

DO FUSION MODULES AND COMPOUND LOSSES REALLY HELP? A MULTI-SEED ABLATION OF A HYBRID CNN–TRANSFORMER MODEL FOR 3D BRAIN TUMOR SEGMENTATION

G. X. Abdukadirova

Fergana State Technical University

M. I. Abduqodirova

Fergana State Technical University

Abstract:

Hybrid CNN–Transformer models for brain tumor segmentation are usually justified by ablation studies in which each variant is trained once, although run-to-run variation can be as large as the reported gains. This study re-examines the components of the HyCTSeg model - a parallel CNN and Swin Transformer encoder with bidirectional cross-attention fusion (CAFF) and a four-component loss (Dice, cross-entropy, focal, Tversky) - with a statistically powered design. Nine variants were each trained with four random seeds on BraTS 2021 (32 independent runs, 150 epochs) on an identical data split, and compared by paired t-tests with Benjamini–Hochberg correction. The multi-seed design reduced the standard deviation to 0.0006–0.0018 Dice. The full model ranked third of five architectures (0.8593 ± 0.0015); the Swin-only encoder was best (0.8645 ± 0.0016) and exceeded the full model by 0.0052 ($p = 0.0096$), and simple additive fusion exceeded it by 0.0032 ($p = 0.0107$), both significant after correction. The differences were concentrated in the enhancing tumor region. The compound loss did not outperform Dice loss alone (-0.0004 , $p = 0.38$). These negative results show that single-run ablations can overstate the value of architectural components and argue for multi-seed evaluation as standard practice.

Keywords: brain tumor segmentation, hybrid CNN–Transformer, feature fusion, cross-attention, compound loss, ablation study, multi-seed evaluation, reproducibility, negative results, BraTS 2021.

INTRODUCTION

Automatic segmentation of gliomas on multimodal MRI is a mature research area supported by the BraTS benchmark [1, 2]. Following the success of convolutional networks [3, 4, 5], a large number of hybrid CNN–Transformer architectures have been proposed, including TransBTS [6], Swin UNETR [7] and TransFuse [8]. New models typically add a component – a fusion module, an attention block or a new

loss term – and justify it by an ablation table in which removing the component lowers the Dice score by a few thousandths.

Such evidence has a well-known weakness. Deep networks trained from different random initializations and data orders produce different results, and in segmentation benchmarks this variation is of the order of 0.001–0.003 Dice [9, 10]. When each ablation variant is trained once, a difference of that size cannot be separated from noise. Surveys of machine learning practice have repeatedly identified single-run comparisons as a major source of irreproducible claims [9, 11].

During the development of HyCTSeg, a hybrid model proposed in the author’s dissertation research, an initial single-run ablation showed differences between variants that were smaller than the estimated run-to-run noise (0.0029). Rather than reporting the most favorable run, the experiment was repeated with a design that has sufficient statistical power.

The aim of this work is to determine, with a multi-seed paired design, whether the two main components of HyCTSeg – the cross-attention fusion module and the compound loss – improve segmentation accuracy compared with simpler alternatives. Three hypotheses were tested: *H1* – the hybrid with CAFF outperforms a single CNN encoder; *H2* – CAFF outperforms concatenation-based fusion; *H4* – the four-component loss outperforms Dice loss alone.

METHODS

Related work

nnU-Net [5] demonstrated that careful configuration of a standard U-Net often matters more than architectural novelty, and later benchmarking studies reached similar conclusions for Transformer-based segmentation models [12]. Fusion of CNN and Transformer features ranges from simple addition or concatenation to attention-based schemes such as the BiFusion module of TransFuse [8]. Compound losses combining Dice [13], cross-entropy, focal [14] and Tversky [15] terms are widely used to address class imbalance, although their benefit over a plain Dice loss has been reported to be task-dependent [16]. Methodological work recommends repeated runs, confidence intervals and correction for multiple comparisons when comparing learning algorithms [9, 11, 17]; these practices are still rare in segmentation papers.

Model and variants

HyCTSeg feeds the four-channel MRI volume X to a 3D residual CNN encoder and a 3D Swin Transformer encoder in parallel. At the bottleneck the two feature maps are fused by CAFF, in which each branch attends to the other:

$$F_{c \rightarrow s} = \text{Attn}(Q_c, K_s, V_s), \quad F_{s \rightarrow c} = \text{Attn}(Q_s, K_c, V_c), \quad (1)$$

followed by projection, a residual connection and layer normalization. A decoder with attention-gated skip connections from the CNN encoder produces the segmentation. Five architecture variants and four loss variants were compared (Table 1); all other settings were held constant.

Table 1. Ablation variants

Code	Group	Description
A_cnn_only	Architecture	CNN encoder only
B_swin_only	Architecture	Swin Transformer encoder only
C_add	Architecture	Parallel encoders, element-wise addition
D_concat	Architecture	Parallel encoders, concatenation
E_caff	Architecture	Parallel encoders, CAFF (full model)
L1_dice	Loss	Dice only
L2_dice_ce	Loss	Dice + cross-entropy
L3_no_tversky	Loss	Dice + cross-entropy + focal
L4_full	Loss	Dice + cross-entropy + focal + Tversky (full model)

Experimental design

Each of the nine variants was trained with four random seeds (1–4). Because E_caff and L4_full denote the same configuration, this gives 32 independent runs. The training/validation split of BraTS 2021 [2] was fixed for all runs (split seed 42), so that only model initialization and augmentation order varied. Each run lasted 150 epochs; the total computation was about 100 GPU-hours on an NVIDIA RTX 5080 (Table 2).

Table 2. Experimental settings

Parameter	Value
Variants	9 (5 architecture + 4 loss)
Seeds per variant	4
Independent runs	32
Epochs per run	150
Data split	Identical for all runs (seed 42)
Primary metric	Best validation mean Dice (WT, TC, ET)
Statistical test	Paired t-test on seed-matched runs
Multiple comparisons	Benjamini–Hochberg FDR correction
Compute	≈100 h, NVIDIA RTX 5080

Statistical analysis

Runs sharing a seed were paired, which removes variance common to both variants. For each comparison the mean paired difference, its 95% confidence interval from the t-distribution with 3 degrees of freedom, and the two-sided p-value were computed. Within the architecture group all 10 pairwise comparisons were corrected with the Benjamini–Hochberg procedure [17]. Analyses were repeated separately for WT, TC and ET.

RESULTS

Precision of the design

Seed-to-seed standard deviations ranged from 0.0006 to 0.0018 Dice, well below the 0.0029 noise level of the original single-run ablation, so differences of about 0.003 became detectable.

Architecture variants

The ranking was B_swin_only (0.8645 ± 0.0016) > C_add (0.8625 ± 0.0006) > E_caff (0.8593 ± 0.0015) ≈ A_cnn_only (0.8592 ± 0.0018) > D_concat (0.8581 ± 0.0007) (Fig. 1a). Paired comparisons of the full model with the alternatives are given in Table 3.

Table 3. Full model (E_caff) versus other architectures, paired over 4 seeds

Comparison	Difference	95% CI	p	Conclusion
E_caff – A_cnn_only	+0.0001	[−0.0018; +0.0020]	0.9081	No difference
E_caff – B_swin_only	−0.0052	[−0.0080; −0.0024]	0.0096	E_caff lower
E_caff – C_add	−0.0032	[−0.0050; −0.0014]	0.0107	E_caff lower
E_caff – D_concat	+0.0012	[−0.0011; +0.0034]	0.1932	No difference

After Benjamini–Hochberg correction, six of the ten architecture comparisons remained significant. The deficits of E_caff relative to B_swin_only and C_add persisted (adjusted p = 0.0215 for both). Among the other significant pairs, C_add exceeded D_concat by 0.0044 (adjusted p = 0.0015) and B_swin_only exceeded A_cnn_only by 0.0053 (adjusted p = 0.0059).

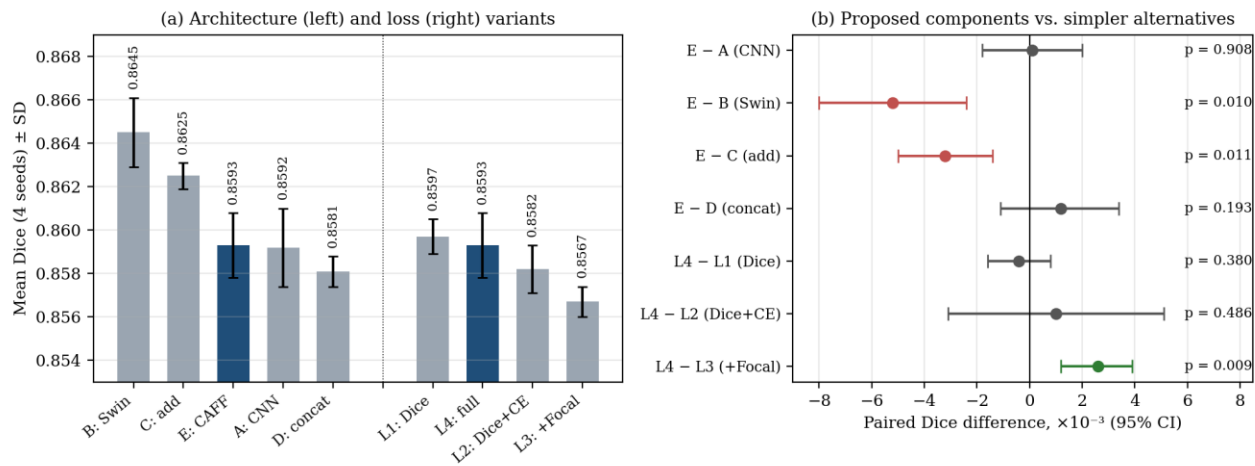


Fig. 1. Mean Dice of all variants (a); paired differences of the proposed components with 95% confidence intervals (b)

Region-wise analysis

The architecture differences were concentrated almost entirely in the enhancing tumor (Table 4). No comparison reached significance for WT or TC.

Table 4. Region-wise paired differences of E_caff (p in parentheses)

Comparison	WT	TC	ET
E_caff - B_swin_only	-0.0031 (0.107)	-0.0041 (0.055)	-0.0084 (0.008)
E_caff - C_add	-0.0009 (0.302)	-0.0027 (0.147)	-0.0061 (0.008)
E_caff - A_cnn_only	+0.0022 (0.060)	-0.0010 (0.571)	-0.0009 (0.064)
E_caff - D_concat	+0.0012 (0.436)	+0.0014 (0.309)	+0.0009 (0.182)

Loss variants

Dice alone gave the highest mean (L1_dice 0.8597 $\pm$ 0.0008), followed by the full compound loss (L4_full 0.8593 $\pm$ 0.0015), Dice + CE (0.8582 $\pm$ 0.0011) and Dice + CE + focal (0.8567 $\pm$ 0.0007). The full loss did not differ from Dice alone (-0.0004, 95% CI [-0.0016; +0.0008], $p = 0.3799$) or from Dice + CE (+0.0010, $p = 0.4859$). Adding the Tversky term to the three-component loss improved the score significantly (+0.0026, 95% CI [+0.0012; +0.0039], $p = 0.0091$), indicating that the focal term had a slightly harmful effect that Tversky partly compensated.

Table 5. Hypothesis testing summary

Hypothesis	Result	Evidence
H1: CAFF hybrid > CNN only	Rejected	+0.0001, $p = 0.908$
H2: CAFF > concatenation	Rejected	+0.0012, $p = 0.193$
H4: compound loss > Dice	Rejected	-0.0004, $p = 0.380$

DISCUSSION

The central result is negative: under a design with adequate power, neither the cross-attention fusion module nor the compound loss improved accuracy over simpler alternatives, and the fusion module was significantly worse than both a single Swin encoder and plain additive fusion. This does not mean that the components are useless in every setting, but it does mean that the gains typically attributed to such components cannot be assumed and must be demonstrated with repeated runs.

A plausible explanation for the fusion result is that attention-based fusion at a single bottleneck level adds parameters and optimization difficulty without adding information: both encoders already see the full input, and a learned mixing of their features tends toward an average of the two rather than a selection of the better one. In a complementary experiment on training-set size, the same hybrid consistently lay

between the pure CNN and pure Swin encoders, which supports this averaging interpretation. The concentration of the deficit in the enhancing tumor, the smallest and most imbalanced region, suggests that the fusion step dilutes fine local detail that the Swin branch alone preserves.

The loss results agree with reports that Dice-based losses are already robust to class imbalance in BraTS [16] and that adding terms can introduce conflicting gradients. The interaction between focal and Tversky terms is a reminder that components of a compound loss should be evaluated jointly rather than assumed to be additive.

The methodological lesson is more general than the specific model. With a single run per variant, the full model could easily have appeared to rank first by chance, since its standard deviation (0.0015) is of the same order as the gaps to its neighbors. Four seeds, pairing on a fixed split and correction for multiple comparisons turned an inconclusive table into clear conclusions at a cost of about 100 GPU-hours, which is modest compared with the cost of an unreproducible claim.

The study has limitations. Four seeds suffice for paired t-tests but not for non-parametric tests; with four pairs the smallest attainable Wilcoxon p-value is 0.125. The ablation used 150 epochs, whereas the main model was trained for 300. Evaluation used the validation split rather than the closed BraTS test set, only one dataset was studied, and self-supervised pre-training, which might benefit the fusion module, was not included.

CONCLUSION

A multi-seed ablation of 32 runs showed that, for 3D brain tumor segmentation on BraTS 2021, the cross-attention fusion module of the HyCTSeg hybrid did not improve accuracy and was significantly inferior to a single Swin Transformer encoder (-0.0052 , $p = 0.0096$) and to additive fusion (-0.0032 , $p = 0.0107$), with the gap concentrated in the enhancing tumor. The four-component loss was not better than Dice loss alone. Hypotheses H1, H2 and H4 were rejected. These results show that single-run ablations can overstate the contribution of architectural components, and they support reporting multiple seeds, paired confidence intervals and corrected p-values as standard practice in medical image segmentation research.

REFERENCES

1. Menze B. H., Jakab A., Bauer S. et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS) // IEEE Transactions on Medical Imaging. – 2015. – Vol. 34, No. 10. – P. 1993–2024.
2. Baid U., Ghodasara S., Mohan S. et al. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification // arXiv:2107.02314. – 2021.
3. Ronneberger O., Fischer P., Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation // MICCAI 2015. – LNCS 9351. – P. 234–241.
4. Çiçek Ö., Abdulkadir A., Lienkamp S. S. et al. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation // MICCAI 2016. – LNCS 9901. – P. 424–432.
5. Isensee F., Jaeger P. F., Kohl S. A. A. et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation // Nature Methods. – 2021. – Vol. 18. – P. 203–211.
6. Wang W., Chen C., Ding M. et al. TransBTS: Multimodal Brain Tumor Segmentation Using Transformer // MICCAI 2021. – LNCS 12901. – P. 109–119.
7. Hatamizadeh A., Nath V., Tang Y. et al. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images // BrainLes 2021. – LNCS 12962. – 2022. – P. 272–284.
8. Zhang Y., Liu H., Hu Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation // MICCAI 2021. – LNCS 12901. – P. 14–24.
9. Bouthillier X., Delaunay P., Bronzi M. et al. Accounting for Variance in Machine Learning Benchmarks // Proceedings of MLSys. – 2021. – Vol. 3. – P. 747–769.

10. Maier-Hein L., Eisenmann M., Reinke A. et al. Why rankings of biomedical image analysis competitions should be interpreted with care // Nature Communications. – 2018. – Vol. 9. – Article 5217.
11. Lipton Z. C., Steinhardt J. Troubling Trends in Machine Learning Scholarship // ACM Queue. – 2019. – Vol. 17, No. 1. – P. 45–77.
12. Isensee F., Wald T., Ulrich C. et al. nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation // MICCAI 2024. – LNCS 15009. – P. 488–498.
13. Milletari F., Navab N., Ahmadi S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation // 3DV. – 2016. – P. 565–571.
14. Lin T.-Y., Goyal P., Girshick R. et al. Focal Loss for Dense Object Detection // IEEE ICCV. – 2017. – P. 2980–2988.
15. Salehi S. S. M., Erdogmus D., Gholipour A. Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks // MLMI 2017. – LNCS 10541. – P. 379–387.
16. Ma J., Chen J., Ng M. et al. Loss odyssey in medical image segmentation // Medical Image Analysis. – 2021. – Vol. 71. – Article 102035.
17. Benjamini Y., Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing // Journal of the Royal Statistical Society: Series B. – 1995. – Vol. 57, No. 1. – P. 289–300.