This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Beyond Visual Realism: The Neuro-Fidelity Index for Validating Geometric Integrity in Synthetic Brain MRI

Won Hee Lee, Yongjin Choi, Yeonwoo Kim, Hyeon Jun Park

Kyung Hee University, Yongin, Republic of Korea
whlee@khu.ac.kr

Abstract. Traditional evaluation metrics for medical image synthesis, such as Fréchet Inception Distance (FID), primarily assess global feature distributions but fail to guarantee the fine-grained anatomical and topological validity required for downstream medical imaging applications. While recent frameworks have introduced Wasserstein-based macroscopic shape comparisons to evaluate anatomical plausibility, they often overlook localized structural continuity. We introduce the Neuro-Fidelity Index (NFI), a multi-dimensional benchmarking framework for synthetic neuroimaging that integrates perceptual sharpness, microscopic topological continuity (Euler characteristics), and regional parcellation consistency into a unified metric using a weighted harmonic mean. The harmonic mean acts as an architectural bottleneck detector, ensuring that a catastrophic failure in any structural dimension is heavily penalized in the final score. We evaluate four representative generative architectures (VAE, HAGAN, LDM, and DDPM) on T1-weighted brain MRIs from a cohort of 639 healthy adults. Through this benchmark, we identify a distinct "perceptual paradox" in latent-space modeling: while the LDM achieves high perceptual (0.887) and anatomical (0.867) similarity, it suffers from severe topological collapse (0.001) driven by geometric discontinuities introduced during latent decoding, resulting in an NFI of 0.002. In contrast, the pixel-space DDPM preserves structural integrity across all scales, achieving a balanced NFI of 0.833. These results demonstrate that the NFI serves as a rigorous, interpretable evaluation framework for medical image synthesis, revealing critical generative failures overlooked by conventional perceptual metrics.

Keywords: Brain, MRI, Synthesis, Diffusion, Fidelity, Generative Models

1 Introduction

Generative AI has fundamentally transformed medical image synthesis, with diffusion models emerging as the state-of-the-art for generating high-resolution neuroimaging data [1, 2]. By leveraging denoising score-matching [3], models such as Denoising Diffusion Probabilistic Models (DDPM) and Latent Diffusion Models (LDM) have achieved unprecedented visual realism, often outperforming Generative Adversarial Networks (GANs) in capturing complex image distributions [4]. These models are now widely adopted as solutions for data scarcity in neuroimaging, providing synthetic datasets for benchmarking, simulation, and data augmentation.

However, a critical “perceptual paradox” has emerged: while these models generate visually realistic MRI volumes, the cortical surfaces extracted from synthetic data often exhibit severe geometric defects. In computer vision, the Fréchet Inception Distance (FID) serves as the standard quality metric, measuring statistical similarity between real and generated feature distributions [5]. However, FID does not explicitly evaluate geometric properties critical for medical imaging, such as topological consistency or anatomical plausibility. While medical imaging employs voxel-wise metrics (PSNR, SSIM) and segmentation overlap (Dice), these remain insufficient for validating surface geometry extracted from volumetric data. Recent work has shown that latent-space compression can introduce microscopic topological artifacts, including self-intersecting surfaces and non-manifold vertices [6, 7], rendering extracted meshes unsuitable for automated morphometric and parcellation pipelines.

In structural neuroimaging, surface integrity is defined by strict biological and geometric constraints. A valid synthetic brain must satisfy three concurrent criteria: (1) perceptual sharpness, representing the absence of visual noise and artifacts; (2) topological integrity, ensuring a manifold surface with a consistent Euler characteristic, free of self-intersections or genus errors [8]; and (3) anatomical plausibility, maintaining morphometric properties (e.g., cortical thickness and subcortical volumes) within normative human biological variance. Topological defects invalidate downstream morphometric analyses, including cortical thickness estimation, surface-based registration, and longitudinal tracking, which are critical tools for diagnosing neurodegenerative diseases and monitoring treatment response [9, 10].

To address these limitations, we propose the Neuro-Fidelity Index (NFI), a unified benchmarking metric that integrates perceptual, topological, and anatomical fidelity via a weighted harmonic mean. Unlike arithmetic means, the harmonic formulation acts as a generative “bottleneck detector”, ensuring that catastrophic failures in any dimension (e.g., topological collapse) yield proportionally lower overall scores. We provide a clear protocol for tuning NFI's structural weights based on localized task priorities, making it highly customizable across different dataset constraints. We benchmark four state-of-the-art generative architectures (VAE, HAGAN, LDM, and DDPM) on T1-weighted brain MRIs from a cohort of 639 healthy adults. Our key contributions are: (1) a rigorous multi-dimensional validation framework that exposes critical failures in latent-space models overlooked by conventional distribution-matching metrics; (2) empirical evidence of the “perceptual paradox”: LDM achieves state-of-the-art FID but near-zero topological fidelity, while pixel-space DDPM maintains geometric integrity; and (3) a diagnostic utility that provides explicit failure-mode tracking to accelerate the architectural development of stable, structure-preserving generative models.

2 Methods

2.1 Dataset and Preprocessing

All experiments were performed on T1-weighted MRI data from the Cambridge Centre for Ageing and Neuroscience (Cam-CAN) cohort [11], consisting of 639 cognitively

healthy participants (318 females; age 18–88 years), establishing a wide adult lifespan baseline for normative anatomical variance. Data acquisition followed approved ethical protocols with informed consent.

Preprocessing included skull removal, bias field correction, and affine registration to the MNI152 standard space at 2 mm isotropic resolution [12, 13], and FreeSurfer intensity normalization [14]. Volumes were cropped from 91×109×91 to 72×96×72 voxels for computational efficiency. Cortical surfaces were reconstructed, parcellated, and structurally validated using FreeSurfer, incorporating automated topology correction to resolve genus errors and enforce a spherical topology (S^2) prior to geometric invariant computation. Parcellations utilized the Desikan-Killiany atlas [15], with subcortical segmentation via FreeSurfer's probabilistic atlas [16]. We extracted 152 morphological features: 68 cortical thickness, 68 cortical surface area, and 16 subcortical volume measures, serving as the reference distribution for NFI calibration.

2.2 The Neuro-Fidelity Index (NFI) Framework

Perceptual Fidelity. To evaluate global feature alignment and visual quality, we compute the FID between real (r) and synthetic (s) feature vectors extracted from a pretrained Inception-v3 network:

$$\text{FID} = \|\mu_r - \mu_s\|_2^2 + Tr(\Sigma_r + \Sigma_s - 2(\Sigma_r\Sigma_s)^{\frac{1}{2}}) \quad (1)$$

where μ and Σ denote feature means and covariances, and $\text{Tr}(\cdot)$ is the matrix trace operator. Because raw FID is unbounded and negatively correlated with quality, direct integration into a unified score is mathematically incompatible. We therefore pass the raw distance through a normalized exponential decay function to derive the perceptual fidelity score $S_{per} \in [0, 1]$:

$$S_{per} = exp(-\frac{FID}{\gamma_{per}}) \quad (2)$$

Here, γ_{per} is a scaling hyperparameter calibrated to the empirical within-cohort baseline variance (γ_{per} = 0.02). Under this formulation, a perfect distributional match (FID $\rightarrow$ 0) yields $S_{per} \rightarrow 1$, while severe structural noise or mode collapse drives $S_{per} \rightarrow 0$, providing a bounded, scale-invariant component for harmonic aggregation.

Topological Integrity. Synthetic cortical surfaces are modeled as closed, two-dimensional manifolds where structural validity requires strict adherence to geometric constraints. Following mesh reconstruction, we evaluate the geometric correctness of the synthetic outputs using the Euler characteristic (χ), defined as $\chi = V - E + F$, where V, E, and F represent the number of vertices, edges, and faces in the mesh, respectively. For a topologically valid hemispheric surface, the target is a genus-0 closed mesh (χ = 2). Across the bilateral and subcortical surface reconstructions of our reference cohort, the empirical distribution yields a normative baseline of μ_{real} = 32.4 ± 14.5. Any deviation reflects topological defects (e.g., handles, holes, or mesh fragmentation) that

invalidate automated parcellation. To quantify distributional drift in structural topology, we compute the topological integrity score $S_{top} \in [0, 1]$ using the 1-Wasserstein distance (W):

$$S_{top} = exp(-\frac{W(\chi_{syn},\ \chi_{real})}{\gamma_{top}}) \tag{3}$$

where $\gamma_{top} = \sigma(\chi_{real})$ is the standard deviation of the reference distribution. This formulation acts as a strict detector of topological collapse; even if the synthetic mean aligns with the baseline, any variance inflation or distribution skewness drives S_{top} exponentially toward zero, exposing mathematically invalid geometry.

Anatomical Fidelity. To enforce biological plausibility, we evaluate the distribution matching of three morphometric feature categories (k) extracted from the Desikan-Killiany parcellation: subcortical volume (s_{vol}), cortical surface area (s_{area}), and cortical thickness (s_{thick}). For each category, we compute the 1-Wasserstein distance (W_k) between the synthetic distribution ($P_{syn,k}$) and the real reference distribution ($P_{real,k}$). We then normalize these distances into bounded scores $s_k \in [0, 1]$ using an exponential decay function:

$$s_k = exp(-\frac{W_k}{\gamma_k}) \tag{4}$$

where γ_k represents a scaling hyperparameter calibrated to the empirical variance of the reference cohort (γ_{vol} = 5643.03 mm^3, γ_{area} = 2505.11 mm^2, γ_{thick} = 2.52 mm). The overall anatomical fidelity score (S_{ana}) is defined as the arithmetic mean of the components:

$$S_{ana} = (s_{vol} + s_{area} + s_{thick})/3 \tag{5}$$

This multi-component aggregation makes the index highly sensitive to anatomical hallucinations, distribution skewness, and mode collapse across distinct structural biomarker units.

2.3 The Unified Neuro-Fidelity Index

The final stage unifies the three sub-metrics (S_{per}, S_{top}, S_{ana}) into a single scalar value. Unlike an arithmetic mean, which allows a high score in one dimension to mask catastrophic failure in another, the NFI utilizes a weighted harmonic mean to act as a strict bottleneck detector:

$$NFI\ =\ (\sum_{i\in\{per,top,ana\}} w_i \cdot S_i^{-1})^{-1} \tag{6}$$

Given that a synthetic volume must concurrently satisfy visual, geometric, and anatomical requirements to be viable, we assign equal weights (w_i = 1/3) by default, yielding the expanded form:

$$NFI\ =\ 3/(S_{per}^{-1} + S_{top}^{-1} + S_{ana}^{-1}) \tag{7}$$

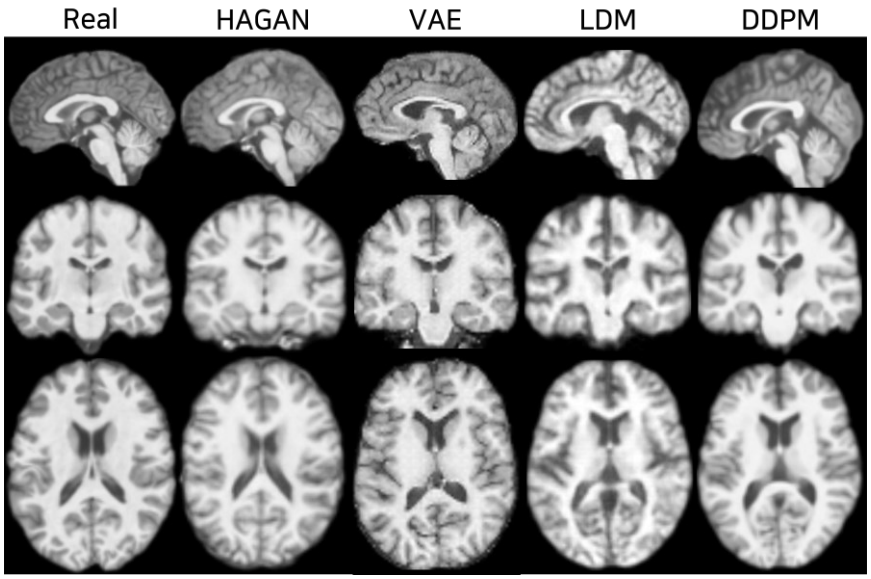

Fig. 1. Representative samples of real and synthetic brain MRI displayed in sagittal, coronal, and axial planes.

The harmonic formulation provides three critical advantages for generative evaluation: (1) if any individual pillar (e.g., S_{top}) approaches zero due to structural collapse, the overall NFI trends to zero regardless of high perceptual realism (S_{per}); (2) because each sub-metric is normalized to [0, 1], the NFI enables standardized benchmarking across disparate generative architectures; and (3) it penalizes high intra-model variance, rewarding balanced models over those that achieve localized perfection at the expense of structural failure. This unified benchmark ensures that a generative architecture is only rated as "high-fidelity" if it accurately preserves the underlying anatomical and topological manifold properties.

3 Experiments

Generative Models. To evaluate the multi-dimensional sensitivity of the NFI, we benchmarked four representative classes of 3D generative architectures, each reflecting a distinct paradigm in manifold learning:

Variational Autoencoder (VAE). A 3D-VAE was implemented to serve as a baseline for distributional mapping [17]. VAEs optimize the evidence lower bound (ELBO) of the data likelihood. However, the reliance on voxel-wise L_2 reconstruction loss typically produces over-smoothed, "blurry" volumes. This model serves as a test case for the sensitivity of S_{per} to high-frequency structural loss.

Hierarchical Anatomical GAN (HAGAN). Representing the adversarial paradigm, HAGAN utilizes a hierarchical discriminator to enforce anatomical realism across multiple spatial scales [18]. While GANs excel at generating sharp textures, they are

historically prone to mode collapse, where the model fails to capture the full biological variance of the Cam-CAN cohort—a failure state specifically monitored by S_{ana}.
Latent Diffusion Model (LDM). We evaluated an LDM that executes the diffusion process within a compressed latent space [2]. Although LDMs achieve state-of-the-art visual results, the non-linear decoding from latent to pixel space can introduce geometric discontinuities. This model is the primary candidate for the "perceptual paradox," where high visual scores (S_{per}) mask underlying topological fragmentation (S_{top}).
Denoising Diffusion Probabilistic Model (DDPM). As a high-fidelity benchmark, we trained a DDPM operating directly in the image space [1]. By modeling the data distribution through a reverse Markovian process without latent compression, the DDPM is hypothesized to preserve the highest degree of topological and anatomical consistency, providing a "ceiling" for the NFI evaluation.

Implementation Details. To ensure a fair comparison, all four generative architectures were trained under a unified experimental framework. The models were implemented in PyTorch using the MONAI framework and trained on an NVIDIA RTX 6000 GPU.

For the diffusion-based models (DDPM and LDM), we employed a 3D U-Net backbone [19] with residual blocks and multi-head spatial attention. The LDM specifically utilized a pre-trained 3D-VAE with a compression factor of 4 to map volumes into the latent space [2]. The HAGAN model utilized a progressive growth strategy for its hierarchical discriminator to maintain stability during adversarial training [18]. The baseline VAE used a symmetrical encoder-decoder with four 3D convolutional layers and a 256-dimensional latent bottleneck [17].

For evaluation, each model synthesized 639 volumes to match the Cam-CAN cohort size. Synthetic volumes were exported in NIfTI format and processed using the identical, automated FreeSurfer *recon-all* pipeline used for the ground-truth data [14]. This pipeline extracted the Euler characteristics for S_{top} and the 152 morphometric features for S_{ana} without manual intervention, ensuring an unbiased structural evaluation.

4 Results

Comparative Performance. Table 1 summarizes performance across four generative architectures on 639 synthetic volumes. While all models produced visually plausible outputs (Fig. 1), their geometric validity varied dramatically. The DDPM achieved the highest NFI (0.833), whereas the LDM exhibited catastrophic topological failure (S_{top} = 0.001) despite superior perceptual quality, resulting in a 417-fold lower NFI (0.002). All pairwise NFI differences are statistically significant ($p < 0.001$, paired t-test with

Table 1. Comparative performance of generative models on the NFI framework.

Model	FID ↓	S_{per} ↑	S_{top} ↑	S_{ana} ↑	Arithmetic ↑	Geometric ↑	NFI ↑
VAE	0.0273	0.255	0.014	0.697	0.322	0.089	0.040
HAGAN	0.0158	0.454	0.040	0.837	0.444	0.197	0.106
LDM	0.0024	0.887	0.001	0.867	0.585	0.027	0.002
DDPM	0.0020	0.905	0.735	0.880	0.840	0.835	0.833

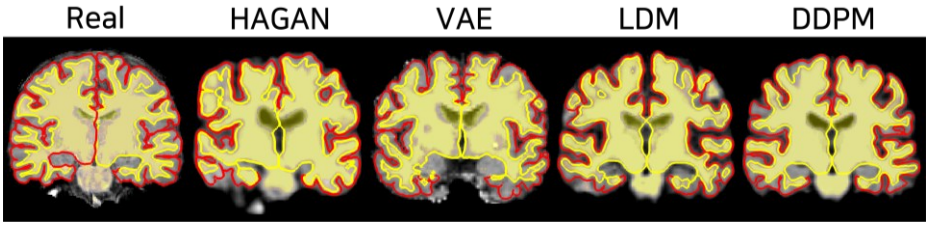

Fig. 2. FreeSurfer surface reconstruction quality. White matter (yellow) and pial surface (red) boundaries overlaid on T1-weighted images.

Bonferroni correction). Bootstrap confidence intervals (n = 1000) confirmed the stability of this benchmark ranking: DDPM (0.831-0.834), HAGAN (0.103-0.108), VAE (0.038-0.042), and LDM (0.002-0.003).

The Perceptual Paradox. Despite near-identical FID scores (LDM: 0.0024 vs. DDPM: 0.0020) and strong anatomical accuracy (S_{ana} = 0.867), LDM's topological integrity collapsed to S_{top} = 0.001. Latent-space compression introduced geometric discontinuities (self-intersections and non-manifold vertices) – visible as white/pial surface boundary errors (Fig. 2) and severe mesh fragmentation (Fig. 3) – that prevent valid surface reconstruction, rendering the generated outputs clinically unusable despite visual plausibility. The harmonic mean formulation correctly penalizes this catastrophic failure (NFI = 0.002), while arithmetic averaging would yield a misleading score of 0.585 (Table 1). This stark contrast between perceptual and geometric fidelity demonstrates that conventional perceptual metrics are insufficient for medical imaging validation.

Anatomical Fidelity. DDPM achieved balanced performance across all dimensions (S_{per} = 0.905, S_{top} = 0.735, S_{ana} = 0.880), establishing the first generative model with NFI > 0.8. Pixel-space diffusion without latent compression preserved manifold structures, yielding a 735-fold topological improvement over LDM (visible in Figs. 2-3). HAGAN demonstrated moderate anatomical accuracy (S_{ana} = 0.837) but poor perceptual quality (S_{per} = 0.454) and weak topological integrity (S_{top} = 0.040), resulting in NFI = 0.106—a 7.9-fold reduction relative to DDPM. VAE exhibited the poorest performance (NFI = 0.040) due to L2-induced blurriness (S_{per} = 0.255) and minimal topological consistency (S_{top} = 0.014), confirming that classical VAE architectures are unsuitable for clinical neuroimaging without substantial modifications.

Weight Sensitivity Analysis. To ensure that the benchmarking capability of the framework is robust to hyperparameter variations, we conducted a weight sensitivity analysis across multiple configurations of w_{per}, w_{top}, and w_{ana} using values in {0.2, 0.33, 0.5} subject to $\sum w_i = 1$.

The structural model rankings remained highly stable across all variations. The DDPM consistently maintained the highest overall performance, with its final index fluctuating tightly between 0.81 and 0.85. In contrast, the LDM remained trapped below 0.01 across all configurations due to the severe bottleneck penalty imposed by the harmonic mean aggregation on its near-zero topological score (S_{top} = 0.001). This stability across diverse parameter spaces confirms that the NFI's diagnostic capability is

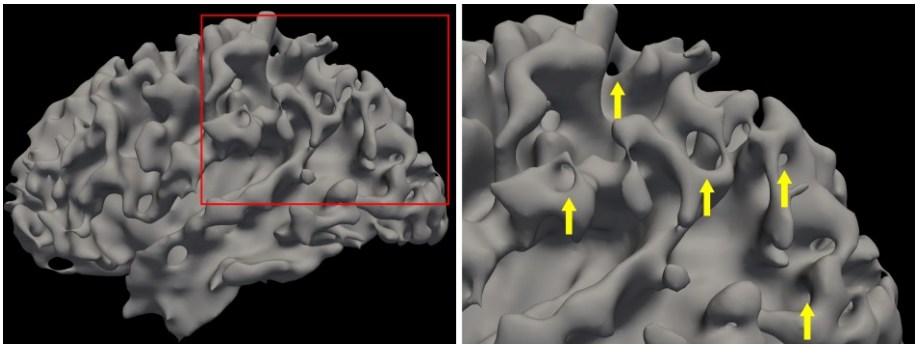

Fig. 3. Topological defects in LDM-generated cortical surfaces revealed by FreeSurfer surface reconstructions.

inherently resilient to user-defined weighting variations. Furthermore, these results support the use of the equal-weighting protocol ($w_{per} = w_{top} = w_{ana} = 1/3$) as a standardized baseline configuration for reproducible medical image synthesis evaluation.

5 Conclusion

We introduce the NFI, a multi-dimensional benchmarking metric that integrates perceptual, topological, and anatomical validation for synthetic brain MRIs. Our evaluation reveals a critical "perceptual paradox" in generative modeling: while latent diffusion models achieve state-of-the-art FID (0.0024), they suffer from severe topological collapse ($S_{top} = 0.001$). Conversely, pixel-space diffusion maintains a balanced structural performance (NFI = 0.833). By leveraging harmonic mean aggregation, the NFI operates as an architectural bottleneck detector, preventing single-dimension structural failures from being masked by isolated distribution-matching metrics. Ultimately, our framework demonstrates that conventional perceptual metrics alone are insufficient for validating the geometric configurations required by medical imaging synthesis. As generative simulation tools continue to evolve, the NFI provides the rigorous, interpretable evaluation framework necessary to ensure synthetic neuroimaging faithfully respects complex biological and structural manifold constraints.

Acknowledgments. This work was supported in part by the National Research Foundation of Korea (NRF), funded by the Korea government (MSIT) (RS-2026-25485605); in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP), funded by the Korea government (MSIT), under grants RS-2024-00509257 (Global AI Frontier Lab) and IITP-2026-RS-2024-00438239 (Information Technology Research Center, ITRC); and in part by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare (RS-2025-02293110).

References

1. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems 33, 6840–6851 (2020)
2. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695. (2022)
3. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems 32, (2019)
4. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems 34, 8780–8794 (2021)
5. 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. Advances in neural information processing systems 30, (2017)
6. Lucas, J., Tucker, G., Grosse, R., Norouzi, M.: Understanding posterior collapse in generative latent variable models. (2019)
7. Wu, J., Peng, W., Li, B., Zhang, Y., Pohl, K.M.: Evaluating the quality of brain MRI generators. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 297–307. Springer, (2024)
8. Ségonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B.: A hybrid approach to the skull stripping problem in MRI. Neuroimage 22, 1060–1075 (2004)
9. Pérez-Millan, A., Borrego-Écija, S., Falgàs, N., Juncà-Parella, J., Bosch, B., Tort-Merino, A., Antonell, A., Bargalló, N., Rami, L., Balasa, M.: Cortical thickness modeling and variability in Alzheimer's disease and frontotemporal dementia. Journal of Neurology 271, 1428–1438 (2024)
10. Li, Y., Wang, Y., Wu, G., Shi, F., Zhou, L., Lin, W., Shen, D., Initiative, A.s.D.N.: Discriminant analysis of longitudinal cortical thickness changes in Alzheimer's disease using dynamic and network features. Neurobiology of aging 33, 427. e415–427. e430 (2012)
11. Shafto, M.A., Tyler, L.K., Dixon, M., Taylor, J.R., Rowe, J.B., Cusack, R., Calder, A.J., Marslen-Wilson, W.D., Duncan, J., Dalgleish, T.: The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. BMC neurology 14, 204 (2014)
12. Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L., Group, B.D.C.: Unbiased average age-appropriate atlases for pediatric studies. Neuroimage 54, 313–327 (2011)
13. Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M.: Fsl. Neuroimage 62, 782–790 (2012)
14. Fischl, B.: FreeSurfer. Neuroimage 62, 774–781 (2012)
15. Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T.: An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968–980 (2006)

16. Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S.: Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355 (2002)
17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
18. 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, 3966–3975 (2022)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, pp. 234–241. Springer, (2015)