ONLS: OPTIMAL NOISE LEVEL SEARCH IN DIFFUSION AUTOENCODERS WITHOUT FINE-TUNING

Zihan Wang
University of Edinburgh
zwang114@ed.ac.uk

ABSTRACT

An ideal counterfactual estimation should achieve balance of precise intervention and identity preservation. Recently, Classifier-Guided Diffusion Model is proven effective to produce realistic and minimal counterfactuals. However, perfect intervention is often challenging to find and requires tedious fine-tuning. In this work, we propose Optimal Noise Level Search (ONLS), which leverages statistics from the guidance to automatically capture the balance without any fine-tuning process or extra network design. We demonstrate that our ONLS could accurately identify the optimal noise level for counterfactual estimation. The optimal per-sample results further contribute to an overall performance enhancement across the dataset. Preprocessing, curated dataset, and code are released on our project page: https://github.com/ImNotPrepared/ONLS.

1 INTRODUCTION

Accurate simulation of aging process holds immense value in clinical and neuroscience research particularly for identifying age-related pathologies (López-Otín et al., 2013; DEV, 2022). A key challenge in this domain is the considerable inter-subject variation, as each individual exhibits a unique aging trajectory (Fernandez et al., 2024). Traditional generative models like GANs (Goodfellow et al., 2020; Liu et al., 2022b; Campello et al., 2022) have been limited in addressing this variability. In contrast, recent advancements in diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b; Chen et al., 2024) have shown promise potential in synthesis tasks (Dhariwal & Nichol, 2021). Diff-SCM (Sanchez & Tsaftaris, 2022) leverages advances in generative energy-based models. The core mechanism initially infers latent variables through a deterministic forward diffusion process. This is followed by an intervention using a reverse diffusion mechanism, guided by the gradients of an anti-causal predictor with respect to the input. This effectively transitions from a state of noise towards the original data distribution. The procedure for recovering manipulable latent spaces from observations implies a connection with autoencoders. This insight provides a perspective to regard the noised image at each timestep t as a latent vector representation.

Prior research (Yang et al., 2022; Kascenas et al., 2023) shows that coarse features are reconstructed in early stages (near $t = T$), while fine-grained features are reconstructed in later stages (near $t = 0$). To utilize intermediate representations with different levels of feature granularity (Liu et al., 2022a; Wang et al., 2024), noise level L is introduced, altering the encoding depth from $t = 0$ to $t = L \times T$ instead of T. Therefore, the decoding process begins with a latent vector at $t = L \times T$, where $0 < L < 1$. We find that an appropriate noise level L is crucial to successful counterfactual estimation. A small L tends to retain original information as the intervention is not sufficient. On the contrary, identity information will be lost or biased given a large L. We assume a proper L facilitates the preservation of advantageous information, serving as prior knowledge to stabilize the manifold and consequently enhance performance. L is treated as a hyperparameter in previous work (Dutt et al., 2023; Sanchez et al., 2022a; Wolleb et al., 2022). However, we observed that the optimal L varies among samples and datasets, making fine-tuning both tedious and time-consuming. We propose a simple yet effective method, **Optimal Noise Level Search (ONLS)**, which requires neither fine-tuning nor additional network design. ONLS contributes to the efficiency and robustness of the model, making it more adaptable and efficient under varying data conditions.

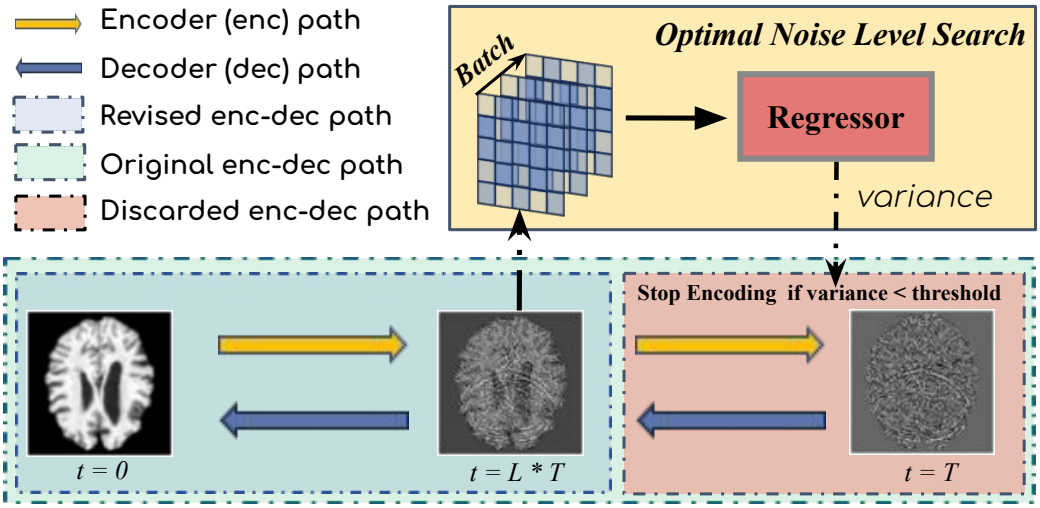

Figure 1: **Mechanism of ONLS:** During inference, the regressor records the variance within the batch during encoding. When the variance drops below the threshold ζ at $t = t_0$, the encoding stops and starts decoding with the guidance of regressor from $t = t_0$ instead of gaussian noise at $t = T$.

2 Methodology and Evaluation

Motivated by two facts: 1) As noise progressively added during encoding, information is 'destroyed' from fine-grained features to coarse features, eventually resulting in pure noise. 2) classifer/regressor essentially is a mapping from class information to labels. A natural assumption is that there exists an optimal balance between precise intervention and identity preservation. At this specific point, class-related information has been removed, allowing for precise intervention. Concurrently, the maximum preservation of identity is ensured, maintaining all non-class-related information intact.

We assume the balance can be found by leveraging class statistics. Subsequently, we incorporate the guidance regressor trained on noisy data into the encoding process. We observe its output trends within a batch as in Fig 2. Notably, when the class information is insufficient, the regressor tends to output uniformly, signifying the variance diminishes in the late encoding stage. The variance reaches minimum at $t = T$, where the class information is entirely eliminated. We thus stop further encoding when the variance drops below the threshold ζ at $t = t_0$, where class information is removed while preserving maximum identity information. Then decoding starts from the optimal representation determined by ONLS from $t = t_0$. Eventually, the model produces the counterfactual result guided by the regressor. Threshold ζ is contingent upon the data and counterfactual feature. The elegance of ONLS lies in its simplicity without any extra design modifications. Detailed validation is in Sec. C.

Method	SSIM(↑)	MSE(↓)	PSNR(↑)
Diff-SCM (Sanchez & Tsaftaris, 2022)	0.69±0.09	0.037±0.015	20.9±3.3
Tian et al. (Xia et al., 2021)	0.71±0.06	0.026±0.023	23.3±2.2
Optimal Noise Level Search (Ours)	**0.74±0.06**	**0.024±0.012**	**24.7±2.5**

Table 1: Quantitative evaluation on ADNI testset. Experiment details can be found in Sec. B, E

3 Conclusion

Our work illustrates the insight of noise level in diffusion autoencoder. Then we introduces ONLS to adeptly balances precision in intervention with identity preservation, thereby enhancing the effectiveness of diffusion autoencoder. The key contribution lies in the automation of the fine-tuning process to identify the perfect amount of intervention, a task marked by its complexity and laborious nature. We believe ONLS can serve as an efficient and accurate method benefiting the community.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

A predictive analytics approach for stroke prediction using machine learning and neural networks. *Healthcare Analytics*, 2:100032, 2022. ISSN 2772-4425.

Víctor Campello, Tian Xia, Xiao Liu, Pedro Sanchez, Carlos Martín-Isla, Steffen Petersen, Santi Seguí, Sotirios Tsaftaris, and Karim Lekadir. Cardiac aging synthesis from cross-sectional data with conditional generative adversarial networks. *Frontiers in Cardiovascular Medicine*, 9:983091, 09 2022. doi: 10.3389/fcvm.2022.983091.

Hao Chen, Jindong Wang, Zihan Wang, Ran Tao, Hongxin Wei, Xing Xie, Masashi Sugiyama, and Bhiksha Raj. Learning with noisy foundation models, 2024.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Raman Dutt, Linus Ericsson, Pedro Sanchez, Sotirios A. Tsaftaris, and Timothy Hospedales. Parameter-efficient fine-tuning for medical image analysis: The missed opportunity, 2023.

Virginia Fernandez, Pedro Sanchez, Walter Hugo Lopez Pinaya, Grzegorz Jacenków, Sotirios A. Tsaftaris, and M. Jorge Cardoso. Privacy distillation: Reducing re-identification risk of diffusion models. In Anirban Mukhopadhyay, Ilkay Oksuz, Sandy Engelhardt, Dajiang Zhu, and Yixuan Yuan (eds.), *Deep Generative Models*, pp. 3–13, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-53767-7.

Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Antanas Kascenas, Pedro Sanchez, Patrick Schrempf, Chaoyang Wang, William Clackett, Shadia S. Mikhael, Jeremy P. Voisey, Keith Goatman, Alexander Weir, Nicolas Pugeault, Sotirios A. Tsaftaris, and Alison Q. O'Neil. The role of noise in denoising models for anomaly detection in medical images. *Medical Image Analysis*, 90:102963, 2023. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2023.102963. URL https://www.sciencedirect.com/science/article/pii/S1361841523002232.

Yijie Li, Hewei Wang, Zhenqi Li, Shaofan Wang, Soumyabrata Dev, and Guoyu Zuo. Daanet: Dual attention aggregating network for salient object detection*. In *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 1–7, 2023. doi: 10.1109/ROBIO58561.2023.10354933.

Xiao Liu, Pedro Sanchez, Spyridon Thermos, Alison Q. O'Neil, and Sotirios A. Tsaftaris. Learning disentangled representations in the imaging domain. *Medical Image Analysis*, 80:102516, 2022a. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102516. URL https://www.sciencedirect.com/science/article/pii/S1361841522001633.

Xiao Liu, Spyridon Thermos, Pedro Sanchez, Alison Q. O'Neil, and Sotirios A. Tsaftaris. Hsic-infogan: Learning unsupervised disentangled representations by maximising approximated mutual information, 2022b.

Carlos López-Otín, Maria A Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. The hallmarks of aging. *Cell*, 153(6):1194–1217, 2013.

Andreas S. Panayides, Amir Amini, Nenad D. Filipovic, Ashish Sharma, Sotirios A. Tsaftaris, Alistair Young, David Foran, Nhan Do, Spyretta Golemati, Tahsin Kurc, Kun Huang, Konstantina S. Nikita, Ben P. Veasey, Michalis Zervakis, Joel H. Saltz, and Constantinos S. Pattichis. Ai in medical imaging informatics: Current challenges and future directions. *IEEE Journal of Biomedical and Health Informatics*, 24(7):1837–1857, 2020. doi: 10.1109/JBHI.2020.2991043.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in Neural Information Processing Systems*, 33:857–869, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Pedro Sanchez and Sotirios A Tsaftaris. Diffusion causal models for counterfactual estimation. *arXiv preprint arXiv:2202.10166*, 2022.

Pedro Sanchez, Antanas Kascenas, Xiao Liu, Alison Q O'Neil, and Sotirios A Tsaftaris. What is healthy? generative counterfactual diffusion for lesion localization. In *MICCAI Workshop on Deep Generative Models*, pp. 34–44. Springer, 2022a.

Pedro Sanchez, Jeremy P. Voisey, Tian Xia, Hannah I. Watson, Alison Q. ONeil, and Sotirios A. Tsaftaris. Causal machine learning for healthcare and precision medicine, 2022b.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Hewei Wang, Yijie Li, Shijia Xi, Shaofan Wang, Muhammad Salman Pathan, and Soumyabrata Dev. Amdcnet: An attentional multi-directional convolutional network for stereo matching. *Displays*, 74:102243, 2022. ISSN 0141-9382. doi: https://doi.org/10.1016/j.displa.2022.102243.

Zihan Wang, Bowen Li, Chen Wang, and Sebastian Scherer. Airshot: Efficient few-shot detection for autonomous exploration, 2024.

Julia Wolleb, Robin Sandkühler, Florentin Bieder, and Philippe C Cattin. The swiss army knife for image-to-image translation: Multi-task diffusion models. *arXiv preprint arXiv:2204.02641*, 2022.

Tian Xia, Agisilaos Chartsias, Sotirios A Tsaftaris, Alzheimer's Disease Neuroimaging Initiative, et al. Consistent brain ageing synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 750–758. Springer, 2019.

Tian Xia, Agisilaos Chartsias, Chengjia Wang, Sotirios A Tsaftaris, Alzheimer's Disease Neuroimaging Initiative, et al. Learning to synthesise the ageing brain without longitudinal data. *Medical Image Analysis*, 73:102169, 2021.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 2022.

Zhipeng Zhao, Huai Yu, Chenwei Lyu, Pengliang Ji, Xiangli Yang, and Wen Yang. Cross-modal 2d-3d localization with single-modal query. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 6171–6174, 2023. doi: 10.1109/IGARSS52108.2023.10282358.

A THEORY

A.1 COUNTERFACTUAL ESTIMATION WITH DIFF-SCM

Following Pearl's Causal Hierarchy (Glymour et al., 2016), estimation of counterfactuals requires three steps: (i) abduction of exogenous noise - forward diffusion with DDIM (Song et al., 2020a); (ii) action - graph mutilation by erasing the edges between the intervened variable and its parents; (iii) prediction - reverse diffusion controlled by the gradients of an anti-causal classifier. Here, we are interested in estimating $x_{\mathrm{CF}}^{(k)}$ based on the observed (factual) $x_{\mathrm{F}}^{(k)}$ for the random variable $\mathbf{x}^f(k)$ after assigning a value $x_{\mathrm{CF}}^{(j)}$ to $\mathbf{x}^{(j)}$, i.e. applying an intervention $do\left(\mathbf{x}^{(j)} = x_{\mathrm{CF}}^{(j)}\right)$, which is equivalent to sample from counterfactual distribution $p_{\mathfrak{B}}\left(\mathbf{x}^{(k)} \mid do\left(\mathbf{x}^{(j)} = x_{\mathrm{CF}}^{(j)}\right); \mathbf{x}^{(k)} = x_{\mathrm{F}}^{(k)}\right)$.

A.2 CONNECTION WITH DIFFUSION MODEL

With diffusion models, abduction can be done with a derivation following (Song et al., 2020a; Li et al., 2023), which bridge a connection between diffusion models and neural ODEs (Chen et al., 2018). They show that one can obtain a deterministic inference system while training with a naturally stochastic diffusion process. One important property is that the sampling process only needs the gradient of an anti-causal predictor w.r.t. the effect when the cause is assigned a specific value.

A.3 CONNECTION WITH AUTOENCODER

Inspired by the fact that the exogenous $\mathbf{u}^{(k)}$ could be considered as latent variable from deep perspective (Pawlowski et al., 2020; Zhao et al., 2023), we can build a connection between diffusion autoencoder(Dhariwal & Nichol, 2021) and counterfactual estimation. The process of abduction and strengthening causal relations equals to encoding and decoding different feature information.

B ALGORITHM

Our theory is substantiated through the task of brain aging synthesis. Given an input brain image x, along with its corresponding actual age a_{org}, our counterfactual task entails simulating the appearance of this brain image at a target age y. The expected output is the image x_{CF}, estimated with counterfactual analysis at age y.

Algorithm 1 Guidance with Regressor

Input: input image x, desired age y, constant gradient scale c, noise level L
Output: counterfactual image $x_{0,\mathrm{CF}}^{(k)} = x_0^{(k)}$

Abduction of Exogenous Noise
for $t = 0$ **to** $L \times T$ **do**
$x_{t+1,\mathrm{F}}^{(k)} \leftarrow \sqrt{\alpha_{t+1}}\left(\frac{x_{t,\mathrm{F}}^{(k)} - \sqrt{1-\alpha_t}\epsilon_\theta(x_{t,\mathrm{F}}^{(k)}, t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t+1}}\epsilon(x_{t,\mathrm{F}}^{(k)}, t)$
end for
$x_{L\times T}^{(k)} = x_{L\times T,\mathrm{F}}^{(k)}$

Generation under Intervention
for $t = L \times T$ **to** 1 **do**
$s_r \leftarrow (y - R(x_t^{(k)}, t)) \cdot \sqrt{R(x_t^{(k)}, t)}$
$\hat{\epsilon} \leftarrow \epsilon_\theta(x_t^{(k)}, t) - s_t c\sqrt{1-\bar{\alpha}_t}\nabla_{x_t^{(k)}} R(x_t^{(k)}, t)$
$x_{t-1}^{(k)} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\alpha_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\epsilon}$
end for
$x_{0,\mathrm{CF}}^{(k)} = x_0^{(k)}$

C VARIANCE VISUALIZATION

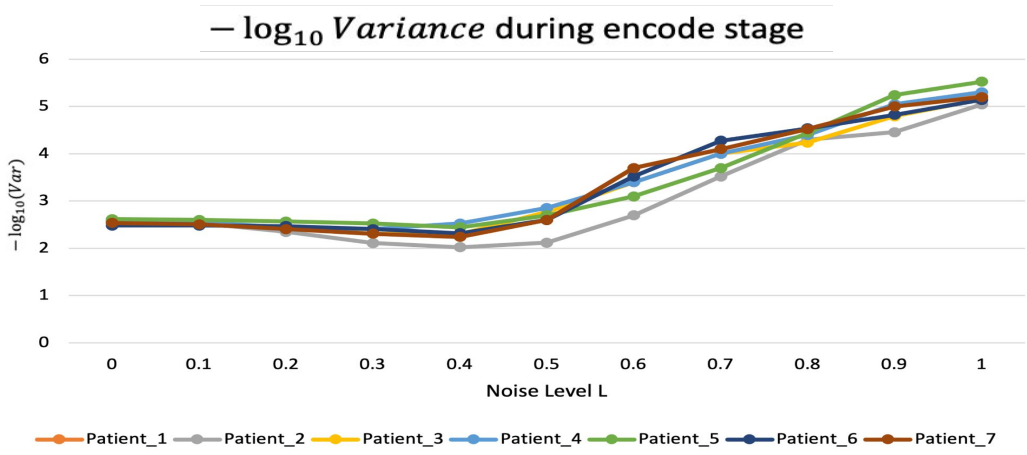

Figure 2: Variance trend as encode progressing from $\mathbf{t}(0)$ to $\mathbf{t}(T)$

We pick 7 random samples and visualize the variance within batch as encoding goes deeper. A similar trend of uniform value output can be observed in Fig. 2. Previous methods adjust noise globally for entire datasets, possibly not ideal for each sample. Our ONLS method automatically finds the optimal noise level for each sample, improving results for most sample pairs and enhancing overall dataset performance. To validate ONLS, we manually tuned L with 0.1 precision from 0.4 to 1.0 (where shows observable changes) and compare the results on our testset in Table 2.

Table 2: Quantitative comparison of different manually set noise level L and ONLS

Metrics	$L = 1.0$	$L = 0.9$	$L = 0.8$	$L = 0.7$	$L = 0.6$	$L = 0.5$	$L = 0.4$	**ONLS**
SSIM	0.69	0.72	0.72	0.71	0.70	0.70	0.67	**0.74**
MSE	0.037	0.031	0.027	0.029	0.032	0.049	0.043	**0.024**
PSNR	20.9	22.8	23.6	22.4	21.7	21.5	18.8	**24.7**

D QUALITATIVE RESULT

Our qualitative result shows a ventricular enlargement and barely no structural differences.

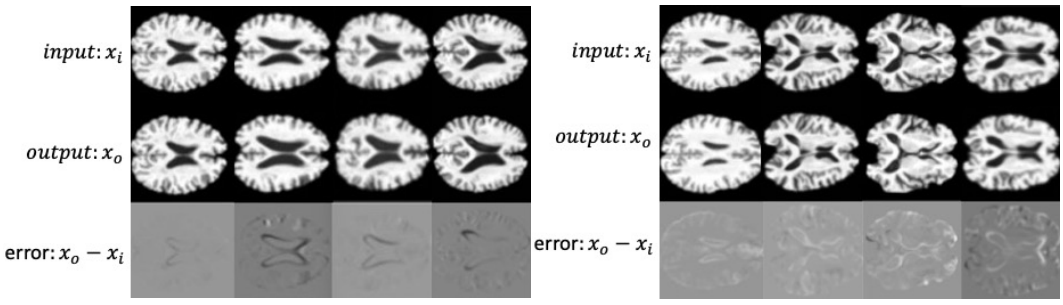

Figure 3: Effect illustration for both **ageing(left)** and **rejuvenation(right)** process.

E EXPERIMENTS SETUP

E.1 PRE-PROCESSING

For a fair comparison, we employed the same methodology as in Xia et al. (2019; 2021). Each image was clipped to the range $[0, \mathbf{V}_{99.5}]$, where $\mathbf{V}_{99.5}$ represents the 99.5% highest intensity value. Subsequently, these values were rescaled within the range $[-1, +1]$. Then we selected the central 60 axial slices, cropping each to a resolution of $[208, 160]$, which was adopted for subsequent evaluations. The pre-processing code can be found on our project website.

We randomly select 112 patience as testset. We first abandon those longitudinal data with severe misalignment problems (i.e. SSIM < 0.6), then choose the age span longer than 2 years to allow obvious changes. Each patient has 30 slices, we feed all as a batch to our network for inference.

E.2 IMPLEMENTATION

ϵ_θ is implemented as an Unet-like network with skip-connections and attention modules (Ronneberger et al., 2015; Wang et al., 2022; Panayides et al., 2020). For anti-causal classification tasks, we use the regressor. For training, we only use cross-sectional data (i.e. For each patient, we only use the images of a certain age to train on) to match real-world scenerio.

E.3 EVALUATION

The strongest validation is the groundtruth longitudinal data, thus we use the longitudinal data cover a limited time span from ADNI dataset to verify the performance. Throughout the experiment, we used standard definition of **mean squared error (MSE)**, **peak signal-to-noise ratio (PSNR)** and **structural similarity (SSIM)** of window length of 11 (following same configuration) to evaluate the closeness of the predicted images to the ground-truth and compare the performance with other methods. Given counterfactual image x_{CF} and groudtruth image x_{gt}, each metric is computed by:

$$\text{Metric}(x_{\text{CF}}, x_{\text{gt}}) \quad \text{where Metric} = \text{MSE, PSNR, SSIM} \tag{1}$$

E.4 HYPERPARAMETERS

For our specific task, we empirically set ζ to $\mathbf{1e^{-4}}$ by fine-tuning on testset to find the optimal value. For the training and network design hyperparameters, we refer readers to our github repository.

F LONG-TERM INTERVENTION

To illustrate ONLS achieves a balance between intervention and preservation, instead of simply neglect the intensity of intervention, we show some results regrading long-term intervention (Sanchez et al., 2022b). Due to lack of in of longitudinal data, we only present qualitative results covering 20-year span below:

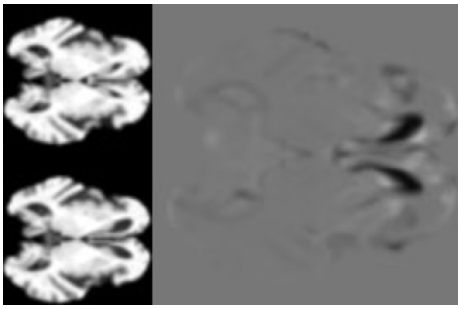

Figure 4: ageing from 57 to 77

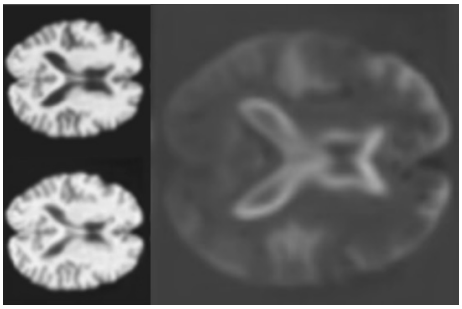

Figure 5: revitalization from 87 to 67