Hemisphere-based Local Feature Fusion from Multimodal Imaging for Interpretable AD Diagnosis

Zibo Zhao[1,†] Yanteng Zhang[2,†,*] Yuxiang Wei[2] Chuanyi Zhang[3] Qiang Liu[1] Vince Calhoun[2]

[1]College of Artificial Intelligence, Chengdu University of Information Technology, Chengdu, P.R.China
[2]Center for Translational Research in Neuroimaging and Data Science
(Georgia State, Georgia Tech, Emory), Atlanta, USA
[3]College of Artificial Intelligence and Automation, Hohai University, Changzhou, P.R.China

***Abstract*—Structural magnetic resonance imaging (sMRI) and positron emission tomography (PET), as the most commonly used imaging modalities for clinical diagnosis of Alzheimer's disease (AD), provide structural and functional information of the brain, respectively. However, multimodal methods still face challenges in AD prediction due to pronounced inter-individual heterogeneity and subtle pathological changes in the imaging manifestations. To address this issue, this work proposed an end-to-end deep learning framework based on the neuroanatomical characteristics of bilateral brain symmetry and the asymmetric distribution of AD pathology. First, brain images are divided into 3D regional patches according to the left and right hemispheres, and the features are encoded via patch CNNs. Subsequently, multi-head attentions are employed to optimize the representation of these local features among brain patch regions in each of the hemispheres. Finally, we developed a Hemisphere-aware Cross Transformer that performs hierarchical feature fusion at both intermodal and interhemispheric levels. Compared to several deep learning models, our proposed network achieved significant improvements in both AD diagnosis and early AD prediction on the ADNI dataset. More importantly, our approach achieves a breakthrough in interpretability, providing critical insights for the exploration of AD patterns in multimodal brain imaging.**

***Index Terms*—Hemisphere-based, Multimodal imaging, Hybrid neural network, Feature fusion, Alzheimer's diagnosis**

I. Introduction

AD, a progressive neurodegenerative disorder, stands as the predominant cause of dementia among the elderly population. This debilitating condition manifests through a spectrum of cognitive impairments, including memory loss, language deficits, executive dysfunction, and emotional disturbances. Despite extensive research, the etiology of AD remains enigmatic, and the medical community continues to grapple with the absence of curative pharmacological interventions [1]. Early intervention, however, has been shown to mitigate the progression of AD, underscoring the importance of accurate and timely diagnosis for patient [2].

In clinical practice, achieving an early and accurate diagnosis of AD remains a significant challenge. The current diagnostic workflow employs a multidimensional evaluation system that integrates various approaches, including neuropsychological scale assessments, cognitive-behavioral evaluations, and neuroimaging examinations. Among these, neuroimaging techniques have become indispensable tools for AD screening and diagnosis due to their advantages in objective visualization. However, given AD's heterogeneous pathological features and progressive nature, coupled with the inherent subjectivity in imaging interpretation, diagnostic results often show substantial variability among different radiologists when evaluating the same imaging data. This diagnostic inconsistency is particularly pronounced during the early stage of AD.

In terms of neuroimage, sMRI, as a non-invasive and radiation-free neuroimaging technique, is widely used in the clinical assessment of AD. sMRI captures characteristic structural alterations, typically revealing cortical thinning, ventricular enlargement, and hippocampal atrophy in AD brain imaging [3]. Meanwhile, 18F-fluorodeoxyglucose PET (FDG PET), a molecular imaging modality, provides direct visualization of cerebral glucose metabolism. This technique detects metabolic abnormalities in AD-vulnerable regions in the frontal lobe and temporal lobe, offering complementary metabolic information that cannot be obtained from sMRI alone [4]. Together, multimodal approaches provide multidimensional evidence for investigating the pathophysiological mechanisms of AD.

Recent advances in deep learning have substantially improved computer-aided diagnosis, particularly in multimodal analysis. Deep neural networks can automatically extract brain features and enhance diagnostic performance. However, brain AD pattern presents unique challenges: pathological changes are often subtle, such as mild atrophy or metabolic alterations, and vary considerably across subjects [5]. These factors complicate accurate feature representation. Single-modality sMRI analysis further struggles to capture such complex changes, limiting CNN-based models. ROI-based approaches offer localized insights but fail to fully characterize progressive, brain-wide alterations. Thus, multimodal imaging has attracted increasing attention for integrating complementary information, which improves whole-brain representation and performance [6]. Yet, effectively fusing heterogeneous multimodal data remains difficult. While attention mechanisms alleviate some limitations, their ability to capture complex

†These authors contributed equally to this work.
*Corresponding author: yzhang129@gsu.edu

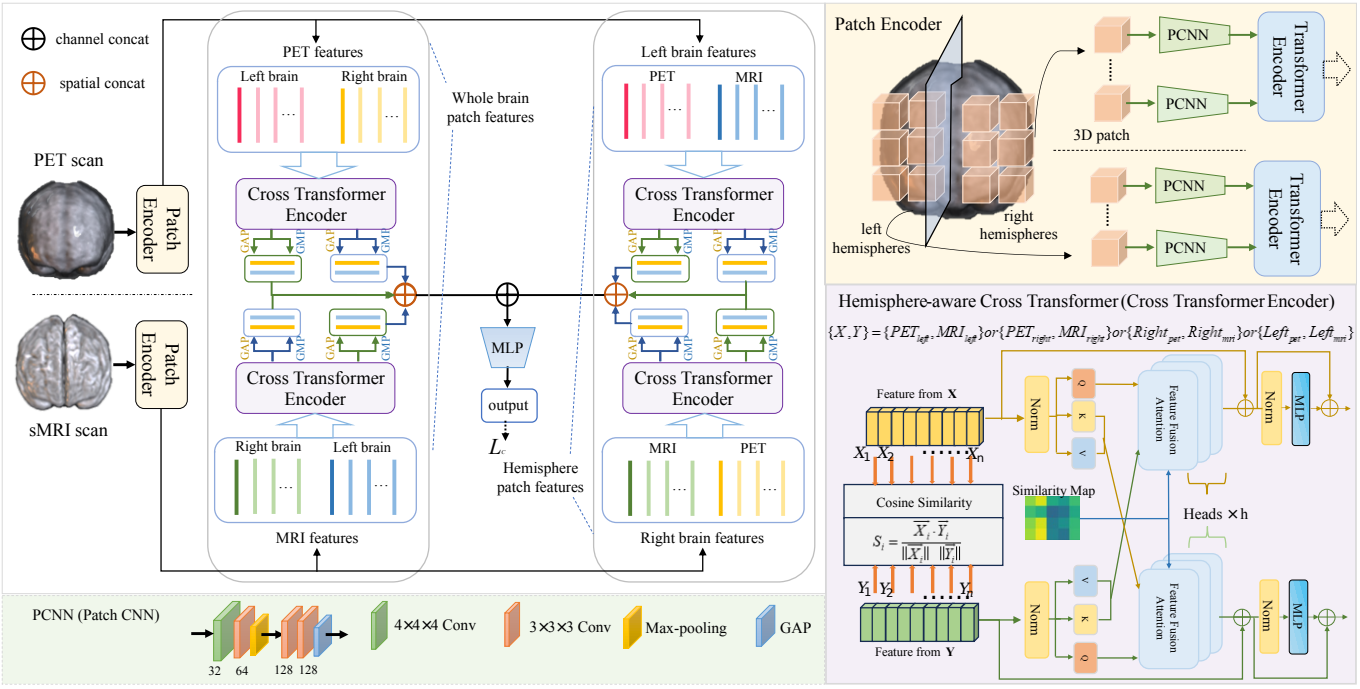

Fig. 1. The overall structure of the proposed hemisphere-based multimodal framework for AD diagnosis. This network employs a patch encoder to extract local features from two brain imaging modalities. Whole-brain multimodal fusion and cross-hemispheric feature fusion are conducted in parallel, capturing complementary information across modalities and hemispheres independently. The fused features are then aggregated and passed through MLP for final output.

and subtle brain patterns is still constrained [7]. Therefore, our study strategically focuses on multiple brain regions and seamlessly integrates patterns from modalities through bi-hemispheric to identify biomarkers, allowing a comprehensive characterization of AD image patterns.

II. Related Works

Deep neural networks play a crucial role in automatically extracting task-relevant features from sMRI [8] [9]. However, single modality imaging analysis still faces challenges, particularly in capturing subtle structural changes during the progression of AD. Due to the complexity of AD imaging pathology, traditional CNN-based diagnostic models exhibit certain limitations in the representation of the brain [10]. Meanwhile, some studies conducted feature extraction methods based on pre-selected regions of interest (ROIs) to analyze changes in specific brain regions [11] [12]. Since AD-related pattern changes typically involve progressive alterations across the whole brain, relying solely on ROIs may not comprehensively reflect the pathological features.

Consequently, multimodal approaches have garnered increasing attention and research interest due to their ability to integrate complementary information for AD assessments. Leveraging multimodal imaging not only enhances CNNs' capability in whole-brain feature extraction but also utilizes complementary information across different modalities, thereby improving the accuracy of AD diagnosis [13] [14]. However, due to the complexity of multimodal data and the heterogeneity among different imaging modalities, traditional CNNs still face challenges in processing such data. In particular, effectively integrating information from multiple modalities to capture the subtle differences in AD-related imaging changes remains a key research focus. The introduction of attention mechanisms has partially alleviated the limitations of CNNs, but their ability to capture complex brain patterns remains constrained [7]. Therefore, future research should focus on developing efficient multimodal deep learning models that not only precisely extract whole brain information but also effectively integrate multimodal information. This approach would enhance the comprehensive characterization of AD-related imaging changes, providing more robust technical support for automated AD diagnosis.

III. Methodology

The proposed framework is illustrated in Fig. 1. Patch CNNs encode brain-local features from all patches within each hemisphere. The set of feature vectors for each hemisphere is then processed separately by a transformer encoder to enhance the representation ability of the vectors. The Hemisphere-aware Cross Transformer is employed to fuse the hemispheric features within each modality and the dual-modal features for each hemisphere. This allows for both intra-modal and inter-hemispheric dependencies to be captured. Finally, the fused features from all components are passed through a three-layer Multi-Layer Perceptron (MLP) for classification.

A. Patch Encoder

The Patch Encoder module integrates 3D Patch CNN (PCNN) and Transformer Encoder to effectively capture both local and global brain features. First, the brain is divided into

3D patches, with each hemisphere (left and right) separately processed by 3D PCNN [9]. Each patch is processed independently, and the extracted features are transformed into a 128-dimensional feature vector. This encoding allows the model to focus on specific spatial regions of the brain, essential for the small spatial size of brain patches.

Then the encoded patch features from each hemisphere are fed into separate Transformer Encoder. Transformer [15] employs *multi-head self-attention* to capture long-range dependencies among brain patches. For each input feature $X \in R^{N\times D}$ (with N patches and feature dimension D), the query, key, and value matrices are obtained via linear projections.

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \tag{1}$$

where $W^Q, W^K, W^V \in R^{D\times d_k}$ are learnable parameters, and d_k is the query/key dimension. The scaled dot-product attention is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

Multi-head attention applies this operation in parallel, and the outputs are concatenated and linearly projected:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(hd_1, \ldots, hd_h)W^O \tag{3}$$

where each head is

$$hd_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{4}$$

and $W^O \in R^{hd_v\times D}$ maps the concatenated output back to dimension D.

B. Hemisphere-aware Cross Transformer

After encoding the two brain modalities with the Patch Encoder, we obtain four feature sets: left and right hemisphere representations for each modality. To integrate features across hemispheres and modalities, we design a Hemisphere-aware Cross Transformer-based fusion module with four variants of Cross Transformer Encoder and feature fusion attention. The encoder leverages cosine similarity between input vectors as a guidance signal, adaptively modulating the fusion of heterogeneous features:

$$S_i = \frac{\vec{X_i}\cdot\vec{Y_i}}{\|\vec{X_i}\|\|\vec{Y_i}\|} \tag{5}$$

where X and Y denote feature vectors from either the two hemispheres of one modality or from different modalities within a hemisphere. The fusion attention is defined as:

$$\text{Attention}(Q, K, V, S) = \text{softmax}\left(\frac{QK^T + S}{\sqrt{d_k}}\right)V \tag{6}$$

where S represents the similarity map.

The first two Cross Transformers fuse left and right hemisphere features within each modality, capturing intra-modal inter-hemispheric relationships and preserving modality-specific information. The other two fuse modality-specific features within each hemisphere, ensuring cross-modal alignment while maintaining hemispheric distinctions. This hierarchical strategy jointly models intra- and cross-modal interactions, enhancing discriminative feature learning.

IV. Experienments

A. Dataset and preprocessing

The brain imaging are obtained from the ADNI(adni.loni.usc.edu). For imaging data, a total of 801 T1-weighted sMRI scans are obtained from the ADNI1 and ADNI2 baseline, including 213 AD subjects, 234 NC subjects, and 354 MCI (219 sMCI and 135 pMCI) subjects. MCI subjects who developed AD within 3 years were labeled as pMCI and those who didn't convert to AD were labeled as sMCI. We split the data into training, validation and test sets, which are 70%, 10% and 20%, respectively. We ensured that the subject IDs are different prevent data leakage [10].

We conducted pre-processing of brain imaging according to standard procedures on the Clinica platform (clinica.run). Specifically, preprocessed sMRI scans undergo ACPC alignment, affine registration to the MNI152 template, and skull stripping [16], followed by intensity correction. For PET scans, they are co-registered according to the corresponding bias-corrected sMRI scan. All preprocessed images are with a resolution of 105×125×105 voxels.

B. Experimental details

The experiments are conducted based on the PyTorch platform. The network trained using the Adam optimizer. The training objective adopts the cross-entropy loss L_c, which measures the discrepancy between predicted probabilities and subject labels. The initial learning rate was set to 1e-4, and after 30 epochs, it was reduced by one-third every 10 epochs until it reached 1e-6. The batch size was set to 12, a total of 80 epochs for training. All Transformer modules consisted of one layer, with an input vector length d=128 for the multi-head attention mechanism and head h=8. The patch size is set to 25, meaning each hemisphere is divided into two patches along the coronal side. The convolutional kernels were first randomly initialized in the AD vs. NC task, and the trained network parameters were then used to initialize the network for the sMCI vs. pMCI task. Diagnostic performance was evaluated by accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under the curve (AUC).

C. Ablation studies

To validate the effectiveness of the proposed architecture and loss functions, we conducted comprehensive ablation studies. ResNet [17] adopts 3D convolution with 18 layers,

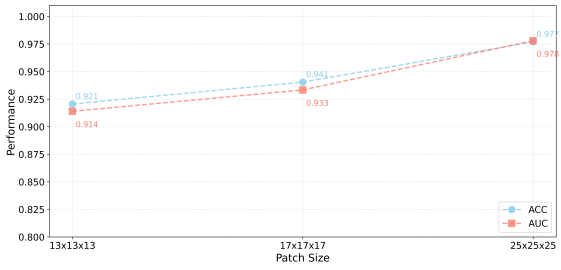

Fig. 2. The impact of different patch sizes on model performance.

TABLE I

AD DIAGNOSTIC PERFORMANCE COMPARISON BASED ON SINGLE MODALITY AND MULTI-MODALITY. FOR THE FEATURE FUSION, T REPRESENTS STANDARD TRANSFORMER ENCODER, F REPRESENTS OUR MODIFIED FUSION WITH OUR MODIFIED CROSS TRANSFORMER.

Method	**Modality**	**Fusion**	AD vs. NC				pMCI vs. sMCI			
			ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
ResNet	MRI	-	88.06	85.10	90.15	0.876	70.75	69.64	72.00	0.708
3DCNN	MRI	-	88.79	83.64	93.08	0.883	74.53	65.18	79.24	0.722
ACNN	MRI	-	88.79	85.31	91.81	0.885	77.35	66.96	82.59	0.748
PCNN	MRI	-	91.04	90.25	91.81	0.910	78.87	51.85	95.45	0.736
LRPCNN	MRI	-	92.22	88.37	95.74	0.920	79.24	76.00	82.14	0.780
ResNet	PET	-	88.92	87.94	91.67	0.885	73.57	58.19	82.07	0.710
3DCNN	PET	-	89.31	86.85	90.52	0.889	77.35	56.25	86.48	0.713
ACNN	PET	-	89.55	85.21	93.16	0.892	79.24	59.14	83.00	0.740
PCNN	PET	-	92.53	91.76	93.16	0.925	80.13	50.00	94.59	0.722
LRPCNN	PET	-	93.28	92.44	95.62	0.929	80.13	62.50	89.18	0.758
MAResNet	PET+MRI	T	91.38	89.06	94.22	0.905	76.61	68.18	81.90	0.750
MACNN	PET+MRI	T	91.65	88.78	94.97	0.908	77.16	65.78	85.30	0.758
PFPCNN	PET+MRI	T	94.86	97.04	92.48	0.947	80.46	64.60	85.48	0.761
TrFPCNN	PET+MRI	F	95.49	94.54	97.14	0.952	80.73	65.78	87.89	0.771
OURS	PET+MRI	F	**95.55**	**97.67**	93.61	**0.956**	**83.01**	68.18	90.84	**0.784**

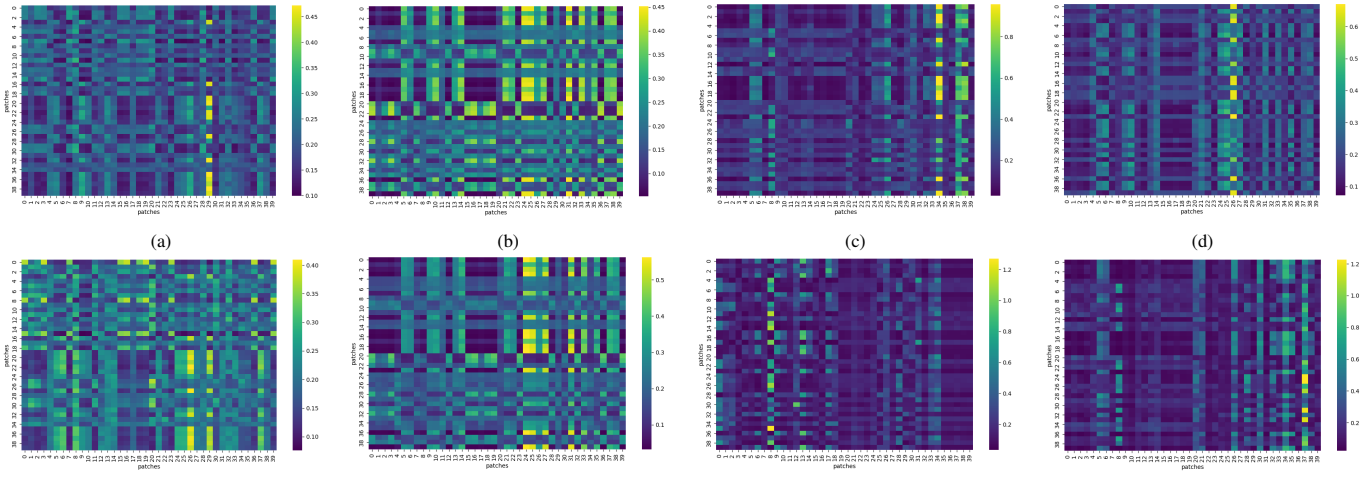

Fig. 3. Subfigures (a)–(d) correspond to the AD vs NC task, and subfigures (e)–(h) correspond to the pMCI vs sMCI task. For each task: (a)/(e) left/right hemisphere correlation for MRI; (b)/(f) left/right hemisphere correlation for PET; (c)/(g) cross-modality correlation for the left hemisphere; (d)/(h) cross-modality correlation for the right hemisphere.

while ACNN incorporates a 3D spatial attention module after the first convolution of 3DCNN. MAResNet and MACNN extend these backbones with integrated attention mechanisms [18], performing whole-brain feature encoding followed by classification. PCNN is a patch-based 3D CNN, and PFPCNN further fuses left–right hemisphere and dual-modal features after PCNN encoding. TrFPCNN introduces a Transformer encoder to directly fuse each modality's hemispheric features before multimodal fusion. In contrast, OURS conducts intra-hemispheric bimodal fusion within each hemisphere, then integrates fused hemispheric representations across hemispheres.

We compared several CNN methods based on single modalities and multimodal as shown in Table I. Patch-based CNN achieved better performance than whole-brain CNN, as it more effectively captures localized structural changes in brain regions, whereas whole-brain CNN tends to obscure fine variations. Moreover, LRPCNN, which learns hemisphere-specific features with non-shared parameters, outperformed PCNN, indicating that despite morphological symmetry, left and right hemispheres still carry distinct structural or functional information beneficial for AD diagnosis. Further, the results show that integrating sMRI and PET with left/right brain patches further improves diagnosis and MCI conversion prediction. By combining multimodal features via Cross Transformer, our framework consistently demonstrated superior performance.

In addition, the patch size in our framework can be flexibly adjusted to accommodate brain imaging data of different sizes generated by various preprocessing methods. Furthermore, the impact of different patch sizes on AD diagnostic performance was explored, with comparative results shown in Fig. 2. The results show that dividing each hemisphere into two patches (patch size=25) along the coronal side produces the best diag-

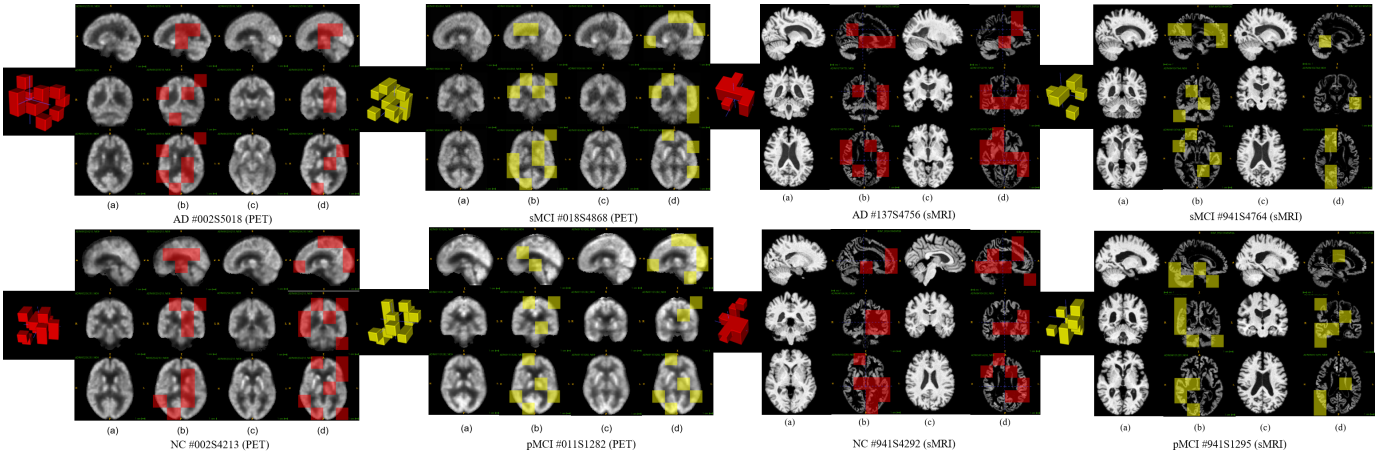

Fig. 4. The attention visualizations of sMRI and PET with AD, NC, sMCI and pMCI subject including. For each subject, images (a) and (c) display the whole brain scans, providing a clearer view of the brain information, while images (b) and (d) highlight the patch regions in brain that the network focuses on. The key regions our model emphasizes include the hippocampus, frontal lobe, and temporal lobe regions. Additionally, there are differences in the areas of focus between the left and right hemispheres.

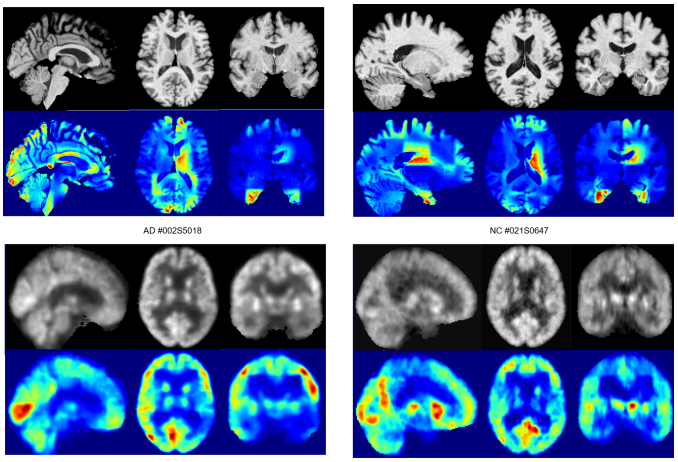

Fig. 5. The top panels shown for the subject of sMRI and PET. The bottom panels are the CAM weights visualization of the heat maps. The highlighted areas in the heatmap correspond to the patch regions of interest, which are associated with the pathological patterns change in sMRI and PET imaging.

nostic performance. The observed effectiveness of this patch size may be related to the underlying anatomical structures, which presents an interesting direction for investigation.

D. Visualization

Interpretability is critical in computer-aided diagnosis. AD-related pathological features are often subtle, scattered across multiple brain regions, and difficult for traditional CAM methods to localize accurately [7]. Moreover, the high dimensionality and structural complexity of brain imaging may cause networks to overemphasize local regions while neglecting other AD-relevant areas [19]. Our method improves interpretability by modeling brain space: patch features from each hemisphere are processed with Transformer encoder to generate attention weights. Fig. 3 shows the correlation matrix of 80 patch regions, where in AD vs. NC, attention concentrates on a few patches, while in the tough MCI task, attention is distributed across broader regions, reflecting increased task complexity.

Fig. 4 highlights the most attended brain regions across categories, including the frontal, temporal, parietal lobes, and hippocampus regions, consistent with known AD biomarkers [20] [21]. Distinct hemispheric focus patterns further validate the method's ability to capture critical features. In MCI, pathological changes are less pronounced than in AD; however, consistent attention to regions such as the hippocampus and frontal lobe indicates their importance [22]. Fig. 5 visualizes the heatmaps in sMRI and PET. These results align with established biomarkers and provide interpretable insights into the model's decision process.

E. Compare with other methods

Many AD studies have been reported to suffer from methodological and data leakage, which can compromise the reliability of their conclusions. Therefore, Table II summarizes several relatively authoritative studies that strictly adhere to multimodal sMRI and PET from the ADNI baseline. Our method achieves superior or competitive performance compared to advanced approaches in multiple diagnostic indices.

V. Conclusion

To improve diagnosis and interpretability, we propose a hemisphere-based hybrid multimodal network that jointly leverages sMRI and PET. By adopting an interleaved fusion strategy, the framework enables parallel learning of local features from both hemispheres across modalities, while capturing long-range dependencies among brain regions and modalities. This design effectively encodes and integrates bi-hemispheric and multimodal information, yielding more comprehensive and discriminative representations for feature learning. In addition, visualization enhances interpretability by highlighting modality- and hemisphere-specific image patterns to support

TABLE II
A COMPARATIVE DESCRIPTION OF SOME CURRENT AD DIAGNOSIS STUDIES USING ADNI BASELINE MULTI-MODAL SMRI AND PET SCANS

Study	AD vs. NC			sMCI vs. pMCI		
	ACC	SEN	SPE	ACC	SEN	SPE
Liu [23]	93.15	93.15	93.57	-	-	-
Miao [24]	94.61	92.92	93.89	-	-	-
Zhang [25]	91.30	-	-	80.30	-	-
Zhang [26]	90.60	95.40	-	75.50	56.0	-
Lin [14]	92.28	90.38	94.37	74.10	73.08	75.0
Huang [13]	90.10	90.85	89.21	72.22	71.25	73.44
Gao [27]	92.0	89.10	94.0	75.30	74.10	77.30
Zhang [28]	92.90	92.30	93.70	81.10	74.20	84.10
OURS	95.55	97.67	93.61	83.01	68.18	90.84

clinical diagnosis. By modeling multimodal interactions at the hemisphere level, the proposed framework learns more discriminative fused representations that better reflect hemispheric asymmetry in AD imaging.

ACKNOWLEDGMENT

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. More details can be found at adni.loni.usc.edu. This work was partly supported by the National Natural Science Foundation of China under Grant 62302149, China Postdoctoral Science Foundation under Grant Number 2025M771578, the National Institutes of Health RF1AG063153.

REFERENCES

[1] M. A. DeTure and D. W. Dickson, "The neuropathological diagnosis of alzheimer's disease," *Molecular neurodegeneration*, vol. 14, no. 1, p. 32, 2019.

[2] L. Robinson, E. Tang, and J. P. Taylor, "Dementia: timely diagnosis and early intervention," *BMJ*, vol. 350, p. h3029, 2015.

[3] J. L. Whitwell, S. A. Przybelski, S. D. Weigand, D. S. Knopman, B. F. Boeve, R. C. Petersen, and C. R. Jack, "3d maps from multiple mri illustrate changing atrophy patterns as subjects progress from mild cognitive impairment to alzheimer's disease," *Brain*, vol. 130, no. 7, pp. 1777–1786, 2007.

[4] A. Myoraku, G. Klein, S. Landau, D. Tosun, and A. D. N. Initiative, "Regional uptakes from early-frame amyloid pet and 18f-fdg pet scans are comparable independent of disease state," *European Journal of Hybrid Imaging*, vol. 6, no. 1, p. 2, 2022.

[5] H. Zhou, L. He, B. Y. Chen, L. Shen, and Y. Zhang, "Multi-modal diagnosis of alzheimer's disease using interpretable graph convolutional networks," *IEEE Transactions on Medical Imaging*, 2024.

[6] R. Sharma, T. Goel, M. Tanveer, C.-T. Lin, and R. Murugan, "Deep-learning-based diagnosis and prognosis of alzheimer's disease: a comprehensive review," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 3, pp. 1123–1138, 2023.

[7] B. Lei, Y. Liang, J. Xie, Y. Wu, E. Liang, Y. Liu, P. Yang, T. Wang, C. Liu, J. Du, X. Xiao, and S. Wang, "Hybrid federated learning with brain-region attention network for multi-center alzheimer's disease detection," *Pattern Recognition*, vol. 153, 2024.

[8] S. Qiu, V. B. Kolachalama, R. Au, E. A. Sartor, J. Yuan, S. H. Auerbach, M.-H. Saint-Hilaire, S. Kedar, A. Swaminathan, Y. J. Alderazi, Y. Zhou, M. Kaku, S. Zhu, B. Dwyer, A. S. Joshi, G. H. Chang, C. Karjadi, X. Zhou, C. Xue, M. I. Miller, and P. S. Joshi, "Development and validation of an interpretable deep learning framework for alzheimer's disease classification," *Brain*, vol. 143, no. 6, pp. 1920–1933, 2020.

[9] C. Lian, M. Liu, J. Zhang, and D. Shen, "Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural mri," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 880–893, 2020.

[10] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, and O. Colliot, "Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation," *Medical Image Analysis*, vol. 63, 2020.

[11] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Landmark-based deep multi-instance learning for brain disease diagnosis," *Medical image analysis*, vol. 43, pp. 157–168, 2018.

[12] K. M. Poloni, R. J. Ferrari, A. D. N. Initiative *et al.*, "A deep ensemble hippocampal cnn model for brain age estimation applied to alzheimer's diagnosis," *Expert Systems with Applications*, vol. 195, p. 116622, 2022.

[13] Y. Huang, J. Xu, Y. Zhou, T. Tong, and X. Zhuang, "Diagnosis of alzheimer's disease via multi-modality 3d convolutional neural network," *Frontiers in Neuroscience*, vol. 13, 2019.

[14] W. Lin, W. Lin, G. Chen, H. Zhang, Q. Gao, Y. Huang, T. Tong, and M. Du, "Bidirectional mapping of brain mri and pet with 3d reversible gan for the diagnosis of alzheimer's disease," *Frontiers in Neuroscience*, vol. 15, 2021.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020, Conference Proceedings.

[16] A. Hoopes, J. S. Mora, A. V. Dalca, B. Fischl, and M. Hoffmann, "Synthstrip: skull-stripping for any brain image," *NeuroImage*, vol. 2022, no. 260, 2022.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, Conference Proceedings, pp. 770–778.

[18] Y. Zhang, Q. Teng, X. He, T. Niu, L. Zhang, Y. Liu, and C. Ren, "Attention-based 3d cnn with multi-layer features for alzheimer's disease diagnosis using brain images," in *2023 45th annual international conference of the ieee engineering in medicine & biology society (embc)*. IEEE, 2023, pp. 1–4.

[19] Z. Xia, G. Yue, Y. Xu, C. Feng, M. Yang, and T. Wang, "A novel end-to-end hybrid network for alzheimer's disease detection using 3d cnn and 3d clstm," in *IEEE 17th International Symposium on Biomedical Imaging*, 2020, Conference Proceedings, pp. 1–4.

[20] E. Hari, E. Kurt, C. Ulasoglu-Yildiz, A. Bayram, B. Bilgic, T. Demiralp, and H. Gurvit, "Morphometric analysis of medial temporal lobe subregions in alzheimer's disease using high-resolution mri," *Brain Structure and Function*, vol. 228, no. 8, pp. 1885–1899, 2023.

[21] K. Ishii, H. Sasaki, A. K. Kono, N. Miyamoto, T. Fukuda, and E. Mori, "Comparison of gray matter and metabolic reduction in mild alzheimer's disease using fdg-pet and voxel-based morphometric mr studies," *European journal of nuclear medicine and molecular imaging*, vol. 32, pp. 959–963, 2005.

[22] S. Kang, S.-W. Kim, J.-K. Seong, A. D. N. Initiative *et al.*, "Disentangling brain atrophy heterogeneity in alzheimer's disease: A deep self-supervised approach with interpretable latent space," *NeuroImage*, vol. 297, p. 120737, 2024.

[23] X. Liu, W. Li, S. Miao, F. Liu, K. Han, and T. T. Bezabih, "Hammf: hierarchical attention-based multi-task and multi-modal fusion model for computer-aided diagnosis of alzheimer's disease," *Computers in Biology and Medicine*, vol. 176, p. 108564, 2024.

[24] S. Miao, Q. Xu, W. Li, C. Yang, B. Sheng, F. Liu, T. T. Bezabih, and X. Yu, "Mmtfn: Multi-modal multi-scale transformer fusion network for alzheimer's disease diagnosis," *International Journal of Imaging Systems and Technology*, vol. 34, no. 1, p. e22970, 2024.

[25] Y. Zhang, K. Sun, Y. Liu, F. Xie, Q. Guo, and D. Shen, "A modality-flexible framework for alzheimer's disease diagnosis following clinical routine," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[26] Z.-C. Zhang, X. Zhao, G. Dong, and X.-M. Zhao, "Improving alzheimer's disease diagnosis with multi-modal pet embedding features by a 3d multi-task mlp-mixer neural network," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 8, pp. 4040–4051, 2023.

[27] X. Gao, F. Shi, D. Shen, and M. Liu, "Task-induced pyramid and attention gan for multimodal brain image imputation and classification in alzheimer's disease," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 1, pp. 36–43, 2022.

[28] Y. Zhang, K. Sun, Y. Liu, and D. Shen, "Transformer-based multimodal fusion for early diagnosis of alzheimer's disease using structural mri and pet," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023, pp. 1–5.