YOLO-UNet: A Hybrid YOLO-Encoder / U-Net-Decoder Network with a Full-Resolution Residual Boundary Branch for Pediatric Brain Tumor Segmentation

Shaohua Tang[0009-0004-0889-8609] and Young-Jin Jung[0000-0003-0169-2865]

Chonnam National University, Yeosu, Republic of Korea
yj@jnu.ac.kr

Abstract. We present YOLO-UNet, a 3D segmentation network for the BraTS-PEDs 2026 challenge (Task 2) that couples a YOLO-style encoder—a CSPDarknet backbone with a path-aggregation feature pyramid (PAFPN)—to a U-Net-style decoder with skip connections. On top of this backbone we introduce a full-resolution residual boundary branch: a lightweight module that extracts high-frequency features directly from the native-resolution input image and adds a zero-initialized residual correction to the decoder logits, which are otherwise produced at reduced resolution and trilinearly upsampled. Because the correction head is zero-initialized, the model output is identical to the interpolation-limited baseline at initialization and subsequently learns only a residual refinement. We further use a region-selective ensemble of ten models: five full-resolution residual models supply the whole-tumor and tumor-core channels, while five base models supply the enhancing-tumor and cystic channels. On the official validation set (91 cases) our method attains Dice scores of 0.940 (whole tumor), 0.939 (tumor core), 0.910 (non-enhancing tumor core) and 0.577 (enhancing tumor), with a normalized surface Dice of 0.664 for whole tumor. Each individual model is compact at 18.95M parameters, although the submitted system ensembles ten of them under four-view test-time augmentation. Code is available at `https://github.com/tshzg/-yolo-unet-brats-peds`.

Keywords: Pediatric brain tumor · 3D semantic segmentation · YOLO · U-Net · Surface Dice · Full-resolution refinement · BraTS

1 Introduction

Pediatric brain tumors are the leading cause of cancer-related death in children, and accurate delineation of tumor sub-regions from multi-parametric MRI is central to diagnosis, treatment planning, and response assessment. Pediatric high-grade gliomas differ markedly from their adult counterparts in anatomical location, imaging appearance, and the relative prevalence of tumor sub-regions such as enhancing tumor, non-enhancing tumor core, and cystic components.

Consequently, models trained on the large adult glioma cohorts of previous BraTS challenges transfer poorly to the pediatric population [4,5]. The BraTS-PEDs 2026 challenge addresses this gap with a standardized, expertly annotated multi-parametric MRI benchmark and a common evaluation protocol [4,5].

This benchmark evaluates two complementary metric families: volumetric overlap (Dice) and surface accuracy (normalized surface Dice, NSD), which capture different failure modes. Dice rewards volumetric agreement but is largely insensitive to small, systematic boundary errors; NSD is evaluated at a strict surface tolerance and directly penalizes sub-voxel boundary displacement that both Dice and the 95th-percentile Hausdorff distance largely ignore. In practice, whole-tumor and tumor-core predictions reach near-ceiling Dice across the leaderboard, so the metric that discriminates methods is surface accuracy.

Encoder–decoder networks with skip connections, exemplified by U-Net [8] and its self-configuring extension nnU-Net [9], remain the dominant paradigm for volumetric medical image segmentation, and object-detection backbones from the YOLO family [7] provide strong, efficient multi-scale feature extractors. A recurring difficulty when combining such backbones with dense segmentation is that the decoder predicts at reduced spatial resolution and relies on band-limited interpolation to recover full resolution, which cannot represent the sub-voxel detail on which surface accuracy depends. Our analysis in Section 3.5 refines this motivation: interpolation accounts for only about 0.1 voxel of localization error, and the larger share of the surface gap is genuine boundary uncertainty. The branch we propose is therefore motivated less by replacing the interpolation than by supplying native-resolution evidence that reduces that uncertainty.

Recovering full-resolution detail in segmentation networks has been approached in several ways: full-resolution residual streams that carry a native-resolution path alongside the downsampling encoder [13], point-based refinement of uncertain boundary locations [14], and post-hoc boundary-correction modules applied to an existing prediction [15]. Our branch differs in being a zero-initialized additive correction to the final logits, which requires no change to the underlying decoder and cannot degrade it at initialization.

Our contributions are:

- **YOLO-UNet**, a hybrid encoder–decoder that pairs a CSPDarknet+PAFPN encoder with a U-Net decoder for 3D tumor segmentation.
- **A full-resolution residual boundary branch** with zero-initialized fusion, which is non-degrading at initialization by construction and empirically improves whole-tumor and tumor-core accuracy.
- **A zero-initialized fusion scheme** that keeps the baseline prediction path intact, so that full-resolution information is added as a correction rather than substituted for the decoder output.
- **A region-selective ensemble** that assigns each tumor sub-region to the model family that segments it best, with the per-family comparison that motivates the assignment reported in Section 3.3.

2 Methods

The challenge provides, for each case, four co-registered MRI modalities and a voxel-wise label map with four foreground classes: enhancing tumor (ET), non-enhancing tumor core (NETC), cystic component (CC), and peritumoral edema (ED). Evaluation is performed on overlapping, nested *regions*: whole tumor (WT), tumor core (TC), and enhancing tumor (ET), together with the cystic (CC) and edema (ED) classes. WT comprises all tumor tissue, TC comprises the enhancing tumor with the non-enhancing core and cystic component, and the enhancing region is ET itself, giving the nesting $\mathrm{ET}, \mathrm{CC} \subset \mathrm{TC} \subset \mathrm{WT}$. We therefore predict these regions in an overlapping multi-label formulation with an independent sigmoid per region channel, rather than as mutually exclusive classes, so that the training objective is aligned with the evaluation regions; the NETC class is recovered as TC minus ET and CC.

2.1 Overview and Architecture

Our network follows an encoder–decoder design. The encoder is a YOLO-style feature extractor: a 3D CSPDarknet backbone, a path-aggregation feature pyramid (PAFPN) with top-down and bottom-up fusion, and a self-attention block (C2PSA) on the deepest level [7]. For a 128^3 input the backbone produces features at strides 2–32; the PAFPN levels P3, P4, P5 operate at strides 8, 16, 32 (spatial sizes 16^3, 8^3, 4^3). The decoder is U-Net-style, progressively upsampling and fusing skip connections down to the stride-2 "stem" feature before predicting region logits. All blocks use InstanceNorm, suited to the batch-size-1 regime of 3D patch training [9]. We term this hybrid *YOLO-UNet*; the model has 18.95M parameters.

2.2 Full-Resolution Residual Boundary Branch

In the baseline decoder, region logits are computed on the stride-2 grid (64^3 for a 128^3 input) by a $1 \times 1 \times 1$ head and then trilinearly upsampled to full resolution. Trilinear interpolation is band-limited: it cannot represent sub-voxel boundary detail that is not already present on the coarse grid.

To recover this detail we add a branch that operates on the native-resolution input image $x \in \mathbb{R}^{4\times 128^3}$—the only tensor in the network that has never been downsampled. Let f be the final decoder feature (64^3) and $\mathrm{pred}(\cdot)$ the baseline $1 \times 1 \times 1$ logit head. The branch computes an upsampled semantic path and a native-resolution detail path, fuses them, and predicts a residual correction Δ:

$$f^{\uparrow} = \mathrm{Up}_{128}(f), \qquad h = E_{fr}(x), \tag{1}$$

$$\Delta = P_{fr}\big(\mathrm{Fuse}([\, f^{\uparrow}; h\,])\big), \tag{2}$$

$$\ell = \mathrm{Up}_{128}\big(\mathrm{pred}(f)\big) + \Delta, \tag{3}$$

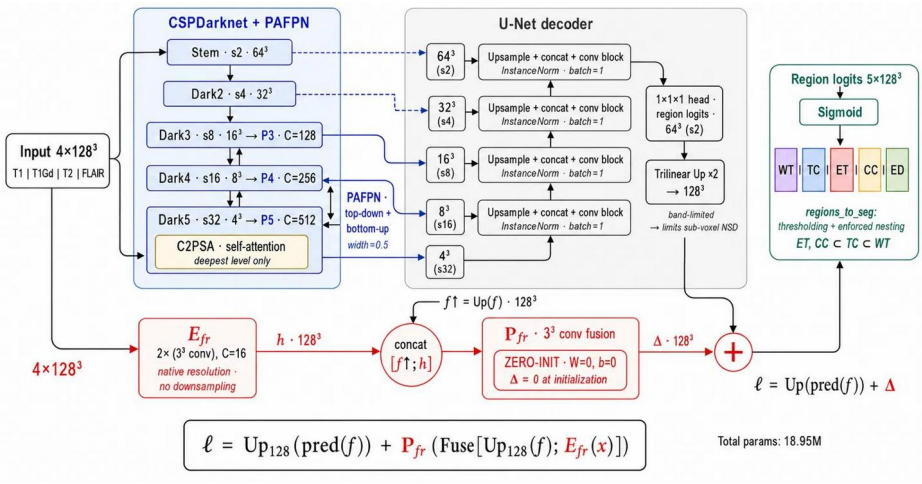

Fig. 1. Overall architecture of YOLO-UNet. A YOLO-style encoder (CSPDarknet + PAFPN, blue) feeds a U-Net-style decoder (grey) whose region logits are produced at stride 2 and trilinearly upsampled. The full-resolution residual branch (red) extracts native-resolution features $E_{fr}(x)$ from the input image and predicts a zero-initialized residual correction Δ that is added to the upsampled baseline logits; the reconciliation step then applies thresholding with enforced nesting ($\mathrm{ET}, \mathrm{CC} \subset \mathrm{TC} \subset \mathrm{WT}$).

where E_{fr} consists of two 3^3 convolutions that stay at 128^3 throughout. Crucially, P_{fr} is zero-initialized ($W = 0$, $b = 0$), so $\Delta \equiv 0$ at initialization and ℓ is bit-identical to the baseline. Training can only add a learned correction on top of the baseline; the band-limited path is preserved as an implicit low-pass prior—important on a dataset of only a few hundred independent training volumes—while the branch learns high-frequency boundary corrections. This design is inspired by zero-convolution residual connections [10,11]. The branch adds only 0.30M parameters (the $3 \times 3 \times 3$ fusion dominates) and about 30% inference time.

Why residual, not replacement. An alternative is to let the branch replace the baseline head ($\ell = \Delta$, random initialization). Replacement bypasses the direct gradient path to the decoder and discards the low-pass prior, and it removes the non-degradation property at initialization. We adopt the residual form for these structural reasons; Section 3.2 reports what we observed when the two were compared during development.

2.3 Region-Selective Ensemble

The task is evaluated on regions {WT, TC, ET, CC, ED} derived from label channels (ET, NETC, CC, ED). We observe that the two model variants favor different regions: models equipped with the full-resolution residual branch and

trained with an additional region-restricted Lovász–IoU objective give the more accurate whole-tumor and tumor-core predictions, whereas the base models retain stronger enhancing-tumor predictions. The cystic component is segmented equally well by both families and is grouped with ET for convenience. The supporting per-family comparison is reported in Section 3.3.

We therefore build a region-selective ensemble over two model families, each a five-fold set: five full-resolution residual models supply the WT, TC and ED channels, and five base models supply the ET and CC channels—ten models in total. For each family, sigmoid probabilities are averaged across its five folds and across test-time flip augmentations; the two families are then combined per region, and the final labeling enforces the nesting $\mathrm{ET}, \mathrm{CC} \subset \mathrm{TC} \subset \mathrm{WT}$ (with NETC obtained as TC minus ET and CC). The fusion does not retrain a meta-model and therefore introduces no additional fitted parameters beyond selecting the better-performing family for each region.

2.4 Training and Inference

Data and preprocessing. We use the four co-registered, skull-stripped mpMRI modalities (T1, T1-Gd, T2, T2-FLAIR) at 1 mm isotropic resolution as a 4-channel input. The training set of 296 cases is split into five folds (237 training, 59 validation per fold); oversampling of cystic-component ($\times 2$) and edema ($\times 1$) cases raises the sampling entries per epoch to 445 (an oversampled draw count, not a count of independent volumes). Each modality is intensity-normalized independently over the non-zero (brain) region, with background held at zero.

Augmentation and sampling. We apply on-the-fly random flips, 90° rotations, scaling, elastic deformation, and intensity augmentations (gamma, additive noise, bias-field simulation, per-channel perturbation). Training uses 128^3 patches at batch size 1, with foreground-biased cropping (0.55), purely negative patches (0.05), and class-aware cropping targeting ET, NETC, CC and ED.

Loss. The training objective is a region-weighted sum of four terms: a binary cross-entropy term, a Tversky term, a per-region boundary term (weight 0.25), and a region-restricted Lovász–IoU term (weight 0.5)—a tractable surrogate for the Jaccard index [12]—applied to the whole-tumor and tumor-core channels. The Tversky term follows the formulation of Salehi et al.,

$$\mathrm{TI} = \frac{\mathrm{TP}}{\mathrm{TP} + \alpha\,\mathrm{FP} + \beta\,\mathrm{FN}}, \tag{4}$$

so that $\alpha = 0.45$ and $\beta = 0.55$ penalize false negatives more heavily and bias the objective toward recall on under-segmented regions. The boundary term is a per-region surface loss that penalizes disagreement in a narrow band around the reference boundary, so that errors near object edges are weighted more heavily than errors in region interiors. The five region channels (WT, TC, ET, CC, ED) are weighted (1.1, 1.15, 1.6, 1.0, 0.5), placing greater emphasis on the smaller

enhancing and core sub-regions, and the boundary term uses per-region weights (1.0, 1.2, 1.6, 0.8, 0.0). Deep supervision is applied at intermediate decoder resolutions with weight 0.20. All losses are computed in the region (overlapping-label) space so that the WT, TC and ET objectives are optimized directly.

Optimization. We use AdamW (learning rate 2×10^{-4}, weight decay 5×10^{-5}) with a cosine-annealing schedule (10-epoch linear warm-up, decaying to 10^{-3} of the peak learning rate) for 250 epochs, with gradient-norm clipping at 1.0. An exponential moving average (EMA) of the weights (decay 0.9995, 2000 warm-up steps) is maintained.

Checkpoint selection. We maintain both raw and EMA weights and periodically evaluate on the held-out fold using a composite focus metric—the mean of per-region Dice and NSD over the whole-tumor, tumor-core and enhancing-tumor regions that determine the ranking. For each fold we retain the single checkpoint (base or EMA) with the highest focus score; the selected epoch varies across folds. The two families are selected independently under the same rule, each contributing five per-fold checkpoints to the ensemble.

2.5 Inference Configuration

Predictions use sliding-window inference with a 128^3 window, a Gaussian importance map, and four-view flip test-time augmentation (original volume plus single-axis flips); each flipped prediction is restored to the reference orientation before aggregation, so the augmentation introduces no interpolation. Sigmoid probabilities are averaged across the four views and the five folds of each family, so that consistent sub-threshold evidence accumulates before the final region-specific threshold—which matters for the small ET and CC components supported by only part of the ensemble. Both families run at overlap 0.50 (raising the base-family overlap to 0.70 gave no net gain).

After fusion, fixed per-region thresholds are applied and the masks reconciled with the anatomical hierarchy: WT defines tumor support; TC is intersected with WT; ET and CC with TC; NETC is TC not claimed by ET or CC; ED is WT outside TC. Where ET and CC overlap, ET is written last and overrides CC. Post-processing is minimal: ET components below 50 voxels and CC components below 30 voxels are removed (the largest-component-only option is disabled, as pediatric tumors may be multifocal), with no size filtering on edema. Output is written in the reference NIfTI geometry. The thresholds and ET filter were chosen primarily on the internal folds, with a few settings checked against official-validation feedback; Table 1 summarizes the final configuration.

2.6 Implementation Details

Models are implemented in PyTorch and trained with mixed use of a single NVIDIA RTX 4080 Super (16 GB) and, for some folds, an NVIDIA A100 (80 GB); all reported inference is performed on the RTX 4080 Super. Each fold is trained

Table 1. Final inference and fusion configuration used for the official validation submission. Region order is WT, TC, ET, CC, ED.

Component	Final setting	Purpose
Model families	5 base + 5 full-res/Lovász	Fold ensemble and region specialization
Patch / overlap	128^3; 0.50 (both families)	Volumetric sliding-window coverage
Window fusion	Gaussian weighting	Suppress patch-border discontinuities
Test-time augmentation	4 flip views	Average orientation-consistent probabilities
Region source	Base: ET, CC; full-res: WT, TC, ED	Use the stronger family per region
Thresholds	0.50, 0.50, 0.45, 0.50, 0.60	Convert fused probabilities to masks
Post-processing	Remove ET components < 50 voxels; CC components < 30 voxels	Reduce isolated false positives
ED filtering	None	Preserve genuinely small components

for 250 epochs. At inference, the complete pipeline—ten models with four-view test-time augmentation—processes the 91-case validation set in approximately 1.4 hours on the RTX 4080 Super. Each model has 18.95M parameters, of which the full-resolution residual branch contributes 0.30M; the base models therefore have 18.65M parameters. The source code for the inference pipeline is available at `https://github.com/tshzg/-yolo-unet-brats-peds`.

3 Results

3.1 Evaluation Protocol

We report two sets of results that are *not directly comparable* and must be read separately. Table 2 reports internal five-fold cross-validation: single models, per-case Dice and NSD averaged over each held-out fold, surface tolerance 1.0 voxel, raw output with no post-processing. Table 4 reports the official Synapse evaluation of the submitted ten-model ensemble on the 91-case validation set, using the organizer-defined global DSC and NSD (`global_dsc_*` / `global_nsd_*`) fields on which BraTS-PEDs is ranked, after the full post-processing of Section 2.5. The two tables differ in cohort, model aggregation, post-processing, metric definition and surface tolerance simultaneously; the difference in absolute values reflects these protocol differences and is not a measure of generalization.

3.2 Ablation: The Full-Resolution Residual Branch

Table 2 compares the baseline decoder against the same decoder augmented with the full-resolution residual branch, across all five folds. Averaged over the

Table 2. Ablation of the full-resolution residual branch across all five folds (internal cross-validation, per-case metrics, NSD tolerance 1.0, best epoch per fold, raw model output).

Configuration	Fold	WT DSC	WT NSD	TC DSC	TC NSD
Baseline	0	0.831	0.760	0.838	0.738
	1	0.793	0.707	0.774	0.681
	2	0.824	0.734	0.814	0.713
	3	0.759	0.703	0.755	0.692
	4	0.757	0.670	0.736	0.638
	mean	0.793	0.715	0.783	0.692
+ full-res residual	0	0.862	0.788	0.845	0.750
	1	0.804	0.723	0.801	0.711
	2	0.828	0.738	0.817	0.715
	3	0.768	0.680	0.745	0.648
	4	0.774	0.685	0.752	0.654
	mean	0.807	0.723	0.792	0.696

five folds, the residual branch improves whole-tumor surface accuracy (WT NSD $0.715 \rightarrow 0.723$) and tumor-core surface accuracy (TC NSD $0.692 \rightarrow 0.696$), with parallel gains in Dice; the branch helps in four of the five folds. In the one fold where it does not help, performance drops modestly rather than catastrophically—a direct consequence of the zero-initialized design, which guarantees the model starts from, and can fall back toward, the baseline.

Scope of the comparison. We note a limitation of this comparison. The residual configuration in Table 2 was trained with the region-restricted Lovász term active, whereas the baseline was not, so the two rows differ in both the architectural branch and one loss term. The reported deltas therefore characterize the joint configuration that was submitted, and we do not attribute them to the branch in isolation. Separating the two factors requires a matched-loss run that we were unable to complete within the challenge timeline, and we identify it as the primary open question for this component.

Effect size. These effects should be read with care. The per-fold TC NSD differences are $+0.012$, $+0.030$, $+0.002$, -0.044 and $+0.016$, so the mean gain of $+0.003$ is smaller than the fold-to-fold spread, and five fold-level means do not support an uncertainty estimate for an effect this size. The Dice differences are more consistent than the surface ones (WT Dice improves on all five folds, WT NSD on four), so we state the surface claim as the weaker of the two; a paired, case-level analysis respecting fold membership is left to future work.

Residual versus replacement fusion. We also examined a replacement variant in which the branch supplies the region logits directly ($\ell = \Delta$, randomly initialized). In preliminary two-fold experiments it improved one fold and degraded the other

Table 3. Per-family out-of-fold performance on the regions that determine the fusion rule (five-fold means, internal cross-validation protocol).

Family	ET DSC	ET NSD	CC DSC	CC NSD	WT NSD	TC NSD
Base	**0.531**	**0.581**	0.433	0.451	0.715	0.692
Full-res / Lovász	0.511	0.560	**0.433**	**0.455**	**0.723**	**0.696**

Table 4. Official validation-set results (91 cases). See Section 3.1 for the metric definition and surface tolerance.

	ET	NETC	CC	ED	TC	WT
Dice	0.577	0.910	0.202	0.000	0.939	0.940
NSD	0.430	0.635	0.211	0.000	0.663	0.664

by a larger margin than any regression seen for residual fusion. As this was not extended to all folds and the predictions were not retained, we report it only qualitatively: our reason for adopting the residual form is primarily structural—preserving the baseline at initialization and keeping a direct gradient path to the decoder—rather than an established margin over replacement.

3.3 Per-Family Comparison Underlying the Fusion Rule

The region-selective assignment of Section 2.3 rests on the two families performing differently on different regions. Table 3 reports this comparison directly, as five-fold out-of-fold means under the protocol of Table 2.

The separation is clearest for the two extremes. The base family is better on enhancing tumor by 0.020 DSC and 0.021 NSD, and the full-resolution/Lovász family is better on whole tumor and tumor core. The cystic component, by contrast, is a tie: the two families score identically to three decimal places. Its assignment to the base family therefore carries no measured advantage, and we group it with ET rather than claiming an empirical basis for that choice. None of these margins is large relative to the fold-to-fold spread reported in Table 2, so the assignment should be read as a reasonable allocation rather than a decisively supported one.

3.4 Official Validation-Set Performance

Table 4 reports our final ten-model region-selective ensemble on the official validation set (91 cases). Whole- and tumor-core Dice reach 0.940 and 0.939 with NSD 0.664 and 0.663; enhancing tumor reaches Dice 0.577.

Edema. The submitted system obtains an ED score of 0.000. Edema was deliberately de-emphasized: it carries the lowest region weight (0.5), a boundary-term weight of zero, and no size filtering. Under the official metric a case with

both reference and prediction empty is excluded from the mean, whereas an empty reference with a non-empty prediction contributes zero; the 0.000 mean therefore reflects the few cases where our pipeline predicted a small amount of edema against an empty reference, not a uniform absence of edema. This is a design trade-off—our configuration was tuned for the regions that dominate the ranking—and a region-specific treatment of edema is the clearest avenue for improvement.

3.5 Surface-Metric Analysis

To characterize the NSD gap we analyze all 91 whole-tumor cases on the validation set (Fig. 2). Whole-tumor Dice is high (mean 0.940) yet WT NSD is markedly lower (mean 0.664, median 0.730), and the distribution is left-skewed: a small tail of hard cases (7% with NSD < 0.3) pulls the mean below the median. Crucially, even among the 72 cases with excellent volumetric overlap (DSC $\geq$ 0.94), the mean WT NSD is only 0.743 and ranges from 0.44 to 0.88.

High volumetric agreement therefore does not by itself imply high surface accuracy. We interpret the residual gap as predominantly one of boundary precision. This interpretation is supported by the instance-level statistics returned by the evaluator: for whole tumor the mean instance-wise F1 is 0.924 with a false-positive rate of 0.011, indicating that spurious or missed lesion instances are rare and that the surface gap is not explained by a detection failure but by sub-voxel boundary displacement.

We further verified, by binning the trained logits against signed distance to the boundary, that they are near-affine (slope ≈ 1.7 for WT), so trilinear upsampling costs only ≈ 0.1 voxel of localization error and the remaining error is genuine boundary uncertainty—motivating a branch that reduces that uncertainty using native-resolution features rather than one that merely changes the interpolation.

3.6 Qualitative Results

Fig. 3 shows representative segmentations spanning a range of difficulty: a large multi-region tumor with enhancing tumor, non-enhancing core and edema (top); a small, compact core lesion (middle); and a small peripheral lesion (bottom). Across all three, the predicted sub-region layout and boundaries closely follow the ground truth, including the fine boundaries of the compact lesion where surface accuracy is most demanding.

4 Discussion

Our results point to a specific characterization of the pediatric segmentation problem under the challenge's evaluation protocol. Across the leaderboard, whole-tumor and tumor-core Dice are near saturation, so volumetric overlap is no longer the discriminating axis; the axis that separates methods is surface accuracy. We

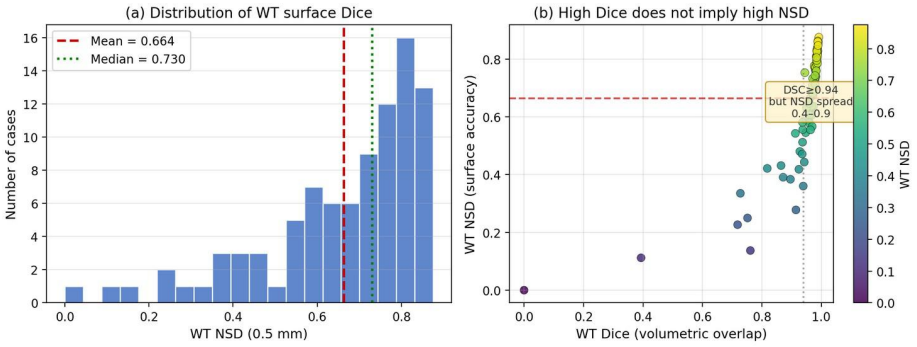

Fig. 2. Surface-metric analysis on the validation set (91 cases). (a) Distribution of WT NSD: mean 0.664, median 0.730, with a left-skewed tail of hard cases. (b) WT Dice vs. WT NSD: even cases with Dice ≥ 0.94 show WT NSD spread across 0.44–0.88, confirming that high volumetric overlap does not guarantee sub-voxel surface accuracy. Surface tolerance as stated in Section 3.1.

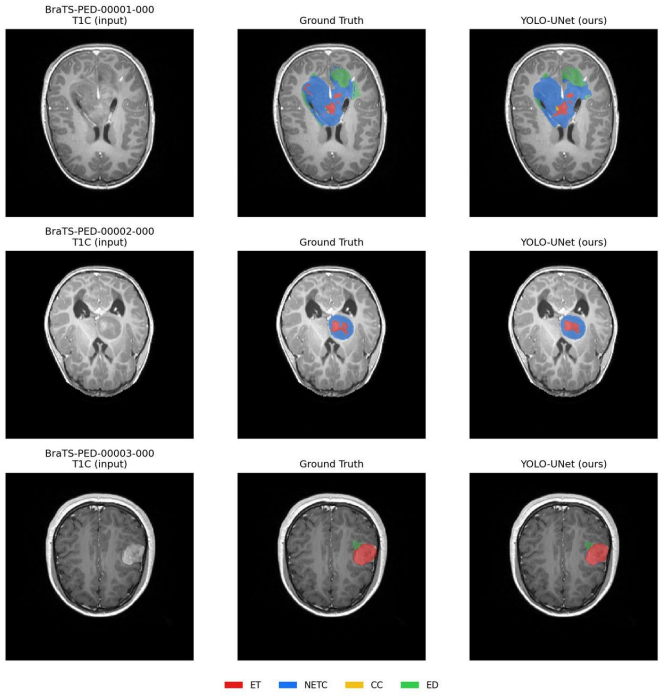

Fig. 3. Qualitative results on three representative cases. Columns: input (T1-Gd), ground truth, and YOLO-UNet prediction. Colors: ET (red), NETC (blue), CC (yellow), ED (green).

therefore designed a module aimed at this axis rather than at overall detection or overlap, which are already strong in our baseline.

Non-degradation at initialization. A recurring risk when adding capacity to a segmentation network on a small dataset is that the addition helps in some folds and hurts in others. The zero-initialized residual formulation (3) removes the most damaging failure mode: *at initialization* the branch output is identically zero, so the model is bit-identical to the baseline and can only add a learned correction. This guarantee applies to the initialization only—it constrains neither the converged solution nor the magnitude of a regression such as the one on fold 3. The band-limited interpolation path is retained as an implicit low-pass prior, valuable when only a few hundred training volumes are available, in contrast to a replacement head, which discards that prior and severs the direct gradient path to the decoder.

The fusion scheme. Given the same native-resolution input and the same capacity, *how* the high-resolution signal is fused determines what guarantees the addition carries: the residual form retains the baseline as the starting point and leaves the decoder's gradient path untouched, while the replacement form does neither. We did not quantify this across all folds and present it as a design argument rather than an empirical result, but it is why we would recommend, when coupling detection-style (YOLO) encoders with boundary-accurate segmentation, adding high-resolution paths as residual corrections rather than replacements.

Negative results. We report two lines that did not yield gains, to save others the effort. First, refining features that had already been downsampled—interpolating a coarse feature map to full resolution and applying further convolutions—did not improve NSD, because interpolation introduces no new high-frequency information. Second, reparameterizing the output as a signed-distance field, or linearizing the logits with respect to distance, produced negligible gains: by binning the trained logits against signed distance we found them already near-affine, so trilinear upsampling costs only ≈ 0.1 voxel of boundary localization. The residual surface error is therefore genuine boundary uncertainty—arising from annotation ambiguity and model error—rather than an interpolation artifact.

Efficiency. Two notions of cost should be distinguished. Each individual model uses 18.95M parameters (the residual branch contributing 0.30M), compact relative to full-resolution baselines of comparable accuracy. The submitted system, however, is an ensemble of ten models under four flip views—40 forward passes per volume—taking approximately 1.4 hours for the 91-case validation set on a single RTX 4080 Super. The compactness claim therefore applies to the architecture, not the ensemble; for tight compute budgets, a single residual model retains the architectural benefit at a fraction of the cost, forgoing the ensembling and test-time augmentation gains.

Generality. Although developed for pediatric tumor segmentation, the full-resolution residual branch is agnostic to the encoder, decoder and target anatomy: it requires only a final decoder feature and the native-resolution input, and can attach to any semantic-segmentation head. The zero-initialization property means it can be added to an already-trained baseline and fine-tuned without degrading the starting point, making it a low-cost, drop-in refinement for other resolution-sensitive segmentation tasks.

Limitations. Our conclusions rest on a comparatively small validation set with fold-to-fold variance, so single-fold comparisons should be read with caution, and five fold-level means do not support an uncertainty estimate for an effect as small as the branch's. The YOLO-style encoder is not ablated against a conventional U-Net or nnU-Net encoder under matched training, and the branch and the Lovász objective were introduced together in the submitted configuration, so these contributions cannot be separated from our results. How much of the validation-set gain transfers to withheld test data remains to be confirmed. Finally, edema and cystic components are far more variable in this pediatric cohort than in adult glioma, and a principled region-specific treatment of them is left to future work.

5 Conclusion

We presented YOLO-UNet, a hybrid YOLO-encoder / U-Net-decoder network for pediatric brain tumor segmentation, together with a lightweight full-resolution residual boundary branch whose zero-initialized fusion cannot degrade the baseline at initialization. On the official BraTS-PEDs 2026 validation set a ten-model region-selective ensemble reaches 0.940 Dice and 0.664 NSD for whole tumor. Our analysis indicates that volumetric overlap is near saturation for the composite regions and the discriminating axis is surface accuracy, and that high Dice does not imply high NSD even on volumetrically well-segmented cases. The principal open questions are separating the branch from the accompanying objective, a case-level paired analysis with proper uncertainty estimation, and a region-specific treatment of edema and cystic components.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00410364). Data used in this publication were obtained as part of the Challenge project through Synapse ID (syn74274097).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015)

2. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., et al.: Advancing the Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**, 170117 (2017)
3. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., et al.: The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. arXiv:2107.02314 (2021)
4. Kazerooni, A.F., Khalili, N., Liu, X., Haldar, D., Jiang, Z., Anwar, S.M., et al.: The Brain Tumor Segmentation (BraTS) Challenge 2023: Focus on Pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). arXiv:2305.17033 (2023). https://doi.org/10.48550/arXiv.2305.17033
5. Kazerooni, A.F., Khalili, N., Liu, X., Gandhi, D., Jiang, Z., Anwar, S.M., et al.: The Brain Tumor Segmentation in Pediatrics (BraTS-PEDs) Challenge: Focus on Pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). arXiv:2404.15009 (2024). https://doi.org/10.48550/arXiv.2404.15009
6. Karargyris, A., Umeton, R., Sheller, M.J., Aristizabal, A., George, J., Wuest, A., et al.: Federated benchmarking of medical artificial intelligence with MedPerf. Nat. Mach. Intell. **5**(7), 799–810 (2023)
7. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO (2023). https://github.com/ultralytics/ultralytics
8. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: MICCAI 2015, LNCS, vol. 9351, pp. 234–241 (2015)
9. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: IEEE CVPR, pp. 770–778 (2016)
11. Zhang, L., Rao, A., Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models. In: IEEE/CVF ICCV, pp. 3836–3847 (2023)
12. Berman, M., Rannen Triki, A., Blaschko, M.B.: The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: IEEE CVPR, pp. 4413–4421 (2018)
13. Pohlen, T., Hermans, A., Mathias, M., Leibe, B.: Full-Resolution Residual Networks for Semantic Segmentation in Street Scenes. In: IEEE CVPR, pp. 3309–3318 (2017)
14. Kirillov, A., Wu, Y., He, K., Girshick, R.: PointRend: Image Segmentation as Rendering. In: IEEE/CVF CVPR, pp. 9799–9808 (2020)
15. Yuan, Y., Xie, J., Chen, X., Wang, J.: SegFix: Model-Agnostic Boundary Refinement for Segmentation. In: ECCV 2020, LNCS, vol. 12357, pp. 489–506 (2020)
16. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. In: Machine Learning in Medical Imaging, LNCS, vol. 10541, pp. 379–387 (2017)