

# 3D MRI SYNTHESIS WITH SLICE-BASED LATENT DIFFUSION MODELS: IMPROVING TUMOR SEGMENTATION TASKS IN DATA-SCARCE REGIMES

Aghiles Kebaili<sup>1</sup>, Jérôme Lapuyade-Lahorgue<sup>1</sup>, Pierre Vera<sup>2</sup> and Su Ruan<sup>1</sup>

<sup>1</sup> LITIS UR 4108, University of Rouen-Normandy, Rouen, 76000, France

<sup>2</sup> CLCC Henri Becquerel, Rouen, 76038, France

## ABSTRACT

Despite the increasing use of deep learning in medical image segmentation, the limited availability of annotated training data remains a major challenge due to the time-consuming data acquisition and privacy regulations. In the context of segmentation tasks, providing both medical images and their corresponding target masks is essential. However, conventional data augmentation approaches mainly focus on image synthesis. In this study, we propose a novel slice-based latent diffusion architecture designed to address the complexities of volumetric data generation in a slice-by-slice fashion. This approach extends the joint distribution modeling of medical images and their associated masks, allowing a simultaneous generation of both under data-scarce regimes. Our approach mitigates the computational complexity and memory expensiveness typically associated with diffusion models. Furthermore, our architecture can be conditioned by tumor characteristics, including size, shape, and relative position, thereby providing a diverse range of tumor variations. Experiments on a segmentation task using the BRATS2022 confirm the effectiveness of the synthesized volumes and masks for data augmentation. Code is available here : <https://github.com/Arksyd96/synthesis-with-slice-based-ldm>

**Index Terms**— Data Augmentation, Diffusion Models, Generative Modeling, MRI

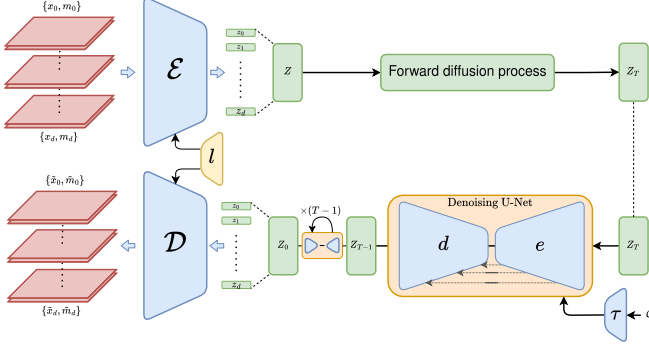
## 1. INTRODUCTION

Deep learning has witnessed remarkable growth in medical imaging, demonstrating its notable effectiveness segmentation tasks across various imaging modalities, including MRI [1, 2]. However, the ongoing challenge of limited access to annotated medical imaging data is a major challenge, primarily due to the rarity of certain pathologies and rigorous medical privacy regulations, consequently leading to a laborious and time-consuming manual delineation of tumor masks by medical professionals. In this context, data augmentation has emerged as an inseparable part of deep learning, enabling models to overcome the limitations associated with a scarcity of training samples and generalize more effectively the data. However, when dealing with com-

plex medical imaging structures, conventional augmentation techniques such as rotations, cropping or noise injection, may introduce deformations, resulting in deviations from the true data distribution. To address these challenges, advanced deep learning-based data augmentation techniques have been proposed, striving to generate synthetic samples that closely resemble real data while preserving the semantic integrity of the medical images [3, 4]. These models also offer privacy preservation and data anonymization.

Generative Adversarial Networks (GANs) [5] have found widespread applications in medical imaging [3, 6] and have been advocated in numerous literature reviews for data augmentation due to their ability to generate realistic images [7]. However, GANs exhibit certain limitations, including learning instability, convergence issues, and the well-documented problem of mode collapse [8], where the generator produces a limited range of samples. In contrast, Variational Autoencoders (VAEs) [9] have been proposed as an alternative to GANs, offering a more stable training process and a more efficient inference procedure. However, VAEs are also known to produce blurry images [3] and are incapable of generating high-resolution images. Recently, diffusion models have emerged as a promising method for image synthesis, offering superior image quality and realism compared to GANs while maintaining a good mode coverage [10]. This has led to the rise of these models and the development of various alternatives, such as the Latent Diffusion Model (LDM) [11]. Although these models provide an attractive solution to the challenge of limited training data, a common issue arises from their high computational cost and demanding memory requirements, making them impractical for 3D medical image synthesis. This holds particularly true for diffusion-based models, which are more resource-intensive, presenting challenges in their integration into clinical routines, especially for real-time tasks like data harmonization or imputation. Beyond this, generative models also require a significant amount of data, limiting their feasibility in medical imaging.

Recent studies have primarily concentrated on image generation or translation [12, 13], which, in the context of tumor segmentation, is insufficient. The importance lies in generating both images and their corresponding tumor masks, as these masks serve as ground truth for segmentation tasks,



**Fig. 1.** Illustration of the proposed architecture. Initially, an MRI volume and its associated 3D mask are decomposed into multiple pairs of 2D slices and masks, denoted as  $\{x_i, m_i\}$ . These are fed into the encoder  $\mathcal{E}$ .  $l(\cdot)$  and  $\tau(\cdot)$  are the positional and condition embedders, respectively.  $c$  represents the tumor features vector. See section 2.2.

adding complexity and cost to the generation process, as we must generalize not only the medical image but also the associated mask. In this study, we introduce a lightweight variant of latent diffusion models (named SBLDM), employing a slice-by-slice approach for the simultaneous generation of medical images and corresponding segmentation masks. Our architecture is trained under data limitations, and we demonstrate its efficacy in augmenting training data for segmentation tasks. Moreover, our model allows precise control of tumor size, shape, and relative position—enabling the generation of a diverse range of tumor variations. This conditioning also serves as regularization to our model. Our evaluation encompasses the quality of generated images, followed by a comprehensive assessment in the context of 3D segmentation tasks using synthesized volumes.

This paper contains three main contributions:

- Proposition of an efficient, slice-by-slice diffusion model for the simultaneous generation of high-quality medical images and associated segmentation masks with tumor feature controlling.
- Highlighting the strength of our approach in data-scarce environments, in contrast to the data-intensive nature of GAN-based architectures.
- Comprehensive evaluation showcasing the effectiveness of our high-quality synthesized MRIs in enhancing segmentation tasks.

## 2. METHOD

### 2.1. Diffusion models

Diffusion models [14, 15] are a subset of generative models based on a forward-and-backward diffusion process. This stochastic process can be thought of as a parameterized

Markov chain with a fixed number of time steps, denoted as  $T$ . During the forward, Gaussian noise is gradually added to an initial data point  $x_0 \sim q(x_0)$ , following a predefined variance scheduler  $\beta_1, \dots, \beta_T$ :

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

During backward, the model is trained to reverse the forward process starting with a Gaussian noise  $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ , and reconstructing it into the initial data distribution  $q(x_0)$  with learned parameters  $\theta$ . This process can be expressed as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t); \sigma_t^2) \quad (2)$$

During training, the model attempts to predict the added noise to  $x_0$ , denoted as  $\epsilon$ , and extracts it at each step from  $x_t$  to recreate the original sample. To learn the parameters  $\theta$ , such that  $p_\theta(x_{t-1}|x_t)$  approximates  $q(x_{t-1}|x_t)$ , maximum likelihood estimation with variational inference is used, that is similar to VAEs [9] maximizing the evidence lower bound (ELBO). The loss function is finally defined as the mean squared error between the added noise  $\epsilon$  and the predicted  $\hat{\epsilon}$ .

Latent diffusion models (LDM) [11] represent a variant of diffusion models that introduces a two-stage process. Initially, data is projected into a lower-dimensional latent space, typically learned through an autoencoder. The diffusion model then operates in this latent space, generating new latent variables. These latents are subsequently transformed back into pixel space using a decoder.

### 2.2. Slice-based latent diffusion model (SBLDM)

We propose a new method in response to the difficulties encountered in training 3D diffusion models, arising from the considerable computational expenses and memory constraints, as well as the need for substantial data quantities to avoid overfitting. Our methodology leverages a 2-dimensional VAE with a positional embedder to encode the volumetric data in a slice-by-slice manner. Decomposing volumes into individual slices enables the construction of a larger 2D dataset with increased variance and greater diversity in modes. The accurate positioning of each slice facilitates the use of 2D autoencoders for 3D volume generation, enhancing the autoencoder’s capacity to focus on individual slices, and leading to improved generalization.

Our architecture is based on a latent diffusion model that jointly generates 3D MRI volumes and the corresponding tumor mask (Figure 1). A positional embedder  $l(\cdot)$  is introduced in our architecture, allowing it to acquire an understanding of the relative position of each slice within the volume. This additional layer of supervision equips the autoencoder with more spatial awareness. The encoder  $\mathcal{E}$  is modeled as conditional distribution  $q_\phi(z_i|x_i, m_i, l(i))$  where  $x_i, m_i \in \mathbb{R}^{W \times H}$  are the image and its corresponding segmentation mask at slice  $i \leq D$ , and  $l(i)$  is the embedding

of the slice  $i$ . The encoder projects samples into a lower-dimensional representation  $z_i \in \mathbb{R}^{W' \times H'}$  that encapsulates common characteristics between the image and its associated segmentation mask for a given slice. On the other hand, the decoder  $\mathcal{D}$  can be seen as conditional joint distribution  $p_\theta(x_i, m_i | z_i, l(i))$  that reconstructs the image and masks pairs given the latent representations and the relative position of the slice. Subsequently, all these individual  $z_i$  are amalgamated to form a 3D latent space, denoted as  $Z = \cup_i z_i \in \mathbb{R}^{W' \times H' \times D}$ . Subsequently, a diffusion model is trained to capture not only the broader latent variable distribution but also the implicit volumetric dimension introduced through the concatenation of the latent representations.

### 2.3. Conditioning on the tumor characteristics

We further propose to control our model based on tumor characteristics, allowing us to control the size, shape, and relative position of the tumor. This conditioning also serves as regularization to our model, improving supervision and mitigating overfitting. Additionally, this conditioning helps address scenarios where models might generate tumor-free volumes. By specifying the position, we reinforce the constraint of adding a visible tumor to the synthesized data. The tumor’s size and shape are quantified through parameters such as voxel volume, surface area, and sphericity. Meanwhile, its relative position is determined by the coordinates of its center of mass  $(x, y, z)$  and its dimensions  $(w, h, d)$ , collectively forming a bounding box around the tumor. To enable this level of control, we leverage a conditioning vector, which is passed through the Multilayer Perceptron  $\tau$  (Figure 1) to encode these parameters into a feature vector. This feature vector is subsequently fused with the main latent representation  $Z$  during the diffusion process using a scale-shift norm.

## 3. EXPERIMENTATIONS

### 3.1. Dataset

We evaluate the efficacy of our proposed method using the publicly available dataset: BRAIn Tumor Segmentation (BRATS2022) [16] proposes multi-modal MRIs with a volume shape of  $240 \times 240 \times 155$  and a voxel resolution of  $1 \times 1 \times 1 \text{ mm}^3$ . The images are skull-stripped and co-registered to the same anatomical template. The ground truth segmentation masks are provided for the tumor core (TC), enhancing tumor (ET), and whole tumor (WT) regions forming three tumor labels. In our experiments, we only consider FLAIR modality and the WT region.

### 3.2. Implementation details

We deliberately limited our training set to only 100 volumes to simulate a data-scarce scenario, and evaluations are made

on another set of 100 volumes. To accommodate memory limitations for comparative methods, all volumes were resized to  $192 \times 192 \times 96$  dimensions. We employ a VAE as an autoencoder with a downsampling factor of 4 and a U-Net [17] for the diffusion. Our architecture excludes attention modules and utilizes only one residual block per resolution. The experiments were conducted on an NVIDIA GeForce RTX A6000 GPU with 48GB of VRAM, using the Adam optimizer. We employed a learning rate of  $1e - 5$ . For the segmentation task, we utilize the nnUNet [18] framework with default settings, including for the standard data augmentation.

### 3.3. Quantitative results

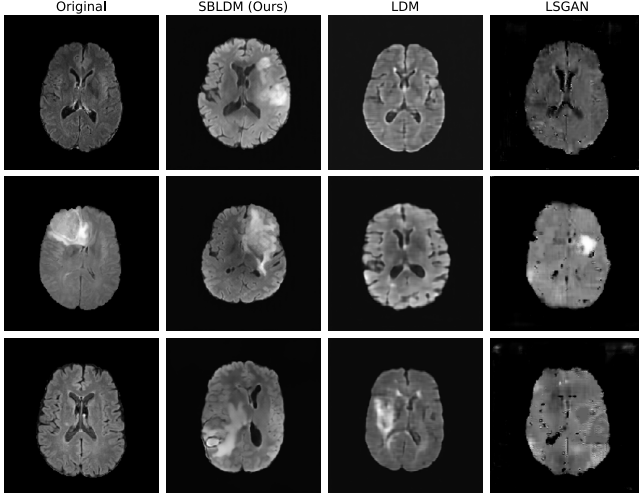
We conducted a quantitative comparison of the generated volumes using our method against two other state-of-the-art techniques. This includes a 3D Least Squares GAN (3D-LSGAN) [19] with a backbone inspired from Deep Convolutional GAN [20] and a 3D version of the original latent diffusion model (3D-LDM) [11], wherein the autoencoder is defined as a VAE-GAN [21] with a downsampling factor of 4. Standard 3D pixel-space diffusion models (DDPM) [15] were excluded due to their memory consumption and unreasonable sampling time. The Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR), as well as the number of parameters and sampling time of each method are chosen for evaluation. Our results demonstrate that our method achieves the highest SSIM of 0.731 and the top PSNR of 21.701 (see Table 1). All accomplished while maintaining an efficient parameter count. Despite its notable efficiency in terms of architecture and sampling time, the 3D-LSGAN experiences the most significant quality impact, primarily due to the data scarcity issue, which makes it impractical for augmentation. GAN-based architectures are notably data-intensive, in contrast to likelihood-based models like 3D-LDM and our approach. To enhance sampling time, we implemented a DDIM sampling scheme [22] with our method, limiting the number of steps to 50. This optimization results in notable time savings, with only a negligible loss in quality.

Methods	PSNR $\uparrow$	SSIM $\uparrow$	#params $\downarrow$	Sampling time $\downarrow$
3D-LSGAN	20.091	0.601	<b>71M</b>	<b>0.002s</b>
3D-LDM	21.034	0.677	728M	13.324s
SBLDM (Ours)	21.466	<b>0.731</b>	159M	30.900s
SBLDM (DDIM)	<b>21.701</b>	0.726	159M	1.965s

**Table 1.** Quantitative performance of the proposed generative models on the BRATS datasets.

### 3.4. Qualitative results

We present a qualitative comparison between samples synthesized using different methods from an axial view in Figure 2. Our method’s samples exhibit a higher level of realism



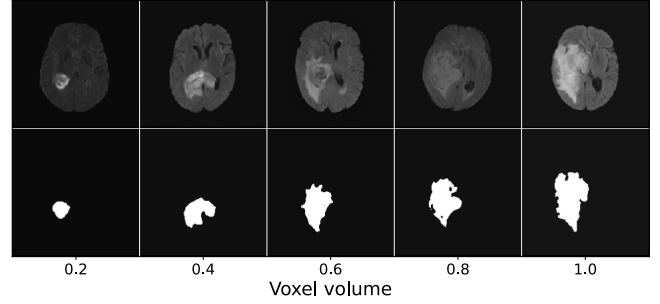
**Fig. 2.** Comparison of images generated using our proposed method and other generative models on the BRATS datasets.

and fine-grained details, being nearly indistinguishable from real volumes. The mode-coverage is further enriched by the conditioning we offer during the generation process. In contrast, the 3D-LDM’s autoencoder trained with only 100 volumes, produces images that are slightly blurrier and lack fine-grained details, indicative of autoencoder underfitting. As for the 3D-LSGAN, the results are subpar, with minimal brain structure and limited details, accompanied by visible image artifacts. Given the scarcity of data, mode collapse is challenging to mitigate at this stage. We do not present images from sagittal and coronal views due to their low quality on the BRATS dataset, rendering them less informative. In Figure 3, we illustrate some conditioned generations of MRIs and corresponding masks using our method. By varying the voxel volume parameter, we generate tumors ranging from small to large while staying within the same brain region. These results demonstrate not only the brain’s variability with each generation but also our method’s successful adherence to size and position constraints.

### 3.5. Using synthetic volumes on a segmentation task

We conducted segmentation model training using synthetically generated data, highlighting their potential as effective data augmentors. Our augmentation pipeline involves data generation from the restricted training set and its combination with the synthetically generated images, thereby creating an augmented dataset, as outlined in [3].

To evaluate our approach, we initially report results with the original restricted dataset and its augmented version with a factor of  $\times 5$ . Then, we enlarge the real dataset with an additional 100 synthesized volumes, generated through each respective method. We also train the nnUNet using only syn-



**Fig. 3.** Figure showcasing volumes with varying tumor Sizes, ranging from 0.0 to 1.0. The tumor position is fixed.

thetic volumes to measure their contribution and individual impact. Our findings, presented in Table 2, underscore the ability of our synthesized images to improve the segmentation task results, achieving a DSC score of 0.815 and IoU of 0.715. This significantly outperforms the corresponding scores for other methods. Notably, the GAN-based architecture appears to deteriorate the results compared to the baseline. This decline can be attributed to the poor quality of the volumes and the occurrence of mode-collapse. Furthermore, we combine both standard augmentation approaches and our synthetic volumes, which leads to even more substantial improvements. We set the number of synthetic volumes to 100 (factor  $\times 2$ ), as we observe that beyond this threshold, the improvement in DSC reaches a plateau. This observation can be attributed to the limited number of modes covered by the synthetic images.

Methods		DSC $\uparrow$	IoU $\uparrow$
Real volumes		0.714 $\pm$ 0.05	0.592 $\pm$ 0.04
Augmented volumes ( $\times 5$ )		0.806 $\pm$ 0.02	0.716 $\pm$ 0.02
Synth only	3D-LSGAN	0.355 $\pm$ 0.13	0.302 $\pm$ 0.11
	3D-LDM	0.529 $\pm$ 0.07	0.401 $\pm$ 0.06
	SBLDM (ours)	<b>0.673<math>\pm</math>0.04</b>	<b>0.551<math>\pm</math>0.04</b>
Real + Synth	3D-LSGAN	0.623 $\pm$ 0.08	0.514 $\pm$ 0.07
	3D-LDM	0.705 $\pm$ 0.03	0.583 $\pm$ 0.02
	SBLDM (ours)	<b>0.815<math>\pm</math>0.02</b>	<b>0.715<math>\pm</math>0.02</b>
Real + SBLDM synth + Aug.		<b>0.834<math>\pm</math>0.01</b>	<b>0.739<math>\pm</math>0.01</b>

**Table 2.** Quantitative performance of the segmentation task in Dice coefficient (DSC), and Intersection-over-Union (IoU).

## 4. CONCLUSION

In this paper, we present a slice-based latent diffusion model for the joint synthesis of 3D MRI volumes and their seg-

mentation masks in data-scarce regimes. Our approach offers practical advantages, requiring fewer computational and memory resources compared to traditional pixel-space and standard latent diffusion models. Our future work will focus on adapting our approach for multi-modal MRI synthesis.

## 5. REFERENCES

- [1] Alexander Selvikvåg Lundervold and Arvid Lundervold, “An overview of deep learning in medical imaging focusing on mri,” *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.
- [2] Tongxue Zhou, Stéphane Canu, Pierre Vera, and Su Ruan, “Latent correlation representation learning for brain tumor segmentation with missing mri modalities,” *IEEE Transactions on Image Processing*, vol. 30, pp. 4263–4274, 2021.
- [3] Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, and Su Ruan, “Deep learning approaches for data augmentation in medical imaging: A review,” *Journal of Imaging*, vol. 9, no. 4, pp. 81, 2023.
- [4] Ying Song, Shuangjia Zheng, Liang Li, Xiang Zhang, Xiaodong Zhang, et al., “Deep learning enables accurate diagnosis of novel coronavirus (covid-19) with ct images,” *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 18, no. 6, pp. 2775–2780, 2021.
- [5] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial networks,” *arXiv preprint arXiv:1406.2661*, 2014.
- [6] Jinbao Wang, Guoyang Xie, et al., “Fedmed-gan: Federated domain translation on unsupervised cross-modality brain image synthesis,” *Neurocomputing*, vol. 546, pp. 126282, 2023.
- [7] Yizhou Chen, Xu-Hua Yang, Zihan Wei, Ali Asghar Heidari, et al., “Generative adversarial networks in medical image augmentation: a review,” *Computers in Biology and Medicine*, p. 105382, 2022.
- [8] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin, “Which training methods for gans do actually converge?,” in *International conference on machine learning*. PMLR, 2018, pp. 3481–3490.
- [9] Diederik P Kingma and Max Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [10] Prafulla Dhariwal and Alexander Nichol, “Diffusion models beat gans on image synthesis,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.
- [11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695.
- [12] Zolnamar Dorjsembe, Hsing-Kuo Pao, Sotavilan Odonchimed, and Furen Xiao, “Conditional diffusion models for semantic 3d medical image synthesis,” *arXiv preprint arXiv:2305.18453*, 2023.
- [13] Lan Jiang, Ye Mao, Xiangfeng Wang, Xi Chen, and Chao Li, “Cola-diff: Conditional latent diffusion model for multi-modal mri synthesis,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 398–408.
- [14] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.
- [15] Jonathan Ho, Ajay Jain, and Pieter Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
- [16] Ujjwal Baid et al., “The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification,” *arXiv preprint arXiv:2107.02314*, 2021.
- [17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [18] Fabian Isensee et al., “nnu-net: Self-adapting framework for u-net-based medical image segmentation,” *arXiv preprint arXiv:1809.10486*, 2018.
- [19] H Xie, RYK Lau, W Zhen, and SP Smolley, “Least squares generative adversarial networks,” *arXiv*, 2016.
- [20] Alec Radford, Luke Metz, and Soumith Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1511.06434*, 2015.
- [21] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, and Ole Winther, “Autoencoding beyond pixels using a

learned similarity metric. corr abs/1512.09300 (2015),”  
*arXiv preprint arxiv:1512.09300*, 2015.

- [22] Jiaming Song, Chenlin Meng, and Stefano Ermon, “De-noising diffusion implicit models,” *arXiv preprint arXiv:2010.02502*, 2020.