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

Supplementary Material
Record of Corrections to the Submitted Version

Vineet Jangir and Bhavesh Sandbhor

Indian Institute of Technology Bombay, Mumbai, India
jangirvineet2@gmail.com

Paper: Seeing Lesions, Not Voxels: A Connected-Component-Weighted Dual-Map Focal Loss and Per-Group Oversampling for Brain Metastases Segmentation and Detection
Submission: 26 — BraTS-METS, MICCAI 2026

Purpose of this document. During revision we audited the submitted manuscript against our training logs, configuration files, commit history, submitted artifacts, and the current challenge specification. That audit found statements in the submitted version that were inaccurate, unsupported, or described behaviour the scored system did not exhibit.

Two of the reviewers' comments (C6Ug W2/W3) identified such errors directly. The remainder we found ourselves, and we record them here rather than silently amending them, because several concern what the reported results actually are — in particular which configuration produced them.

Each entry states what the submitted version claimed, what we found to be the case, and where the camera-ready version now stands.

A. Corrections affecting how the reported results should be interpreted

These are the most consequential. Each changes what the reported numbers represent.

A1. The five ensembled folds were not trained under a single configuration

Submitted version claimed: Section 2.4 described "the trainer used for the results reported here, `nnUNetTrainer_CCwFocalDualMapInstrumented`", implying that all five ensembled folds were trained under the method described in Sections 2.5–2.6.

What we found: Verified against trainer sources, launch scripts, and run READMEs, the five folds differ substantially:

Fold 0 contains no connected-component weighting of any kind. Fold 1 contains the false-negative map but not the false-positive map. The per-group oversampling scheme that the paper's title and abstract foreground was used

Fold	Trainer	FN map	FP map	Oversampling	Patch size
0	`RCfocal`	absent	absent	RC ×2, class-wide	112×160×128
1	`fold1`	present	absent	RC ×2 + small-fg	112×160×128
2	`CCwFocalDualMap`	present	present	RC ×3, class-wide	160×224×192
3	`CCwFocalDualMap`	present	present	RC class-wide	112×160×128
4	`CCwFocalDualMap`	present	present	per-group 5/5/2	112×160×128

in fold 4 only. Fold 2 additionally ran with non-unit per-region coefficients (FP 1.44/1.32/1.37/2.00, GT 1.08/1.09/1.09/1.00), contradicting the submitted claim that "the fold reported here runs with all coefficients at 1.0".

Camera-ready: Section 2.4 now reports per-fold composition in full as a table. The reported ensemble is identified as an average over a progressively developed method rather than five instances of one configuration, and this is listed as a limitation in Section 4.3. Two consequences are stated explicitly: the ensemble numbers are numbers for a mixture, and folds 0 and 1 function as intermediate configurations. Also corrected: the submitted patch size ($112 \times 160 \times 128$) did not hold for fold 2, which ran $160 \times 224 \times 192$.

A2. The postprocessing step described was not applied to the scored submissions

Submitted version claimed: Section 2.7 stated that connected components below $5\,\mathrm{mm}^3$ were removed independently within each region, and that "this threshold is applied identically in local validation and portal inference, avoiding a common, quiet source of unexplained local/portal gaps."

What we found: We labelled every region (26-connectivity) across 30 cases in each shipped submission archive. Single-voxel components ($1.0\,\mathrm{mm}^3$) survive in all of them — 43 sub-$5\,\mathrm{mm}^3$ components in the five-fold portal submission, 30 in the three-fold, and 126 in the fold-3 solo submission. The filter was applied in neither local validation nor portal inference. Notably, the fold-3 solo submission's own configuration declares `minimum_component_mm3 = 5.0`, so the filter was requested and did not execute. The container that does apply this filter was written on 2026-07-31, after the portal submissions, and its own README states it must not be described as bit-identical to submission 9774288.

Camera-ready: Section 2.7 as submitted has been removed. A merged Inference section now states that the reported validation results were produced without connected-component filtering, that the submitted container applies a $5\,\mathrm{mm}^3$ filter, and that test-set results produced by the container are therefore not directly comparable to the validation numbers reported in this paper. The motivation for including the filter in the container — the false-positive burden among the smallest predicted components, visible in Table 6 of the main paper — is stated there.

A3. Data provenance of the quality-control corrections is unresolved

Submitted version claimed: "folds 3 and 4 were trained on post-QC-correction data; folds 0, 1, and 2 were trained on pre-correction data."

What we found: The corrections were applied to the raw label tree on 2026-07-25, but the trainer consumes preprocessed volumes and precomputed weight maps built from those labels. Unless the preprocessing regeneration chain was re-run before a fold launched, that fold trained on pre-correction labels while evaluation scored against corrected references. Folds 0–3 all began training before 2026-07-25 and are therefore pre-correction. Fold 4 began on 2026-07-29; whether the regeneration chain ran before that launch could not be verified from the available records.

Camera-ready: Section 2.1 now states folds 0–3 as pre-correction and fold 4 as unverified, and lists this as a provenance limitation in Section 4.3. The submitted version's assignment of folds 3 and 4 to post-correction data was incorrect. Also corrected: the submitted version did not explain what the QC correction was. Section 2.1 now describes it: an organizer-released workbook listing 115 re-annotated training cases, of which 108 matched by filename and 107 passed validation and were applied; the corrections are substantive relabelings, not boundary adjustments.

B. Corrections to claims about what the method does

B1. The method does not operate on lesion instances

Submitted version claimed: that the pipeline was "organized around one idea: make the loss and the sampler see lesion instances, not voxels", and in the conclusion that it treats "lesion instances, not voxels, as the unit of both loss and sampling."

What is actually the case: the loss and the sampler both remain voxel-wise. What the method changes is the weight attached to each voxel and the probability with which cases containing certain groups are drawn — both derived from connected-component statistics. No stage of the pipeline treats a lesion instance as an atomic unit.

Camera-ready: the abstract, Section 1, Section 4.2 and the conclusion have been rewritten to describe the method as instance-aware weighting of a voxel-wise objective. The title is retained, as a compressed statement of the method's intent — pushing the network toward lesions it would otherwise not learn from — but the body no longer claims instance-level operation.

B2. The loss does not equalize lesion contributions

Submitted version claimed: a term "that equalizes a lesion's loss contribution regardless of its volume."

What is actually the case: weights are capped at $w_{\max} = 254$, so the contribution gap is compressed, not removed. Table 4 of the main paper demonstrates this directly: rc_small's weighted share moves only $\sim$1.1$\times$ precisely because the cap binds long before parity is approached.

Camera-ready: the claim now reads that the weighting compresses rather than equalizes, and Section 2.5 states the cap's effect explicitly with Table 4 as the evidence.

B3. "Symmetric" overstated the false-positive map's behaviour

The submitted abstract described penalizing hallucinated components "on the same footing" as missed ones. The same cap applies to both maps, and fold 2 ran with non-unit per-region coefficients, so exact parity did not hold in practice. The camera-ready describes the second map as counterbalancing rather than symmetric.

C. Corrections against the challenge specification

C1. Segmentation scoring threshold

The submitted version stated in Section 1 and Section 2.8 that DSC and NSD are computed only for lesions at or above 275 mm^3. The current specification states that segmentation evaluation metrics apply to lesions larger than 27 mm^3. Both sections have been corrected.

C2. Detection metric

The submitted version referred to detection scoring by "lesion-wise F1 and an AUC over F1 at multiple thresholds" (Section 1) and "F1, AUC-F1" (Section 2.8). The specification defines detection scoring by F1. No AUC-F1 metric is defined. Both references have been corrected.

C3. Provenance of the 275 mm^3 tier boundary

The submitted Section 2.2 described the three size tiers as "matching the thresholds the challenge itself uses to separate its detection-only and segmentation-scored regimes". The challenge defines one threshold, so this claim had no counterpart in the specification. The 275 mm^3 value was taken from an earlier version of the challenge description, under which segmentation was scored only above that volume. When the specification changed to a single 27 mm^3 threshold, we retained 275 mm^3 as an analytical tier boundary, because the 27–275 mm^3 band contains lesions we found difficult to detect and wanted to track separately from both smaller and larger ones.

Camera-ready: the claim of correspondence with challenge thresholds is removed. The abstract, Section 1, Section 2.2 and Section 2.8 now state that only the 27 mm^3 boundary is the challenge's and that the 275 mm^3 boundary is ours, with the reasoning above given in Section 1.

C4. Cohort registration spaces

The submitted Section 2.1 described "650 non-UCSD cases, co-registered to a common space (SRI24)". The correct breakdown of the 1,296 training cases is:

Space	Cohorts	n
SRI24	Duke 37, NCI 35, Missouri 22, WashU 39, Yale 195	328
Native	UCSF	322
Native	UCSD	646
		Total 1296

Of the 650 non-UCSD cases, 328 are SRI24-registered and 322 (UCSF) are in native space. Section 2.1 has been corrected.

C5. The single-fingerprint justification did not test what it claimed

The submitted Section 2.1 justified using one nnU-Net preprocessing fingerprint on the grounds that the anisotropy ratio ($\sim$1.16$\times$) is "well under nnU-Net's 3$\times$ override threshold". The $\sim$1.16$\times$ figure is the post-resample median spacing ratio taken from the generated plans file, whereas nnU-Net's criterion concerns the spread of input spacings, where the native-space acquisitions reach approximately 6.4$\times$. The cited number therefore did not test the condition it was invoked for. Separately, the $\sim$1.4% resampling volume-loss measurement was computed on an 8-case sample from UCSD only, and covers none of the 322 UCSF native-space cases.

Camera-ready: Section 2.1 presents the single fingerprint as a practical decision taken over a heterogeneous pool containing a minority of strongly anisotropic native acquisitions, and states both the measurement's limited scope and the anisotropy issue as limitations rather than as a passed threshold test.

D. Corrections to reported figures

D1. Table 4 omitted one of twelve groups

The submitted Table 4 (loss attribution) and its accompanying figure reported eleven of the twelve groups; snfh_medium was absent and the raw shares summed to approximately 0.989. The missing row is:

Group	Raw share	Weighted share	Effect
snfh_medium	0.0113	0.0350	3.1$\times$ up

With it, the table sums to 1.000. The camera-ready table is complete.

D2. Oversampling presence figures are simulated, not measured

The submitted oversampling table was captioned "Measured effect of per-group RC oversampling on case-presence rate within the training pool." The figures come from a simulated pass over the dataloader's sampling behaviour, not from instrumentation of a live training pool. The camera-ready text now reads "simulated effect", and the source of the figures is stated. The values themselves are unchanged.

E. Results added during revision

The following were not in the submitted version and are not corrections, but are recorded here for completeness as they were surfaced by the same audit.

E1. Within-fold configuration comparisons (fold 3)

Recovered from a coordination record covering fold 3's manual, evidence-gated tuning, each change judged against a full 259-case evaluation:

Epoch	Change	Outcome
506	FP map RC coefficient 1.0 → 1.25	RC instance precision 0.078–0.40 → 0.5417 at recall 0.619
550	GT small-size coefficient 1.0 → 1.25	Small-lesion recall fell: WT 0.440 → 0.344, TC 0.454 → 0.372, ET 0.425 → 0.343 (∼20% fewer TP); RC precision 0.542 → 0.299
577	GT small-size reverted to 1.0	RC FP coefficient retained

The epoch-506 result is the only direct evidence we have that the false-positive map moves detection metrics. The epoch-550 result is reported in the camera-ready as a second negative result: reweighting the loss further toward small lesions reduced small-lesion detection consistently across three independent regions. These are within-fold comparisons at different training epochs, not controlled ablations with matched schedules.

E2. Oversampling factor derivation

The factors (rc_small 5.0, rc_medium 5.0, rc_large 2.0) were set by reading the measured case-presence column of Table 1 (main paper) directly, where the three RC tiers sit at 2.70%, 2.70% and 8.95% against a 29.9% floor cleared by every other group, and confirmed by simulated sampling. Two formula-based schemes were tried first: reciprocal weighting, which gave netc_large (29.9% presence) 2.48× while giving rc_small (2.70%, eleven times rarer) only 3.0×; and presence brackets, which could not express the intended allocation without reverse-engineering bracket edges around three specific groups.

E3. Limitation added: no training-set evaluation

We never evaluated the model on the training set and therefore had no instrument for detecting overfitting. This is most consequential where the method is most aggressive: per-group oversampling re-presents the same ∼50 distinct rc_small and rc_medium lesions up to five times per epoch, the configuration in which memorization is most plausible. The evaluation was omitted to conserve compute under resource constraints; all model-quality decisions were taken on validation-set metrics alone. It can be run against the existing checkpoints under the same protocol if required.

F. Material removed for length

The following were removed to bring the body within the page limit in effect at each revision stage, following Reviewer C6Ug's identification of material that neither shapes the submitted system nor supports a reported number: storage throughput and dataloader worker-count timings; GPU idle time during evaluation; the export-order bug anecdote; the size-conditional ensemble described as explored but not applied; the per-class volumetric contribution table computed on the non-UCSD subset only; and figures that re-plotted data already tabulated. The "Relation to prior instance-aware losses" discussion was condensed accordingly at each stage; a discussion of prior BraTS-METS challenge outcomes was added in its place.

A qualitative results figure was considered in response to Reviewer BKrD's point 8, but was not added: the challenge does not release ground-truth labels for the official 179-case validation set, so a prediction-versus-ground-truth comparison on those cases cannot be constructed. Our response to BKrD states this directly.