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.

RESIN: Residual Synthesis for Healthy-Tissue Inpainting in Brain MRI

Quan Huu Cap

AI Development Department, Aillis, Inc., Tokyo, Japan
quan.cap@aillis.jp

Abstract. Brain-MRI analysis pipelines often assume healthy anatomy and can degrade in the presence of pathology. The BraTS 2026 Inpainting challenge asks models to fill masked regions, while evaluating reconstruction inside a withheld healthy region. Most prior entries use encoder-decoder U-Nets that gather context through downsampling, potentially sacrificing fine spatial detail. We propose RESIN, a three-dimensional, norm-free adaptation of SRResNet that processes the volume entirely at the input resolution. We evaluate it using patient-grouped splits, paired case-level comparisons and the official scorer. Despite using 5.43M parameters versus the U-Net's 15.07M, RESIN outperforms the tuned U-Net on SSIM, PSNR and RMSE on both held-out folds and remains better in every void-size quartile. Because its receptive field spans the scored regions, these results support full-resolution processing rather than locality as the source of its advantage. Across paired ablations, annealing the learning rate to zero is the strongest training intervention, exceeding the gains from added capacity, test-time augmentation and alternative loss terms without increasing inference cost. An ensemble of three RESIN models attains SSIM 0.8436, PSNR 24.657 and MSE 0.00541 on the official validation set, surpassing the published 2023 and 2024 challenge winners on all three metrics, though not the 2025 winner.

Keywords: Brain MRI · Inpainting · Image Synthesis · Residual Networks · Ensembling · BraTS

1 Introduction

Brain-MRI analysis pipelines are typically developed on healthy anatomy and degrade in the presence of pathology. Replacing a lesion with plausible healthy tissue restores the assumptions such pipelines rely on, and the BraTS inpainting challenge[1] formalizes this need: given a native T1 (t1n) volume with a region zeroed and a binary mask of that region, an algorithm must fill the mask with tissue that is anatomically plausible and intensity-consistent with the surrounding brain [8]. The mask supplied at inference is the *union* of a dilated tumor region and a separate patch placed over genuinely healthy tissue, so the input usually has two holes. An algorithm must fill both, yet evaluation looks *only*

[1] We participate in the challenge under the Synapse team name `TheCapybara`.

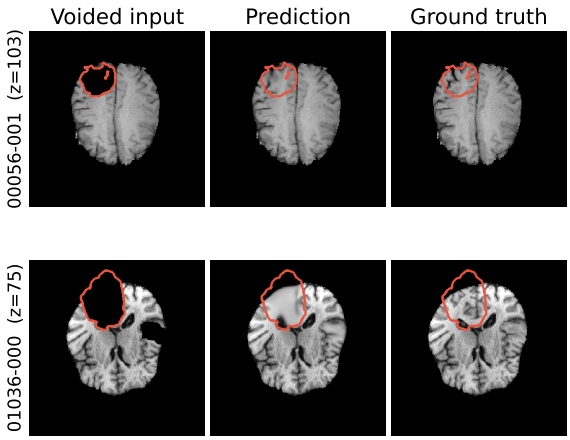

Fig. 1. Task structure and representative outputs for an easy case (top) and a hard case (bottom). Only the healthy sub-region (red outline) is scored. The small void is filled consistently with the surrounding cortex, whereas the larger void recovers coarse anatomy but loses fine interior texture. In the bottom row, the tumor void at left is filled but not graded. All panels share one intensity window.

inside the hidden healthy patch, where the withheld ground truth is real healthy tissue. Faithful synthesis of healthy tissue is therefore what is rewarded, while the tumor region is filled but never graded (Fig. 1).

Every previous edition of this task has been won by a U-Net [16] that predicts a single tissue value per voxel, trained with a combination of L_1 and SSIM losses [19,20]. This established approach combines two design choices; we retain the first and revisit the second. The first is to predict a single best estimate for each voxel, rather than to sample one of the many healthy images that could plausibly fill the hole. This suits the challenge's full-reference metrics, which reward a fill for being numerically close to the withheld ground truth, not for looking realistic, so a generative model that synthesizes sharp but subtly different tissue is penalized by these very metrics [5]. We therefore retain the deterministic, single-estimate formulation.

The second decision is the network itself, and this is the one we revisit. A U-Net gathers context by repeatedly downsampling the volume, on the assumption that a large hole cannot be filled from its rim alone; but every downsampling step discards the fine texture that the scoring metrics reward. We propose **RESIN** (**RE**sidual **S**ynthesis for **IN**painting), a three-dimensional adaptation of the SRResNet super-resolution network [9]. RESIN is a deep stack of residual convolutional blocks that carries the volume at a single, full spatial resolution from input to output, with no downsampling, no upsampling and no internal normalization, so every layer preserves fine detail. We designed it from the intuition that a healthy fill depends mainly on the tissue immediately around the hole, so that preserving fine local detail should matter more than gathering

distant context. RESIN's receptive field in fact spans the scored voids, so our results support full-resolution processing but do not show that distant context is unnecessary. To our knowledge, prior entries to this task all downsample at some point; none uses a full-resolution residual backbone.

On both held-out folds we evaluate, RESIN outperforms a U-Net three times its size on every ranked metric, and its margin does not shrink as the hole grows. Our contribution is a challenge-specific adaptation of a known backbone, backed by paired empirical evidence that full-resolution residual processing outperforms the established U-Net design. Alongside the model we examine how the challenge metric shapes solution design, and we compare every design choice case by case rather than on averages alone. Inference code and trained weights are available at https://github.com/huuquan1994/brats2026_task4_inpainting.

2 Methods

2.1 Data

We use the BraTS 2023 Local Synthesis training set (1251 cases) and validation set (219 cases), built on the BraTS-GLI 2023 collection [8,1,12,4,2,3]. All volumes are native T1, $240 \times 240 \times 155$ at 1 mm isotropic, skull-stripped and co-registered to the SRI24 atlas [15]. Pre-processing of the released data was performed by the organizers using CaPTk [13] and the FeTS tool [14]; skull-stripping follows Thakur et al. [17]. We compute evaluation metrics with the official BraTS `inpainting` package[2] [8]. No additional public or private data was used: every model reported here is trained only on the released training set.

Each training case provides ground-truth and voided volumes and a union mask with disjoint dilated-tumor and healthy components. The voided volume is zero inside the union mask and unchanged outside it.

Validation split. Cases share patient identifiers across timepoints, so a naive random split leaks a patient across the boundary. We use stratified grouped k-fold cross-validation ($k = 5$), grouping on patient identifier and stratifying on a quantile bin of healthy-mask volume, so no patient contributes to both sides of a split and the fold-to-fold distribution of scored-region size is matched. Every choice of architecture, loss, learning-rate schedule and checkpoint was made on fold 0 (250 held-out cases). The official validation set was used only to choose the composition of the final cross-fold ensemble, not to train models or select checkpoints.

2.2 Intensity normalization

Raw intensities vary substantially across cases, so we use the same per-case transform as the evaluation: clip the voided input to its 0.5th and 99.5th percentiles,

[2] https://pypi.org/project/inpainting/

scale to $[0, 1]$, then map affinely to $[-1, 1]$. Predictions are inverted to the original scale before writing. This aligns the model input with the scorer and is less sensitive to outliers than normalization by the per-case maximum. This input-side intensity transform is distinct from normalization inside the networks: the U-Net uses instance normalization, whereas RESIN removes SRResNet's batch normalization.

2.3 Architectures

We train two backbones, both taking the normalized voided volume and binary union mask as input channels and predicting the complete volume in normalized intensity space. The prediction is restricted to the mask at compositing time.

The U-Net has four resolution levels with 48, 96, 192 and 384 channels, instance normalization, PReLU activations, dropout 0.2, and strided and transposed convolutions. Built with MONAI [6], it has 15.07M parameters and an effectively global receptive field. The generator we propose, **RESIN**, adapts the SRResNet super-resolution network [9] to volumetric inpainting. We keep its residual backbone, an initial convolution feeding residual blocks with a global skip to a final projection, and depart from the original in four ways. We remove its upsampling stages so the network remains at the input resolution. We remove batch normalization because batch statistics are unreliable at our batch size of two; the norm-free design also preserves feature range and saves memory [10]. We replace the 2D convolutions with volumetric ones and reduce the head and tail kernels from 9^3 to 3^3 to control computational cost. Finally, we deepen the stack from 16 to 24 residual blocks, providing enough receptive field to span the scored voids. The result has width 64, PReLU activations and 5.43M parameters, and carries the volume at full resolution throughout.

2.4 Objective

We train with a composite objective

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_{\text{SSIM}} \big(1 - \text{SSIM}\big), \tag{1}$$

with $\lambda_1 = \lambda_{\text{SSIM}} = 1$, SSIM being the structural similarity index [18]. Both terms are restricted to the healthy sub-mask, the region the challenge actually scores, rather than the full union.

2.5 Training

Volumes are cropped to the brain bounding box then randomly cropped to $192 \times 192 \times 144$ (U-Net) or $160 \times 160 \times 144$ (RESIN). RESIN uses the smaller crop because full-resolution processing requires more GPU memory. Augmentation is axis flips and 90° rotations, each with probability 0.5. We use Adam [7] ($\beta = (0.9, 0.999)$), batch size 2, bf16 mixed precision, and run 280 epochs with a cosine schedule [11] decaying the learning rate from 10^{-4} to zero over the whole run.

2.6 Inference

The network is applied to the full $240 \times 240 \times 155$ volume with Gaussian-weighted sliding-window inference at each model's own training crop size (window batch 1, 25% overlap), plus optional test-time augmentation (TTA). Our TTA runs the same network on eight flipped copies of the input, one for each subset of the three spatial axes, flips each prediction back, and averages the eight predictions in normalized space before de-normalization and clipping. The submitted system is an equally weighted ensemble of three RESIN models, one per held-out fold (0, 1 and 2), each with $8\times$ flip TTA. Each member is de-normalized through *its own* percentile scale and the members are averaged in original-intensity space. Predictions are written back only inside the supplied mask, all other voxels copied unchanged:

$$\hat{y} = (1 - m) \odot y_{\text{voided}} + m \odot \tilde{f}(x), \tag{2}$$

where m is the supplied union mask and $\tilde{f}(x)$ is the complete predicted intensity volume, not a residual, mapped back to the original scale and clipped at zero. Thus out-of-mask voxels remain bit-identical to the input, and outputs preserve the input geometry and affine.

Inference cost. Timed with the submitted code on a single GPU in fp32, inference takes 7.3 s per case for one model, 55 s with $8\times$ TTA and 166 s for the submitted three-member ensemble with TTA. Peak GPU memory is 4.75 GiB because members run sequentially.

3 Results

3.1 What the ranking rewards

The challenge ranks submissions by an equally weighted rank-sum over SSIM, PSNR and RMSE, each ranked per case across participants [8]. PSNR and RMSE are not independent: because the ground-truth intensity range is fixed across participants for each case, they produce the same ranking. Specifically, the scorer computes PSNR as $10 \log_{10}((\max - \min)^2/\text{MSE})$, where the range is taken from the ground truth and is therefore common to all submissions for that case. The effective score is therefore $\text{rank}_{\text{SSIM}} + 2\,\text{rank}_{\text{RMSE}}$, giving distortion twice the weight of structural similarity. We use this score to select checkpoints and ensemble weights. We report RMSE on our held-out folds; the official validation leaderboard instead returns MSE, which produces the same ordering and is quoted unchanged in Table 3.

3.2 What moves the score

Each comparison in Table 1 is paired on case because variation in case difficulty exceeds the measured effects. Annealing the learning rate to zero is the

Table 1. Design choices evaluated on held-out fold 0 (250 cases), official metrics inside the healthy mask, each paired against its own control. Positive Δ is better; sorted by ΔSSIM.

Change (vs. control)	ΔSSIM	ΔPSNR (dB)
Learning-rate anneal to zero (vs. constant)	+0.0146	+0.51
8× flip test-time augmentation (vs. none)	+0.0038	+0.27
U-Net width, 48 vs. 32 channels	+0.0037	+0.42
Fourier L_1 loss term (vs. none)	+0.0001	+0.05
Multi-scale SSIM loss (vs. single scale)	−0.0051	−0.15

Table 2. Backbone comparison on two held-out folds (250 cases each), paired on case using official metrics inside the healthy mask. Top: fold scores, with Δ = RESIN minus U-Net. Bottom: ΔSSIM by scored-void-size quartile (fold-0 median volumes in parentheses).

Fold 0	SSIM ↑	PSNR (dB) ↑	RMSE ↓
U-Net	0.8406	22.473	0.0725
RESIN	**0.8522**	**23.456**	**0.0672**
Δ	+0.0116	+0.98	−0.005
Fold 1			
U-Net	0.8066	21.421	0.0823
RESIN	**0.8378**	**23.120**	**0.0713**
Δ	+0.0312	+1.70	−0.011

ΔSSIM by void size	Q1 (1.5k vox)	Q2 (10.5k)	Q3 (34k)	Q4 (87k)
Fold 0	+0.0082	+0.0188	+0.0149	+0.0062
Fold 1	+0.0175	+0.0404	+0.0380	+0.0291

largest lever, adding 0.0146 SSIM at no inference cost. Test-time augmentation and width provide smaller gains. The Fourier L_1 term changes SSIM by only +0.0001, whereas multi-scale SSIM reduces it by 0.0051; neither auxiliary loss was retained.

We rank runs with the organizers' scorer rather than training-time SSIM. The latter is computed in each run's internal normalization scale, so comparisons across runs can reverse the official ordering.

3.3 RESIN versus the U-Net

Both backbones use the same schedule, normalization, loss and epoch budget. RESIN (5.43M parameters) beats the U-Net (15.07M) on all three ranked metrics on both folds (Table 2). Their training crops differ: the U-Net uses 192×192×144, whereas RESIN uses 160 × 160 × 144. Larger crops helped in our U-Net sweep, so this asymmetry disadvantages RESIN rather than explaining its gain. For a fixed recipe, the fold-to-fold spread is 0.034 SSIM, larger than any effect in Table 1; absolute scores are therefore compared only within a fold.

Table 3. Official validation results (219 cases). Our rows use 8× flip TTA; bold marks the best of our submissions. Prior-winner values are published results on the same set [20]. The leaderboard reports MSE rather than RMSE.

	SSIM ↑	PSNR ↑	MSE ↓
Ours: U-Net	0.8193	23.316	0.00650
Ours: single RESIN	0.8422	24.554	0.00552
Ours: RESIN ensemble (2 folds)	0.8435	24.646	0.00543
Ours: RESIN ensemble (3 folds)	**0.8436**	**24.657**	0.00541
Ours: RESIN ensemble (3 folds) + 4 U-Nets	0.8416	24.575	**0.00535**
2023 winner [19]	0.812	21.446	0.009
2024 winner	0.841	23.381	0.007
2025 winner [20]	0.873	24.996	0.00476

The SSIM advantage is +0.0116 on fold 0, comparable to the strongest intervention in Table 1, and +0.0312 on fold 1. The absolute magnitudes differ, but the direction is consistent across both paired evaluations despite RESIN using about one third as many parameters.

RESIN also wins every void-size quartile on both folds, with no loss of advantage on the largest voids. Since its receptive field spans the scored regions (Section 2.3), this supports the value of full-resolution processing rather than a benefit from restricting context. The quartile analysis is important because it tests the main concern about avoiding downsampling: if RESIN lacked sufficient context, its margin should collapse on the largest regions, but it remains positive in both folds. The comparison establishes RESIN as the stronger complete configuration but does not isolate which architectural difference is responsible.

3.4 Ensembling and the final system

Our final system averages, with equal weight, the predictions of three RESIN models (the folds 0, 1 and 2 members, each with 8× flip TTA), combined in the original intensity space. Averaging independent same-family members reduces variance: on the validation set a second fold member adds +0.0013 SSIM over a single model, and a third gives a further small gain (Table 3). Adding U-Net members slightly improves MSE but worsens SSIM and PSNR, so the submitted system contains RESIN models only. Concretely, adding four U-Nets to the three RESIN members lowers MSE by about 1% but costs 0.0020 SSIM and 0.08 dB PSNR. Ensemble composition was selected on the official validation set because no single fold can fairly evaluate members trained on its cases.

3.5 Validation leaderboard

A single RESIN already exceeds the published 2023 and 2024 winners on all three metrics. Our submitted three-fold ensemble reaches SSIM 0.8436, PSNR 24.657 and MSE 0.00541, the best of our submissions on SSIM and PSNR. The

seven-member variant has slightly lower mean MSE but worse SSIM and PSNR. This does not contradict the per-case rank equivalence in Section 3.1: averages of MSE and PSNR weight difficult cases differently. Our final system remains below the 2025 winner by 0.029 SSIM, 0.34 dB PSNR and 12% MSE. That system used five generated healthy masks per case rather than the single released mask, an orthogonal source of training variation that we did not explore.

4 Discussion

Two findings stand out. Within a fixed architecture, annealing the learning rate to zero was the strongest intervention and added no inference cost. Across architectures, RESIN beat a U-Net three times its size on both folds and retained its advantage on large voids. Because RESIN's receptive field spans the scored regions, this result supports preserving full-resolution features rather than restricting context. Together, these findings suggest that optimization schedule and spatial representation matter more for this task than added capacity or more elaborate loss functions. This conclusion concerns the complete configurations tested, rather than a single isolated architectural choice.

A second lesson is methodological. Variation in case difficulty is much larger than the effects under study, making paired comparisons essential. Runs must also be ranked with the official scorer because training-time SSIM uses run-specific intensity scales and can invert their ordering. Without pairing, a small change in case mix could overwhelm the effects in Table 1; without official scoring, even the direction of a comparison can be wrong. The evaluation protocol is therefore part of the method, not only a reporting choice.

Limitations. Model selection used one held-out fold, so its effect sizes may not generalize. The backbone comparison covers two complete configurations that also differ in normalization, dropout, residual structure, parameter count and training crop; it therefore does not isolate the effect of downsampling. Development compute was also asymmetric, with four U-Net folds and three RESIN folds trained, although the backbone comparisons pair corresponding folds. Performance still falls sharply with hole size ($r = -0.75$), and ensembling does little for the worst cases. Moreover, the benchmark grades only the hidden healthy sub-mask. Synthesis inside the actual lesion is never scored and remains unvalidated. A high benchmark score therefore demonstrates recovery of withheld healthy anatomy, but cannot establish that tissue synthesized inside a real lesion is anatomically correct; no counterfactual healthy target exists there. Finally, because meningioma training data are disallowed, our models see only glioma appearance and the resulting domain shift is unmeasured.

Acknowledgments. Data used in this publication were obtained as part of the Challenge project through Synapse ID (syn74274097).

Disclosure of Interests. The author has no competing interests to declare.

References

1. Baid, U., Ghodasara, S., Mohan, S., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
2. 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
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-LGG collection. The Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF
4. 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
5. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6228–6237 (2018). https://doi.org/10.1109/CVPR.2018.00652
6. Cardoso, M.J., et al.: MONAI: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
8. Kofler, F., et al.: The brain tumor segmentation (BraTS) challenge: Local synthesis of healthy brain tissue via inpainting. arXiv preprint arXiv:2305.08992 (2023)
9. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4681–4690 (2017). https://doi.org/10.1109/CVPR.2017.19
10. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1132–1140 (2017). https://doi.org/10.1109/CVPRW.2017.151
11. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (ICLR) (2017)
12. 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
13. Pati, S., et al.: The cancer imaging phenomics toolkit (CaPTk): Technical overview. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pp. 380–394. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-46643-5_38
14. Pati, S., et al.: The federated tumor segmentation (FeTS) tool: An open-source solution to further solid tumor research. Physics in Medicine and Biology **67**(20), 204002 (2022). https://doi.org/10.1088/1361-6560/ac9449
15. Rohlfing, T., Zahr, N.M., Sullivan, E.V., Pfefferbaum, A.: The SRI24 multichannel atlas of normal adult human brain structure. Human Brain Mapping **31**(5), 798–819 (2010). https://doi.org/10.1002/hbm.20906

16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28
17. Thakur, S., Doshi, J., Pati, S., et al.: Brain extraction on MRI scans in presence of diffuse glioma: Multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training. NeuroImage **220**, 117081 (2020). https://doi.org/10.1016/j.neuroimage.2020.117081
18. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
19. Zhang, J., Chen, K., Weng, Y.: Synthesis of healthy tissue within tumor area via U-Net. In: Brain Tumor Segmentation, and Cross-Modality Domain Adaptation for Medical Image Segmentation. pp. 233–240. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-76163-8_21
20. Zhang, J., Weng, Y., Chen, K.: Robust 3D brain MRI inpainting with random masking augmentation. In: Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. pp. 102–109. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-16370-7_9