CONTROLLABLE LATENT DIFFUSION-BASED 3D BRAIN TUMOR SEGMENTATION: WITH SYNTHETIC LABEL GENERATION AND DETAILED VARIANCE MAP

Xinyang Liu[1], Pengfei Guo[2], Daniel Capellán-Martín[1,3], Zhifan Jiang[1], Holger R. Roth[2], Austin Tapp[1], Maria J. Ledesma-Carbayo[3], Syed Muhammad Anwar[1,4], Marius George Linguraru[1,4]

[1]Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Hospital, Washington DC
[2]NVIDIA, Santa Clara, CA
[3]Biomedical Image Technologies, ETSI Telecomunicación, Universidad Politécnica de Madrid, and CIBER-BBN, ISCIII, Madrid, Spain
[4]School of Medicine and Health Sciences, George Washington University, Washington DC 20010, USA

2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI) | 979-8-3315-2052-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ISBI60581.2025.10981256

ABSTRACT

Approaches based on the Denoising Diffusion Probabilistic Model (DDPM) have shown promise for directly generating segmentation maps from medical images. However, denoising in the original image space limits the application of DDPM to 2D images. We present a latent diffusion model-based segmentation method (LDM-seg) to directly generate multi-label segmentation maps from 3D medical images, such as multisequence magnetic resonance imaging (MRI). A distinctive aspect of our approach is utilizing ControlNet to apply MRI as a conditioning factor to control the generation process. Trained and validated on the BraTS 2023 Adult Glioma dataset, we show LDM-seg outperforms state-of-the-art methods, including nnU-Net and MedNeXt. In addition to segmentation, the method can be used to generate an unlimited number of realistic brain tumor masks, which are typically required as conditions for generating synthetic brain MRI with tumors. Further, the method can also produce a detailed variance map of predicted segmentations.

Index Terms— latent diffusion model, ControlNet, segmentation, BraTS, magnetic resonance imaging

1. INTRODUCTION

With wide intrinsic heterogeneity in appearance, shape, and histology, brain tumors such as glioblastomas (GBM) and diffuse astrocytic gliomas are challenging to diagnose and treat, although they represent the most common malignant primary tumors of the central nervous system in adults [1]. Automated and quantitative imaging analysis tools for accurate segmentation of brain tumors from magnetic resonance imaging (MRI) can aid clinicians in the diagnosis and treatment of these tumors.

With the vast advancements in deep learning techniques, there has been tremendous success in automatic segmentation of brain tumors. In this field, a prominent effort is the Brain Tumor Segmentation (BraTS) Challenge, which is an ongoing annual event that has been held since 2012 [1]. The winning method of BraTS 2020 was based on nnU-Net [2], which analyzes the training dataset and automatically configures a matching U-Net-based [3] segmentation pipeline. The winning method of BraTS 2021 was based on nnU-Net with some optimizations of the network's pipeline [4]. An ensemble of three different methods, including nnU-Net, was introduced at BraTS 2022 and ranked first. The winning method of BraTS 2023 Adult Glioma sub-challenge [5] also used an ensemble of three different methods: standard nnU-Net, Swin UNETR [6] and the winning method of BraTS 2021. Specifically, they incorporated methods to augment data with synthetic images. Based on the past experiences, nnU-Net has been used as the baseline of these winning methods because of its accuracy, robustness, and ease of use. It is also clear that an ensemble of different architectures is a winning strategy.

Recently, diffusion models have achieved impressive results in image synthesis. Unlike generative adversarial networks, diffusion models do not face issues such as training instabilities and mode-collapse. The latent diffusion model (LDM) is a diffusion model variation employing pretrained autoencoders to better model image distributions in the latent space, while greatly reducing the need for excessive computational resources [7]. In medical imaging, generative models have been used to generate synthetic healthy brain MRI [8], abdominal CT with tumors [9], and brain MRI with tumors [5,10,11]. The synthetic brain MRI with tumors can be utilized for data augmentation to enhance tumor segmentation accuracy [5,12]. To generate such synthetic brain MRI with tumors, synthetic brain tumor label maps are often required as conditional inputs [5,10-12].

Besides generating MR images, diffusion models are used to directly generate segmentation maps with MRIs as conditioning priors [13-18]. Previous approaches relied on the Denoising Diffusion Probabilistic Model (DDPM) [19], which performs the diffusion process in the original image

space. MRIs were treated as conditions that were concatenated [13] or dynamically encoded [14] with the segmentation maps during the denoising process. While showing promising results, these approaches usually work only with 2D MRI slices because of the high computational resources required by DDPM.

In this work, we propose to use LDM to directly generate 3D multi-label segmentation maps of adult glioma from multisequence MRI. A unique feature of our approach is to use ControlNet [20] to impose MRI as conditions during the generation. ControlNet adds precise and specific control during text-to-image generation [20]. Trained and validated on the BraTS 2023 Adult Glioma datasets, we show that our LDM-seg method outperforms the two state-of-the-art (SOTA) non-diffusion-based methods, namely nnU-Net [2] and MedNeXt [21]. In addition to providing accurate segmentations, we show that LDM-seg can generate unlimited brain tumor labels, which can be used as conditions for generating synthetic brain images with lesions [5,10-12]. Because of the stochastic nature of diffusion models, once trained, multiple variants of segmentations can be generated to form a detailed variance map, a byproduct of LDM-seg for better interpretability of segmentation results.

2. METHODS

2.1. Method Overview

The overview of our LDM-based segmentation method is depicted in Fig. 1. It includes three separate phases. The first step (red box) is to train a compression model using only the ground truth labels. The compression model ensures that any label image can be encoded ($\mathcal{E}$) into a latent space and accurately reconstructed through a decoder ($\mathcal{D}$). Based on the pretrained compression model, the second step (green box) is to train an unconditional diffusion model in the latent space also using only the ground truth labels. Combined with the compression model, the unconditional diffusion model is supposed to generate random but realistic brain tumor masks. Based on the pretrained compression model and the unconditional diffusion model, the third step (yellow box) is to train a ControlNet using ground truth labels as input to $\mathcal{E}$ and multisequence 3D MR images as conditions. The ControlNet is attached to each encoder level of the denoising U-Net. The parameters of the pretrained denoising U-Net are frozen, and the ControlNet is trained on a trainable copy of the denoising U-Net with zero convolution layers. For inference, the noisy segmentation map is denoised in the latent space with testing MR images as conditions. It is then upscaled to the image space with pretrained $\mathcal{D}$ to generate the final segmentation. Each inference creates a slightly different segmentation prediction.

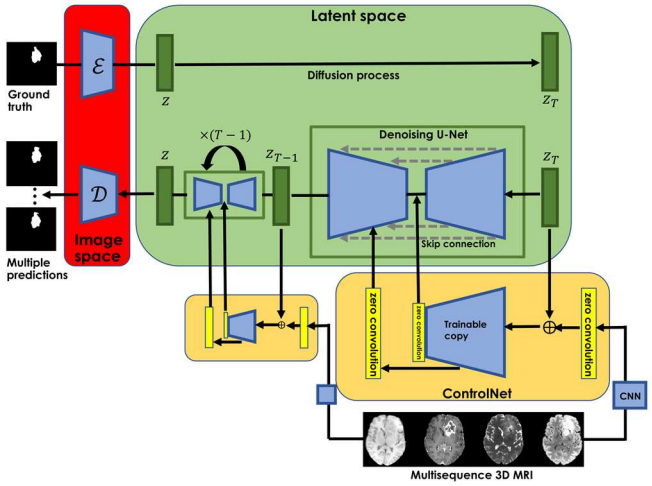

Fig. 1. Method overview. Training the compression model (red box) and the unconditional diffusion model (green box) only require ground truth segmentations. MRI are used as conditions when training a ControlNet (yellow box) based on pretrained compression model and the unconditional diffusion model.

2.2. Data

The publicly available dataset of the Adult Glioma sub-challenge of BraTS 2023 [1] was used. For each case, multisequence MR images were considered including T1-weighted, post-contrast T1-weighted, T2-weighted, and T2 Fluid Attenuated Inversion Recovery (FLAIR). All images were pre-processed following the standard pipeline, i.e., co-registered to the same anatomical template, interpolated to the same resolution (1 mm^3 with image size of 240×240×155), and skull stripped. Ground truth annotations of brain tumors were provided by expert neuroradiologists.

For all methods discussed in this work, training and validation were performed using the BraTS 2023 training dataset (1,251 cases), and validation dataset (219 cases), respectively. As a preprocessing step, we centrally cropped the images and labels to a spatial resolution of 244×244×144. The resulting segmentations from LDM-seg were padded to the spatial resolution of 240×240×155 to be able to submit to the Synapse (www.synapse.org) for validation.

2.3. Compression Model Training

As shown in Fig. 1, training the LDM can be decomposed into training a compression model and training a diffusion model in the latent space. Our perceptual compression model consists of an autoencoder trained by the combination of $\mathcal{L}_1$ loss, Kullback-Leible regularization, perceptual loss and a patch-based adversarial loss. Given a 3D segmentation map $x \in \mathbb{R}^{H\times W\times D}$, the encoder $\mathcal{E}$ encodes x into a latent representation $z = \mathcal{E}(x)$, and the decoder $\mathcal{D}$ reconstructs the segmentation map from the latent z, giving $\tilde{x} = \mathcal{D}(z) =$

$\mathcal{D}(\mathcal{E}(x))$, where $z \in \mathbb{R}^{h \times w \times d}$. The encoder downsamples the image by a factor $f = H/h = W/w = D/d$.

For training the compression model, we set the downsampling factor to be $f = 4$, i.e., two downsamplings with number of channels to be [32, 96, 192]. The output of the encoder was a latent representation of size 61×61×36. The training data was augmented with random flips along the three axes. This augmentation was not used for training the diffusion model and the ControlNet. The model was trained with the Adam optimizer, a learning rate of 10^{-5} and a batch size of 2. The model was trained for 250 epochs.

2.4. Unconditional LDM Training

As probabilistic models designed to learn a data distribution $p(x)$, diffusion models gradually denoise a normally distributed variable, which can be viewed as learning the reverse process of a fixed Markov Chain of length T. Let x_t be a noisy version of the input x, training a diffusion model can be considered as training an equally weighted sequence of denoising autoencoders $\epsilon_\theta(x_t, t) = 1 \ldots T$ to predict a denoised variant of x_t, by minimizing the objective

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),}\,[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2], \quad (1)$$

with t uniformly sampled from $\{1, \ldots, T\}$.

For LDM with pretrained compression model, the input x is downsampled to a low-dimensional latent space in which high-frequency details are absent. Compared with the original image space, this space can train in a lower dimension and thus computationally more efficient. The objective now is

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),}\,[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2]. \quad (2)$$

As in [7], we employed a U-Net architecture for $\epsilon_\theta(\cdot, t)$ with layers of the encoder skip connected with layers of the decoder. Because $\mathcal{E}$ and the diffusion process are fixed, z_t are determined during training, and samples from $p(z)$ are decoded to image space through $\mathcal{D}$.

For training the unconditional diffusion model, we used a scaled linear beta scheduler with T=1,000 steps. The model was trained using a mean squared error (MSE) loss and a batch size of 2. We used the Adam optimizer and a learning rate of 10^{-4}. The model was trained for 250 epochs.

2.5. ControlNet Training

By adding spatial conditioning controls, ControlNet [20] has significantly enhanced the ability to customize pretrained diffusion models. ControlNet preserves the generative capabilities of the pretrained model by locking its parameters, while making a trainable copy of the model's encoding layers. The locked and trainable copies are interconnected using zero convolution layers, i.e., 1×1 convolution with both weight and bias initialized to zero. The convolution weights progressively grow from zero and ensure no harmful noise could affect finetuning of the original diffusion model.

As shown in Fig. 1, the noisy segmentation in the latent space z_t and the four MRI sequences are input to the ControlNet. The trainable copy of the encoder from the denoising U-Net is trained with the latent representations of ground truth labels together with MR images, which was encoded to the same latent space with a convolutional network that was trained jointly with the ControlNet. The output features of ControlNet are incorporated with the middle blocks and the decoder of the denoising U-Net through zero convolution layers.

For training the ControlNet, images were range scaled from [0, 99.5] intensity percentile to [-1, 1]. The model was trained with an MSE loss and a batch size of 2. We used the Adam optimizer and a learning rate of 10^{-4}. The model was trained for 700 epochs. The scaling factor (i.e., $1/\mathrm{SD}(z)$) we obtained during training the diffusion model was used in training the ControlNet.

For inference, the testing MRIs were used as input to the trained ControlNet, and the same scheduler with T=1,000 steps was used for generating the output segmentation map. Our implementation was based on MONAI Generative Models [22], running on a high-performance computing cluster node with NVIDIA H100 GPUs. The hyper-parameters used in this study were empirically determined.

An advantage of diffusion-based methods is that once trained, the model can generate infinite number of plausible variants of predictions, which can then be ensembled for better performance. The LDM-seg was trained with 3 channels, representing the 3 labels of adult glioma considered in BraTS 2023. Each channel was considered as a binary segmentation, and the output of each channel contained positive floating numbers. To ensemble multiple predictions, we averaged them and thresholding with 0.5 as was in [13].

3. EXPERIMENTS AND RESULTS

3.1. Brain Tumor Label Generation

We randomly split the BraTS2023 Adult Glioma training dataset (only ground truth label maps) into Dataset A (1,001 label maps) and Dataset B (250 label maps). We trained the compression model and unconditional diffusion model using Dataset A. Using the trained models, we generated 250 synthetic brain tumor label maps and compared them with Dataset B using Fréchet inception distance (FID), as shown in Table 1 [23]. As a reference, we compared 250 real brain tumor maps (randomly selected from Dataset A) with Dataset B. To interpret the difference in FID scores, we further compared 60 real Adult Glioma label maps (randomly selected from Dataset B) with label maps from a different cohort, i.e., the BraTS2024 Sub-Sahara-Africa Adult Glioma training dataset, which contains 60 cases, each with the defined 3 labels. Table 1 shows our method can generate

realistic synthetic adult glioma label maps. Figure 2 shows examples of LDM generated glioma label maps.

Table 1. Fréchet inception distance between datasets I and II.

Dataset I	Dataset II	FID
250 LDM generated glioma (trained using Dataset A)	250 real glioma (Dataset B)	1.90
250 real glioma (randomly selected from Dataset A)	250 real glioma (Dataset B)	1.96
60 real SSA glioma	60 real glioma	2.64

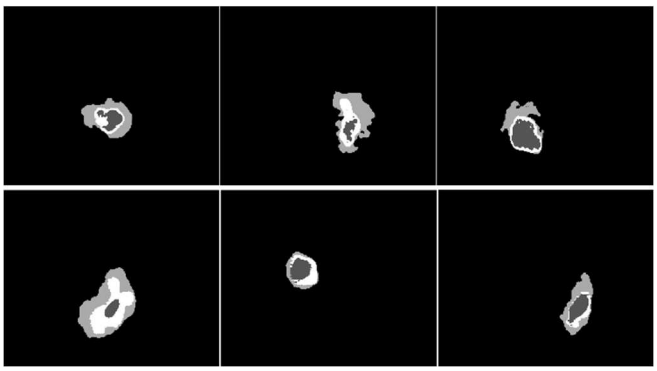

Fig 2. Examples of real Adult Glioma label maps (top) and generated label maps (bottom).

3.2. Segmentation

We compared our segmentation results with and two non-diffusion-based methods: nnUNet [2] and MedNeXt [21]. All methods were trained on the BraTS2023 Adult Glioma training dataset using all 1,251 cases (i.e., no cross-validation to be consistent with how we trained the LDM-seg) and validated on the validation dataset (219 cases) by submission through the Synapse evaluation platform. Following BraTS2023, we used lesion-wise Dice similarity coefficient (LW-DSC) and lesion-wise 95% Hausdorff distance (LW-HD95) to evaluate segmentation on 3 sub-regions of tumor, i.e., the enhancing tumor (ET), the tumor core (TC), and the whole tumor (WT).

Tables 2 and 3 show the segmentation results. LDM-seg outperformed nnU-Net and MedNeXt significantly for WT segmentation. LDM-seg's better performance, especially on LW-HD95, may be caused by its much lower number of false positives (Table 3). The LDM-seg results were generated using an ensemble of 30 predictions. Based on our experiments, the performance of LDM-seg improved with a greater number of predictions used for ensemble, although its performance plateaued after ensemble of 30 predictions. Figure 3 shows LDM-seg's result qualitatively and a detailed variance map, indicating where the model is uncertain with its predictions.

Table 2. Lesion-wise Dice score (best in bold) and p-value between LDM-seg and other.

Model	ET	TC	WT	Mean
nnU-Net	0.725±0.318 (0.537)	0.835±0.224 (0.353)	0.838±0.210 (0.031)	0.799
MedNeXt	**0.764**±0.295 (0.139)	**0.843**±0.219 (0.116)	0.816±0.230 (<0.01)	0.808
LDM-seg	0.736±0.269	0.823±0.234	**0.867**±0.168	**0.809**

Table 3. Lesion-wise HD95 (best in bold), p-value between LDM-seg and other, and number of false positives for WT.

Model	ET	TC	WT	Mean	FP_WT
nnU-Net	67±120 (<0.01)	27±71 (0.470)	38±79 (<0.01)	44	0.228
MedNeXt	50±107 (0.612)	24±67 (0.824)	47±87 (<0.01)	40	0.297
LDM-seg	**46**±105	**23**±70	**23**±59	**31**	**0.087**

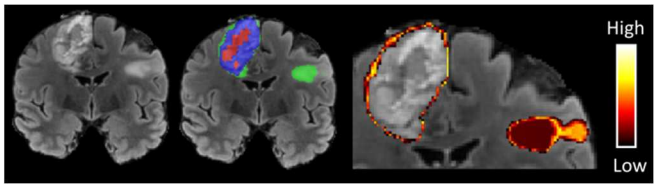

Fig. 3. Original T2-FLAIR (left) overlaid with LDM-seg's result (middle; red: ET, red+blue: TC, all labels: WT). Our method generates detailed variance map (right) illustrating areas with less certainty (i.e., high variance).

4. DISCUSSION AND CONCLUSION

A limitation of having a compression model is that some image details will be lost due to the encoder-decoder architecture. Another limitation of the LDM approach is the slow inference time compared with conventional methods. In the future, this issue may be mitigated by using consistency models [24]. In conclusion, we presented LDM-seg, a method to directly generate 3D multi-label segmentation maps from multisequence MRI. We showed LDM-seg outperformed SOTA methods including nnU-Net and MedNeXt on the BraTS 2023 Adult Glioma datasets. This method can be used to generate unlimited number of realistic tumor masks. It can also produce detailed variance maps along with segmentation. Although demonstrated with brain tumor segmentation, LDM-seg can be extended to other segmentation tasks and modalities. Given these benefits, LDM-seg may be considered as one of the baseline models for ensemble to achieve SOTA performance in a brain tumor segmentation challenge, such as BraTS.

5. ETHICAL COMPLIANCE

The data used for this study is from the publicly available datasets of the Adult Glioma sub-challenge of BraTS 2023 [1]. Ethical approval was not required as confirmed by the license attached with the open-access data.

6. ACKNOWLEDGMENTS

Partial support for this work was provided by the NIH National Cancer Institute award UH3 CA236536, the Spanish Ministerio de Ciencia e Innovación, the Agencia Estatal de Investigación, NextGenerationEU funds, under grants PDC2022-133865-I00 and PID2022-141493OB-I00, and EUCAIM project co-funded by the European Union (Grant Agreement #101100633). The authors gratefully acknowledge the Universidad Politécnica de Madrid (www.upm.es) for providing computing resources on the Magerit Supercomputer.

7. REFERENCES

[1] U. Baid, S. Ghodasara, M. Bilello, et al., "The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogeomic classification," *arXiv:2107.02314*, 2021.

[2] F. Isensee, P.F. Jaeger, S.A. Kohl, et al., "nnU-Net: method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203-211, 2021.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *arXiv:1505.04597*, 2015.

[4] H.M. Luu, and S.H. Park, "Extending nn-unet for brain tumor segmentation," in *Proceedings of MICCAI Brainlesion Workshop*, 2021, pp. 173-186.

[5] A. Ferreira, N. Solak, J. Li, et al., "How we won BraTS 2023 adult glioma challenge? Just faking it! Enhanced synthetic data augmentation and model ensemble for brain tumor segmentation," *arXiv:2402.17317*, 2024.

[6] A. Hatamizadeh, V. Nath, Y. Tang, et al., "Swin unetr: swin transformers for semantic segmentation of brain tumors in mri images," in *Proceedings of MICCAI Brainlesion Workshop*, 2021, pp. 272-284.

[7] R. Rombach, A. Blattmann, D. Lorenz, et al., "High-resolution image synthesis with latent diffusion models," *arXiv:2112.10752*, 2021.

[8] W.H.L. Pinaya, P.-D. Tudosiu, J. Dafflon, et al., "Brain imaging generation with latent diffusion models," in *Proceedings of MICCAI Workshop on Deep Generative Models*, 2022, vol. 13609.

[9] Q. Chen, X. Chen, H. Song, et al., "Towards generalizable tumor synthesis," in *Proceedings of CVPR*, 2024.

[10] J. Wolleb, R. Sandkuhler, F. Bieder, and P.C. Cattin, "The Swiss army knife for image-to-image translation: multi-task diffusion models," *arXiv:2204.02641*, 2022.

[11] Q. Liu, E. Fuster-Garcia, I.T. Hovden, et al., "Treatment-aware diffusion probabilistic model for longitudinal MRI generation and diffuse glioma growth prediction," *arXiv:2309.05406*, 2023.

[12] V. Fernandez, W.H.L. Pinaya, P. Borges, et al., "Can segmentation models be trained with fully synthetically generated data?" in *Proceedings of International Workshop on Simulation and Synthesis in Medical Imaging*, 2022, pp. 79-90.

[13] J. Wolleb, R. Sandkuhler, F. Bieder, et al., "Diffusion models for implicit image segmentation ensembles," *arXiv:2112.03145*, 2021.

[14] J. Wu, R. Fu, H. Fang, et al., "MedSegDiff: medical image segmentation with diffusion probabilistic model," *arXiv:2211.00611*, 2022.

[15] Z. Xing, L. Wan, H. Fu, et al., "Diff-UNet: a diffusion embedded network for volumetric segmentation," *arXiv:2303.10326*, 2023.

[16] J. Zhao and S. Li, "Learning reliability of multi-modality medical images for tumor segmentation via evidence-identified denoising diffusion probabilistic models," in *Proceedings of MICCAI*, 2023.

[17] Y. Fu, Y. Li, S.U. Saeed, et al., "A recycling training strategy for medical image segmentation with diffusion denoising models," *Journal of Machine Learning for Biomedical Imaging*, vol. 2, pp. 507-546, 2023.

[18] T. Ren, A. Sharma, J.H. Rivera, et al., "Re-DiffiNet: modeling discrepancies in tumor segmentation using diffusion models," *arXiv:2402.07354*, 2024.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *arXiv:2006.11239*, 2020.

[20] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv:2302.05543*, 2023.

[21] S. Roy, G. Koehler, C. Ulrich, et al., "MedNeXt: transformer-driven scaling of ConvNets for medical image segmentation," in *Proceedings of MICCAI*, 2023.

[22] W.H.L. Pinaya, M.S. Graham, E. Kerfoot, et al., "Generative AI for medical imaging: extending the MONAI framework," *arXiv:2307.15208*, 2023.

[23] M. Heusel, H. Ramsauer, T. Unterthiner, et al., "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *arXiv:1706.08500*, 2018.

[24] Y. Song, P. Dhariwal, M. Chen, and I. Sutskeverl, "Consistency models," *arXiv:2303.01469*, 2023.