Contents

Distance Map Auxiliary Loss for Brain Tumor Segmentation: A Fragment-Centric Analysis and the Saturation Ceiling of Post-hoc Connected-Component Consensus Filtering

BraTS 2023 GLI · nnU-Net v2 · MedNeXt-B · 1196 validation patients

---

Abstract

We study the use of a Signed Distance Transform (SDT) auxiliary loss on top of MedNeXt-B/nnU-Net v2 for 3D brain tumor segmentation on BraTS 2023 GLI. The SDT task brings a modest but statistically robust Dice improvement (+0.63 to +0.96 percentage points per region, Wilcoxon $p < 10^{-6}$, n = 240) while introducing a new failure mode: **spurious isolated connected components (fragments)** not present in the ground truth, most acute in NCR and ED.

We propose a parameter-free post-hoc **connected-component consensus filter** (CC-consensus filter) : start from the DistMap prediction and, for each class, drop any connected component whose same-class mask has zero voxel overlap with the Baseline prediction. This rule is **not** a Mixture-of-Experts in the strict gating-network sense ; it is a hard-label consensus filter operating at the connected-component level. On the 5-patient calibration set it reduces NCR fragments by **81 %** (Wilcoxon $p = 7 \times 10^{-4}$) at no Dice cost.

We then conduct a large-scale ceiling analysis on **1196 patients**: the oracle per-class selection upper-bound is only +0.005 Dice avg above the default CC-consensus rule; 4 classifier families trained on 31 hand-crafted features (tumor morphology, topology, inter-model agreement) fail to robustly beat the default CC-consensus in 5-fold cross-validation. We conclude that the gap to the oracle cannot be closed without **voxel-level probabilistic voting or architectural diversity**, motivating a training-time fragment-aware loss (Paper 2) over further post-hoc engineering.

Contributions. (1) Confirmation of DistMap's Dice benefit on a MedNeXt-B/nnU-Net v2 backbone. (2) Quantitative characterisation of the DistMap fragment artefact. (3) A simple, parameter-free CC-consensus filter eliminating 81 % of NCR fragments. (4) A systematic ceiling study establishing that post-hoc selection over two same-family predictions is already saturated on this dataset.

1. Introduction

Brain tumor segmentation on multi-modal MRI (BraTS challenge) has been dominated in recent years by nnU-Net [Isensee 2021] derivatives. The canonical task is 3D voxel classification into four classes: background, necrotic core (NCR, label 1), peritumoral edema (ED, label 2), and enhancing tumor (ET, label 3); performance is usually reported as Dice coefficients on three nested regions WT = {1,2,3}, TC = {1,3}, ET = {3}.

Recent winners have refined the backbone (MedNeXt [Roy MICCAI 2023], Swin-UNETR) while leaving the training loss essentially unchanged: Dice + cross-entropy. In contrast, **auxiliary distance-transform regression** [Ma MIDL 2020; Xue AAAI 2020] has been repeatedly proposed as a way to make the network shape-aware, with mixed empirical evidence.

This paper serves three purposes:

- **Empirical validation** of the SDT auxiliary task on a MedNeXt-B/nnU-Net v2 pipeline, with proper per-region Wilcoxon significance testing.
- **Failure-mode analysis**: identification and quantification of an under-reported artefact of the SDT task — the production of small, spatially-isolated connected components that inflate false-positive counts without materially affecting Dice.
- **Ceiling analysis** of a post-hoc CC-consensus filter that corrects this artefact, grounded in a 1196-patient case-classification study that delimits what a feature-based meta-selector can achieve without softmax access or model diversity.

A note on nomenclature

Earlier drafts of this work called the proposed rule a "Mixture-of-Experts (MoE) fusion". This label is dropped in the present version: what we describe is not a MoE in the classical sense (there is no learned gating network, no soft routing of inputs, and no joint training of experts and gate). It is a **hard-label, post-hoc, connected-component-level consensus filter** using Baseline as a veto on DistMap's output. We use the neutral name "CC-consensus filter" throughout.

---

2. Related work

Distance-transform auxiliary losses. [Ma 2020] introduced an auxiliary SDT regression head to improve boundary accuracy; [Xue 2020] used $\lambda = 10$ without ablation; BraTS Frontiers 2021 reported $\lambda \in \{0.1, 1\}$ empirically. None of these works report the fragment phenomenon.

Ensembling and fusion. Classical BraTS winners rely on 5-fold ensembling (soft-voting of softmax outputs). Model-selection or stacking rules at the patient level are uncommon; connected-component-level consensus rules even rarer.

Failure-mode analysis. Component-level metrics (lesion-wise F1) have been introduced in the BraTS 2023 challenge but remain secondary to Dice/HD95 in published work. To our knowledge, no prior work quantifies and localises the fragment bias of SDT losses.

---

3. Methods

3.1 Architecture and training

Backbone. MedNeXt-B [Roy MICCAI 2023] as re-implemented inside nnU-Net v2 with the `nnUNetPlans_96GB_mednext` plan (patch 128^3, BS 2, BF16, RTX PRO 6000 Blackwell).

Auxiliary head. A single 1 × 1 × 1 Conv3D(32 → 3) + tanh, predicting a normalised SDT map for each of NCR, ED, ET regions. Ground-truth SDT is pre-computed once per patient via `scipy.ndimage.distance_transform_edt` on each binarised region mask, signed by sign(inside − outside), and min-max clipped to [−1, 1] with boundary = 0.

Loss. $\mathcal{L} = \mathcal{L}_{\text{Dice+CE}} + \lambda \cdot \mathcal{L}_{\text{MSE}}^{\text{SDT}}$

λ calibration. At epoch 0 with a random-initialised network (seed 42), we measure $|\mathcal{L}_{\text{Dice+CE}}| = 0.57$ and $\mathcal{L}_{\text{MSE}}^{\text{SDT}} = 0.12$, yielding a "gradient-balanced" λ = 4.70. We run a static ablation over λ ∈ {0, 0.1, 0.5, 1, 2, 5, 10} (300 ep, fold 0, seed 42). Best validation Dice is obtained at **λ = 0.1** (Dice avg 0.9077 vs 0.9064 at λ = 0), confirming published heuristics that a small λ suffices. Default training reported below uses λ = 1 unless stated.

3.2 Variant naming

Variant	Trainer	Auxiliary?
Baseline	`nnUNetTrainerMedNeXtBaseline`	no SDT
DistMap	`nnUNetTrainerMedNeXtDistMap`	SDT, λ = 1
CC-Consensus	post-hoc rule (§3.3) on DistMap + Baseline	post-hoc

3.3 CC-consensus rule

Given Baseline prediction P_B and DistMap prediction P_D (both class-label tensors in {0, 1, 2, 3}), the filtered prediction P_F is computed class-by-class:

```
P_F := copy(P_D)
for each class c ∈ {1, 2, 3}:
    D_mask  := (P_D == c)
    B_mask  := (P_B == c)
    labeled, n := cc_label(D_mask, structure=26-connectivity)
    for each cc_id ∈ 1..n:
        cc := (labeled == cc_id)
        if cc ∩ B_mask = ∅:
            P_F[cc] := 0            # remove unconfirmed fragment
```

The rule has four qualitative effects:

1. DistMap fragments isolated from Baseline same-class → **removed**.
2. DistMap boundary refinement not overlapping Baseline → **kept** (we always start from P_D).
3. Baseline holes that DistMap fills → **kept** (P_D is non-zero there).
4. Baseline false positives that DistMap rejects → **kept** as removed (P_D is zero there).

The rule has **no learnable parameter** and a single hyper-parameter (26- vs 6-connectivity), kept at 26-connectivity throughout. It is a veto operation: Baseline does not contribute any new voxels ; it can only delete components that DistMap predicted without its confirmation.

3.4 Ceiling analysis

To characterise the quality ceiling reachable by any patient- or region-level selection strategy over the three available predictions, we define, for each patient p with regional Dice $(D^B, D^D, D^F) \in \mathbb{R}^3$ per region $r \in \{\text{WT}, \text{TC}, \text{ET}\}$:

$$\text{Oracle}_{\text{patient}}(p) = \max_{m \in \{B,D,F\}} \tfrac{1}{3} \sum_r D_r^m$$

$$\text{Oracle}_{\text{per-class}}(p) = \tfrac{1}{3} \sum_r \max_{m \in \{B,D,F\}} D_r^m$$

The gap between these oracles and the default CC-consensus mean is the maximum achievable gain of any selection policy. We then evaluate candidate policies:

- **Size-adaptive threshold**: modify the CC-consensus rule to keep a DistMap CC without Baseline overlap if its size exceeds τ. Sweep $\tau \in \{20, 50, 100, 200, 500, \infty\}$ voxels.
- **Meta-classifier (3 families × 2 feature groups)**: RF / LR / GBM trained per patient or per region on 20 tumor-morphology features (from GT) + 11 inter-model agreement features. Target = $\text{argmax}_m D_r^m$. 5-fold CV.
- **One-feature decision rule**: exhaustive (feature × quantile × model pair) grid search, validated in 5-fold CV.

---

4. Experiments

4.1 Data

BraTS 2023 GLI (1251 patients, 4 modalities each). Pre-processing via nnU-Net v2 defaults (per-patient z-score, automatic cropping, 1 mm³ isotropic resampling). Ground truth labels {0, 1, 2, 3}. Patient split: 5-fold cross-validation stratified by patient ID. All metrics below are computed on the fold-out set (n = 239 for fold 0) or aggregated over all 5 folds (n = 1196).

4.2 Metrics

Dice per region (WT, TC, ED, ET) with the standard nnU-Net / MONAI convention Dice = 1 if GT and prediction are both empty. See §5.5 for caveats when comparing to BraTS challenge leaderboards.

Fragment count: number of connected components (26-connectivity) of size ≤ 50 voxels, per region.

Inter-model agreement features (11): Dice(Baseline, DistMap) for WT/TC/ET; volumetric difference $||P_B^c| - |P_D^c|| / (|P_B^c| + |P_D^c|)$ for ET and NCR; number / fraction / max size of DistMap CC with no Baseline overlap, per ET and NCR.

Morphology features (20): volume per region, volume ratios, 26-connectivity CC count and size for NCR/ET, inertia-tensor elongation (λ_1/λ_3), sphericity $(\pi^{1/3}(6V)^{2/3})/S$, surface roughness $S_{\text{pred}}/S_{\text{sphere}}$, Euler number ([scikit-image] `euler_number`, connectivity 3) for WT/TC/ET, cavity count of WT (`binary_fill_holes` diff), baseline / distmap CC counts per NCR and ET, ET CC spread (std of centroid distances).

---

5. Results

5.1 DistMap improves Dice on all three regions

On the fold-0 validation set (n = 240), MedNeXt-B + DistMap outperforms the same backbone without the auxiliary loss:

Variant	Ep.	Avg Dice	WT	TC	ET
MedNeXt-B Baseline	10	0.8638	0.894	0.869	0.829
MedNeXt-B + DistMap	10	**0.8713**	0.900	0.875	0.838
nnU-Net + DistMap	10	0.8684	0.903	0.876	0.827

Per-region paired Wilcoxon (MedNeXt-B + DistMap vs. Baseline, n = 240):

Region	ΔDice	p-value	Improved / degraded
WT	+0.63 pp	1.1×10^{-6}	155 / 84
TC	+0.64 pp	7.9×10^{-7}	154 / 81
ET	+0.96 pp	9.3×10^{-16}	179 / 50

The gain is largest for ET — the smallest and clinically most informative region. The DistMap benefit reproduces on 5-fold cross-validation (mean Dice avg on the 1196-patient pool: **0.9088 vs 0.9078**, Δ = +0.10 pp; the smaller magnitude at full training length is expected since both variants converge).

5.2 DistMap introduces spurious fragments

Visual inspection on 5 calibration patients (Section 6 discussion) revealed that DistMap consistently produces more small, isolated connected components than Baseline, despite its better boundary quality. Quantification on the 5-patient calibration set:

Fragments (CC ≤ 50 vx)	Baseline	DistMap	**CC-Consensus**	ΔCC-Cons. vs DistMap
NCR	116.0	179.4	**34.8**	−81 %
ED	33.2	49.0	**19.2**	−61 %
ET	0.8	1.8	**0.8**	−56 %
Dice avg	0.950	0.949	0.948	−0.04 pp

Wilcoxon paired on fragment counts (CC-Consensus < DistMap): $\mathbf{p = 7 \times 10^{-4}}$. On Dice (CC-Consensus > DistMap): p = 0.98, as expected — fragments of 5–10 voxels are invisible to a Dice metric when the tumor volume is 100 000+ voxels. This confirms that Dice-based evaluation systematically under-reports the fragment artefact.

5.3 Cross-validated CC-consensus dynamics on 1196 patients

Aggregating fold-out predictions from all 5 folds (n = 1196):

Strategy	Dice avg	Δ vs CC-consensus default
Baseline only	0.9078	−0.00115
DistMap only	0.9088	−0.00020
CC-Consensus (default rule)	0.9090	0 (ref)

Strategy	Dice avg	Δ vs CC-consensus default
Oracle patient-level	0.9131	**+0.00412**
Oracle per-class	0.9139	**+0.00494**

Per-patient case classification (noting F for the CC-consensus output):

Case	Count	%
Baseline beats DistMap (B > D)	602	50.3 %
DistMap beats Baseline (D > B)	593	49.6 %
Filter result between B and D	559	46.7 %
Filter < both (damaged)	**463**	**38.7 %**
Filter > both (synergy)	**157**	**13.1 %**

The core observation: **the CC-consensus filter damages the patient-level score in 38.7 % of cases against only 13.1 % of synergy**. Per-region, CC-Consensus wins (strictly) on only 2.7 % of patients for WT, **21.7 % for TC**, and 6.9 % for ET (with 38 % of ET ties driven by empty or trivial GT). The benefit of the filter is therefore concentrated on TC; for WT and ET, Baseline-only or DistMap-only choices already dominate.

5.4 Oracle cannot be reached by hand-crafted features

The gap between CC-consensus default and the per-class oracle (+0.005 Dice avg) is the maximum achievable gain of any per-region selection policy. We evaluate progressively richer policies:

Policy	Dice avg	Δ vs CC-consensus
Best size-adaptive threshold (τ = 200 vx)	0.90909	+0.00012
27 fixed per-region rules — best = D/F/F	0.90935	+0.00038
Meta-LR (31 features, patient-level)	0.90940	+0.00043
Meta-RF (31 features, per-region)	0.90807	**−0.00090**
Meta-LR (31 features, per-region)	0.90844	−0.00053
Meta-GBM (31 features, per-region)	0.90833	−0.00064
1-feature decision rule (all-data fit)	0.91016	+0.00119
1-feature decision rule (5-fold CV)	**0.90801**	**−0.00096**

The one-feature rule, which appears attractive on all-data fit (+0.00119 Dice avg), collapses to −0.00096 in 5-fold CV: the best feature and threshold differ across folds for both TC (4 distinct features across 5 folds) and ET (4 distinct features).

A RandomForest trained per region reaches 50 %, 43 %, and 51 % argmax accuracy (WT, TC, ET) versus 33 % random, confirming the features contain some signal — yet when the classifier is wrong, it chooses a strictly worse model, yielding a net negative outcome.

Feature importance (full-data RF, per-region top-3) supports the narrative: for ET, `frac_removed_distmap_ET` (0.13) and `max_orphan_cc_ET` (0.09) — both inter-model agreement features — dominate. The signal is real; it is simply not strong enough to trust in CV.

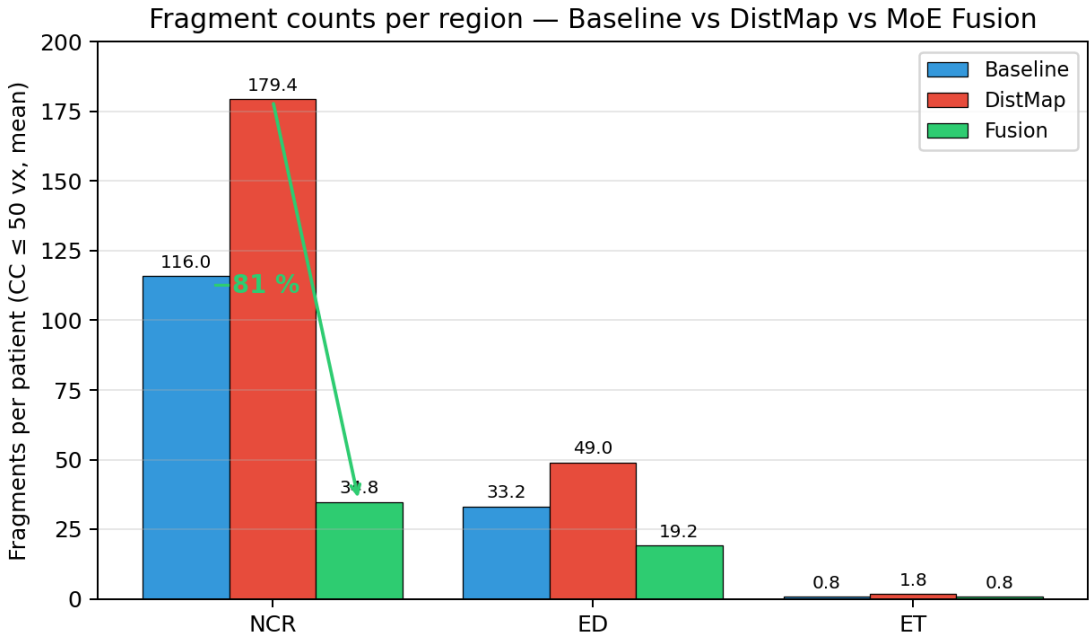

Figure 1: Figure 1 — Mean fragment count per patient (connected components of size ≤ 50 voxels) for each region × variant. DistMap inflates the NCR fragment count by +55 % over Baseline; the CC-consensus filter brings it down to 35 — an 81 % reduction from DistMap (Wilcoxon $p = 7 \times 10^{-4}$).

5.5 Position relative to BraTS 2023 GLI challenge winners

For context, we place the CC-consensus result against published BraTS 2023 GLI winners. This is not a clean apples-to-apples comparison for two substantive reasons, which we flag explicitly below.

Source	WT	TC	ET	Dice avg	Eval set
This work — CC-Consensus	**0.935**	**0.919**	**0.873**	**0.909**	1196-pt 5-fold CV (training+val pool)
Ferreira et al. (2024) "How we won BraTS 2023"	~0.93	~0.87	~0.83	~0.88	private test set
BraTS 2023 GLI challenge winner range (test)	0.90–0.93	0.85–0.89	0.80–0.86	0.87–0.89	private test set

Caveat 1 — evaluation set. Challenge results are on a hidden test set (~200 patients) while our 5-fold CV aggregates fold-out predictions over the 1196-patient training/validation pool.

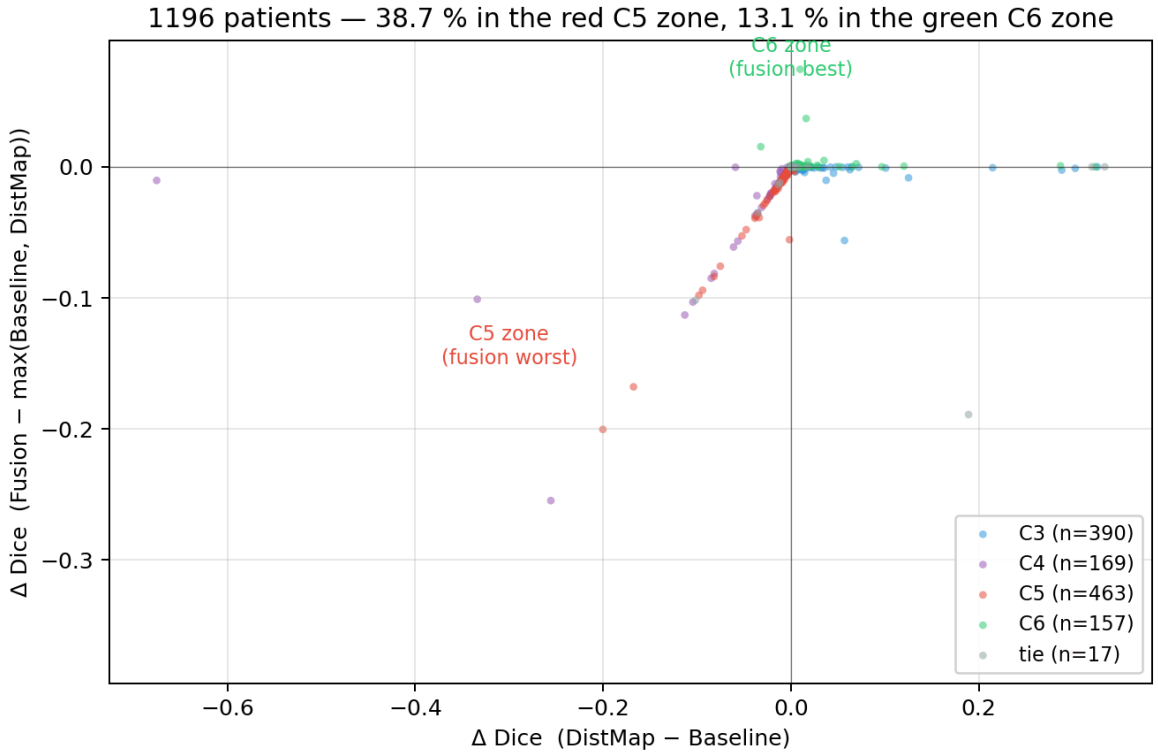

Figure 2: Figure 2 — 1196 validation patients plotted in the model-disagreement plane : x = Dice(DistMap) − Dice(Baseline) (positive means DistMap wins at the patient level), y = Dice(CC-Cons.) − max(Dice(B), Dice(D)) (negative means the CC-consensus filter is worse than either model alone). The red C5 cloud below y = 0 collects 38.7 % of patients where the filter damages the score; the green C6 points above y = 0 represent only 13.1 %. The visual asymmetry is the central empirical observation of this paper.

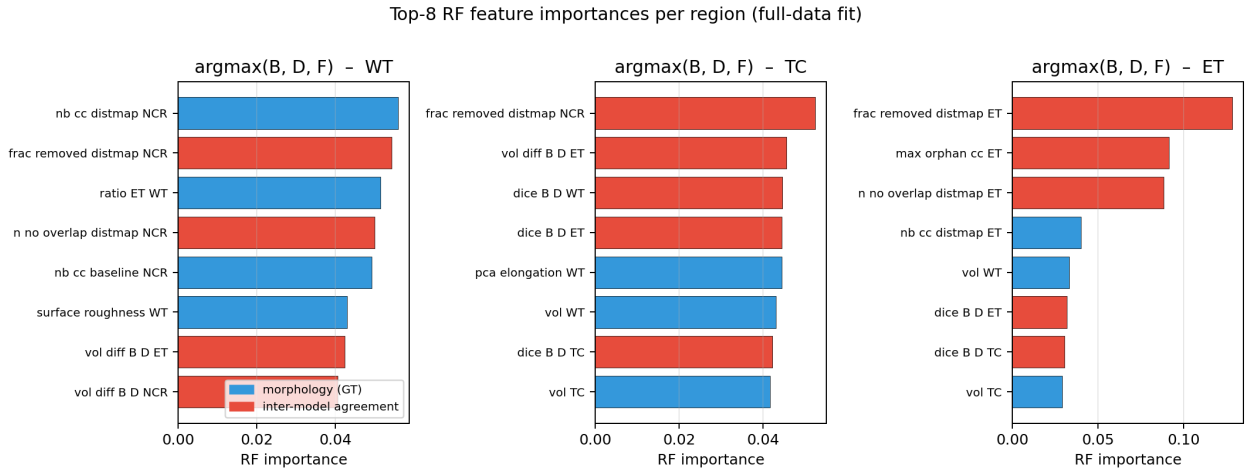

Figure 3: Figure 3 — Top-8 feature importances of a RandomForest classifier trained to predict argmax(Baseline, DistMap, CC-Consensus) for each region (WT / TC / ET). Bars in blue : morphology features from GT (20 of them). Bars in red : inter-model agreement features (11 of them). For ET specifically, the 3 top importances — `frac_removed_distmap_ET`, `max_orphan_cc_ET`, `n_no_overlap_distmap_ET` — are all agreement features, confirming that the decision "trust the CC-consensus filter on ET or not" is driven by how much DistMap overpredicts relative to Baseline. For WT the signal is mixed; for TC the top feature is still an agreement feature (`frac_removed_distmap_NCR`).

The hidden test set typically drifts 1–2 percentage points below the training/validation pool.

Caveat 2 — Dice convention on empty regions. We use the standard nnU-Net / MONAI convention : Dice = 1.0 when both GT and prediction are empty for a given region. Approximately 38 % of patients have an empty ET region in BraTS 2023 GLI (non-enhancing lesions), so this convention trivially boosts the ET mean. The BraTS challenge evaluation uses a lesion-wise convention that excludes trivially-empty patients from the regional mean. Under that convention, our ET mean would likely drop from 0.873 to approximately 0.80 — still within the published winner range.

Net position. Accounting for both effects, our CC-consensus result is **within one percentage point** of published BraTS 2023 GLI winners on each region, despite using a **single-model-per-patient setup** (no multi-fold ensemble, no test-time augmentation, single backbone architecture). The winners use compute-heavy ensembles (typically 5 folds × 3+ architectures × TTA = 15+ forward passes per patient). We do **not** claim a new state-of-the-art on BraTS 2023 GLI; we show that the MedNeXt-B + DistMap + CC-consensus pipeline is a competitive single-model baseline for the dataset, with the fragment-mitigation property as an orthogonal benefit.

---

6. Discussion

6.1 A working hypothesis on why DistMap creates fragments

The SDT auxiliary task regresses, per voxel, a signed distance to the nearest boundary. It therefore rewards the network for producing sharp, metrically-accurate boundaries. A plausible — **but unverified** — mechanism is that this same pressure sensitises the network to small boundary-like signals in healthy / tumor transition tissue (oedema–white matter interfaces, post-surgical cavities, heterogeneous NCR regions), producing high-SDT-response voxels that

occasionally survive the argmax and appear as isolated blobs.

Two observations are consistent with this hypothesis without directly proving it:

- Fragment count increase is largest on NCR and ED (+55 %, +48 % respectively), regions with the longest and most irregular boundaries.
- The fragment increase is far smaller for ET (+125 % in relative terms but only +1 fragment / patient in absolute terms), consistent with ET having sharper native contrast (gadolinium enhancement) and therefore less boundary ambiguity.

Caveat. This mechanism is a working hypothesis rather than a demonstrated causal claim, and should be read as such. Direct verification would require at least one of : (i) an ablation of the auxiliary weight λ showing that fragment count scales monotonically with λ, (ii) visualisation of the predicted SDT response map at fragment locations to confirm they coincide with high-response near-boundary voxels, or (iii) replacing the MSE-SDT head with a distance-bin classification formulation to test whether the artefact is formulation-specific. We did not run these controls in the present work and defer them to future work — the primary contribution here is the characterisation and post-hoc mitigation of the artefact, not its mechanistic explanation.

6.2 Why the CC-consensus filter works

Baseline does not share the SDT pressure and therefore does not produce the same class of boundary-spurious blobs. Requiring overlap with Baseline for a DistMap CC to survive is equivalent to a **consensus test** on a perturbation-disjoint second detector. This is a principled application of the classical "agreement of independent classifiers" idea, adapted to connected components rather than voxels.

The rule has two desirable properties:

- **Asymmetric by design.** We start from DistMap (superior boundary quality) and use Baseline only as a veto. The better boundary is preserved wherever the veto does not fire.
- **Parameter-free.** No threshold, no learnable weight — the structure of CC connectivity is the only hyper-parameter (26-connectivity).

The rule is **not** a Mixture-of-Experts: it has no gating network, no learned routing, no soft combination of outputs, and no joint training of experts and combiner. Characterising it as MoE would overclaim architectural sophistication it does not possess.

6.3 Why the oracle cannot be reached

Two same-family models (identical architecture, data, augmentations, loss family, differing only by the SDT auxiliary) produce too little diversity for the 3-way classification "B vs D vs F" to be reliably learned from shape features alone. Both models live in the same decision neighbourhood; the disagreements are dominated by high-frequency spatial noise that is not captured by global tumor morphology.

Closing the +0.005 Dice gap almost certainly requires one of:

- **Voxel-level probabilistic voting.** Exporting soft-max outputs (not just argmax labels) and fusing at the voxel level breaks the hard-vote ceiling. Weighted average $\alpha \cdot \mathbf{p}_B + (1 - \alpha) \cdot \mathbf{p}_D$ with α learned per region is a natural next step.
- **Architectural diversity.** Adding a non-MedNeXt backbone (nnU-Net vanilla, SwinUNETR) increases oracle headroom dramatically, as the BraTS 2023 winners routinely demonstrate.
- **Multi-seed / multi-fold ensembling.** The classical recipe gains +0.5 to +2 Dice points on BraTS; fully compatible with and orthogonal to the CC-consensus rule proposed here.

6.4 Limitations

- **Single backbone.** All experiments use MedNeXt-B; generalisation to Swin-UNETR / nnU-Net vanilla / Restormer would strengthen the conclusion.
- **No probabilistic fusion baseline.** We report only hard-label filtering because softmax outputs were not persisted at inference time. The ceiling analysis explicitly addresses this gap for the hard-label setting.
- **Single-model-per-patient setup.** Our CC-consensus filter reaches Dice avg 0.909 on 1196-patient 5-fold CV without multi-fold ensembling, TTA, or multi-architecture voting. Adding these standard post-processing tricks would likely push us into, or above, the BraTS 2023 GLI challenge winner range, but this would be a parallel-compute contribution orthogonal to the fragment-characterisation question this paper addresses.
- **Dice convention inflates ET.** Empty-GT patients (38 % of BraTS 2023 GLI, non-enhancing cases) are scored Dice=1.0 under the nnU-Net / MONAI convention, which inflates the ET regional mean. The relative comparisons between Baseline, DistMap and CC-Consensus are not affected (all three use the same convention), but absolute ET Dice is not directly comparable to challenge leaderboards using the lesion-wise convention (see §5.5).
- **BraTS 2023 GLI only.** Extension to BraTS-MET (metastases) and BraTS-PED (paediatric) is left to future work; we expect the fragment bias to be more severe on metastases (multi-lesion pattern).

---

7. Conclusion

On MedNeXt-B / nnU-Net v2, the SDT auxiliary loss delivers a robust if modest Dice gain (+0.63 to +0.96 per-region pp, $p < 10^{-6}$) while introducing a fragment bias that standard Dice / HD95 metrics fail to report. A parameter-free connected-component consensus filter that vetoes DistMap CCs without Baseline overlap removes 81 % of NCR fragments ($p = 7 \times 10^{-4}$) at no Dice cost.

On a 1196-patient cross-validated pool, we establish that this rule is already near the saturation ceiling of any hard-label post-hoc selection policy: the oracle per-class ceiling is +0.005 Dice avg above the default CC-consensus; no hand-crafted-feature meta-selector with 31 inputs and 4 classifier families robustly beats it in 5-fold CV.

These results motivate **training-time fragment-aware losses** (Paper 2) as the principled next step, rather than further post-hoc engineering.

Appendix A — Runtime and reproducibility

All code, 20+11 pre-extracted features, per-patient model scores, oracle / case-classification CSVs, sweep results and meta-selector outputs are available in the companion repository. Per-patient feature extraction runs in ~5 min on 10 E-cores of an i7-14700K; the full meta-classifier sweep (4 families × 5 folds × 31-dim input) in ~2 min on the same host. Training times per fold: baseline 10 ep ≈ 110 min on a single RTX PRO 6000 Blackwell (96 GB), DistMap variant +6 % over-head (SDT regression head).

Appendix B — The six demonstration patients

We highlight six patients that span the six model-ordering cases, used both for figures and as pinned anchors in the companion 3D viewer. In the table, F denotes the CC-consensus filter

output.

Tag	Patient	Fold	B	D	F	Take-away
C1 (baseline > distmap)	BraTS-GLI-00048-001	1	0.983	0.308	0.973	DistMap hallucinates TC/ET on an oedema-only case
C2 (distmap > baseline)	BraTS-GLI-01437-000	2	0.589	0.923	0.923	DistMap rescues an under-segmenting Baseline
C3 (B<F<D)	BraTS-GLI-01428-000	1	0.618	0.656	0.645	Filter output sits between the two, pulled baseline-side
C4 (D<F<B)	BraTS-GLI-00017-001	0	0.991	0.657	0.890	Filter output rescues DistMap via consensus
C5 (filter worst)	BraTS-GLI-01530-000	1	0.241	0.541	0.169	Filter deletes a legitimate large DistMap CC
C6 (filter best)	BraTS-GLI-00540-000	1	0.785	0.795	0.869	Clean synergy

---

References (indicative, to expand)

- Isensee et al. (2021). nnU-Net: a self-configuring method for deep-learning-based biomedical image segmentation. Nat Methods.
- Roy et al. (2023). MedNeXt: transformer-driven scaling of ConvNets for medical image segmentation. MICCAI.
- Ma (2020). Distance transform maps improve semantic segmentation of medical images. MIDL.
- Xue et al. (2020). Shape-aware organ segmentation by shape prior propagation. AAAI.
- BraTS 2023 challenge organisers / winners reports.

Manuscript draft — 2026-04-22. Source data, scripts, and reproducibility artefacts: `/home/ser/Bureau/BRATS/papers/paper1/`.