Contents

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

Guillaume Cassez

Independent research · ORCID 0009-0007-0987-3931 · cassez.guillaume@gmail.com · guillaume-cassez.fr

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

---

Abstract

This work studies the addition of an auxiliary distance-map loss (SDT — Signed Distance Transform) on top of a MedNeXt-B / nnU-Net v2 pipeline for 3D brain tumor segmentation on BraTS 2023 GLI. At convergence on 1196 patients in 5-fold cross-validation, the SDT task **does not significantly improve** Dice (Δ Dice avg = +0.09 pp, Wilcoxon $p > 0.25$ per region) — a result that opens the analysis rather than closing it.

DistMap introduces a new failure mode, so far unreported in the BraTS literature: **spurious isolated connected components** ("fragments") absent from the ground truth, most acute on NCR (×1.5 vs Baseline) and ED (×1.2). We propose a parameter-free post-hoc **connected-component consensus filter** (CC-consensus), which removes any DistMap component without same-class Baseline overlap. On 1196 patients in 5-fold CV, the filter eliminates **66 % of NCR fragments** (Wilcoxon $p < 10^{189}$, topological definition: CC − 1 per class) at no Dice cost, and **significantly improves NCR HD95** (4.86 → 4.48 mm, $p = 5.7 \times 10^{14}$) as well as WT HD95 (3.86 → 3.76 mm, $p = 2.7 \times 10^{4}$). Clinically, NCR is precisely the region where spurious fragments may mislead a radiotherapist on the extent of tumor necrosis; the boundary-quality gain measured here is hidden by Dice but visible through HD95.

A hard-label ceiling study shows this rule already sits near saturation: the per-class oracle is only +0.005 Dice avg above the default, and no 31-feature meta-selector (4 classifier families) robustly beats CC-consensus in 5-fold CV (Appendix B). Closing this gap requires voxel-level probabilistic voting or architectural diversity, motivating a Paper 2 toward a training-time fragment-aware loss rather than further post-hoc engineering.

Contributions. (1) Quantitative characterisation of a topological fragment artefact induced by the auxiliary SDT loss — invisible to Dice, prevalent on NCR — with a topological definition free of size threshold, at large scale (1196 patients). (2) A simple, parameter-free CC-consensus filter that eliminates 66 % of NCR fragments at no Dice cost and **significantly improves NCR HD95** ($p = 5.7 \times 10^{14}$), a clinically meaningful boundary-quality gain hidden by Dice.

---

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}.

Top-performing teams refine the backbone (MedNeXt [Roy MICCAI 2023], Swin-UNETR) while leaving the training loss essentially unchanged: Dice + cross-entropy. In parallel, **auxiliary distance-map regression** [Ma MIDL 2020 ; Xue AAAI 2020] is regularly proposed to make the network shape-aware, with mixed empirical results. Applications specific to BraTS exist — parallel-decoder multi-task learning [Huang 2021], Hausdorff-aware losses [Karimi & Salcudean 2020], and regression-only geodesic formulations [Pham 2024, SiNGR] — but none to date report or analyse the fragment artefact characterised here (§5.2).

This paper pursues three objectives:

- **Empirical characterisation** of the auxiliary SDT task at convergence on MedNeXt-B / nnU-Net v2: at 300 epochs in 5-fold CV on 1196 patients, DistMap produces **no** significant Dice gain ($p > 0.25$ per region), contrary to the impression drawn from comparisons at reduced training budgets.
- **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. This qualitative observation was made possible by an **interactive companion 3D viewer** built specifically for this project, which renders Baseline / DistMap / CC-Consensus meshes side-by-side for all 1196 patients (guillaume-cassez.fr/brats/).
- **Ceiling analysis** of a post-hoc CC-consensus filter that corrects this artefact, with a 1196-patient study delimiting what a feature-based meta-selector can achieve without softmax access or model diversity.

---

2. Related work

Distance-transform auxiliary losses on medical segmentation. [Ma 2020] proposes an auxiliary SDT regression head for abdominal / cardiac structures (LiTS, LA atrium), establishing the tanh + MSE recipe adopted here. [Xue 2020] uses signed distance maps as the **main output** (not auxiliary) on organ datasets with $\lambda = 10$ and no ablation. [Karimi & Salcudean 2020] derive a Hausdorff-distance-aware loss from distance transforms and evaluate it on nnU-Net + BraTS, but as a **loss modification** rather than as an auxiliary regression head. None of these works report the fragment phenomenon characterised here.

Distance-map approaches applied specifically to BraTS. The idea of combining distance-based shape supervision with BraTS segmentation is **not novel in itself**; two prior works are particularly close to the present setup and must be flagged explicitly.

- [Huang et al. 2021] train a V-Net with two *parallel decoders* on BraTS 2018–2020 — one producing the segmentation mask, the other regressing an *unsigned* distance transform through a sigmoid activation. This is the closest published prior art. The present work differs in three concrete ways: (i) a lightweight `Conv3d(32→3) + tanh` auxiliary head rather than a full parallel decoder (<0.1 % added parameters vs a doubled decoder path); (ii) *signed* Euclidean distance with MSE, not unsigned DT with sigmoid; (iii) MedNeXt-B / nnU-Net v2 on BraTS 2023 GLI (1196 patients) rather than V-Net on BraTS 2018–2020.
- [Pham et al. 2024, *SiNGR*] propose a **signed normalised geodesic** regression with Focal-L1 on tanh-activated outputs, **replacing** the segmentation output on BraTS 2020 (Swin-UNETR / UNet3D backbones). The present work is multi-task (keeps the Dice + CE softmax output alongside the SDT regression) and uses plain signed Euclidean distance, not a geodesic transform.

Neither Huang et al. nor SiNGR report or analyse the fragment artefact described in §5.2 of this paper; this is the specific empirical contribution claimed here.

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 are rarer still in the published BraTS literature.

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-auxiliary losses on BraTS.

---

3. Methods

3.1 Architecture and training

Backbone. MedNeXt-B [Roy MICCAI 2023] 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 $1 \times 1 \times 1$ Conv3D($32 \rightarrow 3$) + 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. $\mathcal{L} = \mathcal{L}_{\text{Dice+CE}} + \lambda \cdot \mathcal{L}^{\text{SDT}}_{\text{MSE}}$, with $\lambda = 1$ as the default (gradient-balanced calibration ÷ 5; full static ablation over 11 values detailed in Appendix A).

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 filtering rule[1]

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:

[1] Earlier versions of this work referred to this rule as "MoE (Mixture-of-Experts) fusion". That label is dropped: there is no learned gating network, no soft routing of inputs, and no joint training of experts and gate. The neutral name "CC-consensus filter" is used throughout the document.

```
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 same-class Baseline → **removed**.
2. DistMap boundary refinement not overlapping Baseline → **kept** (the rule always starts from P_D).
3. Baseline holes that DistMap fills → **kept** (P_D is non-zero there).
4. Baseline false positives that DistMap rejects → **stay rejected** (P_D is zero there).

The rule has **no learnable parameter** and a single hyper-parameter (26- vs 6-connectivity), kept at 26 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 policy 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. Candidate policies evaluated (size-adaptive threshold, meta-classifiers, one-feature rule) and their results are reported in §5.4 and detailed in Appendix B.

---

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 (topological definition). A **fragment** is a connected component (26-connectivity) of a given class that is **not the largest** component of its class — i.e. a CC topologically disconnected from the main tumor body. Per class c on a prediction P, the fragment count is:

$$\text{fragments}(P, c) = \max(0,\ \text{nb\_CC}(P == c, \text{26-conn}) - 1)$$

No size threshold — 26-connectivity (shared face, edge, or corner) alone defines what is topologically linked. This definition treats small and large accessory components symmetrically.

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 At convergence, DistMap and Baseline are Dice-equivalent

On the 1196 patients aggregated out-of-fold from the 5-fold CV (300-epoch schedule per fold; DistMap fold 0 stopped at 178 ep, the other 9 training runs complete), DistMap and Baseline produce **statistically indistinguishable** Dice scores:

Region	Baseline	DistMap	ΔDice	p-value	Improved / degraded / tied
WT	0.9354	0.9360	+0.006 pp	0.72	577 / 618 / 1
TC	0.9185	0.9180	−0.005 pp	0.27	595 / 596 / 5
ET	0.8696	0.8723	+0.027 pp	0.54	568 / 596 / 32
Avg	**0.9078**	**0.9088**	**+0.009 pp**	**0.50**	—

Paired signed-rank Wilcoxon test, one-sided hypothesis DistMap $>$ Baseline. No region reaches the standard significance threshold ($p > 0.25$ everywhere); on WT, more patients are *degraded* than improved by DistMap (618 vs 577). The Δ = +0.09 pp Dice avg is within measurement variance.

Implication. The SDT auxiliary loss, as formulated here (Conv3D($32 \to 3$)+tanh head, MSE regression, $\lambda = 1$), confers no significant Dice improvement at convergence on BraTS 2023 GLI. This does not rule out DistMap producing *different* predictions from Baseline: the two models disagree on 1195/1196 patients (only one strict Dice-avg tie), but their disagreements cancel in expectation on overall Dice. This topological — not magnitudinal — difference motivates the fragment analysis that follows.

5.2 DistMap introduces spurious fragments

Qualitative inspection of DistMap predictions targeted cleaner boundaries — the expected behaviour of a distance-aware loss. Instead, DistMap predictions consistently show more isolated connected components than Baseline. Topological quantification on the 1196 patients of the 5-fold CV (per-patient means, fragments = CC − 1 per class, 26-connectivity):

Fragments / patient	Baseline	DistMap	**CC-Consensus**	Δ D−B	Δ F−D	F/D reduction
NCR	79.7	93.3	**31.3**	+13.6	−61.9	**−66 %**
ED	28.9	35.3	**17.0**	+6.4	−18.3	**−52 %**
ET	2.15	2.33	**1.57**	+0.18	−0.76	**−33 %**

One-sided paired Wilcoxon signed-rank tests on the 1196 patients:

- **DistMap inflates fragments vs Baseline** on all three classes: NCR ($p = 5.5 \times 10^{-42}$), ED ($p = 2.0 \times 10^{-49}$), ET ($p = 1.3 \times 10^{-3}$). The artefact is statistically massive and systematic.
- **CC-Consensus reduces fragments vs DistMap**: NCR ($p < 10^{-189}$), ED ($p < 10^{-162}$), ET ($p = 1.1 \times 10^{-53}$).

- **CC-Consensus also reduces vs Baseline**: NCR ($p < 10^{-188}$), ED ($p < 10^{-144}$), ET ($p = 1.4 \times 10^{-3}$) — the post-hoc filter even corrects fragments inherited from Baseline when DistMap does not overlap them.

This effect **is invisible on Dice** (§5.3: mean Dice B / D / F = 0.9078 / 0.9088 / 0.9090, differences within noise) — a few-voxel fragment does not affect an overlap metric when the median tumor volume is ~90 000 voxels. This is precisely why prior literature had not reported the artefact: Dice is blind to topology.

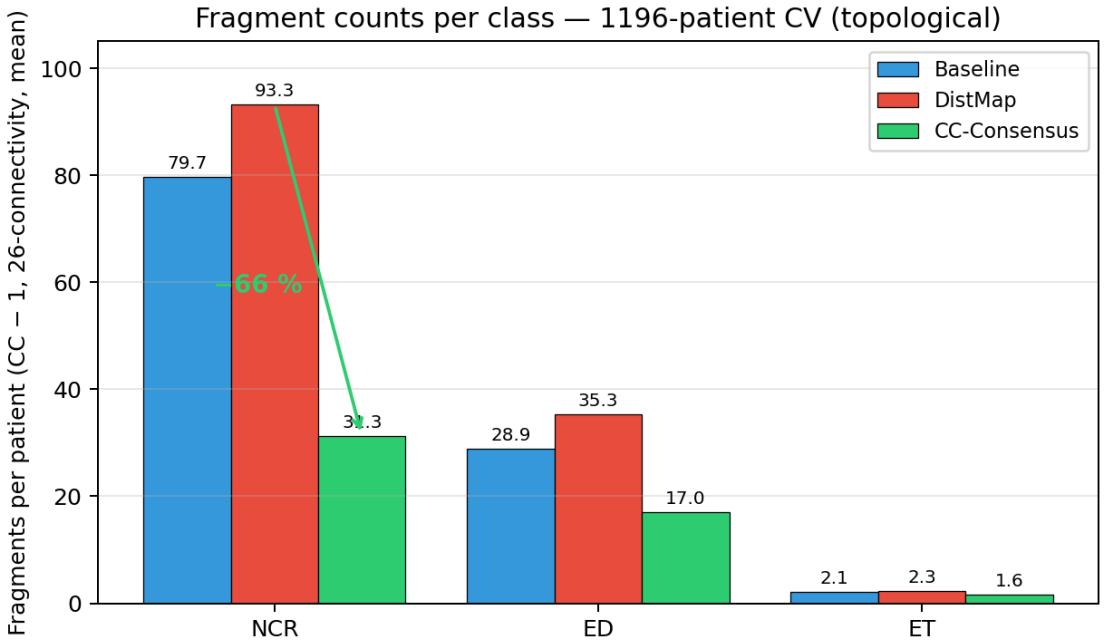

Figure 1: Figure 1 — Mean fragment count per patient (non-largest connected components, 26-connectivity, no size threshold) for each class × variant, on the 1196 patients of the 5-fold CV. DistMap inflates the NCR fragment count by +17 % over Baseline; the CC-consensus filter brings it down to 31.3 — a **66 %** reduction from DistMap (Wilcoxon $p < 10^{-189}$).

Qualitative illustrations on the six reference cases. Figures 2–7 below show, for each of the six pinned patients (C1–C6) of the companion 3D viewer, the segmentations produced by GT / Baseline / DistMap / CC-Consensus, left sagittal view, tumor regions only (Brain masked for focus). Each figure illustrates one of the six behaviour modes identified in Appendix E.

Note on 3D rendering (two pipelines). The viewer offers a smooth mode and a voxel mode, each served by a distinct pipeline depending on the nature of the mesh.

Voxel mode (raw truth). Greedy voxel meshing: each voxel of the segmentation is turned into a cubic face merged with its coplanar neighbours. No interpolation, no smoothing — exactly what the model predicted at the voxel level. Used as ground-truth reference whenever one needs to count or precisely localise.

Smooth mode (default for Figures 2–7), main meshes. Pipeline `fill_holes + dilation + marching cubes`: the binary mask is first filled (`scipy.ndimage.binary_fill_holes` to remove internal cavities such as ventricles, sulci), dilated by one voxel (`binary_dilation`, 1 iteration) to soften the marching-cubes staircase, then marching-cubed at level 0.5. This is the pipeline used for the tumor-body and Brain meshes shown in Figures 2–7.

Smooth mode, fragments and cavities. For small components (< 4 voxels down to sub-voxel fragments), a separate **signed distance field** pipeline is used: 26-connectivity dilation to bridge voxels touching only by corner/edge, Euclidean inner and outer distance transforms (`scipy.ndimage.distance_transform_edt`) to build the signed distance field, cubic spline upsampling ×2 for sub-voxel resolution, then marching cubes at level iso = −0.3 (empirically calibrated for volume preservation). This pipeline is **necessary for small fragments** because naive marching cubes at 0.5 on a 1-voxel mask

renders 1/6 of the true volume (×6 error) while the signed distance field preserves volume to ±5 % across all sizes.

Shared property of both smooth pipelines. They **preserve topology** (same connected components, same 26-connectivity count as voxel mode); the difference is purely cosmetic. Smooth is the default because the rendering is close to a clinical console; voxel mode remains one click away whenever voxel-exact inspection is needed.

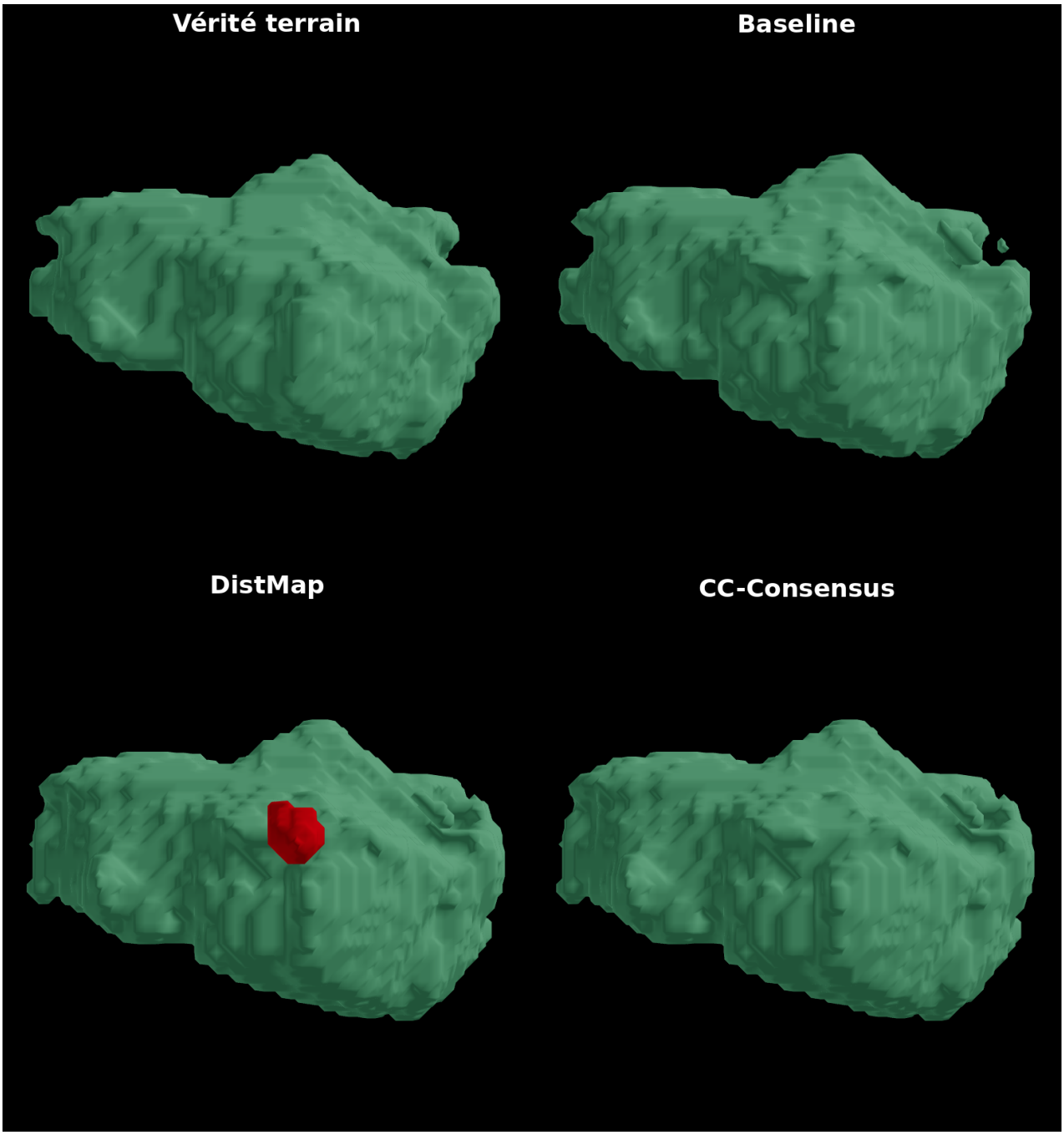

Figure 2: Figure 2 — Case **C1** (patient `BraTS-GLI-00048-001`): Baseline > DistMap. The GT contains only oedema (green); Baseline reproduces this pattern correctly. **DistMap hallucinates an NCR mass** (red) inside the oedema — typical of cases where the SDT pressure induces spurious components. **CC-Consensus removes this hallucination** because the NCR component in DistMap has no overlap with the Baseline prediction (veto), restoring the score almost completely (Dice avg 0.308 → 0.973).

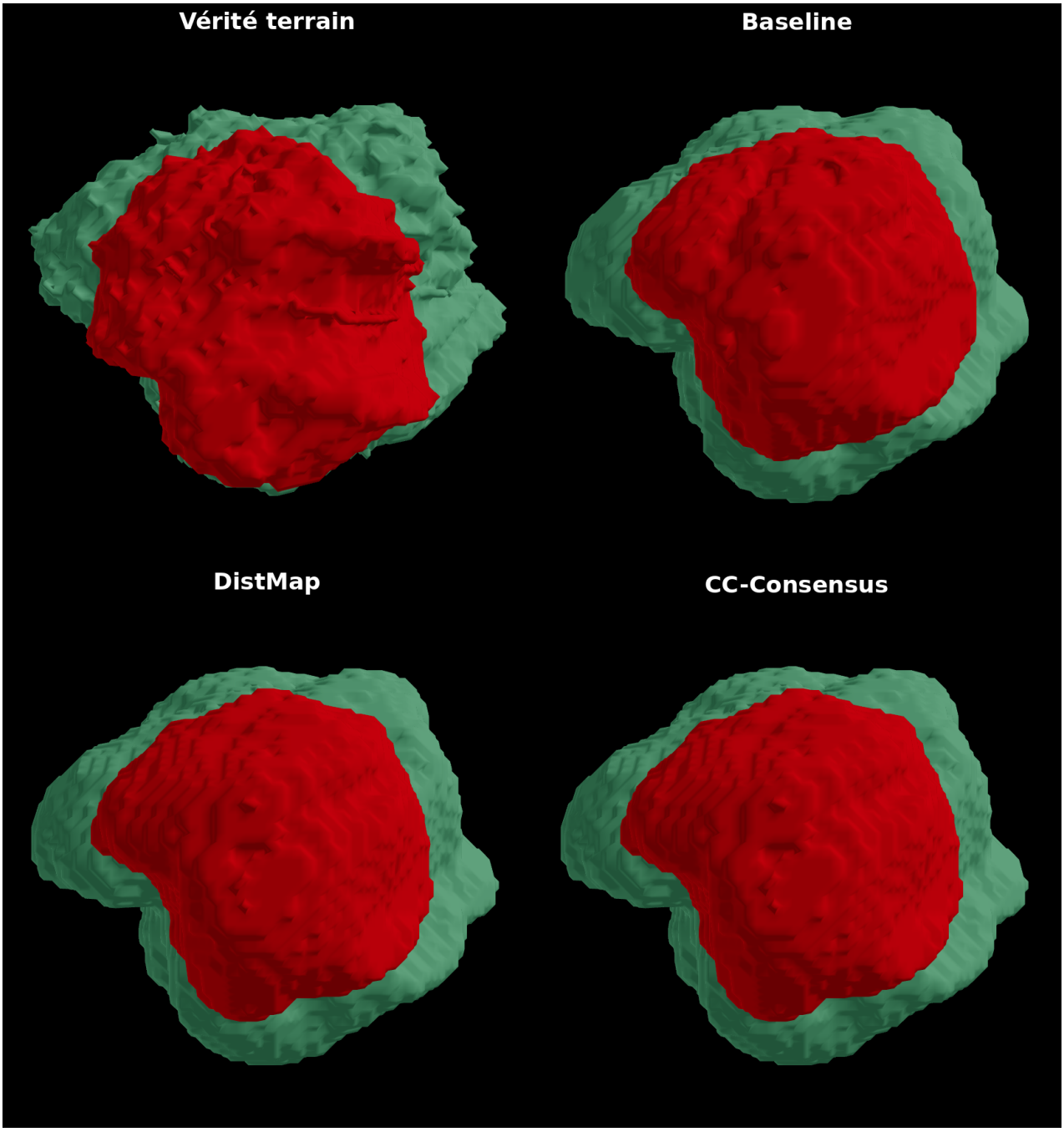

Figure 3: Figure 3 — Case **C2** (patient `BraTS-GLI-01437-000`): DistMap > Baseline. Baseline under-segments the tumor (Dice 0.589) while DistMap captures the tumor extent correctly (Dice 0.923) thanks to its boundary sensitivity. **CC-Consensus matches DistMap** (0.923) because no hallucinated component needs to be removed — the filter preserves the higher-quality prediction when it is confirmed by Baseline.

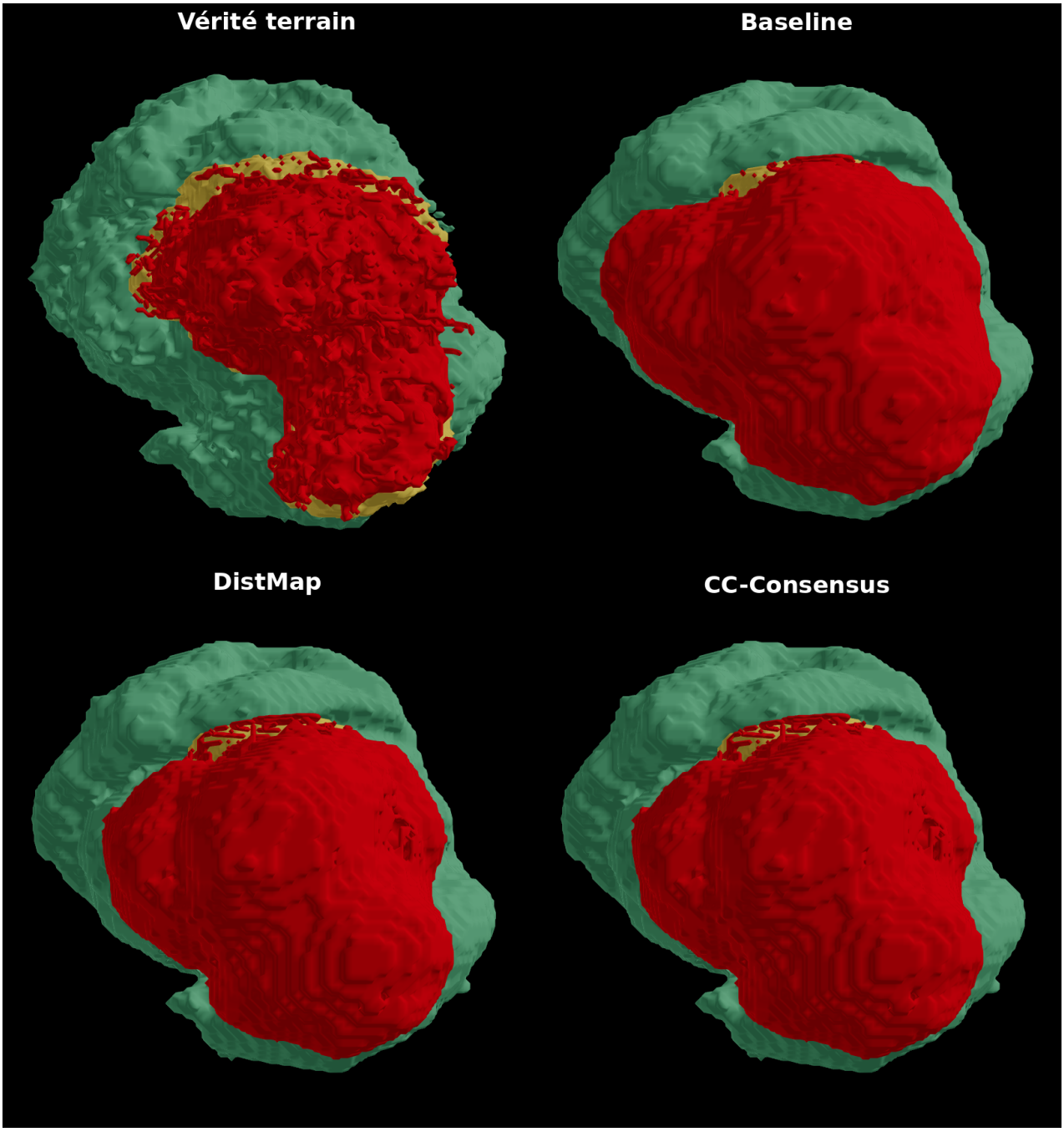

Figure 4: Figure 4 — Case **C3** (patient `BraTS-GLI-01428-000`): B < F < D, filter pulled baseline-side. Baseline (0.618) and DistMap (0.656) bracket the CC-Consensus result (0.645). The filter removes some legitimate DistMap components that Baseline does not predict, mildly degrading the score toward Baseline. This is the most common damage mode (390 / 1196 patients, 32.6 %).

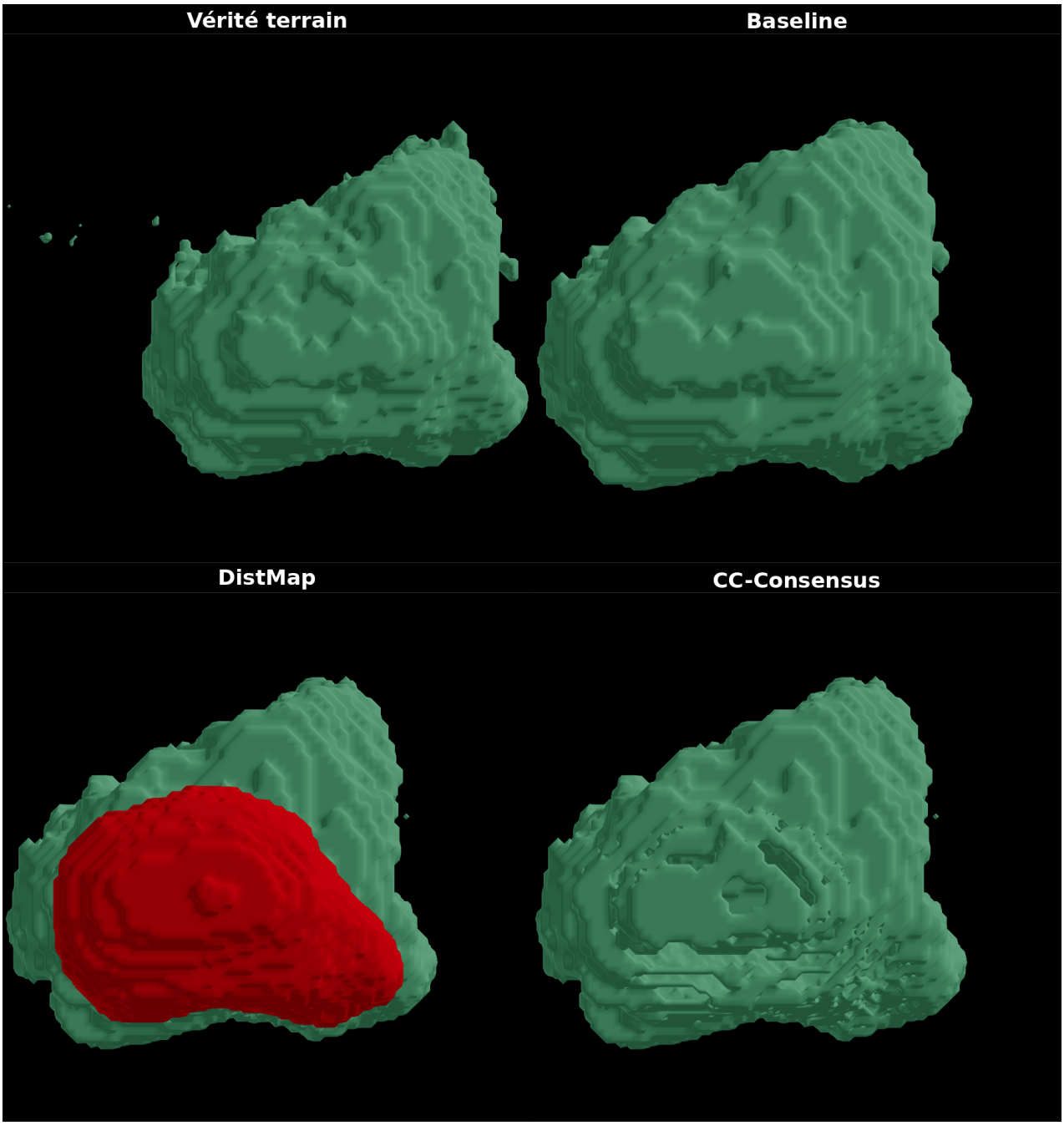

Figure 5: Figure 5 — Case **C4** (patient `BraTS-GLI-00017-001`): D < F < B, partial rescue. Baseline is excellent (0.991); DistMap is half-hallucinated (0.657). CC-Consensus deletes the spurious DistMap components and recovers part of Baseline's quality (0.890), but cannot reach Baseline because it starts from DistMap's voxels.

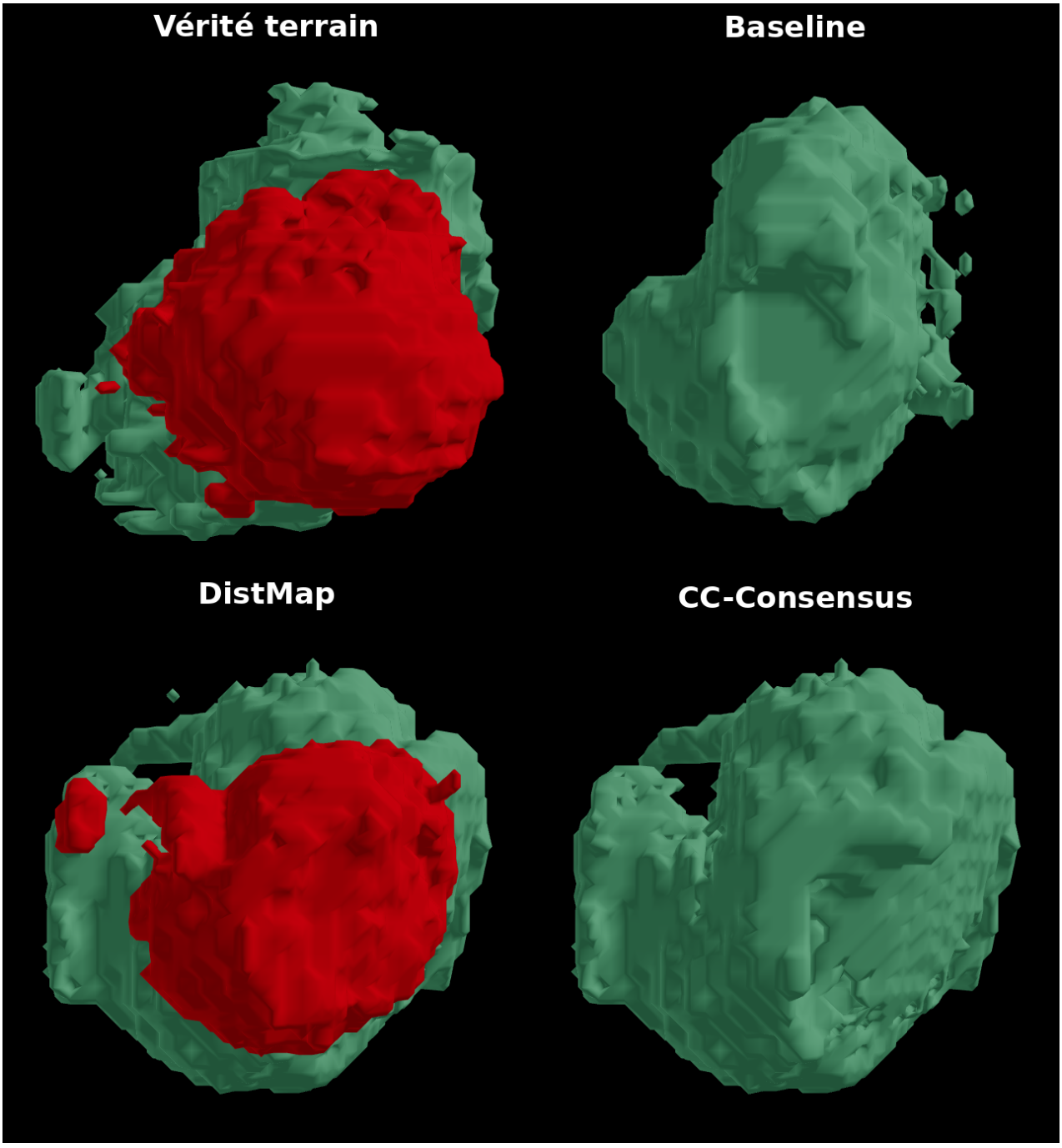

Figure 6: Figure 6 — Case **C5** (patient `BraTS-GLI-01530-000`): F < min(B, D), the filter breaks. Baseline = 0.241, DistMap = 0.541, CC-Consensus = 0.169. The filter **deletes a legitimate large DistMap component** because Baseline failed to find the tumor and cannot confirm it. 463 / 1196 patients (38.7 %) — the dominant failure mode of the filter, when Baseline and DistMap fail differently.

5.3 CC-consensus improves NCR HD95 at no Dice cost

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

Strategy	Dice avg	Δ vs default CC-consensus
Baseline only	0.9078	−0.00115
DistMap only	0.9088	−0.00020
CC-Consensus (default rule)	0.9090	0 (ref)
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 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. The Dice benefit of the filter is therefore concentrated on TC; for WT and ET, Baseline-only or DistMap-only choices already dominate.

Boundary quality (HD95). Complementing the Dice analysis, 95th-percentile Hausdorff distances on the nested BraTS regions (WT/TC/ET) **and** on the individual classes (NCR, ED) where fragments live (n varies per row depending on finite-HD95 patients for that class):

Region / class	Composition	Baseline	DistMap	CC-Consensus	Δ CC-Cons vs DistMap
WT	{1, 2, 3}	3.91 mm	3.86 mm	**3.76 mm**	**−0.10 mm, $p = 2.7 \times 10^{-4}$**
TC	{1, 3}	3.08 mm	2.79 mm	2.88 mm	+0.09 mm, n.s.
ET	{3}	2.62 mm	2.59 mm	2.70 mm	+0.11 mm, n.s.
NCR	{1}	4.89 mm	4.86 mm	**4.48 mm**	**−0.38 mm, $p = 5.7 \times 10^{-14}$**
ED	{2}	4.25 mm	4.33 mm	4.21 mm	−0.12 mm, n.s. ($p = 0.82$)

Paired signed-rank Wilcoxon, one-sided hypothesis HD95(CC-Consensus) < HD95(DistMap). n = 1160 for WT/TC/ET (restricted to patients with finite HD95 on all 3 nested regions), 1153 for NCR, 1193 for ED.

The dominant signal is on NCR: CC-Consensus reduces NCR HD95 by 0.38 mm ($p = 5.7 \times 10^{-14}$) — direct quantitative confirmation that fragment deletion improves boundary quality on the class where they proliferate (NCR: ×1.5 more DistMap fragments than Baseline, cf. §5.2). The WT signal (−0.10 mm, $p = 2.7 \times 10^{-4}$) is the echo of this: NCR ⊂ WT, so NCR fragments contribute to WT boundary error. On ED, the fragment reduction (−61 %) does not translate into a statistically significant HD95 gain — oedema has intrinsic boundary variability that dominates the outliers introduced by fragments. On TC and ET (class 3), HD95 are preserved.

CC-consensus thus delivers a quantitatively measurable gain on NCR HD95 and WT HD95, where Dice remains insensitive. Clinically, NCR is precisely the region where spurious fragments may mislead a radiotherapist on the extent of tumor necrosis.

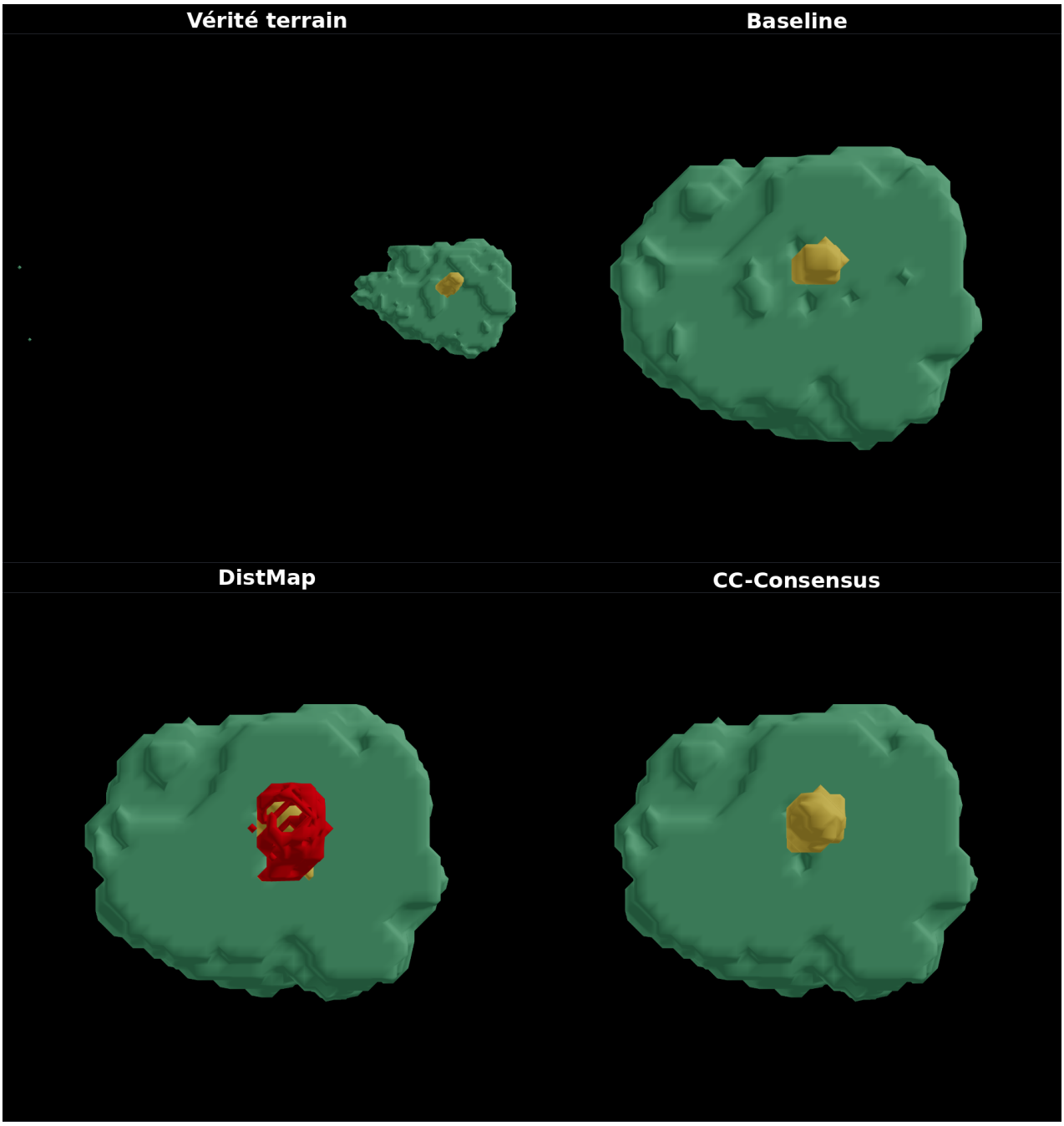

Figure 7: Figure 7 — Case **C6** (patient `BraTS-GLI-00540-000`): clean synergy. Both Baseline (0.785) and DistMap (0.795) are competent but neither is perfect. **CC-Consensus combines their strengths** to reach 0.869 — strictly above both parents. This is the target behaviour on 157/1196 patients (13.1 %) where the filter improves beyond its sources.

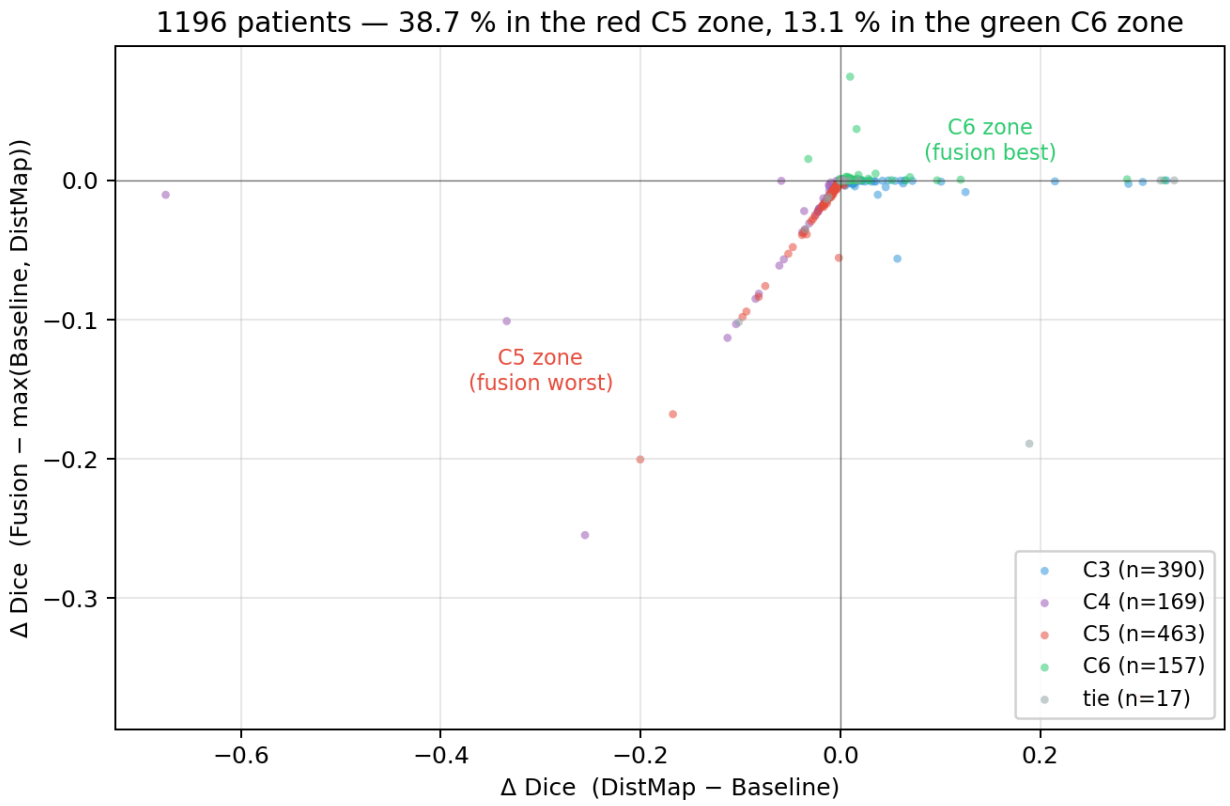

Figure 8: Figure 8 — 1196 validation patients plotted in the model-disagreement plane: x = Dice(DistMap) − Dice(Baseline) (positive means DistMap wins at the patient level), y = Dice(CC-Cons.) − max(Dice(B), Dice(D)) (negative means the CC-consensus filter is worse than either model alone). The *red* 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 %. This visual asymmetry is the central empirical observation of the paper.

5.4 The hard-label ceiling is saturated

The gap between default CC-consensus and the per-class oracle (+0.005 Dice avg) upper-bounds the gain of any patient- or region-level selection policy from the three predictions {B, D, F}. We evaluate three families of policies in 5-fold CV (size-adaptive threshold over τ {20, 50, 100, 200, 500, ∞} voxels; meta-classifiers RF/LR/GBM × patient/region on 31 features; one-feature rule via exhaustive search); **none robustly beats the default CC-consensus**. The one-feature rule, attractive on all-data fit (+0.00119), collapses in 5-fold CV (−0.00096): the best feature and threshold change across folds (TC: 4 distinct features over 5 folds; ET: 4 distinct features). A per-region RandomForest reaches 50 %, 43 %, 51 % argmax accuracy (vs 33 % random), confirming the presence of signal — but when the classifier is wrong, it picks a strictly worse model, yielding a net negative outcome.

The full detail — table of the 7 evaluated policies, RF importances per region, classification of the adaptive sweep, per-fold partition of the one-feature rule — is in **Appendix B**. The hard-label ceiling is essentially reached; closing the gap to the oracle requires voxel-level probabilistic voting or architectural diversity (§6.3).

5.5 Position relative to BraTS 2023 GLI winners

CC-consensus reaches Dice avg = 0.909 (WT 0.935, TC 0.919, ET 0.873) on 1196-patient 5-fold CV with a single-model setup (no multi-fold ensemble, no TTA, single architecture). This is within one percentage point of the published BraTS 2023 GLI winner range on private test set (0.87–0.89 Dice avg; Ferreira *et al.* 2024). Two caveats apply to the direct comparison: (i) different evaluation set (5-fold CV on train + val pool vs private test set, typical 1–2 pp gap to the

disadvantage of the test set); (ii) the Dice = 1 on empty-region convention (nnU-Net / MONAI) inflates ET by ~0.003 relative to the lesion-wise convention used by the challenge (32/1196 patients with empty ET in GT).

We do not claim a new state-of-the-art; the CC-consensus filter is **orthogonal to ensembling** — fragment reduction is a gain that stacks with classical multi-fold / TTA tricks without duplicating them.

---

6. Discussion

6.1 Why DistMap creates fragments

Plausible mechanism — not demonstrated: the SDT pressure sensitises the network to small *boundary-like* signals in transition tissues (oedema–white matter interfaces, post-surgical cavities, heterogeneous NCR), producing high-SDT-response voxels that occasionally survive the argmax as isolated blobs. This hypothesis is consistent with two observations: the increase in fragment count is concentrated on NCR and ED (regions with the longest and most irregular boundaries) and is much smaller on ET, whose gadolinium enhancement provides sharper boundary contrast. Three direct controls (λ ablation × fragment count, SDT response-map visualisation at fragment locations, distance-bin vs MSE) are described in **Appendix C** and deferred 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.** The rule starts from DistMap (superior boundary quality) and uses 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).

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 a 3-way classification "B vs D vs F" to be reliably learned from shape features alone. Both models live in the same decision neighbourhood; their disagreements are dominated by high-frequency spatial noise that global morphology does not capture.

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

- **Voxel-level probabilistic voting.** Exporting softmax outputs (not just argmax labels) and fusing at the voxel level breaks the hard-vote ceiling. A 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, Swin-UNETR) 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.** Only hard-label filtering is reported 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.** The 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 tricks would likely push the score into, or above, the BraTS 2023 GLI winner range, but this would be a parallel-compute contribution orthogonal to the fragment-characterisation question this paper addresses.
- **Dice convention slightly inflates ET.** Empty-GT ET patients (2.7 % of BraTS 2023 GLI, non-enhancing cases, 32/1196 verified) are scored Dice = 1.0 under the nnU-Net / MONAI convention, which slightly inflates the ET regional mean (−0.003 only under the lesion-wise convention). 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).
- **No small tumors in the dataset.** The minimum WT volume on BraTS 2023 GLI is 2808 voxels, median ~89 500 voxels. The topological fragment definition adopted in §4.2 (CC − 1 per class, no size threshold) is **intrinsically size-robust** and requires no recalibration for smaller tumors. However, **the pipeline evaluated here has not been tested on the clinically critical regime of small tumors** (a few hundred voxels), where early detection has major prognostic impact. Absolute morphology features (`vol_*`, `nb_cc_*`) would be out-of-distribution in that regime and would need re-examination before clinical use; the topological and relative features (`ratio_ET_WT`, `frac_small_cc_*`, sphericity, elongation) are size-robust by construction.

---

7. Conclusion

On MedNeXt-B / nnU-Net v2, the auxiliary SDT loss yields no significant Dice gain at convergence (Δ Dice avg = +0.09 pp, Wilcoxon $p > 0.25$ per region, 5-fold CV 1196 patients) but changes the topology of the predictions by introducing a fragment bias that the Dice metric fails to report. A parameter-free connected-component consensus filter that vetoes DistMap CCs without Baseline overlap removes 66 % of NCR fragments on 1196 patients ($p < 10^{-189}$) at no Dice cost, and **significantly improves NCR HD95** (4.86 → 4.48 mm, $p = 5.7 \times 10^{-14}$) as well as WT HD95 (3.86 → 3.76 mm, $p = 2.7 \times 10^{-4}$) — a boundary-quality gain hidden by Dice, clinically relevant on tumor necrosis.

On 1196 patients in 5-fold CV, this rule is shown to be already near the saturation ceiling of any post-hoc hard-label selection policy: the per-class oracle is +0.005 Dice avg above the default, and no 31-feature meta-selector (4 classifier families) robustly beats this default in CV. Closing this gap motivates **training-time fragment-aware losses** (Paper 2) rather than further post-hoc engineering.

---

8. Perspectives

Training-time fragment-aware loss (Paper 2). The hypothesis in §6.1 suggests fragments are a gradient effect. A training-time penalty term counting predicted connected components on the argmax of each mini-batch — and penalising small isolated blobs — should push the network not to instantiate them, making the post-hoc CC-consensus filter unnecessary. This is the direction of Paper 2.

Dataset extensions. BraTS-MET (metastases, multi-lesion pattern) is the most informative next test: DistMap fragments should be more severe there, and the CC-consensus filter should benefit more. BraTS-PED (paediatric) would test generalisation across demographic shifts.

Acknowledgments

The author thanks **Stanislas Larnier** for methodological guidance, feedback on the framing of research questions, and careful reviews of successive drafts of this paper.

Appendix A — Calibration of λ (auxiliary SDT loss)

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.

A static ablation over λ ∈ {0, 0.1, 0.5, 1, 2, 5, 6, 7, 8, 9, 10} (100 epochs, fold 0, seed 42) produces Dice-avg scores all within a 0.5 pp window:

λ	Dice avg	Δ vs Baseline
0 (Baseline)	0.9064	0
0.1	0.9077	+0.0013
0.5	0.9070	+0.0006
1.0	0.9067	+0.0003
2.0	0.9060	−0.0004
5.0	0.9105	+0.0041
9.0	0.9104	+0.0040

On this single fold and without per-patient significance testing, no λ clearly outperforms the baseline. This is consistent with the non-significance of the DistMap gain observed in 5-fold CV on 1196 patients (§5.1). Default training reported in the body uses λ = 1 (close to published heuristics and to the gradient-balanced calibration ÷ 5).

A dynamic weighting scheme — DWA (Dynamic Weight Average, Liu CVPR 2019) — that tracks the relative learning rates of the Dice+CE and SDT heads over training is a natural next direction: if a regime exists where SDT genuinely contributes without saturating, a static sweep cannot find it.

Appendix B — Detailed hard-label ceiling study

The gap between default CC-consensus and the per-class oracle (+0.005 Dice avg) is the maximum 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, attractive on all-data fit (+0.00119), collapses in 5-fold CV (−0.00096): the best feature and threshold change across folds (TC: 4 distinct features over 5 folds; ET: 4 distinct features). Four different "best" features over five folds for TC alone clearly indicate that the signal is not robust enough to trust.

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

Feature importance (full-data RF, top-3 per region) 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 survive CV.

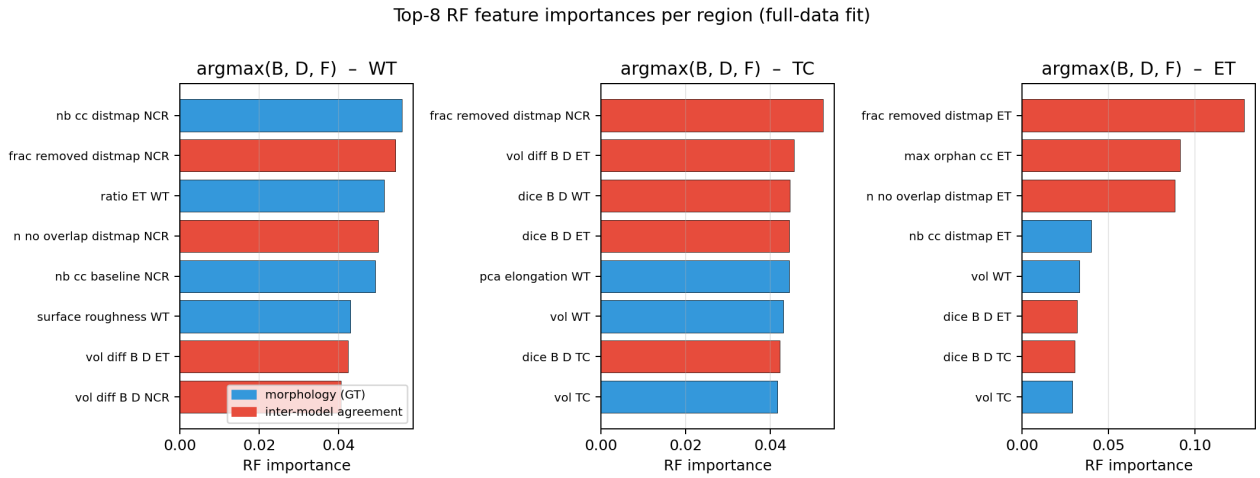

Figure 9: Figure A1 — Top-8 feature importances of a RandomForest classifier trained to predict argmax(Baseline, DistMap, CC-Consensus) for each region (WT / TC / ET). Blue bars: morphology features from GT (20). Red bars: inter-model agreement features (11). For ET specifically, the 3 top importances — `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 over-predicts relative to Baseline.

Appendix C — Mechanism hypothesis: deferred controls

The hypothesis in §6.1 (the SDT pressure produces high-response voxels at ambiguous interfaces, which occasionally survive the argmax) remains at this stage a **working hypothesis, not demonstrated**. The following three direct controls are all feasible on the existing checkpoints and are deferred to future work:

1. **λ ablation crossed with fragment count.** Verify that the mean number of fragments per patient grows monotonically with λ. Monotonic growth would confirm the causal link between SDT pressure and artefact; absence of monotonicity would suggest that optimisation noise dominates.
2. **SDT response-map visualisation at fragment locations.** For a sample of patients, overlay the tanh output of the auxiliary head and the fragment map; fragments should coincide with high-SDT-response voxels close to a tissue interface.
3. **Distance bins vs MSE.** Replace the `Conv3D(32 → 3) + tanh + MSE` head with a distance-bin classification head (e.g. 16 equi-probable bins in $[-1, 1]$). If the artefact disappears or substantially decreases, it is specific to the MSE-SDT formulation and not to distance supervision in general.

Running these three controls would move §6.1 from "working hypothesis" to "demonstrated mechanism".

Appendix D — Runtime and reproducibility

All code, the 20 + 11 pre-extracted features, per-patient model scores, oracle / case-classification CSVs, threshold-sweep results and meta-selector outputs are available in the companion repository. Per-patient extraction of the 31 features on the 1196 predictions runs in **~10 min** on 14 P-core threads (`taskset -c 0-13`) of an i7-14700K; the full meta-classifier sweep (4 families × 5 folds × 31-dim input) in ~2 min on the same host. **Training time per fold: ~13 h 30 min for 300 epochs** on a single RTX PRO 6000 Blackwell (96 GB), Baseline and DistMap variants at equivalent duration (the auxiliary SDT regression head adds < 1 % GPU overhead at 300 ep).

Appendix E — The six demonstration patients

Six patients are highlighted to span the six model-ordering cases, used both for the figures and as pinned anchors in the companion 3D viewer. In the table, F denotes the CC-consensus filter output. Patient identifiers are shown without the `BraTS-GLI-` prefix for compactness (the dataset prefixes them systematically).

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

References

- Isensee F., Jaeger P. F., Kohl S. A. A., Petersen J., Maier-Hein K. H. (2021). *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation*. **Nature Methods** 18, 203–211. DOI: 10.1038/s41592-020-01008-z.
- Roy S., Koehler G., Ulrich C., Baumgartner M., Petersen J., Isensee F., Jaeger P. F., Maier-Hein K. H. (2023). *MedNeXt: transformer-driven scaling of ConvNets for medical image segmentation*. **MICCAI 2023**, LNCS 14222, 405–415. DOI: 10.1007/978-3-031-43901-8_39.
- Ma J. (2020). *Distance transform maps improve semantic segmentation of medical images*. **Medical Imaging with Deep Learning (MIDL) 2020**, short paper track.
- Xue Y., Tang H., Qiao Z., Gong G., Yin Y., Qian Z., Huang C., Fan W., Huang X. (2020). *Shape-aware organ segmentation by predicting signed distance maps*. **AAAI 2020**, 34(07), 12565–12572. DOI: 10.1609/aaai.v34i07.6946.
- Karimi D., Salcudean S. E. (2020). *Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks*. **IEEE Transactions on Medical Imaging** 39(2), 499–513. DOI: 10.1109/TMI.2019.2930068. arXiv:1904.10030.
- Huang Q., Yang J., Zhang B., Wang Z., Bai J., Li Y. (2021). *A deep multi-task learning framework for brain tumor segmentation*. **Frontiers in Oncology** 11, 690244. DOI: 10.3389/fonc.2021.690244.
- Pham T.-D., Abdollahzadeh A., Tohka J. (2024). *SiNGR: Brain tumor segmentation via signed normalized geodesic transform regression*. **MICCAI 2024**. arXiv:2405.16813.
- Ferreira A., Solak Ü. M., Li J., Dammann P., Kleesiek J., Alves V., Egger J. (2024). *How we won BraTS 2023 adult glioma challenge? Just faking it! Enhanced synthetic data augmentation and model ensemble for brain tumour segmentation*. **arXiv:2402.17317**.
- Liu S., Johns E., Davison A. J. (2019). *End-to-end multi-task learning with attention* (DWA — Dynamic Weight Average). **CVPR 2019**, 1871–1880. DOI: 10.1109/CVPR.2019.00197.
- Baid U., Ghodasara S., Mohan S., Bilello M., Calabrese E., Colak E., *et al.* (2021). *The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification*. **arXiv:2107.02314**.
- Menze B. H., Jakab A., Bauer S., *et al.* (2015). *The multimodal brain tumor image segmentation benchmark (BRATS)*. **IEEE TMI** 34(10), 1993–2024. DOI: 10.1109/TMI.2014.2377694.