HD-TTA: Hypothesis-Driven Test-Time Adaptation for Safer Brain Tumor Segmentation

Kartik Jhawar[1,2][0000−0001−7725−1877] and Lipo Wang[1,2⋆][0000−0002−4257−7639]

[1] Institute for Digital Molecular Analytics and Science, Nanyang Technological University, Singapore 636921
[2] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798
ELPWang@ntu.edu.sg

Abstract. Standard Test-Time Adaptation (TTA) methods typically treat inference as a blind optimization task, applying generic objectives to all or filtered test samples. In safety-critical medical segmentation, this lack of selectivity often causes the tumor mask to spill into healthy brain tissue or degrades already-accurate predictions. We propose Hypothesis-Driven TTA, a novel framework that reformulates adaptation as a dynamic decision process. Rather than forcing a single optimization trajectory, our method generates intuitive competing geometric hypotheses: compaction (is the prediction noisy? trim volumetric artifacts) versus inflation (is the valid tumor under-segmented? safely inflate to recover). An unsupervised representation-guided selector then identifies the safest outcome via intrinsic texture consistency. Additionally, a pre-screening Gatekeeper prevents negative transfer by skipping adaptation on confident cases. We validate this proof-of-concept on a cross-domain binary brain tumor segmentation task, applying a model trained on adult BraTS 2023 gliomas to unseen pediatric and more challenging meningioma target domains. HD-TTA improves safety-oriented outcomes (95^{th} percentile Hausdorff Distance (HD95) and Precision) over several state-of-the-art representative baselines. Under severe domain shift (BraTS-MEN), HD-TTA demonstrates its robustness by reducing HD95 by approximately 6.4 mm and improving Precision by over 4%, while maintaining comparable Dice scores. These results demonstrate that resolving the safety-adaptation trade-off via explicit hypothesis selection is a viable, robust path for safe clinical model deployment. The code is available at `https://github.com/kartikvi01091896/HD-TTA`.

Keywords: Test-Time Adaptation · Brain Tumor Segmentation · Dynamic Inference

1 Introduction

Deep neural networks have become strong performers for medical image segmentation when the training and test images share similar scanners and acquisition protocols. In brain MRI, robust U-Net style pipelines such as nnU-Net

⋆ Corresponding author.

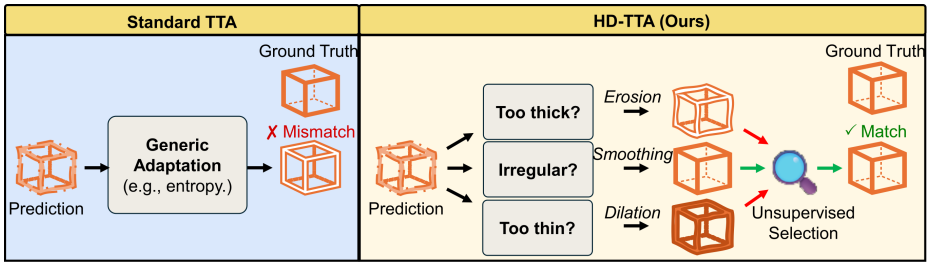

Fig. 1. Comparison of Standard TTA versus our proposed HD-TTA framework.

are often a strong starting point [7]. However, clinical deployment rarely satisfies the matched-distribution assumption. Differences in scanner vendor, sequence parameters, preprocessing, or patient cohorts can shift image statistics and pathology appearance, leading to missed tumor regions, false positive islands, and boundary leakage. These failures are not only about average overlap, they directly affect whether outputs are reliable for downstream use.

Test-time adaptation (TTA) improves inference robustness without target labels. Approaches range from classic TTA [18] to online parameter-update methods using entropy minimization [19], often stabilized by sample filtering (SAR [16], EATA [15]) or teacher-student frameworks [20]. Recent state-of-the-art methods introduce proxy supervision via self-training (IST [13]), feature-statistics alignment (TCA [22]), or segmentation-oriented evaluation objectives (TEGDA [23]).

Despite success in general vision, many TTA methods remain *blind* to specific medical risks. While some approaches incorporate priors to constrain adaptation [8,21], they often incur high computational overhead or assume fixed strategies that fail on heterogeneous tumors. Recent divide-and-conquer frameworks enable specialized adaptation but rely on interactive user clicks [10], unsuitable for label-free deployment. Crucially, most automated methods still attempt to adapt every / filtered test case(s) using a generic objective. In safety-critical segmentation, this lack of selectivity causes negative transfer: unnecessary updates on easy cases reduce Precision by hallucinating isolated tumor islands in healthy tissue, while generic objectives under severe shift produce confident but anatomically implausible boundary leakage. Since overlap-based metrics like Dice and Recall hide these harms, we prioritize 95^{th} percentile Hausdorff Distance (HD95) to capture severe boundary errors and Precision to evaluate false-positive suppression in healthy tissue.

We propose **Hypothesis-Driven Test-Time Adaptation (HD-TTA)**, a decision-oriented framework that prioritizes *selective* and *safe* refinement rather than unconditional optimization. Given a test volume, HD-TTA produces a baseline prediction, decides whether refinement is warranted, generates a small set of competing geometrically meaningful hypotheses, and selects the safest output using unsupervised plausibility signals (see the conceptual illustration in Fig. 1). In short, HD-TTA is the novel unsupervised framework capable of dynamically

switching strategies during test time adaptation, preventing the catastrophic failure (negative transfer) that occurs when standard inflation methods are applied to stable cases. We validate HD-TTA on the distinct Brain Tumor Segmentation (BraTS) 2023 Pediatric (PED) [6, 9] and Meningioma (MEN) [11, 12] target-domains using a nnU-Net v2 [7] backbone trained on BraTS 2023 GLI (glioma) dataset [1–4, 14]. Crucially, we apply the unchanged optimization framework with strictly fixed hyperparameters across domains to demonstrate out-of-the-box robustness. Our focus is on demonstrating consistent relative improvements in safety-critical metrics (HD95 and Precision) under distribution shifts, rather than comparison with fully supervised in-domain leaderboard results.

The technical contribution of HD-TTA is hypothesis-conditioned parallel logit refinement combined with unsupervised hypothesis selection under frozen backbone parameters. The specific proxy losses used for compactness and diffusion are instantiations of this framework and can be replaced. The Gatekeeper and selector are modular components. While this work focuses on binary tumor segmentation, the hypothesis-driven refinement framework is general and can be extended to multi-class or other structured prediction tasks.

2 Method

As shown in Fig. 2, HD-TTA is an orthogonal refinement layer atop a frozen nnU-Net v2 backbone (pre-trained on BraTS-GLI). Given a 3D multi-modal MRI volume target-domain test case x with initial logits z_0 and prediction P_0, adaptation optimizes only the test-sample logits z. The backbone remains frozen; probabilities $P = \text{sigmoid}(z)$ are computed at each step solely for loss evaluation and mask generation. The pipeline executes three stages: (1) a Gatekeeper to prevent unnecessary updates, (2) parallel generation of competing geometric Hypotheses, and (3) Representation-Guided Selection to output the safest correction.

2.1 Selective Adaptation via the Gatekeeper

To mitigate negative transfer on already accurate predictions, a lightweight Gatekeeper assesses the confidence of P_0. It flags a sample for hypothesis-conditioned refinement if its volume is critically small (< 300 voxels, risking tumor collapse) or if its uncertainty ratio (proportion of the predicted tumor volume where $0.3 < P_0 < 0.7$) exceeds 5%. Confident samples skip adaptation. Notably, this uncertainty ratio measures structural ambiguity over raw logit confidence. Confidently-wrong, crisp predictions yield low uncertainty and intentionally skip adaptation, preserving the baseline rather than amplifying errors.

2.2 Hypothesis-Conditioned Refinement

Two intuitive competing hypotheses for the proof-of-concept binary brain tumor segmentation task are as follows:

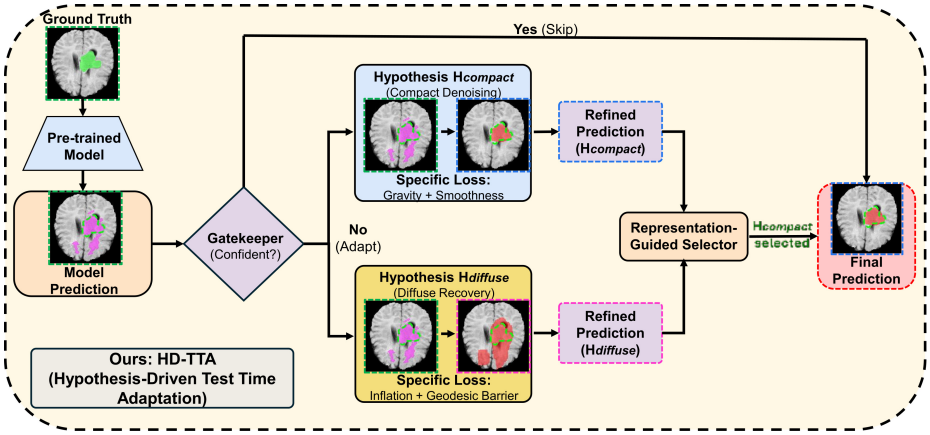

Fig. 2. Overview of the proposed HD-TTA framework. The inference pipeline consists of a selective Gatekeeper, parallel hypothesis generation (e.g., $H_{compact}$ winning to trim spurious false positives, versus $H_{diffuse}$), and an unsupervised representation-guided selector.

Hypothesis 1: Compact Denoising (H_{compact}). This hypothesis addresses over-segmentation and spurious noise islands. A loss function that enforces smoothness and spatial compactness:

$$\mathcal{L}_{compact} = \lambda_{ent}\mathcal{H}(P) + \underbrace{\lambda_{TV}||\nabla P||_1 + \lambda_{grav}\mathcal{V}(P)}_{\textbf{Specific to } H_{compact}} + \lambda_{anc}||z - z_0||^2 \quad (1)$$

Here, λ represents the weighting coefficients, $\mathcal{H}(P)$ is the entropy, and $||\nabla P||_1$ is standard Total Variation (TV) to smooth jagged borders [17]. The crucial term is the gravity loss $\mathcal{V}(P)$, which minimizes the spatial variance of the foreground coordinates, effectively pulling outlier pixels toward the tumor centroid. Formally, $V(P) = \sum_i P_i \|x_i - \mu\|^2$, where $\mu = \frac{\sum_i P_i x_i}{\sum_i P_i}$ is the probability-weighted foreground centroid. An anchor term penalizes deviation from the baseline logits.

Hypothesis 2: Diffuse Recovery (H_{diffuse}). This hypothesis addresses under-segmentation by encouraging controlled growth. To prevent uncontrolled leakage into healthy tissue, we replace standard TV with a geodesic barrier [5]:

$$\mathcal{L}_{diffuse} = \lambda_{ent}\mathcal{H}(P) + \underbrace{\lambda_{geo}\sum(g \cdot |\nabla P|) + \lambda_{inf}(-\mu_P)}_{\textbf{Specific to } H_{diffuse}} + \lambda_{anc}||z - z_0||^2 \quad (2)$$

The inflation term $-\mu_P = -\frac{1}{N}\sum_i P_i$ drives foreground expansion, constrained by a geodesic barrier g derived from image gradients. Formally, $g = \exp(-\alpha\|\nabla I\|)$, where large image gradients produce $g \approx 0$, halting expansion at anatomical boundaries, preventing leakage into the skull or ventricles.

2.3 Representation-Guided Selection

To autonomously select the safest hypothesis without ground truth, we utilize an intrinsic texture consistency signal. This is a proof-of-concept selector; however, other selectors based on morphology or test-sample perturbation stability can be swapped in without changing the overall HD-TTA framework. Since blind expansion carries a high risk of boundary leakage (increasing HD95), we treat H_{compact} as the safe default and require H_{diffuse} to pass a strict texture validation check to be selected. We define the high-confidence "tumor core" ($P_0 > 0.8$) and the expansion candidate region $\Delta = H_{diffuse} \setminus \text{Core}$ (i.e., the pixels added by the inflation hypothesis). We evaluate the consistency of this candidate region against the core:

$$S_{\text{rep}} = \exp\left(-0.5 \cdot \left(\frac{|\mu_{\Delta} - \mu_{\text{core}}|}{\gamma(\sigma_{\text{core}} + \epsilon)}\right)^2\right) \tag{3}$$

where μ and σ denote the mean and standard deviation of intensity, and $\gamma = 1.5$ serves as a tolerance scaling factor. The model selects H_{diffuse} only if the newly recruited pixels share the intrinsic intensity signature of the core ($S_{\text{rep}} > 0.95$); otherwise, it reverts to the safe default, H_{compact}.

3 Experiment and Results

3.1 Implementation Details.

We train a nnU-Net v2 backbone on the BraTS 2023 GLI source domain (official training split, 1251 cases) for 250 epochs using Dice loss and polynomial decay. For our binary segmentation proof-of-concept, we merge multi-class labels into a single whole-tumor mask for both training and evaluation. Target domains include the BraTS 2023 PED (99 cases) and BraTS 2023 MEN (144 cases, randomly sampled from the training split to match BraTS-PED's scale, enabling directly comparable aggregate statistics across target domains), evaluated without target-tuned optimization[3]. All methods optimize logits using frozen backbones, incurring no extra memory overhead. Classical TTA [18] (augmentation implemented via nnU-Net's mirroring) is uniformly active across all baselines for fairness. Baselines are adapted to use Instance Normalization, matching nnU-Net. This ensures a fair comparison: all methods operate under identical normalisation conditions. All experiments run on an RTX A5000 GPU (CUDA 12.2). The Gatekeeper flags 23.6% (PED) and 99.3% (MEN) of cases, yielding average per-case overheads of 5.17s and 21.74s, respectively. Flagged borderline cases undergo 1000 logit optimization steps (Adam, lr=0.1) per hypothesis, adding average additional latency of 21.91s (PED) and 21.89s (MEN) per flagged case.

[3] **Data use.** Data used in this publication were obtained as part of the Brain Tumor Segmentation (BraTS) Challenge project through Synapse ID: syn51156910 (CC-BY-NC 4.0; required BraTS citations included). Ethics approval and consent waived for this public, anonymized data.

Table 1. Comparison of HD-TTA against state-of-the-art representative TTA methods on BraTS-PED and BraTS-MEN target sets. Values are Mean ± Std. Best values are in **bold**. $^{\#}$ denotes statistically significant improvement ($p < 0.05$) over the best existing method (TCA) via a one-sided paired Wilcoxon signed-rank test with Holm–Bonferroni correction. In our nnU-Net v2 configuration, TEGDA produced unchanged output masks relative to Classic TTA; we therefore report identical metrics and discuss this observation in the text.

Method	Dice (%) ↑	HD95 (mm) ↓	Prec (%) ↑
BraTS-PED			
nnU-Net v2 (2021) [7]	86.47 ± 11.81	8.20 ± 12.33	82.90 ± 15.58
Classic TTA (2021) [18]	86.95 ± 11.56	6.59 ± 9.75	83.44 ± 15.44
SAR (2023) [16]	85.89 ± 12.35	7.58 ± 10.35	81.12 ± 16.26
IST (2024) [13]	86.19 ± 15.20	7.19 ± 12.88	85.33 ± 15.00
TCA (2025) [22]	**87.43±11.13**	6.39 ± 9.58	84.62 ± 14.97
TEGDA (2025) [23]	86.95 ± 11.62	6.59 ± 9.80	83.44 ± 15.52
HD-TTA (Ours)	87.26 ± 11.33	**5.35±6.91**	**86.32±13.34**$^{\#}$
BraTS-MEN			
nnU-Net v2 (2021) [7]	12.81 ± 22.15	73.50 ± 31.47	13.77 ± 24.95
Classic TTA (2021) [18]	13.20 ± 22.96	72.19 ± 32.42	14.63 ± 26.41
SAR (2023) [16]	12.98 ± 22.43	73.44 ± 31.77	13.41 ± 24.38
IST (2024) [13]	2.51 ± 10.90	96.00 ± 16.15	3.01 ± 13.68
TCA (2025) [22]	**13.26±23.16**	70.96 ± 32.48	15.36 ± 27.66
TEGDA (2025) [23]	13.20 ± 23.04	72.19 ± 32.53	14.63 ± 26.50
HD-TTA (Ours)	10.57 ± 19.47	**64.55±30.51**$^{\#}$	**19.64±35.22**$^{\#}$

Evaluation metrics include volume-level Dice similarity (in %), the HD95 (in mm), and precision (in %). The representation-guided selector (Sec. 2.3) uses: tumour core defined as $P_0 > 0.8$, expansion region $\Delta = H_{\text{diffuse}} \setminus \text{Core}$, consistency score S_{rep} by Eq. 3 with $\gamma = 1.5$, and selection threshold $S_{\text{rep}} > 0.95$. All hyperparameters, including loss coefficients in Section 2.2 ($H_{compact}$: $\lambda_{ent} = 10$, $\lambda_{TV} = 0.5$, $\lambda_{grav} = 50$, $\lambda_{anc} = 50$; $H_{diffuse}$: $\lambda_{ent} = 2$, $\lambda_{geo} = 50$, $\lambda_{inf} = 5$, $\lambda_{anc} = 0.1$), Gatekeeper thresholds in Section 2.1, and the Selection thresholds in Section 2.3, were established based on geometric intuition during initial method development to ensure stable optimization and define safe operational boundaries. Crucially, they were kept strictly fixed for all experiments without any target-domain tuning.

3.2 Quantitative and Qualitative Comparison.

Table 1 compares HD-TTA against representative TTA methods, ranging from purely averaging predictions to online adaptation using proxy supervision. Across both datasets, HD-TTA consistently achieves the best safety-oriented objectives. On BraTS-PED, nnU-Net v2 performs strongly (Dice 86.47%), making aggressive adaptation detrimental. Blind methods (SAR, IST) fail on borderline cases due to background dominance biasing entropy minimization, and IST amplifies errors under shift. While the strongest competitor, TCA improves Dice (best

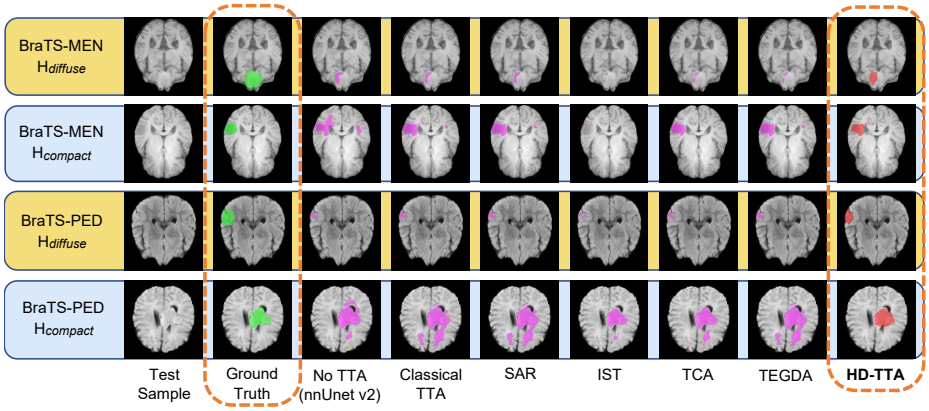

Fig. 3. Qualitative comparison on representative BraTS-MEN (Rows 1-2) and BraTS-PED cases (Rows 3-4). In Row 1, entropy-based and other methods fail to initiate growth, but HD-TTA selects H_{diffuse} to relatively recover missed tumor volume (red). In Row 2, HD-TTA selects H_{compact} to cleanly eliminate the spurious false-positive island, an artifact (in right hemisphere) even TCA fails to fully suppress. In Row 3, the backbone under-segments, prompting HD-TTA to select H_{diffuse} to recover the boundary where other methods fail. In Row 4, the backbone over-segments. While blind methods exacerbate this, our Gatekeeper flags the instability and selects H_{compact} to restore a plausible shape (notably, IST attempts but fails to remove this artifact entirely).

on both datasets) by aligning feature statistics, its global alignment objective does not consistently minimize worst-case boundary risk (HD95 6.39 mm). In contrast, HD-TTA achieves the lowest boundary error (HD95 5.35 mm, a 16.3% improvement over TCA) and high Precision (86.32%, +1.7% over TCA), while maintaining comparable Dice. This also confirms that our Gatekeeper successfully prevents negative transfer on easy cases, a capability missing in "always-adapt" baselines. Despite high standard deviations typical of heterogeneous target cohorts [23], HD-TTA avoids metric trade-offs, consistently improving both safety metrics (lowering HD95 and raising Precision) across diverse datasets.

HD-TTA's generalizability is pronounced on the challenging BraTS-MEN dataset (severe domain gap from BraTS-GLI in tumor location and appearance), where fragmented tumors yield universally low Dice, making Precision and HD95 critical for safety evaluation. Most baselines struggle to recover the missed tumor, yielding high HD95 scores (> 70 mm). HD-TTA substantially improves over the strongest competitor (TCA) in safety metrics, reducing HD95 by approximately 6.4 mm (70.96 $\rightarrow$ 64.55 mm) and improving Precision by over 4% (15.36 $\rightarrow$ 19.64%). Notably, TEGDA [23] results mirror Classical TTA. In our binary per-case reimplementation, TEGDA's parameter updates did not alter the final binarized masks. We attribute this to differences from TEGDA's original streaming multi-class setting, including binary evaluation, per-case adaptation, and Instance Normalization in nnU-Net v2, rather than a general limitation of

Table 2. Ablation study on the BraTS-PED dataset. Values are Mean ± Std. Best values are in **bold**.

Method	Dice (%) ↑	HD95 (mm) ↓	Prec (%) ↑
w/o Gatekeeper	85.38 ± 10.91	6.19 ± 6.87	**90.19 ± 14.15**
w/o Edge Map	87.22 ± 11.43	5.44 ± 7.08	86.70 ± 12.86
only H_{compact}	87.26 ± 11.33	5.35 ± 6.97	86.32 ± 13.34
only H_{diffuse}	82.18 ± 18.17	9.31 ± 13.21	76.83 ± 23.20
Full HD-TTA	**87.26 ± 11.38**	**5.35 ± 6.95**	86.32 ± 13.40

TEGDA. Despite universally low Dice here, HD-TTA's superior boundary adherence effectively suppresses false positives while recovering valid tumor, aligning with clinical safety goals. Visual analysis (Fig. 3) confirms this, demonstrating that HD-TTA succeeds where baselines fail. Overall, the quantitative and qualitative evidence consistently supports the central claim of HD-TTA: rather than blindly optimizing every test case under a single objective, flagging borderline cases and selectively choosing between competing geometric hypotheses leads to safer boundary behavior. While Dice remains comparable to strong baselines, HD-TTA consistently improves boundary reliability (HD95) and false-positive control (Precision), which are critical in safety-sensitive medical segmentation.

3.3 Ablation Study.

Table 2 details the impact of each component on the BraTS-PED dataset. First, regarding hypothesis selection, we observe that forcing the inflation strategy (only H_{diffuse}) is dangerous on stable cases, causing significant boundary leakage (HD95 spikes to 9.31 mm). In contrast, Full HD-TTA successfully identifies this risk and converges to the performance of the safe H_{compact} strategy (HD95 5.35 mm), proving that the selector correctly rejects unsafe updates. Second, the Gatekeeper prevents negative transfer, while hypothesis selection primarily acts as a safeguard against unstable inflation. Blindly adapting every case (w/o Gatekeeper) forces H_{compact} onto already-accurate predictions. While this aggressive erosion artificially inflates Precision by eliminating all false positives, it degrades overall overlap (Dice 85.38%) and boundary accuracy (HD95 6.19 mm) by deleting valid tumor tissue. Thus, on stable domains like BraTS-PED, safety improvements arise primarily from the Gatekeeper's selective suppression rather than frequent inflation. Finally, removing the edge map constraint worsens boundary adherence (HD95 increases to 5.44 mm), confirming its role as a necessary physical guardrail against leakage.

4 Conclusion

We presented HD-TTA, a decision-oriented inference framework that replaces blind per-case optimization with selective refinement and hypothesis-conditioned updates. In short, HD-TTA operates on a clear directive: flag borderline cases,

hypothesize tailored strategies, and make an unsupervised selection of the "safest" correction. Using a strong nnU-Net v2 backbone trained on BraTS 2023 GLI, we evaluated HD-TTA without any target-domain tuning on BraTS-PED and more challenging BraTS-MEN. We observed consistent improvements in safety-oriented metrics, achieving lower boundary error (HD95) and higher Precision relative to representative TTA baselines, while maintaining comparable Dice. These results along with the qualitative analysis support the core principle that explicitly structured, competing refinement hypotheses can better control failure modes such as under-segmentation, boundary leakage and spurious islands under domain shift. For multi-class extension, the framework instantiates K hypothesis pairs $\{H_k^{\text{compact}}, H_k^{\text{diffuse}}\}$ with class-conditional selection scores S_{rep}^k, inter-class exclusivity constraints, and modality-specific geodesic barriers (T1-Gd for core, FLAIR for oedema). Fundamentally, this unsupervised principle of diagnosing risks and tailoring remedies generalizes to other modalities, such as temporal adaptation in videos. Clinically, HD-TTA can automatically correct anomalous scans, generating safe pseudo-labels for continuous semi-supervised retraining.

Acknowledgments. This research is supported by the Ministry of Education, Singapore, under its Research Centre of Excellence award to the Institute for Digital Molecular Analytics & Science, NTU (IDMxS, grant: EDUNC-33-18-279-V12).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. The cancer imaging archive **7** (2017)
2. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification (2021), https://arxiv.org/abs/2107.02314
3. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection (2017). https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF
4. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 170117 (2017)
5. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. International journal of computer vision **22**(1), 61–79 (1997)
6. Fathi Kazerooni, A., et al.: BraTS-PEDs: Results of the multi-consortium international pediatric brain tumor segmentation challenge 2023 (2024), https://arxiv.org/abs/2407.08855
7. Isensee, F., Jaeger, P.F., Kohl, S.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)

8. Karani, N., Erdil, E., Chaitanya, K., Konukoglu, E.: Test-time adaptable neural networks for robust medical image segmentation. Medical Image Analysis **68**, 101907 (2021)
9. Kazerooni, A.F., Khalili, N., Liu, X., et al.: The brain tumor segmentation (brats) challenge 2023: Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds) (2024), https://arxiv.org/abs/2305.17033
10. Kim, J., Kwon, H., Kweon, H., Jeong, W., Yoon, K.J.: Dc-tta: Divide-and-conquer framework for test-time adaptation of interactive segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23279–23289 (2025)
11. LaBella, D., Adewole, M., Alonso-Basanta, M., et al.: The asnr-miccai brain tumor segmentation (brats) challenge 2023: Intracranial meningioma (2023), https://arxiv.org/abs/2305.07642
12. LaBella, D., Baid, U., et al.: Analysis of the BraTS 2023 intracranial meningioma segmentation challenge. Machine Learning for Biomedical Imaging **4** (2025), https://arxiv.org/abs/2405.09787
13. Ma, J.: Improved self-training for test-time adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23701–23710 (2024)
14. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
15. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. In: International conference on machine learning. pp. 16888–16905. PMLR (2022)
16. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=g2YraF75Tj
17. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena **60**(1-4), 259–268 (1992)
18. Shanmugam, D., Blalock, D., Balakrishnan, G., Guttag, J.: Better aggregation in test-time augmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1214–1223 (2021)
19. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=uXl3bZLkr3c
20. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7201–7211 (2022)
21. Yang, C., Guo, X., Chen, Z., Yuan, Y.: Source free domain adaptation for medical image segmentation with fourier style mining. Medical Image Analysis **79**, 102457 (2022)
22. You, L., Lu, J., Huang, X.: Test-time correlation alignment. In: International Conference on Machine Learning. pp. 72700–72729. PMLR (2025)
23. Zhou, Y., Wu, J., Liao, W., Zhang, S., Zhang, S., Wang, G.: Tegda: Test-time evaluation-guided dynamic adaptation for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 628–637. Springer (2025)