A Controlled Comparison of U-Net Decoder Architectures and Loss Functions for Brain Tumor Segmentation on BraTS 2020

S.U. Ravi Kumar Chavali
Research Scholar
Rayalaseema University, Kurnool
Andhra Pradesh, India

P.V. Kumar
School of Engineering
Anurag University
Hyderabad, Telangana, India

ABSTRACT
A matched-condition comparison of four U-Net family decoder architectures (U-Net, UNet++, MAnet, LinkNet) crossed with four standard segmentation losses (Dice, Focal, Tversky, Focal-Tversky) is conducted on the BraTS 2020 dataset, producing a 16-configuration matrix. All configurations share an identical ResNet34 encoder pre-trained on ImageNet, identical preprocessing, augmentation, optimizer schedule, and held-out 38-case test set. Statistical significance is assessed via paired bootstrap resampling with 10,000 iterations on per-case Dice. Mean Dice across all 16 configurations falls in the narrow range 0.792 to 0.826. UNet++ paired with Dice loss ranks first (mean 0.826; WT 0.851, TC 0.799, ET 0.828), but the top five configurations are statistically indistinguishable (all p greater than 0.32, with 95 percent confidence intervals spanning zero). LinkNet with Dice attains mean Dice 0.815, within 0.011 of the top, while requiring 2.7 times less inference time (53.9 versus 145.4 ms per case) and 18 percent fewer parameters than UNet++ with Dice. Worst-case analysis reveals three failure patterns: edema over-segmentation in atypical presentations, necrotic-core and edema confusion, and missed multi-focal disease. The findings suggest that improvements claimed in prior literature may be partly attributable to unreported variation in experimental conditions, and that LinkNet offers a favorable accuracy-efficiency trade-off for clinical deployment.

General Terms
Experimentation, Measurement, Performance.

Keywords
Brain tumor segmentation, BraTS 2020, U-Net, deep learning, multi-modal MRI, comparative evaluation, statistical significance, computational efficiency.

1. INTRODUCTION
1.1 Clinical Context
Gliomas are the most common primary malignant brain tumors in adults, with an annual incidence of roughly six per 100,000 [19]. Manual delineation of tumor sub-regions on multi-modal MRI by expert neuroradiologists, the current clinical standard, takes 30 to 60 minutes per case and shows inter-rater Dice agreement of only 0.74 to 0.85 [16]. Following BraTS convention, performance is evaluated on three hierarchical sub-regions: Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET). The Brain Tumor Segmentation (BraTS) challenge series [16, 4, 3] has standardized datasets and protocols, and U-Net family methods have come to dominate the task.

1.2 The Current State of the Field
The U-Net of Ronneberger et al. [20] and its extensions, including Attention U-Net [18], UNet++ [23], V-Net [17], LinkNet [5], and various Res-U-Net hybrids, form a busy family of architectures, each claiming improvements over a baseline. nnU-Net [10] demonstrated that careful preprocessing and training of a standard U-Net can match or exceed elaborate architectures. Transformer-based approaches [6, 8] have also been proposed. However, cross-paper comparison is difficult, since different studies use different preprocessing, augmentation, and training schedules, so the marginal contribution of any individual architectural innovation is unclear. Recent work [12], nnU-Net Revisited, shows that under rigorous matched evaluation many architectures fail to outperform a properly tuned nnU-Net baseline. The more granular question of how individual U-Net family decoders and losses compare under matched conditions has not been systematically addressed for brain tumor segmentation.

1.3 Contributions
This gap is addressed with a 16-configuration matched-condition comparison on BraTS 2020. All configurations share an identical ResNet34 encoder, preprocessing, augmentation, optimizer schedule, and held-out test set, isolating the effects of decoder topology and loss-function choice from confounds. The contributions are: (C1) the first head-to-head comparison of four U-Net decoders crossed with four standard losses on BraTS 2020 under fully matched conditions; (C2) an empirical null result, in which paired bootstrap testing shows the top five configurations to be statistically indistinguishable (all p greater than 0.32, 95 percent confidence intervals spanning zero); (C3) an accuracy-efficiency Pareto analysis showing that LinkNet attains accuracy within 0.011 of the top-ranked configuration while running 2.7 times faster; (C4) a sub-region-level and failure-mode analysis identifying three clinically distinct failure patterns; and (C5) a reproducible methodological framework, in which preprocessing scripts, training code, evaluation scripts, and all 16 trained checkpoints are publicly released.

2. RELATED WORK
2.1 U-Net and the Encoder-Decoder Paradigm
The U-Net of Ronneberger et al. [20], with its encoder-decoder structure and skip connections, has dominated medical image segmentation for nearly a decade and was extended to 3D in [7]. The Dice loss introduced with V-Net [17] became the standard objective for class-imbalanced sub-region segmentation. The original U-Net topology and the V-Net Dice loss are adopted here as the baseline reference.

2.2 Architectural Extensions of U-Net
Several U-Net variants address specific weaknesses: Attention U-Net [18] adds attention gates on skip connections, UNet++ [23]

uses nested dense skip connections, Res-U-Net [9] integrates residual blocks in the encoder, and LinkNet [5] swaps concatenation for element-wise addition. Each was validated against vanilla U-Net on a different task. No prior study has systematically compared these variants on a shared brain-tumor benchmark under matched conditions, a gap this study addresses.

2.3 Configuration-Centric Approaches

A complementary line of work argues that disciplined preprocessing and training matters as much as architectural innovation. nnU-Net [10], a self-configuring framework using a standard U-Net, matched or outperformed specialized methods on the Medical Segmentation Decathlon and BraTS. Its BraTS-specific variant [11] won the BraTS 2020 challenge (Dice 0.890 / 0.851 / 0.820). This motivates the matched-condition design adopted here.

2.4 Transformer-Based Segmentation

Transformer-based approaches such as TransUNet [6] and Swin UNETR [8] have been proposed, with the latter achieving competitive performance on BraTS 2021 (Dice 0.926 / 0.878 / 0.858). However, nnU-Net Revisited [12] showed that most transformer models do not outperform properly tuned CNN baselines under matched compute. Transformer comparison lies outside the scope of this 2D matched-condition study.

2.5 Surveys and Methodological Critiques

Recent surveys [15, 14, 1] consistently identify three open problems: (i) lack of standardized evaluation protocols, (ii) limited reporting of computational cost alongside accuracy, and (iii) insufficient failure-mode analysis. Reference [12] reinforces these concerns regarding non-standardized validation. The present work addresses all three: matched-condition comparison (Section 3), computational cost analysis (Section 4.4), and failure-mode analysis (Section 4.5).

3. METHODOLOGY

The full experimental pipeline is summarized in Figure 1.

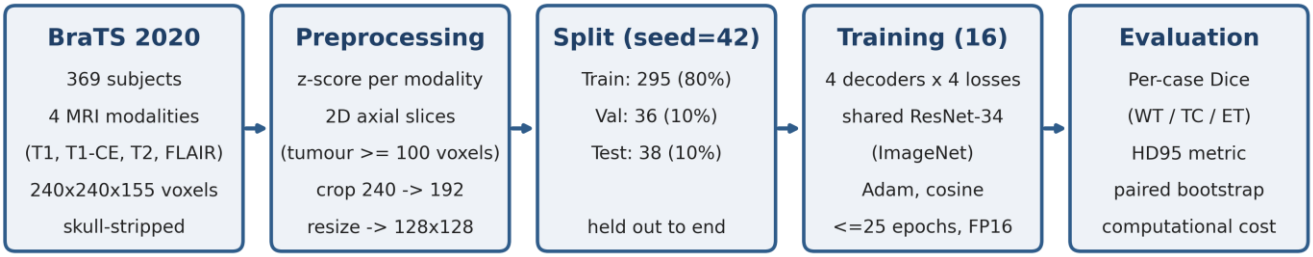

Fig. 1: Experimental pipeline overview: from raw BraTS 2020 volumes through preprocessing, deterministic split, and matched-condition training of all 16 configurations, to evaluation on the held-out test set.

3.1 Dataset and Split

The BraTS 2020 dataset [16, 4, 3] is used: 369 cases of high-grade and low-grade glioma, each with four co-registered MRI sequences (T1, T1-CE, T2, FLAIR) at 240 x 240 x 155 voxels, skull-stripped and registered to a common template. Each subject is annotated by expert neuroradiologists with three tumor labels (necrotic and non-enhancing core, peritumoral edema, enhancing tumor). The native label space is remapped from {0, 1, 2, 4} to a contiguous range {0, 1, 2, 3}, and cases are partitioned 80:10:10 (295 train / 36 validation / 38 test) with seed 42. The test split is held out entirely until final evaluation.

3.2 Preprocessing

Each MRI volume is preprocessed independently per modality: z-score normalization over non-zero (foreground) voxels; 2D axial slice extraction (slices with at least 100 tumor voxels); center-crop 240 to 192 and resize to 128 x 128 (bilinear for images, nearest-neighbor for masks); and four-channel stacking of modalities. Preprocessed slices are stored as float16 arrays, totaling approximately 30,000 tumor-bearing slices across all splits.

3.3 Architectures

Four U-Net family decoder architectures are compared, all built on a shared ResNet34 encoder pre-trained on ImageNet [9], isolating the comparison to decoder design: U-Net [20] (canonical encoder-decoder with single skip connections, the baseline); UNet++ [23] (nested dense skip pathways); MAnet (positional and channel attention in the decoder); and LinkNet [5] (element-wise addition instead of concatenation for efficiency). All implementations use the segmentation-models-pytorch library with four input channels and four output classes.

3.4 Loss Functions

Each architecture is crossed with four losses: Dice [17] (overlap-based, field-standard); Focal [13] (gamma = 2, alpha = 0.25); Tversky [21] (alpha = 0.3, beta = 0.7, emphasizing recall); and Focal-Tversky [2] (gamma = 1.333).

3.5 Training Setup

All 16 configurations are trained under identical conditions with the Adam optimizer (learning rate 1 x 10^-4, cosine annealing) for at most 25 epochs with early stopping (patience 5) on validation mean Dice. Batch sizes match GPU memory: 32 for U-Net and LinkNet, 16 for UNet++ and MAnet; the batch-size variation is acknowledged as a confound and is expected to be small [22]. Augmentation is limited to random horizontal and vertical flipping (p = 0.5) to avoid augmentation-by-architecture interactions. Mixed-precision (FP16) training is used via automatic casting and gradient scaling. All experiments run on a single NVIDIA T4 GPU under Google Colaboratory free-tier limits, with checkpoint-and-resume enabling multi-day continuation. For two MAnet configurations (Dice and Focal-Tversky) whose mixed-precision training induced NaN corruption of decoder batch-normalization running statistics, these statistics are recomputed after training via 200 forward-only passes, with the trained weights left unchanged.

3.6 Evaluation Metrics

Per-case Dice is reported for WT, TC, and ET (aggregating intersection and union within each case, giving statistically independent samples), together with the 95th-percentile Hausdorff Distance (HD95), computational-cost metrics (parameter count, GFLOPs per slice, milliseconds per case on the T4 at batch 32), and paired bootstrap resampling (10,000

iterations) for statistical significance with 95 percent confidence intervals and two-tailed p-values.

4. RESULTS

4.1 Overall Performance Across Configurations

Mean Dice and HD95 for all 16 configurations are reported in Table 1. Mean Dice ranged from 0.792 (MAnet with Focal) to 0.826 (UNet++ with Dice), a spread of only 0.034 across all 16. The mean-Dice ranking by architecture family (averaging across losses) is UNet++ (0.821) greater than U-Net (0.814) greater than LinkNet (0.808) approximately equal to MAnet (0.804), with no architecture dominating across all losses. Validation-versus-test gaps were small (approximately 0.010 to 0.030), indicating good generalization. Figure 2 visualizes mean test Dice as a 4 x 4 heatmap, with UNet++ and Dice highlighted as best. The narrow color spread is striking: the majority of configurations fall within plus or minus 0.010 of the median, suggesting that architecture and loss choice have limited impact on aggregate Dice under matched conditions.

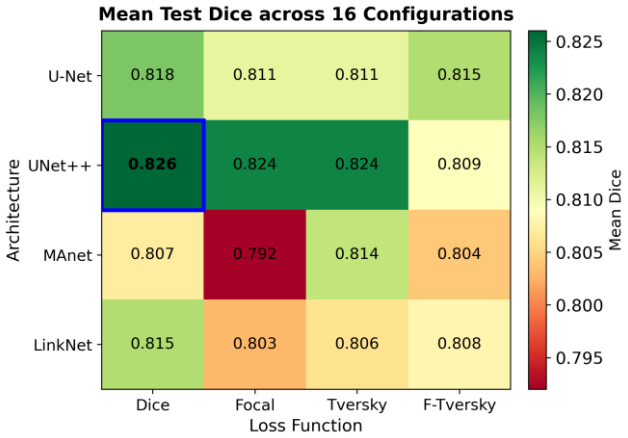

Fig. 2: Mean test Dice across the 16 (architecture x loss) configurations. UNet++ with Dice (0.826) is outlined in blue as the top-ranked configuration.

Table 1: Test-set performance for all 16 (architecture x loss) configurations on BraTS 2020. Best values per column are in bold. WT: Whole Tumor; TC: Tumor Core; ET: Enhancing Tumor; HD95 in pixels; Val = validation-set mean Dice.

Architecture	Loss	WT	TC	ET	Mean	WT	TC	ET	Val
U-Net	Dice	0.855	0.794	0.806	0.818	3.36	2.63	2.86	0.834
U-Net	Focal	0.851	0.789	0.792	0.811	3.25	3.03	3.28	0.827
U-Net	Tversky	0.859	0.775	0.799	0.811	3.57	2.92	3.15	0.836
U-Net	Focal-Tversky	0.857	0.791	0.799	0.815	3.94	2.52	2.88	0.827
UNet++	**Dice**	**0.851**	**0.799**	**0.828**	**0.826**	**4.45**	**3.46**	**2.84**	**0.841**
UNet++	Focal	0.861	0.799	0.813	0.824	3.30	2.50	2.82	0.835
UNet++	Tversky	0.858	0.796	0.817	0.824	3.83	2.98	2.85	0.834
UNet++	Focal-Tversky	0.854	0.778	0.794	0.809	4.54	3.73	3.07	0.832
MAnet	Dice	0.844	0.774	0.804	0.807	4.34	3.52	3.14	0.826
MAnet	Focal	0.822	0.766	0.788	0.792	3.31	2.96	2.90	0.827
MAnet	Tversky	0.855	0.788	0.800	0.814	3.08	2.76	2.96	0.838
MAnet	Focal-Tversky	0.844	0.775	0.793	0.804	5.01	4.05	3.87	0.826
LinkNet	Dice	0.851	0.787	0.807	0.815	3.82	3.10	2.64	0.825
LinkNet	Focal	0.853	0.775	0.782	0.803	3.39	3.24	3.17	0.811
LinkNet	Tversky	0.853	0.778	0.786	0.806	3.48	3.10	3.01	0.814
LinkNet	Focal-Tversky	0.850	0.788	0.786	0.808	4.46	3.11	3.11	0.820

Dice metrics (higher is better) on the left; HD95 (lower is better) on the right.

4.2 Per-Sub-Region Performance

The sub-region breakdown reveals patterns invisible in the aggregate (Figure 3). All configurations achieved high WT Dice (0.822 to 0.861), indicating that gross localization is architecture-independent. The greatest variability occurred in Tumor Core (0.766 to 0.799) and Enhancing Tumor (0.782 to 0.828). The best configuration differed per sub-region: WT best was UNet++ with Focal (0.861), TC best UNet++ with Dice (0.799), and ET best UNet++ with Dice (0.828). Counter-intuitively, the asymmetric losses (Tversky, Focal-Tversky) designed for class imbalance did not consistently outperform symmetric Dice on the most imbalanced sub-region (ET); UNet++ with Dice achieved the highest ET Dice across all 16.

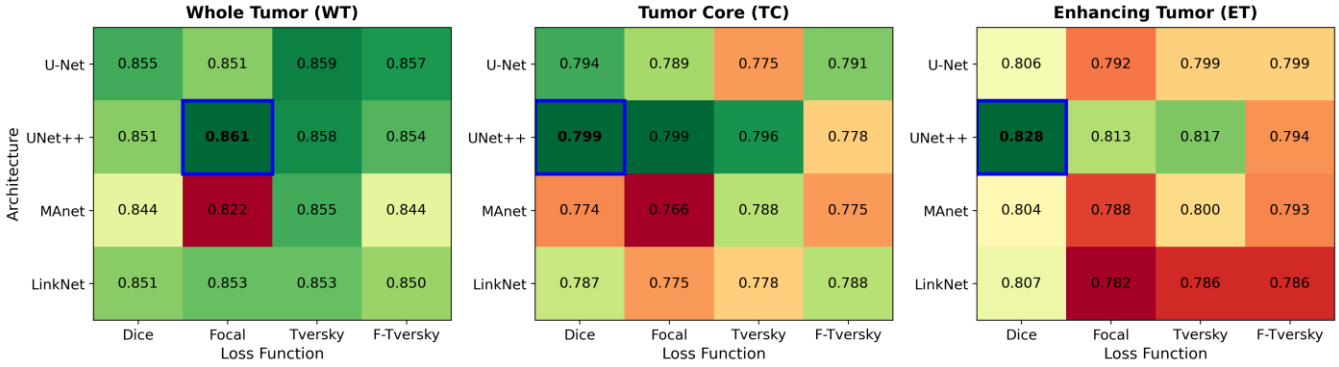

Fig. 3: Per-sub-region Dice across all 16 configurations. Left: Whole Tumor (WT). Middle: Tumor Core (TC). Right: Enhancing Tumor (ET). Sub-region-best cells are outlined in blue.

4.3 Statistical Significance

Paired bootstrap resampling (10,000 iterations) on per-case Dice yielded the comparisons in Table 2 and Figure 7. Relative to the top-ranked UNet++ with Dice, the differences are: versus U-Net with Focal-Tversky (delta = +0.0060, CI [-0.0093, +0.0267], p = 0.544); versus UNet++ with Tversky (delta = +0.0068, CI [-0.0053, +0.0273], p = 0.594); versus UNet++ with Focal (delta = +0.0073, CI [-0.0066, +0.0271], p = 0.430); and versus U-Net with Dice (delta = +0.0093, CI [-0.0050, +0.0311], p = 0.329). None reached significance at alpha = 0.05, and all 95 percent confidence intervals include zero. The top five configurations are therefore statistically indistinguishable in mean per-case Dice on this test set. This is the central empirical finding: under matched conditions, the choice of decoder architecture and loss function among standard U-Net variants does not yield statistically significant accuracy differences on test data.

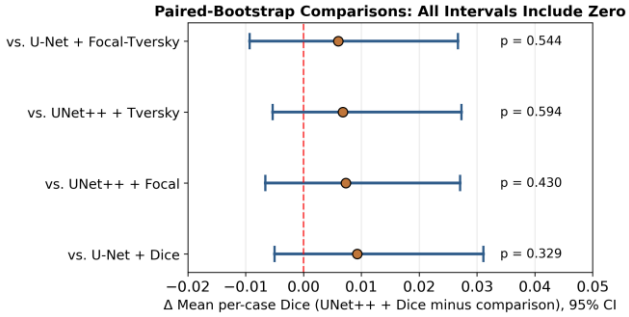

Fig. 7: Paired-bootstrap comparison of the top configuration against the next four. Every 95 percent confidence interval crosses the dashed zero line, so no difference is significant.

The absence of significance is best read as an upper bound on the effect size rather than proof of exact equivalence. With 38 independent test cases, the paired bootstrap can resolve mean-Dice differences of roughly 0.03, whereas the observed gaps among the top five configurations are all below 0.01; any true difference is thus smaller than this test can detect at the available sample size.

Table 2: Pairwise bootstrap comparison of the top-ranked configuration (UNet++ with Dice) against the next four ranked configurations. Delta is the mean per-case Dice difference (UNet++ with Dice minus comparison); the 95 percent CI is from 10,000 bootstrap iterations on per-case scores; the p-value is two-tailed. The last column indicates whether the 95 percent CI excludes zero

Configuration A	Configuration B	Delta Mean Dice	95% CI	p-value	Significant?
UNet++ + Dice	U-Net + Focal-Tversky	+0.0060	[-0.0093, +0.0267]	0.544	No
UNet++ + Dice	UNet++ + Tversky	+0.0068	[-0.0053, +0.0273]	0.594	No
UNet++ + Dice	UNet++ + Focal	+0.0073	[-0.0066, +0.0271]	0.430	No
UNet++ + Dice	U-Net + Dice	+0.0093	[-0.0050, +0.0311]	0.329	No

4.4 Computational Analysis

While the top configurations are statistically tied on Dice, they differ substantially in cost (Table 3, Figure 6): U-Net (24.44 M parameters, 3.96 GFLOPs per slice, 67.3 ms per case), UNet++ (26.08 M, 9.26 GFLOPs, 145.4 ms), MAnet (31.79 M, 4.22 GFLOPs, 75.9 ms), and LinkNet (21.78 M, 2.77 GFLOPs, 53.9 ms). Three observations follow. First, the nested skip pathways of UNet++ add about 6 percent to parameters but more than double the inference FLOPs and time over the baseline U-Net, a disproportionate overhead for a marginal accuracy gain. Second, MAnet has the largest parameter count yet ranks outside the top five, and is Pareto-dominated by smaller alternatives. Third, LinkNet has the smallest parameter count, the lowest FLOPs, and the highest throughput, with a mean Dice (0.815) matching U-Net with Focal-Tversky and only 0.011 below the top. For inference-constrained deployment, LinkNet offers the most favorable accuracy-efficiency trade-off, as Figure 6 makes visually explicit.

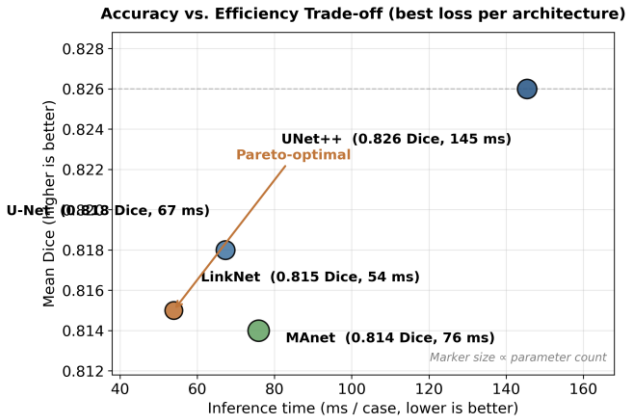

Fig. 6: Accuracy versus efficiency for the best loss of each architecture. LinkNet lies on the Pareto frontier, combining the lowest inference time with near-top accuracy; marker area is proportional to parameter count.

Table 3: Computational cost per architecture: parameter count, GFLOPs per 128 x 128 slice, mean per-case inference time on a single NVIDIA T4 GPU at batch size 32, and throughput in cases per second. The loss function does not change these costs within an architecture.

Architecture	Params (M)	GFLOPs / slice	ms / case	Cases / s
U-Net	24.44	3.96	67.3	14.85
UNet++	26.08	9.26	145.4	6.88
MAnet	31.79	4.22	75.9	13.17
LinkNet	21.78	2.77	53.9	18.56

4.5 Qualitative Results and Failure-Mode Analysis

Figure 4 compares the top-accuracy configuration (UNet++ with Dice) against the efficiency-optimal one (LinkNet with Dice) on three representative test cases spanning easy, medium, and hard difficulty tertiles. The two produce visually similar segmentations across difficulty levels, consistent with the statistical equivalence established in Section 4.3. Figure 5 analyzes the three lowest-performing test cases under the top model with per-pixel error maps. Three failure patterns emerge: (1) over-segmentation of peritumoral edema in atypical glioma presentations (Case 277, Dice 0.578), in which substantial non-tumor tissue is labeled as edema, suggesting a reliance on FLAIR intensity that breaks down for out-of-distribution presentations; (2) misclassification of edema as necrotic core (Case 331, Dice 0.709), in which the model over-extends necrotic labels into edema regions, reflecting the well-known difficulty of the necrotic-edema boundary; and (3) under-detection of multi-focal disease (Case 309, Dice 0.720), in which the dominant mass is captured but portions of a spatially separated lesion component are missed. These distinct failure modes motivate the architectural enhancements pursued in follow-up work.

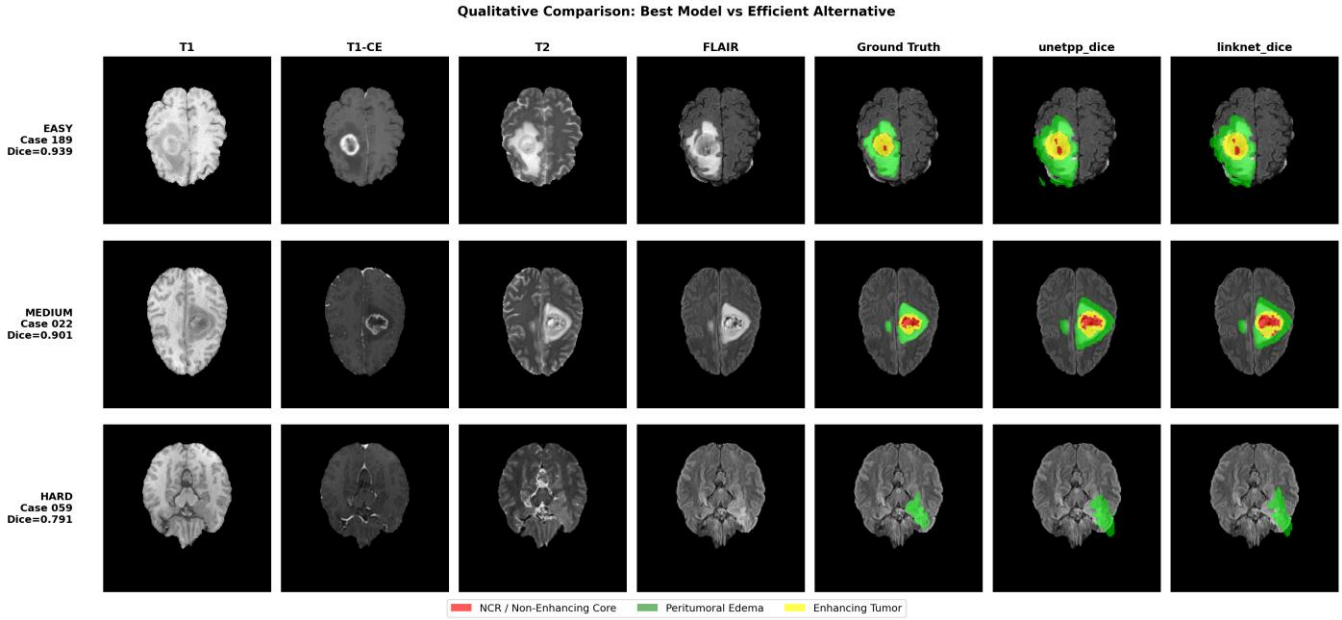

Fig. 4: Qualitative comparison of the top-accuracy (UNet++ with Dice) and efficiency-optimal (LinkNet with Dice) configurations on three representative test cases (easy: Case 189, Dice 0.939; medium: Case 022, Dice 0.901; hard: Case 059, Dice 0.791). Overlay colors: red = necrotic and non-enhancing core, green = peritumoral edema, yellow = enhancing tumor.

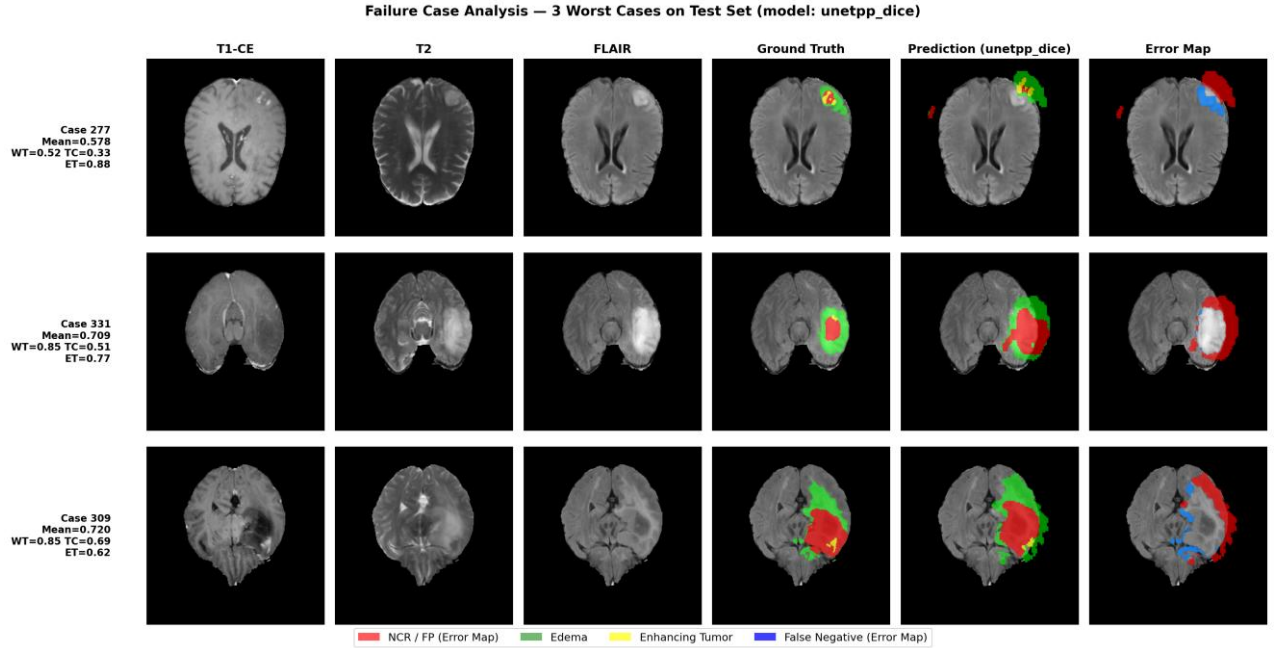

Fig. 5: The three worst-performing test cases under UNet++ with Dice. Columns: T1-CE, T2, FLAIR, ground-truth overlay, prediction overlay, and per-pixel error map (red = false positive, blue = false negative).

5. DISCUSSION

5.1 Interpreting the Null Result

The null result is not the absence of effect but an upper bound on effect size: any true difference between the top configurations is smaller than test-set sampling noise can detect at this sample size. This is consistent with [12]: when preprocessing, training, encoder backbone, and augmentation are held constant, the marginal effect of varying decoder topology among U-Net variants, or of varying loss among standard formulations, is too small to detect on a moderately sized test set. It should be emphasized that this does not contradict prior literature claiming improvements from Attention U-Net, UNet++, or LinkNet; rather, it indicates that such improvements are not separable from experimental-condition effects at this scale.

5.2 Accuracy-Efficiency Trade-offs

The top configurations are statistically indistinguishable on Dice but differ substantially in cost. UNet++ requires 2.2 times the inference time of U-Net and 2.7 times that of LinkNet, without a significant accuracy gain. LinkNet attains accuracy within 0.011 of the top-ranked configuration while using the fewest parameters and the lowest inference time. For inference-bound scenarios such as high-throughput screening, edge deployment, or limited memory, LinkNet is Pareto-optimal within this matrix.

5.3 Sub-region Sensitivity

Despite the aggregate equivalence, per-sub-region rankings vary: UNet++ with Focal is best for WT (0.861), while UNet++ with Dice leads on both TC (0.799) and ET (0.828). For applications focused on a specific sub-region, for example ET delineation for radiotherapy targeting, aggregate-best model selection may still be optimal within this matrix; however, within-sub-region rankings are not subjected to significance testing here, and rigorous sub-region claims would require larger test sets than the 38 cases available.

5.4 Limitations

Four limitations bound these conclusions. First, the comparison is restricted to 2D axial slices due to free-tier compute constraints; 3D context is known to improve performance, particularly at sub-region boundaries varying along the inferior-superior axis, so the findings apply to 2D decoder choice rather than to 3D more broadly. Second, the 38-case test set is sufficient for large effects but insufficient for subtle architectural differences; larger evaluation cohorts or cross-validation would tighten the confidence intervals of Section 4.3. Third, each configuration was trained with a single initialization, and multi-seed evaluation would yield more robust conclusions at three to five times the compute cost. Fourth, batch size varies across architectures (16 versus 32) due to memory, an effect expected to be small [22] but not ruled out, and the restricted augmentation (flipping only), while avoiding augmentation-by-architecture confounding, may not generalize to stronger augmentation regimes.

6. CONCLUSION

A systematic, matched-condition evaluation of four U-Net family decoder architectures crossed with four standard segmentation losses for BraTS 2020 brain tumor segmentation has been presented. Three central findings emerge. First, paired bootstrap testing reveals the top five configurations to be statistically indistinguishable (all p greater than 0.32, 95 percent confidence intervals spanning zero), suggesting that prior architectural and loss-function improvement claims may be partly attributable to unreported variation in experimental conditions. Second, LinkNet attains accuracy within 0.011 of the top-ranked configuration while requiring 2.7 times less inference time and 18 percent fewer parameters than UNet++ with Dice, which is directly relevant for inference-constrained deployment. Third, the qualitative analysis identifies three failure patterns, namely edema over-segmentation in atypical presentations, necrotic-core and edema confusion, and missed multi-focal disease, that motivate concrete architectural targets for future work. In forthcoming companion work, a novel enhanced architecture addressing these failure modes is proposed. The complete pipeline, preprocessing, training, and evaluation code, and all 16 trained checkpoints are released, to enable reproducibility and to provide a methodological template for future comparative studies in medical image segmentation.

7. ACKNOWLEDGEMENTS

The authors thank the organizers of the BraTS challenge for making the BraTS 2020 dataset publicly available, and the Kaggle community for hosting an accessible distribution. Computational experiments were conducted using the Google Colaboratory free-tier GPU service. This work received no specific external funding.

8. DATA AND CODE AVAILABILITY

The BraTS 2020 dataset is publicly available via Kaggle. Code and trained checkpoints will be made publicly available upon acceptance.

9. REFERENCES

[1] Z. U. Abidin, R. A. Naqvi, A. Haider, H. S. Kim, D. Jeong, and S. W. Lee, "Recent deep learning-based brain tumor segmentation models using multi-modality magnetic resonance imaging: A prospective survey," Frontiers in Bioengineering and Biotechnology, vol. 12, art. 1392807, 2024.

[2] N. Abraham and N. M. Khan, "A novel focal Tversky loss function with improved attention U-Net for lesion segmentation," in IEEE International Symposium on Biomedical Imaging (ISBI), pp. 683-687, 2019.

[3] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, et al., "The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification," arXiv:2107.02314, 2021.

[4] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, et al., "Advancing the Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," Scientific Data, vol. 4, art. 170117, 2017.

[5] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in IEEE Visual Communications and Image Processing (VCIP), 2017.

[6] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, et al., "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv:2102.04306, 2021.

[7] O. Cicek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in MICCAI, pp. 424-432, Springer, 2016.

[8] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in Brainlesion (BrainLes 2021), vol. 12962, pp. 272-284, Springer, 2022.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.

[10] 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.

[11] F. Isensee, P. F. Jaeger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, "nnU-Net for brain tumor segmentation," arXiv:2011.00848, 2020.

[12] F. Isensee, T. Wald, C. Ulrich, M. Baumgartner, S. Roy, K. Maier-Hein, and P. F. Jaeger, "nnU-Net Revisited: A call for rigorous validation in 3D medical image segmentation," in MICCAI, 2024. arXiv:2404.09556.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in IEEE International Conference on Computer Vision (ICCV), pp. 2980-2988, 2017.

[14] Z. Liu, L. Chen, L. Tong, F. Zhou, Z. Jiang, et al., "Deep learning based brain tumor segmentation: A survey," arXiv:2007.09479, 2020.

[15] T. Magadza and S. Viriri, "Deep learning for brain tumor segmentation: A survey of state-of-the-art," Journal of Imaging, vol. 7, no. 2, art. 19, 2021.

[16] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," IEEE Transactions on Medical Imaging, vol. 34, no. 10, pp. 1993-2024, 2015.

[17] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in 4th International Conference on 3D Vision (3DV), pp. 565-571, IEEE, 2016.

[18] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, et al., "Attention U-Net: Learning where to look for the pancreas," in Medical Imaging with Deep Learning (MIDL), 2018.

[19] Q. T. Ostrom, G. Cioffi, K. Waite, C. Kruchko, and J. S. Barnholtz-Sloan, "CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2014-2018," Neuro-Oncology, vol. 23, Supplement 3, pp. iii1-iii105, 2021.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234-241, Springer, 2015.

[21] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," in Machine Learning in Medical Imaging (MLMI), vol. 10541, pp. 379-387, Springer, 2017.

[22] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le, "Don't decay the learning rate, increase the batch size," in International Conference on Learning Representations (ICLR), 2018.

[23] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in Deep Learning in Medical Image Analysis (DLMIA), vol. 11045, pp. 3-11, Springer, 2018.