This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

# Generalizable Brain Tumor Segmentation with Self-Training and Tumor-Aware Deformations

Henrique Zan Grande[1], Jeovane Honorio Alves[2], Rayson Laroca[1], and Andre Gustavo Hochuli[1]

[1] Pontifícia Universidade Católica do Paraná (PUCPR), Programa de Pós-Graduação em Informática (PPGIa), Curitiba, Paraná, Brazil
{henrique.zgrande,rayson,aghochuli}@ppgia.pucpr.br
[2] University of Luxembourg, SEDAN - SnT, Luxembourg, Luxembourg
jeovane.alves@uni.lu

**Abstract.** This work presents an approach to the Generalizability Across Tumors (BraTS-GoAT) task of the BraTS 2026 Challenge, which focuses on robust segmentation of brain tumor sub-regions across a heterogeneous patient population. The proposed method employs the nnU-Net framework with a large residual encoder architecture, integrating a semi-supervised learning technique with pseudo-labels generated from the unlabeled training data and a tumor-aware deformable augmentation that locally deforms the lesion while preserving the surrounding anatomy. We evaluate the individual contributions of each component, as well as their combination, using varying proportions of the most confident pseudo-labeled cases. The best-performing arrangement for the generalization task achieves Dice and NSD scores of 0.881 and 0.473 for Whole Tumor, 0.817 and 0.490 for Tumor Core, and 0.775 and 0.533 for Enhancing Tumor on the BraTS-GoAT validation set, improving over the labeled-only baselines across all tumor regions and confirming that self-training and the proposed augmentation are complementary. Our source code is publicly available at https://github.com/Henrique-zan/brats-goat-2026/.

**Keywords:** BraTS-GoAT · Brain tumor segmentation · nnU-Net · Semi-supervised learning · Data augmentation · Domain generalization.

## 1 Introduction

Brain tumors vary considerably in appearance, size and location, making automatic segmentation from Magnetic Resonance Imaging (MRI) images a long-standing challenge. Accurate delineation of the Whole Tumor, Tumor Core, and Enhancing Tumor is essential for diagnosis, treatment planning, and therapy assessment [4,16]. In this vein, the Brain Tumor Segmentation (BraTS) challenge has driven advances in this field by providing a large annotated benchmark [3,5].

Most state-of-the-art methods for BraTS are based on deep encoder-decoder architectures derived from U-Net and its volumetric extension, 3D U-Net [7,18].

Building on this architectural family, nnU-Net provides a self-configuring segmentation pipeline that adapts preprocessing, network design, training, and inference to the target dataset [10]. Its robustness and reproducibility have established it as a reference framework for developing and evaluating medical image segmentation methods [2,6]. Nevertheless, maintaining reliable performance across distinct tumor types, acquisition protocols, imaging sites, and patient populations remains challenging.

The Generalizability Across Tumors (BraTS-GoAT) task shifts the emphasis from in-distribution accuracy to generalization across heterogeneous tumor distributions [5,12]. Its data include multiple tumor types, acquisition sites, and patient populations, thereby penalizing methods that overfit site-specific image characteristics or a narrow range of tumor morphologies. This setting motivates the use of unlabeled training data through semi-supervised learning and the introduction of anatomy-aware augmentations that increase morphological diversity without producing unrealistic global deformations [13]. Together, these strategies can broaden the effective training distribution and improve robustness to previously unseen cases.

To address this generalization challenge, we extend nnU-Net with three complementary components: a region-based teacher built from the Large Residual Encoder preset, a self-training pipeline that uses confident pseudo-labels generated by a modified nnU-Net teacher [19], and a tumor-aware deformable augmentation. The first two components strengthen pseudo-label generation and knowledge transfer, whereas the proposed augmentation constitutes the main methodological contribution by increasing tumor-shape variability while preserving the surrounding anatomy.

Based on these observations, this work is guided by the following research questions:

- **RQ1**. What is the impact of incorporating unlabeled BraTS-GoAT data through self-training on segmentation generalization across heterogeneous tumor distributions?
- **RQ2**. How does tumor-aware deformable augmentation complement self-training in improving segmentation robustness and generalization?

To answer these questions, we conduct a systematic ablation study in which each component is evaluated independently before being integrated into the complete pipeline. Experiments are conducted using a patient-level holdout to prevent data leakage, and the best-performing arrangement is selected based on validation results from the official BraTS-GoAT challenge platform. The results show that self-training improves overall performance and that the proposed augmentation provides additional gains for Whole Tumor (WT) and Tumor Core (TC), while largely preserving the improvements obtained for Enhancing Tumor (ET).

The remainder of this paper is organized as follows. Section 2 describes the challenge dataset and evaluation regions. Section 3 presents the proposed teacher–student framework, tumor-aware augmentation, and training procedure. Section 4 reports the quantitative and qualitative results, while Section 5 an-

alyzes the contribution of each component. Finally, Section 6 summarizes the main findings and outlines directions for future work.

## 2 Dataset Specification

We use the official BraTS-GoAT dataset from the BraTS 2026 Challenge [5], a multi-institutional collection designed to evaluate the generalizability of brain tumor segmentation methods across tumor types, acquisition sites, and patient populations. Each case comprises four MRI modalities: native T1-weighted (T1n), contrast-enhanced T1-weighted (T1c), T2-weighted (T2w), and T2-FLAIR (T2f), used as four input channels. All volumes are skull-stripped, assigned to an anatomical template, and resampled to an isotropic resolution of $1\text{mm}^3$ ($240 \times 240 \times 155$ voxels), following the standard BraTS preprocessing pipeline [3,4]. Voxel intensities are normalized independently for each modality using nnU-Net default z-score normalization within the non-zero brain region.

The dataset comprises a labeled training set, an unlabeled training pool, and an official validation set without public ground truth, evaluated through the challenge platform. The labeled set includes fully annotated adult gliomas [4], partially annotated meningiomas [14], and brain metastases [17], while the remaining cases constitute the unlabeled pool. The validation set further includes BraTS-Africa [1] and pediatric tumors [8,9], introducing previously unseen tumor types and patient populations. Annotations follow the standard BraTS protocol: necrotic/non-enhancing tumor core (NCR/NET, label 1), peritumoral edema (ED, label 2), and enhancing tumor (ET, label 3). Performance is evaluated on the derived regions Whole Tumor (WT $= 1 \cup 2 \cup 3$), Tumor Core (TC $= 1 \cup 3$), and Enhancing Tumor (ET $= 3$). Fig. 1 shows the four MRI modalities and the corresponding segmentation for a representative subject. The dataset contains 1,351 labeled, 1,138 unlabeled, and 451 validation volumes (patients).

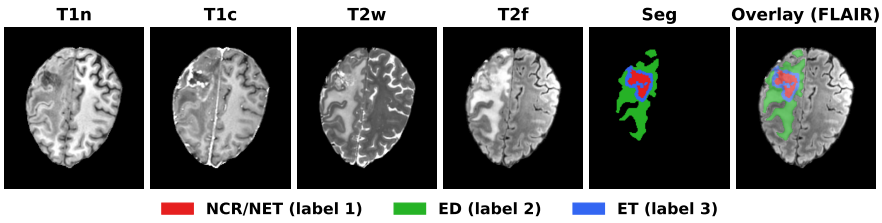


**Fig. 1.** Example MRI modalities and segmentation masks.

## 3 Method

Fig. 2 summarizes the proposed pipeline. A supervised model (teacher) is first trained on the labeled data and then used to generate pseudo-labels for the unlabeled cases. The most confident predictions are merged with the labeled data to

improve robustness and generalization. The second model architecture (student) employs an unmodified large residual encoder nnU-Net [11] and is trained with the proposed tumor-aware deformable augmentation to enhance its representations.

The following sections describe each component in detail. The complete implementation, including all hyperparameters, source code, and trained model weights, is publicly available for research purposes[1].

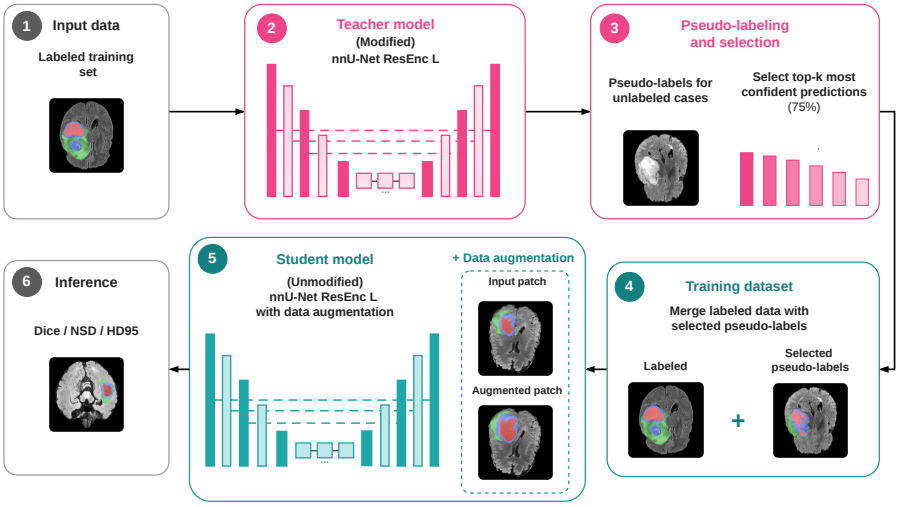


**Fig. 2.** Overview of our nnU-Net-based self-training pipeline for BraTS-GoAT.

### 3.1 Supervised Model (Teacher)

To maximize the quality of the pseudo-labels, the teacher model is based on nnU-Net v2 [10] using the 3D full-resolution configuration with the Large Residual Encoder preset proposed in [11]. Following Satushe et al. [19], we further replace the default instance normalization with group normalization, incorporate axial attention into the decoder, and augment the loss with a Hausdorff-distance term to improve boundary delineation. The model is trained in a region-based manner, predicting the three overlapping tumor regions (WT, TC, and ET) with sigmoid outputs. The teacher is trained with a patient-level five-fold cross-validation, which keeps each patient within a single fold and thus avoids leakage. The final prediction averages the five resulting models to reduce variance and potential bias.

Although more computationally demanding than the standard nnU-Net, this architecture provides greater capacity to model complex anatomical and morphological variations. Its strong performance in the BraTS 2025 Generalization

[1] https://github.com/Henrique-zan/brats-goat-2026/

Challenge [19] further motivated its selection as the teacher model for pseudo-label generation, used in the subsequent training sequence.

### 3.2 Semi-Supervised Model (Student)

To leverage the unlabeled training data described in Section 2, we adopt a self-training strategy [15,20], in which a teacher model trained on labeled data generates pseudo-labels that are subsequently incorporated into a new training stage.

In our approach, the supervised teacher model described in Section 3.1 predicts the three overlapping tumor regions (WT, TC, and ET). Cases with empty Whole Tumor predictions are discarded, and the remaining pseudo-labeled cases are ranked by confidence, calculated as the mean, across the three regions, of the average foreground probability within each predicted region. The top 75% most confident cases, corresponding to confidence scores of at least 97.22%, are retained and merged with the labeled training set. This confidence-based filtering reduces the impact of noisy pseudo-labels while substantially expanding the effective training set.

We selected the retention fraction by comparing 25%, 50%, 75%, and 100% of the confidence-ranked cases, each trained identically to the student models except for an 80%/20% split adopted only for this comparison, and evaluated on the BraTS-GoAT validation set. As shown in Table 1, the 75% setting offered the best trade-off between the number of retained cases and label reliability, and was therefore adopted for the final model.

**Table 1.** BraTS-GoAT validation results for the four pseudo-label retention fractions, taken after discarding the empty-prediction cases. Dice and NSD are reported as percentages, whereas HD95 is in millimeters. For all metrics, results are presented as mean ± standard deviation over all validation cases.

| Retention | Metric | WT | TC | ET | Avg |
|---|---|---|---|---|---|
| 25% 268 cases | Dice (↑) | 86.3 ± 20.8 | 80.2 ± 28.1 | 76.2 ± 30.9 | 80.9 |
| | NSD (↑) | 45.3 ± 20.4 | 47.8 ± 28.3 | 52.2 ± 26.6 | 48.4 |
| | HD95 (↓) | 19.3 ± 69.5 | 26.4 ± 82.2 | 44.9 ± 114.6 | 30.2 |
| 50% 536 cases | Dice (↑) | 86.9 ± 20.0 | 80.2 ± 28.0 | 76.3 ± 30.7 | 81.1 |
| | NSD (↑) | 46.1 ± 20.3 | 48.3 ± 27.9 | 52.5 ± 26.4 | 49.0 |
| | HD95 (↓) | 17.0 ± 63.4 | 26.1 ± 80.7 | 45.2 ± 114.2 | 29.4 |
| 75% 803 cases | Dice (↑) | **87.7** ± 18.5 | **81.3** ± 26.5 | **77.3** ± 29.8 | **82.1** |
| | NSD (↑) | **46.6** ± 20.0 | **49.0** ± 27.5 | **53.4** ± 25.9 | **49.7** |
| | HD95 (↓) | 15.3 ± 58.6 | 21.8 ± 71.6 | **40.9** ± 108.8 | 26.0 |
| 100% 1,071 cases | Dice (↑) | 87.4 ± 18.6 | 80.8 ± 26.7 | 76.8 ± 30.0 | 81.7 |
| | NSD (↑) | 46.2 ± 19.9 | 48.4 ± 27.4 | 52.7 ± 26.0 | 49.1 |
| | HD95 (↓) | **14.7** ± 56.3 | **21.4** ± 69.9 | 41.1 ± 108.9 | **25.8** |

Unlike the teacher model, the student employs the unmodified large residual encoder nnU-Net v2 architecture [10,11] without the additional architectural modifications introduced for pseudo-label generation. The student is trained on the combined labeled and pseudo-labeled dataset using the proposed tumor-aware deformable augmentation described in Section 3.3.

Training uses a patient-level 95%/5% (training/validation) holdout to prevent data leakage and ensure unbiased model selection. The small validation split maximizes the amount of labeled data available for training, which is particularly beneficial in the semi-supervised setting. Final performance is evaluated independently on the official BraTS-GoAT validation set.

This design yields a computationally efficient student while leveraging the knowledge captured by the higher-capacity teacher through self-training. Consequently, the student benefits from reliable pseudo-labels, improving generalization without requiring the same model complexity as the teacher.

### 3.3 Tumor-Aware Deformable Augmentation

The proposed tumor-aware deformable augmentation increases tumor morphological variability while preserving anatomical characteristics. It combines localized elastic deformations and isotropic scaling to expose the network to a broader range of plausible tumor morphologies, thereby improving generalization.

The deformation is restricted to the lesion and a narrow surrounding region using a Gaussian-smoothed weighting map derived from the dilated tumor mask, ensuring a smooth transition to the unchanged tissue. For each sampled patch containing a sufficiently large tumor, the transform applies either a smooth elastic deformation or an isotropic scaling around the tumor center of mass, each with equal probability. Smaller lesions are excluded to avoid unrealistic deformations and interpolation artifacts. Image intensities and segmentation labels are warped using the same displacement field, with linear and nearest-neighbor interpolation, respectively, preserving geometric consistency. For computational efficiency, the deformation is computed only within an expanded tumor bounding box, where an edge-fade window enforces zero displacement at the boundaries, preventing boundary artifacts.

Table 2 summarizes the augmentation parameters, and Fig. 3 shows representative examples.

**Table 2.** Parameters of the tumor-localized deformable augmentation.

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| Application probability | 0.5 | Dilation iterations | 8 |
| Minimum tumor size (voxels) | 50 | Halo smoothing | 5.0 |
| Elastic field smoothness | 4.0 | ROI margin (voxels) | 20 |
| Maximum elastic displacement | 9.0 | Edge-fade width (voxels) | 10 |
| Scale range | 0.65–1.45 | | |

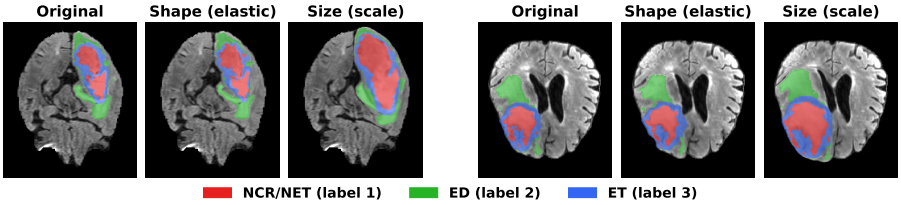


**Fig. 3.** Deformable augmentation examples (original, shape, size).

### 3.4 Training Procedure

Both the teacher and the student are trained with stochastic gradient descent, a polynomial learning-rate schedule, deep supervision, and a combined Dice and cross-entropy loss. For the students, the patch size ($160 \times 192 \times 160$) and batch size (3) are automatically determined by the nnU-Net framework using rule-based heuristics that depend on the dataset and hardware characteristics, and remain fixed across all student experiments, ensuring a fair comparison across all configurations. Each student is also trained for 1000 epochs. The teacher instead uses a fixed patch size ($128 \times 128 \times 128$) and batch size (5), augments the loss with a Hausdorff-distance term, and is trained for 250 epochs. For all models, the epoch with the highest validation Dice score is selected for inference. All experiments are performed on an NVIDIA H100 GPU with 80 GB of memory.

## 4 Results

To fairly evaluate the proposed approach and answer the research questions posed in Section 1, we define an experimental protocol that isolates the contribution of each pipeline component. We first evaluate the supervised teacher model to assess the quality of the generated pseudo-labels. Next, we compare the student trained only on labeled data with the student trained via self-training (**RQ1**). We then evaluate the impact of the augmentation alone, applied to the student trained only on labeled data. Finally, we evaluate the complete proposed approach that combines self-training with the tumor-aware deformable augmentation (**RQ2**).

As the validation ground truth is not publicly available, predictions are generated locally for the validation cases and submitted to the challenge platform, which returns the official Dice, NSD, and HD95 metrics. For each metric, we report the mean and standard deviation across predictions for 451 validation cases.

Table 3 summarizes the ablation results. The first configuration is the supervised teacher evaluated directly on the validation set. The second is the student trained only on labeled data. The third adds confidence-filtered pseudo-labels through self-training. The fourth applies the tumor-aware deformable augmentation to the labeled-only student. The fifth is the submitted model, which combines self-training with the augmentation.

**Table 3.** BraTS-GoAT validation results for the teacher and the four student configurations, evaluated through the submission platform. Dice and NSD are reported as percentages, whereas HD95 is in millimeters. For all metrics, results are presented as mean ± standard deviation over all validation cases.

| Model | Metric | WT | TC | ET | Avg |
|---|---|---|---|---|---|
| Teacher Labeled Data Only | Dice (↑) | 86.7 ± 20.3 | 80.2 ± 28.4 | 77.3 ± 29.8 | 81.4 |
| | NSD (↑) | 45.5 ± 20.1 | 48.3 ± 27.8 | 53.0 ± 25.9 | 48.9 |
| | HD95 (↓) | 16.9 ± 63.2 | 25.0 ± 79.1 | 41.0 ± 109.3 | 27.6 |
| Student Labeled Data Only | Dice (↑) | 86.6 ± 20.9 | 81.1 ± 27.4 | 76.0 ± 31.4 | 81.2 |
| | NSD (↑) | 46.8 ± 20.9 | 48.6 ± 27.6 | 52.1 ± 26.5 | 49.1 |
| | HD95 (↓) | 19.3 ± 69.5 | 25.6 ± 82.1 | 46.8 ± 117.3 | 30.6 |
| Student w/ Pseudo-Labels | Dice (↑) | 87.8 ± 18.5 | 80.3 ± 27.7 | **77.5** ± 29.1 | 81.9 |
| | NSD (↑) | 46.9 ± 19.9 | 48.4 ± 27.6 | 53.4 ± 25.7 | 49.5 |
| | HD95 (↓) | 15.6 ± 60.8 | 23.3 ± 75.2 | **39.2** ± 106.9 | 26.0 |
| Student w/ Tumor-Aware Aug. | Dice (↑) | 87.3 ± 19.2 | 81.5 ± 26.6 | 76.9 ± 30.5 | 81.9 |
| | NSD (↑) | **47.3** ± 21.1 | **49.9** ± 28.2 | **53.5** ± 26.8 | **50.2** |
| | HD95 (↓) | 15.2 ± 58.4 | **18.7** ± 65.3 | 41.9 ± 111.0 | **25.3** |
| Student (submitted) w/ Pseudo-Labels and Tumor-Aware Aug. | Dice (↑) | **88.1** ± 17.8 | **81.7** ± 25.8 | **77.5** ± 29.5 | **82.4** |
| | NSD (↑) | **47.3** ± 19.9 | 49.0 ± 27.3 | 53.3 ± 25.7 | 49.9 |
| | HD95 (↓) | **14.0** ± 56.0 | 21.5 ± 71.6 | 40.7 ± 108.9 | 25.4 |

As a complementary analysis, Table 4 reports the mean Dice scores obtained on validation splits derived from the challenge-released training data and used for model selection. The left block corresponds to Table 3 and includes the five-fold teacher average together with results on the 95%/5% student holdout (68 cases), whereas the right block corresponds to Table 1 and reports results using an 80%/20% holdout (271 cases). Since the two blocks are based on different validation protocols, their results should be interpreted only through within-block comparisons. Moreover, as these splits are derived from the training data, the reported scores reflect in-distribution performance and should not be interpreted as estimates of generalization to the official challenge validation set.

**Table 4.** Internal validation results for the evaluated models and pseudo-label retention fractions, as computed by nnU-Net at the end of training. Dice scores are reported as percentages and averaged across all validation cases.

| Model | Avg. Dice (↑) |
|---|---|
| Teacher, Labeled Data Only | 89.8 |
| Student, Labeled Data Only | 81.3 |
| Student w/ Pseudo-Labels | 80.9 |
| Student w/ Tumor-Aware Aug. | 80.9 |
| Student (submitted) | 81.2 |

| Retention | Avg. Dice (↑) |
|---|---|
| 25% pseudo-labels | 82.4 |
| 50% pseudo-labels | 82.0 |
| 75% pseudo-labels | 82.2 |
| 100% pseudo-labels | 82.1 |

## 5 Discussion

Comparing the supervised teacher with the student trained only on labeled data reveals that the higher-capacity teacher provides only marginal performance differences despite its substantially greater computational complexity. While the teacher achieves slightly better average Dice and HD95, the standard nnU-Net student obtains comparable results, even outperforming the teacher on some individual metrics and tumor regions. This observation suggests that the increased architectural complexity is not justified for direct deployment. Instead, its primary value lies in producing reliable pseudo-labels that transfer knowledge to a simpler and more efficient student through self-training.

Among the student models, incorporating pseudo-labels improves the average Dice score from 81.2% to 81.9% while reducing the average HD95 from 30.6 to 26.0 mm, demonstrating that self-training effectively exploits the unlabeled data (**RQ1**). The tumor-aware deformable augmentation improves boundary metrics most clearly: applied to the labeled-only student, it achieves the best average NSD (50.2%) and HD95 (25.3 mm) among all configurations. Combined with self-training, the submitted model attains the best average Dice (82.4%), with average NSD and HD95 of 49.9% and 25.4 mm, confirming that the augmentation contributes complementary boundary gains (**RQ2**).

Region-wise, the largest gains are observed for Whole Tumor, where the complete approach achieves the highest Dice (88.1%), NSD (47.3%), and the lowest HD95 (14.0 mm). Tumor Core also benefits from the combined strategy, improving all three evaluation metrics over both the baseline and the self-training-only configuration. For Enhancing Tumor, self-training provides the largest improvement, increasing Dice from 76.0% to 77.5% and reducing HD95 by more than 7 mm, while the proposed augmentation preserves these gains. Overall, the results indicate that self-training and the proposed tumor-aware augmentation are complementary: pseudo-labels primarily expand the effective training distribution (**RQ1**), whereas the augmentation improves robustness to morphological variability, particularly for the larger tumor regions (**RQ2**).

Fig. 4 presents predictions from the submitted model for two representative validation cases, overlaid on the FLAIR images. As the validation annotations are unavailable, these examples support only a qualitative inspection of the predicted NCR/NET, ED, and ET regions, rather than a direct visual comparison with the ground truth.

## 6 Conclusions

This paper presented an approach to the BraTS-GoAT task of the BraTS 2026 Challenge that combines confidence-filtered self-training with tumor-aware deformable augmentation. Self-training expands the effective training set using high-confidence pseudo-labels, while the proposed augmentation increases the diversity of plausible tumor shapes and sizes without applying global deformations to the surrounding anatomy.

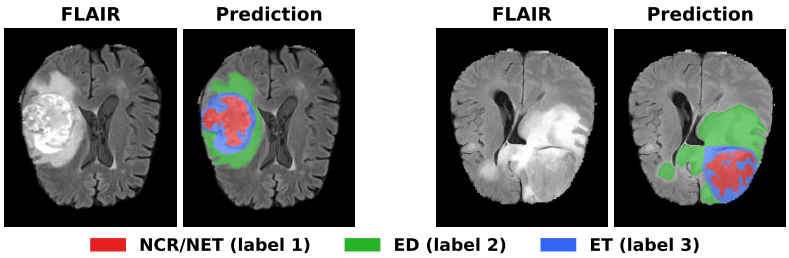


**Fig. 4.** Qualitative segmentation results of the submitted model on the validation set.

On the official validation set, the combined approach improves both overlap and boundary metrics relative to the labeled-only student, increasing the average Dice score from 81.2% to 82.4% and reducing the average HD95 from 30.6 to 25.4 mm. The results indicate that the two components are complementary, with the augmentation providing its clearest additional gains for WT and TC.

Future work will investigate iterative self-training, teacher ensembles, more informative pseudo-label ranking criteria, post-processing strategies, and more detailed analyses across MRI modalities, tumor types, and acquisition domains. We also plan to explore additional methods for further improving segmentation performance.

**Acknowledgments.** The authors acknowledge the financial support of the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES), Finance Code 001; the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), under Grants No. 302909/2022-2 and 444192/2024-7; the *Financiadora de Estudos e Projetos* (FINEP), for the High-Performance Computing (HPC) infrastructure provided by the *Centro Integrado de Soluções em Inteligência Artificial* (CISIA) at the *Pontifícia Universidade Católica do Paraná* (PUCPR); and the *Fundação Araucária*, in partnership with the *Secretaria da Ciência, Tecnologia e Ensino Superior do Estado do Paraná*(SETI-PR), under Grant No. 653/2025 (FA/UNIVERSAL).

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

## References

1. Adewole, M., Rudie, J.D., Gbadamosi, A., et al.: The brain tumor segmentation (BraTS) challenge 2023: Glioma segmentation in sub-saharan africa patient population (BraTS-Africa). arXiv preprint arXiv:2305.19369 (2023)
2. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., et al.: The medical segmentation decathlon. Nature Communications **13**(1), 4128 (2022)
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 preprint arXiv:2107.02314 (2021)

4. 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. Scientific Data **4**, 170117 (2017)
5. Bakas, S., et al.: BraTS 2026 cluster of challenges: (Apr 2026). `https://doi.org/10.5281/zenodo.19714728`
6. Bancerek, M., Rudzki, P., Nalepa, J.: Segmentation of pre and post-treatment brain metastases using nnU-Nets. In: Bakas, S., Dennis, E., Astaraki, M., Baid, U., et al. (eds.) Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. pp. 161–172. Springer Nature Switzerland, Cham (2026)
7. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 424–432 (2016)
8. Fathi Kazerooni, A., 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)
9. Fathi Kazerooni, A., 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)
10. 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)
11. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K.H., Jaeger, P.F.: nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 488–498 (2024)
12. Karargyris, A., et al.: Federated benchmarking of medical artificial intelligence with medperf. Nature Machine Intelligence **5**(7), 799–810 (2023)
13. Krinski, B.A., Ruiz, D.V., Laroca, R., Todt, E.: DACov: A deeper analysis of data augmentation on the computed tomography segmentation problem. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **11**(5), 1724–1741 (2023)
14. LaBella, D., et al.: The ASNR-MICCAI brain tumor segmentation (BraTS) challenge 2023: Intracranial meningioma. arXiv preprint arXiv:2305.07642 (2023)
15. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: International Conference on Machine Learning (ICML) – Workshop on Challenges in Representation Learning (2013)
16. 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 Transactions on Medical Imaging **34**(10), 1993–2024 (2015)
17. Moawad, A.W., et al.: The brain tumor segmentation (BraTS-METS) challenge 2023: Brain metastasis segmentation on pre-treatment MRI. arXiv preprint arXiv:2306.00838 (2023)
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). vol. 9351, pp. 234–241 (2015)
19. Satushe, V., Arora, M., Metkar, S.: Ensemble-based generalization for brain tumor segmentation using nnU-Net variants and Swin UNETR. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 538–548 (2026)
20. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves ImageNet classification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10687–10698 (2020)