International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

BW-Dice: A Boundary-Weighted Dice Loss for Brain Tumor Sub-Region Segmentation

S. U. Ravi Kumar Chavali[1], P. V. Kumar[2]

[1]Research Scholar, Rayalaseema University, Kurnool, Andhra Pradesh, India
[2]School of Engineering, Anurag University, Hyderabad, Telangana, India

Abstract

Brain tumor segmentation from multi-modal MRI is dominated by U-Net family architectures paired with Dice-based losses. In prior matched-condition work on BraTS 2020, it was established that the top five configurations of decoder architecture and standard loss function are statistically indistinguishable on aggregate mean Dice, and three distinct failure patterns were identified — edema over-segmentation, NCR/edema boundary confusion, and missed multi-focal disease. Motivated by the NCR/edema boundary failure mode, a Boundary-Weighted Dice (BW-Dice) loss is proposed, a generalization of soft multiclass Dice that applies per-pixel Gaussian weighting near ground-truth class boundaries, with a class-specific multiplier emphasizing the NCR-edema interface. BW-Dice reduces to standard multiclass Dice when the boundary emphasis strength $\lambda = 0$, providing a principled generalization rather than a competing alternative. BW-Dice is evaluated on the UNet++ architecture under identical training conditions to the prior baseline, spanning four hyperparameter configurations and three random seeds, with a controlled ablation removing the class-specific term. Across all configurations, BW-Dice significantly improves Tumor Core segmentation ($\Delta = +0.012$, $p < 10^{-3}$) while significantly degrading Enhancing Tumor segmentation ($\Delta = -0.018$, $p < 10^{-3}$), with aggregate mean Dice statistically indistinguishable from vanilla Dice ($\Delta = -0.001$, $p = 0.97$). The controlled ablation demonstrates that this regional pattern persists without class-specific weighting, indicating it is a general property of boundary-emphasis losses rather than an artifact of the design choice. Class-specific NCR-edema emphasis significantly reduces the magnitude of ET degradation but does not significantly change TC improvement. Implications for future loss-function design in medical image segmentation with hierarchical class structure are discussed.

Keywords: Brain Tumor Segmentation, BraTS 2020, Boundary-Weighted Loss, Dice Loss, U-Net, Medical Image Analysis, Statistical Significance Testing, Class Imbalance

1. Introduction

Accurate delineation of glioma sub-regions from multi-modal magnetic resonance imaging (MRI) is essential for diagnosis, radiotherapy planning, and treatment monitoring. The BraTS challenge series [2, 3, 4] has standardized this problem into three hierarchical sub-regions of clinical interest: Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET), evaluated by Dice similarity coefficient and 95th-percentile Hausdorff distance (HD95).

U-Net family architectures paired with Dice-based losses dominate the field [5, 6, 7, 9]. Architectural refinements — attention gates [8], nested skip pathways [7], and transformer encoders [20, 21] — have been reported to improve segmentation accuracy on BraTS. Cross-study comparisons are however complicated by unreported variation in preprocessing, augmentation, and training schedule, making the marginal contribution of any individual innovation difficult to isolate [19].

Motivation from prior work. In prior matched-condition analysis on BraTS 2020 [1], 16 configurations spanning four U-Net decoders (U-Net, UNet++, MAnet, LinkNet) with four segmentation losses (Dice, Focal, Tversky, Focal-Tversky) were compared under strictly identical training conditions. Two findings motivate the present work. First, the top five configurations were statistically indistinguishable on aggregate mean Dice (all $p > 0.32$, 95% CIs spanning zero), with UNet++ + Dice achieving the highest mean at 0.826. Second, qualitative failure-mode analysis identified three distinct failure patterns on low-performing cases: edema over-segmentation in atypical presentations, NCR/edema boundary confusion, and missed multi-focal disease. These findings suggest that further gains within the standard loss family are unlikely under matched conditions, and that targeted interventions on specific failure modes may be a more productive direction.

The present work. Motivated by the NCR-edema boundary confusion failure pattern, a Boundary-Weighted Dice (BW-Dice) loss is proposed, a generalization of soft multiclass Dice loss that applies per-pixel Gaussian weighting near ground-truth class boundaries, with a class-specific multiplier emphasizing the NCR-edema interface. BW-Dice reduces to standard multiclass Dice when the boundary emphasis strength $\lambda = 0$, providing a principled generalization rather than a competing alternative.

BW-Dice is evaluated on the UNet++ architecture under identical training conditions to the prior baseline, spanning four hyperparameter configurations and three random seeds. A controlled ablation isolates the contribution of class-specificity. Across all configurations, BW-Dice significantly improves TC segmentation ($\Delta = +0.012$, $p < 10^{-3}$) while significantly degrading ET segmentation ($\Delta = -0.018$, $p < 10^{-3}$), with aggregate mean Dice statistically indistinguishable from vanilla Dice ($\Delta = -0.001$, $p = 0.97$). Class-specific NCR-edema emphasis significantly reduces the magnitude of ET degradation but does not change TC improvement.

The contributions are as follows.

(C1) BW-Dice, a class-aware boundary-weighted Dice loss targeting the NCR-edema failure mode identified in [1].

(C2) Empirical characterization of the per-region effect of BW-Dice under matched conditions with multi-seed replication and paired bootstrap significance testing.

(C3) Controlled ablation demonstrating that the regional effect pattern is inherent to boundary emphasis rather than an artifact of class-specific weighting.

(C4) Analysis of implications for loss-function design on tasks with hierarchical class structure.

Section 2 reviews related loss functions. Section 3 defines BW-Dice and its gradient. Section 4 describes the experimental protocol and results. Section 5 discusses implications and limitations. Section 6 concludes.

2. Related Work

Overlap-based segmentation losses. The Dice similarity coefficient, adapted as a differentiable loss by Milletari et al. [6], is the standard objective for medical image segmentation. Tversky loss [11] generalizes Dice via asymmetric weights α, β on false positives and false negatives, with Dice recovered at $\alpha = \beta = 0.5$. Focal-Tversky [12] adds an image-level focusing exponent γ. These form a continuous family (Dice $\subset$ Tversky $\subset$ Focal-Tversky) with progressively more hyperparameters. The prior work [1] compared all four members with four U-Net variants under matched conditions and found no configuration statistically significantly better than UNet++ + Dice on aggregate mean Dice.

Distribution-matching losses. Focal loss [10], introduced for dense object detection, modulates cross-entropy by $(1 - p_t)^\gamma$ to concentrate gradient on hard-to-classify pixels. In segmentation, hard pixels correlate with — but are not equivalent to — class boundaries. Composite losses combining Dice with cross-entropy or Focal have shown consistent aggregate gains [13, 17], though rarely subjected to matched-condition significance testing.

Boundary-emphasis losses. Explicit boundary weighting dates to the original U-Net [5], which used a pre-computed weight map based on distance to cell boundaries. Subsequent boundary-focused approaches include the boundary loss of Kervadec et al. [14] operating on level-set representations, Hausdorff-distance-based losses [15] directly optimizing HD95 approximations, and active-contour losses [16] applying level-set regularization to predicted boundaries. BW-Dice differs from these in three respects: (i) it operates on the standard soft Dice formulation rather than requiring new mathematical structures; (ii) it preserves the gradient direction of Dice at every pixel while altering only gradient magnitude near boundaries; and (iii) it incorporates class-specific boundary weighting driven by a specific failure-mode analysis rather than treating all boundaries uniformly.

Region-hierarchical losses for BraTS. The BraTS evaluation metric operates on hierarchical regions (WT ⊇ TC ⊇ ET) rather than raw class labels. The BraTS variant of nnU-Net [18] trains directly on these derived regions, winning BraTS 2020. The present approach retains raw four-class training to preserve comparability with methods trained on raw classes, and to avoid conflating loss-design innovation with evaluation-metric mimicry. Combining boundary weighting with region-hierarchical training is a natural direction for future work.

3. Methodology

3.1 Boundary-Weighted Dice Loss

Let $\hat{y}_c(p) \in [0, 1]$ denote the predicted probability that pixel p belongs to class c, obtained via softmax over class logits, and let $y_c(p) \in \{0, 1\}$ denote the corresponding one-hot ground-truth label. Standard multiclass Dice loss on the foreground classes C = {NCR, edema, ET} is given by Equation (1), where ε prevents division by zero. Background (class 0) is excluded from the average following common practice, so as not to dilute gradient signal by the dominant background class.

$$L_{\text{Dice}} = 1 - \frac{1}{|C|}\sum_{c}\frac{2\sum_{p}\hat{y}_c(p)\,y_c(p) + \varepsilon}{\sum_{p}\hat{y}_c(p) + \sum_{p}y_c(p) + \varepsilon} \quad (1)$$

Boundary-Weighted Dice (BW-Dice) generalizes this by introducing a per-pixel spatial weight $w(p) \geq 1$ that multiplicatively scales each pixel's contribution to both numerator and denominator, as given in Equation (2). Setting $w(p) = 1$ for all p recovers vanilla Dice, so BW-Dice is a strict generalization.

$$L_{\text{BW-Dice}} = 1 - \frac{1}{|C|}\sum_{c}\frac{2\sum_{p}w(p)\,\hat{y}_c(p)\,y_c(p) + \varepsilon}{\sum_{p}w(p)\,\hat{y}_c(p) + \sum_{p}w(p)\,y_c(p) + \varepsilon} \quad (2)$$

3.2 Weight Map Construction

Given a ground-truth label map, a single distance transform to the nearest class-to-class boundary is computed, and pixels whose nearest boundary is of the NCR-edema type are identified. The weight is defined in Equation (3), where d(p) is the Euclidean distance in pixels from p to the nearest class boundary, $\lambda \geq 0$ controls boundary emphasis strength, $\sigma > 0$ controls Gaussian spread, and $m_k(p)$ is a class-specific multiplier: $m_k(p) = m_{NE}$ if the nearest boundary is of NCR-edema type, else $m_k(p) = 1$. The weight is bounded above by $1 + \lambda \cdot m_{NE}$ at NCR-edema boundary pixels, decaying with Gaussian shape into region interiors. Weight maps are pre-computed per slice during data loading using the Euclidean distance transform of SciPy, taking approximately 1 ms per 128 × 128 slice.

$$w(p) = 1 + \lambda \cdot m_k(p) \cdot \exp\left[-\frac{d(p)^2}{2\sigma^2}\right] \quad (3)$$

3.3 Gradient Behavior

Differentiating with respect to a single prediction $\hat{y}_c(q)$ and denoting $N_c = \Sigma_p\, w(p)\, \hat{y}_c(p)\, y_c(p)$ and $D_c = \Sigma_p\, w(p)\, [\hat{y}_c(p) + y_c(p)]$, Equation (4) is obtained.

$$\frac{\partial L_c}{\partial \hat{y}_c(q)} = -\, w(q)\, \frac{2\, y_c(q)\,(D_c + \varepsilon) - (2N_c + \varepsilon)}{(D_c + \varepsilon)^2} \quad (4)$$

Two properties follow. First, the gradient scales linearly with w(q): boundary pixels contribute proportionally more to parameter updates than interior pixels. Second, the sign and structure of the gradient are identical to vanilla Dice; boundary weighting alters only the magnitude of gradients, not their direction. In this sense BW-Dice separates what the model should learn (unchanged from Dice) from how strongly each pixel contributes to that learning (scaled by w(p)).

3.4 Hyperparameter Choices

The value $m_{NE} = 2.0$ is fixed throughout, providing a 2× multiplicative boost to NCR-edema boundaries relative to other boundaries. This choice is motivated by the specific NCR-edema failure mode identified in [1] and held constant across all experiments. The two remaining hyperparameters are swept: $\lambda \in \{1, 2, 3\}$ (boundary emphasis strength) and $\sigma \in \{3, 5\}$ (Gaussian spread in pixels). At 128 × 128 resolution, $\sigma = 3$ corresponds to an approximately 6-pixel-wide boundary emphasis band on each side of a boundary; $\sigma = 5$ extends this to approximately 10 pixels.

3.5 Experimental Protocol

Dataset and split. BraTS 2020 [2, 3, 4], 369 cases split 295 / 36 / 38 for training / validation / test with seed 42, matching [1]. After preprocessing (z-score normalization per modality, center crop 240 → 192, resize to 128 × 128, tumor-bearing slice extraction with ≥ 100 tumor voxels), 2,242 test slices remain.

Architecture. UNet++ with ResNet-34 encoder pre-trained on ImageNet, via segmentation-models-pytorch [22]. UNet++ is chosen as the primary architecture based on its top-performing status in the matched-condition analysis [1].

Training. All conditions match [1]: Adam optimizer (lr 10^{-4}, cosine annealing), up to 25 epochs with early stopping (patience 5) on validation mean Dice, batch size 16, random flip augmentation, FP16 mixed precision. Random seeds are explicitly controlled for multi-seed replication.

Evaluation. Per-slice Dice for WT, TC, ET on all 2,242 test slices, matching the evaluation protocol of [1] for direct comparability. HD95 is computed per-slice on slices with non-empty prediction and ground truth. Statistical significance is assessed via paired bootstrap (10,000 iterations) on per-slice Dice, reporting mean difference, 95% CI, and two-tailed p-value. Bootstrap comparisons are made against the UNet++ + Dice baseline from [1], re-evaluated in the same session to eliminate implementation drift.

Experimental matrix. Four hyperparameter configurations from the sweep $(\lambda, \sigma) \in \{(1,3), (2,3), (3,3), (2,5)\}$ are reported, three seeded replications (42, 123, 2024) of the best-performing configuration, and one ablation with $m_{NE} = 1$ (uniform boundary weighting) at the best-performing configuration.

4. Results

4.1 Reproduction of the Vanilla Dice Baseline

To ensure comparability, the prior UNet++ + Dice checkpoint is first re-evaluated under the same per-slice evaluation protocol used throughout this work. This reproduction closely matches the reported values in [1]: mean test Dice 0.826 vs. reported 0.826 ($\Delta = 0.0001$), with per-region values WT = 0.851 vs. 0.851, TC = 0.799 vs. 0.799, ET = 0.828 vs. 0.828. All subsequent comparisons use this re-evaluated baseline to eliminate implementation drift.

4.2 Hyperparameter Sweep

Table 1 reports test-set results across four BW-Dice hyperparameter configurations, together with the vanilla Dice baseline. All BW-Dice configurations produce positive per-slice TC deltas and negative ET deltas relative to the baseline. The configuration ($\lambda = 2$, $\sigma = 5$) achieves the highest test mean Dice (0.833) and the largest TC gain (+0.028), and is designated the primary configuration for subsequent multi-seed analysis. It is noted that this ranking is based on a single training run per configuration and that the differences between BW-Dice configurations are of the same order as inter-seed variance (Section 4.3).

Table 1: Test-Set Per-Slice Dice Across BW-Dice Hyperparameter Configurations. Numbers Are Means Over 2,242 Test Slices. All BW-Dice Configurations Use $m_{ne} = 2.0$. Best Row in Bold.

Configuration	λ	σ	Mean Dice	WT	TC	ET
UNet++ + Dice (baseline)	—	—	0.826	0.851	0.799	0.828
BW-Dice	1	3	0.828	0.858	0.808	0.818
BW-Dice	2	3	0.826	0.855	0.809	0.814
BW-Dice	3	3	0.824	0.854	0.803	0.815
BW-Dice	**2**	**5**	**0.833**	**0.853**	**0.827**	**0.820**

Paired bootstrap testing (Table 2) shows that only the (2, 5) configuration achieves a statistically significant aggregate mean Dice improvement ($\Delta = +0.007$, $p = 0.003$). All configurations show significant TC improvement, and all show significant ET degradation. WT changes are marginal to non-significant across configurations.

Table 2: Paired Bootstrap (10,000 Iterations) vs. UNet++ + Dice Baseline. A Check Mark Indicates the 95% CI Excludes Zero.

Configuration	Δ Mean [95% CI]	Δ WT	Δ TC	Δ ET
BW-Dice (1, 3)	+0.002 [−0.003, +0.008]	+0.006 ✓	+0.009	−0.009 ✓
BW-Dice (2, 3)	−0.000 [−0.005, +0.005]	+0.004	+0.010 ✓	−0.014 ✓
BW-Dice (3, 3)	−0.002 [−0.007, +0.004]	+0.003	+0.004	−0.013 ✓
BW-Dice (2, 5)	**+0.007 [+0.002, +0.013] ✓**	**+0.002**	**+0.028 ✓**	**−0.008 ✓**

4.3 Multi-Seed Replication

To assess robustness of the (2, 5) configuration to random initialization, it is retrained with three seeds (42, 123, 2024). Test-set results (Table 3) show that the aggregate mean Dice improvement observed in Section 4.2 does not replicate across seeds: multi-seed test mean Dice is 0.825 ± 0.003, statistically indistinguishable from the vanilla Dice baseline (0.826, $\Delta = -0.001$, signal-to-noise ratio 0.27).

However, the per-region trade-off pattern is highly reproducible. TC improvement is significant across all three seeds ($\Delta = +0.012$, S/N ratio 3.83), and ET degradation is significant across all three ($\Delta = -0.018$, S/N ratio 2.74). WT shows a modest improvement (S/N ratio 2.21). The single-run aggregate gain in Section 4.2 is interpreted as within-seed variance; the reproducible finding across seeds is the regional trade-off.

Table 3: Multi-Seed Test Results for BW-Dice ($\lambda = 2$, $\sigma = 5$). S/N Is Signal-to-Noise Ratio; p-Values Are From Paired Bootstrap (10,000 Iterations) vs. UNet++ + Dice Baseline.

Seed	Mean Dice	WT	TC	ET
42	0.829	0.857	0.811	0.819
123	0.823	0.856	0.807	0.805
2024	0.824	0.853	0.814	0.805
Mean ± SD	**0.825 ± 0.003**	**0.855 ± 0.002**	**0.811 ± 0.003**	**0.810 ± 0.007**
Δ vs. baseline	−0.001	+0.004	+0.012	−0.018
S/N ratio	0.27	2.21	3.83	2.74
p-value	0.97	0.03	$< 10^{-3}$	$< 10^{-3}$

4.4 Ablation: Removal of Class-Specific Weighting

To isolate the contribution of the NCR-edema class-specific multiplier, the (2, 5) configuration is retrained with $m_{NE} = 1$, applying uniform boundary weighting to all class-to-class boundaries. Table 4 shows both the ablation configuration and the class-specific counterpart at seed 42.

Uniform boundary weighting still produces the qualitative TC+/ET− pattern: WT +0.005 (significant), TC +0.008 (marginal, $p = 0.10$), ET −0.013 (significant). Adding the NCR-edema multiplier significantly reduces ET degradation ($\Delta = +0.004$ over uniform, $p = 0.023$) but does not significantly change TC ($\Delta = +0.004$, $p = 0.32$) or WT ($\Delta = +0.001$, $p = 0.47$). It is concluded that the regional trade-off is a general property of boundary-emphasis losses on hierarchical class structures, and that class-specific NCR-edema weighting acts primarily by mitigating the ET degradation rather than by amplifying the TC gain.

Table 4: Ablation — Uniform vs. Class-Specific Boundary Weighting at ($\lambda = 2$, $\sigma = 5$), Seed 42. All Values Are Per-Slice Test Dice. A Check Mark Indicates the Delta Is Statistically Significant.

Configuration	Mean	WT	TC	ET
Uniform ($m_{NE} = 1$)	0.826	0.856	0.807	0.815
Class-specific ($m_{NE} = 2$)	0.829	0.857	0.811	0.819
Δ (class-specific − uniform)	**+0.003**	**+0.001**	**+0.004**	**+0.004 ✓**

4.5 Qualitative Analysis

Figure 1 illustrates the BW-Dice weight map on a representative test slice containing all three tumor sub-regions. The weight approaches its maximum ($1 + \lambda \cdot m_{NE} = 5.0$ at $\lambda = 2$, $m_{NE} = 2$) precisely at the NCR-edema interface, decays with the Gaussian envelope into region interiors, and returns to unity in background pixels.

Figure 1: BW-Dice Weight Map w(p) on a Representative Test Slice. Left: Ground-Truth Label Map (NCR: Red, Edema: Green, ET: Yellow). Right: Computed Weight Map at $\lambda = 2$, $\sigma = 5$, $m_{ne} = 2$. Peak Weights Trace the NCR-Edema Interface Where w(p) = 5.0; Interior Pixels Have w = 1.

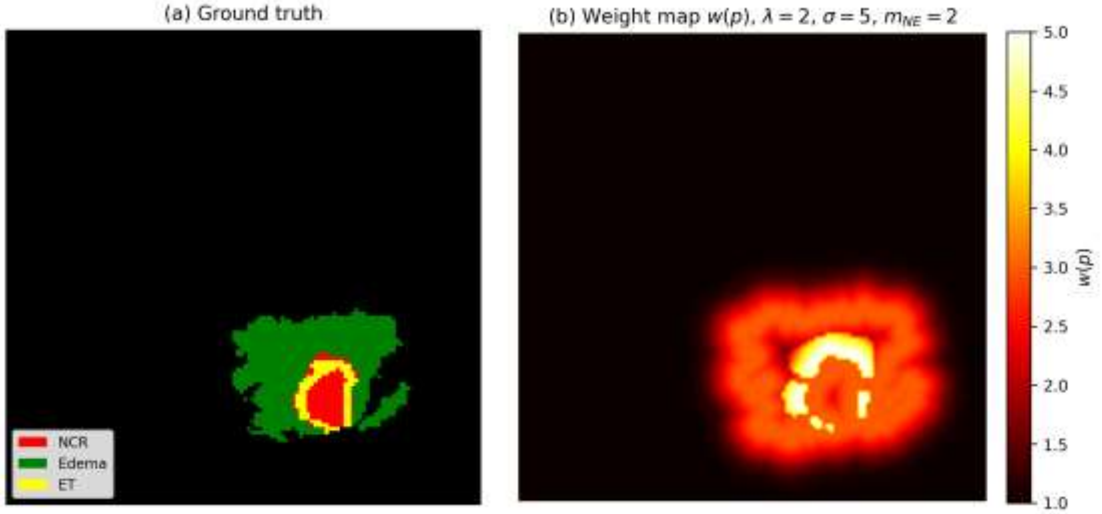
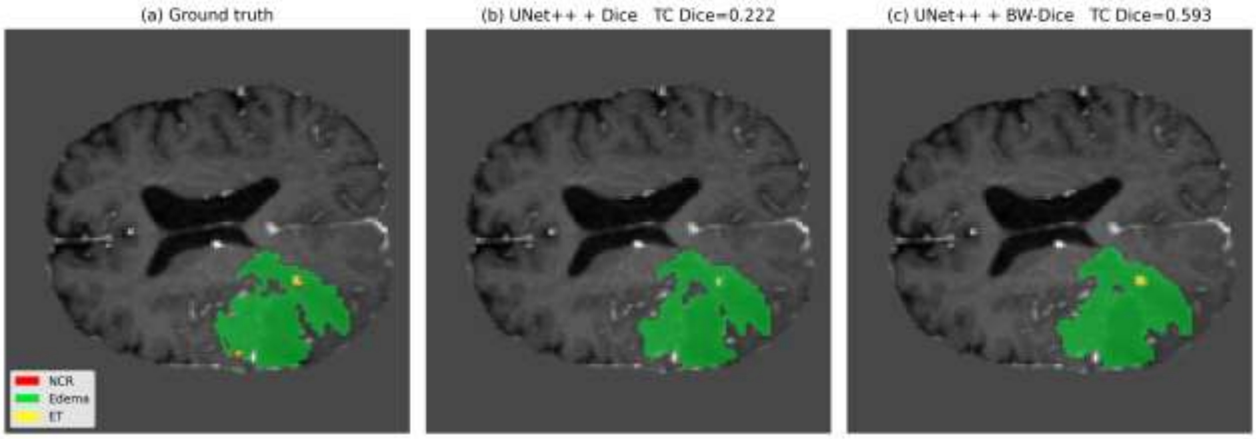


Figure 2 compares vanilla Dice and BW-Dice (2, 5) predictions on test slice 175. The tumor is predominantly edema (green) with small NCR (red) and enhancing tumor (ET, yellow) components. Vanilla Dice (b) misses the ET component almost entirely, whereas BW-Dice (c) recovers a small ET prediction in the upper region of the tumor, improving per-slice TC Dice from 0.22 to 0.59. This slice illustrates the mechanism operating strongly; average TC improvement across all 2,242 test slices is +0.012 (Section 4.3).

Figure 2: Qualitative Comparison on Test Slice 175. (a) Ground Truth. (b) UNet++ + Dice Prediction: TC Dice = 0.222. (c) UNet++ + BW-Dice (2, 5): TC Dice = 0.593. BW-Dice Recovers the Small ET Component That Vanilla Dice Misses.



5. Discussion

5.1 The Reproducible Finding: Regional Trade-off, Not Aggregate Improvement

The results support a specific, statistically robust finding: BW-Dice systematically shifts UNet++ segmentation performance from Enhancing Tumor toward Tumor Core, with the trade-off approximately balancing on aggregate. The TC gain ($\Delta = +0.012$, S/N 3.83) and ET loss ($\Delta = -0.018$, S/N 2.74) reproduce across three random seeds. The aggregate mean Dice ($\Delta = -0.001$, $p = 0.97$) is statistically indistinguishable from vanilla Dice.

This finding is consistent with — and extends — the null result of the prior matched-condition analysis [1], which showed that within the standard loss family, no single loss achieves statistically significant aggregate improvement over Dice on UNet++. BW-Dice, though outside that family, similarly fails to improve aggregate mean Dice under matched conditions. Where BW-Dice differs is in the distribution of performance across sub-regions rather than the aggregate mean.

5.2 Interpreting the Regional Trade-off

Two mechanisms are proposed that plausibly explain the observed trade-off, and these are complementary rather than exclusive.

Scale mismatch between boundary emphasis and class size. ET is the smallest of the three foreground classes on BraTS 2020; on average slices, ET occupies a spatial extent comparable to or smaller than the Gaussian boundary emphasis kernel ($\sigma = 5$ pixels, spanning approximately 10 pixels on each side of a boundary at 128-pixel resolution). For a class whose interior is small relative to the boundary emphasis zone, weighting the boundary 5× heavier than the interior effectively over-emphasizes boundary pixels relative to interior structure. The gradient signal driving interior consistency is dominated by boundary corrections, destabilizing the learning of the class.

Class-uniform gradient scaling on a class-heterogeneous task. In its multiclass form, BW-Dice applies the same per-pixel weight to all classes. Boundary pixels of NCR and edema receive the same weight boost as boundary pixels of ET. But because ET boundaries constitute a much larger fraction of the total pixels of ET than NCR boundaries do of the pixels of NCR, ET experiences proportionally greater gradient contribution from boundary pixels. The class-uniform weighting is thus non-uniform in its effective learning signal per class.

The ablation (Section 4.4) provides evidence for the interpretation: uniform boundary weighting (without NCR-edema emphasis) still produces the qualitative TC+/ET− pattern, indicating that the trade-off arises from boundary emphasis itself rather than from the class-specific NCR-edema multiplier. Adding class-specificity reduces ET degradation but does not amplify TC gain — consistent with the interpretation that the NCR-edema multiplier primarily shifts weight away from ET boundaries, mitigating the ET-specific over-emphasis without redirecting benefit to TC.

5.3 Practical Implications

For BraTS 2020 and similar hierarchical medical segmentation tasks, the findings suggest that uniform boundary emphasis is best applied selectively rather than globally across sub-regions. Two practical directions are noted. **Region-specific boundary emphasis** could be applied to WT and TC while explicitly excluded from ET, if ET size is small relative to the emphasis kernel; and **adaptive emphasis scale** in which the Gaussian spread σ is adapted per class based on typical class size, scaling emphasis kernels down for small classes. Neither direction is a straightforward drop-in modification; both are natural follow-ups to the present work.

5.4 Limitations

Single architecture. The experiments use only UNet++. Preliminary attempts to evaluate BW-Dice on LinkNet under FP16 mixed precision revealed training instabilities specific to the element-wise-additive skip connections of LinkNet, which could not be resolved within the available compute budget. Cross-architecture generalization is future work.

2D processing. Following [1], the method operates on 2D axial slices. 3D context is known to improve BraTS performance, particularly at inter-slice sub-region boundaries. Whether the observed trade-off pattern extends to 3D remains an open question.

Single-dataset evaluation. Findings are specific to BraTS 2020. Generalization to other tumor datasets (BraTS 2021, BraTS-METS, glioma-specific cohorts) requires additional evaluation.

Slice-level bootstrap. Slice-level bootstrap is reported for consistency with [1]. Slice-level testing is anti-conservative relative to case-level testing due to intra-patient slice correlations; case-level bootstrap on 38 test cases would produce wider confidence intervals but consistent effect direction.

Fixed multiplier. The class-specific NCR-edema multiplier is fixed at $m_{NE} = 2.0$ throughout. A systematic sweep of this parameter, or extension to arbitrary class-pair emphases, may reveal further design space.

6. Conclusion

BW-Dice was proposed, a boundary-weighted generalization of multiclass Dice loss with class-specific emphasis on the NCR-edema interface, motivated by a failure mode identified in prior matched-condition analysis of BraTS 2020. Under identical training conditions, multi-seed replication, and paired bootstrap significance testing, BW-Dice does not improve aggregate mean Dice ($\Delta = -0.001$, $p = 0.97$) but produces a highly reproducible regional trade-off: Tumor Core segmentation improves ($\Delta = +0.012$, $p < 10^{-3}$) while Enhancing Tumor degrades ($\Delta = -0.018$, $p < 10^{-3}$). A controlled ablation demonstrates that this trade-off is inherent to boundary emphasis rather than an artifact of class-specific weighting; class-specificity primarily reduces the magnitude of ET degradation. The trade-off is interpreted as a scale mismatch between the boundary emphasis kernel and the smallest sub-region class, and region-specific or adaptive emphasis scales are identified as promising directions for future loss-function design on tasks with hierarchical class structure.

Data Availability

The BraTS 2020 dataset is publicly available via the Center for Biomedical Image Computing and Analytics (CBICA) at the University of Pennsylvania. Preprocessing pipeline, training code, model checkpoints, and evaluation scripts to reproduce all results in this paper are available from the authors on reasonable request.

References

1. S. U. Ravi Kumar Chavali and P. V. Kumar, "A Controlled Comparison of U-Net Decoder Architectures and Loss Functions for Brain Tumor Segmentation on BraTS 2020," 2025 (Accepted for publication).
2. B. H. Menze, A. Jakab, S. Bauer, et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," IEEE Transactions on Medical Imaging, vol. 34, no. 10, pp. 1993–2024, 2015.
3. S. Bakas, H. Akbari, A. Sotiras, et al., "Advancing the Cancer Genome Atlas Glioma MRI Collections with Expert Segmentation Labels and Radiomic Features," Scientific Data, vol. 4, article 170117, 2017.

4. S. Bakas, M. Reyes, A. Jakab, et al., "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge," arXiv:1811.02629, 2018.
5. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in MICCAI, pp. 234–241, 2015.
6. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in 3DV, pp. 565–571, 2016.
7. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in DLMIA, pp. 3–11, 2018.
8. O. Oktay, J. Schlemper, L. Le Folgoc, et al., "Attention U-Net: Learning Where to Look for the Pancreas," in MIDL, 2018.
9. A. Chaurasia and E. Culurciello, "LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation," in IEEE VCIP, pp. 1–4, 2017.
10. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in ICCV, pp. 2980–2988, 2017.
11. S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks," in MLMI, pp. 379–387, 2017.
12. N. Abraham and N. M. Khan, "A Novel Focal Tversky Loss Function with Improved Attention U-Net for Lesion Segmentation," in IEEE ISBI, pp. 683–687, 2019.
13. S. A. Taghanaki, Y. Zheng, S. K. Zhou, et al., "Combo Loss: Handling Input and Output Imbalance in Multi-Organ Segmentation," Computerized Medical Imaging and Graphics, vol. 75, pp. 24–33, 2019.
14. H. Kervadec, J. Bouchtiba, C. Desrosiers, et al., "Boundary Loss for Highly Unbalanced Segmentation," in MIDL, 2019.
15. D. Karimi and S. E. Salcudean, "Reducing the Hausdorff Distance in Medical Image Segmentation with Convolutional Neural Networks," IEEE Transactions on Medical Imaging, vol. 39, no. 2, pp. 499–513, 2020.
16. X. Chen, B. M. Williams, S. R. Vallabhaneni, et al., "Learning Active Contour Models for Medical Image Segmentation," in CVPR, pp. 11632–11640, 2019.
17. 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," Nature Methods, vol. 18, no. 2, pp. 203–211, 2021.
18. F. Isensee, P. F. Jaeger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, "nnU-Net for Brain Tumor Segmentation," in Brainlesion (BrainLes 2020), pp. 118–132, 2021.
19. F. Isensee, T. Wald, C. Ulrich, et al., "nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation," in MICCAI, 2024.
20. J. Chen, Y. Lu, Q. Yu, et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv:2102.04306, 2021.
21. A. Hatamizadeh, V. Nath, Y. Tang, et al., "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," in Brainlesion (BrainLes 2021), pp. 272–284, 2022.
22. P. Iakubovskii, "Segmentation Models Pytorch," GitHub repository, 2019. https://github.com/qubvel/segmentation_models.pytorch