The available record states that volumes were cropped or resized to 128 x 128 x 128, the encoder used channel widths 8 through 128, training used Adam with an initial learning rate of 5 x 10^-4, batch size 2, 100 epochs, dropout between 0.1 and 0.3, TensorFlow/Keras, and an NVIDIA A100 with 40 GB memory. Adam is a standard adaptive optimizer [37]. AdamW separates weight decay from the adaptive update and is a reasonable alternative only if its decay coefficient is reported [38]. Focal Tversky loss is another documented option for rare lesion regions [39], but it should not be listed unless it is actually used.

The source is internally inconsistent about channel count: it lists four BraTS modalities but later describes a three-channel 128-cubed tensor. This revision defines four input modalities. If one sequence was omitted in the actual experiment, the model must be retrained or the text and figures must identify the exact three-sequence configuration. The record also does not state the random seed, patient split, augmentation probabilities, loss weights, validation criterion, graph-cut parameters, software versions, checkpoint policy, or inference-time resampling.

C. Required Replication Protocol

A defensible replication should use a predeclared patient-level split or five-fold cross-validation. Every fold should retain the same preprocessing, epoch budget, and model-selection rule. Hyperparameters must be tuned only on validation data. At least three seeds should be reported for the final configuration. Per-case WT, TC, and ET metrics should be stored, and comparisons should use paired tests or bootstrap confidence intervals. Competition rankings can change with metric aggregation and missing reporting; biomedical challenge results require careful interpretation [40].

Reference masks may contain annotation variability. STAPLE provides a probabilistic framework for combining multiple segmentations when multiple raters are available [41], and generalized overlap measures clarify how label combinations affect evaluation [42]. Those methods do not replace BraTS reference labels; they motivate explicit treatment of annotation uncertainty. Probability calibration should be checked on validation data, since modern neural networks can be overconfident [43]. Deep ensembles offer one practical uncertainty estimate at increased compute cost [44].

D. Metrics

$$Dice = 2TP / (2TP + FP + FN), \quad IoU = TP / (TP + FP + FN) \quad (12)$$

$$Sensitivity = TP / (TP + FN), \quad Specificity = TN / (TN + FP) \tag{13}$$

$$Precision = TP / (TP + FP), \quad Accuracy = (TP + TN) / (TP + TN + FP + FN) \tag{14}$$

HD95 is the 95th percentile of the bidirectional boundary-to-boundary distance distribution in millimeters. Dice and IoU should be computed per subject and per target, then summarized. Empty ET masks require a declared convention. Accuracy and specificity should not be treated as primary endpoints on a background-dominated volume.

TABLE II MINIMUM REPRODUCIBILITY RECORD

Record	**Required content**
Data	Case identifiers or split hash; modality list; label mapping
Preprocessing	Crop, interpolation, normalization, augmentation, leakage controls
Model	Layer configuration, parameters, initialization, software commit
Training	Seed, optimizer, learning-rate schedule, epochs, checkpoint rule
Refinement	q, eta, beta, edge scales, component threshold
Evaluation	Per-case WT/TC/ET values, empty-mask rule, confidence intervals
Resources	GPU, memory, training time, inference time, parameter count, FLOPs

6. RESULTS

A. Source-Reported Aggregate Values

Table III reproduces the aggregate values written in the supplied manuscript. They are not re-estimated from predictions. The table removes the draft's unsupported bars for Modified U-Net, Z-Net, and Mask R-CNN, since exact values, citations, training conditions, and test partitions were not supplied. The only baseline with numerical values in the prose is U-Net.

TABLE III SOURCE-REPORTED RESULTS; NOT INDEPENDENTLY RE-ESTIMATED

Metric	**U-Net**	**Proposed**	**Delta**
Dice	0.85143	0.98381	+0.13238
IoU	0.821	0.968	+0.147
Sensitivity	0.805	0.956	+0.151
Specificity	0.830	0.972	+0.142
Accuracy	0.845	0.979	+0.134
Recall	0.823	0.967	+0.144
Precision	0.838	0.974	+0.136
HD95 (mm)	7.2	3.8	-3.4

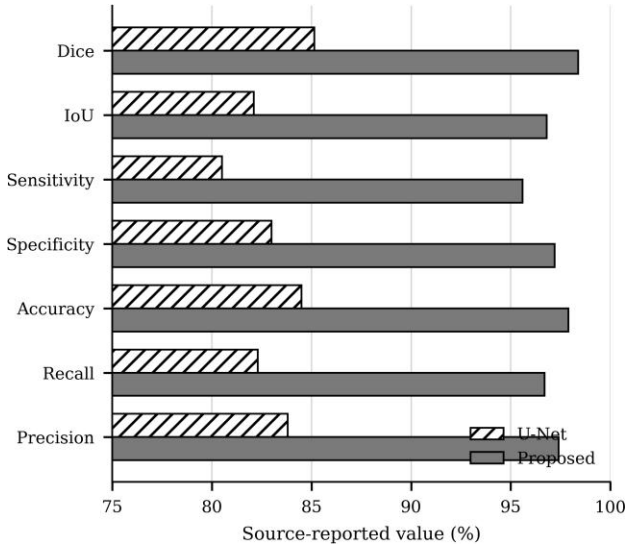

Fig. 2. Source-reported aggregate overlap and voxel-classification metrics. Values were transcribed from the supplied manuscript and were not recomputed.

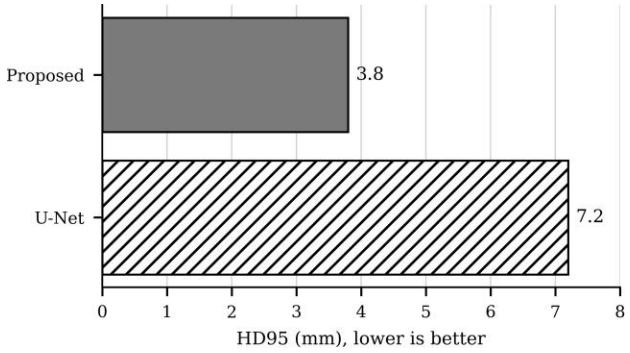

Fig. 3. Source-reported HD95 comparison. Lower values indicate smaller boundary discrepancy.

B. Internal Consistency Checks

For one binary confusion matrix, Dice and IoU satisfy IoU = Dice/(2-Dice). The proposed Dice value 0.98381 predicts IoU approximately 0.96814, which agrees with the reported 0.968 after rounding. The U-Net Dice value 0.85143 predicts IoU approximately 0.741, not the reported 0.821. The discrepancy could result from different target regions, macro versus micro averaging, or transcription error. The paper cannot select among those explanations without per-case records.

$$IoU = Dice / (2 - Dice) \qquad (15)$$

Sensitivity and recall are the same quantity when they use the same positive class and aggregation. The source reports 0.956 sensitivity and 0.967 recall for the proposed model, and 0.805 versus 0.823 for U-Net. Those pairs cannot both describe the same aggregation. They are preserved in Table III to maintain traceability, but a final experimental revision should keep one term or explain different targets and averaging rules.

The reported HD95 improvement from 7.2 to 3.8 mm is directionally consistent with better boundary localization. A pooled number does not show whether gains occur in WT, TC, or ET, and it does not reveal failures on small enhancing lesions. Source-reported accuracy and specificity are high, but both are influenced by the dominant background. Dice, per-region sensitivity, precision, and HD95 should drive the main comparison.

C. Evidence Required for Comparative Claims

The reported values are markedly above many published BraTS results, so they require a complete audit trail: patient split identifiers, unchanged baseline training, raw prediction masks, metric code, and confidence intervals. Without those items, the document can state that the source record reports the values, but it cannot claim state-of-the-art performance, clinical readiness, or statistical superiority. Parameter count, FLOPs, training time, and inference time were not present, so efficiency remains an architectural intention rather than a measured result.

7. DISCUSSION

A. Architectural Interpretation

The coherent model has a clear division of labor. Robust conditioning standardizes input range and preserves spatial correspondence. Modality gates learn scale-dependent contributions. Residual inception blocks retain local multiscale evidence, bottleneck context supplies long-range interaction at manageable token count, and guided skip gates control high-resolution features. Fuzzy graph-cut refinement introduces an interpretable tradeoff between posterior confidence and spatial coherence. This division is stronger than the original draft's sequence of disconnected algorithms.

The design is not a claim that every component is new. Residual, inception, attention, transformer, FCM, and graph-cut mechanisms are established and cited near their use. The integration can be considered a contribution only if an ablation holds the split, preprocessing, optimizer, and evaluation constant. A minimum ablation would compare: base 3-D U-Net; plus residual inception; plus modality gating; plus bottleneck context; plus guided skips; and plus graph-fuzzy refinement. Each row should report WT, TC, and ET with uncertainty and computational cost.

B. Modality Fusion and Missing Data

A softmax gate can reveal relative modality use, but its weights are not causal explanations. Removing one sequence at inference changes the input distribution and may invalidate the learned gate. A missing-modality study should train or adapt the model under modality dropout and evaluate every clinically relevant subset. Filling an absent sequence with zeros without training for that state can create a recognizable but artificial pattern. The current source record evaluates neither missing data nor cross-site transfer.

C. Clinical and Reporting Interpretation

Segmentation masks can support volume measurement and target review, but the present experiment does not evaluate lesion-level detection, tumor grading, survival prediction, treatment response, or clinical decision impact. The word detection should therefore refer only to spatial localization implicit in the segmentation mask. A clinical claim would require external data, reader studies, failure analysis, inference-time constraints, and a predefined workflow showing how a radiologist reviews or corrects the output.

Boundary refinement has an asymmetric risk. Removing a tiny false-positive island may improve precision, yet removing a true small enhancing focus can be clinically consequential. Component filtering should use voxel spacing to define a physical-volume threshold, and results should be reported before and after refinement. The graph term should be inspected on low-contrast boundaries, multifocal disease, postoperative cavities, and cases with very small ET.

D. Statistical Validity and Failure Modes

Aggregate scores hide distribution shift and difficult cases. Per-case plots should stratify by tumor grade, lesion size, acquisition site when available, and presence of ET. Bootstrap confidence intervals can summarize metric uncertainty; paired tests can compare two models on the same cases. Repeated seeds are necessary to separate architecture effects from stochastic optimization. The test set must remain untouched until the model and all thresholds are fixed.

Expected failure modes include incomplete skull stripping, extreme bias field, motion, unusual postoperative anatomy, missing or corrupted modalities, tiny disconnected ET, and intensity patterns outside the training range. Confidence calibration and ensemble disagreement can flag uncertain volumes, but uncertainty is not a substitute for error detection. A quality-control interface should display the four modalities, probability maps, final labels, component volumes, and any case-level warnings.

E. Reproducibility Limitations

The present revision is constrained by the supplied document. No executable code, training log, subject split, predictions, model checkpoint, or metric export was available. The numerical section is consequently a checked transcription and consistency audit, not an independent experiment. The four-versus-three modality conflict, the Dice-IoU mismatch for U-Net, and the sensitivity-recall mismatch must be resolved from original experiment files before submission.

A second limitation is that resizing from 240 x 240 x 155 to 128 cubed changes physical geometry and may compress small lesions. A reproducible study should state whether the operation is cropping, resampling, padding, or a combination, give the final voxel spacing, and compute HD95 after restoring predictions to the reference geometry. Nearest-neighbor interpolation is required for labels. Reporting only tensor dimensions is insufficient.

A third limitation is the absence of measured efficiency. The title uses the word efficient, yet memory, throughput, FLOPs, and parameter count are not reported. Efficiency should be established relative to a defined baseline on the same hardware and input size. Bottleneck attention may reduce token cost, but the modality stems and 3-D feature maps can still dominate memory. Mixed precision, sliding-window inference, and patch overlap must be documented if used.

8. CONCLUSION

This paper has reformulated the supplied draft as a coherent IEEE-style study of multilevel-fusion 3-D brain-tumor segmentation. The MLF-3D pipeline combines bounded multimodal conditioning, scale-dependent modality fusion, residual inception encoding, bottleneck context, guided skip decoding, and fuzzy graph-cut refinement. Three algorithms and seven propositions clarify what the system guarantees: bounded and order-preserving conditioning, convex fusion, normalized probabilities, an identity gradient path, and monotone block-coordinate refinement under stated assumptions. None of these results is presented as a proof of clinical accuracy or global convergence.

The source-reported performance values were reproduced in monochrome tables and figures, then checked for algebraic consistency. The proposed Dice and IoU values agree after rounding, but the U-Net Dice-IoU pair and both sensitivity-recall pairs require correction or an explanation of aggregation. The absence of case-level predictions, split identifiers, repeated runs, code, and resource measurements prevents a defensible state-of-the-art or efficiency claim. Completing the replication record in Table II, reporting WT, TC, and ET separately, and running a fixed ablation are the necessary next steps before peer-review submission.

References

1. B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," IEEE Trans. Med. Imag., vol. 34, no. 10, pp. 1993-2024, Oct. 2015, doi: 10.1109/TMI.2014.2377694.
2. S. Bakas et al., "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," Sci. Data, vol. 4, Art. no. 170117, Sep. 2017, doi: 10.1038/sdata.2017.117.
3. F. Isensee, P. F. Jager, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, "nnU-Net for brain tumor segmentation," in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, LNCS 12659, 2021, pp. 118-132, doi: 10.1007/978-3-030-72087-2_11.
4. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. MICCAI, LNCS 9351, 2015, pp. 234-241, doi: 10.1007/978-3-319-24574-4_28.
5. O. Cicek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in Proc. MICCAI, LNCS 9901, 2016, pp. 424-432, doi: 10.1007/978-3-319-46723-8_49.
6. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in Proc. 4th Int. Conf. 3D Vision, 2016, pp. 565-571, doi: 10.1109/3DV.2016.79.
7. S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," IEEE Trans. Med. Imag., vol. 35, no. 5, pp. 1240-1251, May 2016, doi: 10.1109/TMI.2016.2538465.
8. M. Havaei et al., "Brain tumor segmentation with deep neural networks," Med. Image Anal., vol. 35, pp. 18-31, Jan. 2017, doi: 10.1016/j.media.2016.05.004.
9. K. Kamnitsas et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," Med. Image Anal., vol. 36, pp. 61-78, Feb. 2017, doi: 10.1016/j.media.2016.10.004.
10. A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in Brainlesion, LNCS 11384, 2019, pp. 311-320, doi: 10.1007/978-3-030-11726-9_28.

11. F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," Nat. Methods, vol. 18, pp. 203-211, Feb. 2021, doi: 10.1038/s41592-020-01008-z.
12. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.
13. C. Szegedy et al., "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.
14. O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," arXiv:1804.03999, 2018.
15. E. K. Rutoh, Q. Z. Guang, N. Bahadar, R. Raza, and M. S. Hanif, "GAIR-U-Net: 3D guided attention inception residual U-Net for brain tumor segmentation using multimodal MRI images," J. King Saud Univ.-Comput. Inf. Sci., vol. 36, no. 6, Art. no. 102086, 2024, doi: 10.1016/j.jksuci.2024.102086.
16. W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "TransBTS: Multimodal brain tumor segmentation using transformer," in Proc. MICCAI, LNCS 12901, 2021, pp. 109-119, doi: 10.1007/978-3-030-87193-2_11.
17. A. Hatamizadeh et al., "UNETR: Transformers for 3D medical image segmentation," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis., 2022, pp. 574-584, doi: 10.1109/WACV51458.2022.00181.
18. A. Hatamizadeh et al., "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in Brainlesion, LNCS 12962, 2022, pp. 272-284, doi: 10.1007/978-3-031-08999-2_22.
19. A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 5998-6008.
20. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in DLMIA/ML-CDS, LNCS 11045, 2018, pp. 3-11, doi: 10.1007/978-3-030-00889-5_1.
21. N. J. Tustison et al., "N4ITK: Improved N3 bias correction," IEEE Trans. Med. Imag., vol. 29, no. 6, pp. 1310-1320, Jun. 2010, doi: 10.1109/TMI.2010.2046908.
22. L. G. Nyul, J. K. Udupa, and X. Zhang, "New variants of a method of MRI scale standardization," IEEE Trans. Med. Imag., vol. 19, no. 2, pp. 143-150, Feb. 2000, doi: 10.1109/42.836373.
23. J. Nalepa, M. Marcinkiewicz, and M. Kawulok, "Data augmentation for brain-tumor segmentation: A review," Front. Comput. Neurosci., vol. 13, Art. no. 83, Dec. 2019, doi: 10.3389/fncom.2019.00083.
24. C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations," in DLMIA/ML-CDS, LNCS 10553, 2017, pp. 240-248, doi: 10.1007/978-3-319-67558-9_28.
25. S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," in MLMI, LNCS 10541, 2017, pp. 379-387, doi: 10.1007/978-3-319-67389-9_44.
26. A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," BMC Med. Imag., vol. 15, Art. no. 29, Aug. 2015, doi: 10.1186/s12880-015-0068-x.
27. D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, no. 9, pp. 850-863, Sep. 1993, doi: 10.1109/34.232073.
28. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images," in Proc. IEEE Int. Conf. Comput. Vis., 2001, pp. 105-112, doi: 10.1109/ICCV.2001.937505.
29. J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," Comput. Geosci., vol. 10, nos. 2-3, pp. 191-203, 1984, doi: 10.1016/0098-3004(84)90020-7.
30. D. L. Pham, C. Xu, and J. L. Prince, "Current methods in medical image segmentation," Annu. Rev. Biomed. Eng., vol. 2, pp. 315-337, 2000, doi: 10.1146/annurev.bioeng.2.1.315.
31. T. Henry et al., "Brain tumor segmentation with self-ensembled, deeply-supervised 3D U-Net neural networks: A BraTS 2020 challenge solution," in Brainlesion, LNCS 12658, 2021, pp. 327-339, doi: 10.1007/978-3-030-72084-1_30.
32. F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "No New-Net," in Brainlesion, LNCS 11384, 2019, pp. 234-244, doi: 10.1007/978-3-030-11726-9_21.
33. J. Lin et al., "CKD-TransBTS: Clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation," IEEE Trans. Med. Imag., vol. 42, no. 8, pp. 2451-2461, Aug. 2023, doi: 10.1109/TMI.2023.3250474.
34. W. Zhang, S. Chen, Y. Ma, Y. Liu, and X. Cao, "ETUNet: Exploring efficient transformer enhanced UNet for 3D brain tumor segmentation," Comput. Biol. Med., vol. 171, Art. no. 108005, Mar. 2024, doi: 10.1016/j.compbiomed.2024.108005.
35. B. Jagadeesh and G. A. Kumar, "Brain tumor segmentation with missing MRI modalities using edge aware discriminative feature fusion based transformer U-Net," Appl. Soft Comput., vol. 161, Art. no. 111709, Aug. 2024, doi: 10.1016/j.asoc.2024.111709.
36. D. An, P. Liu, Y. Feng, P. Ding, W. Zhou, and B. Yu, "Dynamic weighted knowledge distillation for brain tumor segmentation," Pattern Recognit., vol. 155, Art. no. 110731, Nov. 2024, doi: 10.1016/j.patcog.2024.110731.
37. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Represent., 2015.
38. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. Int. Conf. Learn. Represent., 2019.
39. N. Abraham and N. M. Khan, "A novel focal Tversky loss function with improved attention U-Net for lesion segmentation," in Proc. IEEE 16th Int. Symp. Biomed. Imag., 2019, pp. 683-687, doi: 10.1109/ISBI.2019.8759329.
40. L. Maier-Hein et al., "Why rankings of biomedical image analysis competitions should be interpreted with care," Nat. Commun., vol. 9, Art. no. 5217, Dec. 2018, doi: 10.1038/s41467-018-07619-7.

41. S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," IEEE Trans. Med. Imag., vol. 23, no. 7, pp. 903-921, Jul. 2004, doi: 10.1109/TMI.2004.828354.
42. W. R. Crum, O. Camara, and D. L. G. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," IEEE Trans. Med. Imag., vol. 25, no. 11, pp. 1451-1461, Nov. 2006, doi: 10.1109/TMI.2006.880587.
43. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Mach. Learn., PMLR 70, 2017, pp. 1321-1330.
44. B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 6402-6413.