

RESEARCH ARTICLE

Diffusional magnetic resonance imaging anonymizing with variational autoencoder

Yunheng Shen¹ | Ling Zheng² | Ruohan Liu¹ | Haoran Feng³ | Hairong Lv¹

¹Department of Automation, Tsinghua University, Beijing, China

²Computer Science Department, Fujian Provincial Maternity and Children's Hospital, Fuzhou, Fujian, China

³Shenzhen International Graduate School, Tsinghua University, Shenzhen, Guangdong, China

Correspondence

Hairong Lv.

Email: lvhairong@tsinghua.edu.cn

Funding information

The National Key Research and Development Program of China, Grant/Award Number: 2022YFF1202400

Abstract

Anonymization is a crucial de-identification technique that protects data privacy while ensuring its utility for model building. Current generative models such as generative adversarial networks and variational autoencoders (VAEs) have been applied to medical image anonymization but mainly focus on general image features, lacking specificity in regions of interest such as lesions. This study proposes a novel framework for brain magnetic resonance imaging anonymization, enabling the handling of lesion region prediction while preserving patient privacy. The framework consists of three stages: pre-training VAEs to represent lesion and non-lesion regions in latent space; fine-tuning these latent representations using a diffusion model conditioned on spatial and temporal features; and generating medical image substitutions through joint decoding of lesion and non-lesion latent representations. The comparative investigation has highlighted the benefits of our proposed methods, achieving a promising privacy–utility balance. In a small number of real sample scenarios, using synthetic samples with an 86% anonymity rate still enhanced the downstream segmentation task by 4.60% and the classification task by 8.75%. Our proposed framework offers significant improvements over existing methods in preserving privacy and maintaining data utility for lesion prediction tasks, which holds potential implications for enhanced privacy practices in medical imaging.

KEYWORDS

anonymization, data privacy, diffusion model, medical image processing, variational auto-encoder

1 | INTRODUCTION

The anonymization of medical images is the process of removing or altering personal identifiers in the images so that the data subject is no longer identifiable [1]. The use of medical images can lead to data breaches, impacting individual privacy, a problem that worsens when data authority is shared among institutions. Consequently, effective anonymization techniques for

medical images are crucial to protect patient privacy while maintaining the clinical utility of images. In this work, we focus on anonymizing the non-lesion anatomical background while preserving lesion characteristics required for downstream tasks.

Regulations, such as the General Data Protection Regulation (GDPR) in Europe [2] and the Health Insurance Portability and Accountability Act (HIPAA) [3] impose additional constraints on data sharing. These

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Quantitative Biology* published by John Wiley & Sons Australia, Ltd on behalf of Higher Education Press.

restrictions hinder the completeness of data in data-driven fields, thereby affecting the progression of model learning. This also encourages a reliable anonymization for medical images.

Several studies have been presented in the literature to tackle the problem of anonymizing medical images using generative models, including generative adversarial networks (GANs) [4], and variational auto-encoders (VAEs) [5]. Fezai et al. [4] proposed a GAN aimed at anonymizing magnetic resonance imaging (MRI) scans of the brain, utilizing data from the Alzheimer's disease neuroimaging initiative (ADNI) [6] and the multimodal brain tumor image segmentation benchmark (BRATS) datasets [7]. Nevertheless, GANs tend to be biased toward dense MRI scans, resulting in generated MRI scans that are not practical for use. Fezai and colleagues [4] presented the denoising VAE calibrated for different MRI equipment, utilizing a convolutional neural network (CNN) [8] to initially identify the equipment brands, which then guides the VAE in generating anonymized brain MRI images. This method is particularly effective for scan-level classification tasks as it maintains a focus on reconstructing entire scans, preserving anatomical structures and features similar to the original images. In a similar study, Li et al. [5] developed an AE-GAN network, enhancing identity removal loss while reducing utility-preserving loss, aimed at anonymizing the chest X-ray images from MIMIC-CXR [9]. However, these methods cannot be applied to lower-level tasks, such as lesion classification.

To address the aforementioned limitations in the context of lesion-centric utility, this paper proposes a lesion-preserving background-anonymizing framework for brain MRI, which includes three stages: (1) Using VAE to pre-train latent representations for the non-lesion image background and lesion regions, respectively; (2) diffusional modal is used to regenerate latent representations of images, enhancing anonymization diversity; and (3) combining latent representations of non-lesion background and lesion regions together for anonymizing images using a pre-trained VAE decoder, thereby varying the image background while preserving features of lesion regions for lesion classification tasks. Additionally, we also designed an evaluation criterion to measure the degree of MRI images being anonymized.

The rest of this paper is organized as follows. Section 2 conducts a variety of comparative and qualitative experiments to assess the proposed anonymization methodology using the BraTS2020 dataset [10]. Section 3 wraps up the paper and explores future outlooks. Section 4.1 introduces the situation of the dataset we used detailed. Section 4.2 provides preliminary knowledge about VAE and diffusion. Section 4.3 reviews related work on medical image anonymization. Section 4.4 introduces our proposed method framework and implementation based on diffusional VAE for medical

image anonymization. Section 4.5 proposes a new metric for evaluating the effectiveness of MRI anonymization.

2 | RESULTS

Experimental results confirm the robustness of the proposed diffusion-based method in anonymizing brain MRI scans. Evaluating the pre-trained VAEs on 20,000 samples demonstrated a high structural similarity index (SSIM) of 98.2% or above, alongside high peak signal-to-noise ratios (PSNR) and low mean squared errors (MSE). In terms of downstream utility, the anonymized images were evaluated on segmentation and classification tasks against real images. For binary tumor segmentation, mixing real and anonymized images improved the mean Intersection-over-Union (mIoU) from 61.4% to 66.00%. In the lesion classification task, anonymized images achieved a comparable accuracy of 95.75% relative to the 99.25% attained with real images. Furthermore, supplementing scarce real datasets with synthetic data further increased accuracy by 8.75%. Visual analysis indicates that while synthetic images differ from originals more than VAE reconstructions, they successfully preserve lesion structures while varying non-lesion regions. Finally, a PSNR-based criterion was used to quantify anonymization success. Our proposed method achieves 86% anonymization without sacrificing the utility of the generated images.

2.1 | Experimental configurations

In traditional VAE structures, the latent representation is usually a one-dimensional vector corresponding to the image level. However, to align our approach with the latent diffusion model, we maintain the latent representation at the pixel level. Specifically, the encoder generates a latent tensor of dimensions $[32 \times 30 \times 30]$. This tensor is divisionally allocated along the first dimension to store both the mean and variance of the Gaussian distribution being modeled. The detailed specifications of each module in the proposed VAE can be found in Tables 1 and 2. Across all modules, leakyrelu is used as the activation function, batch normalization is applied for normalization, and the self-attention unit utilizes a four-head multi-head self-attention configuration.

For implementing the diffusion model, we designate the total number of time steps T as 1000 and use the U-Net as the network backbone. The U-Net architecture is mainly completed by 3 downsampling blocks, a bottom convolutional block, and 3 upsampling blocks. Each downsampling block containing a bottleneck structure composed of two convolutional layers and a

TABLE 1 Composition of variational auto-encoder modules.

Module type	Quantity	Channel changes	Stride
Downsampling module	3	1–4–8–16	2
Bottleneck convolution module	1	16–32	1
Upsampling module	3	16–8–4–1	2

TABLE 2 Parameters of variational auto-encoder modules.

Module type	Conv layer	Transpose conv layer	Residual connection	Activation	Normalization	Attention block
Downsampling module	2	N	Y	2	2	N
Bottleneck convolution module	1	N	N	N	N	N
Upsampling module	1	1	Y	2	2	N/Y ^a

^aFor the two complete VAEs used for feature extraction, the upsampling modules do not contain attention blocks; for the independently trained decoder, the upsampling modules include attention blocks.

downsampling convolutional layer with a stride of 2. After each downsampling, there is a self-attention module for learning spatial correlation information. The bottom convolutional block is a bottleneck structure composed of three convolutional layers. The structure of the upsampling module is similar to the mirror image of the downsampling module, but its input includes the same-sized intermediate output from the downsampling process through skip connections. The training process of the above model was conducted on Nvidia-V100, using the Adam optimizer with a learning rate of 1e-4.

2.2 | Assessment of pre-trained VAE

In this work, VAE has been used to encode lesion and non-lesion regions into latent representations, respectively, thereby being jointly decoded to obtain recovered images. Its robustness has a significant impact on the performance of image anonymization. In this section, a visual comparison and the metrics of image generation, including SSIM, PSNR, and MSE, are carried out for assessing the goodness of pre-trained VAE on 20,000 brain MRI samples.

Table 3 presents that the pre-trained VAEs have overall achieved 98.2% or above of the SSIM metric while preserving the high PSNR and low MSE metrics. These results could guarantee the stability of pre-trained VAEs. Clearly speaking, visual comparisons between real and decoded MRI images for the non-lesion, lesion, and whole images are demonstrated in Figure 1A–C, respectively. It is evident that the images are effectively reconstructed, which suggests that a satisfactory latent distribution has been successfully acquired.

TABLE 3 Variational auto-encoder reconstruction performance on BraTS2020 slices.

Region	SSIM ↑	PSNR ↑	MSE ↓
Lesion	0.9960	33.3	0.0005
Surrounding	0.9830	27.9	0.0016
Whole image	0.9820	30.7	0.0009

2.3 | Analysis for segmentation task

In order to demonstrate how anonymized medical images can be useful, a series of experiments has been carried out. These experiments focus on comparing both real and synthetic images at different grades, including high-grade glioma (HGG) and low-grade glioma (LGG). For the segmentation task on 2D axial slices, we perform binary tumor-versus-background segmentation and compare models trained with real versus anonymized images. The metric of intersection-over-union (mIoU) is used for this comparison. To be more general, different combinations of real and synthetic images are planned to illustrate the utility of anonymized counterparts.

From Table 4, numerous observations can be drawn and further elaborated upon. The training dataset consists of images that are all anonymized, achieving 49.65% mIoU. Although it showed the lowest performance, mixing with real images in a half-by-half manner improved mIoU to 66.00% with the difference of 16.35%. The training dataset, full of real images, obtained the best mIoU at 73.01%. However, the scarcity of real images hugely impaired the segmentation performance, only getting 61.40% when the number of training data was cut in half.

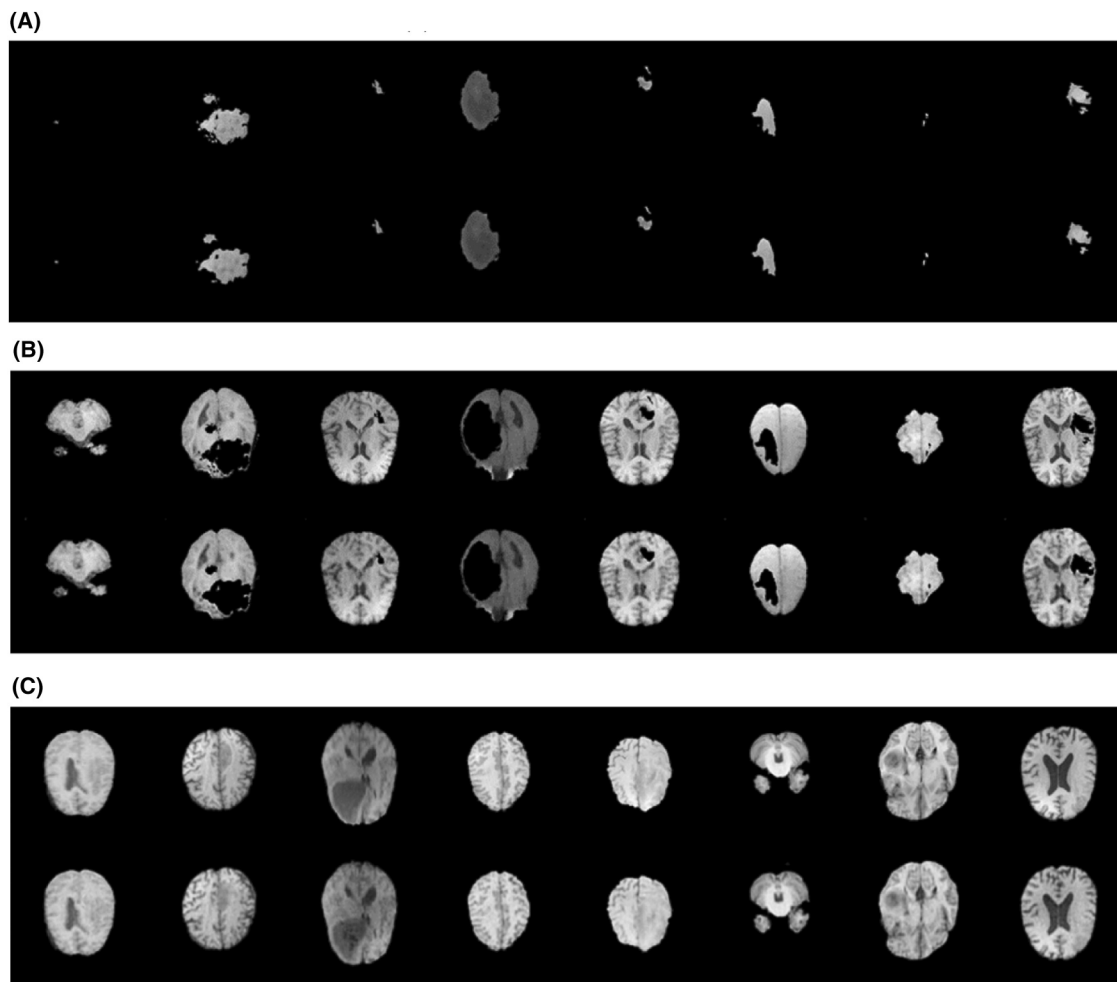


FIGURE 1 Visual comparisons between real and decoded images for non-lesion, lesion, and whole images by variational auto-encoder. (A) Non-lesion regions. (B) Lesion regions. (C) Whole images.

TABLE 4 Segmentation performance (mIoU) on different data settings.

Training set	Real HGG	Real LGG	Synthetic HGG	Synthetic LGG	mIoU (%)
Synthetic only	0	0	1000	1000	49.65
Real (1000)	500	500	0	0	61.40
Real + synthetic	500	500	500	500	66.00
Real (2000)	1000	1000	0	0	73.01

Note: "Synthetic only" contains 2000 synthetic anonymous images without any real images. "Real (1000/2000)" contains 1000/2000 real images. "Real + Synthetic" contains 1000 real and 1000 anonymous images each. Under each setting, the images of both HGG and LGG categories are split evenly.

Abbreviations: HGG: high-grade glioma; LGG: low-grade glioma.

2.4 | Analysis for classification task

We also examined the anonymized images on the lesion classification task with a ResNet18 classifier. Unlike performance in the segmentation task, with the same amount of training samples, anonymized images had a comparable accuracy of 95.75% to real ones of 99.25%. Other results observed in Table 5 demonstrated that the addition of anonymized images to the training dataset by the same number can improve the

classification performance. In this case, 8.75% of its accuracy was enhanced. In addition, the limited density of the training dataset continues to pose challenges for models that rely on data input.

2.5 | Qualitative comparison

To visually demonstrate the effectiveness of anonymized images, we randomly chose a total of 10 different

TABLE 5 Classification performance (accuracy) on different training settings.

Training set	Real HGG	Real LGG	Synthetic HGG	Synthetic LGG	Accuracy (%)
Real (4000)	2000	2000	0	0	99.25
Synthetic only	0	0	2000	2000	95.75
Real (200)	100	100	0	0	61.00
Real + synthetic	100	100	100	100	69.75
Real (400)	200	200	0	0	71.25

Note: "Synthetic only" contains 4000 synthetic anonymous images without any real images. "Real (200/400/4000)" contains 200/400/4000 real images.

"Real + Synthetic" contains 200 real and 200 anonymous images each. Under each setting, the images of both HGG and LGG categories are split evenly.

Abbreviations: HGG: high-grade glioma; LGG: low-grade glioma.

groups of samples for the purposes of this presentation. The first two columns in Figure 2 indicate the original MRI and lesion region, respectively. The following columns signify the results of five medical image generations. It is obvious to observe that the non-lesion regions were varying while the structure and features of the lesion regions were preserved.

Quantitatively, the similarity between the generated and real images was assessed (see Table 6), showing that while the synthetic images differ from the originals (lower SSIM/PSNR compared to VAE reconstructions), they maintain reasonable structural characteristics.

2.6 | Self-designed evaluation

We adopted a PSNR-based criterion for quantifying anonymization performance: a synthetic image $\hat{x}_i$ is considered anonymized if $\text{PSNR}(\hat{x}_i, x_i) - \text{PSNR}(\hat{x}_i, x_j) < 7.36$, where x_j is an image from another patient.

The proposed method was compared with several traditional augmentation/anonymization methods: Gaussian blur, horizontal flipping, random crop and resize, and Gaussian noise addition (with ratios of 0.1 and 0.5). Figure 3 displays qualitative comparisons. Table 7 reports the proportion of synthetic images meeting the anonymization criterion and the average PSNR values.

The results demonstrate that the proposed method achieves a higher anonymization rate than most traditional augmentation techniques, without sacrificing the utility of the generated images.

3 | CONCLUSIONS

This study proposed an alternative framework for MRI image anonymization based on VAE bridging with a diffusion model, consisting of three modules: pre-training VAE encoder-decoder models for lesion and non-lesion regions, respectively, which is followed by performing a diffusion process for the latent

representation of lesion regions, combining with the latent non-lesion regions to generate anonymized MRI images, capable of dealing with region-level image classification tasks. The experimental study illustrated the availability and efficacy of this proposed anonymized approach. Additionally, this work can be further extended to 3D MRI anonymization when taking a consecutive sequence of MRI images. Although the non-lesion regions in BraTS are anonymized, achieving stronger de-identification of the overall anatomical contour (e.g., skull and cranial shape) remains challenging and will be pursued in future work.

As the framework depends on lesion masks to partition lesion versus non-lesion latents, these masks can be obtained automatically from pretrained segmentation models to enable scalability. However, the accuracy of the pre-trained models may affect the training of subsequent anonymization models, and we will study the effects of different pre-trained segmentation models in our future work, and attempt to jointly train segmentation models with anonymization models. We also caution that rare or distinctive lesion morphologies may themselves carry identification risk. In the future work, we will explore lesion-style augmentation and morphology regularization (while preserving clinical plausibility) to reduce identifiability without compromising lesion-centric utility.

4 | MATERIALS AND METHODS

4.1 | Benchmarking datasets

The experiments were conducted using the publicly available BraTS2020 dataset, which is widely used in the field of medical image analysis, especially for brain MRI segmentation tasks [7, 11]. The dataset contains multi-modal MRI scans and segmentation labels from 369 patients, with each patient's MRI scan comprising a 3D volume of size $240 \times 240 \times 155$. For model training and evaluation, we extracted axial slices to obtain single-channel grayscale images of size 240×240 .

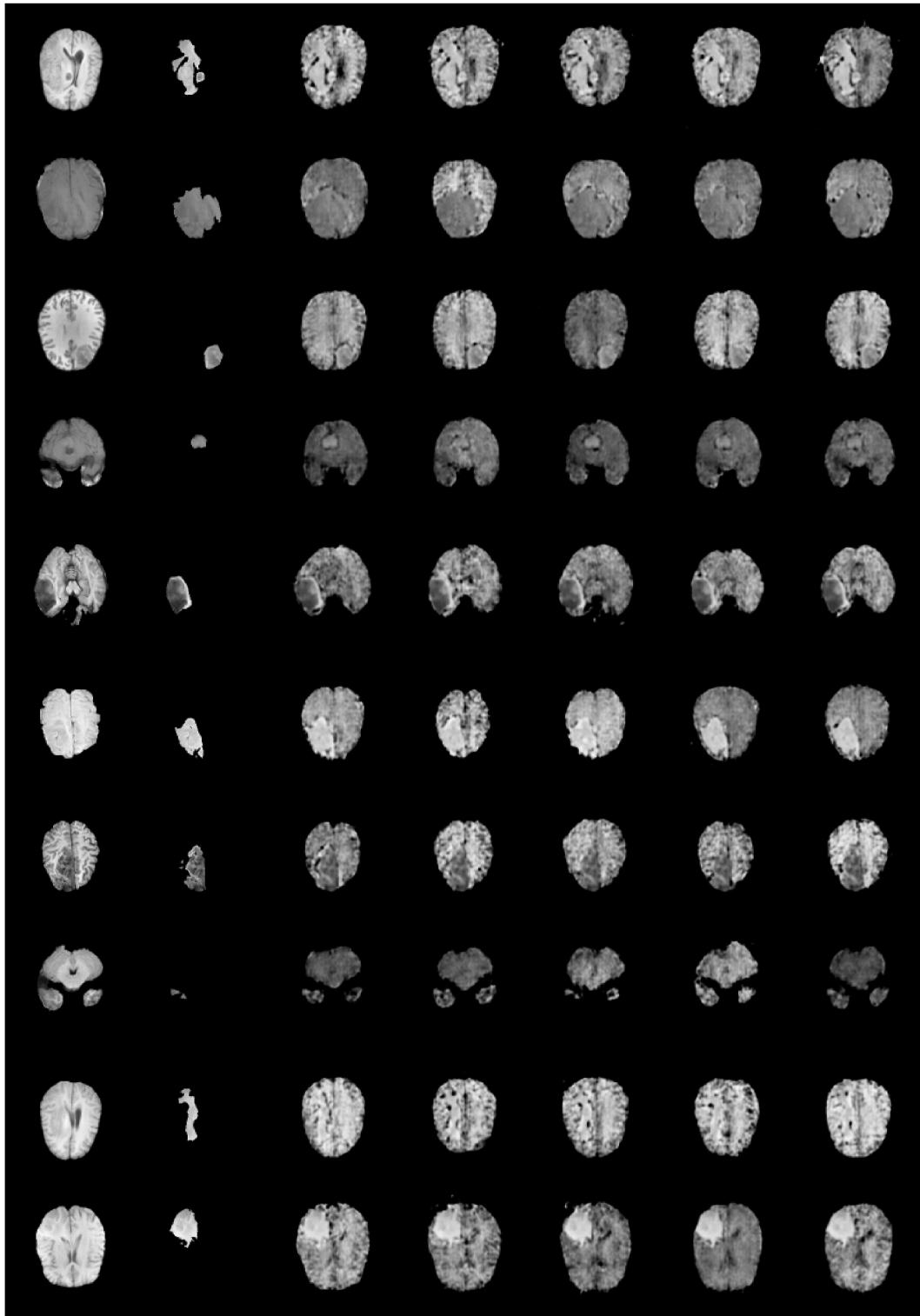


FIGURE 2 Sample virtual magnetic resonance imaging (MRI) images generated from a single real slice. Leftmost: original; second: lesion mask; subsequent columns: different virtual samples.

TABLE 6 The evaluation compares with the original image on different regions (mean $\pm$ std).

Region	SSIM $\uparrow$	PSNR $\uparrow$	MSE $\downarrow$
Whole	0.7466 $\pm$ 0.0016	16.8765 $\pm$ 0.0511	0.0206 $\pm$ 0.0002
Lesion	0.9969 $\pm$ 0.00002	37.7245 $\pm$ 0.0126	0.000169 $\pm$ 0.0000005
Surrounding	0.7678 $\pm$ 0.0016	16.9125 $\pm$ 0.0515	0.0204 $\pm$ 0.0002
Diff (L-S)	0.2291 $\pm$ 0.0016	20.8120 $\pm$ 0.0498	-0.0202 $\pm$ 0.0002

**FIGURE 3** Qualitative comparison of anonymization/augmentation methods: (left to right) original, proposed, Gaussian blur, flip, crop resize, Gaussian noise (0.1), Gaussian noise (0.5).**TABLE 7** Anonymization evaluation results.

Method	Anonymization rate (%)	PSNR (to real)	PSNR (to others)	PSNR difference $\downarrow$
Proposed method	86	17.65	14.93	2.72
Gaussian blur	1	33.01	14.57	18.44
Horizontal flip	61	20.47	13.96	6.50
Crop resize	81	15.71	11.90	3.81
Gaussian noise (0.1)	51	20.00	12.93	7.07
Gaussian noise (0.5)	100	6.02	5.28	0.73

4.2 | Preliminary

4.2.1 | VAE basis

Autoencoder (AE) is a self-supervised neural network designed to encode input data into a compressed representation in the latent space and subsequently decode it to reconstruct the initial input. Nonetheless, the AE lacks restrictions on the encoding space, which frequently leads to a discontinuous latent space. As a result, this can generate unpredictable and nonsensical outputs during sampling.

To remedy this bottleneck, VAE employs the Gaussian distribution to constrain latent vectors, enabling smooth interpolation between inputs and latent vectors and vice versa. Given the observed image input X_v and latent vector Z_v , the purpose of VAE is to obtain the following posterior distribution through the Bayes theorem:

$$p(Z_v | X_v) = \frac{p(X_v | Z_v)p(Z_v)}{p(X_v)} \quad (1)$$

where $p(X_v | Z_v)$ is the likelihood, indicating a decoder while $p(X_v)$ is the normalizing constant.

Real-world data distribution complexities make the posterior distribution $p(Z_v | X_v)$ inherently complicated, posing significant challenges to direct representation. The AE takes a pivotal step by assuming that the latent space, once encoded, adheres to a multivariate Gaussian distribution. Under this presumption, the latent space encoding merely needs to capture the mean and variance for each dimension of the Gaussian distribution, thereby allowing for the efficient recording of Gaussian distribution characteristics at every position within the forward pass. Through the application of the reparameterization trick, the VAE can successfully sample the true latent space variables from the encoded latent space, thereby enabling the reconstruction of the original input. During training, the VAE endeavors to refine this assumption by minimizing the disparity from the target distribution, adjusting its learning process accordingly, as its loss function is intricately structured to encourage closer approximation.

$$L_v = L_{\text{rec}} + L_{\text{KL}} \quad (2)$$

$$L_{\text{rec}} = \text{MSE}(X_v, \widehat{X}_v) = \frac{1}{n} \sum_n (x_i - \widehat{x}_i)^2 \quad (3)$$

$$L_{\text{KL}} = D_{\text{KL}}(X_v, \widehat{X}_v) = \int_{X_v} q(Z_v | X_v) \cdot \log \left(\frac{q(Z_v | X_v)}{p(Z_v | X_v)} \right) dX_v \quad (4)$$

By integrating the L2 loss, which provides a measure of the reconstruction error, together with the KL

loss, which assesses the discrepancy in distribution approximation, into the loss function, the VAE is capable of improving the quality of restored images. Additionally, this approach aids in approximating the latent space distribution to resemble a multivariate Gaussian distribution more effectively as the training progresses.

4.2.2 | Diffusion model

The diffusion model is designed to learn the underlying distribution of input data, thereby enabling the generation of new samples that conform to the same distribution. Compared with classical generative models such as AEGAN, diffusion models demonstrate superior performance in both sample quality and training stability. Its central innovation lies in the introduction of a forward diffusion process and a corresponding reverse generative process, which together form a Markov chain of noise perturbation and denoising.

In the forward process, the input data x_0 is gradually perturbed by Gaussian noise over T steps, eventually transforming into nearly pure Gaussian noise. At each step $t \in \{1, 2, \dots, T\}$, noise is added according to the conditional distribution

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I) \quad (5)$$

where $\beta_t \in (0, 1)$ is a pre-defined variance schedule controlling the magnitude of injected noise at step t . By iteratively applying this process, the data distribution converges to an isotropic Gaussian distribution. This process also admits a closed-form solution that allows direct sampling of x_t from the original input x_0 without explicitly traversing all intermediate states:

$$q(x_t | x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) I) \quad (6)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. This formulation shows that the forward process is equivalent to a gradual interpolation between the data distribution and Gaussian noise.

The reverse process $p_\theta(x_{t-1} | x_t)$ is parameterized by a neural network and aims to gradually remove noise in order to reconstruct the input distribution. Since the true reverse transition $q(x_{t-1} | x_t)$ is intractable, it is approximated as a Gaussian distribution with mean and variance predicted by the neural network:

$$p_\theta(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, c), \Sigma_\theta(x_t, t, c)) \quad (7)$$

where μ_θ and Σ_θ denote the estimated mean and variance, and c represents the conditioning information.

Similar to VAE, the training objective of the diffusion model is to maximize the log-likelihood function by optimizing the variational lower bound (VLB):

$$L_{\text{VLB}} = D_{\text{KL}}(q(x_T | x_0) | p_\theta(x_T)) + \sum_{t=2}^T L_t - \log p_\theta(x_0 | x_1) \quad (8)$$

where $L_t = D_{\text{KL}}(q(x_{t-1} | x_t, x_0) | p_\theta(x_{t-1} | x_t))$, and the $q(x_{t-1} | x_t, x_0)$ can be derived using Bayes theorem.

In practice, most implementations fix the variance Σ_θ and only learn the mean. Training is performed by optimizing a variational bound L_t , which reduces to predicting the noise component ϵ that was added in the forward process:

$$\mathcal{L}_{\text{sml}} = E_{t, x_0, \epsilon} [\|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2] \quad (9)$$

where ϵ_θ is the neural network prediction of the Gaussian noise ϵ , θ denotes the learnable parameters of the diffusion network. By learning to predict ϵ at arbitrary time steps, the model implicitly learns the data distribution and is able to generate high-quality samples by reversing the diffusion trajectory step by step.

4.3 | Related work

4.3.1 | Generative model as medical image synthesis and augmentation

Generative models have been increasingly adopted for medical image synthesis and augmentation, primarily aiming to address data scarcity and class imbalance. Early approaches focused on GAN. For example, Shin et al. [12] proposed a conditional GAN framework for brain MRI augmentation that leverages segmentation labels to map label maps into realistic images. Uzunova et al. [13] extended this by introducing tumor feature injection to model lesion-induced tissue deformations, enhancing realism. Ge et al. [14] further developed pairwise GAN to exploit cross-modality consistency for improved brain MRI synthesis. Similar GAN-based augmentation strategies have been applied to retinal imaging [15–17] and pulmonary lesion imaging [18, 19], broadening the scope of application.

Recently, diffusion models have emerged as a powerful alternative, demonstrating better image quality, diversity, and stability in medical image augmentation tasks [20]. In their study, Zhang et al. [21] utilized the single-image denoising diffusion model (SinDDM) and its variation, FewDDM, on lung ultrasound images, showing that it produced higher image quality than

single-image GAN and enhanced multi-class classification results. Morão et al. [22] introduced a conditional denoising diffusion generative model (cDDGM) for producing counterfactual MR images under different imaging scenarios, thereby remarkably enhancing segmentation robustness in the face of domain shifts. Montoya-del-Angel et al. [23] presented MAM-E, a system that employs a text-prompted conditional diffusion method to produce complete digital mammograms and artificial lesions, allowing for manageable and superior quality enhancements in mammographic imaging.

Although these studies highlight the potential of generative models to enhance data availability and refine downstream analysis, they predominantly concentrate on data utility. Privacy considerations are frequently neglected, creating a risk that synthetic images might still embed patient-specific features and allow for re-identification when cross-referenced between datasets with multiple central points. Maximizing the data utility while preserving data privacy remains problematic.

4.3.2 | Generative model as medical image anonymization

To the best of our knowledge, there have been a countable number of studies that have explored the use of generative models for anonymizing medical images. The initial techniques for anonymizing MRI images concentrated on blurring or removing facial characteristics. DeepPrivacy [24] pioneered the use of GAN for facial anonymization by replacing the facial area with a synthetic, naturally posed face while preserving the surrounding context. Its successor, DeepPrivacy2 [25], enhanced this method to support high-resolution and full-body anonymization. Additionally, CIAGAN [26] introduced conditional identity anonymization, refining the distinction between identity and expression. Although these methods can compromise the seen image features, impeding observability, re-identification from the imaged anatomical properties would be practically possible [27].

4.3.3 | Limitations of existing work

Across prior studies on generative models for medical image synthesis and augmentation, the emphasis has largely been on improving downstream utility, whereas privacy considerations are rarely examined, synthetic images can still carry patient-specific anatomical signatures. Conversely, existing anonymization efforts in above section have predominantly targeted facial identity, achieving de-identification by altering or replacing visible facial regions, however they do not address re-identification via anatomical properties.

Motivated by these gaps, we proposed a novel framework for bridging VAE and a diffusion model, aiming at maximizing the utility of medical images while minimizing the leakage of data privacy. Furthermore, we adapt an existing anonymization evaluation metric from tabular data research to the context of medical image augmentation, in order to rigorously assess the privacy-preserving effectiveness of our proposed method.

4.4 | Medical image anonymization

As demonstrated in Figure 4, our proposed methodology for anonymizing medical images consists of a detailed and well-structured framework. The core of our approach is centered on the latent diffusion model. Our process commences with the use of a AE, which projects medical images into a latent space. This projection not only reduces the image size but also enhances feature differentiation. Once the images are mapped into this latent space, we proceed with training the diffusion model using these latent vectors, enabling the model to grasp the underlying data distribution. In the final step, we use the diffusion model to sample random latent vectors that symbolize non-lesion regions. These sampled vectors are then combined with the original latent vectors corresponding to the lesion regions. This synthesis results in anonymized images that preserve the crucial characteristics of the lesion regions, thus achieving effective anonymization while maintaining the indispensable features of the lesions.

Algorithm 1 offers a detailed overview of the pipeline as a whole. In the following sections, we will thoroughly examine both the basic concepts and the complex nuances of each element that makes up this framework. Notably, we shall investigate three core components: the VAE, the diffusion model, and the methodology employed to generate anonymized images. Each element will be individually scrutinized, with a focus on elaborating its functions and mechanisms, to provide a thorough understanding of its contributions to the overall system.

Algorithm 1 Medical image anonymization pipeline with VAE and diffusion model.

- 1: Input: Gaussian noise $\mathbf{z}$, lesion-only image $\mathbf{x}_{\text{lesion}}$
- 2: Output: Virtual image $\hat{\mathbf{x}}$
- 3: VAE pre-training process:
- 4: Train VAE₁ on lesion-free images
- 5: Train VAE₂ on lesion-only images
- 6: Further train the decoder D₃ with concatenated features from frozen VAE₁ and VAE₂ encoder
- 7: Diffusion model training process:

- 8: Train the diffusion model in the latent space of VAE₁ using Gaussian noise with frozen encoder E₁, conditioned on the class label **C**, the lesion mask **M** and the slice depth **D**
- 9: Medical image anonymization:
- 10: Sample noise $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- 11: Denoise **z** with diffusion model conditioned on the class label **C**, the lesion mask **M** and the slice depth **D**, to obtain virtual non-lesion regions features **L_d**
- 12: Encode the lesion-only image **x_{lesion}** with VAE₂ to get lesion regions features **L₂**
- 13: Concatenate the features of lesion and non-lesion regions $\mathbf{f}_{concat} = [\mathbf{L}_d; \mathbf{L}_2]$
- 14: Decode **f_{concat}** with Decoder D₃ to finally generate virtual anonymized image **x̂**

4.4.1 | Pre-training of VAE

The principal concept of our approach is to maintain critical lesion details while altering non-essential aspects. To accomplish this, we propose a region-separation framework that treats the lesion regions and its non-lesion regions separately. As depicted in Figure 4, we initially isolate the lesion regions from the

adjacent tissues in brain MRI scans. This strategy enables us to distinguish key lesion features from the contextual background, ensuring the synthesized images remain clinically useful and better anonymized.

Figure 4 demonstrates that the proposed framework consists of a pair of entirely developed VAE models, identified as VAE₁ and VAE₂, along with an extra decoder module, referred to as D₃. The overall VAE structure is illustrated in Figure 5.

- VAE₁ is dedicated to encoding and reconstructing the non-lesion region.
- VAE₂ is applied to the lesion region.
- Decoder D₃, which shares the same structure as the decoder in a standard VAE, takes as input the concatenated latent representations from both VAE₁ and VAE₂, and reconstructs the full brain MRI image.

Formally, given an input image x , we first split it into lesion region x_l and non-lesion region x_s . The encoding-decoding processes are defined as follows:

$$z_l = E_1(x_l), \hat{x}_l = D_1(z_l), \quad (10)$$

$$z_s = E_2(x_s), \hat{x}_s = D_2(z_s), \quad (11)$$

where E_1, D_1 denote the encoder and decoder of VAE₁, and E_2, D_2 denote the encoder and decoder of VAE₂.

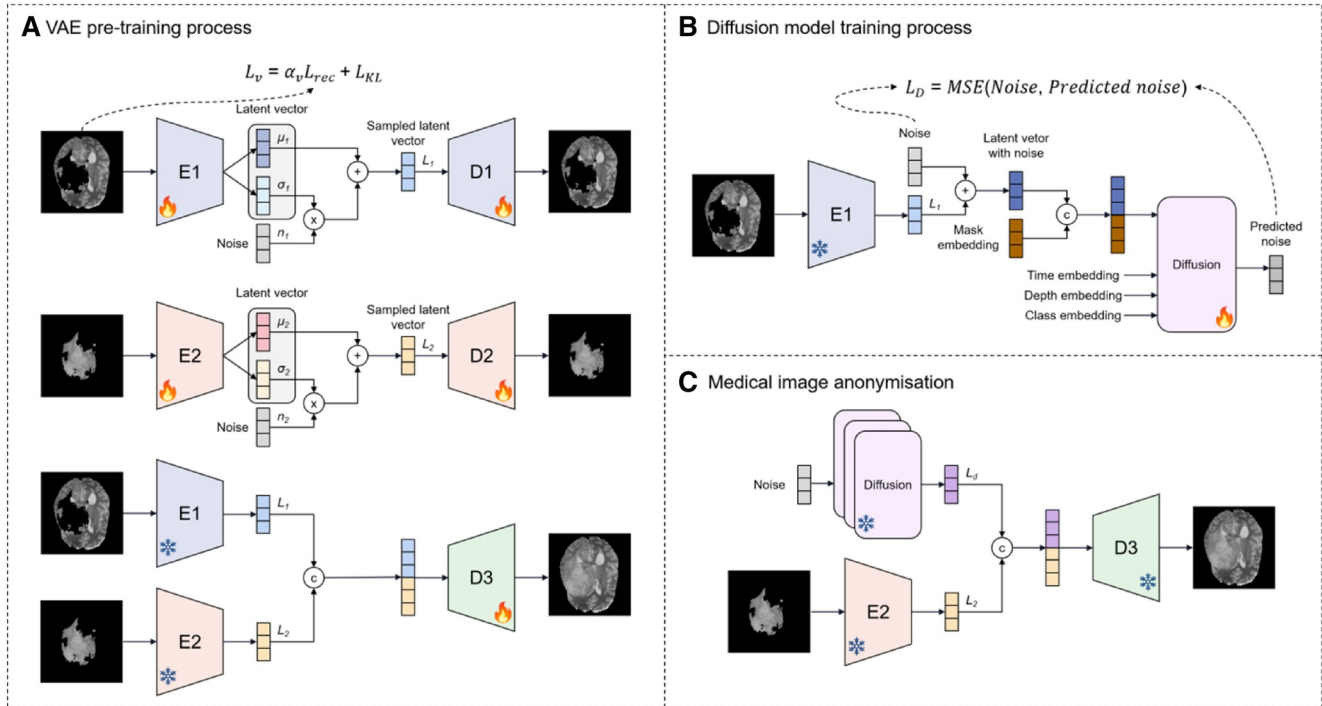


FIGURE 4 The pipeline of the proposed diffusion model-based VAE, including three stages: (A) VAE pre-training process: Training the VAE that encodes the spatial information of medical images into latent space vectors; (B) diffusion model training process: After obtaining the VAE model trained in the previous step, we use the intermediate output of the VAE to train the diffusion model, obtaining a diffusion model that generates new latent space vectors from pure noise; and (C) medical image anonymization: After the training of the first two steps is completed, freeze the model parameters, and integrate the modules of the first two parts to generate the final anonymized images. VAE, variational auto-encoder.

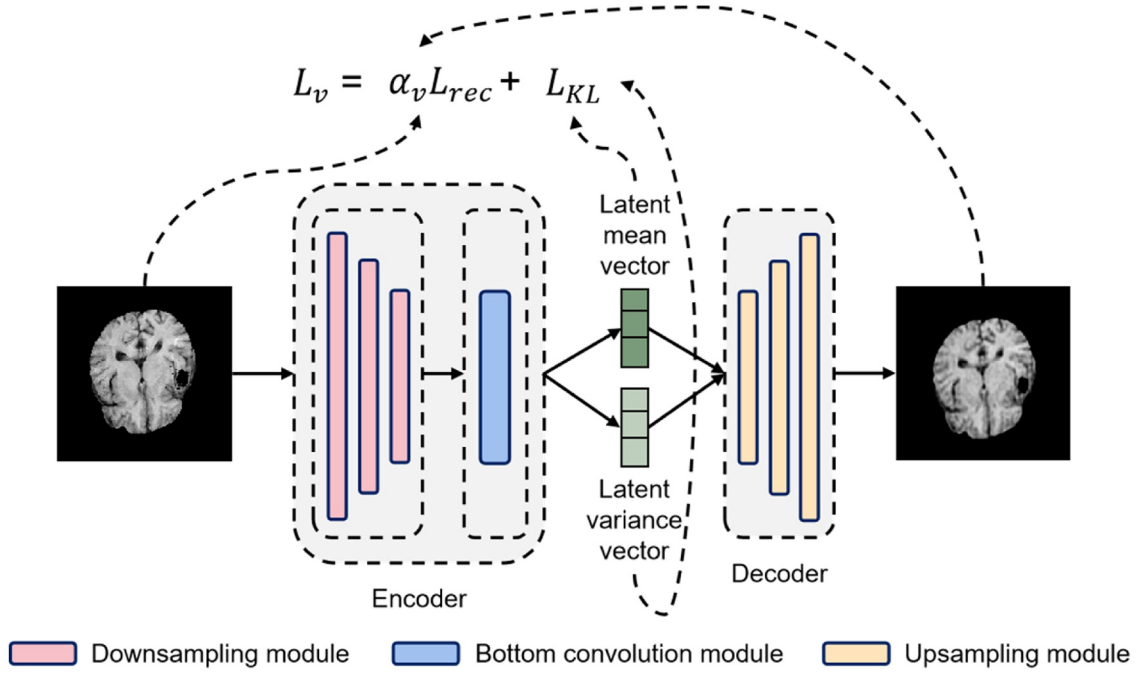


FIGURE 5 The detail of variational auto-encoder architecture.

Finally, the fused reconstruction is obtained by the following:

$$\hat{x} = D_3([z_l, z_s]), \quad (12)$$

where $[z_l, z_s]$ denotes the concatenation of lesion and non-lesion latent representations, and D_3 is the independent decoder ensuring continuity at the fusion boundary.

The latent codes from the two VAE serve different purposes:

- The lesion latent representation z_l is preserved without further modification, ensuring that clinically relevant lesion features remain intact.
- The non-lesion latent representation z_s is subsequently fed into a latent diffusion model, which selectively modifies non-essential background features, achieving the goal of anonymization.
- Decoder D_3 integrates both latent codes and reconstructs the complete brain MRI, avoiding discontinuity that would occur if fusion were performed directly in image space.

In order to balance the reconstruction loss and the KL divergence during training, we introduce a variable hyperparameter α_v to control their relative contributions. The final training objective is defined as follows:

$$L_v = \alpha_v L_{rec} + L_{KL} \quad (13)$$

To ensure that the model focuses more on reconstruction in the early stage of training and gradually shifts attention to the latent distribution in the later

stage, we adopt an iterative scheduling strategy for α_v . Specifically, its initial value is set to 16, and it is halved every 10 training epochs until it reaches 1:

$$\alpha_v^{(t+1)} = \max\left(\frac{\alpha_v^{(t)}}{2}, 1\right). \quad (14)$$

This strategy guarantees that the training process prioritizes reconstruction loss at the beginning while progressively incorporating the KL divergence constraint in later stages, leading to more stable and realistic image generation.

4.4.2 | Training of diffusion model in latent space

Directly applying diffusion processes in the original image space is computationally expensive and often redundant, as high-dimensional pixel representations contain perceptually irrelevant details. To address this issue, we adopt the latent diffusion framework, where the input non-lesion image x_l is first encoded into a compact latent representation z_l through a pretrained autoencoder $E_1(\cdot)$. The diffusion process is then performed in this latent space, which significantly reduces computational costs while preserving essential semantic information. For simplicity, the following text uses z_0 to represent the original hidden vector and z_t to represent the vector after t steps of noise addition.

In the forward diffusion process, Gaussian noise is gradually added to the latent representation across T steps:

$$q(z_t | z_{t-1}) = \mathcal{N}(z_t; \sqrt{1 - \beta_t} z_{t-1}, \beta_t I), \quad (15)$$

with the closed-form solution

$$q(z_t | z_0) = \mathcal{N}(z_t; \sqrt{\bar{\alpha}_t} z_0, (1 - \bar{\alpha}_t) I), \quad (16)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$.

The reverse process aims to recover z_0 from noisy latent z_t by gradually denoising. This is parameterized by a conditional U-Net, which takes as input the noisy latent z_t , the diffusion time step t , and additional domain-specific conditions. Specifically, we introduce two auxiliary conditions beyond conventional class label c : a two-dimensional lesion mask M that provides structural information of lesion regions, and a slice depth parameter D that encodes the relative spatial position within the volumetric scan. These conditional signals are embedded through learnable encoders and injected into the U-Net at multiple scales, ensuring that the model is aware of lesion morphology and global spatial context during denoising. The detailed U-Net of diffusion model is illustrated in Figure 6.

The reverse transition is thus modeled as follows:

$$\begin{aligned} p_\theta(z_{t-1} | z_t, c, M, D) \\ = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t, c, M, D), \Sigma_\theta(z_t, t, c, M, D)). \end{aligned} \quad (17)$$

In practice, we fix the variance Σ_θ and let the U-Net predict the noise component ϵ added in the forward process. The training objective is defined as the mean squared error (MSE) between the true noise and the predicted noise:

$$\mathcal{L}_D = E_{t, z_0, \epsilon} [\|\epsilon - \epsilon_\theta(z_t, t, c, M, D)\|_2^2]. \quad (18)$$

During inference, the model starts from Gaussian noise in the latent space and iteratively applies the learned reverse process until obtaining a denoised latent $\hat{z}_0$, which is then decoded back into the image space by $D_3(\cdot)$. This procedure enables efficient generation of high-quality medical images that preserve both global anatomical consistency and local lesion details.

4.4.3 | Generation of anonymized MRI images

During inference, anonymized MRI images are generated by integrating lesion-specific information with anatomically plausible structures derived from the diffusion model. As illustrated in Figure 4, the process begins with random noise, which is iteratively denoised by the diffusion model with the conditional class label c , lesion mask M , and slice depth D , to produce a latent representation L_d that captures the global anatomical structure of healthy brain tissue. In parallel, lesion-only MRI images are passed through the encoder E_2 to obtain latent codes L_2 , which represent the spatial distribution and morphological characteristics of lesions.

To synthesize a complete MRI image, the diffusion-derived latent vector L_d is concatenated with the lesion representation L_2 , thereby forming a joint latent space that contains both structural and lesion information. This combined latent representation is then decoded by D_3 to generate a full-resolution MRI image. The resulting image preserves realistic anatomical details while seamlessly embedding lesion regions.

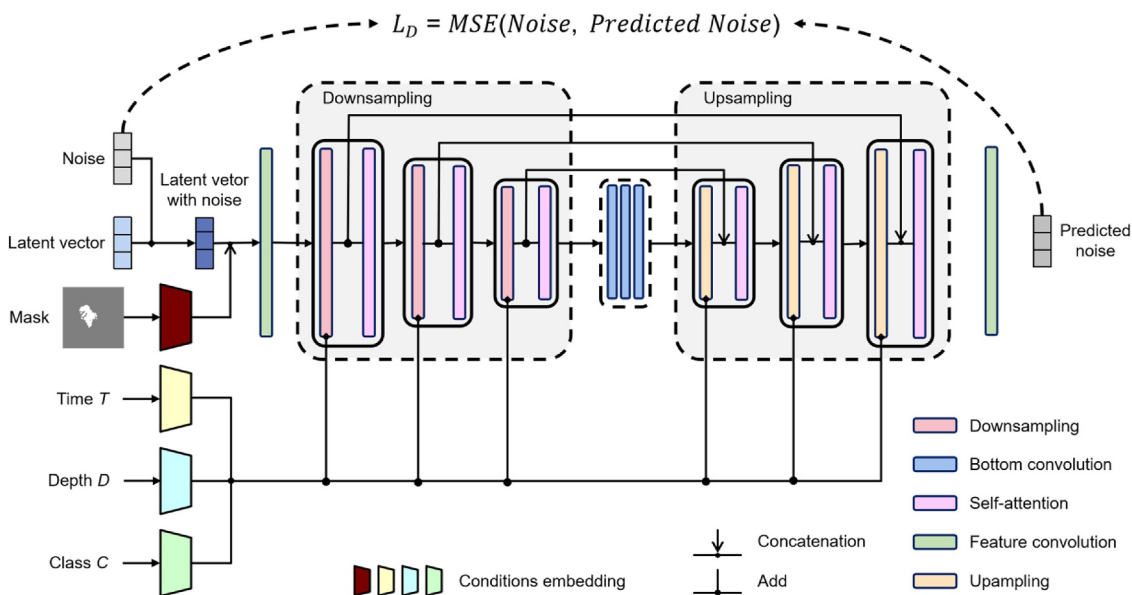


FIGURE 6 The detailed U-net architecture of diffusion model.

By disentangling healthy structures from lesion information and recombining them only at the decoding stage, the pipeline ensures that the generated images are both anatomically coherent and anonymized while still retaining clinically relevant lesion features. This inference mechanism enables the controlled synthesis of anonymized MRI data suitable for downstream applications such as model training and evaluation.

4.5 | Criterion of evaluating validation of anonymized MRI

In addition, we also construct an evaluation criterion to measure how much privacy MRI images preserve.

Yoon et al. [28] proposed using weighted Euclidean distance as a similarity criterion in their research on anonymizing multidimensional laboratory report data, as shown in Equation (19):

$$\begin{aligned} r_i &= \min_{x \in D/x_i} \|W \cdot (x_i - x_j)\| \\ \hat{r}_i &= \min_{x \in \hat{D}/\hat{x}_i} \|W \cdot (x_i - \hat{x}_j)\| \\ \mathcal{A} &= 1_{\hat{r}_i < \epsilon r_i} \end{aligned} \quad (19)$$

where r_i represents the similarity index of the real sample i , $\hat{r}_i$ represents the similarity index of the synthetic sample of i , D is the real dataset, $\hat{D}$ is the synthetic dataset, and W is the weighted matrix, ϵ is the anonymity threshold parameter and $\mathcal{A}$ indicates whether the sample meets the anonymity.

The anonymity of the synthetic samples is based on the above similarity. Specifically, if the distance between the synthetic sample and other samples is less than ϵ times the distance between the real sample, it is considered that the sample has good anonymity.

Inspired by this idea, we propose to use PSNR instead of the weighted Euclidean distance. PSNR is an important indicator for measuring the similarity between two images, and its formula is as follows:

$$\text{PSNR}(x, y) = 20 * \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}(x, y)} \right) \quad (20)$$

Moreover, PSNR also contains the MSE, so we adopt the PSNR difference to define the distinguishability between two pairs of images, and use it as an evaluation criterion for data anonymity, as shown in Equation (21):

$$\begin{aligned} \text{Anony}(\hat{x}_i) &= \text{PSNR}(\hat{x}_i, x_i) - \text{PSNR}(\hat{x}_i, x_j) \\ &= 20 \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}(\hat{x}_i, x_i)} \right) - 20 \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}(\hat{x}_i, x_j)} \right) \\ &= 20 \log_{10} \frac{\text{MSE}(\hat{x}_i, x_j)}{\text{MSE}(\hat{x}_i, x_i)} \end{aligned} \quad (21)$$

where x_i and x_j are two real medical images, and $\hat{x}_i$ is the synthetic image generated using x_i . The larger the PSNR difference, the closer the synthetic image is to its original real image, indicating poorer anonymity; whereas the smaller the PSNR difference, the harder it is to distinguish the source of the original image from the synthetic image, indicating better anonymity.

In addition, to make it easier to determine the threshold of evaluation criterion for data anonymity, we propose the assumption that “the probability of different images being distinguished as the original source of a certain synthetic image is inversely proportional to its MSE with the synthetic image,” which means the following:

$$\begin{aligned} p_i &= \frac{\frac{1}{\text{MSE}(\hat{x}_i, x_i)}}{\frac{1}{\text{MSE}(\hat{x}_i, x_i)} + \frac{1}{\text{MSE}(\hat{x}_i, x_j)}} \\ &= \frac{\text{MSE}(\hat{x}_i, x_j)}{\text{MSE}(\hat{x}_i, x_i) + \text{MSE}(\hat{x}_i, x_j)} \end{aligned} \quad (22)$$

Substitute Equation (22) into the Equation (21), we can finally obtain the following:

$$\begin{aligned} \text{Anony}(\hat{x}_i) &= 20 \log_{10} \frac{\text{MSE}(\hat{x}_i, x_j)}{\text{MSE}(\hat{x}_i, x_i)} \\ &= 20 \log_{10} \frac{p_i}{1 - p_i} \end{aligned} \quad (23)$$

At this stage, the threshold for the PSNR difference can be determined by specifying a discriminant probability p within the interval $(0, 1)$. In this paper, we set $p = 0.7$, resulting in $\text{Anony}_{\text{threshold}} = 7.36$. Specifically, when $\text{PSNR}(\hat{x}_i, x_i) - \text{PSNR}(\hat{x}_i, x_j) < 7.36$, the synthesized image is considered to exhibit a satisfactory level of data anonymity. Conversely, if the difference exceeds this threshold, the anonymity is deemed insufficient.

AUTHOR CONTRIBUTIONS

Yunheng Shen: Conceptualization; methodology; validation; investigation; formal analysis; writing—original draft; writing—review and editing. **Ling Zheng:**

Methodology; data curation; writing—original draft; writing—review and editing. **Ruohan Liu:** Visualization; data curation. **Haoran Feng:** Formal analysis; data curation. **Hairong Lv:** Supervision; conceptualization; funding acquisition; resources.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (Grant No. 2022YFF1202400).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The BraTS dataset can be accessed at the MICCAI website.

ETHICS STATEMENT

This article does not contain any studies with human or animal materials performed by any of the authors.

REFERENCES

- [1] Sood A, Pattnaik T, Malhotra R, Nayyar V, Narayan B, Mishra D, et al. Medical imaging privacy: a systematic scoping review of key parameters in dataset construction and data protection. *J Med Imag Radiat Sci*. 2025;56(5):101914.
- [2] Regulation P. Regulation (EU) 2016/679 of the European Parliament and of the council. Regulation. 2016; 679(2016):10–3.
- [3] Act A. Health insurance portability and accountability act of 1996. Public Law. 1996;104:191.
- [4] Fezai L, Urruty T, Bourdon P, Fernandez-Maloigne C, Initiative ADN. Deep anonymization of medical imaging. *Multimed Tool Appl*. 2023;82(6):9533–47.
- [5] Li H, Ayache N, Delingette H. Generative medical image anonymization based on latent code projection and optimization. In: 2025 IEEE 22nd international symposium on biomedical Imaging (ISBI). IEEE; 2025. p. 1–4.
- [6] Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, et al. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics*. 2005;15(4):869–77.
- [7] Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imag*. 2014;34(10):1993–2024.
- [8] LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. *Neural Comput*. 1989;1(4):541–51.
- [9] Johnson AE, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*. 2019;6(1):317.
- [10] Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, et al. The multimodal brain tumor image segmentation benchmark (BraTS) 2020 challenge. In: Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), BrainLes Workshop. Lecture Notes in Computer Science. Springer; 2020.
- [11] Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby J, et al. Segmentation labels and radiomic features for the preoperative scans of the TCGA-LGG collection [Data Set]. The Cancer Imaging Archive. Version; 2017.
- [12] Shin HC, Tenenholtz NA, Rogers JK, Schwarz CG, Senjem ML, Gunter JL, et al. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In: Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction With MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings. Vol. 11037 of Lecture Notes in Computer Science. Springer; 2018. p. 1–11.
- [13] Uzunova H, Ehrhardt J, Handels H. Generation of annotated brain tumor MRIs with tumor-induced tissue deformations for training and assessment of neural networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV. Vol. 12264 of Lecture Notes in Computer Science. Springer; 2020. p. 501–11.
- [14] Ge C, Gu IYH, Jakola AS, Yang J. Enlarged training dataset by pairwise GANs for molecular-based brain tumor classification. *IEEE Access*. 2020;8:22560–70.
- [15] Mahapatra D, Bozorgtabar B, Shao L. Pathological retinal region segmentation from OCT images using geometric relation based augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020. p. 9611–20.
- [16] Diaz-Pinto A, Colomer A, Naranjo V, Morales S, Xu Y, Frangi AF. Retinal image synthesis and semi-supervised learning for glaucoma assessment. *IEEE Trans Med Imag*. 2019;38(9):2211–8.
- [17] Zhao H, Li H, Maurer-Stroh S, Cheng L, Wong WK, See SA. Synthesizing retinal and neuronal images with generative adversarial nets. *Med Image Anal*. 2018;49:14–26.
- [18] Motamed S, Rogalla P, Khalvati F. Data augmentation using generative adversarial networks (GANs) for GAN-based detection of pneumonia and COVID-19 in chest X-ray images. *Inform Med Unlocked*. 2021;27:100779.
- [19] Wang Q, Zhang X, Zhang W, Gao M, Huang S, Wang J, et al. Realistic lung nodule synthesis with multi-target co-guided adversarial mechanism. *IEEE Trans Med Imag*. 2021;40(9):2343–53.
- [20] Liu Y, Wang W, Zhou H, Zhang Y, Li X. A review of diffusion models in microscopic biomedical imaging. *Front Med*. 2025; 12:1551894.
- [21] Zhang Y, Liu Z, Kim I, Hwang S. SinDDM and FewDDM: single-image and few-shot denoising diffusion models for medical image generation. In: Proceedings of the 3rd Conference on Medical Imaging With Deep Learning. Vol. 225 of Proceedings of Machine Learning Research; 2023. p. 882–900.
- [22] Morão I, Shapira N, Baumgartner CF, Ball T, Bortsova G, Welling M, et al. Counterfactual medical image generation using conditional denoising diffusion models. 2024. Preprint at arXiv:241023835.
- [23] Montoya-del Angel D, Qiu H, Valverde S, Oliver A, Martí R. MAM-E: text-prompted diffusion models for mammogram and lesion synthesis. 2023. Preprint at arXiv:231109822.
- [24] Hukkelås H, Lindseth F, Lindseth F. DeepPrivacy: a generative adversarial network for face anonymization. In: International Symposium on Visual Computing. Springer; 2019. p. 565–78.
- [25] Hukkelås H, Mester R, Lindseth F. DeepPrivacy2: towards realistic full-body anonymization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 14395–404.
- [26] Maximov M, Elezi I, Leal-Taixé L. CIAGAN: conditional identity anonymization generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 5447–56.

- [27] Steeg K, Bohrer E, Schäfer SB, Vu VD, Scherberich J, Windfelder AG, et al. Re-identification of anonymised MRI head images with publicly available software: investigation of the current risk to patient privacy. *eClinicalMedicine*. 2024;78:102930.
- [28] Yoon J, Drumright LN, Van Der Schaar M. Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE J Biomed Health Inform*. 2020;24(8): 2378–88.

How to cite this article: Shen Y, Zheng L, Liu R, Feng H, Lv H. Diffusional magnetic resonance imaging anonymizing with variational autoencoder. *Quantitative Biology*. 2026;e70033. <https://doi.org/10.1002/qub2.70033>