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.

Mask-Guided 3D U-Net with Component-Wise Tiled Inference for Healthy Brain Tissue Inpainting

Lingyu Guo[1]

James Watt School of Engineering, University of Glasgow, Glasgow, United Kingdom
3119118G@student.gla.ac.uk

Abstract. This paper presents the final solution of team *waterdog* for the BraTS 2026 healthy brain tissue image inpainting task. We constructed a compact mask-guided 3D U-Net that takes a voided T1-weighted MRI volume and its binary reconstruction mask as two input channels. Training employed target-centred three-dimensional cropping, online relocation of healthy-tissue masks, spatial and intensity augmentation, and an in-mask loss combining L1, mean squared error (MSE), and local three-dimensional structural similarity (SSIM). During inference, connected mask components were covered using overlapping tiles. Four test-time views were averaged, followed by Gaussian-weighted fusion of adjacent tiles. Hard composition ensured that observed voxels remained unchanged. The selected model achieved an MSE of 0.012794, a PSNR of 27.083 dB, and an SSIM of 0.8723 on the internal validation set. On the 219-case organiser-provided validation set, it achieved an MSE of 0.008 ± 0.005, a PSNR of 22.340 ± 3.963 dB, and an SSIM of 0.798 ± 0.119. Because the two evaluation protocols use different intensity normalisation and spatial support, internal results guided model selection and organiser-reported results describe final submission performance. The Dockerised system runs reproducibly on one GPU. Qualitative inspection nevertheless shows residual smoothing of fine anatomical detail.

Keywords: Brain MRI · image inpainting · 3D U-Net · BraTS · test-time augmentation

1 Introduction

Brain magnetic resonance imaging (MRI) provides strong soft-tissue contrast, but pathology, acquisition artefacts, surgical resection, or deliberate benchmark masking may obscure or remove tissue in selected regions. Image inpainting estimates plausible content within these masked regions from the remaining anatomy. In BraTS 2026 Task 4, the input is a skull-stripped T1-weighted volume containing one or more voids and a binary inpainting mask. The output must retain the input geometry and observed voxels outside the mask.

Simple interpolation is insufficient. Healthy tissue includes grey and white matter, the ventricles and sulci, and cerebrospinal fluid. Its boundaries are continuous but non-uniform. The model needs local texture, three-dimensional context, and the ability to handle unseen mask shapes and positions. BraTS provides a common benchmark for this task [6]. Its metrics include MSE, PSNR, and SSIM.

We used a conventional 3D U-Net, which extends U-Net to volumetric data and passes spatial detail through encoder–decoder skip connections [11,2]. Instead of proposing another architecture, we focused on turning this compact model into a reproducible pipeline by matching mask-aware training, full-volume inference, and the required output format. The main contributions are:

- A two-channel, mask-guided 3D U-Net whose final composition changes only the unknown voxels and copies the known image regions exactly.
- Online relocation of healthy-tissue masks to create more varied training samples without selecting known abnormal tissue as a target.
- Full-volume inference based on connected components, overlapping tiles, four flip views, and Gaussian fusion.
- An offline Docker implementation using one selected model state.

2 Related Work

U-Net combines multi-scale context with high-resolution skip features and has become a standard design in biomedical imaging [11]. Its three-dimensional extension enables direct learning of volumetric context [2]. nnU-Net further demonstrates that preprocessing, patch selection, augmentation, and inference design may be as important as a new architecture [5]. BraTS has also driven robust three-dimensional methods for heterogeneous brain MRI [9,1]. Recent healthy-tissue restoration research indicates that supervised U-Net models can be effective when training and inference pipelines are aligned [16].

Mask-aware image inpainting methods include partial convolution and gated convolution. Partial convolution excludes invalid pixels during feature updates [7], whereas gated convolution learns a soft feature selection mechanism for irregular missing regions [15]. These operations control which features are updated, but a full three-dimensional implementation would require more memory and development time. We instead add the binary mask as a second input channel to ordinary 3D convolutions. This adapts standard 3D U-Net components without introducing a specialised inpainting operator. The loss also includes SSIM, which measures local luminance, contrast, and structure [14].

3 Method

3.1 Data, preprocessing, and formulation

The challenge training dataset contains 1,251 cases. Each has a complete T1 image (`t1n`), a voided version, and healthy and unhealthy masks, all sized $240 \times$

240×155. With seed 2023, we assigned 1,000 cases to training and 251 to internal validation. The split was performed at the patient-case level, so no case appeared in both subsets, preventing patient-level data leakage. The organiser-provided validation set contains 219 voided T1 volumes and their binary inpainting masks.

Before cropping, negative intensities are set to zero. Each volume is then divided by its positive maximum and mapped from $[0, 1]$ to $[-1, 1]$. The maximum is retained to restore the original scale. Training uses $128 \times 128 \times 112$ crops centred on healthy targets and generates two samples per case per epoch. For the second sample, the healthy mask is relocated with probability 0.9 and may be axially rotated by multiples of 90° and flipped. Candidate locations must lie in positive-intensity brain tissue and not overlap the unhealthy mask after five dilations. This increases mask diversity without reconstructing known tumour tissue as healthy tissue. Intensity augmentation nonlinearly changes contrast as $x' = x^{\gamma}$ after clipping intensities to $[0, 1]$, with $\gamma \sim U(0.85, 1.15)$. Values below one brighten intermediate intensities, whereas values above one attenuate them. Additional spatial augmentation uses axial 90° rotations and independent flips. Such targeted augmentation can improve generalisation with limited medical data [12].

Let X represent the complete T1 volume, M_h the healthy-tissue mask, and M_u the unhealthy-region mask. The input is voided over their union, $X_v = X \odot [1 - (M_h \vee M_u)]$, whereas conditioning and supervision use only M_h to avoid treating tumour voxels as healthy targets. Let $M \in \{0, 1\}^{D \times H \times W}$ denote the binary mask used by the current stage, where $M_i = 1$ identifies a voxel to reconstruct and $M_i = 0$ identifies an observed voxel. During training, $M = M_h$. During organiser-provided validation inference, M is the current connected component of the binary inpainting mask supplied with the voided image. The network receives the concatenated normalised image and mask:

$$P = f_\theta([X_v, M]). \tag{1}$$

The prediction is inserted by hard composition,

$$\hat{X} = (1 - M) \odot X_v + M \odot P. \tag{2}$$

Thus, the binary mask conditions the network and controls hard composition; during training it also restricts the supervised loss. All observed voxels outside the current mask are preserved exactly. The distinction between healthy masks used for training and organiser-provided inpainting masks used for validation is considered when interpreting the results in Section 5.

3.2 Network and training objective

The U-Net has two input channels and one output channel. The encoder uses 32, 64, and 128 channels, followed by a 256-channel bridge and symmetric decoder. Each block has two reflection-padded $3 \times 3 \times 3$ convolutions, instance normalisation, and PReLU. Stride-two convolutions downsample; transposed convolutions upsample before concatenation with skip features. Dropout3D with probability

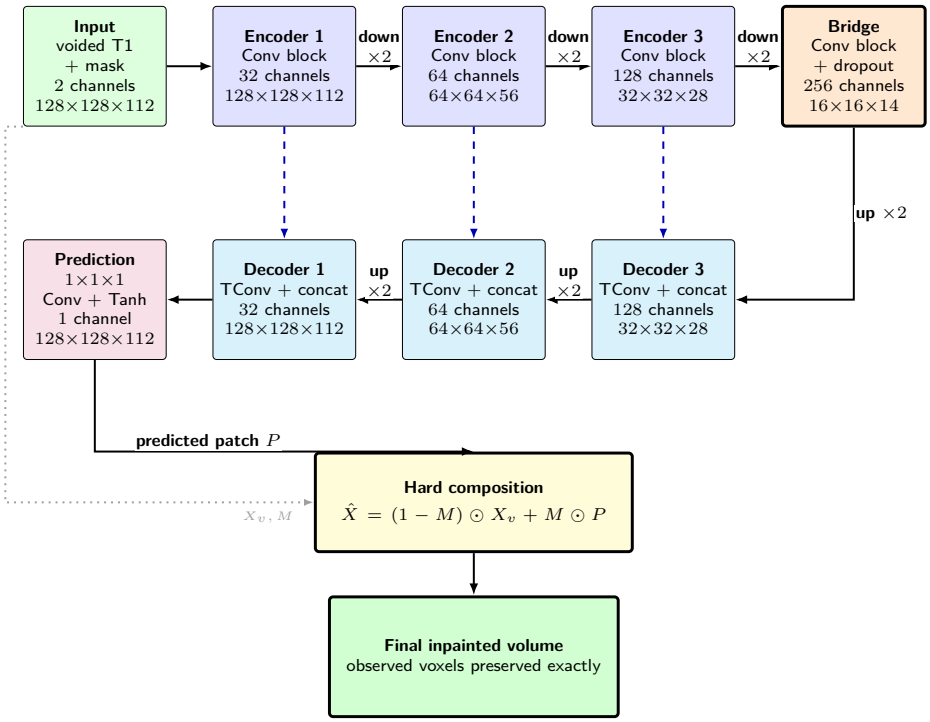

Fig. 1. Architecture of the mask-guided 3D U-Net. A convolutional block contains two reflection-padded $3 \times 3 \times 3$ convolutions, each followed by instance normalisation and PReLU. Encoder blocks are followed by stride-two convolutional downsampling. Decoder blocks use a $2 \times 2 \times 2$ transposed convolution, concatenation with the corresponding encoder feature map, and a convolutional block. Dropout3D with probability 0.2 is applied in the bridge and decoder blocks. The predicted patch P is inserted only inside mask M through hard composition, while observed voxels from X_v are copied unchanged.

0.2 is applied in the deeper path. A final $1 \times 1 \times 1$ convolution and hyperbolic tangent output the normalised prediction. The model has approximately 5.8 million trainable parameters. Instance normalisation stabilises per-volume appearance, whereas PReLU learns the negative activation slope [13,4].

The voxel loss is calculated only within M. Let $N_M = \sum_i M_i$; then

$$\mathcal{L}_1 = \frac{1}{N_M} \sum_i M_i |\hat{X}_i - X_i|, \tag{3}$$

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{N_M} \sum_i M_i (\hat{X}_i - X_i)^2. \tag{4}$$

L1 mitigates isolated intensity outliers, MSE rewards voxel-wise accuracy, and SSIM provides a complementary local structural signal. Local 3D SSIM

uses a 7^3 average-pooling window over the target and its immediate boundary. The complete objective is

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{\text{MSE}} + \lambda_s (1 - \text{SSIM}_{7^3}). \tag{5}$$

The initial optimisation weights are $(1, 0.1, 1)$. A 15-epoch refinement increases the MSE weight to give $(1, 1, 1)$. Because refinement resumes from an already converged checkpoint while increasing the contribution of MSE, the learning rate is reduced from 3×10^{-5} to 5×10^{-6} to avoid abrupt parameter updates. Training uses AdamW [8], weight decay 10^{-5}, mixed precision, batch size one, and gradient clipping at 1.0. The initial phase also uses ReduceLROnPlateau. Experiments use PyTorch [10] and PyTorch Lightning [3].

3.3 Component-wise full-volume inference

Figure 2 summarises the submitted pipeline. We process each connected component of the organiser-provided inpainting mask independently. Overlapping tiles of size $128 \times 128 \times 112$ cover each component with 50% overlap. The network receives a normalised voided tile and its component mask, consistent with the two-channel training interface.

Each tile uses four views: the original and one flip along each spatial axis. Predictions are flipped back and averaged. Overlapping tiles are fused with a separable Gaussian importance map whose standard deviation is one sixth of the crop size and minimum weight is 10^{-3}. If $p_k(i)$ and $w_k(i)$ are the prediction and weight of tile k at voxel i, the fused result is

$$P(i) = \frac{\sum_{k:i \in k} w_k(i) p_k(i)}{\sum_{k:i \in k} w_k(i)}. \tag{6}$$

The implementation verifies mask coverage, clips the normalised output, restores the original scale, and applies Equation 2. The input affine and NIfTI header geometry remain unchanged.

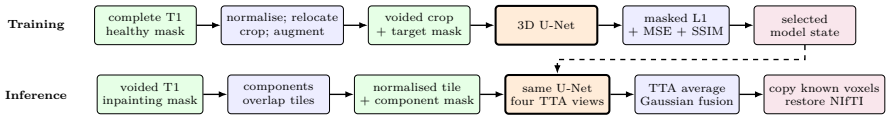

Fig. 2. Final *waterdog* pipeline. The submitted system uses the single model state selected at refinement epoch 8 rather than a model ensemble.

4 Experiments and Results

4.1 Evaluation protocol and model selection

Internal metrics are calculated within the healthy target mask on the fixed 251-case split. MSE and mean absolute error (MAE) use the $[-1, 1]$ representation. Because this interval has width two, PSNR uses a squared data range of four. SSIM uses our local three-dimensional implementation. The organiser-provided evaluator instead clips both the prediction and hidden target using the 0.5th and 99.5th intensity percentiles of the corresponding voided T1 volume, maps them to $[0, 1]$, and computes metrics within the hidden target mask.[1] It also evaluates complete volumes rather than the deterministic crops used for internal model selection. The two protocols therefore assess the same reconstruction task but their absolute values are not directly comparable. Model selection used only the internal split; organiser-reported scores were used for final reporting.

Table 1 compares the final model candidates. The 15-epoch refinement reduced MSE from 0.013244 to 0.012794 (3.40%), decreased MAE by 1.25%, increased PSNR by 0.123 dB, and increased SSIM by 0.00155. Refinement epoch 11 had slightly higher SSIM than epoch 8 but worse compound loss, MAE, MSE, and PSNR. We therefore selected epoch 8 instead of averaging two highly related models. Initial and refinement losses are not directly comparable because their MSE weights differ.

The validation curve was no longer improving consistently near the end of training and instead moved up and down between epochs. This suggests that running many more epochs with the same settings would give limited benefit. The internal and organiser-reported scores also use different normalisation, spatial support, and SSIM implementations. Their numerical gap therefore reflects both model generalisation and protocol differences and should not be interpreted as a direct estimate of either factor alone.

[1] Organiser-provided metric implementation: `https://github.com/BraTS-inpainting/inpainting/blob/main/inpainting/challenge_metrics_2023.py`.

Table 1. Internal validation on 251 held-out cases. Lower MAE/MSE and higher PSNR/SSIM are better.

Model state	Loss	MAE	MSE	PSNR (dB)	SSIM
Initial phase, epoch 117	0.198043	0.067457	0.013244	26.960	0.8707
Refinement, epoch 6	0.208029	0.067111	0.012892	26.987	0.8720
Refinement, epoch 8	**0.207119**	**0.066611**	**0.012794**	**27.083**	0.8723
Refinement, epoch 11	0.207394	0.067031	0.012799	27.052	**0.8724**

Table 2. Statistics recorded for the selected model. Training values are epoch means over randomly augmented crops. Internal and organiser-provided validation use different intensity ranges, spatial support, and metric implementations; their values are therefore not directly comparable. The organiser-provided set contains 219 cases.

Protocol	L1/MAE	MSE	PSNR	SSIM
Training crops	0.065494	0.010862	–	0.8819
Internal validation	0.066611	0.012794	27.083	0.8723
Organiser validation	–	0.008 ± 0.005	22.340 ± 3.963	0.798 ± 0.119

Figure 3 deliberately includes non-ideal cases. Example A follows the target intensities closely with little residual error. Examples B and C recover the broad distribution but smooth some sulci and tissue boundaries, especially at internal anatomical edges. Voxel-wise regression may move towards a local mean when a large void removes structural context. The images support the numerical result but cannot establish clinical plausibility without expert review. Error maps show only the healthy target; voxels outside the red contour are not evaluated.

4.2 Inference-time ablation

We tested three inference configurations on all 251 complete internal volumes using the organiser's full-volume intensity normalisation and masked metric implementation, rather than the crop-level protocol in Table 1. The selected refinement checkpoint (epoch 8), the fixed JSON file recording the 251-case validation split, crop size, overlap, and tile batch size remained unchanged. Only flip TTA and fusion weights varied.

Compared with inference without TTA, four-view averaging reduces MAE by 2.83% and MSE by 3.85%, while increasing PSNR by 0.270 dB and SSIM by 0.0035. This effect is consistent at the case level: MAE improves in 229 of 251 cases, whereas MSE, PSNR, and SSIM improve in 213 cases. Gaussian weighting yields only marginal improvements compared with uniform fusion. For MSE, 164 cases show identical results because many target regions are covered by only one tile, causing normalisation to cancel the fusion weights. Consequently, the primary gains during inference arise from flip-based TTA.

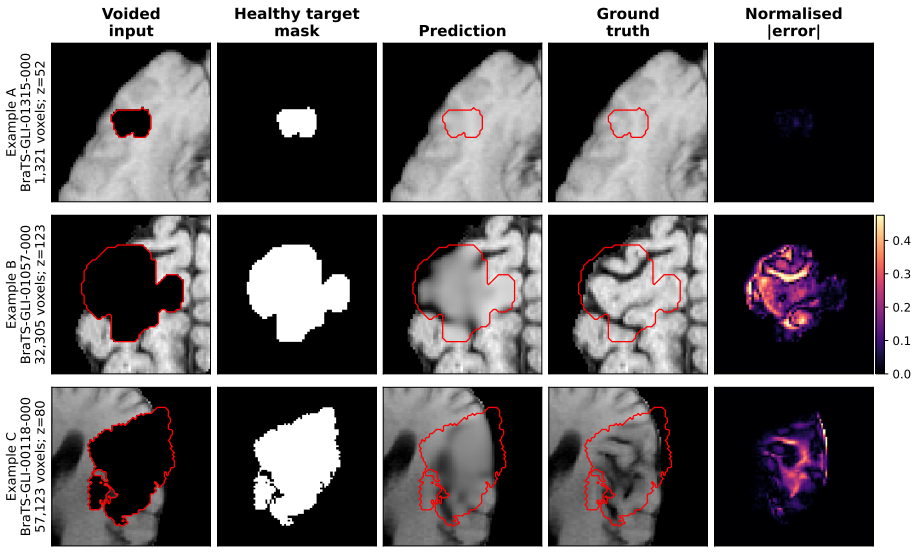

Fig. 3. Qualitative results on three cases from the internal validation split. Red contours indicate the healthy target regions used for evaluation. Normalised absolute-error maps are restricted to those masks and share one colour scale. Organiser-provided validation examples are not included because their reference images are hidden.

Table 3. Inference-time ablation on the fixed 251-case internal validation set using the organiser's full-volume metric implementation.

TTA	Fusion	MAE ↓	MSE ↓	PSNR ↑	SSIM ↑
None	Gaussian	0.055401	0.008451	22.166	0.794180
Four views	Uniform	0.053868	0.008133	22.433	0.797481
Four views	Gaussian	**0.053832**	**0.008126**	**22.436**	**0.797686**

4.3 Deployment and reproducibility

The source code and inference instructions are available at `https://github.com/lingyuguo7/waterdog-brats2026-task4`.

The final code, selected model state, and dependencies were packaged into an offline Linux/AMD64 GPU Docker image. It reads NIfTI files from `/input`, writes flat outputs to `/output`, and requires no network connection. We generated all 219 organiser-provided validation predictions and verified the filename suffix, finite values, $240 \times 240 \times 155$ shape, geometry, and unchanged voxels outside the mask. The main settings needed to reproduce the submitted system are collected in Table 4.

Table 4. Configuration used for the final submission.

Item	Setting
Split	1,000 training / 251 validation; seed 2023
Training input / target	voided T1 + healthy mask / complete healthy region
Validation input / output	voided T1 + inpainting mask / composed T1 volume
Crop / batch	$128 \times 128 \times 112$ / 1
Channels	32, 64, 128; bridge 256
Model / hardware	5.8M parameters / RTX 4090, 24 GB
Regularisation	Dropout3D 0.2; weight decay 10^{-5}
Sampling	2 samples/case; relocation probability 0.9
Refinement	15 epochs; LR 5×10^{-6}; loss weights 1:1:1
Inference	tile batch 2; 50% overlap; Gaussian fusion; four views
Final model	refinement epoch 8; single model state

5 Discussion

Each processing stage addresses a task constraint. The mask marks the unknown region, the masked loss prevents observed background from dominating training, and hard composition copies voxels outside the inpainting mask. Mask relocation adds target locations, while the dilated unhealthy mask stops tumour tissue from becoming a healthy target. These are design reasons, not separate causal findings, because we did not run a controlled training-stage ablation that isolated each component.

The inference ablation gives a clearer result: flip averaging consistently improved the metrics, whereas Gaussian weights added little over uniform fusion. Connected-component coverage avoids unnecessary tiles between distant mask regions, but its individual contribution was not ablated.

The examples show behaviour hidden by summary metrics. Broad local intensities are restored, but fine sulci and tissue boundaries weaken in Examples B and C, possibly because large voids leave insufficient context for voxel-wise regression. SSIM adds a local structural signal but cannot guarantee globally plausible anatomy. Partial or gated convolution [7,15], attention, or coarse-to-fine prediction may sharpen boundaries, although sharpness alone does not prove clinical accuracy.

Several limitations remain. We changed sampling, augmentation, and loss design, but did not complete a training-component ablation or a controlled comparison with a single-channel 3D U-Net. At inference, we tested TTA and fusion, but not overlap or connected-component tiling. One split and one seed also leave partition and optimisation variance unknown. Internal selection used crop-level metrics, whereas organiser scoring used complete volumes. Training conditioned on healthy masks, while organiser-provided inference used inpainting masks; this mismatch may partly explain the online SSIM gap. Finally, only three internal cases were reviewed without expert scores, and validation reference volumes re-

main hidden. Future work should unify the evaluation protocol, test more overlap ratios and patient splits, and add expert assessment.

6 Conclusion

Team *waterdog* developed a practical mask-guided 3D U-Net for healthy brain tissue image inpainting in the BraTS 2026 challenge. The final pipeline incorporates target-centred cropping, mask relocation, an in-mask L1–MSE–SSIM objective, and component-wise overlapping inference. Four-view TTA reduced full-volume MSE by 3.85%, whereas Gaussian-weighted fusion yielded only marginal additional improvement over uniform fusion. The 15-epoch refinement at 5×10^{-6} improved internal crop-level MSE by 3.40%, and the model state from refinement epoch 8 was ultimately selected. Future work requires more complete training-stage ablation and closer alignment between internal and organiser-provided evaluation.

Data use and ethics. This study used de-identified data distributed by the BraTS challenge; no additional patient data were collected, and no clinical decisions were made. All data use adhered to the challenge access conditions.

Acknowledgements. We thank the BraTS 2026 organisers for providing the data, baseline, evaluation tools, and submission infrastructure. We also thank our supervisor, Ali Gooya, for the guidance and feedback provided throughout the project.

References

1. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: 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
2. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: Learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Lecture Notes in Computer Science, vol. 9901, pp. 424–432. Springer (2016). https://doi.org/10.1007/978-3-319-46723-8_49
3. Falcon, W., Team, T.P.L.: Pytorch lightning (2019). https://doi.org/10.5281/zenodo.3828935
4. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1026–1034 (2015). https://doi.org/10.1109/ICCV.2015.123
5. 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**, 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z

6. Kofler, F., Meissen, F., Steinbauer, F., Graf, R., et al.: The brain tumor segmentation (BraTS) challenge 2023: Local synthesis of healthy brain tissue via inpainting. arXiv preprint arXiv:2305.08992 (2023). https://doi.org/10.48550/arXiv.2305.08992
7. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Computer Vision – ECCV 2018. Lecture Notes in Computer Science, vol. 11215, pp. 89–105. Springer (2018). https://doi.org/10.1007/978-3-030-01252-6_6
8. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7
9. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., 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
10. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. vol. 32, pp. 8024–8035 (2019)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28
12. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of Big Data **6**, 60 (2019). https://doi.org/10.1186/s40537-019-0197-0
13. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016). https://doi.org/10.48550/arXiv.1607.08022
14. 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
15. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4471–4480 (2019). https://doi.org/10.1109/ICCV.2019.00457
16. Zhang, J., Weng, Y., Chen, K.: U-net based healthy 3d brain tissue inpainting. arXiv preprint arXiv:2507.18126 (2025). https://doi.org/10.48550/arXiv.2507.18126