CS 191W: Generating Anatomically Realistic 3D Brain MRIs with GAN-Based Models

Lara Selin Seyahi

June 7, 2025

1 Introduction

Magnetic resonance imaging (MRI) plays a vital role in advancing our understanding of neurodegenerative diseases, such as HIV-associated neurocognitive disorder (HAND). HAND is associated with a range of anatomical changes, particularly with atrophy in the gray matter, in areas such as the anterior cingulate gyrus, lateral temporal cortex, motor and sensory areas, the parietal and frontal lobes [1]. These changes in brain structure makes MRI crucial for both research and clinical applications, including the development of AI models for diagnosis and disease progression modeling, as large, diverse datasets are essential for effectively training, validating, and generalizing AI models in medical imaging [2, 3]. However, the collection of real MRI datasets is often constrained by factors such as patient privacy, limited recruitment, and high costs, resulting in a scarcity of available data. Thus, there is a growing need for reliable methods to generate synthetic 3D brain MRIs that can supplement and enhance real-world datasets.

Generative models that can synthesize realistic 3D MRI data are therefore highly valuable, as they can augment limited clinical datasets, preserve patient privacy, and facilitate the development and benchmarking of machine learning algorithms for neuroimaging in HIV and related disorders. In this project, I focused on evaluating the performance of four state-of-the-art generative models (HAGAN, VAEGAN, AlphaGAN, and AlphaWGAN) for the task of generating 3D brain MRIs. By systematically assessing these models, I aimed to determine their effectiveness in producing anatomically accurate and diverse samples that could meaningfully support neuroimaging research on HAND and other neurological conditions.

It was found that quantitatively, the VAEGAN model performed the best. The next stages of this project will focus on evaluating the performance of a diffusion based model and investigating the potential of generative models for generating disease-specific 3D MRIs for HIV-associated neurocognitive disorders.

2 Background

Generative Adversarial Networks (GANs) are a class of deep learning models that leverage a unique framework known as adversarial training. In this setup, two neural networks (the generator and the discriminator) are trained through adversarial training[4].

- The generator learns to create synthetic data (such as 3D brain MRIs) from random noise, attempting to produce samples that are indistinguishable from real data.
- The discriminator, on the other hand, tries to distinguish between genuine samples from the training set and those generated by the generator.

Through this adversarial interplay, the generator progressively learns to create more realistic data, as it attempts to "fool" the discriminator, while the discriminator improves at detecting fakes. This process enables GANs to model complex data distributions and synthesize realistic data samples [4].

In this project, I focus on four GAN variants for generating 3D brain MRI samples:

- **HAGAN** [5]: The Hierarchical Amortized GAN (HAGAN) is a GAN architecture that generates high-resolution 3D medical images by training on low-resolution whole images and high-resolution sub-volumes. This approach enables efficient memory use and anatomically consistent full-image synthesis (Sun et al., 2022)

- **VAEGAN** [6]: The Variational Autoencoder GAN (VAEGAN) combines the VAE structure and adversarial loss. The model is composed of an encoder, which encodes the input into two latent vectors (mean and variance), creating an underlying distribution for the data. The decoder decodes latent vectors back to image. The decoder of the VAE structure also acts as the generator. The discriminator from GAN tries to tell apart real images from images produced by the generator/decoder.

- **AlphaGAN** [7]: alphaGAN is a generative adversarial network framework that uses automated, fully differentiable neural architecture search to optimize both the generator and discriminator networks. By formulating the search as a bi-level minimax optimization, AlphaGAN efficiently discovers high-performing GAN architectures within a vast search space, improving model quality with minimal computational resources.

- **AlphaWGAN** [8]: AlphaWGAN incorporates the Wasserstein-GAN that minimizes a reasonable and efficient approximation of the EM distance.

3 Methods

This project focuses on generating 3D brain MRIs using four generative models (HAGAN, VAEGAN, AlphaGAN, and AlphaWGAN) and systematically evaluating their performance.

3.1 Model Implementation and Sample Generation

All model implementations and evaluation pipelines were based on the open-source `MRIAnatEval` repository [11], which provides pre-trained weights, model architectures, and evaluation scripts for a range of 3D generative models used in neuroimaging.

To generate synthetic 3D brain MRI samples for each model, I wrote custom scripts that leveraged the `MRIAnatEval` codebase. These scripts automated the process of model loading, sample generation, and output saving for HAGAN, VAEGAN, AlphaGAN, and AlphaWGAN. Model training and inference were performed on the Stanford Sherlock high-performance computing cluster. The generated volumes were saved in NIfTI format for standardized downstream evaluation. Ten samples per model has been generated.

3.2 Evaluation Metrics

To assess the quality and anatomical realism of the generated MRIs for each model, two quantitative metrics were employed:

- **3D Multi-Scale Structural Similarity Index Measure (3D MS-SSIM):** This metric extends the traditional SSIM, which measures structural similarity between two images, to three dimensions and multiple spatial scales. It also incorporates ”the variations of viewing conditions”. A higher 3D MS-SSIM indicates that the generated image more closely resembles real MRIs in terms of structural details and perceptual features [9].

- **3D Maximum Mean Discrepancy (3D MMD):** MMD is a statistical test that quantifies the difference between the distributions of real and generated samples. A lower MMD value means the generated images have a distribution more similar to the real MRIs, indicating higher fidelity of the synthesis process [10].

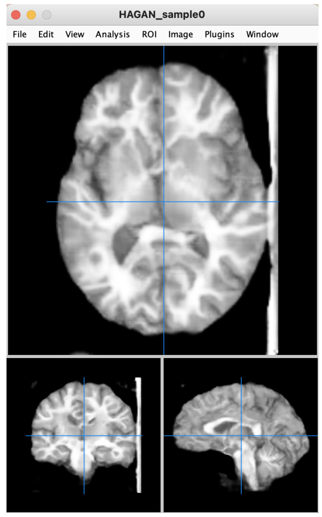

Figure 1: HAGAN Generated Sample

3.3 Data and Preprocessing

The same indices were used for selecting ten real samples across all evaluations to ensure that each set of generated samples was compared to an identical set of real images. The evaluation was conducted on a dataset containing 1,500 NIfTI files, with real MRI data sourced from Wu et al. [12]. Data processing in the study included skull stripping to remove non-brain tissue and affine registration to the SRI atlas to ensure consistent anatomical alignment across scans. The further pre-processing pipeline in my evaluation script included the following steps: intensity clipping, normalization, zeroing low intensities, and, cropping and padding. This standardized preprocessing ensured fair and consistent evaluation across real and generated 3D MRI samples.

4 Results

Table 1: Performance of Generative Models on 3D MRI Synthesis (10 Samples Per Model). Higher SSIM and lower MMD indicate better performance.

Model	**3D SSIM** $\uparrow$	**3D MMD** $\downarrow$
HAGAN	0.561 ($\pm$ 0.00)	775,113 ($\pm$ 3,255)
VAEGAN	0.819 ($\pm$ 0.00)	165,021 ($\pm$ 10,511)
ALPHAGAN	0.049 ($\pm$ 0.00)	2,187,918 ($\pm$ 14,031)
ALPHAWGAN	0.086 ($\pm$ 0.00)	2,794,725 ($\pm$ 26,173)

The samples have been visualized using Mango, which is a viewer for medical research images. The generated sample images per model and Mango's sample image have been added as figures (Fig.1 - 5).

5 Discussion

VAEGAN demonstrated the strongest overall performance according to the evaluation metrics, achieving the highest structural similarity (SSIM $\approx$ 0.82) and the lowest distributional difference (MMD $\approx$ 165,000). HAGAN performed moderately well, with a reasonable SSIM ($\approx$ 0.56) and a low MMD ($\approx$ 775,000); while hierarchical amortization appears beneficial, its performance did not surpass that of VAEGAN.

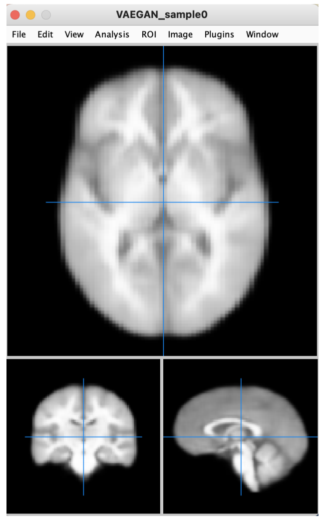

Figure 2: VAEGAN Generated Sample

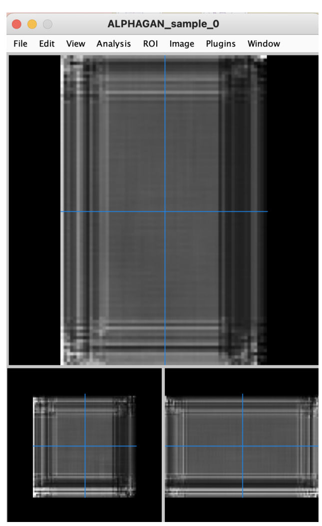

Figure 3: ALPHAGAN Generated Sample

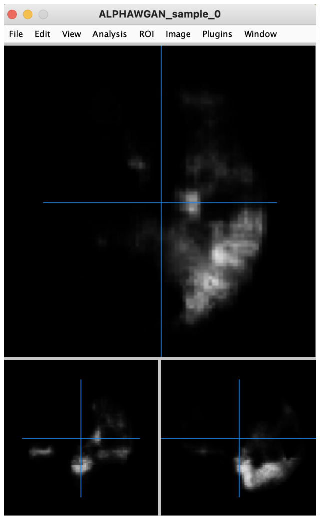

Figure 4: ALPHAWGAN Generated Sample

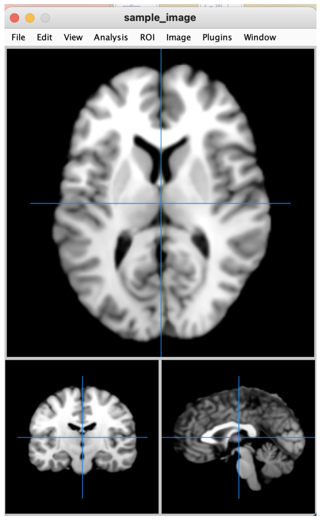

Figure 5: Sample Image from Mango

Interestingly, the visual appearance of the generated images tells a somewhat different story than the numerical results. Images produced by VAEGAN often appear blurrier and smoother, while HAGAN-generated images are noticeably sharper and may seem more perceptually realistic or detailed. This discrepancy can be explained by the well-documented tendency of VAEs to produce smoother outputs, "Variational autoencoders (VAEs) are powerful generative modelling methods, however they suffer from blurry generated samples and reconstructions compared to the images they have been trained on" [13]. In contrast, although HAGAN produces sharper and more detailed images, these details might not perfectly align with the distribution or pixel statistics measured by SSIM and MMD.

AlphaGAN and AlphaWGAN substantially underperformed, with very low SSIM values ($\approx$ 0.05–0.09) and high MMD scores (over 2 million), indicating that the synthetic MRIs generated by these models lacked anatomical realism and failed to match the distribution of real data. These results suggest the possibility of implementation errors or that not all adversarial architectures generalize well to 3D MRI generation tasks.

Overall, VAEGAN produced the best results among the tested models according to the quantitative metrics. However, further improvements may be possible with diffusion-based models, such as PCGM, which could potentially offer better anatomical fidelity and diversity in generated 3D brain MRIs.

6 Future Directions

The next stages of this project will focus on evaluating the performance of a diffusion-based model for 3D brain MRI generation. In addition, future work will investigate the potential of generative models to synthesize disease-specific 3D MRIs, with a particular emphasis on imaging patterns associated with HIV-associated neurocognitive disorders. These directions aim to further improve the anatomical realism and clinical utility of synthetic MRIs, and to explore how advanced generative modeling can contribute to research and diagnosis in neurodegenerative diseases.

References

[1] Ances, B. M., and Hammoud, D. A., *Neuroimaging of HIV-associated neurocognitive disorders (HAND)*, Current Opinion in HIV and AIDS, 9(6):545–551, 2014. https://doi.org/10.1097/COH.0000000000000112

[2] Greenspan, H., Van Ginneken, B., & Summers, R. M., *Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique*, IEEE Transactions on Medical Imaging, 35(5):1153–1159, 2016. https://doi.org/10.1109/TMI.2016.2553401

[3] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A. W., Van Ginneken, B., & Sánchez, C. I., *A survey on deep learning in medical image analysis*, Medical Image Analysis, 42:60–88, 2017. https://doi.org/10.1016/j.media.2017.07.005

[4] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y., *Generative Adversarial Nets*, Advances in Neural Information Processing Systems, 2014. https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[5] Sun L, Chen J, Xu Y, Gong M, Yu K, Batmanghelich K. *Hierarchical Amortized GAN for 3D High Resolution Medical Image Synthesis*, IEEE J Biomed Health Inform, 2022. https://pmc.ncbi.nlm.nih.gov/articles/PMC9413516/

[6] Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O., *Autoencoding beyond pixels using a learned similarity metric*, International Conference on Machine Learning (ICML), 2016. http://proceedings.mlr.press/v48/larsen16.html

[7] Tian, Y., Shen, L., Su, G., Li, Z., & Liu, W., *AlphaGAN: Fully Differentiable Architecture Search for Generative Adversarial Networks*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021. arXiv:2006.09134 [cs.CV]. https://doi.org/10.48550/arXiv.2006.09134

[8] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein Generative Adversarial Networks*, Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR 70:214–223, 2017. https://proceedings.mlr.press/v70/arjovsky17a.html

[9] Z. Wang, E. P. Simoncelli, and A. C. Bovik, *Multiscale structural similarity for image quality assessment*, Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2003. https://ieeexplore.ieee.org/document/1292216

[10] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, *A Kernel Two-Sample Test*, Journal of Machine Learning Research, 13:723–773, 2012. http://jmlr.org/papers/v13/gretton12a.html

[11] Jiaqi Wang, et al., *MRIAnatEval: Anatomically-Informed Evaluation for 3D MRI Synthesis*, GitHub Repository, 2024. https://github.com/jiaqiw01/MRIAnatEval

[12] J. Wu, W. Peng, B. Li, Y. Zhang, and K. M. Pohl, *Evaluating the Quality of Brain MRI Generators*, Proceedings of MICCAI 2024, 2024. arXiv:2409.08463. https://arxiv.org/abs/2409.08463

[13] C. Bredell, S. Nader, and C. M. Bishop, *Explicitly Minimizing the Blur Error of Variational Autoencoders*, arXiv preprint arXiv:2304.05939, 2023. https://arxiv.org/abs/2304.05939