Pre- and Post-Treatment Brain Metastases Segmentation Using nnU-Net with Post-Processing for BraTS 2026

Haobin Liu[1] and Xin Wang[1,2(✉)]

[1] College of Software, Jilin University, Changchun, China
[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China
w_x@jlu.edu.cn

Abstract. Brain metastases exhibit high inter-lesion variability in size, enhancement pattern, and post-treatment appearance, making volumetric segmentation of both pre- and post-treatment cases the central challenge of the BraTS 2026 Task 1 (Brain Metastases). We build a pragmatic pipeline on a 5-fold nnU-Net ResEnc-L ensemble, in which each fold is trained independently for 1,000 epochs with the standard Dice + cross-entropy loss on 1,296 four-modality training cases. This ensemble is followed by a rule-based post-processing cascade tuned for the lesion-wise Dice similarity coefficient (LW-DSC), a detection-oriented metric that behaves very differently from the traditional global Dice. The final pipeline reaches an LW-DSC of **0.733** / **0.751** / **0.713** / **0.549** on the enhancing tumour (ET), tumour core (TC), whole tumour (WT), and resection cavity (RC) sub-regions on the official validation leaderboard. Rather than trusting these leaderboard gains, we audit every post-processing stage with a five-fold out-of-fold (OOF) analysis with no model-training leakage over all 1,296 training cases, scored with the official BraTS evaluation code (`BraTS_evaluation`): it confirms two stages as robust, per-fold-consistent improvements while the third improves only the leaderboard and does not reproduce out-of-fold. We further provide a mechanistic analysis of the LW-DSC metric that explains why recall-recovering post-processing carries low risk whereas component deletion does not, and we report thirteen negative results spanning loss engineering, alternative backbones, and inference-time settings, several of which run counter to widely held intuitions. Source code is released under Apache-2.0 at https://github.com/hornbeamliu/brats2026-met.

Keywords: Brain metastases segmentation · nnU-Net · Post-processing · Lesion-wise Dice · BraTS 2026

1 Introduction

Brain metastases are the most common intracranial neoplasms in adults, and accurate volumetric assessment underpins radiosurgery planning and treatment-response monitoring, where manual RANO-BM measurements under-detect small

lesions [14,21,28,1]. Within the long-running BraTS benchmark [18,5,3,4,2,27,11], the metastases sub-challenge (BraTS-MET) [20,17] targets the pre- and post-treatment setting [6] and, in its 2026 edition, poses three difficulties that separate it from prior glioma tasks: (i) cases aggregated from eight institutions, mixing SRI24-registered and native-space scans, produce strong inter-site heterogeneity; (ii) a resection-cavity (RC) label that is extremely rare, present in only about one in eight training cases, creates severe class imbalance; and (iii) the ranking metric is a *lesion-wise* Dice (LW-DSC) that averages Dice per connected component, rather than the conventional global-volume Dice, and is therefore dominated by the detection of small ($< 27\,\mathrm{mm}^3$) lesions.

In this short paper we describe our BraTS 2026 Task 1 pipeline (team `bin_JLU`). We keep the architecture and training deliberately close to the strongest publicly documented nnU-Net baseline for BraTS glioma [10,9,24,7] and concentrate our engineering effort on inference-time post-processing designed for the LW-DSC metric. The pipeline couples a 5-fold nnU-Net ResEnc-L ensemble with a rule-based cascade whose two robust stages are a rule-based clean-up (`RuleClean`) and a resection-cavity boundary expansion (`BoundaryExpand`); a third, outside-brain false-positive suppression (`brainF70`), was part of the submitted Docker container but does not survive our out-of-fold audit and is reported as a negative result (Sections 3.2 and 4). Each stage is a separate script acting on the ensembled volume, so its LW-DSC contribution can be measured in isolation. On the official Synapse validation leaderboard the pipeline places all three tumour-side sub-regions (ET, TC, WT) in the leading tier (Section 3); this is an evaluation on the *validation* set, not a hidden-test or final-ranking result.

Rather than relying on the leaderboard alone, we audit each stage with a five-fold out-of-fold analysis over all 1,296 training cases with no model-training leakage (Section 3.2): this confirms `RuleClean` and `BoundaryExpand` as robust across folds, whereas the `brainF70` gain appears only on the leaderboard and falls within out-of-fold variance. Finally, we document thirteen unsuccessful attempts (Section 4) across loss engineering, alternative backbones, and inference-time settings; several run counter to common intuition: notably, finer sliding-window overlap and imbalance-oriented losses can each degrade performance, so future participants need not revisit them.

2 Methods

2.1 Dataset

We use the BraTS-MET 2025 training dataset [17] as released for the 2026 challenge: pre- and post-treatment multi-parametric MRI of brain metastases aggregated from eight institutions, mixing SRI24 [23]-registered and native-space acquisitions. Each case provides four modalities — post-contrast T1-weighted (T1c), native T1-weighted (T1n), T2-weighted FLAIR (T2-FLAIR), and T2-weighted (T2w) — and a 5-class annotation: background (0), non-enhancing tumour core (NETC, 1), surrounding non-enhancing FLAIR hyperintensity (SNFH,

2), enhancing tumour (ET, 3), and resection cavity (RC, 4). The evaluation regions are TC = NETC ∪ ET and WT = NETC ∪ SNFH ∪ ET.

The 2026 release provides all four modalities for every one of the **1,296** training cases, encoded in the nnU-Net [9] four-channel format (`_0000`–`_0003` = T1c, T1n, T2-FLAIR, T2w). Every training case therefore carries a native T2w volume, and we perform no T2w synthesis or zero-imputation.

Validation is done on the official 179-case Synapse validation set, in which every case likewise provides all four modalities including a native T2w. Both the training and the official validation data used in this work are therefore complete-modality; the inference-time fallback described in Section 2.5 is a safeguard for the hidden test phase and is never exercised on the data reported here. All local out-of-fold (OOF) ablations reported in Section 3.2 are computed on the 1,296 training cases using each case's held-out fold's prediction, and are therefore independent of the Synapse validation submission budget.

2.2 Backbone architecture

We adopt the residual encoder variant of nnU-Net (ResEnc-L) introduced in the "nnU-Net Revisited" study [10]. Its residual blocks follow the design of ResNet [8], and its feature maps are normalised with instance normalisation [25] as is standard in nnU-Net. We use the official `nnUNetPlannerResEncL` planner and the resulting `nnUNetResEncUNetLPlans` configuration. All network hyperparameters, including patch size, batch size, number of pooling stages and per-stage channel widths, are set automatically by the planner from the dataset fingerprint and are not manually tuned. We use the `3d_fullres` configuration exclusively; no cascade or low-resolution stage is trained. The final architecture consists of a residual encoder-decoder with deep supervision at all resolutions, four input channels, and five output channels corresponding to background plus the four foreground classes.

2.3 Training

Each of the five folds is trained *independently* with the standard nnU-Net trainer `nnUNetTrainer_ResEncL` for a full 1,000 epochs. We use SGD with Nesterov momentum $\mu = 0.99$ and weight decay 3×10^{-5}, and the standard nnU-Net polynomial learning-rate schedule with initial learning rate $\eta_0 = 0.01$ and power 0.9; adaptive optimisers such as Adam [12] and AdamW [15] are known to underperform SGD on nnU-Net's compound loss and were not considered here. The optimisation objective is the standard nnU-Net compound loss $\mathcal{L} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}}$, where the Dice term follows the formulation popularised by V-Net [19], with deep supervision at every decoder scale, without any loss re-weighting or focal / top-k modification (see Section 4 for negative results on these alternatives). Data augmentation follows the nnU-Net default recipe (random mirroring, elastic deformation, rotation, scaling, additive Gaussian noise, Gaussian blur, multiplicative brightness / contrast, low-resolution simulation, and gamma correction). Training is performed with automatic mixed precision (fp16). We train each fold on

2×NVIDIA RTX A6000 (48 GB VRAM) using the standard batch = 2 split that the nnU-Net planner selects for `3d_fullres`.

2.4 Post-processing cascade

The cascade is executed as three independent scripts run in sequence on the 5-fold ensemble prediction. We deliberately keep them as separate stages rather than fuse into a single entry point, so that each stage's effect on the LW-DSC can be quantified separately (Section 3.2).

Stage 1: rule-based clean-up (`RuleClean`). The clean-up stage applies four rules to the argmaxed segmentation volume, in the order and with the exact semantics specified in Algorithm 1. All four operate on 26-connected components rather than isolated voxels, and every removed voxel is relabelled to background. `ET-min10` and `RC-min10` remove connected components of ET and RC whose physical volume is below 10 mm^3, respectively; the 10 mm^3 threshold was selected because it is well below the challenge's 27 mm^3 small-lesion evaluation cut-off yet still removes the long tail of single-voxel spurious predictions. `SNFH-adj2mm` removes an entire SNFH connected component unless at least one of its voxels lies within 2 mm of a TC voxel; because retention is all-or-nothing per component, the interior of a retained edema component is never eroded. This mirrors the clinical observation that peritumoural FLAIR hyperintensity should be attached to the tumour body. `NETC-adj2mm` applies the same whole-component rule to NETC with ET as the anchor, enforcing the BraTS-MET annotation guideline that NETC "must be enclosed by enhancing tumour"; the 2 mm tolerance absorbs slight boundary imprecision at the ET/NETC interface. Applied in isolation to the raw 5-fold ensemble, `RuleClean` reaches an LW-DSC average across ET / TC / WT / RC of **0.6823** on the Synapse validation set, a +0.0305 gain over the un-post-processed ensemble baseline.

Stage 2: resection-cavity boundary expansion (`BoundaryExpand`). This stage addresses a specific, systematic failure of the trained ensemble: it *under-segments* the resection cavity, predicting RC components whose contours are truncated inside the true cavity (raw RC lesion-wise DSC is only 0.367). LW-DSC on the RC channel is particularly sensitive to this boundary shrinkage, because a case with a small predicted RC component that is a subset of a larger ground-truth RC is penalised much more strongly under lesion-wise matching than under global Dice. We therefore grow each predicted RC component outward by a fixed margin in that case's own native voxel grid, the same space in which the network makes its prediction, using a background-restricted 3D ring dilation. The operation is thus motivated by the observed under-segmentation rather than by any target physical distance, and it is region-protected (never entering ET/NETC/SNFH). Let R_0 denote the set of voxels labelled RC after Stage 1, and let P_{RC} denote the ensembled RC softmax probability. Starting from

Algorithm 1 `RuleClean` (Stage 1). All operations act on 26-connected components; every removed voxel is relabelled to background. Labels: 1=NETC, 2=SNFH, 3=ET, 4=RC; the tumour core is TC $= \{1, 3, 4\}$.

Require: argmax segmentation S; voxel spacing (s_x, s_y, s_z) in mm

function RemoveSmall(S, ℓ, τ)
  $\tau_{\text{vox}} \leftarrow \max\big(1,\ \text{round}(\tau/(s_x s_y s_z))\big)$
  for all 26-connected components C of $\{v : S[v] = \ell\}$ **do**
    if $|C| < \tau_{\text{vox}}$ **then**
      $S[C] \leftarrow 0$ ▷ delete whole component
    end if
  end for
  return S
end function

function KeepIfTouch(S, ℓ, A, d) ▷ A: anchor mask
  $r \leftarrow \max\big(1,\ \text{round}(d/\min(s_x, s_y, s_z))\big)$
  for all 26-connected components C of $\{v : S[v] = \ell\}$ **do**
    if $\text{dilate}(C, r) \cap A = \varnothing$ **then**
      $S[C] \leftarrow 0$
    end if
  end for
  return S
end function

$S \leftarrow$ RemoveSmall($S,\ 3,\ 10\,\text{mm}^3$) ▷ `ET-min10`
$S \leftarrow$ RemoveSmall($S,\ 4,\ 10\,\text{mm}^3$) ▷ `RC-min10`
$S \leftarrow$ KeepIfTouch($S,\ 2,\ \{v : S[v] \in \text{TC}\},\ 2\,\text{mm}$) ▷ `SNFH-adj2mm`
$S \leftarrow$ KeepIfTouch($S,\ 1,\ \{v : S[v] = 3\},\ 2\,\text{mm}$) ▷ `NETC-adj2mm`
return S

$R^{(0)} = R_0$, we iterate for $i = 1, 2, 3$:

$$R^{(i)} = R^{(i-1)} \cup \Big\{ v \in \partial R^{(i-1)} \;\Big|\; \text{label}(v) = \text{background} \wedge P_{\text{RC}}(v) > 0.10 \Big\}, \quad (1)$$

where $\partial R^{(i-1)}$ is the 26-connected boundary shell of $R^{(i-1)}$. The gate on P_{RC} is a real, active safeguard in the submitted pipeline: a background voxel is admitted into RC only where the ensembled RC probability exceeds 0.10, so the expansion can never grow into regions the network considers confidently non-RC. It is deliberately set as a permissive lower bound rather than a tight threshold. The observation that replacing 0.10 by 0 yields near-identical outputs is the expected behaviour of such a lower bound: voxels immediately adjacent to a predicted RC wall almost always already carry $P_{\text{RC}} > 0.10$, so the gate binds only where the model is *not* confident (precisely the cases it is meant to block) and is otherwise rarely called upon. Together with the region protection (Stage 2 never enters ET/NETC/SNFH) and a per-component cap that limits growth to at most 3× the original component volume, the gate is one of three independent constraints that keep the operator conservative rather than a blind dilation.

More aggressive settings ($i = 5$ dilations with no gate) are catastrophic in our setting; see Section 4. In the five-fold out-of-fold audit of Section 3.2, this stage produces a small but sign-consistent RC gain on every fold and leaves the ET/TC/WT channels untouched, confirming it as a safe, if modest, operator.

Stage 3: outside-brain false-positive suppression (`brainF70`). A third stage targets RC components placed partly outside the brain. Exploiting the skull-stripped T1c (intensity 0 outside the parenchyma), it defines a brain mask $B = \{v : \mathrm{T1c}(v) > 0\}$ and deletes any RC component whose brain fraction $|C \cap B|/|C|$ is below 0.70; no brain-extraction model is needed. It was part of the submitted Docker container but does not survive the five-fold out-of-fold audit (Section 3.2). We therefore exclude it from the recommended cascade, which is `RuleClean` followed by `BoundaryExpand`, and report it as a negative result (Section 4).

2.5 Inference and ensembling

Inference uses the standard nnU-Net sliding-window predictor with Gaussian importance weighting at the default step size of 0.5 (50% overlap); a sweep at 0.25 with three ensemble weightings was systematically worse (Section 4). The five per-fold softmax volumes are fused by arithmetic mean and decoded with `argmax`; weighted averaging gave no improvement. Test-time mirroring along all three axes is enabled, as in the nnU-Net default. The Task 1 environment provides a single 24 GB A10G GPU, so the container loads the five fold models one at a time, accumulating softmax on disk before fusion; this is the configuration behind the Section 3.1 results.

For the hidden test phase the container defines a deterministic missing-modality fallback: a case lacking native T2w reuses its T2-FLAIR for the `_0003` channel (the closest T2-weighted contrast) rather than zero-filling. All cases reported here provide native T2w (Section 2.1), so this fallback is never exercised.

3 Results

We report lesion-wise Dice similarity coefficient (LW-DSC) for each of the four evaluation sub-regions (ET, TC, WT, RC), following the challenge protocol [17,16,22]. The primary summary statistic is the arithmetic mean of the four per-region LW-DSC values (denoted `avg4`). We complement the primary metric with per-region small-instance F_1 where relevant.

3.1 Main validation results

Table 1 summarises the score of our final submission on the official Synapse validation set (179 cases). This submission is the configuration behind our Docker container: the 5-fold ResEnc-L ensemble described in Section 2.5 followed by the

Table 1. Main validation results on the official Synapse validation set (179 cases); "Final pipeline" is the submitted Docker configuration and avg4 is in bold. See the text for the `brainF70` caveat.

Submission	ET	TC	WT	RC	avg4
Final pipeline	0.7328	0.7510	0.7131	0.5487	**0.6864**

post-processing cascade of Section 2.4. It includes the `brainF70` stage, which we retain here only to reproduce the exact submitted leaderboard score; as shown by the out-of-fold audit in Section 3.2, `brainF70` is not a reliable component, and our recommended cascade is `RuleClean` followed by `BoundaryExpand`.

The resection cavity remains the sole critical deficit relative to our own tumour-side numbers; we discuss the gap in Section 5.

3.2 Ablation of the post-processing cascade

Consuming the Synapse validation submission budget for a full ablation of the three post-processing stages would have been prohibitive. We therefore evaluate the cascade on the complete five-fold OOF predictions produced by the nnU-Net trainer: for each fold we score its held-out validation split with that fold's own single model, so every one of the 1,296 training cases is scored exactly once by a model that never saw it. This gives an estimate with no model-training leakage on two orders of magnitude more cases than a single Synapse submission would allow. Scoring uses the official BraTS evaluation code (`BraTS_evaluation`), which wraps the Panoptica lesion-wise evaluator in its `mets` configuration (lesion-volume threshold 27 mm^3, overlap 0.2); regions absent from the ground truth are skipped rather than scored, exactly as on the leaderboard. The resulting OOF absolute values are not directly comparable to the ensemble leaderboard numbers of Section 3.1 (they come from single models and a different case population), so we read Table 2 only for the *sign and per-fold consistency* of each stage's increment, never against the leaderboard scale.

Three findings survive this larger audit. First, `RuleClean` is the dominant and unambiguous contributor: it increases avg4 by +0.0640 and RC by +0.1244, with a positive per-fold avg4 increment on all five folds (range +0.052 to +0.082). Second, `BoundaryExpand` is a small but consistent RC operator: +0.0080 RC (+0.0021 avg4), non-negative on every fold and with no change to the ET/TC/WT channels. Third, `brainF70` does *not* replicate its leaderboard behaviour: overall it changes avg4 by −0.0006 and RC by −0.0020, is triggered on only 6 of the 1,296 cases, and reverses sign across folds — removing a true false positive on fold 0, but deleting a genuine resection cavity on fold 2. For reference, on the leaderboard the `RuleClean`→full increment is +0.0164 RC (0.5323 → 0.5487); our audit attributes the durable part of this to `BoundaryExpand` and the remainder to leaderboard noise from a single `brainF70` deletion. Because this net effect lies within fold-to-fold noise rather than a reliable regression, we exclude `brainF70` from the recommended cascade and report it as a negative result

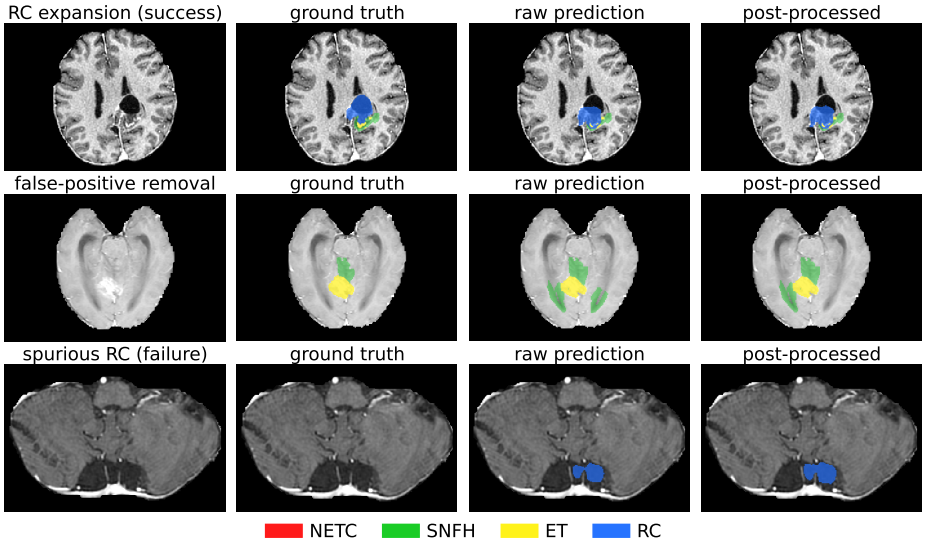

Fig. 1. Raw-versus-post-processed comparison on three fold-0 held-out cases (columns, left to right: T1c input, ground truth, raw ensemble, post-processed with `RuleClean` + `BoundaryExpand`). The first column shows the T1c input, titled with the phenomenon that row illustrates. Top: `BoundaryExpand` grows an under-segmented RC cavity toward the ground-truth extent. Middle: `RuleClean` removes a spurious sub-threshold component. Bottom: a failure case — the raw model predicts a spurious RC on a baseline scan (no GT cavity) that the cascade does not suppress. Overlays: NETC (red), SNFH (green), ET (yellow), RC (blue).

(Section 4); it was present in the submitted Docker container (Table 1) only to reproduce the leaderboard score, and is in any case inert on all but a small number of cases. Beyond lesion-wise DSC, Table 2 also reports the additional official metrics requested for a complete assessment (per-region lesion-wise NSD, HD95, and large- and small-instance F1) for the raw ensemble and each cumulative stage. `RuleClean` improves NSD, HD95, and large-instance F1 in every region and roughly halves HD95, confirming that it sharpens both overlap and boundary fidelity; it does lower small-instance F1 (e.g. ET 0.360 $\rightarrow$ 0.277), the direct and intended consequence of deleting sub-10 mm^3 components, a cost outweighed on the size-dominant lesion-wise and large-instance metrics. `BoundaryExpand` improves RC HD95 (112.3 $\rightarrow$ 109.5 mm) alongside its RC DSC gain, so the overlap improvement does not come at the expense of surface fidelity.

Figure 1 shows the raw versus post-processed output on three fold-0 cases (a success, a false-positive removal, and a failure).

Table 2. Post-processing ablation on the full five-fold OOF (1,296 cases; official BraTS-METS/Panoptica evaluator, per-region means over ground-truth-present cases). Stages are cumulative: `RuleClean`, then `BoundaryExpand` (*recommended* cascade), then `brainF70` (*submitted*; a negative result, Section 4). `BoundaryExpand` and `brainF70` modify only RC, so ET/TC/WT are unchanged after `RuleClean`. avg4 is the mean over the four regions; ↑/↓: higher/lower is better.

Metric	Configuration	ET	TC	WT	RC	avg4
DSC ↑	Raw	0.6163	0.6395	0.5968	0.3667	0.5548
	+ `RuleClean`	0.6643	0.6799	0.6400	0.4911	0.6188
	+ `BoundaryExpand`	0.6643	0.6799	0.6400	0.4991	0.6209
	+ `brainF70`	0.6643	0.6799	0.6400	0.4971	0.6203
NSD ↑	Raw	0.690	0.700	0.618	0.296	0.576
	+ `RuleClean`	0.739	0.739	0.655	0.381	0.629
	+ `BoundaryExpand`	0.739	0.739	0.655	0.377	0.628
	+ `brainF70`	0.739	0.739	0.655	0.374	0.627
HD95 (mm) ↓	Raw	90.7	86.4	94.7	170.9	110.7
	+ `RuleClean`	72.4	71.8	79.6	112.3	84.0
	+ `BoundaryExpand`	72.4	71.8	79.6	109.5	83.3
	+ `brainF70`	72.4	71.8	79.6	113.2	84.2
Large-inst. F1 ↑	Raw	0.752	0.761	0.765	0.070	0.587
	+ `RuleClean`	0.792	0.794	0.799	0.087	0.618
	+ `BoundaryExpand`	0.792	0.794	0.799	0.089	0.619
	+ `brainF70`	0.792	0.794	0.799	0.088	0.618
Small-inst. F1 ↑	Raw	0.360	0.368	0.302	0.021	0.263
	+ `RuleClean`	0.277	0.279	0.223	0.017	0.199
	+ `BoundaryExpand`	0.277	0.279	0.223	0.017	0.199
	+ `brainF70`	0.277	0.279	0.223	0.017	0.199

4 Failed attempts

On the training side, a Dice+top-k CE loss ($k = 10\%$) improved ET and TC by under a point each but degraded RC by more than two, and a routing variant feeding only the RC channel from the top-k model reproduced the failure. A MedNeXt-L backbone added as a sixth ensemble member degraded `avg4` under both an equal 1/6 and a 0.7/0.3 ResEnc-L/MedNeXt weighting. Two Focal-Tversky variants also failed: $\gamma = 0.75$ collapsed ET and RC Dice to zero within a hundred epochs, and a milder γ ran to epoch 613 without exceeding the DC+CE baseline. We were thus unable to compensate RC's extreme rarity through the loss alone at this dataset scale.

On the post-processing side, two aggressive volume filters (removing RC components below 100 or 300 voxels) degraded RC by up to seven points, and a five-iteration ungated ring dilation produced the study's largest RC regression by extending into skull and CSF; these bracket the conservative, gated $i = 3$ `BoundaryExpand`. `brainF70` (Stage 3, deleting RC components with brain frac-

tion < 0.70) helped one leaderboard case but was net-negative and sign-reversing across the five OOF folds (Section 3.2), so it is excluded from the recommended cascade. Lowering the RC softmax threshold to 0.40 traded precision for recall at a net loss; an SNFH $< 10\,\mathrm{mm}^3$ filter was neutral and dropped; four region-swap fusion variants gave nothing once an axial I/O bug was fixed; and a 0.25 sliding-window sweep (75% overlap, three weightings) was consistently worse than 0.5 (`avg4` 0.6795–0.6796 vs 0.6823), which we attribute to over-averaging at patch boundaries.

5 Discussion

Post-processing carries more weight under LW-DSC than under global Dice, but only in one direction. Prior BraTS write-ups report that post-processing beyond light small-blob filtering is nearly neutral [13,26]; under voxel-wise (global-volume) Dice this is expected, as a single small component is negligible as a voxel fraction. Lesion-wise DSC changes this: it matches components case-by-case and averages Dice *per lesion*, so that same component becomes an entire term in the mean, creating a sharp asymmetry. For a 200-voxel ground-truth RC lesion, growing a partial prediction from 80 to 160 voxels lifts that lesion's Dice from 0.57 to 0.89, whereas 40 spurious voxels on an otherwise perfect match cost only $1.0 \rightarrow 0.91$ — yet deleting the component outright scores it 0. Recall-recovering operations therefore sit on the safe side of this asymmetry, while deletion is the single most expensive event.

Our experiments trace out exactly this safe band. Confidence-gated boundary growth (`BoundaryExpand`) recovers RC recall with a sign-consistent gain at no cost to the other channels, whereas both departures from the band (over-deletion by volume filters and ungated over-expansion) regress RC (Sections 3.2 and 4). The practical rule for LW-DSC post-processing is thus narrow but robust: recover recall conservatively under a probability gate, and never delete a component unless it is confidently a false positive.

Remaining RC gap. Our final RC LW-DSC of 0.549 remains well below the top validation submissions. We conjecture the highest-yielding RC levers lie on the training side (foreground oversampling, a connected-component loss on RC) and detection side (an RC-specialised head) rather than in post-processing; a dedicated RC pathway is our first-priority extension.

Acknowledgments. We thank the BraTS 2026 organisers for curating the challenge datasets, and the Synapse team for hosting the evaluation infrastructure.

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

References

1. Andrearczyk, V., et al.: Automatic detection and multi-component segmentation of brain metastases in longitudinal MRI. Scientific Reports **14** (2024). https://doi.org/10.1038/s41598-024-78865-7

2. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
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-GBM collection. The Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q
4. 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. The Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF
5. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., et al.: Advancing the Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data **4**, 170117 (2017). https://doi.org/10.1038/sdata.2017.117
6. Bancerek, M., Rudzki, P., Nalepa, J.: Segmentation of pre- and post-treatment brain metastases using nnU-Nets. In: Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. pp. 161–172. LNCS, Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-16365-3_15
7. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: MICCAI. LNCS, vol. 9901, pp. 424–432. Springer (2016). https://doi.org/10.1007/978-3-319-46723-8_49
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
9. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
10. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jäger, P.F.: nnU-Net revisited: a call for rigorous validation in 3D medical image segmentation. In: MICCAI. pp. 488–498. LNCS, Springer (2024). https://doi.org/10.1007/978-3-031-72114-4_47
11. Karargyris, A., Umeton, R., Sheller, M.J., Aristizabal, A., George, J., Wuest, A., Pati, S., et al.: Federated benchmarking of medical artificial intelligence with MedPerf. Nature Machine Intelligence **5**, 799–810 (2023). https://doi.org/10.1038/s42256-023-00652-2
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
13. Krikorian, W., Purwar, A.: Automated segmentation for the brain tumor segmentation (BraTS) metastases 2025 challenge using multi-architectural deep learning. In: Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. pp. 173–181. LNCS, Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-16365-3_16
14. Lamba, N., Wen, P.Y., Aizer, A.A.: Epidemiology of brain metastases and leptomeningeal disease. Neuro-Oncology **23**(9), 1447–1456 (2021). https://doi.org/10.1093/neuonc/noab101
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
16. Maier-Hein, L., et al.: Metrics reloaded: recommendations for image analysis validation. Nature Methods **21**(2), 195–212 (2024). https://doi.org/10.1038/s41592-023-02151-z

17. Maleki, N., et al.: Analysis of the MICCAI brain tumor segmentation — metastases (BraTS-METS) 2025 lighthouse challenge: Brain metastasis segmentation on pre- and post-treatment MRI. arXiv preprint arXiv:2504.12527 (2025). https://doi.org/10.48550/arXiv.2504.12527
18. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694
19. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 3DV. pp. 565–571 (2016). https://doi.org/10.1109/3DV.2016.79
20. Moawad, A.W., Janas, A., Baid, U., Ramakrishnan, D., Saluja, R., Ashraf, N., Maleki, N., Jekel, L., et al.: The brain tumor segmentation (BraTS-METS) challenge 2023: Brain metastasis segmentation on pre-treatment MRI. arXiv preprint arXiv:2306.00838 (2023). https://doi.org/10.48550/arXiv.2306.00838
21. Ostrom, Q.T., Wright, C.H., Barnholtz-Sloan, J.S.: Brain metastases: epidemiology. Handbook of Clinical Neurology **149**, 27–42 (2018). https://doi.org/10.1016/B978-0-12-811161-1.00002-5
22. Reinke, A., et al.: Understanding metric-related pitfalls in image analysis validation. Nature Methods **21**(2), 182–194 (2024). https://doi.org/10.1038/s41592-023-02150-0
23. Rohlfing, T., Zahr, N.M., Sullivan, E.V., Pfefferbaum, A.: The SRI24 multichannel atlas of normal adult human brain structure. Human Brain Mapping **31**, 798–819 (2010). https://doi.org/10.1002/hbm.20906
24. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: MICCAI. LNCS, vol. 9351, pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28
25. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: the missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
26. Umeh, F., Yordanov, N.Y., Maleki, N., Amiruddin, R., Moawad, A., Pytlarz, M., Chukwurah, C., Aboian, M.: Taking advantage of MONAI and DiNTS frameworks to develop a state-of-the-art algorithm for automatic segmentation of brain metastases. In: Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. pp. 182–189. LNCS, Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-16365-3_17
27. de Verdier, M.C., et al.: The 2024 brain tumor segmentation (BraTS) challenge: glioma segmentation on post-treatment MRI. arXiv preprint arXiv:2405.18368 (2024). https://doi.org/10.48550/arXiv.2405.18368
28. Wang, J.Y., et al.: Stratified assessment of an FDA-cleared deep learning algorithm for automated detection and contouring of metastatic brain tumors in stereotactic radiosurgery. Radiation Oncology **18**, 61 (2023). https://doi.org/10.1186/s13014-023-02246-z