Joint learning framework of cross-modal synthesis and diagnosis for Alzheimer's disease by mining underlying shared modality information ——Supplementary Material——

Contents

A. Network architectures

A.1. Discriminator

We use two modality discriminators, D_{m} and D_{p}, to discriminate the authenticity of synthesized images. In Fig. S1, both discriminators are designed based on PatchGAN [8] using six convolutional layers with $3 \times 3 \times 3$ kernel. The first five convolution layers have a stride of size 2, and the last layer's stride is set to 1 to generate the final feature map of size $8 \times 8 \times 8$. Both discriminators are trained using the adversarial loss $\mathcal{L}_{\mathrm{GAN}}$ in the form of LSGAN loss [13], as described in Eq. (9) of the manuscript.

A.2. Self-attention block

The self-attention mechanism [26] is widely used in various visual [6, 9, 18] and language tasks [5, 21] because it has been shown to effectively learn global interactions (*i.e.* relations between distant object parts) [14]. Here, we introduce a fully convolution-implemented self-attention block at the bottleneck of the 3D Unet-based synthesis model in Fig. 3 of the manuscript.

For an N-heads self-attention block, we use $\boldsymbol{F} \in \mathbb{R}^{D\times H\times W\times C_{in}}$ and $\boldsymbol{G} \in \mathbb{R}^{D\times H\times W\times C_{out}}$ to represent the input and output features. Let $\boldsymbol{f}_{rst} \in \mathbb{R}^{C_{in}}$, $\boldsymbol{g}_{rst} \in \mathbb{R}^{C_{out}}$ denote the corresponding vector of voxel (r, s, t). Multi-head self-attention can be decomposed into two steps [15]:

$$
\begin{aligned}
&\textbf{Step I}: \boldsymbol{q}_{rst}^{(l)} = \boldsymbol{W}_q^{(l)}\boldsymbol{f}_{rst}, \boldsymbol{k}_{rst}^{(l)} = \boldsymbol{W}_k^{(l)}\boldsymbol{f}_{rst}, \boldsymbol{v}_{rst}^{(l)} = \boldsymbol{W}_v^{(l)}\boldsymbol{f}_{rst},\\
&\textbf{Step II}: \boldsymbol{g}_{rst} = \underset{l=1}{\overset{N}{\mathrm{Concat}}}(head_1, head_2, \ldots, head_N),\\
&\text{where} \quad head_l = \sum_{a,b,c\in\mathcal{N}_K(r,s,t)} \underset{\mathcal{N}_k(r,s,t)}{\mathrm{Softmax}}\left(\frac{\boldsymbol{q}_{rst}^{(l)}\boldsymbol{k}_{abc}^{(l)\,T}}{\sqrt{d}}\right)\boldsymbol{v}_{abc}^{(l)}.
\end{aligned}
\tag{S1}
$$

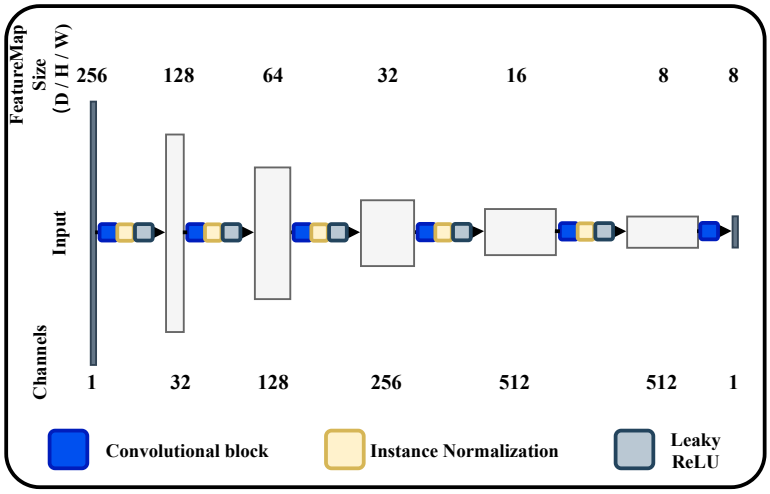


Figure S1. Illustration of our discriminator.

In **Step I**, $1 \times 1 \times 1$ convolutions are first conducted to project the input feature as query, key, and value. $\boldsymbol{W}_q^{(l)}$, $\boldsymbol{W}_k^{(l)}$, and $\boldsymbol{W}_v^{(l)}$ are the projection matrices for the queries, keys, and values, respectively. In **Step II**, $\mathcal{N}_K(r, s, t)$ represents a local region of voxels with spatial extent K centered around (r, s, t), and d is the feature dimension of $\boldsymbol{q}_{rst}$; this step comprises the calculation of the attention weights and aggregation of the value matrices, which refers to gathering local features.

B. Pre-processing details

The T1 weighted MRI images used in this paper are from ADNI, AIBL, and NACC three datasets, while the 18-FDG PET images are only from the ADNI dataset. All modality data from different datasets were obtained in NIFTI format, and the MRI and PET data used in the ADNI dataset have been gone through the full ADNI pre-processing pipelines. Our cross-modal synthesis network is not evaluated on the external datasets since the PET image scanning protocol in the other datasets differs from that in the ADNI dataset. For example, the PET images available in AIBL are amyloid PET images, with a slice thickness of 2.0 or 3.0 mm for majority, while we used FDG PET images in ADNI for our experiments.

All MRI and PET in different datasets are preprocessed according to the corresponding modality preprocessing pipelines presented in Fig. S2. For MRI, all scans are preprocessed in six operations: (1) anterior commissure (AC) posterior commissure (PC) alignment, (2) intensity correction, (3) skull stripping, (4) registration to MNI152 template space (ICBM 2009c Nonlinear Symmetric template, McGill University, Canada), (5) clipping the outliers, and (6) normalization. The PET images are preprocessed only by the last four operations.

The first three operations in the MRI preprocessing pipeline are implemented via Freesurfer 5.3.0 [7], and the skull stripping operation for PET modality is implemented by s3 algorithms [11] based on ANTs 2.1.0 [1] software. The alignment to standard space is achieved via the FLIRT 6.0 tool available within the FSL package (Wellcome Center, University of Oxford, UK) for both modalities. In the fifth operation, we clip outliers for MRI and PET modalities based on [20]'s method. In addition, we encountered many failure cases when preprocessing on our institute-owned data platform, especially for 1.5 Tesla MRI images of poor quality. Common preprocessing failures [3, 4] include incomplete skull stripping, registration failure, image artifacts, and others. To address this issue, we conducted a manual visual quality assurance process and carefully removed the failed samples before proceeding with our experiments. Then, we clip all the MRI preprocessed by the previous operations to the range: $[0, 130]$ and $[0, 1.7]$ for the PET by statistical analysis. All voxels smaller than the minimum value in the specified range are assigned the corresponding minimum value. Voxels larger than the maximum value are also operated similarly. To maintain the original data properties as much as possible, we use min-max normalization of all the data to the range $[-1, 1]$ in the sixth operation. Finally, we fill the surrounding area with black, resulting in the final image size of $256 \times 256 \times 256$.

For better reproducibility, we now make all the subject Ids involved in our experiment publicly available at https://github.com/thibault-wch/Joint-Learning-for-Alzheimer-disease.

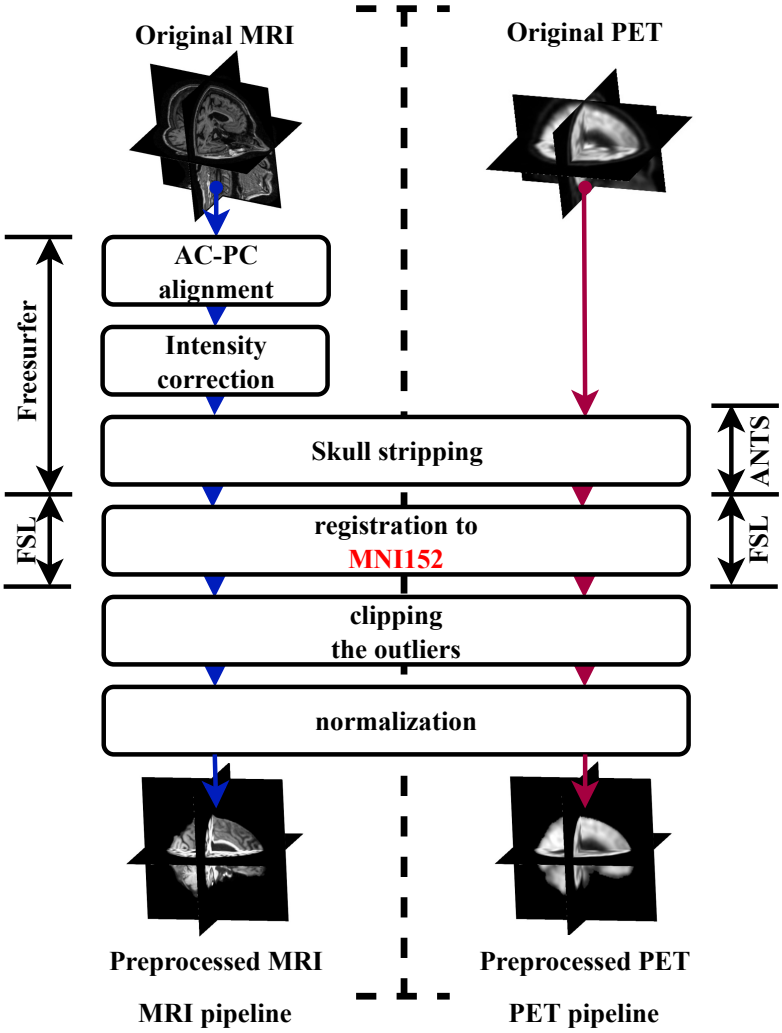


Figure S2. Illustration of pre-processing pipelines for MRI and PET.

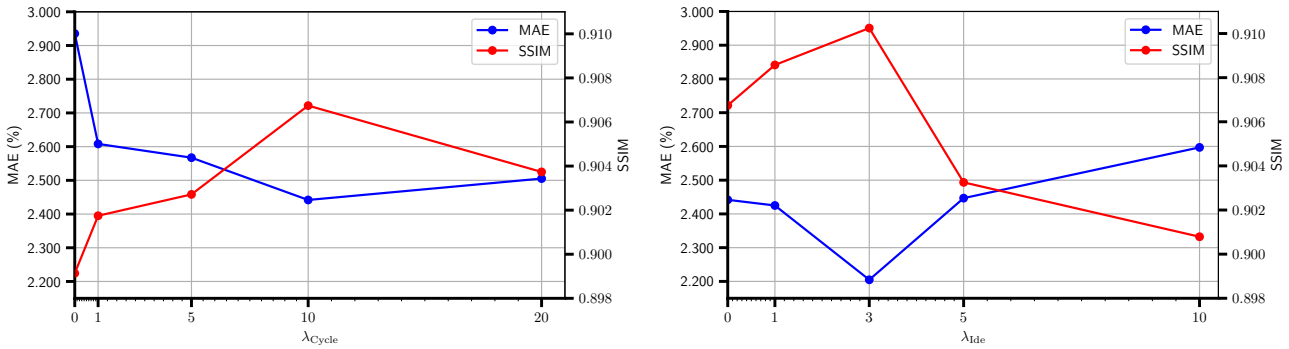
C. More experimental results

C.1. Hyperparameter selection for loss function

We progressively adjusted the loss weights according to image quality metrics on the ADNI validation set in a greedy manner instead of computationally expensive grid searching. Fig. S3 shows the MAE and SSIM between synthesized and real images in the validation set; we note that a better image quality corresponds to lower MAE and higher SSIM. We first fixed the weight for the adversarial loss $\mathcal{L}_{\text{GAN}}$ as 1, and then determined the weight λ_{Cycle} for the cycle consistency loss $\mathcal{L}_{\text{Cycle}}$, followed by the weight λ_{Ide} for the identity loss $\mathcal{L}_{\text{Ide}}$. By adjusting λ_{Cycle} in Fig. S3(a), we find that $\lambda_{\text{Cycle}} = 10$ gives the best performance among all candidates. With fixed $\lambda_{\text{Cycle}} = 10$, we further adjust λ_{Ide} in Fig. S3(b) and find that $\lambda_{\text{Ide}} = 3$ gives the best performance among all candidates. Furthermore, in our experiment, the weight λ_{Cls} of the diagnosis loss $\mathcal{L}_{\text{Cls}}$ was set to 1 for joint training.

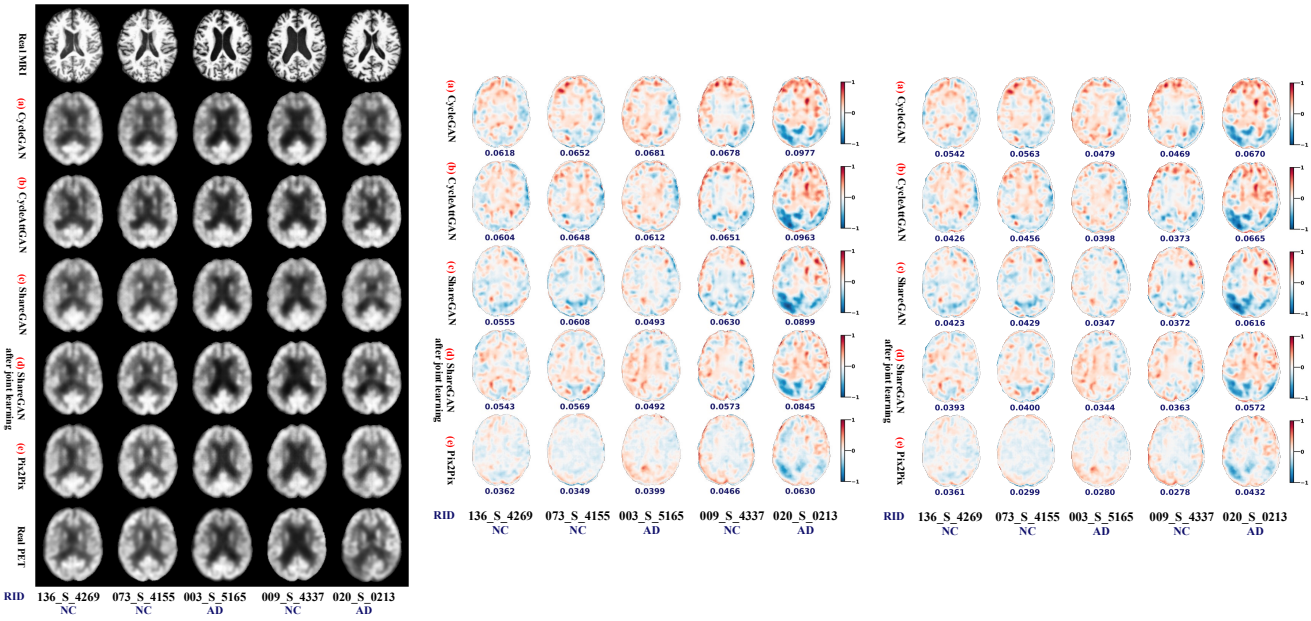
C.2. More cross-modal synthesized results

We randomly select five subjects from the ADNI dataset and present cross-modal synthesized images, along with the corresponding Standardized Uptake Value Ratio (SUVR) and pixel error maps in Fig. S4. Note that the pons [24] is selected as the reference region for calculating the SUVR. The synthesized images in Fig. S4(a) validate that our cross-modal synthesis network can produce more realistic and reasonable images. Of note, both SUVR and pixel error maps in Figs. S4(b) and S4(c)



(a) (b)

Figure S3. The effect of the weights λ_{Cycle} and λ_{Ide} in our method on the ADNI validation set. (a) The effect of the weight λ_{Cycle}, **(b)** The effect of the weight λ_{Ide}.



(a) (b) (c)

Figure S4. Cross-modal synthesized results and SUVR error maps of the proposed and other comparative networks on the ADNI testing set. (a) More cross-modal synthesized results; **(b)** The SUVR error map ($\boldsymbol{x}_p - \boldsymbol{x}_p^*$) results between the real PET images $\boldsymbol{x}_p$ and synthesized PET images $\boldsymbol{x}_p^*$; note that the value presented here is mean absolute SUVR error. **(c)** The pixel error map ($\boldsymbol{x}_p - \boldsymbol{x}_p^*$) results between the real PET images $\boldsymbol{x}_p$ and synthesized PET images $\boldsymbol{x}_p^*$; note that the value presented here is mean absolute pixel error.

consistently indicate that our network can effectively learn pixel-level corresponding information as well as the relationships between different brain regions. Furthermore, we can use the SUVR processing as a post-processing step for our synthesized PET images, allowing radiologists to select different reference regions and calculate SUVR maps for specific analysis. In addition, we also present the pixel error map of the Fig. 7 from the manuscript in Fig. S5. The findings align with the previous observations.

However, in some more complex images, even the supervised Pix2Pix method has difficulty in modeling some molecular functional information from the structural MRI image, such as the subject with RID 020_S_0213. Therefore, we should focus on more than the authenticity of synthesized images; that is, the synthesis task can be more meaningful to make the synthesized image have more discriminative information related to the downstream diagnosis task.

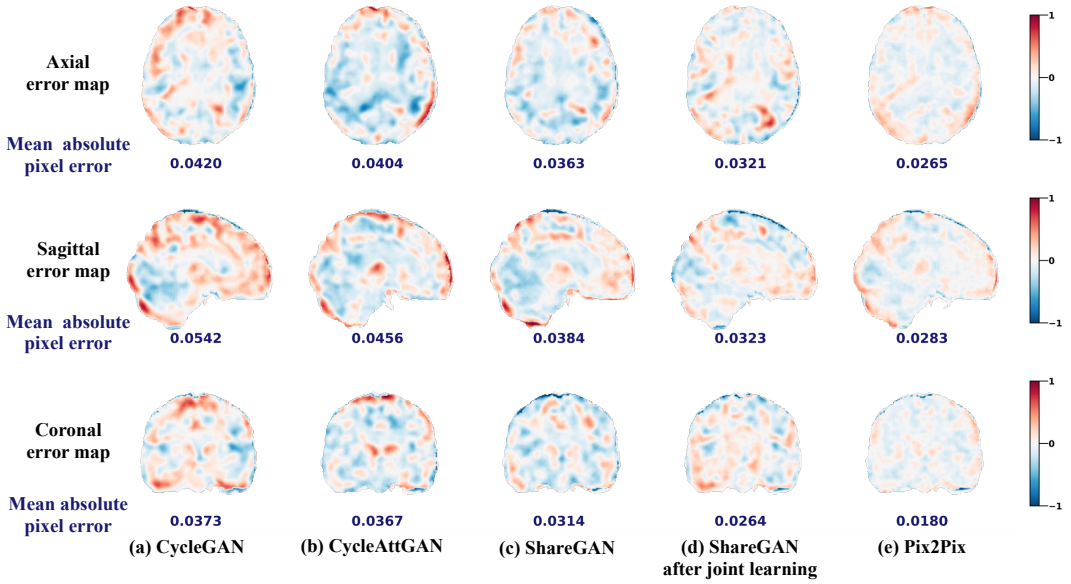


Figure S5. The pixel error map ($\boldsymbol{x}_p - \boldsymbol{x}_p^*$) results between the real PET images $\boldsymbol{x}_p$ and synthesized PET images $\boldsymbol{x}_p^*$.

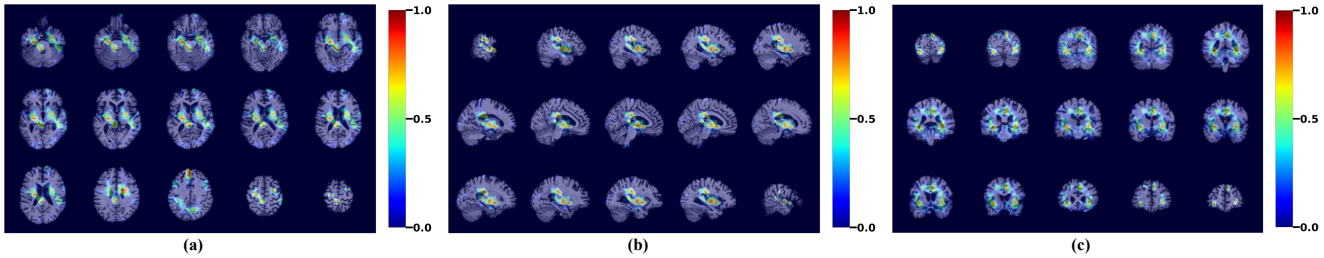


Figure S6. Grad-CAM interpretable results for three different slice-dividing directions of one brain randomly selected from the AIBL testing set; RID: 181 (NC). (a) Axial planes, **(b)** Sagittal planes, and **(c)** Coronal planes.

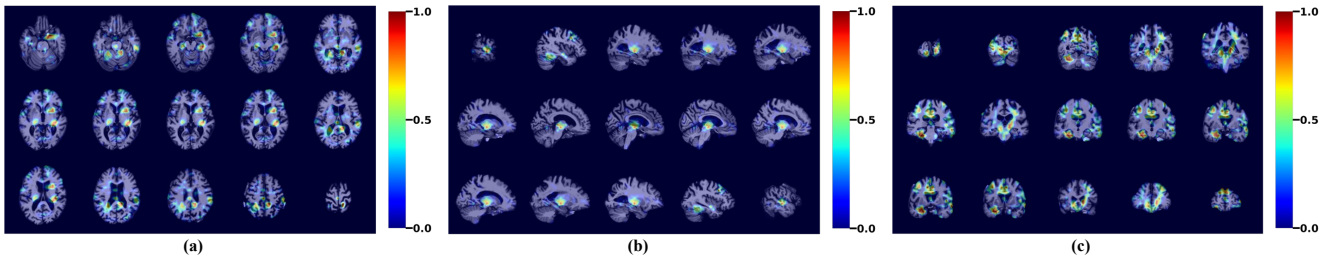


Figure S7. Grad-CAM interpretable results for three different slice-dividing directions of one brain randomly selected from the NACC testing set; RID: 5369 (AD). (a) Axial planes, **(b)** Sagittal planes, and **(c)** Coronal planes.

C.3. More interpretable results

The proposed diagnosis network trained on ADNI dataset is independently tested on the external AIBL and NACC datasets. Figs. S6 and S7 present Grad-CAM interpretable results for three different slice-dividing directions of one brain randomly selected from external AIBL and NACC datasets. We discover similarities in the ROIs from slice to slice due to the use of a shared 2D backbone and the proposed slice-shift module.

Moreover, following [19], we average all slices' heat maps along the trained slice-dividing direction and combine the brain automatic anatomical labeling atlas 3 [23] to present the focused regions of interest for the three subjects used in the manuscript. The results shown in Fig. S8 demonstrate that our 2D convolution-based diagnosis networks trained by different

directions largely focus on the same regions around the hippocampus, consistent with our previous findings. However, we would like to note that slight differences exist among the three planes due to different trained slice-dividing directions and individual differences.

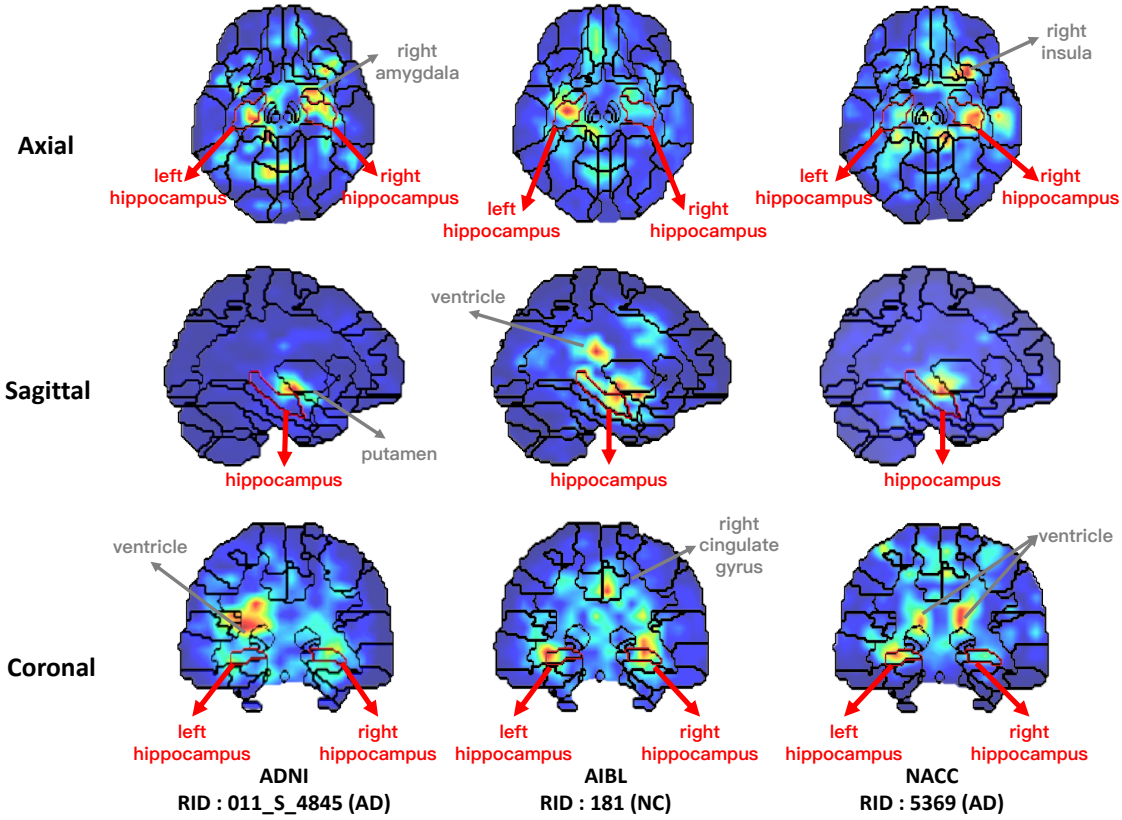


Figure S8. Grad-CAM averaged slices' heat maps derived from three models trained by different slice-dividing directions for the three subjects used in the manuscript.

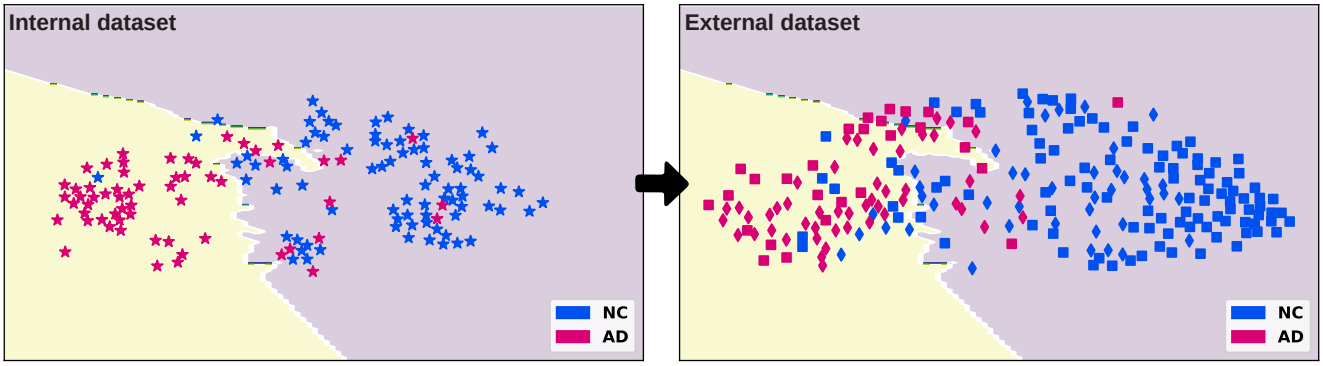
C.4. Visualization of data using t-SNE

Here, we perform t-distributed stochastic neighbor embedding (t-SNE) [25] to visualize the features before the last fully-connected layer of our diagnosis network on the internal ADNI and external AIBL and NACC testing sets in Figs. S9(a) and S9(b), respectively. We observed that our method could effectively classify subjects into two categories (NC and AD) using a shared decision boundary from internal to external generalization; here, following [22], we used k-NN ($k = 5$) to generate the decision boundary over the projected 2d internal ADNI testing data for the illustration of the generalization.

D. Detailed comparison with related works

Here, we present the comparison between related works [2, 10, 12, 16, 17] of leveraging PET to improve AD diagnosis from MRI and ours in Fig. S10. Related works usually contain two steps: the first is to synthesize the missing PET modality images, and the second is to use both synthesized PET and original MRI for so-called multi-modal diagnosis. However, during the entire workflow, the information for the synthesized PET images is all derived from the input MRI images, so no additional information is introduced for AD diagnosis; the so-called AD multi-modal diagnosis is essentially single-modal.

Unlike existing methods that take the synthesized PET images as an independent modality, our framework considers the synthesized PET images to be task-specific data augmentation of the input MRI images and feeds either synthesized PET or real MRI into the same diagnosis network to mine the underlying shared information between them. In addition, we jointly train the cross-modal synthesis and AD diagnosis networks to supervise each other, providing more discriminative diagnosis information in the synthesized images and improving the performance of AD diagnosis.



(a) Visualization of internal ADNI testing data

(b) Visualization of external AIBL and NACC testing data

Figure S9. The fully-connected layer outputs of our diagnosis network from different testing sets are embedded in a 2d plot using t-SNE. The color (blue versus red) is used to distinguish NC from AD, whereas a unique symbol shape is used to represent individuals derived from the same cohort. The '★' symbol indicates ADNI testing data, the '■' symbol indicates AIBL testing data, and the '♦' symbol indicates NACC testing data. Note that the decision boundary is generated by k-NN ($k = 5$) over the projected 2d internal ADNI testing data.

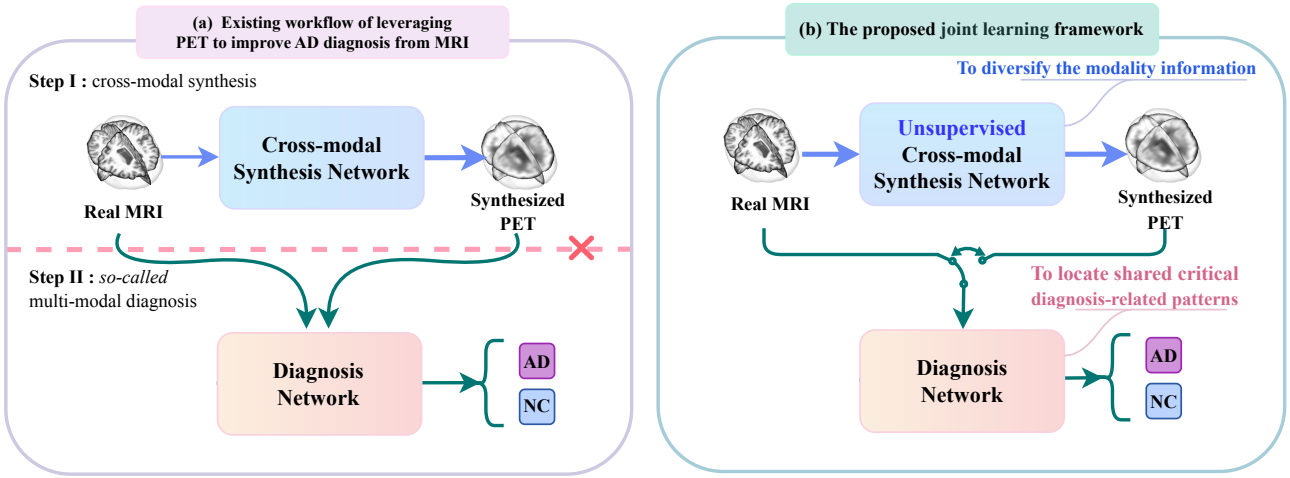


Figure S10. The comparison between related works of leveraging PET to improve AD diagnosis from MRI and our joint learning framework. (**a**) Related works separate cross-modal synthesis and diagnosis tasks, and use both so-called different modalities for AD diagnosis. (**b**) The proposed framework jointly learns unsupervised cross-modal synthesis and diagnosis tasks by mining shared modality information, and uses either modality for diagnosis.

References

[1] Brian B Avants, Nick Tustison, Gang Song, et al. Advanced normalization tools (ANTs). *Insight Journal*, 2(365):1–35, 2009.

[2] Sergio Campos, Luis Pizarro, Carlos Valle, Katherine R Gray, Daniel Rueckert, and Héctor Allende. Evaluating imputation techniques for missing data in ADNI: A patient classification study. In *Iberoamerican Congress on Pattern Recognition*, pages 3–10. Springer, 2015.

[3] Keith S Cover et al. Reproducibility of hippocampal atrophy rates measured with manual, FreeSurfer, AdaBoost, FSL/FIRST and the MAPS-HBSI methods in Alzheimer's disease. *Psychiatry Research: Neuroimaging*, 252:26–35, 2016.

[4] Mahsa Dadar, Vladimir S Fonov, D Louis Collins, Alzheimer's Disease Neuroimaging Initiative, et al. A comparison of publicly available linear MRI stereotaxic registration techniques. *NeuroImage*, 174:191–200, 2018.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 2018.

[6] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020.

[7] Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, 2012.

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017.

[9] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR-modulated detection for end-to-end multi-modal understanding. In *ICCV*, pages 1780–1790, 2021.

[10] Rongjian Li, Wenlu Zhang, Heung-Il Suk, Li Wang, Jiang Li, Dinggang Shen, and Shuiwang Ji. Deep learning based imaging data completion for improved brain disease diagnosis. In *MICCAI*, pages 305–312. Springer, 2014.

[11] Jana Lipková et al. Personalized radiotherapy design for glioblastoma: Integrating mathematical tumor models, multimodal scans, and bayesian inference. *Transactions on Medical Imaging*, 38(8):1875–1884, 2019.

[12] Yunbi Liu, Ling Yue, Shifu Xiao, Wei Yang, Dinggang Shen, and Mingxia Liu. Assessing clinical progression from subjective cognitive decline to mild cognitive impairment with incomplete multi-modal neuroimages. *Medical Image Analysis*, 75:102266, 2022.

[13] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2794–2802, 2017.

[14] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *NeurIPS*, pages 23296–23308, 2021.

[15] Xuran Pan, Chunjiang Ge, Rui Lu, Shiji Song, Guanfu Chen, Zeyi Huang, and Gao Huang. On the integration of self-attention and convolution. In *CVPR*, pages 815–825, June 2022.

[16] Yongsheng Pan, Mingxia Liu, Chunfeng Lian, Yong Xia, and Dinggang Shen. Spatially-constrained fisher representation for brain disease identification with incomplete multi-modal neuroimages. *Transactions on Medical Imaging*, 39(9):2965–2975, 2020.

[17] Yongsheng Pan, Mingxia Liu, Yong Xia, and Dinggang Shen. Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data. *Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6839–6853, 2021.

[18] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *ICCV*, pages 367–376, October 2021.

[19] Shangran Qiu et al. Multimodal deep learning for Alzheimer's disease dementia assessment. *Nature Communications*, 13(1):3404, 2022.

[20] Shangran Qiu, Prajakta S Joshi, Matthew I Miller, Chonghua Xue, Xiao Zhou, Cody Karjadi, Gary H Chang, Anant S Joshi, Brigid Dwyer, Shuhan Zhu, et al. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain*, 143(6):1920–1933, 2020.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[22] Francisco Caio M Rodrigues, Roberto Hirata, and Alexandru Cristian Telea. Image-based visualization of classifier decision boundaries. In *SIBGRAPI*, pages 353–360. IEEE, 2018.

[23] Edmund T Rolls, Chu-Chung Huang, Ching-Po Lin, Jianfeng Feng, and Marc Joliot. Automated anatomical labelling atlas 3. *NeuroImage*, 206:116189, 2020.

[24] Elina Thibeau-Sutre, Mauricio Diaz, Ravi Hassanaly, Alexandre Routier, Didier Dormont, Olivier Colliot, and Ninon Burgos. ClinicaDL: An open-source deep learning software for reproducible neuroimaging processing. *Computer Methods and Programs in Biomedicine*, 220:106818, 2022.

[25] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.