Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluating Boundary-Aware Losses for U-KAN on BraTS 2023 Brain Tumor Segmentation

Mourad Hikki

m.hikki.ced@uca.ac.ma

Cadi Ayyad University (UCA)

Aziz Darouichi

Cadi Ayyad University (UCA)

Abdessamad Elboushaki

Cadi Ayyad University (UCA)

Research Article

Keywords:

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11040029/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Evaluating Boundary-Aware Losses for U-KAN on BraTS 2023 Brain Tumor Segmentation

Mourad Hikki[1,2*], Aziz Darouichi[1] and Abdessamad Elboushaki[1]

[1]L2IS Lab, Faculty of Science and Technology (FST), Cadi Ayyad University (UCA), P.O. Box 549, Ave. Abdelkarim Elkhattabi, Marrakech, 40000, Morocco.
[2]LAMIGEP, EMSI Marrakech, Marrakech, Morocco.

*Corresponding author(s). E-mail(s): m.hikki.ced@uca.ac.ma;

Abstract

Background: Boundary-aware loss functions are intended to improve segmentation by penalizing spatial errors that may be insufficiently represented by region-based objectives. Whether they provide meaningful and reproducible benefits when incorporated into U-KAN for brain tumor segmentation remains unclear.

Methods: We compared five loss configurations in separate 2D and 3D U-KAN pipelines using multimodal MRI from the BraTS 2023 adult glioma dataset: a region-based baseline, fixed Boundary Loss, Staged Boundary Loss, a Generalized Surface Loss composite, and standalone Boundary Difference-over-Union loss. Each configuration was trained with three seeds in each pipeline, producing 30 runs. Performance was assessed using regional and macro Dice, the 95th-percentile Hausdorff distance (HD95), run-level consistency, and patient-level paired comparisons against the corresponding baseline.

Results: In 2D, Staged Boundary Loss achieved the highest mean macro Dice (0.8139 versus 0.8001) and reduced mean HD95 (22.55 versus 34.81 mm). Patient-level comparisons favored this configuration, but its overlap improvement was concentrated in whole-tumor segmentation and it exceeded the baseline in only one of three matched seed comparisons. In 3D, Staged Boundary Loss produced only small descriptive improvements in macro Dice (0.89794 versus 0.89698) and HD95 (3.645 versus 3.736 mm), with no significant favorable patient-level difference after multiplicity correction. The other boundary-aware configurations did not consistently outperform the baselines, and standalone Boundary Difference-over-Union exhibited substantial instability in one 2D run.

Conclusions: Boundary-aware training did not provide a consistent advantage over region-based objectives. Staged Boundary Loss was the most promising configuration, but its benefits were pipeline-dependent and not reliably reproduced across training runs.

Keywords: Brain tumor segmentation, boundary-aware loss, U-KAN, BraTS 2023, medical image segmentation, deep learning

1 Introduction

Artificial intelligence provides tools for transforming complex data into information that can support scientific investigation and practical decision-making. In medicine, these tools have been applied to image classification, abnormality detection, image registration, and the delineation of anatomical structures. Deep learning enables such systems to learn representations directly from examples, reducing reliance on manually designed image features. Among these applications, image segmentation provides a link between visual information and quantitative analysis by assigning tissue or structure labels to individual pixels or voxels. The resulting delineations allow image data to be represented through spatial measurements rather than visual inspection alone [1].

Brain tumor segmentation illustrates both the potential and the difficulty of this approach. Gliomas exhibit substantial variation in location, size, shape, and internal appearance, while the boundaries between abnormal and surrounding tissue can be difficult to identify, [2]. Multimodal magnetic resonance imaging (MRI) provides complementary information for distinguishing tumor-associated edema, necrotic or nonenhancing core, and enhancing tumor. However, manually delineating these regions throughout a three-dimensional examination is time-consuming and subject to disagreement between observers. The Brain Tumor Segmentation (BraTS) initiative has supported research in this area through shared imaging datasets, expert annotations, and common evaluation tasks, providing a basis for comparing automated segmentation methods [3–5].

Deep learning approaches to segmentation have benefited from encoder–decoder architectures that combine contextual information with spatial detail. U-Net established an influential design in which a contracting path extracts increasingly abstract features and an expanding path restores spatial resolution, with skip connections linking corresponding stages [6]. Volumetric extensions such as 3D U-Net and V-Net established end-to-end architectures for three-dimensional segmentation and highlighted the relevance of Dice-based objectives under foreground–background imbalance [7, 8]. U-KAN extends this design by incorporating Kolmogorov–Arnold Network components into intermediate image representations [9]. Its combination of a familiar U-shaped structure with learnable nonlinear transformations makes it a suitable backbone for investigating the effects of different training objectives. Accordingly, this study adopts U-KAN as a common architecture for comparing boundary-aware loss formulations in brain tumor segmentation.

Architecture alone, however, does not determine segmentation behavior. During training, the loss function defines how discrepancies between predictions and reference annotations contribute to model updates. Cross-entropy and Dice-based objectives emphasize voxelwise classification and regional overlap, respectively. Boundary-aware objectives introduce additional sensitivity to the spatial arrangement of errors, for example through distance-based penalties or target-size-dependent formulations intended to emphasize boundary mismatch. These approaches offer a rationale for improving delineation beyond regional overlap, but they do not constitute a single interchangeable family: some complement regional objectives, whereas others are designed for standalone use [10, 11].

Whether this additional boundary emphasis produces consistent benefits requires empirical evaluation. A large comparative study of medical segmentation losses found that no evaluated objective performed best across all investigated tasks, although compound losses were generally robust [12]. Recent validation work has likewise emphasized that apparent improvements in medical image segmentation depend on the quality of the baseline, dataset, computational budget, and evaluation protocol [13]. For U-KAN brain tumor segmentation, the relevant question is therefore not simply whether a boundary-aware configuration achieves a higher average score, but whether any improvement extends across tumor regions, boundary measurements, and repeated training runs. Examining these aspects together also allows favorable averages to be distinguished from inconsistent gains or occasional severe failures.

This study compares five loss configurations using slice-based two-dimensional and volumetric three-dimensional U-KAN pipelines on the labeled BraTS 2023 glioma collection. The configurations comprise a cross-entropy–Dice baseline, fixed-weight Boundary Loss, a staged Boundary Loss schedule, a Generalized Surface Loss composite, and standalone Boundary Difference-over-Union loss. Each configuration is evaluated using three training seeds in each pipeline, yielding 30 runs. The baseline uses binary cross-entropy in 2D and multiclass cross-entropy in 3D. Comparisons are conducted separately within each pipeline because their data cohorts and target representations differ.

The study makes three contributions:

- A systematic evaluation of five loss configurations in both slice-based 2D and volumetric 3D U-KAN pipelines, comprising 30 independently trained runs.
- A repeated-seed analysis of mean performance, run-to-run variability, and matched seed-level differences, distinguishing isolated gains from improvements that recur across training runs.
- A multilevel assessment combining region-wise Dice and boundary-distance measurements with patient-level paired comparisons against the corresponding regional baseline.

Together, these analyses assess whether boundary-aware training offers consistent improvements under the evaluated conditions.

2 Related Work

2.1 U-shaped and KAN-based segmentation networks

U-Net established an influential encoder–decoder design in which a contracting path learns increasingly abstract feature representations and an expanding path restores spatial resolution, with skip connections linking corresponding stages [6]. The 3D U-Net and V-Net adaptations established closely related volumetric baselines, replacing two-dimensional operators with three-dimensional processing and, in the case of V-Net, directly optimizing a Dice-based objective [7, 8]. Building on this framework, U-KAN retained the multiscale encoder–decoder structure and skip connections while introducing tokenized Kolmogorov–Arnold network (KAN) transformations into intermediate feature representations [9]. The original model was evaluated on breast ultrasound, gland histology, and colonoscopy segmentation datasets, and subsequent work extended this direction to other medical imaging settings. Implicit U-KAN2.0 combined MultiKAN layers with second-order neural ordinary differential equations and was evaluated on multiple 2D datasets and one 3D medical segmentation dataset [14]. Ali *et al.* applied U-KAN to multiclass nuclei segmentation in histopathology using an adaptive sliding-window strategy [15]. Together, these studies demonstrate the application of U-KAN and its extensions across different modalities and anatomical targets.

Within brain tumor segmentation, an initial study adapted U-KAN to volumetric multimodal MRI and introduced an attention-enhanced UKAN-SE variant on BraTS 2024 [16]. This research line was subsequently developed into UKAN-EP, which combines a 3D U-KAN backbone with efficient channel attention, pyramid feature aggregation, and dynamic loss weighting [17]. Other modern U-shaped approaches include Swin UNETR, which uses a hierarchical transformer encoder for multimodal 3D brain-tumor MRI segmentation [18]. Independently, ResKANNet incorporated KAN modules into a residual U-shaped architecture with multiscale attention for 3D brain tumor segmentation [19]. These studies establish KAN-enhanced and other modern U-shaped networks as relevant design families for this task. Building on that foundation, the present study retains U-KAN as the common architecture within each pipeline and investigates how the choice of loss configuration affects segmentation performance.

2.2 Boundary-aware segmentation objectives

Region-based objectives optimize voxel classification or foreground overlap, whereas boundary-aware objectives introduce spatial information about contour errors. Kervadec *et al.* express a boundary objective using predicted probabilities and a ground-truth distance map, and combine it with a regional loss [10]. Generalized Surface Loss (GSL) uses a normalized distance-based formulation to target surface accuracy [20]. Boundary Difference-over-Union (Boundary DoU) instead adapts an overlap objective according to the boundary-to-region relationship [11]. In glioma segmentation, boundary-sensitive behavior has also been pursued through hierarchical

attention and hybrid losses, while recent 3D medical segmentation work has introduced explicit hard and soft boundary constraints within knowledge distillation [21, 22].

These formulations motivate, but do not guarantee, improved brain tumor segmentation with U-KAN. Their behavior may depend on output encoding, the treatment of absent regions, and the balance between regional and boundary terms. We therefore compare a regional baseline, fixed and staged Boundary Loss configurations, a GSL-based composite, and standalone Boundary DoU.

3 Materials and Methods

3.1 Study design and data

The study uses the adult glioma training collection of the ASNR–MICCAI BraTS 2023 challenge, within the BraTS challenge framework [3–5, 23]. Inputs contain four MRI modalities: native T1-weighted, contrast-enhanced T1-weighted, T2-weighted, and T2-FLAIR images. Predictions are assessed for whole tumor (WT), tumor core (TC), and enhancing tumor (ET). WT contains edema and tumor core; TC comprises non-enhancing/necrotic core and enhancing tumor; ET is the enhancing component.

Five loss configurations were evaluated using both a slice-based 2D pipeline and a volumetric 3D pipeline, with three training-seed labels (42, 123, and 3407) per configuration, yielding 30 runs. The test exports contain 126 scan cases from 126 unique patient identifiers for every 2D run. For every 3D run, they contain 250 scan cases from 242 unique patient identifiers because some patients contributed scans from more than one timepoint. Accordingly, the scan-level descriptive analyses use 250 cases, whereas the patient-level paired analyses aggregate repeated timepoints and use 242 patients.

Table 1 summarizes the cohort composition and case-level splits used by the two pipelines. Although both pipelines derive from the same labeled BraTS collection, their preprocessing and split procedures produce different training, validation, and test cohorts; therefore, performance comparisons are made within each pipeline rather than as a direct comparison of dimensionality.

Figure 1 summarizes the overall experimental design used in this study: the five loss configurations are evaluated in separate 2D slice-based and 3D volumetric U-KAN pipelines with three training seeds each, producing 30 runs in total, followed by checkpoint selection, test-set inference, metric computation, and run-level and patient-level analyses.

3.2 Preprocessing and network configurations

The 2D pipeline processes axial slices, ordered as T1c, T1n, T2-FLAIR, and T2w input channels. Nonzero intensities are z-score normalized separately for each modality and slice, and slices are zero-padded to 256×256. Training and validation preprocessing retains slices with at least one tumor label. Spatial augmentation applies random quarter-turn rotations and horizontal flips jointly to images, masks, and, where required, distance maps.

Table 1 Cohort composition and case-level splits.

Characteristic	2D U-KAN pipeline	3D U-KAN pipeline
Source cohort	BraTS 2023 adult-glioma training collection	Same
Total labeled scan cases	1,251	1,251
MRI modalities	T1c, T1n, T2-FLAIR, T2w	T2w, T2-FLAIR, T1n, T1c
Target representation	Three overlapping regions: WT, TC, ET	Four exclusive classes: background, NCR/NET, edema, ET
Training cases	999	875
Validation cases	126	125
Held-out test cases	126 scans from 126 patients	250 scans from 242 patients
Repeated timepoints in test set	None identified	8 patients contributed two scans
Split construction	Case-level split; 10% validation and 10% test	Case-level split; nominal 70/10/20
Experimental runs	5 losses × 3 seeds = 15	5 losses × 3 seeds = 15

The 2D and 3D pipelines use the same BraTS 2023 source cohort but different case-level splits and target representations. Their results are therefore not direct dimensionality comparisons.

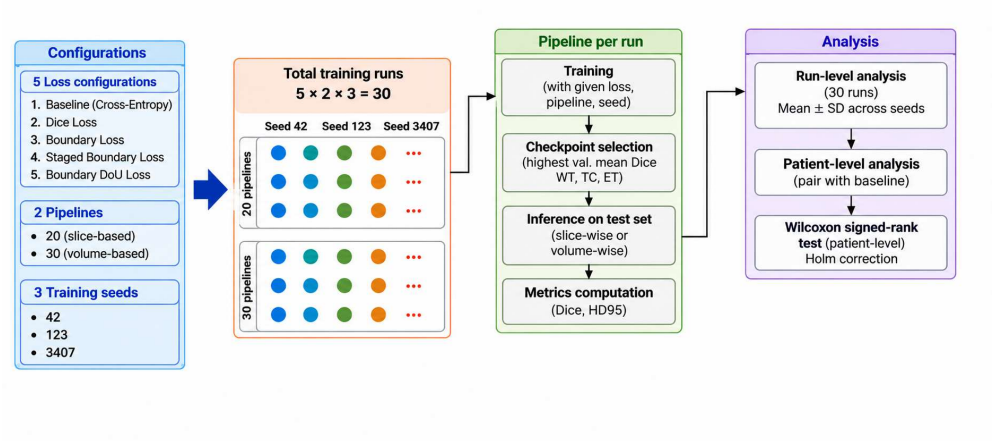


Fig. 1 Overall experimental design. Five loss configurations are evaluated in two pipelines with three training seeds, producing 30 runs. Each run includes training, checkpoint selection, test-set inference, and metric computation, followed by run-level and patient-level analyses.

The 2D U-KAN has four input channels, embedding widths $[128, 160, 256]$, and three independent sigmoid outputs representing WT, TC, and ET. These overlapping targets are optimized directly. The architecture combines convolutional encoder/decoder operations with tokenized KAN blocks.

The 3D implementation replaces the corresponding spatial operators with volumetric operations while retaining the U-KAN layout and embedding widths $[128, 160, 256]$. Preprocessing crops to the nonzero image bounding box, performs modality-wise z-score normalization, and targets 1-mm isotropic spacing. The modality order is T2w, T2-FLAIR, T1n, and T1c. Training uses $128 \times 128 \times 128$ patches with augmentation. Four softmax channels represent background, non-enhancing/necrotic core, edema, and enhancing tumor. WT, TC, and ET are derived from the exclusive class labels

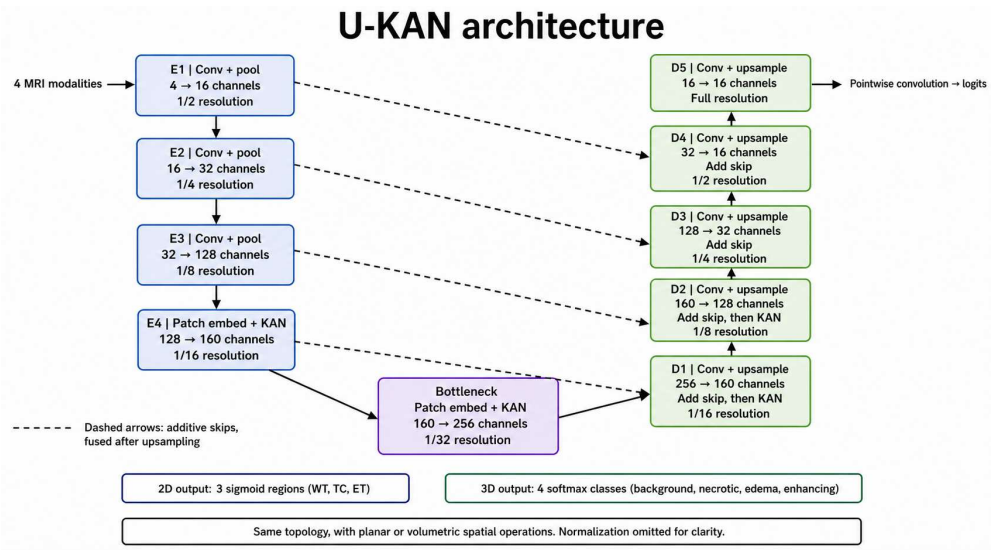


Fig. 2 Shared topology of the 2D and 3D U-KAN implementations evaluated in this study. Boxes show channel transitions and spatial resolution relative to the input along each axis. Decoder boxes group convolution, upsampling, skip addition, and, in D1 and D2, KAN processing. Dashed arrows denote additive skips fused after upsampling; D5 has no skip. Spatial operators and output encodings are pipeline-specific. Normalization layers are omitted from the schematic for clarity.

for evaluation. Thus, a similarly named loss is adapted to a multilabel task in 2D and a multiclass task in 3D.

Figure 2 summarizes the shared topology. Three convolutional encoder stages followed by max pooling produce 16, 32, and 128 channels at one-half, one-quarter, and one-eighth of the input spatial resolution. Two stride-two patch embeddings introduce tokenized KAN stages with 160 and 256 channels, reaching one-thirty-second resolution at the bottleneck. Each KAN block applies layer normalization followed by three sequential KANLinear transformations, each paired with a depthwise convolution, batch normalization, and ReLU, within a residual connection. Token stages additionally use output layer normalization before restoring spatial feature maps.

Five decoder stages restore full resolution with 160, 128, 32, 16, and 16 channels. Each stage applies a two-convolution block followed by bilinear upsampling in 2D or trilinear upsampling in 3D. The first four stages then add the matching encoder features element-wise; fusion therefore occurs after channel conversion and upsampling. The first two decoder stages additionally apply KAN blocks after fusion. Encoder skips E1, E2, E3, and E4 connect to D4, D3, D2, and D1, respectively; D5 has no encoder skip. A pointwise convolution produces three logits for the 2D regional outputs or four logits for the 3D class outputs, followed by the corresponding sigmoid or softmax activation when probabilities are needed.

3.3 Loss configurations

Let p_{bci} and g_{bci} denote a predicted probability and binary target for sample b, channel c, and spatial location i. All equations below describe the implemented adaptations. Smoothing is $\epsilon = 10^{-5}$.

3.3.1 Regional baseline

For 2D, the regional objective is

$$L_{\text{reg}}^{2D} = 0.5L_{\text{BCE}} + 1 - \frac{1}{B}\sum_{b}\frac{2\sum_{c,i} p_{bci}g_{bci} + \epsilon}{\sum_{c,i} p_{bci} + \sum_{c,i} g_{bci} + \epsilon}. \tag{1}$$

The soft Dice term is computed over all channels and pixels per sample. For 3D, binary cross-entropy is replaced by multiclass cross-entropy, and the Dice term is averaged separately over samples and all four classes, including background:

$$L_{\text{reg}}^{3D} = 0.5L_{\text{CE}} + 1 - \frac{1}{BC}\sum_{b,c}\frac{2\sum_{i} p_{bci}g_{bci} + \epsilon}{\sum_{i} p_{bci} + \sum_{i} g_{bci} + \epsilon}. \tag{2}$$

Accordingly, the baseline is BCE–Dice in 2D and CE–Dice in 3D, not an identical objective across dimensions.

3.3.2 Fixed and Staged Boundary Loss Configurations

The signed distance map ϕ_{bc} is positive outside the target and nonpositive inside, using the implemented convention

$$\phi = \mathrm{d}_{\text{out}}(1 - g) - (\mathrm{d}_{\text{in}} - 1)g, \tag{3}$$

where the distance transforms use the training pixel/voxel grid; absent target channels receive a zero map. The boundary composite is

$$L_{\text{B}}(t) = \alpha(t)L_{\text{reg}} + [1 - \alpha(t)]\,\text{mean}_{b,c,i}(\phi_{bci}p_{bci}). \tag{4}$$

The fixed Boundary Loss configuration uses $\alpha = 0.5$. The staged Boundary Loss configuration begins with the regional objective alone, progressively introducing the boundary term. For zero-indexed epoch t,

$$\alpha(t) = \begin{cases} 1, & 0 \le t < 10, \\ 1 - 0.5(t-9)/20, & 10 \le t < 30, \\ 0.5 - 0.3(t-29)/15, & 30 \le t < 45, \\ 0.2, & t \ge 45. \end{cases} \tag{5}$$

The resulting boundary-term weight, $1 - \alpha(t)$, is shown in Fig. 3.

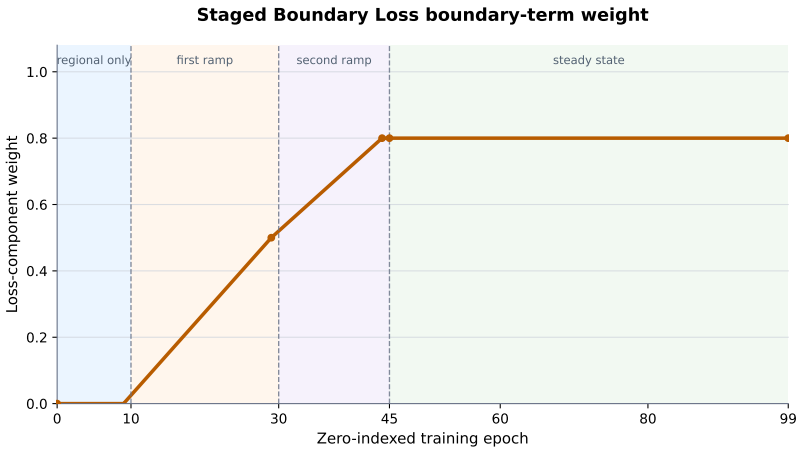


Fig. 3 Boundary-term weight $1 - \alpha(t)$ used by the staged Boundary Loss configuration over the 100-epoch training budget. The boundary term is introduced after epoch 9, reaches weight 0.5 by epoch 29, reaches weight 0.8 by epoch 44, and remains at 0.8 thereafter.

3.3.3 Generalized surface composite

For the set $\mathcal{V} = \{(b, c) : \sum_i g_{bci} > 0\}$, the implemented surface term is

$$L_{\text{surf}} = \frac{1}{|\mathcal{V}|} \sum_{(b,c)\in\mathcal{V}} \left[1 - \frac{\sum_i [\phi_{bci}(1 - g_{bci} - p_{bci})]^2}{\sum_i \phi_{bci}^2 + \epsilon}\right]. \tag{6}$$

It is combined as $L_{\text{GSL}} = 0.5L_{\text{reg}} + 0.5L_{\text{surf}}$. Empty target channels are excluded only from the surface term; the regional component remains active.

3.3.4 Standalone Boundary DoU

For each sample–channel pair, let $I = \sum_i p_i g_i$, $A = \sum_i p_i^2 + \sum_i g_i^2$, S be the target area/volume, and C_∂ its boundary count. The implementation uses

$$\beta = \min\left\{0.8,\ 2\left(1 - \frac{C_\partial + \epsilon}{S + \epsilon}\right) - 1\right\}, \tag{7}$$

$$L_{\text{DoU}} = \frac{1}{|\mathcal{V}|} \sum_{(b,c)\in\mathcal{V}} \frac{A - 2I + \epsilon}{A - (1 + \beta)I + \epsilon}. \tag{8}$$

Boundary counts use axial neighbors (four in 2D and six in 3D). Crucially, the implemented adaptation excludes target-empty channels. In the independent-sigmoid 2D setting, an absent TC or ET channel therefore receives no direct loss gradient for that sample. This differs from asserting that all background pixels are ignored: background pixels within a target-present channel still contribute through A. Shared features also continue to receive gradients from other valid channels.

Table 2 Main optimization, training, and inference settings.

Setting	2D pipeline	3D pipeline
Optimizer	AdamW	AdamW
Learning rate	10^{-4} ordinary; 10^{-2} KAN	10^{-4} ordinary; 10^{-2} KAN
Weight decay	10^{-4}	10^{-4}
Learning-rate schedule	Cosine annealing; $\eta_{\min} = 10^{-5}$	Polynomial decay; exponent 0.9
Epochs / iterations	100 epochs	100 epochs; 500 iterations/epoch
Batch size	96 slices	4 patches
Early stopping	Patience 10 epochs	Patience 10 epochs
Input size	256×256 axial slices	128^3 patches
Checkpoint selection	Best validation mean Dice	Best validation mean Dice
Inference	Sigmoid threshold 0.5	Sliding window; 50% overlap; Gaussian blending; test-time mirroring

Both pipelines used five loss configurations and three training seeds. Learning rates are reported separately for ordinary and KAN parameters. Detailed preprocessing and augmentation settings are described in the Methods.

4 Experimental Protocol and Evaluation

4.1 Optimization and checkpoint selection

The 2D runs use AdamW, learning rate 10^{-4} for ordinary parameters and 10^{-2} for KAN parameters, weight decay 10^{-4}, and cosine annealing with minimum learning rate 10^{-5}. The maximum budget is 100 epochs. All runs use batch size 96 and early-stopping. Checkpoint selection tracks the highest validation mean WT/TC/ET Dice.

The 3D configurations use AdamW with the same parameter-group learning rates and weight decay, polynomial learning-rate decay with exponent 0.9, batch size 4, and at most 100 epochs. Each epoch contains 500 training iterations. Validation uses 100 iterations every two epochs. The training selects best checkpoints by mean validation WT/TC/ET Dice.

4.2 Inference and postprocessing

The 2D evaluation exports use raw-volume ground truth, probability threshold 0.5 for each region. Slice predictions are restored to subject volumes. Full-volume inference is distinct from the tumor-positive slice distribution used in training and validation; MRI-empty slices may be skipped by the inference implementation, but tumor-negative brain slices are not equivalent to MRI-empty slices.

The 3D prediction configuration specifies sliding-window inference with 128^3 windows, 50% overlap, Gaussian blending, mixed precision, and mirroring along the three spatial axes. Labels are obtained by argmax and converted to WT/TC/ET.

The principal optimization, training, input, checkpoint, and inference settings are summarized in Table 2.

4.3 Metrics and aggregation

For binary prediction P and reference G, overlap is assessed using

$$\mathrm{Dice}(P, G) = \frac{2|P \cap G|}{|P| + |G|}. \tag{9}$$

We report per-region Dice and macro Dice, computed within each scan case across WT, TC, and ET and then averaged over scan cases in the descriptive performance tables.

Surface error is reported as the 95th-percentile Hausdorff distance (HD95), in millimeters, under each pipeline's native convention.

4.4 Analysis strategy

Descriptive performance tables show the mean and sample standard deviation across the three run-level scan-cohort means ($n = 3$, denominator $n - 1$ for the variance). Matched seed-label differences expose whether an average ranking is consistent across runs. All runs, including the severe Boundary DoU failure, remain in the primary analysis.

4.4.1 Paired patient-level comparisons

For each endpoint, we first average a scan's scores across the three seeds, then average multiple scan timepoints within the same patient. This gives one score per patient and loss configuration, without averaging predictions or constructing an ensemble. For patient i and loss ℓ, the paired difference is

$$d_{i\ell} = \frac{1}{|T_i|} \sum_{t \in T_i} \frac{1}{3} \sum_{s \in \{42,123,3407\}} \big(y_{its\ell} - y_{its0}\big), \tag{10}$$

where T_i is the patient's set of scan timepoints and 0 denotes the within-pipeline baseline. Each patient receives equal weight: 126 patients in 2D and 242 in 3D, rather than treating scan timepoints or seed–patient observations as independent replicates.

Two-sided Wilcoxon signed-rank tests compare each of the four nonbaseline configurations with BCE–Dice in 2D or CE–Dice in 3D. We use SciPy 1.17.1, Pratt handling of zero differences, the tie-corrected normal approximation, and continuity correction. Differences are rounded to 12 decimal places only for ranking. The test concerns signed ranks; a location-shift interpretation assumes symmetric paired differences. It is not a test of equality of means. Holm correction is applied to all 32 comparisons within each pipeline (four losses $\times$ WT, TC, ET, and macro values for both Dice and HD95), with adjusted $p < 0.05$ as the decision rule.

Mean paired differences are accompanied by unadjusted 95% percentile bootstrap confidence intervals, using 20,000 patient resamples and fixed random seeds. Bootstrap resampling retains the paired scores and holds the fitted models fixed. These intervals are not simultaneous intervals and need not agree with the signed-rank decisions,

Table 3 Test performance under native pipeline protocols.

Loss configuration	WT Dice	TC Dice	ET Dice	Macro Dice	Macro HD95
2D U-KAN: 126 patients; three runs					
Baseline	0.839 ± 0.079	0.785 ± 0.044	0.776 ± 0.023	0.800 ± 0.045	34.81 ± 16.57
Fixed boundary	0.864 ± 0.036	0.728 ± 0.010	0.766 ± 0.023	0.786 ± 0.021	29.10 ± 8.98
Staged Boundary Loss	0.878 ± 0.012	0.786 ± 0.022	0.778 ± 0.001	0.814 ± 0.004	22.55 ± 4.31
Generalized surface	0.870 ± 0.006	0.761 ± 0.023	0.779 ± 0.007	0.803 ± 0.005	23.09 ± 5.71
Boundary DoU	0.874 ± 0.002	0.471 ± 0.393	0.516 ± 0.437	0.620 ± 0.277	71.09 ± 38.91
3D U-KAN: 250 scan cases from 242 patients; three runs					
Baseline	0.919 ± 0.006	0.913 ± 0.005	0.859 ± 0.002	0.897 ± 0.004	3.74 ± 0.13
Fixed boundary	0.907 ± 0.013	0.891 ± 0.012	0.833 ± 0.017	0.877 ± 0.014	4.16 ± 0.46
Staged Boundary Loss	0.920 ± 0.004	0.911 ± 0.007	0.863 ± 0.003	0.898 ± 0.005	3.64 ± 0.31
Generalized surface	0.916 ± 0.006	0.909 ± 0.006	0.857 ± 0.009	0.894 ± 0.006	4.13 ± 0.54
Boundary DoU	0.917 ± 0.005	0.898 ± 0.008	0.857 ± 0.004	0.891 ± 0.004	4.22 ± 0.31

Values are mean ± sample SD across three runs. Dice is on a 0–1 scale (higher is better); HD95 is in mm (lower is better). The baseline is BCE–Dice in 2D and CE–Dice in 3D. The 2D and 3D panels are not a controlled dimensionality comparison.

which assess a different aspect of the difference distribution. All inference is retrospective and conditional on the evaluated fitted models; it does not estimate uncertainty over new training runs or establish causal loss effects under the protocol deviations. No minimum clinically meaningful difference was prespecified. Nonsignificance is not interpreted as equivalence, and significance is not interpreted as clinical benefit.

5 Results

5.1 2D segmentation

Staged Boundary Loss has the highest mean 2D macro Dice, 0.8139±0.0035, compared with 0.8001 ± 0.0447 for BCE–Dice (Table 3). The average difference is 0.0139, or 1.39 Dice percentage points. Its mean native-protocol HD95 is 22.55 mm, compared with 34.81 mm for the baseline, a difference of −12.26 mm. Generalized surface loss has a similar mean HD95 (23.09 mm), but only a small mean Dice increase over baseline (0.8033 versus 0.8001). Fixed Boundary Loss lowers mean HD95 to 29.10 mm while reducing macro Dice to 0.7859. Thus, improved mean surface distance does not necessarily coincide with improved overlap.

In the 2D pipeline, the apparent advantage of staged Boundary Loss over BCE–Dice is concentrated in whole-tumor segmentation. Mean WT Dice increases from 0.8391 to 0.8775. TC Dice changes from 0.7851 to 0.7859, and ET Dice from 0.7760 to 0.7783. An overall average therefore obscures substantial differences between subregions. Standalone Boundary DoU has the lowest macro Dice (0.6205) and largest run SD (0.2769), driven by a severe TC/ET failure in one run rather than a corresponding loss of WT performance.

5.2 3D segmentation

Staged Boundary Loss also has the highest mean 3D macro Dice, but only narrowly: 0.89794 versus 0.89698 for CE–Dice (Table 3). This is an absolute increase of 0.00096, or approximately 0.10 Dice percentage points. Mean HD95 decreases from 3.736 to

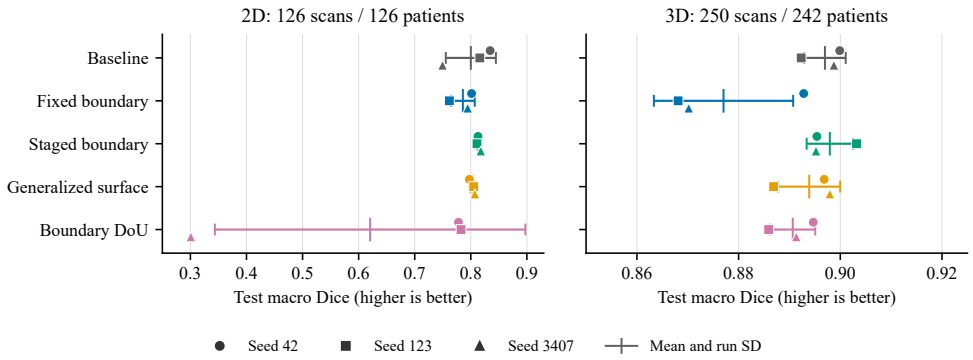


Fig. 4 Every run's scan-weighted test macro Dice. Shapes identify seed labels; central ticks and horizontal whiskers show the mean and sample SD across three runs, not confidence intervals. The panels use different horizontal scales and different test cohorts.

3.645 mm, a difference of 0.091 mm. These descriptive differences do not establish a practically meaningful gain.

The remaining boundary-aware configurations have lower mean macro Dice and higher mean HD95 than the 3D baseline. Generalized surface loss reaches 0.8939 Dice and 4.13 mm HD95, Boundary DoU reaches 0.8906 and 4.22 mm, and fixed Boundary Loss reaches 0.8770 and 4.16 mm. The highest mean ET Dice occurs with the staged Boundary Loss configuration (0.8626), but its TC Dice (0.9113) is slightly below baseline (0.9127). No boundary-aware configuration uniformly dominates the baseline across all regions.

5.3 Run-level consistency and sensitivity

The highest mean does not imply consistent improvement. Staged Boundary Loss exceeds the baseline's macro Dice in only one of three matched seed-label comparisons in each pipeline (Fig. 4, Table 4). In 2D, the positive mean difference is driven by seed 3407, where baseline Dice is 0.7496 and staged Boundary Loss Dice is 0.8179. Baseline runs with seeds 42 and 123 both exceed their corresponding staged Boundary Loss runs. The staged Boundary Loss configuration has lower native HD95 in all three 2D comparisons, although the same protocol limitations still apply.

For 3D, the staged Boundary Loss Dice difference is positive only for seed 123. Two of three staged Boundary Loss runs have lower HD95, and one has higher HD95. The result is better characterized as a small, inconsistent average advantage than as reliable superiority.

5.4 Paired patient-level comparisons

For the 2D fitted models, staged Boundary Loss improves seed-averaged macro Dice by 1.39 percentage points (95% CI 0.98 to 1.84; Holm-adjusted $p < 0.001$), with positive differences in 94 of 126 patients (Table 5). The mean complete-case macro HD95 difference is -12.25 mm (95% CI -13.65 to -10.86; adjusted $p < 0.001$). The regional Dice result is concentrated in WT; the staged Boundary Loss TC and ET

Table 4 Matched seed differences for Staged Boundary Loss.

Pipeline	Seed	Δ Dice (pp)	Δ HD95
2D	42	−2.164	−11.00
2D	123	−0.501	−0.56
2D	3407	+6.826	−25.23
3D	42	−0.449	−0.05
3D	123	+1.089	−0.38
3D	3407	−0.352	+0.16

Differences are Staged Boundary Loss minus baseline. Dice differences are percentage points and HD95 differences are mm. Positive Dice and negative HD95 favor Staged Boundary Loss.

comparisons do not pass the multiplicity correction (Table 6). The corresponding regional HD95 comparisons are reported in Table 7. Fixed Boundary Loss reduces macro Dice, while the generalized surface composite has no significant macro Dice difference; both reduce macro HD95. Boundary DoU reduces macro Dice and increases macro HD95 when all three runs are retained.

In 3D, the patient-weighted staged Boundary Loss macro Dice difference is 0.16 percentage points (95% CI −0.12 to 0.51; adjusted $p = 1.000$), and its macro HD95 difference is −0.17 mm (95% CI −0.47 to 0.11; adjusted $p = 0.446$). Neither comparison provides evidence of a difference after correction. Fixed Boundary Loss has lower macro Dice and a positive signed-rank HD95 shift. No boundary-aware configuration has a significant favorable macro Dice or macro HD95 signed-rank comparison in 3D.

The mean difference and the Wilcoxon signed-rank test summarize different aspects of patient-level change: the mean reflects the magnitudes of the changes, whereas the test evaluates their directions and ranks. For example, 3D Boundary DoU has a negative mean macro Dice difference (−0.56 percentage points), but a positive median difference (0.09 percentage points) and no significant signed-rank difference after correction. Its adverse mean is therefore not a uniform patient-level shift. Likewise, some HD95 mean intervals cross zero despite significant signed-rank tests. We retain both summaries rather than interpret rank-test significance as a confidence statement about the mean.

5.5 Severe oversegmentation with 2D Boundary DoU

The seed-3407 Boundary DoU run achieves WT Dice 0.8736 but TC Dice 0.0178 and ET Dice 0.0110. Mean TC precision is 0.0091 with sensitivity 0.9139; mean ET precision is 0.0056 with sensitivity 0.8739. This combination indicates extensive false-positive prediction, rather than simple failure to detect foreground. We refer to this as a region-specific oversegmentation failure, not complete network collapse. The other two 2D Boundary DoU runs have macro Dice 0.7781 and 0.7826, so the severe failure is not shared by all seeds.

Table 5 Patient-level macro comparisons.

Pipeline	Loss vs. baseline	Δ Dice [95% CI] (pp)	p_H	n_{HD}	Δ HD95 [95% CI] (mm)	p_H
2D	Fixed boundary	−1.42 [-1.97, -0.85]	< 0.001	121	−5.59 [-7.09, -4.03]	< 0.001
2D	Staged Boundary Loss	+1.39 [+0.98, +1.84]	< 0.001	121	−12.25 [-13.65, -10.86]	< 0.001
2D	Generalized surface	+0.32 [-0.23, +0.90]	1.000	121	−11.71 [-13.15, -10.32]	< 0.001
2D	Boundary DoU	−17.96 [-18.80, -17.04]	< 0.001	121	+36.55 [+34.54, +38.55]	< 0.001
3D	Fixed boundary	−1.94 [-2.35, -1.53]	< 0.001	242	+0.35 [-0.04, +0.69]	< 0.001
3D	Staged Boundary Loss	+0.16 [-0.12, +0.51]	1.000	242	−0.17 [-0.47, +0.11]	0.446
3D	Generalized surface	−0.33 [-0.75, +0.05]	0.276	242	+0.33 [-0.09, +0.77]	0.029
3D	Boundary DoU	−0.56 [-1.01, -0.15]	1.000	242	+0.46 [+0.21, +0.74]	0.925

Scores are averaged over seeds and within-patient timepoints. Differences are loss minus baseline; positive Dice and negative HD95 favor the alternative. Dice uses 126 patients in 2D and 242 in 3D; n_{HD} gives complete-case HD95 counts. Confidence intervals are unadjusted patient-bootstrap intervals for the mean difference. p_H denotes two-sided Wilcoxon p-values Holm-adjusted across all 32 comparisons within each pipeline.

Table 6 Regional patient-level Dice comparisons.

Pipeline	Loss vs. baseline	Δ WT	p_H	Δ TC	p_H	Δ ET	p_H
2D	Fixed boundary	+2.45	< 0.001	−5.72	< 0.001	−0.98	< 0.001
2D	Staged Boundary Loss	+3.84	< 0.001	+0.08	1.000	+0.24	1.000
2D	Generalized surface	+3.06	< 0.001	−2.43	< 0.001	+0.35	0.168
2D	Boundary DoU	+3.52	< 0.001	−31.39	< 0.001	−26.01	< 0.001
3D	Fixed boundary	−1.23	< 0.001	−2.06	< 0.001	−2.53	< 0.001
3D	Staged Boundary Loss	+0.09	< 0.001	−0.05	1.000	+0.45	0.578
3D	Generalized surface	−0.33	1.000	−0.39	0.353	−0.25	0.005
3D	Boundary DoU	−0.16	0.013	−1.39	0.569	−0.12	1.000

Values are mean differences in percentage points, with scores averaged across seeds and repeated timepoints. Positive mean differences favor the alternative, but the signed-rank test does not test the mean. All regional and macro Dice/HD95 comparisons share the same 32-test correction family within a pipeline.

5.6 Qualitative comparison

To complement the aggregate metrics, Figs. 5 and 6 show matched axial slices from the 2D and 3D test volumes, respectively. Before inspecting the images, test cases in each pipeline were ranked using macro Dice from the seed-123 baseline model. Two cases were then selected independently within each pipeline: the case with a score closest to the 25th percentile and the case with a score closest to the median. Ties were resolved by case ID. The same case and slice are shown for every loss configuration within that pipeline. This is a retrospective, reproducible illustration rule rather than a representative sample or evidence of population-wide superiority.

The displayed slice is the axial slice with the largest WT reference area. T1c is shown for anatomical contrast, while FLAIR is used behind filled exclusive-class overlays. The colors denote edema (blue), necrotic/non-enhancing core (orange), and enhancing tumor (pink).

Across these examples, all configurations localize the gross tumor in the same anatomical region. The visible differences are concentrated at the outer edema boundary and at the internal necrotic/enhancing class interfaces, rather than in the location of the lesion. The 2D lower-quartile case makes the larger spread between methods visible, whereas the 2D median and both 3D examples show more similar class overlays.

Table 7 Regional patient-level HD95 comparisons.

Pipeline	Loss vs. baseline	Δ WT	p_H	Δ TC	p_H	Δ ET	p_H
2D	Fixed boundary	−2.04	0.013	−7.62	< 0.001	−7.49	< 0.001
2D	Staged Boundary Loss	−8.87	< 0.001	−16.27	< 0.001	−11.64	< 0.001
2D	Generalized surface	−11.48	< 0.001	−12.44	< 0.001	−11.26	< 0.001
2D	Boundary DoU	−15.90	< 0.001	+60.50	< 0.001	+64.23	< 0.001
3D	Fixed boundary	+0.41	< 0.001	+0.29	< 0.001	+0.35	< 0.001
3D	Staged Boundary Loss	−0.28	0.080	−0.05	1.000	−0.18	1.000
3D	Generalized surface	+0.81	0.037	+0.09	1.000	+0.08	0.136
3D	Boundary DoU	+0.77	1.000	+0.41	1.000	+0.19	1.000

Values are mean HD95 differences in mm and Holm-adjusted signed-rank p-values. Negative differences favor the alternative. Complete-case counts are 126 WT, 125 TC, and 121 ET in 2D, and 242 for every region in 3D. Empty-mask conventions remain pipeline-specific. The correction family is the same as in Tables 5 and 6.

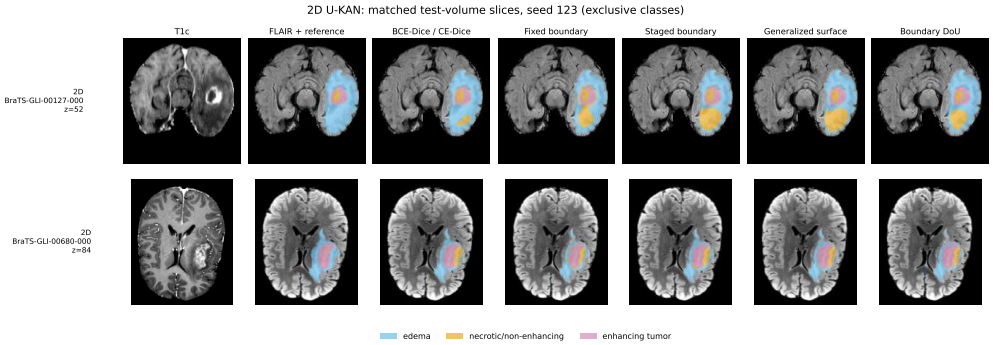


Fig. 5 Qualitative comparison of the five 2D U-KAN loss configurations on two held-out test volumes, all using seed 123. Cases are selected by the seed-123 baseline macro Dice (lower-quartile and median cases within the 2D pipeline). Each row uses one identical axial slice selected by the largest reference WT area. Columns show T1c, FLAIR with the reference overlay, and predictions from BCE–Dice, fixed Boundary Loss, staged Boundary Loss, Generalized Surface Loss, and Boundary DoU. Filled colors denote edema (blue), necrotic/non-enhancing core (orange), and enhancing tumor (pink).

These observations are illustrative and should not be generalized beyond the selected cases; they also do not replace the quantitative per-region and run-level analyses.

6 Discussion

6.1 Do boundary-aware objectives improve this backbone?

The results do not support a general claim that adding a boundary-aware objective meaningfully improves U-KAN segmentation. In the 2D pipeline, staged Boundary Loss and the generalized surface composite are promising for mean surface distance, while the staged Boundary Loss configuration also has the highest average overlap. Patient-level tests support differences for these particular fitted models, but that overlap ranking remains sensitive to one low-performing baseline run. Patient-level



Fig. 6 Qualitative comparison of the five 3D U-KAN loss configurations on two held-out test volumes, all using seed 123. Cases are selected by the seed-123 baseline macro Dice (lower-quartile and median cases within the 3D pipeline). Each row uses one identical axial slice selected by the largest reference WT area. Columns show T1c, FLAIR with the reference overlay, and predictions from CE–Dice, fixed Boundary Loss, staged Boundary Loss, Generalized Surface Loss, and Boundary DoU. Filled colors denote edema (blue), necrotic/non-enhancing core (orange), and enhancing tumor (pink).

significance after seed averaging is compatible with inconsistent gains between individual training runs: these analyses address different sources of variation. In the 3D pipeline, the regional baseline remains competitive; staged Boundary Loss has no significant favorable macro Dice or HD95 comparison after correction, and its small descriptive advantage is not consistent across seed labels.

The contrast between fixed and staged Boundary Loss suggests that the weighting schedule deserves attention alongside the choice of boundary formulation. Beginning with a regional objective may aid optimization before emphasizing spatial errors.

6.2 Interpreting the Boundary DoU failure

The failure is compatible with inadequate direct supervision for absent channels in the implemented 2D adaptation. A slice can contain WT while containing neither TC nor ET. With valid-channel filtering, its WT loss remains active, but its TC and ET loss terms are omitted. Seeing such a slice therefore does not itself provide a direct penalty for predicting the absent regions. The independent sigmoid outputs provide no softmax competition to suppress those channels. In contrast, multiclass softmax probabilities are coupled, so loss on another valid class can affect an absent class's logit.

This explanation has limits. On slices with TC or ET present, background pixels in those channels are still penalized, and shared-feature learning can influence predictions on other slices. Excluding absent channels therefore does not make failure inevitable, nor does it explain by itself why one seed failed and two did not. Tumor-positive slice sampling and target-empty validation exclusion may also contribute to a mismatch between model selection and deployment on whole volumes.

The evidence does not establish that predictions are accurate whenever tumor is present and fail only on tumor-free slices. That claim requires slice-stratified evaluation, including WT-only, TC-without-ET, and tumor-negative slices. Such an analysis is not available in the present aggregate exports. Likewise, retaining standalone DoU

is a valid experimental choice, but a failure in this adapted setting is not evidence that the original formulation fails universally. We report the observed failure and its implementation context rather than silently discarding the run or modifying its loss after observing the test outcome.

6.3 Limitations

The study has several limitations. First, the experiments use one dataset and one architectural family. The findings consequently describe the behavior of the implemented loss configurations with U-KAN on BraTS 2023 and may not generalize to other architectures, imaging datasets, or clinical populations. Second, the 2D and 3D pipelines differ in cohort composition, target representation, preprocessing, sampling, and evaluation conventions. Their numerical results therefore support only within-pipeline comparisons, not a controlled comparison between 2D and 3D segmentation.

7 Conclusion

This study examined whether boundary-aware loss functions improve U-KAN brain tumor segmentation by comparing five loss configurations in separate 2D and 3D pipelines across three training runs. Overall, the results do not demonstrate a consistent advantage for boundary-aware training over the corresponding region-based baselines. Staged Boundary Loss achieved the highest mean macro Dice in both pipelines and produced favorable boundary-distance results in the 2D experiments, but these improvements were not consistently reproduced across runs and were small in the 3D pipeline. The remaining boundary-aware configurations generally matched or underperformed the baselines, while standalone Boundary DoU showed substantial instability in 2D.

These findings indicate that introducing boundary information into the training objective does not by itself guarantee more accurate or reliable segmentation. The formulation and weighting schedule of the loss, its treatment of absent tumor regions, and its interaction with the segmentation pipeline all appear to influence the outcome. For U-KAN segmentation on BraTS 2023, conventional region-based losses therefore remain strong reference objectives.

Building on these findings, future work could explore adaptive and region-specific boundary weighting to balance overlap and boundary accuracy across tumor subregions. The staged Boundary Loss configuration provides a promising starting point for investigating training schedules that adjust boundary emphasis as learning progresses. Extending this evaluation to additional imaging datasets and segmentation architectures would further characterize the settings in which boundary-aware objectives offer the greatest benefit.

List of abbreviations

ASNR American Society of Neuroradiology
BCE binary cross-entropy
CE cross-entropy

ET enhancing tumor
FLAIR fluid-attenuated inversion recovery
GSL Generalized Surface Loss
HD95 95th-percentile Hausdorff distance
KAN Kolmogorov–Arnold Network
MRI magnetic resonance imaging
NCR/NET necrotic/non-enhancing tumor
TC tumor core
T1c contrast-enhanced T1-weighted MRI
T1n native T1-weighted MRI
T2-FLAIR T2-weighted fluid-attenuated inversion recovery MRI
T2w T2-weighted MRI
WT whole tumor

Acknowledgements

Not applicable.

Author contributions

MH contributed to methodology, investigation, software, data curation, and writing the original draft. AD and AE contributed to methodology, conceptualization, writing, review and editing, and supervision. All authors read and approved the final manuscript.

Funding

No specific funding was received for this study.

Data availability

The BraTS 2023 dataset analyzed in this study is publicly available through Synapse at https://www.synapse.org/Synapse:syn51156910/wiki/6212822 and is cited through the corresponding challenge data record [23]. The source data were not generated by the authors.

Declarations

Ethics approval and consent to participate

This study used publicly available anonymized neuroimaging data from the BraTS 2023 challenge dataset. No additional ethics approval was required as the dataset was previously approved by the institutional review boards of the contributing institutions and all patient identifiers were removed prior to public release. Consent to participate was obtained by the original data collectors at the contributing institutions.

Consent for publication

Not applicable. This study uses publicly available anonymized data with no identifiable patient information.

Competing interests

The authors declare that they have no competing interests.

References

[1] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A Survey on Deep Learning in Medical Image Analysis. Medical Image Analysis. 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.

[2] Davis ME. Epidemiology and Overview of Gliomas. Seminars in Oncology Nursing. 2018;34(5):420–429. Adult Gliomas. https://doi.org/10.1016/j.soncn.2018.10.001.

[3] Baid U, et al. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. arXiv preprint arXiv:210702314. 2021;arXiv:2107.02314. [eess.IV].

[4] Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging. 2015;34(10):1993–2024. https://doi.org/10.1109/TMI.2014.2377694.

[5] Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, et al. Advancing The Cancer Genome Atlas Glioma MRI Collections with Expert Segmentation Labels and Radiomic Features. Scientific Data. 2017;4:170117. https://doi.org/10.1038/sdata.2017.117.

[6] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing; 2015. p. 234–241. Available from: http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[7] Cicek O, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. vol. 9901 of Lecture Notes in Computer Science. Springer International Publishing; 2016. p. 424–432. Available from: https://doi.org/10.1007/978-3-319-46723-8_49.

[8] Milletari F, Navab N, Ahmadi SA. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). IEEE; 2016. p. 565–571. Available from: https://doi.org/10.1109/3DV.2016.79.

[9] Li C, Liu X, Li W, Wang C, Liu H, Liu Y, et al. U-KAN Makes Strong Backbone for Medical Image Segmentation and Generation. Proceedings of the AAAI Conference on Artificial Intelligence. 2025 Apr;39(5):4652–4660. https://doi.org/10.1609/aaai.v39i5.32491.

[10] Kervadec H, Bouchtiba J, Desrosiers C, Granger E, Dolz J, Ben Ayed I. Boundary loss for highly unbalanced segmentation. In: Cardoso MJ, Feragen A, Glocker B, Konukoglu E, Oguz I, Unal G, et al., editors. Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning. vol. 102 of Proceedings of Machine Learning Research. PMLR; 2019. p. 285–296. Available from: https://proceedings.mlr.press/v102/kervadec19a.html.

[11] Sun F, Luo Z, Li S. Boundary Difference over Union Loss for Medical Image Segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Springer Nature Switzerland; 2023. p. 292–301. Available from: http://dx.doi.org/10.1007/978-3-031-43901-8_28.

[12] Ma J, Chen J, Ng M, Huang R, Li Y, Li C, et al. Loss Odyssey in Medical Image Segmentation. Medical Image Analysis. 2021;71:102035. https://doi.org/10.1016/j.media.2021.102035.

[13] Isensee F, Wald T, Ulrich C, Baumgartner M, Roy S, Maier-Hein K, et al. nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Springer Nature Switzerland; 2024. p. 488–498. Available from: https://doi.org/10.1007/978-3-031-72114-4_47.

[14] Cheng CW, Zhao Y, Cheng Y, Montoya-Zegarra JA, Schönlieb CB, Aviles-Rivero AI. Implicit U-KAN2.0: Dynamic, Efficient and Interpretable Medical Image Segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. 15970. Springer Nature Switzerland; 2025. p. 304–314. Available from: https://doi.org/10.1007/978-3-032-05141-7_30.

[15] Ali U, Qi J, Rashid A, Musaddiq MH. U-KAN for Multi-Nuclei Segmentation Using an Adaptive Sliding Window Approach. International Journal of Imaging Systems and Technology. 2026;36(1):e70283. https://doi.org/10.1002/ima.70283.

[16] Tang T, Chen Y, Shu H.: 3D U-KAN Implementation for Multi-Modal MRI Brain Tumor Segmentation. Available from: https://arxiv.org/abs/2408.00273.

[17] Chen Y, Tang T, Kim T, Shu H. UKAN-EP: enhancing U-KAN with efficient attention and pyramid aggregation for 3D multi-modal MRI brain tumor segmentation. BMC Medical Imaging. 2025 Nov;25(1). https://doi.org/10.1186/s12880-025-02053-w.

[18] Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images.

In: BrainLes 2021. Springer Nature; 2022. p. 272–284. Available from: https://doi.org/10.1007/978-3-031-08999-2_22.

[19] Li Z, Kang Z, Wang Y, Kang X, Yang S, Zhao Q. ResKANNet: A Residual Kolmogorov-Arnold Network with Multi-Scale Attention for Brain Tumor Segmentation. Neurocomputing. 2026;664:132062. https://doi.org/10.1016/j.neucom.2025.132062.

[20] Celaya A, Riviere B, Fuentes D.: A Generalized Surface Loss for Reducing the Hausdorff Distance in Medical Imaging Segmentation. arXiv. Available from: https://arxiv.org/abs/2302.03868.

[21] Kong D, Liu X, Wang Y, Li D, Xue J. 3D hierarchical dual-attention fully convolutional networks with hybrid losses for diverse glioma segmentation. Knowledge-Based Systems. 2022;237:107692. https://doi.org/10.1016/j.knosys.2021.107692.

[22] Yu X, Teng L, Zhang D, Zheng J, Chen H. Attention correction feature and boundary constraint knowledge distillation for efficient 3D medical image segmentation. Expert Systems with Applications. 2025;262:125670. https://doi.org/10.1016/j.eswa.2024.125670.

[23] Bakas S, Ujjwal Baid, Farahani K, Albrecht J, Eddy J, Bergquist T, et al.: The International Brain Tumor Segmentation (BraTS) Cluster of Challenges. Zenodo. Available from: https://zenodo.org/record/7837974.