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.

A Mask-Gated 3D U-Net with Void-Size Conditioning for Healthy Tissue Inpainting in Brain MRI

Navid Toufan[1(✉)] and Ahmed Ashraf[1]

Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Canada
toufann@myumanitoba.ca

Abstract. Magnetic resonance imaging (MRI) is not limited to just diagnosis, but also can be employed in treatment planning, brain parcellation, segmentation or brain tumor modeling. However, these downstream applications cannot work properly if the tumor is present in the image. The BraTS 2026 local inpainting challenge asks for the synthesis of healthy brain tissue in the region of interest. We propose a 3D UNet-Like model for this task that addresses three properties specific to inpainting tasks. First, the features inside the void have no information, so the skip connection is gated by a valid mask, which is downsampled alongside the feature maps, to prevent the voided region from contributing to the decoder stage. Second, reconstruction quality approximately drops with the size of the void region, so the network is conditioned on the void volume by feature-wise linear modulation, allowing the network to adapt its behavior across cavity scales. Third, provided masks are derived from real tumor geometry, therefore by placing them in random positions they may extend beyond the brain boundary, where the target is background, not brain tissue. A background suppression loss teaches the model the boundary rather than filling background with synthetic tissue. The model is trained with a composite objective function. On the official validation set of 219 cases, our method achieves an SSIM of 0.832, a PSNR of 23.366 dB, and an MSE of 0.006. The code is available at `https://github.com/Navid-Toufan/NT_BraTS_Inpainting`.

Keywords: Brain MRI · Image Inpainting · Healthy Tissue Synthesis · 3D UNet · Mask-Gated Skip Connections · Feature-wise Linear Modulation · BraTS

1 Introduction

Magnetic resonance imaging (MRI) is one of the leading imaging modalities for investigating soft tissues and organs in order to diagnose diseases and lesions along with monitoring therapy response and tumor progression [1]. MRI is regularly used for both detection of brain tumors and also for characterizing the degree of aggressiveness and heterogeneity of disease [2,3,4]. It is worth mentioning that applications of MRI scans are not limited to just diagnosis, but also can

be employed in treatment planning [2], brain parcellation and segmentation [5,6], brain tumor modeling [7,8], and downstream analysis pipelines [9].

However, one of the challenges in MRI scans of patients with a brain tumor is that the lesion interferes with most of the aforementioned applications. It would therefore be clinically useful to develop algorithms for generating estimates of what the tumor-affected area of the brain would look like in the absence of a tumor. Acknowledging the clinical significance of such algorithms, the Brain Tumor Segmentation (BraTS) Local Inpainting Challenge was launched in recent years [9]. By synthesizing plausible healthy brain tissue in the lesion region, it enables normal parcellation and segmentation, supports treatment planning, handles local artifacts and opens avenues for further research on brain–tumor modeling.

Image inpainting is a common image processing task that helps to fill in missing or damaged regions of an image. Various inpainting approaches include traditional methods and deep-learning-based methods. The traditional techniques fall into two main families: diffusion-based methods, which fill the hole by propagating structure from the boundary via PDEs, often from a variational formulation [10,11], and patch-based methods, which copy similar patches from the known image [12]. Hybrid methods combine both by restoring structure and texture separately [13].

With the emergence of deep learning techniques, many new approaches including CNN-based such as encoder–decoder architectures [14], fully convolutional networks [15], and UNet-based models [16] have been proposed. Generative models including generative adversarial networks (GANs) [17,18], variational autoencoders (VAEs) [19] and diffusion models [20,21] are also used in inpainting tasks. Alongside these, transformer-based models are among well-known methods for inpainting tasks [22,23].

2 Prior Work on Tumor Inpainting

For tumor inpainting as well, feed-forward UNet-like models have repeatedly proven to be a strong choice for this task. Zhang et al. [24] proposed a UNet-like model with random masking augmentation, ranking first in the BraTS 2023 inpainting challenge. The same group's follow-up model [25] again secured first place in the 2024 challenge, and their most recent model [26], trained with a combined mean absolute error (MAE) and structural similarity index measure (SSIM) loss, ranked first once more in the 2025 challenge. This indicates that feed-forward U-Net architectures remain a consistently strong and promising approach for inpainting tasks.

GAN-based models have shown their effectiveness in inpainting tasks. Zeineldin and Mathis-Ullrich [27] used a 3D Pix2Pix GAN to inpaint healthy brain tissue in BraTS 2023, combined with an ensemble segmentation model. Ha et al. [28] proposed PSegGAN, a pseudo-segmentation-guided GAN for BraTS 2025 brain tissue inpainting. The framework employs three networks: a generator to reconstruct masked volumes, a discriminator for perceptual realism, and a seg-

mentation network that provides anatomical guidance using SynthSeg pseudo-segmentation maps. More recently, VAEs have mainly appeared as a component within latent diffusion models. Biller et al. [29] adopted a pretrained MAISI VAE to compress 3D T1 volumes into a compact latent space, on which a latent diffusion model conditioned on tissue segmentations and tumor concentrations performed healthy tissue inpainting for the BraTS 2025 challenge.

Diffusion models are also a well known technique for image inpainting. Durrer et al. [30] trained a 2D DDPM on slices where healthy tissue was cropped out, then substituted the diseased slices in the original 3D volume with the inpainted healthy slices, achieving second place in the BraTS 2023 challenge. Dai et al. [31] proposed a conditional DDPM trained exclusively on healthy tissue, paired with a multi-stage DDIM sampling scheme that fuses unmasked context into each denoising stage to improve structural coherence. Ferreira et al. [32] addressed the high computational cost of diffusion-based inpainting by proposing a conditional IDDPM with a hybrid ResNet-Swin-UNet denoising backbone, where the encoder/decoder stages combine ResNet blocks with Swin Transformer blocks, and an accelerated sampling strategy, achieving over 10 times faster sample generation while remaining competitive with standard diffusion baselines.

Transformer-based components have also been incorporated into recent inpainting architectures. Koch et al. [33] trained an unconditional 3D diffusion transformer, built from Hourglass Transformer blocks and efficient attention variants, directly on the full-resolution pixel space, applying it to inpainting without any inpainting-specific training. Sarıtaş et al. [34] proposed Local2Global, a UNet-like architecture in which each of the four encoder stages progressively shifts from convolution to attention, moving from local patterns to full global context, so as to balance anatomical coherence with the computational cost of processing 3D volumes.

3 Methodology

3.1 Overview

In this work, the model needs the T1-weighted (T1n) MRI volume x whose region of interest has been voided and a binary mask m showing the region to be filled. Then the model is trained to predict the voided region with synthesized healthy tissue while leaving the non-voided region unchanged. We address this with a 3D UNet-like architecture with a combination of loss functions. In order to increase training diversity, we generate additional synthesis masks from the provided healthy and unhealthy masks.

3.2 Data Preparation and Mask Augmentation

We use the BraTS-Local-Inpainting dataset [9], which builds upon the BraTS-GLI 2023 dataset [35,3,4]. It consists of 1251 cases for the training set and 219 cases for the validation set. Each case of the training set contains a T1n MRI

volume and its voided version along with unhealthy and healthy masks and a combination of both masks, while the validation set contains only the T1n-voided volume and a binary mask. The MRI volumes are in NIfTI (Neuroimaging Informatics Technology Initiative) format with a shape of $240 \times 240 \times 155$ voxels.

The model receives two inputs: the T1n volume with a healthy region zeroed out, and the binary mask of that region. We void a healthy region rather than the tumor because training requires ground-truth intensities inside the hole, which are only available for healthy tissue. A fixed $208 \times 208 \times 144$ patch is extracted around each voided region by expanding the mask's bounding box symmetrically to the patch size.

Each original T1n is intensity-normalized by clipping negative values to zero and dividing by the maximum voxel value of the volume, mapping it to the range [0,1]. Finally, it is linearly mapped to the [-1,1] range to match the network's Tanh output and provide zero-centered inputs for the sake of gradient stability and optimization.

To expose the model to more diverse samples, we generated 4 more synthesis masks from the provided healthy and unhealthy masks. Every generated mask is flipped, independently for each of the three axes, with probability 0.5, and is then given two rotations whose angles are sampled uniformly from $[0°, 360°)$, one about the (D, W) axes and one about the (W, H) axes. Of the two masks built from the healthy shape, one is resized to a random volume in 180k–320k voxels and the other to 320k–650k; the two built from the tumour shape instead receive a per-axis scale factor drawn independently from $[0.7, 1.4]$. The mask is then put at a random site no closer than five voxels to the tumour.

3.3 Model Architecture

Fig. 1 illustrates our 3D UNet-like model architecture. The voided T1 and the binary inpainting mask patches are stacked as a 2-channel input and passed through four encoder stages (32, 64, 128, and 256 channels), a bottleneck (512 channels), and four reconstruction decoder stages. Each stage is a block of two 3D convolutions with instance normalization and ReLU. Downsampling is done with max pooling and upsampling with transposed convolutions. The final $1\times1\times1$ convolution and Tanh activation produce the predicted volume, which is then composited with the original voided scan so that only the masked voxels are replaced.

Standard convolutions and skip connections apply the same operation uniformly across valid and hole regions alike, allowing voided values to propagate into the network's computations and degrade reconstruction quality. To address this, mask-aware mechanisms have been proposed, such as gated convolutions [17], which learn a per-channel soft gate, multiplied onto the feature map at every convolutional layer throughout the network.

We instead introduce a gated valid-mask skip connection from each encoder level to the corresponding decoder level, to prevent the participation of the voided region values in the reconstruction process, which may otherwise degrade

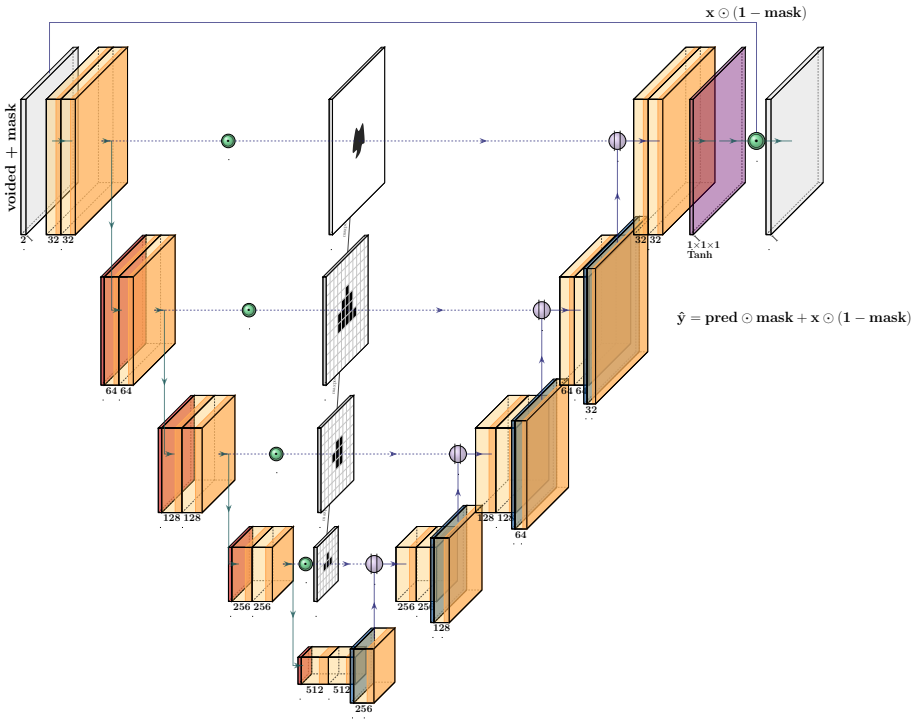

Fig. 1. The 3D UNet-like model architecture used in this work.

the quality of the reconstruction. At the first level, we multiply the encoder feature map by the valid-mask $(1 - m)$ before it is concatenated into the decoder. At each subsequent level, the valid-mask is first max-pooled from the previous level to match the current resolution, then multiplied by the encoder feature map before concatenation, keeping the valid-mask and feature resolutions consistent throughout the network. We downsample the valid-mask by max pooling: a coarse voxel is marked void only if every child voxel is void, so the cavity gets smaller as we go deeper and the gate has its strongest effect at the first levels. Formally, let v_l denote the valid mask at encoder level l, with $v_0 = 1 - m$ at full resolution and $v_{l+1} = \text{maxpool}_2(v_l)$. Each skip connection is gated as $\text{skip}_l \leftarrow \text{skip}_l \odot v_l$, where $\odot$ is the element-wise product.

3.4 Loss Functions

For this problem, the loss function consists of five components: MAE, SSIM, the gradient difference loss (GDL), the perceptual loss and the background loss.

Let Y denote the ground-truth volume, $\hat{Y}$ the predicted volume, and m the binary inpainting mask ($m_i = 1$ for hole voxels). Let $\mathcal{M} = \{i : m_i = 1\}$ be the

set of masked voxels. The MAE loss is considered a voxel-wise reconstruction loss to measure the average absolute intensity between the predicted and ground truth voxels:

$$\mathcal{L}_{MAE} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left| \hat{Y}_i - Y_i \right|. \tag{1}$$

However, the MAE loss has no notion of neighboring voxels, structure and texture or in other words, MAE works well at low frequencies like overall tone, brightness and color while missing high frequencies resembling edges, sharpness and fine details which can result in blurriness. In order to mitigate some of these issues and blurriness, SSIM loss [36] is used to measure the structural similarity capturing the local contrast, texture and structure information. Suppose μ, σ^2, and $\sigma_{\hat{Y}Y}$ denote local mean, variance, and covariance computed via a box window:

$$\mathrm{SSIM}_i = \frac{(2\mu_{\hat{Y}}\mu_Y + C_1)(2\sigma_{\hat{Y}Y} + C_2)}{(\mu_{\hat{Y}}^2 + \mu_Y^2 + C_1)(\sigma_{\hat{Y}}^2 + \sigma_Y^2 + C_2)}, \tag{2}$$

$$\mathcal{L}_{SSIM} = 1 - \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \mathrm{SSIM}_i \,. \tag{3}$$

Another well-known strategy to mitigate blurriness is to use the GDL loss [37] to penalize the difference between the gradient of the prediction and the ground truth, rather than the intensity difference, which can result in sharper predictions in terms of edges and local structure. For each axis $d \in \{D, W, H\}$, let $\nabla_d Y_i = Y_{i+e_d} - Y_i$ denote the finite-difference gradient, and let $\mathcal{M}_d \subseteq \mathcal{M}$ be the mask voxels along that axis:

$$\mathcal{L}_{GDL} = \frac{1}{3} \sum_{d \in \{D,W,H\}} \frac{1}{|\mathcal{M}_d|} \sum_{i \in \mathcal{M}_d} \left| \nabla_d \hat{Y}_i - \nabla_d Y_i \right|. \tag{4}$$

Perceptual loss compares the scans in a feature space that tracks texture and structure, rather than voxel-wise intensity. Because VGG16 is two-dimensional, we apply it to at most 16 axial slices per batch that intersect the mask. Each one-channel slice is scaled to $[0, 1]$, copied to three channels, and normalised with the ImageNet mean and standard deviation, then passed through a frozen ImageNet-pretrained VGG16 [38]. We take activations ϕ_l from three layers, $\mathcal{F} = \{\mathrm{relu1_2}, \mathrm{relu2_2}, \mathrm{relu3_3}\}$ (indices 3, 8 and 15 of `vgg16.features`), which are the last ReLU of each of the first three convolutional blocks; deeper layers encode semantics rather than the local texture needed here:

$$\mathcal{L}_{perc} = \frac{1}{|\mathcal{F}|} \sum_{l \in \mathcal{F}} \left\| \phi_l(\hat{Y}_s) - \phi_l(Y_s) \right\|_1, \tag{5}$$

where $\hat{Y}_s$ and Y_s are the sampled slices. The feature maps are left unmasked, because compositing already makes the prediction and the ground truth the same outside the cavity.

The BraTS inpainting challenge protocol allows healthy masks to overlap with out-of-brain background by up to 25% [9]. This can cause the model to

learn to fill background regions with synthetic tissue, an undesired behavior. To address this, we introduce a background loss that penalizes non-background predictions wherever the ground truth is background. We define $\mathcal{B} = \{i \in \mathcal{M} : Y_i \leq v_{bg} + \tau\}$ as the mask voxels whose ground truth is background ($v_{bg} = -1$, $\tau = 10^{-3}$):

$$\mathcal{L}_{BG} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left| \hat{Y}_i - v_{bg} \right|. \tag{6}$$

The total loss combines all five components:

$$\mathcal{L} = \lambda_{MAE}\mathcal{L}_{MAE} + \lambda_{GDL}\mathcal{L}_{GDL} + \lambda_{BG}\mathcal{L}_{BG} + \lambda_{SSIM}\mathcal{L}_{SSIM} + \lambda_{perc}\mathcal{L}_{perc}. \tag{7}$$

3.5 Void-Size Conditioning

By analyzing our predictions, we noticed that the reconstruction quality highly depends on the size of the region needed to be filled. By grouping the voided regions into small, medium, and large based on their size, we noticed that SSIM drops significantly from small to medium and large voided areas. This indicates that voxels in the interior of a large cavity are far away from any surrounding healthy tissue, so the model has progressively less context to reconstruct the region. To address this issue, we introduce a mechanism to condition the model on the size of the voided area rather than training the model with different mask sizes. The number of voxels inside the cavity is passed through a Fourier feature embedding and a small multilayer perceptron as a learned description of the mask size. The resulting vector modulates the feature maps at the bottleneck and the decoder stages by a per-channel scale and shift using Feature-wise Linear Modulation (FiLM) [39]. The modulation is initialized to zero like an unconditional model, and the network only learns to adapt its behavior to the size of the voided area.

4 Results

4.1 Evaluation Metrics

To evaluate the performance of the model, there are three metrics defined by the BraTS Challenge: structural similarity index measure (SSIM), peak signal to noise ratio (PSNR), and mean squared error (MSE). The evaluation is restricted to the voided regions of healthy masks due to the availability of ground truth only within healthy tissue. Each of these metrics focuses on different aspects of the reconstruction quality. Differently from SSIM, PSNR and MSE are more sensitive to intensity differences.

4.2 Experimental Setup

The model was trained using the Adam optimizer with an initial learning rate of 1×10^{-4}, decayed following a cosine schedule to 1×10^{-6} over 100 epochs,

with a batch size of 2 and mixed precision training on two NVIDIA RTX A4000 GPUs. Volumes were scaled by their maximum intensity to $[0, 1]$ and then to $[-1, 1]$, and cropped to $208 \times 208 \times 144$ centered on the inpainting mask, a size chosen so that every validation cavity is fully contained without truncation. The loss weights were set to $\lambda_{MAE} = 1.0$, $\lambda_{GDL} = 0.1$, $\lambda_{BG} = 0.3$, $\lambda_{SSIM} = 0.25$ and $\lambda_{perc} = 0.1$.

Masks are regenerated at every epoch rather than fixed in advance, so over the training schedule the model is exposed to numerous distinct generated masks.

4.3 Quantitative Results on the Local Validation Split

We hold out 10% of the training set (125 cases) as a local validation split. Table 1 compares our model with the two official challenge baselines, a 3D Pix2Pix network and a 3D autoencoder, using the released pretrained weights on the same cases. Our model improves SSIM by 0.050 over Pix2Pix3D and by 0.052 over the autoencoder.

Table 1. Comparison with official baselines on the local validation split (mean ± std).

Model	SSIM ↑	PSNR (dB) ↑	MSE ↓
Pix2Pix3D baseline [9]	0.820 ± 0.122	19.72 ± 2.39	0.011 ± 0.007
AutoEncoder baseline [9]	0.818 ± 0.136	19.59 ± 2.50	0.011 ± 0.006
Ours (30 epochs)	0.852 ± 0.102	21.50 ± 2.72	0.007 ± 0.006
Ours (100 epochs)	$\mathbf{0.870 \pm 0.094}$	$\mathbf{22.38 \pm 2.87}$	$\mathbf{0.006 \pm 0.005}$

4.4 Quantitative Results on the Official Validation Set

Table 2 reports our model's performance on the BraTS 2026 Local Inpainting challenge validation set, computed by the challenge online platform over the 219 validation cases. Our model achieved a mean SSIM of 0.832 ± 0.098, a mean PSNR of 23.366 ± 4.082 and a mean MSE of 0.006 ± 0.005.

Table 2. Performance on the BraTS Local Inpainting online validation set (mean ± std over 219 cases).

Validation on	Cases	SSIM ↑	PSNR (dB) ↑	MSE ↓
Validation Dataset	219	0.832 ± 0.098	23.366 ± 4.082	0.006 ± 0.005

By grouping the voided regions into small, medium and large based on their size, we can see in Table 3 that performance consistently degrades as the voided region size grows, confirming that voxels inside mask region far from the cavity boundary are harder to reconstruct.

Table 3. Performance broken down by void size, based on tertiles of the cavity voxel count (mean ± std).

Void size	Cases	Range (voxels)	SSIM ↑	PSNR (dB) ↑	MSE ↓
Small	73	51,517–193,427	0.877 ± 0.103	24.75 ± 4.91	0.005 ± 0.004
Medium	73	193,953–299,683	0.824 ± 0.093	23.43 ± 3.81	0.006 ± 0.004
Large	73	300,696–644,098	0.795 ± 0.078	21.92 ± 2.67	0.008 ± 0.005
All	219	51,517–644,098	0.832 ± 0.098	23.36 ± 4.08	0.006 ± 0.005

4.5 Ablation Study

Eight variants were trained with the same 30-epoch cosine schedule, crop $208 \times 208 \times 144$, and batch size 2, on the 125-case local validation split. Evaluation uses the same cases for every variant, and differences are given as paired per-case changes (Table 4). These numbers isolate each component in one particular order, under a shorter schedule, and with masks generated on the fly, so each run sees a different random draw. Every configuration changes the objective and, for the architectural terms, the search space itself, so a solution trained with only a subset of components is not the full model with those components taken out. The values should therefore be read as the effect of adding a component at a given point, not as its contribution to the final model.

Table 4. Component-wise ablation on the local validation split (125 cases). Each row adds one component to the row above. Δ is the paired change in SSIM; n is the number of cases (out of 125) with a higher SSIM than under the configuration of the row above.

Configuration	Epochs	SSIM ↑	PSNR ↑	MSE ↓	Δ SSIM	n
3D U-Net, masked MAE	30	0.806	19.71	0.011	—	—
+ SSIM	30	0.806	19.82	0.011	+0.001	76
+ mask augmentation	30	0.856	21.61	**0.007**	**+0.049**	**117**
+ GDL	30	**0.859**	**21.68**	**0.007**	+0.004	78
+ background	30	0.851	21.38	0.008	−0.008	37
+ perceptual	30	0.850	21.27	0.008	−0.001	60
+ mask gating	30	0.844	21.24	0.008	−0.005	40
+ void-size cond.	30	0.852	21.50	**0.007**	+0.008	95

Mask augmentation. Mask augmentation is by far the most important component, contributing +0.049 SSIM and improving 117 of 125 cases. Generating additional masks at a range of sizes and positions exposes the model to far more varied inpainting geometries than the provided masks alone.

Void-size conditioning. Void size conditioning improves SSIM by 0.008, and the gain increases with the size of the region being filled, from +0.001 for the smallest quarter of cases to +0.013 for the largest. The mechanism therefore helps most

where the model has least surrounding tissue to work from, which is what it was designed to do.

Loss terms. Adding the GDL term yields a modest improvement of +0.004, whereas neither the SSIM nor perceptual terms produce any measurable benefit, and the perceptual term also raises training time by 10%. As explained in Sect. 5, these terms were meant to push the model away from overly smooth predictions; the ablation confirms they only partially succeed, since the metric being optimized still favors smoothness.

Background suppression. Introducing $\mathcal{L}_{BG}$ trades 0.008 SSIM for a drop in tissue synthesised outside the brain, from 6.4% to 2.1% of extracranial mask voxels, in all 70 cases with substantial overlap. The SSIM cost follows from the network not being able to pin down the boundary tightly enough to penalise only the exterior, so the penalty leaks inward and darkens genuine tissue within three voxels of the edge to 88% of its true intensity. Sweeping λ_{BG} over $\{0, 0.1, 0.3\}$ gives SSIM of 0.859, 0.854, 0.851 and extracranial synthesis of 6.4%, 3.4%, 2.1%. We keep $\lambda_{BG} = 0.3$ and note that intensity-based metrics penalise behaviour that is anatomically preferable.

Mask gating. Under this training schedule, mask gating did not improve SSIM (-0.005, improving only 40 of 125 cases). Gating is applied only along the skip path by design, since the encoder must preserve the cavity's shape so the decoder can tell which region it is reconstructing. Its limited benefit may stem from the lack of renormalisation: the feature map is multiplied by the valid mask without rescaling for the reduced support, unlike partial convolutions.

Computational cost. The loss terms add nothing at inference; they only affect training time. The architectural additions are also cheap: the network has 22,576,193 parameters without void-size conditioning and 22,852,865 with it. Mask gating adds no parameters, since it is an element-wise product with the downsampled valid mask. Inference takes 293 ms per case at 5.7 GB without gating, 302 ms at 6.5 GB with gating, and 313 ms at 6.5 GB with both, measured on a single NVIDIA RTX A4000 at crop $208 \times 208 \times 144$, batch size 1, single precision. Reconstruction is therefore one forward pass, unlike the many network evaluations needed for iterative diffusion sampling.

4.6 Qualitative Results

Fig. 2 shows representative validation cases ranked by SSIM: the best, median, and worst case, each shown as four axial slices with the inpainted region outlined in green. In the best case, the reconstructed tissue is nearly indistinguishable from the surrounding anatomy. The median case still recovers plausible tissue structure, though with some loss of fine detail near the cavity boundary. The worst case corresponds to a large cavity in the validation set, where the reconstruction remains anatomically plausible but is visibly blurrier, consistent with the loss of context noted in Sect. 3.5 for large voids. Fig. 3 illustrates the effect of the background-suppression loss, with green outlining the inpainting mask and red the portion extending beyond the brain boundary, where the ground truth is background. Without $\mathcal{L}_{BG}$ the model synthesises tissue-like intensity in the

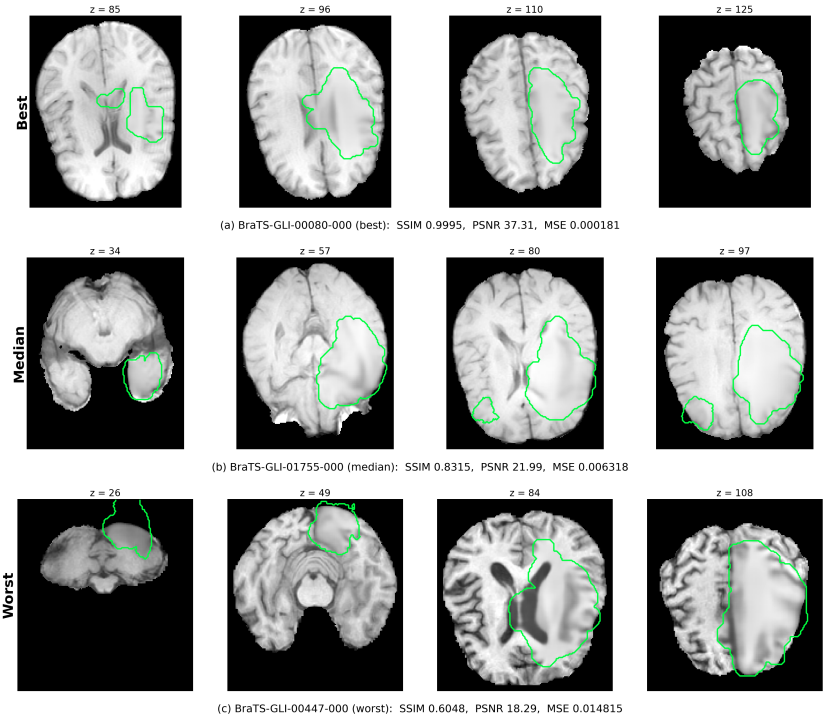

Fig. 2. Qualitative comparison of the best, median, and worst validation cases ranked by SSIM. The voided region is outlined in green.

extracranial region; adding the term suppresses it. Across the local validation split, tissue synthesised outside the brain falls from 6.4% to 2.1% of out-of-brain mask voxels, reduced in all 70 cases with substantial overlap.

5 Discussion

In this work we showed that a feed-forward UNet-like model trained with a composite loss function can still be a competitive choice for the local inpainting task, and it needs only one forward pass to fill a voided region while diffusion-based approaches need many iterative sampling steps. Our gated valid-mask skip connection uses the same idea as the mask-aware methods in Sect. 3.3, with one difference: the feature map is multiplied by the valid-mask and is not renormalised by how many voxels remain valid.

The most noticeable limitation of our reconstructions is blurriness. As we can see in Fig. 2, the inpainted regions are smoother than the surrounding tissue and the fine details visible outside the mask are missing inside it. This is not specific to our model, as the same effect has been reported by previous submissions and

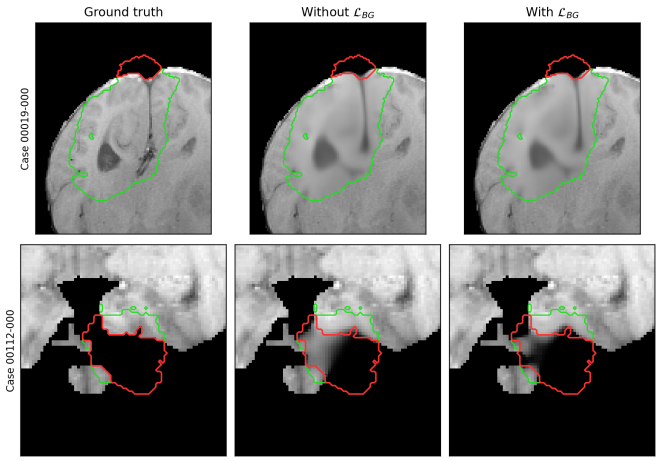

Fig. 3. Effect of the background-suppression loss on out-of-brain tissue synthesis.

by the challenge organizers [26,9]. In our opinion the reason goes beyond the choice of the loss function. Since there is no ground truth inside the cavity, the real texture is only one of many plausible options, and a prediction close to the average of those options is the one that stays closest to whichever texture happens to be correct. This also explains why the problem becomes worse for larger cavities, as shown in Table 3, because the voxels in the middle of a large void are far from any real tissue. Moreover, the metrics used for ranking measure how close the prediction is to the ground truth, and a smooth prediction is exactly what scores best under this kind of measure. It has already been observed that fidelity metrics correlate weakly with human judgment for generation tasks and can mislead research toward overly smooth results [23]. Our SSIM, GDL and perceptual terms were added to push the model away from the smooth solution, but they can only reduce this behavior and not remove it, because the metric being optimized still rewards it. Our model outputs a single reconstruction and gives no measure of how uncertain it is, even though many textures are plausible inside the cavity.

For future work we plan to explore generative formulations that sample one plausible texture instead of averaging over all of them, which we believe is the most direct way to address the blurriness discussed above. Overall, our model achieved an SSIM of 0.832, a PSNR of 23.366 dB and an MSE of 0.006 on the 219 cases of the official validation set, which shows that a single-pass UNet-like model remains a practical approach for healthy tissue inpainting in brain MRI.

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

References

1. Bruno, F., Arrigoni, F., Mariani, S., Splendiani, A., Di Cesare, E., Masciocchi, C., Barile, A.: Advanced magnetic resonance imaging (MRI) of soft tissue tumors: Techniques and applications. La Radiologia Medica **124**(4), 243–252 (2019). https://doi.org/10.1007/s11547-019-01035-7
2. Villanueva-Meyer, J.E., Mabray, M.C., Cha, S.: Current clinical brain tumor imaging. Neurosurgery **81**(3), 397–415 (2017). https://doi.org/10.1093/neuros/nyx103
3. Menze, B.H., Jakab, A., Bauer, S., 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
4. Bakas, S., Akbari, H., Sotiras, A., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data **4**(1), 170117 (2017). https://doi.org/10.1038/sdata.2017.117
5. Fischl, B.: Freesurfer. NeuroImage **62**(2), 774–781 (2012). https://doi.org/10.1016/j.neuroimage.2012.01.021
6. Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M.: FSL. NeuroImage **62**(2), 782–790 (2012). https://doi.org/10.1016/j.neuroimage.2011.09.015
7. Ezhov, I., Mot, T., Shit, S., Lipkova, J., Paetzold, J.C., Kofler, F., et al.: Geometry-aware neural solver for fast bayesian calibration of brain tumor models. IEEE Transactions on Medical Imaging **41**(5), 1269–1278 (2021)
8. Ezhov, I., Scibilia, K., Franitza, K., Steinbauer, F., Shit, S., Zimmer, L., et al.: Learn-morph-infer: A new way of solving the inverse problem for brain tumor modeling. Medical Image Analysis **83**, 102672 (2023)
9. Kofler, F., Meissen, F., Steinbauer, F., et al.: The brain tumor segmentation (BraTS) challenge: Local synthesis of healthy brain tissue via inpainting (2024). https://doi.org/10.48550/arXiv.2305.08992
10. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of SIGGRAPH (2000)
11. Chan, T.F., Shen, J.: Mathematical models for digital inpainting. Journal of Visual Communication and Image Representation **12**(4), 417–424 (2001)
12. Criminisi, A., Pérez, P., Toyama, K.: Region filling and completion by exemplars. IEEE Transactions on Image Processing **13**(9), 1200–1212 (2004). https://doi.org/10.1109/TIP.2004.833105
13. Bertalmio, M., Vese, L., Sapiro, G., Osher, S.: Simultaneous structure and texture image inpainting. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. 707–712 (2003)
14. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2536–2544 (2016)
15. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6721–6729 (2017)
16. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2149–2159 (2022)

17. 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 (ICCV). pp. 4471–4480 (2019). https://doi.org/10.1109/ICCV.2019.00457
18. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. In: International Conference on Learning Representations (ICLR) (2021)
19. Zheng, C., Cham, T.J., Cai, J.: Pluralistic image completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1438–1447 (2019)
20. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11461–11471 (2022)
21. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022)
22. Li, W., Lin, Z., Zhou, K., Qi, L., Wang, Y., Jia, J.: MAT: Mask-aware transformer for large hole image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10758–10768 (2022). https://doi.org/10.1109/CVPR52688.2022.01049
23. Wang, Q., He, S., Su, M., Zhao, F.: Image inpainting methods: A review of deep learning approaches. Symmetry **18**(1), 94 (2026). https://doi.org/10.3390/sym18010094
24. 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 (MICCAI Challenges BraTS 2023 and CrossMoDA 2023). Lecture Notes in Computer Science, vol. 14669, pp. 233–240. Springer (2024). https://doi.org/10.1007/978-3-031-76163-8_21
25. Zhang, J., Weng, Y., Chen, K.: U-Net based healthy 3D brain tissue inpainting (2025). https://doi.org/10.48550/arXiv.2507.18126
26. Zhang, J., Weng, Y., Chen, K.: Robust 3D brain MRI inpainting with random masking augmentation. In: MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025. Lecture Notes in Computer Science, vol. 16377, pp. 102–109. Springer (2026). https://doi.org/10.1007/978-3-032-16370-7_9
27. Zeineldin, R.A., Mathis-Ullrich, F.: Ensemble learning and 3D Pix2Pix for comprehensive brain tumor analysis in multimodal MRI. In: Brain Tumor Segmentation, and Cross-Modality Domain Adaptation for Medical Image Segmentation (MICCAI Challenges BraTS 2023 and CrossMoDA 2023). Lecture Notes in Computer Science, vol. 14669, pp. 24–34. Springer (2024). https://doi.org/10.1007/978-3-031-76163-8_3
28. Ha, J., Park, J.S., Oh, J., Crandall, D.: PSegGAN: Pseudo-segmentation-guided GANs for brain tissue inpainting. In: MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025. Lecture Notes in Computer Science, vol. 16377, pp. 168–180. Springer (2026). https://doi.org/10.1007/978-3-032-16370-7_15
29. Biller, V., Zimmer, L., Erdur, C., Nagar, S., Rückert, D., Bubeck, N., Weidner, J.: A biophysically-conditioned generative framework for 3D brain tumor MRI synthesis. In: MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025. Lecture Notes in Computer Science, vol. 16377, pp. 91–101. Springer (2026). https://doi.org/10.1007/978-3-032-16370-7_8

30. Durrer, A., Cattin, P.C., Wolleb, J.: Denoising diffusion models for inpainting of healthy brain tissue. In: Brain Tumor Segmentation, and Cross-Modality Domain Adaptation for Medical Image Segmentation (MICCAI Challenges BraTS 2023 and CrossMoDA 2023). Lecture Notes in Computer Science, vol. 14669, pp. 35–45. Springer (2024). https://doi.org/10.1007/978-3-031-76163-8_4
31. Dai, Y., Bi, Y., Navab, N., Jiang, Z.: Context-aware healthy brain inpainting: A multi-stage DDIM approach for the BraTS 2025 challenge. In: MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025. Lecture Notes in Computer Science, vol. 16377, pp. 158–167. Springer (2026). https://doi.org/10.1007/978-3-032-16370-7_14
32. Ferreira, A., Luijten, G., Hinrichs-Puladi, B., Kleesiek, J., Alves, V., Egger, J.: Achieving over 10x faster sample generation with conditional denoising diffusion. In: MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025. Lecture Notes in Computer Science, vol. 16377, pp. 123–132. Springer (2026). https://doi.org/10.1007/978-3-032-16370-7_11
33. Koch, A., Aydin, O.U., Hilbert, A., Frey, D.: Local brain tumour inpainting using diffusion transformers. In: MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025. Lecture Notes in Computer Science, vol. 16377, pp. 133–145. Springer (2026). https://doi.org/10.1007/978-3-032-16370-7_12
34. Sarıtaş, E., Öksüz, İ., Ekenel, H.K.: Local2global: UNet with hierarchical attention mechanisms for improved MR image inpainting. In: MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025. Lecture Notes in Computer Science, vol. 16377, pp. 110–122. Springer (2026). https://doi.org/10.1007/978-3-032-16370-7_10
35. Baid, U., Ghodasara, S., Mohan, S., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification (2021). https://doi.org/10.48550/arXiv.2107.02314
36. 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
37. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. In: International Conference on Learning Representations (ICLR) (2016). https://doi.org/10.48550/arXiv.1511.05440
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015). https://doi.org/10.48550/arXiv.1409.1556
39. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32, pp. 3942–3951 (2018). https://doi.org/10.1609/aaai.v32i1.11671