Self-supervised Longitudinal Neuroimaging Augmentation for Continuous Staging of Alzheimer's Disease

Haoyuan Liu[1], Mingzhao Tong[2], Shiaofen Fang[2], and Jingwen Yan[1]✉

[1] Department of Biomedical Engineering and Informatics, Indiana University Indianapolis, Indianapolis IN 46202, USA
[2] Department of Computer Science, Indiana University Indianapolis, Indianapolis IN 46202, USA
jingyan@iu.edu

Abstract. Traditional diagnostic groups in Alzheimer's disease provide only a coarse estimate of disease progression, limiting their utility for early detection and disease monitoring. Recent efforts have sought to learn continuous disease staging from longitudinal neuroimaging biomarkers. However, such data are often limited in practice, as most subjects in typical cohorts have only one single visit. To address this challenge, we propose a self-supervised data augmentation framework based on a residual generative adversarial network (GAN) to synthesize follow-up visits for baseline-only subjects. The model incorporates a monotonic brain change constraint to generate biologically plausible follow-up visits without the need of longitudinal supervision. The synthesized visits showed overall good agreement with real longitudinal follow-up visits in terms of direction of change. When combined with real longitudinal observations, the augmented data lead to consistent and measurable improvements in downstream continuous staging, as evidenced by significantly enhanced temporal preservation across real follow-up visits in held-out test data. Source code is available through https://github.com/StochasticAgCl/GAN_SLOPE.

Keywords: Continuous Staging · GAN · Data Augmentation

1 Introduction

Alzheimer's disease (AD) is the most common progressive neurodegenerative disorder, accounting for 60–80% of dementia cases worldwide [2]. Current clinical classification schemes in AD rely heavily on amyloid PET abnormality (e.g., amyloid-positive vs. amyloid-negative) [8,6]. However, such categorization provides only a coarse estimate of disease progression, collapsing a biologically continuous process into a small number of discrete groups. As a result, individuals within the same category may exhibit substantial heterogeneity in underlying amyloid burden and progression stage [11,3], limiting the clinical utility of these groupings for precise disease monitoring and evaluation of treatment effects.

Recent efforts in disease progression modeling seek to move beyond coarse diagnostic grouping. Approaches like event-based models (EBMs) provide a more refined staging by estimating the order in which predefined imaging biomarkers become abnormal [5,15,14]. However, EBMs remain fundamentally discrete and exploratory with limited sensitivity to subtle changes and limited capability in staging new patients. Moreover, EBMs were primarily designed for cross-sectional analysis without explicit evaluation of estimated staging scores. More recently, deep learning–based approaches have been introduced to model AD progression as a fully continuous process from neuroimaging data. In particular, SLOPE provides a parametric framework for continuous AD staging and enables direct staging of new patients [12]. Beyond that, SLOPE introduced evaluation metrics to assess whether the estimated staging scores preserve the temporal ordering of follow-up visits. Such evaluation is essential to ensure that the learned staging score is trustworthy and biologically meaningful. Despite these improvements, SLOPE relies heavily on the supervision of longitudinal follow-up visits, which however is often very sparse in observational disease cohorts [9,4].

To address this limitation, we propose a self-supervised longitudinal data augmentation framework using a residual generative adversarial network (GAN). This model generates plausible near-future amyloid PET patterns to augment single-visit subjects during training. By incorporating the biological prior of irreversible amyloid accumulation [7,13], the generator infers the most likely direction of disease progression without explicit longitudinal supervision. The generated samples effectively densify the training data and enhance the learning of continuous disease staging. We evaluate the framework using directional consistency between generated samples and real follow-up observations. Furthermore, we train SLOPE on the augmented dataset and demonstrate that continuous staging consistency is significantly improved on held-out longitudinal cohorts.

2 Methods

2.1 SLOPE

SLOPE is an autoencoder-based framework that learns a low-dimensional embedding of neuroimaging data [12]. A directional loss is incorporated to enforce that longitudinal follow-up visits of the same subject move consistently along a common progression direction in the latent space. The learned embedding is further projected using UMAP [10], and each subject's relative position along a fitted principal curve is defined as a continuous staging score ranging from 0 to 1. To evaluate robustness, SLOPE introduces the violation ratio, which measures the percentage of consecutive follow-up visits whose estimated staging scores decrease instead of increase.

2.2 Residual-GAN for Longitudinal Augmentation

To address limited longitudinal data, we propose a self-supervised generative framework to generate plausible future amyloid states for single-visit samples.

As shown in Fig. 1, the generator is implemented as a residual network that predicts a progression residual from an imaging feature vector and random noise. This residual is added to the input via a skip connection, allowing the model to learn incremental disease progression rather than reconstructing the entire sample. To ensure biological plausibility, we impose a non-negative activation constraint on the residual using a scaled softmax function. This enforces monotonic increases in amyloid-related biomarkers, reflecting the irreversible nature of amyloid accumulation in AD.

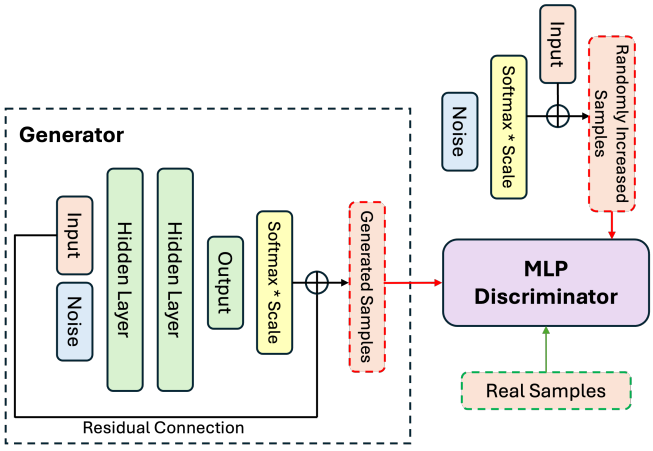

Fig. 1. Overview of the architecture of residual-GAN

The objective of the generator is to produce plausible future visits that are indistinguishable from real samples while remaining distinguishable from directionally monotonic but biologically incoherent progression examples. To construct such reference examples, we generate some synthetic samples $\hat{x}_{rand}$ by adding a random non-negative residual to real samples. Here, z is sampled from a standard Gaussian distribution. α represents the increasing scale and is treated as a hyperparameter.

$$\hat{x}_{rand} = x + \alpha \cdot Softmax(z) \tag{1}$$

These samples satisfy the monotonic constraint but exhibit unstructured feature changes, serving as negative examples that do not reflect realistic amyloid progression. During adversarial training, the discriminator learns to distinguish real samples from both generator outputs and randomly monotonic samples. Consequently, the generator is encouraged to produce future states that resemble real follow-up observations and deviate from arbitrary monotonic shifts, thereby learning structured and biologically consistent progression patterns. Note that regional amyloid SUVR measurements across longitudinal follow-up visits are

intrinsically noisy and may fluctuate over time, without strictly satisfying monotonic progression. Therefore, the goal of the proposed model is not to reconstruct noisy future observations exactly, but rather to generate biologically plausible future visits that remain consistent with the underlying progression pattern while closely approximating the measured data.

Taken together, the final objective of the whole GAN (D_θ, G_ϕ) can be formulated as below, and will be optimized via Monte-Carlo sampling during the training stage.

$$\min_{\phi}\max_{\theta}\mathbb{E}_x[\log\sigma(D_\theta(x))]+\mathbb{E}_x[\log(1-\sigma(D_\theta(G_\phi(x))))]+\mathbb{E}_x[\log(1-\sigma((D_\theta(\hat{x}_{rand})))] \tag{2}$$

3 Experiment

3.1 Dataset

We obtained amyloid PET data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. From the processed PET measurements, we extract a total of 68 bilateral features corresponding to the amyloid accumulation at different cortical regions. In total, 1,161 subjects were included for the analysis, in which 393 subjects have only one single visit and the remaining subjects contain multiple follow-up visits. At baseline, there are 409 cognitively normal (CN) subjects, 531 with mild cognitive impairment (MCI), and 188 with AD. An additional 33 subjects without a diagnosis were included, as diagnostic labels were not required for model training. During the follow-up period, 47 CN subjects converted to MCI, 11 CN subjects converted to AD, and 94 MCI subjects converted to AD.

To evaluate the effectiveness of the proposed method, subjects with multiple visits were randomly divided into two equal groups. The first half of the multi-visit subjects, together with all single-visit subjects, were used for GAN training and hyperparameter tuning. Augmented samples were generated for the single-visit subjects to expand the training set for SLOPE. The remaining 50% of the multi-visit subjects were reserved as a strictly held-out testing set and were not involved in any model training or hyperparameter selection.

3.2 Model training

During GAN training, all samples are treated as cross-sectional, and no longitudinal information is used. The generator learns to produce plausible progression solely from imaging features under the imposed biological constraints. After convergence, the trained generator is applied to all single-visit subjects to perform two-step forward augmentation, producing two pseudo-future follow-up visits. The augmented dataset is then used to train SLOPE with fixed hyperparameters provided by the original paper [12]. SLOPE first learns latent representations from the augmented data, after which UMAP is applied for dimensionality reduction. A spline-based principal curve is subsequently fitted along the latent

manifold to obtain continuous staging scores ranging from 0 to 1. We determined the best hyperparameters for GAN using Optuna's multi-objective optimization [1] by simultaneously minimizing generation loss and violation ratio of downstream SLOPE staging within the training set. Final performance evaluation is conducted exclusively on a strictly held-out test set.

3.3 Evaluation Strategy

Generation Quality We evaluate generation quality using subjects with substantial amyloid progression in real follow-up visits (final Centiloid > 10 and total amyloid increase > 20), including 78 training samples and 85 test subjects. For each selected subject, the generated sample is compared to the real follow-up visit with the closest total feature change, defined as the sum of regional amyloid variations. Similarity is assessed on change vectors relative to baseline to evaluate progression direction. We used cosine similarity and Pearson correlation to quantify directional consistency, and Fréchet Inception Distance (FID) to measure distributional similarity. Randomly monotonically increasing reference vectors and directional similarities between consecutive real visits are included as baselines.

To assess robustness to stochastic optimization, we repeat the full GAN training process using three different random seeds. For each seed, the generator is retrained from scratch, augmented samples are regenerated, and SLOPE is subsequently retrained on the augmented dataset. Final performance metrics are computed on the held-out test set for each run. Results are reported as the average performance across the three independent trials.

Temporal preservation across follow-up visits We further evaluate whether augmentation improves continuous staging consistency by comparing violation ratios on both training and test multi-visit subjects. A lower violation ratio indicates better preservation of longitudinal ordering. Results from SLOPE trained with and without augmented longitudinal samples are compared to quantify the impact of generated samples. The violation ratio is defined as

$$r = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K_i - 1} \mathbf{1}\big[p_{i,k+1} < p_{i,k} - \tau\big], \tag{3}$$

where $p_{i,k}$ denotes the inferred staging score of sample $x_{i,k}$ from subject i, K_i is the number of visits for subject i, and τ is a tolerance threshold to ignore minor stage reversals.

4 Results

4.1 Quality of Augmented Samples

We compare the proposed GAN against two references: randomly monotonically increased samples with matched amyloid magnitude and progression between

real consecutive follow-up visits. The random baseline tests whether improvements arise from trivial perturbations, while real consecutive visits provide a reference for how consistent true longitudinal progression is in practice.

As shown in Fig. 2, generated samples exhibit higher directional agreement with real follow-up visits than the random baseline in both Pearson correlation (Fig. 2A) and cosine similarity (Fig. 2B), with consistent performance across training and test sets. Directional similarity between real consecutive visits is relatively low. This is expected as comparisons between augmented samples and real visits are restricted to those with substantial amyloid change, whereas most real consecutive follow-up intervals exhibit minimal progression regardless of final amyloid status. Because amyloid changes over short follow-up intervals are often small and close to the measurement noise level, directional similarity between real consecutive visits is inherently low and highly variable. Therefore, real–real similarity should not be interpreted as an upper bound of achievable consistency.

Consistent with these findings, the generated samples also achieve lower FID scores (train: 0.169, test: 0.186) compared to randomly increased samples (train: 0.428, test: 0.420), indicating closer distributional alignment with real progression data.

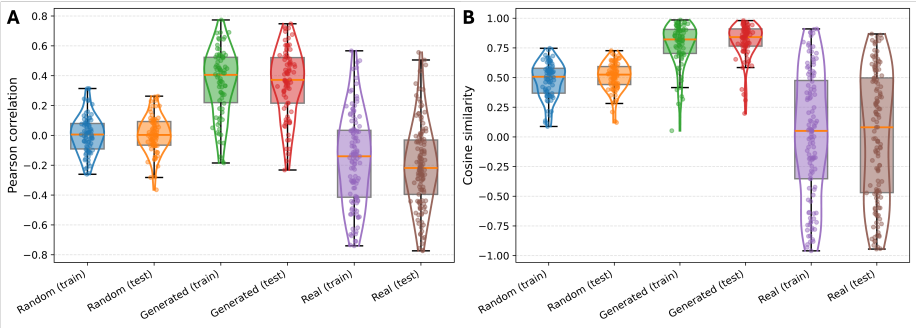

Fig. 2. Similarity between generated and real progression directions, benchmarked against the similarity observed between real consecutive visits and those generated under a random monotonicity assumption.

4.2 Temporal preservation of staging scores

We first evaluated whether generative augmentation improves the temporal consistency of continuous staging at the diagnostic group level. SLOPE was trained on training set with additional augmented longitudinal samples, and the resulting embedding are visualized using UMAP. The relative position of samples on the fitted curve was calculated as staging score, ranging from 0 to 1 (Fig. 3A). In Fig. 3B, we observed that the generated samples are well integrated within

the distribution of real samples. Notably, few augmented samples are positioned at the beginning of the curve, a region dominated by cognitively normal samples with minimal amyloid pathology. This is expected, as our generative model aims to synthesize a future visit with moderate amyloid increase. Fig. 3C shows a clear separation of CN and AD samples in the test set. CN samples are primarily located on the left side of the UMAP embedding, corresponding to lower staging scores, while AD patients are positioned toward the end of the curve with staging scores approaching 1. This separation of diagnosis groups aligns with the expected disease progression pattern and provides additional evidence for the biological plausibility of the learned staging scores. Fig. 3D shows a few samples with augmented and real follow-up visits, both showing progressive staging score along follow-up visits.

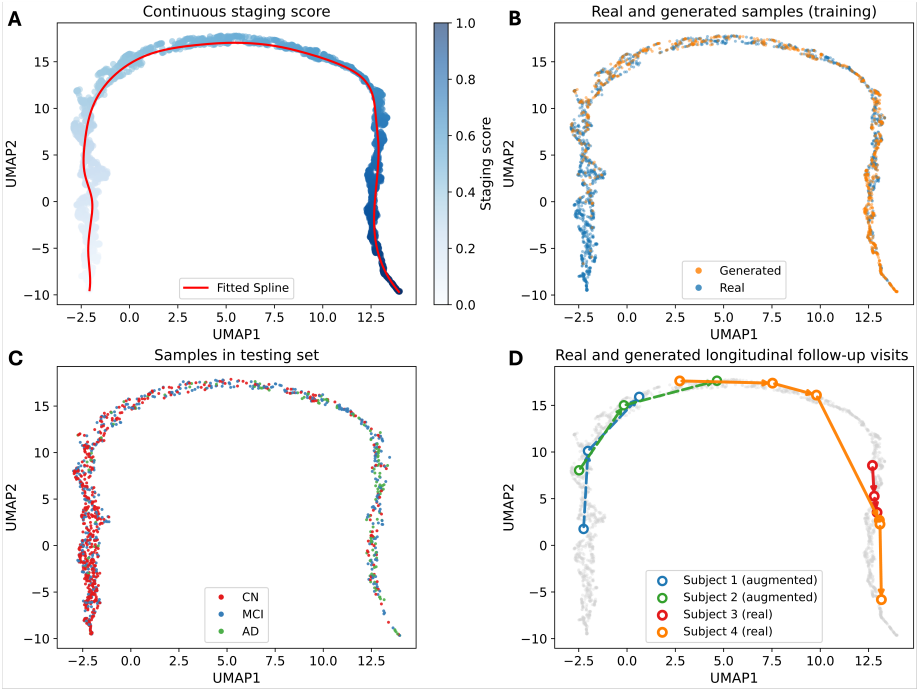

Fig. 3. UMAP embedding derived from SLOPE trained with augmented longitudinal data. A) Embedding of training samples colored by estimated staging score, which was calculated as the relative position on the fitted line. B) Embedding colored by real and generated samples in the training set. C) Embedding colored by diagnosis groups in the testing set. D) Progression trajectories from 2 subjects with generated follow-ups and 2 test subjects with real follow-ups. Dashed lines indicate trajectories of generated follow-ups, whereas solid lines represent trajectories derived from real longitudinal observations in the testing set.

To compare SLOPE trained with and without augmented data, we evaluated their derived staging scores using the violation ratio, defined as the percentage of consecutive follow-up visits with reversed staging scores. A lower violation ratio indicates a more trustworthy staging score where the temporal sequence of visits is preserved. As shown in Fig. 4, incorporating augmented follow-up samples consistently reduces the violation ratio across various tolerance thresholds (where minor reversals below the threshold are ignored). This improvement is evident in both training and testing set. Notably, at a 0.05 tolerance threshold, augmentation reduces violation ratio from approximately 18% to 10%. The shaded regions represent performance variance across three independent runs with different random seeds, showing that the original SLOPE model exhibits higher variability than the GAN-augmented version. These results suggest that augmentation enhances the stability of continuous staging and that the improvement is robust to stochastic initialization. Together, these findings demonstrate that biologically constrained generative augmentation enhances the temporal consistency of downstream continuous staging under sparse longitudinal supervision.

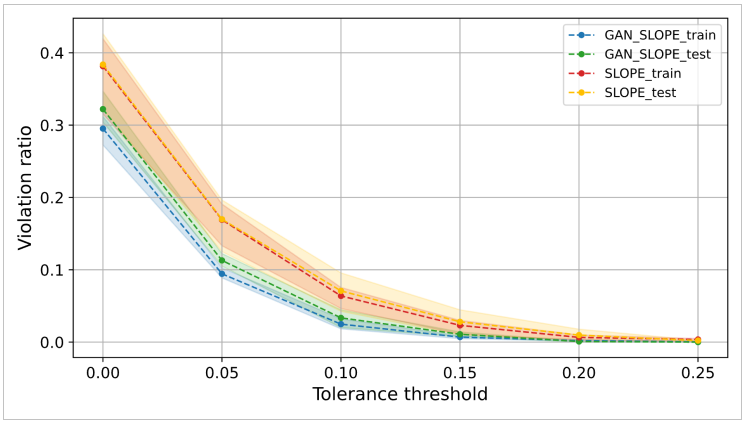

Fig. 4. Violation ratio across different tolerance thresholds. Shaded areas represent the variance in performance measured over three runs with different random seeds, while dotted lines represent the average performance across the seeds.

5 Conclusion

In this study, we proposed a biologically constrained generative augmentation framework to enhance continuous disease staging in Alzheimer's disease under sparse longitudinal supervision. The residual GAN is trained cross-sectionally to generate plausible future samples guided by the monotonicity of amyloid accumulation, and these synthetic samples are incorporated into SLOPE to strengthen

staging inference. Extensive evaluation demonstrates that the generated samples exhibit structured consistency with real longitudinal changes and, more importantly, significantly reduce temporal violations in continuous staging on held-out test subjects across multiple random seeds. These findings suggest that principled generative augmentation can effectively improve the robustness and stability of continuous disease staging.

Acknowledgments. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: usc.edu.

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

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2623–2631 (2019)
2. Better, M.A.: Alzheimer's disease facts and figures. Alzheimers Dement **19**(4), 1598–1695 (2023)
3. Cody, K.A., Sokołowski, A., Johns, E., Guerra, L.M., Winer, J.R., Young, C.B., Younes, K., Dumitrescu, L., Archer, D.B., Durant, A., et al.: Characterizing amyloid and tau positron emission tomography-based stages across the clinical continuum. Alzheimer's & Dementia **22**(1), e71017 (2026)
4. Fischer, L., Parker, D., Maboudian, S., Fonseca, C., Tato-Fernández, C., Annen, L., Arunachalam, P., Bacci, J.R., Barboure, M., Capelli, S., et al.: Longitudinal biomarker studies in human neuroimaging: capturing biological change of alzheimer's pathology. Alzheimer's Research & Therapy (2025)
5. Fonteijn, H.M., Modat, M., Clarkson, M.J., Barnes, J., Lehmann, M., Hobbs, N.Z., Scahill, R.I., Tabrizi, S.J., Ourselin, S., Fox, N.C., et al.: An event-based model for disease progression and its application in familial alzheimer's disease and huntington's disease. NeuroImage **60**(3), 1880–1889 (2012)
6. Höglund, K., Kern, S., Zettergren, A., Börjesson-Hansson, A., Zetterberg, H., Skoog, I., Blennow, K.: Preclinical amyloid pathology biomarker positivity: effects on tau pathology and neurodegeneration. Translational psychiatry **7**(1), e995–e995 (2017)
7. Jack, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q.: Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. The Lancet Neurology **9**(1), 119–128 (2010)
8. Jack Jr, C.R., Bennett, D.A., Blennow, K., Carrillo, M.C., Dunn, B., Haeberlein, S.B., Holtzman, D.M., Jagust, W., Jessen, F., Karlawish, J., et al.: Nia-aa research framework: toward a biological definition of alzheimer's disease. Alzheimer's & dementia **14**(4), 535–562 (2018)

9. Jack Jr, C.R., Knopman, D.S., Jagust, W.J., Petersen, R.C., Weiner, M.W., Aisen, P.S., Shaw, L.M., Vemuri, P., Wiste, H.J., Weigand, S.D., et al.: Update on hypothetical model of alzheimer's disease biomarkers. Lancet neurology **12**(2), 207 (2013)
10. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
11. Prosser, L., Sudre, C.H., Oxtoby, N.P., Young, A.L., Malone, I.B., Manning, E.N., Pemberton, H., Walsh, P., Barkhof, F., Biessels, G.J., et al.: Biomarker pathway heterogeneity of amyloid-positive individuals. Alzheimer's & Dementia **20**(12), 8503–8515 (2024)
12. Tong, M., Mehfooz, F., Zhang, S., Wang, Y., Fang, S., Saykin, A.J., Wang, X., Yan, J., for the Alzheimer's Disease Neuroimaging Initiative: Learning a continuous progression trajectory of amyloid in alzheimer's disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring **18**(1), e70309 (2026)
13. Villemagne, V.L., Burnham, S., Bourgeat, P., Brown, B., Ellis, K.A., Salvado, O., Szoeke, C., Macaulay, S.L., Martins, R., Maruff, P., et al.: Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic alzheimer's disease: a prospective cohort study. The Lancet Neurology **12**(4), 357–367 (2013)
14. Young, A.L., Marinescu, R.V., Oxtoby, N.P., Bocchetta, M., Yong, K., Firth, N.C., Cash, D.M., Thomas, D.L., Dick, K.M., Cardoso, J., et al.: Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. Nature communications **9**(1), 4273 (2018)
15. Young, A.L., Oxtoby, N.P., Daga, P., Cash, D.M., Fox, N.C., Ourselin, S., Schott, J.M., Alexander, D.C.: A data-driven model of biomarker changes in sporadic alzheimer's disease. Brain **137**(9), 2564–2577 (2014)