BRAINEDITOR: STRUCTURE-DISENTANGLED BRAIN MRI SYNTHESIS VIA NATURAL LANGUAGE PROMPTED DIFFUSION IMAGE EDITING

Jialin Li[1,†], Dong Wei[2,†], Jiaming Qiu[3,†], Yiwei Ren[1], Pujin Cheng[1,4], Junyan Lyu[1], Terry Tao Ye[1], Yefeng Zheng[2,5,], Xiaoying Tang[1,4,*]*

[1]Electronic and Electrical Engineering, Southern University of Science and Technology, China
[2]Jarvis Research Center, Tencent YouTu Lab, China
[3]Harbin Institute of Technology, China
[4]Electronic and Electrical Engineering, University of Hong Kong, China
[5]Medical Artificial Intelligence Lab, Westlake University, Hangzhou, China

ABSTRACT

Deep learning has advanced medical image synthesis models to tackle data shortages and assist in disease prediction and analysis. In this work, we introduce **BrainEditor**, a novel approach to structure-disentangled brain MRI synthesis via text-prompted diffusion image editing. Unlike most existing methods that often alter multiple anatomical structures in a linked manner, our approach allows precise control over individual structures. This enables the simulation of potentially diverse and complex disease patterns in clinical practice. By integrating textual prompts into the image synthesis process, we enhance BrainEditor's flexibility for user interaction. To achieve the disentangled synthesis, we generate artificially structure-disentangled training image pairs through Poisson image editing, and link text descriptions to image progressions by quantizing volumetric change rates of brain structures into limited intervals. Experiments on the public OASIS-2 dataset show that BrainEditor accurately synthesizes high-quality images following textual instructions and improves the performance of downstream tasks through data augmentation. The code is available at https://github.com/LIKP0/BrainEditor.

***Index Terms*—** Disentangled image synthesis, brain MRI, Alzheimer's disease

1. INTRODUCTION

By synthesizing realistic images for data augmentation, deep synthesis methods tackle the challenge of insufficient training data for deep learning based medical image analysis in healthcare. Recently, many studies have shown convincing results for brain MRI synthesis. Xia et al. [1] generated brain aging images. Similarly, 4D-DANI-Net [2] employed age as a condition factor to synthesize longitudinal MRI data, presenting brain structure alterations associated with aging and dementia. ADESyn [3] simulated the continuous brain changes through different stages of Alzheimer's disease (AD) [4]. DiDiGAN [5] proposed a smooth manifold to realize seamless structure transformation of images from healthy individuals to AD patients. However, these studies assumed synchronous changes for all involved brain structures associated with disease progression. For example, 4D-DANI-Net can generate general AD progressions with synchronous ventricular enlargement and cortical atrophy based on age. However, the pattern of brain structural changes varies among patients [6–8], e.g., some may present only mild while others may show severe ventricular enlargement with the same extent of cortical atrophy. The synchronous assumption makes it difficult to synthesize diverse progression images precisely. Moreover, different pathological changes may manifest at various stages of disease progression, and a single pattern cannot effectively simulate the complexity of these processes.

To address these challenges, we introduce **BrainEditor**, a novel approach to *structure-disentangled* medical image syn-

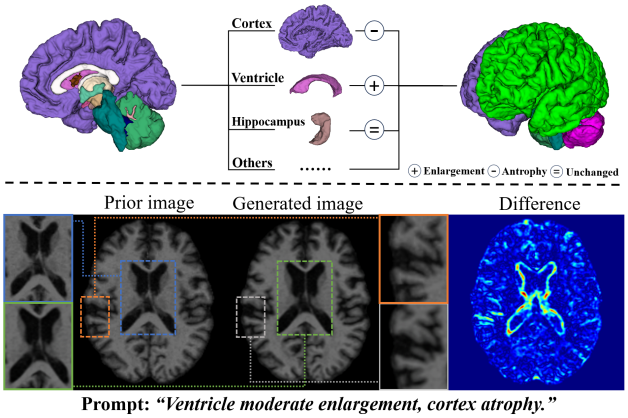

Fig. 1. Top: Unlike most existing methods, our disentangled approach, BrainEditor, modifies individual structures while ensuring harmony in the synthesized image. Bottom: BrainEditor edits prior images following textual instructions.

† Contributed equally. * Corresponding authors.

thesis. BrainEditor enables precise control over the editing of individual anatomical structures. In addition, it integrates text prompts to enhance controllability. We evaluate our approach on the public OASIS-2 dataset [9]. Results demonstrate that BrainEditor can generate high-quality images following textual instructions. Moreover, by generating augmented data, it can improve the performance of downstream tasks.

Figure 1 highlights the advantages of our approach. Most existing methods generate various anatomical structures simultaneously, and adjusting one structure often leads to linked changes in others. This forces image synthesis to follow a single pattern for a trained model. As a result, multiple models must be trained in order to synthesize various combinations of disentangled, heterogeneous structural variations. In contrast, our approach allows separate, precise editing of individual structures and naturally merges them in the synthesized image (Fig. 1 top). In addition, it incorporates textual prompts for synthesis control, enabling flexible generation of different combinations of disentangled structural changes by composing corresponding instructions (Fig. 1 bottom). To achieve the structure-disentangled synthesis with the precise control, this work addresses two main challenges:

How to achieve structure-disentangled synthesis? In other words, how can we edit a specific brain structure without affecting others? With limited training data, we generate pairs of structure-disentangled images using Poisson image editing [10] for training, where only a specific structure exhibits pathological changes from a prior to a target image, with other structures unchanged. These images facilitate the model in learning stable control over individual structures, thereby achieving disentanglement.

How to use natural language to control image editing precisely? Inspired by previous work [11], we link textual descriptions to image variations by calculating and quantifying the volumetric change rates of brain structures. For instance, given an MRI image as reference, the prompt "mild ventricle enlargement" describes a progression image with a (0%, 10%] increase in ventricle volume concerning the reference.

The main contributions of our work are as follows. (1) To the best of our knowledge, we present the first work on structure-disentangled brain MRI synthesis based on Poisson image editing. (2) By quantifying the volumetric change rates of brain structures, we propose a novel approach to text-to-image mapping for disentangled, precise editing. This enables more convenient and user-friendly image synthesis.

2. METHODS

We perform the task of brain MRI synthesis by prompted image editing, where a prior image is modified according to the natural language instructions to yield a synthesized image. Below, we first introduce how to link the image transformations with textual descriptions. Next, we introduce a method to produce structure-disentangled image pairs for training. which is then improved with Poisson image editing. Lastly, we describe the model training process.

Bridging Texts and Images. Inspired by previous research [11], we quantify volume change rates in brain structures between prior and target images to compose text descriptions, such as "ventricle enlarges by 5.48%". However, the exact numbers are challenging for the models to learn, likely due to the deficiency of text encoders in understanding numbers (e.g., 5.48% and 5.49% are similar) and the insufficient samples for each rate. Therefore, we quantize the change rates into limited intervals for each structure of interest and describe each interval with a uniform phrase, to ensure sufficient training samples for each interval and reduce learning difficulty. For a structure of interest, the volume change rate R between the target and prior images is calculated as follows:

$$R = (V_t - V_p)/V_p \times 100\%, \tag{1}$$

where V_t and V_p are the volumes of the structure in the target and prior images. We obtain V_t and V_p from the segmentation masks produced by FreeSurfer [12]. R is then mapped to an extent-describing text according to predefined rules for that structure. For example, a 3% ventricle enlargement between a pair of target and prior images is described by "mild". The mapping rules can be adjusted to accommodate different anatomical structures for their characteristic change patterns.

Artificially Disentangled Training Images. We use longitudinal MRI data for training, including multiple scans of a patient at different time points. These data are grouped into image pairs consisting of a prior and a target image, where the former was captured earlier than the latter. The changes in the structures of interest between a pair are computed to generate a qualitative textual prompt, as described above. Thus, the model is trained to edit the prior image following the prompt and produce the target image. However, we observe that, after training, the model cannot synthesize disentangled images with separate modifications to different structures. We conjecture this is because most images in the experimental dataset present synchronous changes to various structures, prohibiting the model from learning each structure decoupled. To address this issue, we propose an approach to producing artificially structure-disentangled training images.

The pipeline is illustrated in Fig. 2(a). For a structure of interest S in the target image, we crop it out and paste it onto the prior image to create an "S-disentangled" image. This process produces a new target image where only the specified structure changes while others remain intact. Note the "copy-n-paste" target image effectively captures the true associated pathological changes of the specified structure. Correspondingly, we compose a prompt that only describes the change in that structure. Thus, the artificially disentangled images enable the model to learn a one-to-one mapping between the textual descriptions and the disentangled structure.

Poisson Image Editing. Our preliminary exploration found that while the copy-n-paste technique is straightforward and effective, it can introduce noticeable artifacts, particularly at

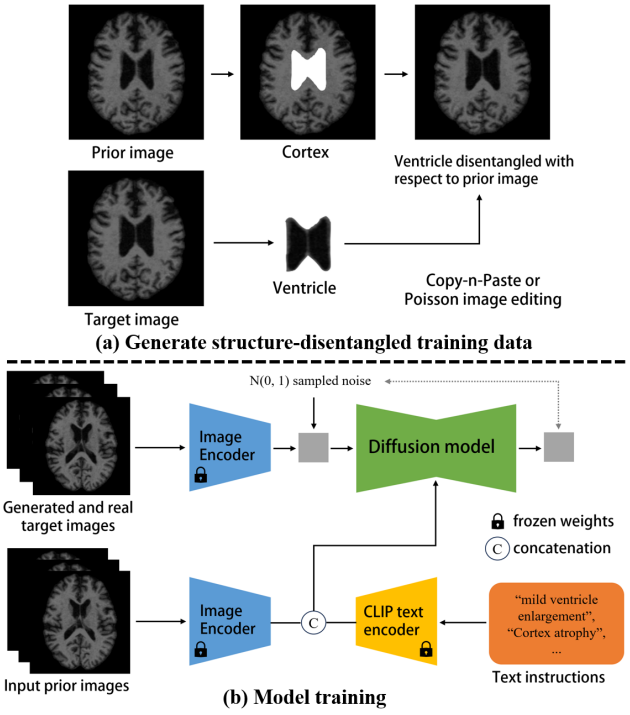

Fig. 2. Overview of the BrainEditor framework. (a) We generate structure-disentangled data with Poisson image editing. (b) We train a conditional latent diffusion model on a mixture of the generated and real data.

the boundaries of structures, leading to a decrease in quality and potential errors in the synthesized images. Therefore, we further improve the artificially disentangled images with Poisson image editing [10] through seamless cloning. Poisson image editing is a classic technique used in image processing to seamlessly blend a source image region into another image by solving Poisson's equation, ensuring smooth transitions and preserving gradient information. In our context (Fig. 2(a)), denote the prior image by f^*, the target image by g, the region (of a structure) to be cloned from g to f^* by Ω and its boundary by $\partial\Omega$. To make the cloned content f smooth (i.e., the content pasted into f^*), Poisson image editing solves the following Poisson equation with Dirichlet boundary conditions [13]:

$$\triangle f = \triangle g \text{ over } \Omega, \text{ with } f|_{\partial\Omega} = f^*|_{\partial\Omega}, \tag{2}$$

where $\triangle$ is the Laplacian operator. Intuitively, Eq. (2) states that f has the same texture as g within Ω, and the same values as f^* on the boundary of Ω. In this way, we obtain artificially disentangled images of improved quality for training.

Training Image Editing Model. We base our model on InstructPix2Pix [14], a state-of-the-art image editing framework that combines a latent diffusion model [15] with a CLIP [16] text condition branch (Fig.2.b). The latent diffusion model enhances the efficiency of diffusion models by operating in the latent space projected by a pretrained variational autoencoder [17], which consists of an encoder ε and a decoder. For a target image x, the diffusion process progressively adds noise to the encoded latent vector $z = \varepsilon(x)$, with the noise level increasing over timesteps $t \in T$. We train a network ϵ_θ to predict the added noise ϵ from the noisy latent vector z_t, with the prior image c_I and text prompt c_T as conditions. The objective function is:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(x),\mathcal{E}(c_I),c_T,\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, t, \mathcal{E}(c_I), c_T\right)\right\|_2^2\right]. \tag{3}$$

We initialize our model with the pretrained weights from Stable Diffusion [15] checkpoints. We find that the CLIP model, pretrained on natural images, is not fully adequate for aligning the medical terminologies of brain structures with the corresponding regions in the MRI images. For example, it cannot correctly associate "ventricle" with the center area of the brain in the image. Our structure-disentangled training data helps address this issue. For example, a pair of ventricle-disentangled images establishes a one-to-one mapping from the term "ventricle" to the specific changes in the images, making the model correctly learn the concept of "ventricle".

The model is trained in two stages. In the first stage, we fine-tune the autoencoder on our training data, significantly improving the quality of reconstructed images. Then, in the second, we fine-tune the diffusion model weights with the autoencoder and CLIP model frozen.

3. EXPERIMENTS AND RESULTS

Dataset and Preprocessing. We use the public OASIS-2 [9] dataset, containing a longitudinal collection of 373 T1-weighted MRI scans of 150 subjects. For each 3D scan, we employ FreeSurfer [12] for registration, skull-tripping, and whole-brain segmentation. Subsequently, 30 center axial slices are extracted, cropped to 192×192 pixels, and resized to 256×256 pixels. This results in 4,620 pairs of real images (after filtering low-quality images), which are divided into training, validation, and testing sets with a ratio of 8:1:1. This work focuses on changes in ventricular enlargement and cortical atrophy, common trends in brain aging and AD. We calculate the volume change rates for the two structures from the segmentation masks, and quantify ventricle enlargement into three intervals: "mild" (0%–10%), "moderate" (10%–20%), and "severe" (>20%). Cortex changes are grouped into only one interval: "atrophy" ($> 0\%$). We generate 7,390 structure-disentangled image pairs from the training set and combine them for training. This allows the model to effectively achieve disentanglement while also learning from the real images with joint anatomical variations.

Implementation. We initialize our model from the pretrained Stable Diffusion v1.5 checkpoint. Both stages are trained on a 40 GB NVIDIA A100 GPU for 250 epochs with the Adam optimizer. In stage 1, we use a fixed learning rate 4.5×10^{-6} and a batch size of 8. In stage 2, we use a fixed learning rate 10^{-4} with a batch size of 32.

Table 1. Quantitative evaluation results of image quality and structural change accuracy.

	PSNR↑	SSIM↑	ACC↑	MAE↓
Comparison to other methods				
DiDiGAN [5]	22.31	0.470	-	-
CycleGAN [18]	29.78	0.858	0.24	9.79
StarGAN [19]	29.27	0.860	0.05	10.97
Pix2Pix [20]	29.95	0.856	0.76	3.52
Ours (Poisson)	32.62	**0.912**	**0.86**	**3.35**
Ablation study				
Not disentangled	**32.86**	**0.912**	0.69	3.57
Copy-n-paste	32.27	0.906	0.70	3.67

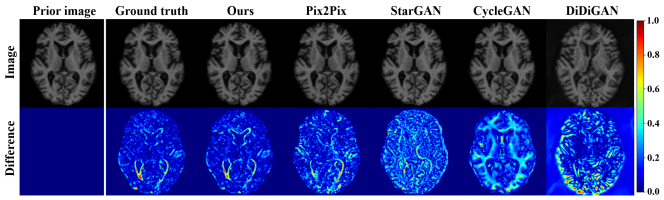

Fig. 3. Example generation results. Our model edits the prior image with the prompt "Cortex atrophy and severe ventricular enlargement." The difference maps show that the structural variations produced by our model closely resemble the ground truth. Best viewed zoomed in with a digital copy.

Evaluation Metrics. For direct evaluation of image quality, we use the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM). To assess whether the model accurately simulates structural changes according to textual prompts, we calculate the accuracy (ACC) of the quantified change interval, and the mean absolute error (MAE) of the volume change rate of the synthesized images evaluated against real target images.

Quality of Synthesized Images. We compare our model with general image translation/editing methods, including CycleGAN [18], StarGAN [19], Pix2Pix [20], and a specialized method in creating representative reference images for AD disease characteristic discovery, DiDiGAN [5]. It is worth noting that all compared methods based on GANs require training multiple models to synthesize images of heterogeneous change intervals. In contrast, our model only needs one model, thanks to the flexibility of textual-description-prompted image editing. Table 1 top charts the quantitative evaluation results. DiDiGAN, CycleGAN, and StarGAN yield bad numbers in ACC and MAE (the segmentation model [21] even fails to segment images synthesized by DiDiGAN). Further investigation reveals that they produce images with almost no effective structural changes (Fig. 3). In comparison, Pix2Pix greatly improves performance, especially in ACC and MAE. Lastly, our model substantially outperforms Pix2Pix in all four metrics. Figure 3 visualizes example results of compared methods. The difference maps indicate that the changes in the ventricle and cortex simulated by our model closely resemble the actual changes.

Table 2. Data augmentation performance on the downstream task of three-class AD classification. AUC: area under the receiver operating characteristic curve.

Metrics	**BaseLine**	**BigAug** [23]	**Pix2Pix** [20]	**HomoAug**	**HeterAug**
F1-score↑	0.924	0.933	0.934	0.931	**0.944**
AUC↑	0.987	0.986	0.990	0.991	**0.993**

Ablation Study. We conduct ablative experiments to validate our proposed structure-disentangled training images. As shown in Table 1 bottom, "Ours (Poisson)" notably boosts the structural editing accuracy in ACC and MAE compared with the "Not disentangled" variant, while maintaining comparable image quality in PSNR and SSIM. This suggests that our structure disentangling with the Poisson image editing enhances the model's understanding of the structures' characteristics and enables precise control.

Data Augmentation for Downstream Task. We also evaluate our model for data augmentation on a downstream task, i.e., three-class classification of AD: cognitively normal, mild cognitive impairment [22], and AD. We synthesize images by editing the real training images illustrating AD progression. Two augmentation settings are used: "HeterAug," which generates images with disentangled changes (e.g., cortical atrophy or ventricular enlargement), and "HomoAug," which combines both changes. We retain the original training images' labels for the synthetic ones. In addition, we compare our method with a baseline model without data augmentation, and two data augmentation methods: Pix2Pix [20] (GAN-based) and BigAug [23] (image transformation based). Table 2 shows the results. By synthesizing disentangled structural progression, HeteAug achieves greater diversity in the generated images than other methods, leading to the greatest improvements in the downstream task.

4. CONCLUSION

We proposed BrainEditor, a conditional latent diffusion model for the structure-disentangled synthesis of brain MRI via text-prompted image editing. Experiments on a public dataset demonstrated its promising performance and flexibility, and validated its disentangling motivation. Future directions include extending BrainEditor for more datasets.

5. ACKNOWLEDGMENT

This study was supported by the National Key Research and Development Program of China (2023YFC2415400); the National Natural Science Foundation of China (T2422012, 62071210); the Guangdong Basic and Applied Basic Research (2024B1515020088); the Shenzhen Science and Technology Program (RCYX20210609103056042); the Guangdong Basic and Applied Basic Research (2021A1515220131); the High Level of Special Funds (G030230001, G03034K003).

6. REFERENCES

[1] Tian Xia, Agisilaos Chartsias, Sotirios A Tsaftaris, and Alzheimer's Disease Neuroimaging Initiative, "Consistent brain ageing synthesis," in *MICCAI*. Springer, 2019, pp. 750–758.

[2] Daniele Ravi et al., "Degenerative adversarial neuroimage nets for brain scan simulations: Application in ageing and dementia," *MIA*, vol. 75, pp. 102257, 2022.

[3] Euijin Jung, Miguel Luna, and Sang Hyun Park, "Conditional GAN with an attention-based generator and a 3D discriminator for 3D medical image generation," in *MICCAI*. Springer, 2021, pp. 318–328.

[4] Philip Scheltens et al., "Alzheimer's disease," *The Lancet*, vol. 397, no. 10284, pp. 1577–1590, 2021.

[5] Siyu Liu et al., "Style-based manifold for weakly-supervised disease characteristic discovery," in *MICCAI*. Springer, 2023, pp. 368–378.

[6] Konstantinos Poulakis et al., "Heterogeneous patterns of brain atrophy in Alzheimer's disease," *Neurobiology of aging*, vol. 65, pp. 98–108, 2018.

[7] Xiaolong Shan et al., "Mapping the heterogeneous brain structural phenotype of autism spectrum disorder using the normative model," *Biological Psychiatry*, vol. 91, no. 11, pp. 967–976, 2022.

[8] Lingyu Liu, Shen Sun, Wenjie Kang, Shuicai Wu, and Lan Lin, "A review of neuroimaging-based data-driven approach for alzheimer's disease heterogeneity analysis," *Reviews in the Neurosciences*, vol. 35, no. 2, pp. 121–139, 2024.

[9] Daniel S Marcus, Anthony F Fotenos, John G Csernansky, John C Morris, and Randy L Buckner, "Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults," *J. Cogn. Neurosci.*, vol. 22, no. 12, pp. 2677–2684, 2010.

[10] Patrick Pérez, Michel Gangnet, and Andrew Blake, "Poisson image editing," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 577–582. 2023.

[11] Deming Wang, "MR image-based measurement of rates of change in volumes of brain structures. Part I: method and validation," *Magnetic resonance imaging*, vol. 20, no. 1, pp. 27–40, 2002.

[12] Bruce Fischl, "Freesurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

[13] Alexander H-D Cheng and Daisy T Cheng, "Heritage and early history of the boundary element method," *Engineering analysis with boundary elements*, vol. 29, no. 3, pp. 268–302, 2005.

[14] Tim Brooks, Aleksander Holynski, and Alexei A Efros, "InstructPix2Pix: Learning to follow image editing instructions," in *CVPR*, 2023, pp. 18392–18402.

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10684–10695.

[16] Alec Radford et al., "Learning transferable visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763.

[17] Diederik P Kingma and Max Welling, "Auto-encoding variational Bayes," *arXiv:1312.6114*, 2013.

[18] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2223–2232.

[19] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *CVPR*, 2018, pp. 8789–8797.

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.

[21] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[22] Angela M Sanford, "Mild cognitive impairment," *Clinics in geriatric medicine*, vol. 33, no. 3, pp. 325–337, 2017.

[23] Ling Zhang et al., "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation," *TMI*, vol. 39, no. 7, pp. 2531–2540, 2020.