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

CarveMix-RC: Addressing Rare-Class Imbalance Through Lesion-Aware Synthetic Augmentation for Brain Metastasis Segmentation

S1. Validation Cohort and Bootstrap Procedure

Statistical and paired bootstrap analysis

The online validation cohort comprised 179 cases. To quantify uncertainty in the observed validation improvements, we performed a case-level paired bootstrap using 10,000 resamples of the case-wise lesion DSC values obtained after inference; segmentation inference itself was not repeated. Because all configurations were evaluated on the same validation cases, case indices were sampled with replacement while preserving the pairing between configurations. For configurations A and B, the mean paired difference for region r was defined as $\Delta_r^{(AB)} = (1/N_r)\,\Sigma_{i=1}^{Nr}\,[DSC_{i,r}^{(B)} - DSC_{i,r}^{(A)}]$, and the 95% confidence interval was obtained from the 2.5th and 97.5th percentiles of the bootstrap distribution. Region-specific analyses included cases with defined lesion-wise DSC (ET: N = 156; TC: N = 156; WT: N = 161; RC: N = 22). Relative to the ResEncL baseline (C2), the final ensemble (C9) improved DSC by +0.055 for ET (95% CI [0.035, 0.078]), +0.053 for TC ([0.033, 0.076]), +0.064 for WT ([0.039, 0.092]), and +0.171 for RC ([0.067, 0.285]); all four intervals exclude zero. The larger RC gain is accompanied by a wider interval because only 22 RC cases were available, so isolated RC component effects are interpreted cautiously. These intervals quantify case-sampling uncertainty in the observed validation-set differences and not variability due to model retraining or stochastic inference.

S2. Configuration Definitions

ID	Configuration	Architecture	RC-W	CM	PP
C1	Baseline	Standard nnU-Net	No	No	No
C2	Baseline	ResEncL	No	No	No
C3	+ RC weighting	ResEncL	Yes	No	No
C4	+ RC weighting + PP	ResEncL	Yes	No	Yes
C5	+ RC-W + CarveMix	Standard nnU-Net	Yes	Yes	No
C6	+ RC-W + CarveMix	ResEncL	Yes	Yes	No
C7	+ RC-W + CarveMix + PP	Standard nnU-Net	Yes	Yes	Yes

ID	Configu-ration	Architec-ture	RC-W	CM	PP
C8	+ RC-W + CarveMix + PP	ResEncL	Yes	Yes	Yes
C9	Final en-semble	Std + ResEncL	Yes	Yes	Yes

Table 1. S1. Definition of the nine evaluated configurations. RC-W: RC-weighted optimization; CM: CarveMix; PP: lesion-aware post-processing.

S3. Paired Bootstrap Relative to the ResEncL Baseline

ID	ET Δ DSC [95% CI]	TC Δ DSC [95% CI]	WT Δ DSC [95% CI]	RC Δ DSC [95% CI]
C1	−0.012 [−0.032, +0.009]	−0.011 [−0.031, +0.009]	+0.003 [−0.025, +0.031]	−0.035 [−0.167, +0.089]
C2	Reference	Reference	Reference	Reference
C3	−0.015 [−0.037, +0.007]	−0.013 [−0.034, +0.009]	+0.003 [−0.022, +0.029]	+0.076 [−0.014, +0.172]
C4	**+0.051 [+0.031, +0.072]**	**+0.049 [+0.030, +0.070]**	**+0.050 [+0.029, +0.073]**	**+0.137 [+0.046, +0.241]**
C5	−0.020 [−0.045, +0.003]	−0.023 [−0.047, +0.001]	+0.002 [−0.028, +0.030]	+0.031 [−0.039, +0.105]
C6	**+0.038 [+0.018, +0.059]**	**+0.037 [+0.016, +0.057]**	**+0.045 [+0.021, +0.070]**	**+0.126 [+0.030, +0.240]**
C7	+0.000 [−0.018, +0.019]	+0.006 [−0.013, +0.024]	**+0.027 [+0.006, +0.049]**	+0.077 [−0.015, +0.179]
C8	**+0.034 [+0.013, +0.054]**	**+0.027 [+0.007, +0.049]**	**+0.042 [+0.015, +0.070]**	**+0.171 [+0.064, +0.289]**
C9	**+0.055 [+0.035, +0.078]**	**+0.053 [+0.033, +0.076]**	**+0.064 [+0.039, +0.092]**	**+0.171 [+0.067, +0.285]**

Table 2. S2. Case-level paired bootstrap differences relative to the ResEncL baseline (C2). Values mean paired lesion-wise DSC differences with 95% percentile CIs. Bold indicates a CI excluding zero.

S4. Component-Level RC Transitions

Comparison	Effect	RC Δ DSC	95% CI
C2 → C3	RC weighting	+0.076	[−0.014, +0.172]
C3 → C4	Post-processing; no CarveMix	+0.061	[−0.036, +0.158]
C3 → C6	CarveMix; no PP	+0.050	[−0.056, +0.162]
C4 → C8	CarveMix; with PP	+0.035	[−0.025, +0.100]
C6 → C8	Post-processing; with CarveMix	+0.045	[−0.046, +0.142]
C8 → C9	Ensembling	≈0.000	[−0.065, +0.065]
C2 → C9	**Full pipeline**	**+0.171**	**[+0.067, +0.285]**

Table 3. S3. Direct paired RC comparisons for the principal component transitions. The wider intervals reflect the limited RC-positive subset (22 eligible RC cases). The isolated component increments are positive but their CIs cross zero; the end-to-end C2→C9 improvement is clearly separated from zero.

S5. Lesion-Size Detection Analysis

Table 4. S4. Lesion-level detection performance of the final ensemble stratified by lesion size. TP/FP/FN and F1 are reported by the official BraTS-METS evaluation; precision and recall are derived as TP/(TP+FP) and TP/(TP+FN). Precision is undefined for small RC lesions because TP = FP = 0.

Region	Lesion size	TP	FP	FN		Precision	Recall	F1
ET	Small	1.31	0.57	5.41		0.456	0.325	0.347
	Large	4.10	0.35	0.55		0.899	0.917	0.896
	All	**4.13**	**0.35**	**2.74**		**0.902**	**0.807**	**0.811**
TC	Small	1.32	0.53	5.36		0.464	0.330	0.354
	Large	4.10	0.33	0.54		0.900	0.921	0.899
	All	**4.14**	**0.33**	**2.68**		**0.904**	**0.814**	**0.818**
WT	Small	0.96	0.79	5.07		0.331	0.241	0.251
	Large	3.79	0.48	0.63		0.886	0.908	0.882
	All	3.87	0.47	2.66		0.899	0.794	0.805
RC	Small	0.00	0.00	1.17		0.000	0.000	0.000
	Large	0.91	0.18	0.18		0.818	0.841	0.817
	All	**0.12**	**0.14**	**0.06**		**0.545**	**0.754**	**0.476**

Large-lesion recall is high for ET, TC, and WT (0.917, 0.921, and 0.908), whereas small-lesion recall decreases to 0.325, 0.330, and 0.241, respectively. RC remains the

most difficult region: small-RC recall is 0, while large-RC recall is 0.841. These results identify small-lesion false negatives as the dominant lesion-detection limitation.

Table 5. S5. Method-defining and training parameters used for RC-aware optimization, CarveMix-RC augmentation, and lesion-aware post-processing.

Parameter	Value	Description
Training epochs	1000	nnU-Net v2 training schedule
Initial learning rate, η_0	0.01	Initial learning rate
Dice-loss weight, w_{Dice}	1.0	Contribution of Dice loss
Cross-entropy weight, w_{CE}	1.0	Contribution of cross-entropy loss
RC class weight, w_{RC}	3.0	Cross-entropy weight assigned to RC
Foreground oversampling probability, p_{fg}	0.66	Probability of foreground-focused patch sampling
CarveMix distance parameter, λ	$\lambda \sim \mathcal{U}(-3,5)$	Signed-distance threshold controlling RC contraction/expansion
Minimum donor RC size, V_{min}^{donor}	50 voxels	Minimum cavity size for donor eligibility
Number of synthetic cases, T	300	Number of generated CarveMix-RC cases
Minimum RC component size, V_{min}^{RC}	50 voxels	Smaller RC components removed during post-processing
Minimum ET component size, V_{min}^{ET}	15 voxels	Smaller ET components removed during post-processing
Component connectivity	26-connected	3D connected-component definition
RC hole filling	Binary hole filling	Applied after RC component filtering

Note. All other preprocessing, augmentation, optimization, and inference settings followed the automatically configured nnU-Net v2 pipeline unless otherwise specified.