Adapting PTransBTS for Pediatric Brain Tumor Segmentation in BraTS-PEDs 2026

Xiangchen Hou[1], Yanjun Peng[1*], and Haitao Yu[1]

College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, Shandong, China
houxiangchen@sdust.edu.cn; yjpeng@sdust.edu.cn; yuhaitaocn@foxmail.com

Abstract. Accurate segmentation of pediatric brain tumor subregions in multi-parametric magnetic resonance imaging is challenging because tumor appearance varies across cases and small enhancing tumor (ET) and cystic component (CC) regions may contain few voxels. We present a task-specific adaptation of PTransBTS for BraTS-PEDs 2026 Task 2. The submitted configuration combines modality-specific stems, a convolution-only backbone path, inherited attention-based pyramid priors, five-class prediction, Dice–Focal deep supervision, and extensive spatial and intensity augmentation. At inference, logits are averaged across eight flip views within each fold model, followed by softmax and probability averaging across five models; threshold-based class-volume filtering and geometry restoration produce the final NIfTI segmentation. On the official Synapse validation set (submission 9772538), global DSCs were 0.5096, 0.9085, 0.2196, 0.0090, 0.9400, and 0.9415 for ET, NET, CC, ED, TC, and WT, respectively. These results show substantially higher overlap for composite regions than for fine-grained classes, with ED and CC remaining the main limitations. Source code is available at https://github.com/hxc236/PTransBTS-PED.

Keywords: Pediatric brain tumor · MRI segmentation · PTransBTS · Deep supervision · Model ensemble

1 Introduction

Automated segmentation of pediatric brain tumors supports quantitative assessment of heterogeneous tumor tissue while reducing the burden and inter-reader variability of manual delineation. The Brain Tumor Segmentation in Pediatrics (BraTS-PEDs) initiative provides expert annotations, standardized labels, and centralized evaluation for this purpose [4,5,6]. In BraTS-PEDs 2026 Task 2, a system receives co-registered native T1-weighted (T1N), contrast-enhanced T1-weighted (T1C), T2-weighted (T2W), and T2-weighted fluid-attenuated inversion recovery (T2F) images. It must identify enhancing tumor (ET), non-enhancing tumor (NET), cystic component (CC), and peritumoral edema (ED); tumor core (TC) and whole tumor (WT) are subsequently derived for evaluation.

* Corresponding author.

Established three-dimensional U-shaped networks remain strong choices for BraTS tasks [2]. Challenge reports based on nnU-Net and model ensembles show that target definitions, augmentation, ensembling, post-processing, and input–output reliability can be as consequential as changes to the backbone [3,1,7]. This observation is particularly relevant to pediatric data: small ET or CC regions are strongly imbalanced, appearances vary across tumor types and treatment states, and the submitted algorithm must preserve spatial metadata while operating without network access.

Our goal is therefore not to introduce another architectural block, but to turn an existing BraTS method into a Task 2 submission. We use PTransBTS [8], a BraTS-MEN 2025 method, as the starting point. We retain its modality-aware input processing, U-shaped backbone, and learnable pyramid priors, while adapting the model to four pediatric labels, five-class training, and the BraTS-PEDs submission contract. The final hyperparameter configuration assigns all feature channels to the convolutional path; consequently, no Transformer channel is active.

This adaptation involves three practical challenges. First, the four tumor labels differ substantially in prevalence and may be very small, requiring multi-scale supervision and strong data perturbation. Second, predictions must remain stable across independently trained folds and spatial orientations. Third, the pipeline must reproduce preprocessing, output geometry, and label conventions exactly. We address these challenges through a deeply supervised training recipe, within-model logit averaging across flip views, cross-model probability averaging, and an offline inference pipeline.

The contributions of this work are:

- a task-specific PTransBTS configuration with modality-specific stems, a convolution-only backbone path, inherited attention-based pyramid priors, and five-class BraTS-PEDs output heads (Sect. 3);
- a task-specific training configuration combining deep supervision, Dice–Focal loss, extensive augmentation, and five-fold model training (Sect. 4); and
- a deployment-oriented inference pipeline using five-model probability ensembling after eight-view flip-based test-time augmentation (TTA) with logit averaging, a fixed class-volume heuristic, and geometry-preserving output restoration (Sect. 4.3).

These are engineering contributions to the final submission pipeline; no individual component is claimed as architecturally novel.

2 Related Work

The nnU-Net framework established a strong reference point for three-dimensional medical image segmentation by configuring preprocessing, network topology, training, and post-processing for each dataset [2]. Its BraTS 2020 variant further showed that region definitions, stronger augmentation, post-processing, and model ensembling can materially affect a challenge submission even when the

Table 1. PTransBTS components and the settings used for BraTS-PEDs 2026 Task 2.

Component	Inherited role	Task 2 configuration
Input processing	Modality-aware feature extraction	Four independent 3D stems for T1N, T1C, T2W, and T2F
Backbone	Multi-resolution encoder–decoder	Six resolutions with channels $(24, 64, 128, 256, 320, 320)$
Parallel paths	Convolutional and Transformer backbone channels	Convolution proportion 1; backbone Transformer channel count 0
Pyramid prior	Coarse-to-fine guidance of decoder features	Learnable $48 \times 4 \times 4 \times 4$ prior refined at four scales
Output	Task-dependent segmentation heads	Five heads for background, ET, NET, CC, and ED

underlying framework is unchanged [3]. Recent BraTS systems continue this pipeline-oriented approach. Capellán-Martín et al. combined predictions from complementary segmentation models and used region-specific post-processing for pediatric and other brain tumor tasks [1], while Li et al. adapted an nnU-Net-based framework to BraTS-PED 2025 [7].

The BraTS-PEDs challenge papers define the pediatric imaging setting and its standardized subregion evaluation [4,5,6]. Their reported challenge outcomes also place ensembles among the commonly used approaches for pediatric tumor segmentation. PTransBTS provides a different starting point by combining modality-aware processing, parallel convolutional and self-attention paths, and learnable pyramid priors [8]. Our work evaluates a task-specific configuration of that model under the 2026 pediatric protocol. The trained configuration allocates all backbone channels to the convolutional path; our positioning is therefore an adaptation and evaluation of the complete submission pipeline rather than a claim about a new attention architecture.

3 PTransBTS-Based Task Adaptation

3.1 Baseline and adaptation scope

PTransBTS combines modality-aware processing, convolutional and self-attention paths, and learnable priors [8]. We adapt PTransBTS rather than redesigning the network. Table 1 distinguishes inherited components from Task 2 settings. The input order is T1N, T1C, T2W, and T2F. Five mutually exclusive output channels represent background, ET, NET, CC, and ED. TC and WT are not predicted independently, as defined in Equation 1:

$$\mathrm{TC} = \mathrm{ET} \cup \mathrm{NET} \cup \mathrm{CC}, \qquad \mathrm{WT} = \mathrm{TC} \cup \mathrm{ED}. \tag{1}$$

3.2 Configured network

The model contains 38,711,460 trainable parameters. Each modality-specific stem applies two 3D convolution–instance-normalization–LeakyReLU blocks. The four feature maps are concatenated and projected using a $1 \times 1 \times 1$ convolution. Five downsampling operations produce a six-resolution representation; transposed convolutions restore the spatial resolution in the decoder.

The architecture allocates channels between parallel convolutional and self-attention paths using a proportion hyperparameter. For all trained checkpoints, the convolution proportion is one; the complementary Transformer proportion is converted to an integer channel count of zero at every stage. Given the modest training cohort and the voxel-wise nature of the task, we favored this convolution-only backbone path because 3D convolutions provide strong locality and translation-equivariant inductive biases for learning tumor boundaries, regional texture, and spatial continuity. Multiscale convolutional U-shaped networks also remain strong reference systems for data-limited biomedical segmentation [2,3]. The inherited attention-based pyramid-prior module was retained to provide coarse-to-fine contextual guidance. This fixed configuration was used consistently across all five folds; however, no Transformer-active alternative was evaluated, so the results do not establish a performance advantage for this configuration over the hybrid parent configuration.

At four decoder resolutions, a learnable prior tensor is progressively refined. The prior module compares the current decoder feature with the prior and injects the resulting coarse-to-fine guidance into the corresponding encoder skip feature. At full resolution, five class-specific heads combine the decoder output with the modality-stem features. Both the prior mechanism and the multi-head design are inherited design choices; their individual effects are not isolated in this work.

4 Training and Inference Pipeline

Figure 1 summarizes the path from four input volumes to the submitted segmentation.

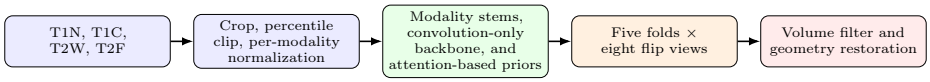

Fig. 1. Overview of the final submission pipeline. Deep supervision and augmentation are used only during training. At inference, view logits are averaged within each model before probabilities are averaged across models.

4.1 Preprocessing and augmentation

A foreground bounding box is computed from the union of nonzero voxels over the four modalities. Each cropped modality is clipped to its 1st and 99th positive-voxel percentiles and independently rescaled to $[0, 1]$. Dimensions smaller than

128 voxels are zero-padded, with the padding offset randomized during training. Training uses random 128^3 crops.

Training applies a random affine transform with probability 0.2, rotation range 0.26 radians per axis, and scale range 0.2. Each spatial axis is independently flipped with probability 0.5, and each orthogonal axis pair is independently rotated by a multiple of 90° with probability 0.5. Intensity perturbations include Gaussian noise, Gaussian smoothing, intensity scaling, and contrast or gamma adjustment; each is applied with probability 0.2.

4.2 Objective and optimization

Deep supervision is applied to the final prediction and four intermediate decoder outputs. Let ℓ_k be the Dice–Focal loss at output k, from highest to lowest spatial resolution. Equation 2 defines the background-excluding training objective:

$$\mathcal{L} = \sum_{k=0}^{4} w_k \ell_k, \qquad w_k = \frac{2^{-k}}{\sum_{j=0}^{4} 2^{-j}}. \tag{2}$$

The Dice and Focal terms are weighted equally. The focal term uses $\gamma = 2$ with no α weighting, and the Dice numerator and denominator smoothing constants are both 10^{-5}. The five folds are trained for 401 epochs using AdamW, initial learning rate 2×10^{-4}, weight decay 10^{-5}, and $\beta = (0.9, 0.95)$. An effective batch size of four is obtained with batch size one and four gradient-accumulation steps. Five warm-up epochs start at 3×10^{-5} and are followed by cosine annealing. The random seed is 42. Training used a single NVIDIA RTX 4090 and required approximately 24 hours per fold, or approximately 120 GPU-hours for all five folds.

4.3 Ensembling, post-processing, and output restoration

Each fold model processes the foreground crop using sliding-window inference with a 128^3 region of interest, overlap 0.6, and sliding-window batch size two. For each model, logits are computed for the native input and all seven non-empty combinations of flips along the three spatial axes. Flipped outputs are restored to the native orientation, and their logits are averaged across the eight views. Softmax is applied to the resulting model-specific mean logits, after which probabilities are averaged across the five models. In a single-case benchmark on an NVIDIA RTX 4070 SUPER, this complete five-model, eight-view configuration required 197.85 seconds and 6,989.7 MiB of peak allocated CUDA memory.

After argmax, a fixed class-volume heuristic is applied. If ET, NET, CC, or ED contains fewer than 33 voxels and is not the largest foreground class, its voxels are reassigned to the largest class. This is a global class-volume rule, not connected-component removal. The threshold was hand-set before the official submission to suppress extremely small predicted class volumes; it was not selected by optimizing out-of-fold metrics or from an anatomical volume criterion.

Table 2. Official BraTS-PEDs 2026 Synapse validation results for submission 9772538.

Metric	ET	NET	CC	ED	TC	WT
Global DSC	0.5096	0.9085	0.2196	0.0090	0.9400	0.9415
Global NSD	0.3907	0.6023	0.2100	0.0264	0.6289	0.6323

Its effect is therefore evaluated separately in Section 5.2. Padding is then removed, the crop is inserted into an array of the original size, and the T1N geometry is copied to an unsigned 8-bit NIfTI output.

5 Experiments and Results

5.1 Dataset and evaluation protocol

The training set used in this study contains 294 scans associated with 262 subject identifiers. For five-fold model training and ensembling, the scan manifest was partitioned into folds of 59, 59, 59, 59, and 58 scans. The partitioning was performed at scan rather than subject level. Twelve longitudinal subjects, comprising 42 scans, occur in more than one fold. The official Synapse evaluation is the primary quantitative result.

Only the BraTS-PEDs 2026 training set was used. We did not use external data, data from previous BraTS editions, or pretrained weights; all fold models were trained from random initialization.

The official evaluation reports global Dice similarity coefficient (DSC) and normalized surface Dice (NSD) for ET, NET, CC, ED, TC, and WT. Table 2 reports the values returned for official submission 9772538; none were estimated from local predictions.

The highest global DSC values were obtained for WT (0.9415) and TC (0.9400), followed by NET (0.9085). Performance was lower for ET (0.5096) and CC (0.2196), while ED remained particularly challenging (global DSC 0.0090 and global NSD 0.0264). These results indicate that the submitted pipeline delineated composite tumor regions more reliably than individual fine-grained tissue classes. These values characterize the complete submitted pipeline and do not isolate the effect of any component.

5.2 Local out-of-fold evaluation and post-processing ablation

For local out-of-fold (OOF) evaluation, each scan is predicted only by the fold model that did not train on it, using eight-view flip TTA but not five-model ensembling because the other models may have seen that scan. We pool all held-out scans and report per-case DSC and NSD as mean ± sample standard deviation over the n reference-present scans for each region (Table 3); this avoids inflation from absent–absent cases. The scan-level folds remain not strictly subject independent (Section 6).

Table 3. Five-fold OOF per-case performance on reference-present scans. Values are mean ± sample SD over the indicated present-only cases; Table 3 uses the 33-voxel post-processing rule.

Region	Present n	DSC ↑	NSD ↑
ET	198	0.6748 ± 0.2742	0.7333 ± 0.2627
NET	292	0.7843 ± 0.2221	0.7035 ± 0.2105
CC	104	0.4150 ± 0.3360	0.5001 ± 0.3268
ED	66	0.3577 ± 0.3056	0.3741 ± 0.2615
TC	294	0.8551 ± 0.1739	0.7111 ± 0.2332
WT	294	0.8777 ± 0.1500	0.7461 ± 0.2033

To isolate the 33-voxel rule, we reuse the same OOF predictions and toggle only this step after argmax. Table 4 reports per-case means with and without the rule, with Δ = with − without; all upstream settings and metric subsets remain fixed. A negative Δ therefore indicates degradation, whereas a positive value indicates improvement.

Table 4. OOF ablation of the 33-voxel heuristic on the same scans as Table 3. Values are present-only per-case means, and Δ = with − without.

Metric	Setting	ET	NET	CC	ED	TC	WT
DSC	Without filter	0.6788	0.7843	0.4279	0.3577	0.8551	0.8777
DSC	With filter	0.6748	0.7843	0.4150	0.3577	0.8551	0.8777
DSC	Δ	-0.0040	0.0000	-0.0128	-0.0000	0.0000	0.0000
NSD	Without filter	0.7403	0.7033	0.5301	0.3824	0.7111	0.7461
NSD	With filter	0.7333	0.7035	0.5001	0.3741	0.7111	0.7461
NSD	Δ	-0.0070	0.0002	-0.0300	-0.0083	0.0000	0.0000

With the filter, DSC decreased for ET ($\Delta = -0.0040$) and CC ($\Delta = -0.0128$), while ED changed only negligibly ($\Delta = -0.0000$); NET, TC, and WT were unchanged to four decimal places. For NSD, the decreases were −0.0070, −0.0300, and −0.0083 for ET, CC, and ED, respectively, whereas NET increased slightly (+0.0002) and TC and WT were unchanged to four decimal places. Thus, on these fixed present-only OOF cases, the rule produced measurable decreases in CC and ET scores, but the small ED DSC change and the unchanged composite scores indicate that it does not explain the overall regional pattern by itself. This paired ablation establishes the rule's effect on these OOF predictions, but it does not show that the rule caused the official ED failure or that its effect generalizes to the organizer-held validation set. The official Synapse values in Table 2 were produced with the filter.

5.3 Interpretation of region-wise metrics

The six reported regions form a hierarchy rather than six independent labels. TC is the union of ET, NET, and CC, while WT additionally includes ED. Consequently, confusing two constituent labels can reduce their individual scores without changing the corresponding composite mask. The high TC and WT scores should therefore be interpreted as evidence that the overall tumor extent was more stable than its internal decomposition, not as evidence of uniform accuracy across tumor subregions.

For every evaluated region, global NSD was lower than global DSC. The gap was largest for NET, TC, and WT, for which overlap remained high while the surface scores were between 0.6023 and 0.6323. Thus, the overlap scores do not remove the need to examine boundary placement. The official report provides submission-level global values rather than a per-case score distribution, so these results do not support estimates of variance, confidence intervals, or the frequency of complete failure on individual cases.

5.4 Qualitative cross-validation example

Figure 2 shows one training-set case selected without using its prediction score. We restricted selection to subjects represented by one scan and selected the case whose WT voxel count was closest to the median of the eligible cases. This deterministic rule selected `BraTS-PED-00060-000`, which belongs to fold 0. The displayed segmentation was produced only by the fold 0 model, for which the scan was held out. This qualitative-only prediction used the native view, sliding-window batch size one, and overlap 0.25; it is not the five-model, eight-view prediction used for official evaluation. The same axial slice is shown for T1C and T2F, followed by the manual annotation and prediction.

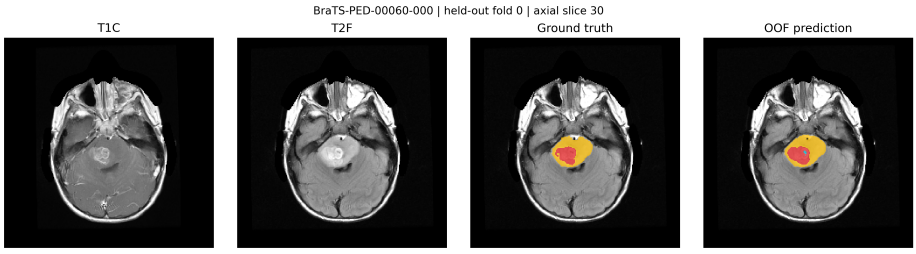

Fig. 2. Qualitative out-of-fold example from the training set. Colors: red, ET; yellow, NET; cyan, CC; green, ED. The fold 0 model did not use the displayed scan for training.

6 Discussion

This work follows the engineering-oriented pattern of successful nnU-Net-based challenge reports: start from a strong segmentation system, state exactly what is

inherited, and document the task-specific decisions that produce the submitted predictions [3,7]. In our case, the central adaptation is the combination of a five-class pediatric target, multi-scale optimization, fold and orientation averaging, fixed class-volume post-processing, and strict geometry restoration. These choices form one pipeline; without controlled ablations, performance cannot be attributed to any single component.

The convolution-only backbone setting is also important for interpreting the method. Although PTransBTS was introduced as a hybrid architecture, the trained configuration contains no Transformer channels in its channel-split backbone. No Transformer-active alternative was evaluated, so the submitted performance cannot be attributed to disabling those channels. The attention-based learnable pyramid prior remains active, but it is inherited from the baseline and is not claimed as a new contribution.

Five-fold ensembling and eight-view TTA aggregate predictions across training splits and spatial orientations. Their effects were not isolated, so the aggregate result should not be interpreted as evidence for either component alone. The 33-voxel filter is deliberately simple and may remove a real, very small class or reassign it to an anatomically implausible class. It should consequently be understood as a submission heuristic rather than a generally validated post-processing method. The OOF effects were sparse and case concentrated, particularly for CC and ET, so the ablation indicates sensitivity rather than systematic degradation. This caution is especially relevant to CC and ET, where a small absolute change in predicted volume can have a large effect on region-wise overlap.

Because the folds were not grouped by subject, local out-of-fold estimates are not strictly subject independent: longitudinal scans from the same subject can occur in both the training and held-out folds. Their correlation may make local cross-validation estimates optimistic. This limitation does not affect the official Synapse evaluation, which was performed on a separate organizer-held validation set.

Two directions warrant further study. First, controlled baseline-to-final ablations could quantify the contribution of each component. Second, evaluation with missing modalities, external institutions, and acquisition shifts could characterize robustness beyond the organizer-defined setting. The official Synapse results reported here characterize the complete submitted pipeline under the challenge protocol.

7 Conclusion

We have adapted PTransBTS to BraTS-PEDs 2026 Task 2 through changes to the target representation, training configuration, inference ensemble, post-processing, and geometry-preserving output restoration. The submitted network uses a convolution-only backbone path with modality-specific stems and inherited attention-based pyramid priors. The work is intended as a transparent account of adapting a task-specific baseline for a pediatric segmentation challenge, not as a claim of architectural novelty. On official evaluation, composite TC and WT regions were segmented substantially more reliably than the individual CC and ED classes.

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Capellán-Martín, D., Jiang, Z., Parida, A., Liu, X., Lam, V., Nisar, H., Tapp, A., Elsharkawi, S., Ledesma-Carbayo, M.J., Anwar, S.M., Linguraru, M.G.: Model ensemble for brain tumor segmentation in magnetic resonance imaging. In: Brain Tumor Segmentation, and Cross-Modality Domain Adaptation for Medical Image Segmentation. Lecture Notes in Computer Science, vol. 14669, pp. 221–232. Springer (2024). https://doi.org/10.1007/978-3-031-76163-8_20
2. 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. Nature Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
3. Isensee, F., Jäger, P.F., Full, P.M., Vollmuth, P., Maier-Hein, K.H.: nnU-Net for brain tumor segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pp. 118–132. Lecture Notes in Computer Science, Springer (2021). https://doi.org/10.1007/978-3-030-72087-2_11
4. Kazerooni, A.F., et al.: The brain tumor segmentation (BraTS) challenge 2023: Focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). arXiv preprint arXiv:2305.17033 (2023), https://arxiv.org/abs/2305.17033
5. Kazerooni, A.F., et al.: The brain tumor segmentation in pediatrics (BraTS-PEDs) challenge: Focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). arXiv preprint arXiv:2404.15009 (2024), https://arxiv.org/abs/2404.15009
6. Kazerooni, A.F., et al.: BraTS-PEDs: Results of the multi-consortium international pediatric brain tumor segmentation challenge 2023. Machine Learning for Biomedical Imaging **3**, 72–87 (2025). https://doi.org/10.59275/j.melba.2025-f6fg
7. Li, X., Xu, Z.Q.J., Ren, Y., Qiu, T., Wang, X.: An advanced nnU-Net framework for BraTS-2025 PED. In: Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. Lecture Notes in Computer Science, vol. 16376, pp. 422–433. Springer (2026). https://doi.org/10.1007/978-3-032-16365-3_38
8. Yu, H., Peng, Y.: PTransBTS: A hybrid transformer integrating priors for brain tumor segmentation. In: Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. Lecture Notes in Computer Science, vol. 16376, pp. 87–98. Springer (2026). https://doi.org/10.1007/978-3-032-16365-3_8