Evaluating the Quality of Brain MRI Generators

Jiaqi Wu†[1][0009−0003−7108−079X], Wei Peng†[1][0000−0002−2892−5764], Binxu Li[1][[0009−0001−8044−591X], Yu Zhang[2][0000−0001−9726−6400], and Kilian M. Pohl[1⋆][0000−0001−5416−5159]

[1] Stanford University, Stanford, CA 94305
[2] Lehigh University, Bethlehem, PA 18015

Abstract. Deep learning models generating structural brain MRIs have the potential to significantly accelerate discovery of neuroscience studies. However, their use has been limited in part by the way their quality is evaluated. Most evaluations of generative models focus on metrics originally designed for natural images (such as structural similarity index and Fréchet inception distance). As we show in a comparison of 6 state-of-the-art generative models trained and tested on over 3000 MRIs, these metrics are sensitive to the experimental setup and inadequately assess how well brain MRIs capture macrostructural properties of brain regions (a.k.a., anatomical plausibility). This shortcoming of the metrics results in inconclusive findings even when qualitative differences between the outputs of models are evident. We therefore propose a framework for evaluating models generating brain MRIs, which requires uniform processing of the real MRIs, standardizing the implementation of the models, and automatically segmenting the MRIs generated by the models. The segmentations are used for quantifying the plausibility of anatomy displayed in the MRIs. To ensure meaningful quantification, it is crucial that the segmentations are highly reliable. Our framework rigorously checks this reliability, a step often overlooked by prior work. Only 3 of the 6 generative models produced MRIs, of which at least 95% had highly reliable segmentations. More importantly, the assessment of each model by our framework is in line with qualitative assessments, reinforcing the validity of our approach. The code of this framework is available via https://github.com/jiaqiw01/MRIAnatEval.git.

1 Introduction

Deep learning could have a significant impact on the analysis of magnetic resonance imaging (MRI) studies for tasks such as classification [1] and identification of biomarkers [3]. Reliably training deep learning models for these tasks requires a larger number of samples [15], while most brain MRI studies are relatively small. Augmenting the training data with brain MRIs produced by generative models (such as shown in Fig. 1) could thus be of great value.

Current generative models, typically based on Generative Adversarial Networks (GANs) [10, 20] or diffusion probabilistic models [13, 17, 23], require a

⋆ Corresponding Author: kpohl@stanford.edu; † indicates equal contribution

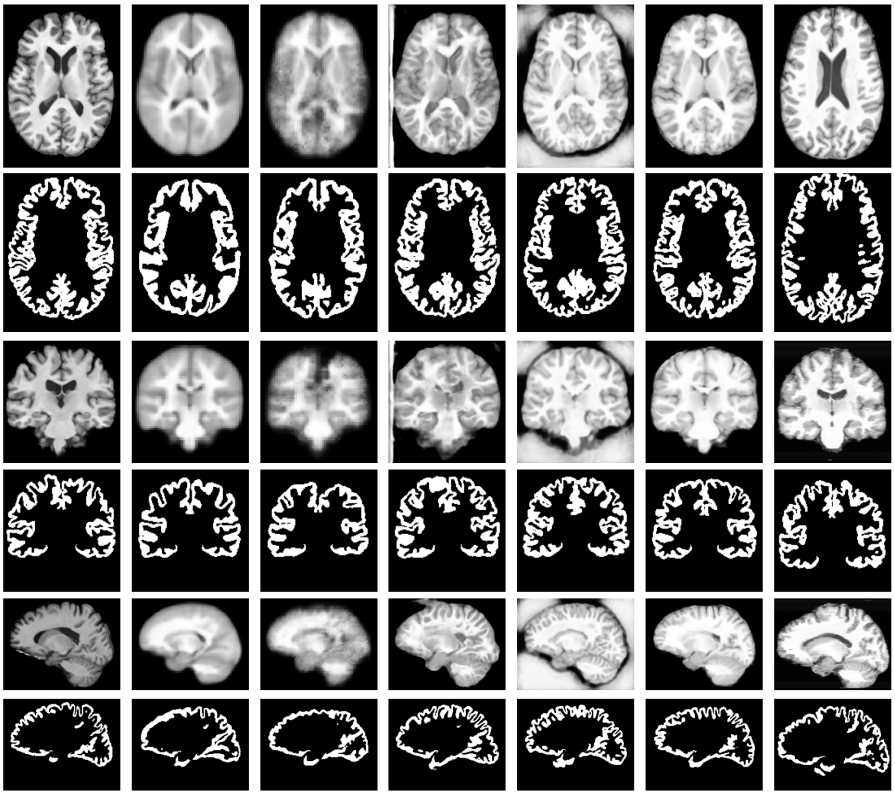

Fig. 1: Axial, coronal, and sagittal view of real and synthetic MRIs produced by six methods. The corresponding gray matter segmentations are produced by *Synthseg*$^+$ [2], which also provides QC scores (listed below method names). Even if the MRIs are of relatively low quality (such as those generated by VAE-GAN, α-WGAN, and MONAI-LDM), the segmentations still look good. However, their QC scores (in italic) is below 0.65 indicating low reliability.

rigorous quality assessment to be useful in brain MRI studies. However, this remains an open issue as the quality of generated MRIs depends on the experimental setup (i.e., the specific implementation of the generative models and the MRIs they are trained and tested on) and the evaluation metrics. Popular metrics used for comparison are those commonly applied to 2D natural images, such as Multi-Scale Structural Similarity (MS-SSIM) [22], Fréchet Inception Distance (FID) [9], and Maximum-Mean Discrepancy (MMD) [7]. However, the outcome of these metrics heavily relies on how they are applied. For example, the MMD score depends on the dimensionality (i.e., per 2D slice or 3D volume) and space (i.e., feature or image space) it is computed over [13, 20]. For FID and MMD,

the choice of feature extractor used by the generator (such as ResNet [8] or Inception Net [21]) can also impact the score [9]. Furthermore, the FID score depends on both image quality and diversity of the generated samples making an interpretation of the score difficult [19]. Even worse, the scores might suggest that the generated MRIs are similar to real ones while the shape of brain regions shown in those MRIs is unrealistic [14]. In [14], we therefore propose to quantify anatomical plausibility, i.e., how well the synthetic MRIs capture properties of brain regions.

Measuring anatomical plausibility is sensitive not only to the processing of the real MRIs and implementation of each method, but also to the automatic MRI segmentations needed for measuring properties of brain regions. To rigorously compare generative models, we therefore propose an evaluation framework that standardizes each of these three components. Novel in standardizing the automatic MRI segmentations is quantitatively assessing their reliability. Automatic segmenters, such as $Synthseg^+$ [2], heavily rely on atlases and therefore can produce a realistic-looking label map even if the MRI is of extremely low quality (such as those produced by VAE-GAN and α-WGAN in Fig. 1). The regional measurements extracted from such a segmentation could suggest that the anatomical plausibility of the MRI is high. To avoid this scenario, our framework quantitatively checks the reliability of the label maps. If more than 5% of the segmentations from an MRI generator are deemed unreliable, we then view the quality of the MRIs created by the generative model as too low for assessment.

We use this framework to compare 3 state-of-the-art GANs and 3 diffusion methods, which are trained and tested on over 3000 structural brain MRIs collected by three studies. Our findings reveal that metrics commonly used on 2D natural images (i.e., MS-SSIM, FID, and MMD) often yield inconclusive findings despite evident qualitative differences. In contrast, our proposed framework effectively captures these differences, providing a more accurate assessment of brain MRI generator performance.

2 Evaluation Framework

Our framework standardizes the processing of the real MRIs, unifies the implementations of both GANs and diffusion models, and measures the anatomical plausibility of the generated MRIs. These three components are now described in further detail.

2.1 Standardized Processing of Real MRIs

As in [14], the processing of the T1-weighted MRIs consists of denoising, bias field correction, skull stripping, intensity normalization between -1 and 1, and affine registration to the SRI atlas [18], which results in MRIs of with 1mm voxel resolution. We pad all MRIs to end up with 144 x 192 x 144 voxels, as some methods [23] need the number of voxels to be dividedable by 3. The resulting MRIs are then used for training and testing the generative models.

2.2 Unified Implementation of Generative Models

For a fair comparison, all generative models should be implemented using the same software platform, for which we choose PyTorch 2.0 [12]. We then choose state-of-the-art 3D MRI generators that we can implement in PyTorch. For GANs, we select VAE-GAN [10], α-WGAN [10], and HA-GAN [20]. With respect to diffusion models, we choose MONAI latent diffusion model (MONAI-LDM) [17], the text-conditioned diffusion model [23] (MedSyn), and the conditional diffusion probabilistic model (cDPM) [13].

2.3 Measuring Anatomical Plausibility

As in [14], we evaluate the anatomical plausibility of the generated MRIs by extracting regional brain measurements from them and comparing their distribution to the measurements extracted from real MRIs. In our case, each MRI is parcellated into 16 subcortical and 33 cortical regions using the automatic segmenter $Synthseg^+$ [2]. Unlike in [14], we check the reliability of those parcellations to ensure that the measures of anatomical plausibility can be trusted.

We assess the reliability based on the quality control (QC) scores of 8 brain regions that $Synthseg^+$ provides with each segmentation (see also the example for one of the scores in Fig. 1). The reliability of an MRI is considered too low for assessment if any of the 8 QC scores are below a pre-defined threshold. We choose the threshold so that 5% of the real MRIs of the test set fail the check. Generative models that produce MRIs whosecorresponding segmentations result in an even higher failure rate are labeled as too unreliable for assessment.

For each MRI generated by those models passing QCs, we record the volume of each brain region based on their parcellation created by $Synthseg^+$. From those scores, we regress out the total intracranial volume to eliminate the impact of brain size on the evaluation. For each region, the distribution of volume scores is then compared to the ones extracted from real MRI of the test set using Cohen's d score [6] as it was done in [14].

3 Comparison of 3D MRI Generators

Our comparison of the six generative models listed in Section 2.2 is based on T1-weighted longitudinal brain MRIs from 1,236 normal controls (age range: 13 to 91 years, 589 male/647 female) pooled from three studies: the Alzheimer's Disease Neuroimaging Initiative [16] (ADNI, 342 controls from ADNI 1, 2, 3 and GO), the National Consortium on Alcohol and Neurodevelopment in Adolescence [4] (NCANDA, 621 controls from NCANDA_PUBLIC_6Y_STRUCTURAL_V01 [11]), and an in-house dataset from SRI International [24] (SRI, 273 subjects; PI: Drs. Pefferbaum and Sullivan). 400 MRIs of 400 subjects then define the test set. They are sampled from the data set so that they uniformly span the age range of the three studies and half of them are female (see [14] for more details). The training set consists of

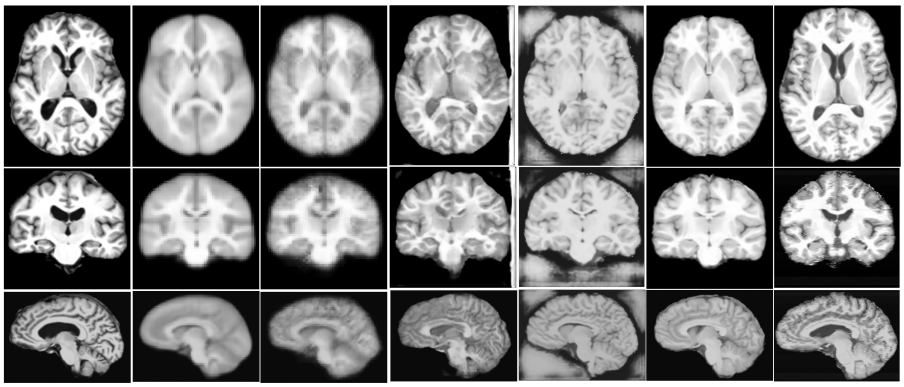

Fig. 2: Real MRI and synthetic MRIs from the 6 generative methods used in our comparison. In general, the MRIs from diffusion models [13, 23] provide greater anatomical details than the GANs.

the remaining 836 subjects of which 3060 T1-weighted MRIs were acquired. All MRIs are processed using the pipeline described in Section 2.1.

After completing training, each of the six models generates 400 MRIs of which a typical example is shown in Fig. 1 & 2. Visually assessing those images reveals that the outputs of VAE-GAN and α-WGAN are blurry, the background of MRIs generated by HA-GAN and MONAI-LDM show artifacts, and those produced by the two diffusion models (MedSyn and cDPM) provide the most detail with respect to brain anatomy. In the remainder of this section, we quantitatively compare the MRIs of the six methods using metrics commonly applied to 2D natural images and 3D MRIs (Section 3.1). The inconclusive findings of that comparison then motivates our evaluation framework of Section 3.2, which is based on anatomical plausibility (see also Section 2.3).

3.1 Evaluation Based on Common Metrics

As examples of metrics commonly used for assessing images, we apply FID, MMD, and MS-SSIM to the real and generated MRIs. For each model, FID and MMD compare the distributions of real and synthetic MRIs in a lower dimensional feature space. To do so, we map all MRIs into this space via a pre-trained encoder. We then document the high impact of the encoder by recoding their scores with respect to three implementations of ResNet [8],i.e., ResNet50 (R50), another version of ResNet50 trained on 23 datasets (R50_23), and ResNet101 (R101) [5]. According to Table 1, the lowest FID and MMD scores for VAE-GAN and α-WGAN (the two models with the lowest quality MRIs based on visual inspection) are recorded with R101, followed by R50_23, and R50. Relative to the other methods, their scores for R50 were only better than MONAI-LDM. However, α-WGAN produces the second-best scores when the encoder is R50_23.

Table 1: Evaluating 400 MRIs of each approach using common metrics. Other than for MS-SSIM, lower scores are considered better with the best score being in bold and the second best underlined. For MS-SSIM, the score closest to one recorded on the real MRIs of the test set (i.e., 0.88) is considered the best.

Model	R101		R50_23		R50		Image	MS-
	FID	MMD	FID	MMD	FID	MMD	MMD	SSIM
VAE-GAN	0.032	0.020	0.081	0.046	0.400	0.210	**131925**	0.91
α-WGAN	0.032	0.020	<u>0.060</u>	<u>0.040</u>	0.480	0.250	<u>203999</u>	**0.88**
HA-GAN	0.035	0.018	0.079	<u>0.040</u>	<u>0.080</u>	<u>0.043</u>	759363	0.78
MONAI-LDM	0.300	0.150	1.420	0.690	1.870	0.940	3314614	0.58
MedSyn	**0.012**	**0.010**	**0.057**	**0.037**	**0.044**	**0.034**	217897	**0.88**
cDPM	<u>0.019</u>	<u>0.014</u>	0.140	0.082	0.130	0.081	586022	0.75

Moreover, the two models produce the best (i.e., lowest) MMD among all the approaches when computed in the image space (Image MMD). Overall, we conclude from these results that even the worst quality images based on visually inspection can produce the best scores.

This observation is confirmed for MS-SSIM, the only metric in this comparison independently computed from the real MRIs. Higher values are generally interpreted as better, which in this case would again point to VAE-GAN (MS-SSIM: 0.91) and α-WGAN (MS-SSIM: 0.88) being the best approaches. However, the score of VAE-GAN is even higher than the score reported on the real MRIs (MS-SSIM: 0.88) pointing towards a lack of diversity among its MRIs and thus a mode collapse. Assuming the quality of MRIs is higher the closer the MS-SSIM score is to the one measured on the real MRI, α-WGAN (and MedSyn) would still be the best approach.

The above observations highlight the limitations of these three evaluation metrics, which are commonly used in the literature. These metrics are overly sensitive to the choice of experimental setup and do not reliably reflect the visual quality of the MRIs.

3.2 MRI-specific Assessments

These shortcomings are the motivation behind our framework for computing anatomical plausibility. The first step in computing that metric is to check that the segmentations extracted from the MRIs are reliable. For each method Table 2 lists the total number of ROI segmentations and corresponding MRIs that failed the check (i.e, QC score < 0.65, which is the case for 4.75% of real MRI). Interestingly, $Synthseg^+$ produces segmentations that look realistic even if the MRIs are of bad quality as in the case of VAE-GAN (see Fig. 1). However, the quality score generated by $Synthseg^+$ clearly points out that most of those segmentations should not be trusted as less than 1% pass quality control according to Table 2. Performing much better is α-WGAN (success rate 83.25%) but the

Table 2: Rate of MRIs passing QC check

	real MRIs	VAE-GAN	α-WGAN	HA-GAN	MONAI-LDM	MedSyn	cDPM
Total failed ROI	19	399	67	0	95	6	11
Total failed MRIs	19	397	67	0	56	6	11
MRIs passing rate	95.25%	0.75%	83.25%	100.00%	86.00%	99.00%	97.25%

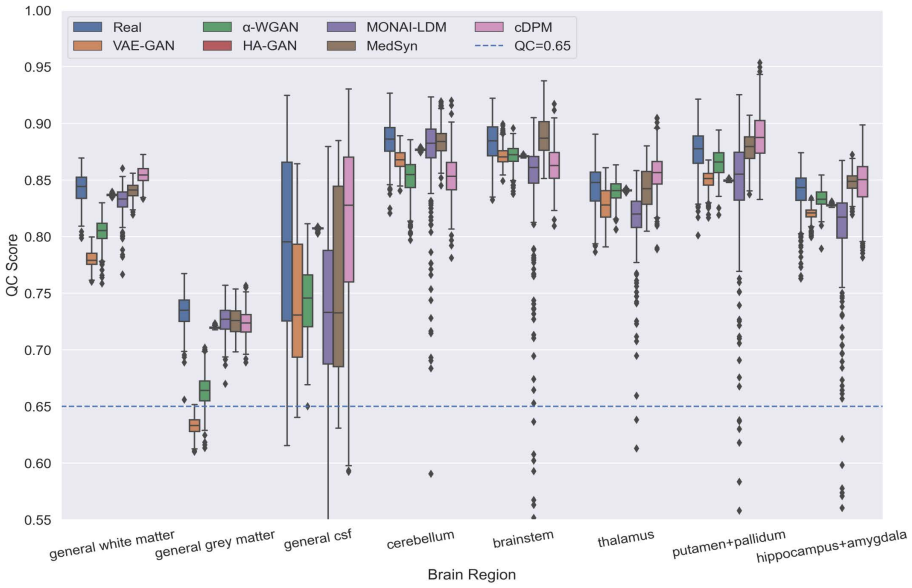

Fig. 3: QC scores of 8 brain regions and 400 MRIs produced by y *Synthseg*$^+$.

quality of its gray matter segmentation often fails to meet the threshold according to the plot shown in Fig. 3. A slightly better success rate has MONAI-LDM with 86%. However, when an MRI produced by this method fails to meet the QC threshold, it is not for a specific region and often involves multiple ones. Only the three models with the best-looking MRIs pass the 95% threshold, which are HA-GAN and the diffusion models cDPM and MedSyn.The gray matter is the only region failing QC for few MRIs generated by cDPM and MedSyn, while HA-GAN is the only model, where all MRIs passed QC. Interestingly enough, the MRIs produced by HA-GAN are visually inferior to those of the two diffusion models so the QC failure rate should not be used as a final metric for assessing anatomical plausibility.

QC detection and ROI Comparison The anatomical plausibility of the MRIs produced by those three methods is summarized by the Cohen's d score between the volume distributions of real and synthetic MRIs for each of the 16 subcortical regions and 33 cortical regions (Table 3). A value closer to 0 indi-

Table 3: Cohen's d of Cortical and Subcortical ROIs with best score in bold

	cerebral WM	lateral ventricle	inferior	cerebellum WM	cerebellum GM	thalamus	caudate	putamen
HA-GAN	0.68	0.10	0.50	-0.35	-0.59	-0.18	0.39	0.53
MedSyn	-0.12	0.69	0.69	**0.22**	0.26	-0.39	-0.31	-0.17
cDPM	**0.00**	**-0.05**	**-0.15**	0.25	**0.23**	**0.05**	**-0.12**	**-0.08**

WM=white matter, GM=gray matter

	pallidum	ventricle 3^{rd}	ventricle 4^{th}	brain-stem	hippo-campus	amygdala	accumbens	cerebrospinal fluid
HA-GAN	1.16	**0.09**	0.92	-0.34	-1.01	-0.04	0.68	0.31
MedSyn	-0.38	0.15	0.34	-0.41	-0.79	-0.76	-0.21	-0.51
cDPM	**0.01**	-0.16	**0.06**	**0.18**	**-0.02**	**0.02**	**-0.07**	**0.04**

	bankssts	caudal AC	caudal MF	cuneus	entorhinal	fusiform	inferior parietal	inferior temporal
HA-GAN	0.25	1.13	0.26	1.09	1.11	2.00	1.08	**0.06**
MedSyn	0.26	0.56	-0.38	0.27	0.95	**-0.11**	-0.09	-0.25
cDPM	**0.04**	**-0.21**	**-0.09**	**0.11**	**0.07**	0.28	**0.00**	0.32

	isthmus cingulate	lateral occipital	lateral orbitofrontal	frontal-pole	lingual	medial orbitofrontal	middle temporal	parahippo-campal
HA-GAN	1.32	0.93	-0.28	1.12	0.24	0.47	-0.38	**-0.09**
MedSyn	**0.07**	-0.41	-0.43	-0.35	-0.72	-0.47	-0.12	-0.14
cDPM	0.11	**0.15**	**-0.06**	**-0.07**	**0.22**	**0.04**	**0.01**	0.13

	central pre	central para	central post	pars opercularis	pars orbitalis	pars triangularis	peri-calcarine	temporal-pole
HA-GAN	-0.26	-1.10	-1.81	0.92	0.73	-0.16	-0.58	-1.00
MedSyn	0.40	0.32	0.38	-0.31	0.16	0.27	-1.37	**0.04**
cDPM	**-0.02**	**0.02**	**-0.02**	**-0.10**	**0.00**	**-0.05**	**0.03**	0.10

	cingulate	precuneus	rostral AC	rostral MF	superior-frontal	superior-parietal	superior-temporal	supra-marginal	transverse temporal
HA-GAN	-0.22	-0.41	1.13	1.02	0.14	1.18	-0.58	-0.95	0.17
MedSyn	**-0.04**	-0.61	**0.11**	0.42	**-0.12**	**0.14**	**-0.07**	0.17	0.14
cDPM	0.09	**0.04**	-0.16	**-0.18**	-0.28	0.21	-0.08	**0.00**	**0.08**

AC=anterior cingulate, MF=middle frontal

cates higher overlap between the distributions and thus anatomical plausibility. Reflecting visual assessment, diffusion models generally produce MRIs of higher anatomical plausibility than HA-GAN, as cDPM has the best Cohen's d score for 38 regions (i.e., 77.5% of regions), followed by MedSyn with 9 regions (18.4%), and HA-GAN with 2 regions (4.1%). Note, that one can also use the anatomical plausibility scores proposed here to exclude the MRIs for aiding the analyzes focusing on regions in which these MRIs receive poor scores (e.g., $|d|>0.8$), such as recorded for 14 regions by HA-GAN.

4 Conclusion

This work not only documents the shortcomings in current assessments of structural brain MRIs produced by generative models but also proposes a framework for solving these issues. The framework standardizes the experimental setup for comparing methods and assessing anatomical plausibility, a metric we previously

introduced in [14]. Unlike in [14], we ensure that the metric returns reliable results by checking that MRI segmentations meet a pre-defined quality threshold. We use this framework to compare six state-of-the-art methods revealing that diffusion models generally produce higher-quality MRIs than generative adversarial networks. More importantly, these assessments align with the visual quality of the MRIs displayed in this article. By creating a reliable assessment for generated MRIs, this framework provides a critical step toward using these images to advance MRI studies.

Acknowledgments. Part of the data set used for this analysis and work was supported by funding from the National Institute of Health (DA057567, AA021681, AA021690, AA021691, AA021692, AA021695, AA021696, AA021697, AA017347, AA010723, AA005965, and AA028840), the DGIST R&D program of the Ministry of Science and ICT of KOREA (22-KUJoint-02), and the Stanford HAI Google Cloud Credits.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdelaziz Ismael, S.A., Mohammed, A., Hefny, H.: An enhanced deep learning approach for brain cancer MRI images classification using residual networks. Artificial Intelligence in Medicine **102**, 101779 (2020)
2. Billot, B., Colin, Magdamo Cheng, Y., Das, S., Iglesias, J.E.: Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets. Proceedings of the National Academy of Sciences (PNAS) **120**(9), 1–10 (2023). https://doi.org/10.1073/pnas.2216399120
3. Bowles, C., Gunn, R., Hammers, A., Rueckert, D.: Modelling the progression of Alzheimer's disease in MRI using generative adversarial networks. Medical Imaging 2018: Image Processing (Mar 2018). https://doi.org/10.1117/12.2293256
4. Brown, S.A., Brumback, T., Tomlinson, K., Cummins, K., Thompson, W.K., Nagel, B.J., De Bellis, M.D., Hooper, S.R., Clark, D.B., Chung, T., et al.: The national consortium on alcohol and neurodevelopment in adolescence (NCANDA): A multisite study of adolescent development and substance use. Journal of Studies on Alcohol and Drugs **76**(6), 895–908 (Nov 2015). https://doi.org/10.15288/jsad.2015.76.895
5. Chen, S., Ma, K., Zheng, Y.: Med3D: Transfer learning for 3D medical image analysis (2019)
6. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. Taylor & Francis (2013)
7. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. The Journal of Machine Learning Research **13**(1), 723–773 (2012)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems. vol. 30, pp. 6629 – 6640 (2017)

10. Kwon, G., Han, C., Kim, D.s.: Generation of 3D brain MRI using auto-encoding generative adversarial networks. In: Medical Image Computing and Computer Assisted Intervention, Lecture Notes in Computer Science. vol. 11766, pp. 118–126. Springer International Publishing (2019)
11. Ouyang, J., Zhao, Q., Adeli, E., Zaharchuk, G., Pohl, K.M.: Self-supervised learning of neighborhood embedding for longitudinal MRI. Medical Image Analysis **82**, 102571 (2022)
12. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. vol. 32, pp. 8026 – 8037 (2019)
13. Peng, W., Adeli, E., Bosschieter, T., Park, S.H., Zhao, Q., Pohl, K.M.: Generating realistic brain MRIs via a conditional diffusion probabilistic model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Lecture Notes in Computer Science. vol. 14227, pp. 14–24. Springer (2023)
14. Peng, W., Bosschieter, T., Ouyang, J., Paul, R., Adeli, E., Zhao, Q., Pohl, K.M.: Metadata-conditioned generative models to synthesize anatomically-plausible 3D brain MRIs (2023), https://arxiv.org/abs/2310.04630
15. Peng, W., Feng, L., Zhao, G., Liu, F.: Learning optimal k-space acquisition and reconstruction using physics-informed neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20794–20803 (2022)
16. Petersen, R.C., Aisen, P.S., Beckett, L.A., Donohue, M.C., Gamst, A.C., Harvey, D.J., Jack, C.R., Jagust, W.J., Shaw, L.M., Toga, A.W., Trojanowski, J., Weiner, M.: Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. Neurology **74**(3), 201–209 (2010)
17. Pinaya, W.H.L., Graham, M.S., Kerfoot, E., Tudosiu, P.D., Dafflon, J., Fernandez, V., Sanchez, P., Wolleb, J., da Costa, P.F., Patel, A., Chung, H., Zhao, C., Peng, W., Liu, Z., Mei, X., Lucena, O., Ye, J.C., Tsaftaris, S.A., Dogra, P., Feng, A., Modat, M., Nachev, P., Ourselin, S., Cardoso, M.J.: Generative AI for medical imaging: extending the MONAI framework (2023), https://arxiv.org/abs/2307.15208
18. Rohlfing, T., Zahr, N.M., Sullivan, E.V., Pfefferbaum, A.: The SRI24 multichannel atlas of normal adult human brain structure. Human Brain Mapping **31**(5), 798–819 (2010)
19. Sajjadi, M.S.M., Bachem, O., Lucic, M., Bousquet, O., Gelly, S.: Assessing generative models via precision and recall. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018), https://proceedings.neurips.cc/paper_files/paper/2018/file/f7696a9b362ac5a51c3dc8f098b73923-Paper.pdf
20. Sun, L., Chen, J., Xu, Y., Gong, M., Yu, K., Batmanghelich, K.: Hierarchical amortized GAN for 3D high resolution medical image synthesis. IEEE Journal of Biomedical and Health Informatics **26**(8), 3966–3975 (2022)
21. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9 (2015)

22. Wang, Z., Simoncelli, E., Bovik, A.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003. vol. 2, pp. 1398–1402 (2003). https://doi.org/10.1109/ACSSC.2003.1292216
23. Xu, Y., Sun, L., Peng, W., Jia, S., Morrison, K., Perer, A., Zandifar, A., Visweswaran, S., Eslami, M., Batmanghelich, K.: MedSyn: Text-guided anatomy-aware synthesis of high-fidelity 3D CT images. IEEE Transactions on Medical Imaging (In press)
24. Zhao, Q., Liu, Z., Adeli, E., Pohl, K.M.: Longitudinal self-supervised learning. Medical Image Analysis **71**, 102051 (2021)