Prior-guided Prototype Aggregation Learning for Alzheimer's Disease Diagnosis

Yueqin Diao[1], Huihui Fang[2](✉), Hanyi Yu[1], Yuning Wang[1], Yaling Tao[1], Ziyan Huang[1], Si Yong Yeo[3,4], and Yanwu Xu[1,5]

[1] School of Future Technology, South China University of Technology, Guangdong, China
[2] College of Computing and Data Science, Nanyang Technological University, Singapore
[3] MedVisAI Lab
[4] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore
[5] Pazhou Lab, Guangdong, China
huihui.fang@ntu.edu.sg

Abstract. Alzheimer's disease (AD) diagnosis faces the challenge of capturing complex patterns of subtle structural and functional changes in neuroimaging and the underutilization of clinical prior knowledge. Current deep learning methods primarily focus on structural magnetic resonance imaging (sMRI) analysis, often overlooking the critical disease concepts that clinicians rely on. To address this limitation, we propose a Prior-guided Prototype Aggregation Learning (PPAL) framework. This framework leverages structured prompts to large language models (LLMs) to extract disease-related anatomical descriptions as clinical prior knowledge and progressively aggregates the visual features of AD and cognitively normal (CN) individuals, bridging the semantic gap between sMRI features and LLM-derived clinical concepts to construct category prototype representations. Meanwhile, we design a slice selection and compression module that adaptively learns the importance of different slices, prioritizing those most critical for AD diagnosis. Ultimately, AD diagnosis is achieved by computing the semantic similarity between MRI slice features and the category prototypes. Experimental results demonstrate that, compared to state-of-the-art 2D slice-based methods, incorporating clinical prior knowledge not only enhances the identification of pathological regions but also shows significant advantages in the zero-shot mild cognitive impairment (MCI) conversion task. The code is available at: https://github.com/diaoyq121/PPAL.

Keywords: Alzheimer's disease · Aggregation attention · Semantic similarity.

1 Introduction

Alzheimer's disease (AD) is a prevalent neurodegenerative disorder marked by memory loss and cognitive decline, ultimately leading to significant impairment

in daily activities [18]. As an irreversible condition with no definitive cure, early diagnosis is essential for timely intervention and effective disease management [16,25]. In recent years, deep learning methods have significantly advanced AD diagnosis by extracting subtle brain morphology and anatomical features from structural magnetic resonance imaging (sMRI) scans [2,3]. Common analytical approaches include voxel-level, patch-level, and slice-level analysis. Voxel-level methods employ 3D convolutional neural networks (CNNs) to extract whole-brain features, incorporating attention mechanisms [7,5] and multi-scale analysis [20] to improve detection of pathological regions. Patch-level methods divide sMRI scans into smaller patches for localized feature learning [24], often combining multi-instance learning with attention mechanisms [17] to enhance performance. However, due to their reliance on 3D CNNs, both voxel- and patch-level methods are computationally intensive and tend to overemphasize local features. In contrast, slice-level methods decompose 3D MR images into 2D slices, enabling more efficient transfer learning and improved interpretability [15,10].

However, existing methods have not fully leveraged the external knowledge from textual priors. This knowledge enriches semantic understanding, aids in deep feature clustering [8], and improves diagnostic accuracy. In natural scenes, CLIP [12] performs text-image alignment using a dual-tower architecture with contrastive learning, excelling in tasks such as classification [6,21], segmentation [9,13], and anomaly detection [19]. Inspired by this, vision-language models (VLMs) like ViLa-MIL [14], BiomedCLIP [22], and MI-Zero [11] have achieved excellent results in various medical diagnosis tasks. Additionally, Fang et al. [4] addressed the scarcity of image-text pairs by combining large language models (LLMs) with medical images through refined prompts. However, the use of image-text semantic similarity in AD diagnosis is still rare, and the domain gap between natural and medical images limits the effectiveness of CLIP-based methods. To address these challenges, we propose a Prior-guided Prototype Aggregation Learning (PPAL) framework, which progressively aggregates image category representations using textual priors. This approach refines text embeddings while bridging the gap between image and text feature representations, providing a more interpretable and reliable framework for AD diagnosis.

The main contributions of our work are as follows: 1) We propose a guidance mechanism based on external textual concept knowledge to progressively aggregate visual features, while dynamically optimizing text prompt expressions by incorporating contextual information, thereby enhancing the semantic alignment between visual and textual modalities; 2) We employ a Slice Selection and Compression (SSC) module to model the importance of slices, enabling the model to focus on critical slices and achieve dynamic fusion; 3) We designed a disease knowledge-guided network framework for AD diagnosis, which enhances both model performance and interpretability by incorporating external knowledge.

2 Method

2.1 Overall

Our core goal is to integrate medical prior knowledge with image features for AD diagnosis via semantic similarity computation. As shown in Figure 1, we first use a LLM to extract disease-related concepts through a QA approach. Then, the **Slice Selection and Compression (SSC)** network assigns importance scores to MRI slices, selecting and aggregating key slices. The **Text-Aggregated Visual (TAV)** representation network iteratively fuse visual features, reducing semantic gaps and enriching textual descriptions. Finally, AD diagnosis is achieved by computing the semantic similarity between image features and the aggregated text prototype embeddings.

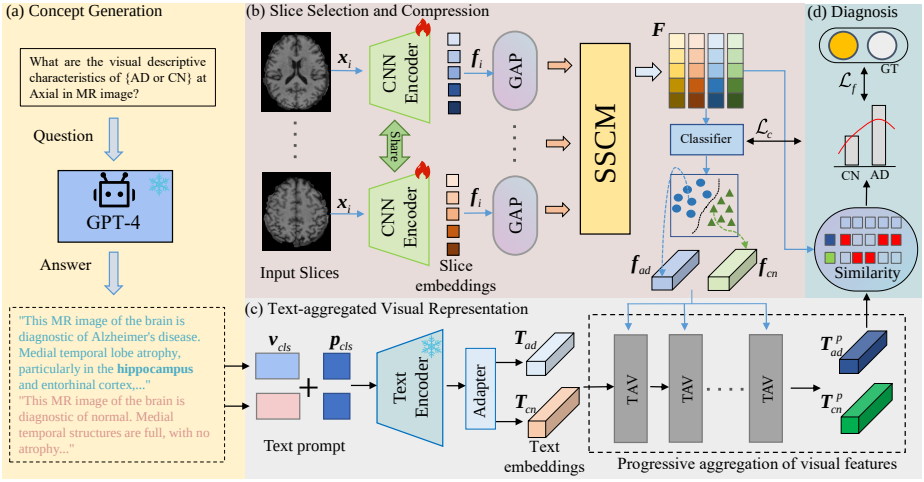

Fig. 1. Pipeline of the proposed PPAL framework. (a) Extract disease-related concepts through a question-and-answer mechanism. (b) Select and integrate key slice information to enhance critical representations. (c) Progressively aggregate visual features using text embeddings to obtain more refined prototype feature representations. (d) Classify based on semantic similarity measurements.

2.2 Slice selection and compression

In Alzheimer's disease diagnosis, clinicians focus on key slices with diagnostic value. Building on this, our study proposes a critical slice selection method that integrates the most relevant slices related to pathological features. We use a shared convolutional encoder based on the pre-trained VGG16 architecture. To ensure compatibility, the original slices are replicated along the channel dimension and resized to 224×224 pixels. The encoder extracts features, followed by

global average pooling to generate a feature embedding for each slice:

$$\boldsymbol{f}_i = \mathrm{GAP}\left(E_{\mathrm{share}}\left(\boldsymbol{x}_i; \theta_{\mathrm{cnn}}\right)\right) \tag{1}$$

Where $\boldsymbol{x}_i \in \mathbb{R}^{C\times H\times W}$ and $\boldsymbol{f}_i$ are the i-th slice and the corresponding feature embedding, respectively, and θ_{cnn} represents the parameters of the shared encoder E_{share}. As shown in Figure 2, to select the important slices, the slice embeddings are concatenated to obtain a feature representation $\boldsymbol{f}_c \in \mathbb{R}^{N\times d}$, and then a multi-head self-attention mechanism is applied to learn the correlations between different slices:

$$\tilde{\boldsymbol{f}}_c = MultiHead\left(\boldsymbol{Q}_c, \boldsymbol{K}_c, \boldsymbol{V}_c\right) \tag{2}$$

where $\boldsymbol{Q}_c = \boldsymbol{f_c} \times {\boldsymbol{W}_q}^c$, $\boldsymbol{K}_c = \boldsymbol{V}_c = \boldsymbol{f}_c \times \boldsymbol{W}_{kv}^c$ and ${\boldsymbol{W}_q}^c \in \mathbb{R}^{d\times d}$, $\boldsymbol{W}_{kv}^c \in \mathbb{R}^{d\times d}$ denote the weight matrix for the linear transformation. Subsequently, the output of the attention layer is added to the original input, and layer normalization is applied to obtain a new feature representation $\boldsymbol{f}'_c$. Following this, slice attention (SA) is applied to dynamically adjust and integrate information from multiple slices, considering both global context and slice relevance. This mechanism ensures that the extracted features are selectively focused on AD-related content. The process can be formally described as follows:

$$\alpha_i = \frac{\exp\left\{\boldsymbol{W}_b^T\left(\tanh\left(\boldsymbol{W}_v \boldsymbol{f}'_{ci}\right) \odot sign\left(\boldsymbol{W}_u \boldsymbol{f}'_{ci}\right)\right)\right\}}{\sum_{j=1}^{N} \exp\left\{\boldsymbol{W}_b^T\left(\tanh\left(\boldsymbol{W}_v \boldsymbol{f}'_{cj}\right) \odot sign\left(\boldsymbol{W}_u \boldsymbol{f}'_{cj}\right)\right)\right\}} \tag{3}$$

$$\beta_i = \frac{\exp\left\{\left(mean\left(\boldsymbol{f}'_{ci}\right)\right)\right\}}{\sum_{j=1}^{N} \exp\left\{\left(mean\left(\boldsymbol{f}'_{ci}\right)\right)\right\}} \tag{4}$$

$$\boldsymbol{F} = \sum_{i=1}^{N} s_i \boldsymbol{f}'_{ci}, s_i = \alpha_i + \beta_i \tag{5}$$

Where $\boldsymbol{W}_u \in \mathbb{R}^{d\times d}$, $\boldsymbol{W}_v \in \mathbb{R}^{d\times d}$, and $\boldsymbol{W}_b \in \mathbb{R}^{d\times 1}$ are trainable weight matrices. $\tanh(\cdot)$ and $\mathrm{sign}(\cdot)$ are activation functions, with s_i representing the score of the i-th slice. We dynamically integrate the information from each slice using the scores s_i to obtain $\boldsymbol{F}$. Subsequently, a classifier is applied to map the integrated features into the classification space for coarse classification, yielding the probabilities for each category:

$$\hat{y}^c = classfier\left(\boldsymbol{F}\right), d_i = softmax\left(\hat{y}^c\right) \tag{6}$$

Based on the class probabilities d_i, we divide the features into two categories: $\boldsymbol{f}_{ad} = d_0 \odot \boldsymbol{F}$ and $\boldsymbol{f}_{cn} = d_1 \odot \boldsymbol{F}$, which are used to enhance the text features for each category. We employ a binary cross-entropy loss function for regularization:

$$\mathcal{L}_c = \mathcal{L}_{bce}\left(\hat{y}^c, y\right) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log\left(\hat{y}_i^c\right) + \left(1 - y_i\right) \cdot \log\left(1 - \hat{y}_i^c\right)\right] \tag{7}$$

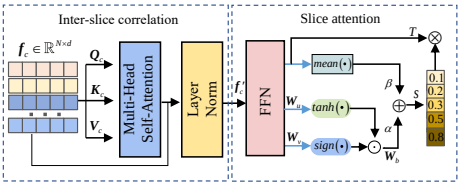

Fig. 2. Slice selection and compression module.

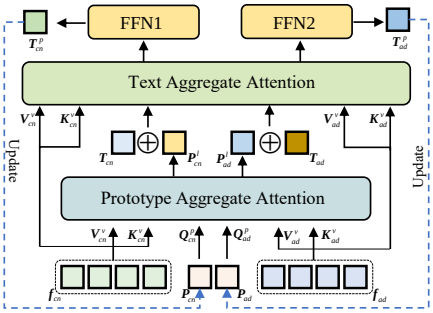

Fig. 3. Text-aggregated visual module.

2.3 Text-aggregated visual representation

Although image-text alignment has been achieved through contrastive learning in natural scenes, directly applying the CLIP model to medical scenarios still faces the challenge of semantic misalignment between image and text, which affects diagnostic performance. To address this challenge, we leverage text-based prior knowledge to progressively aggregate visual features, thus obtaining prototype embeddings for AD and CN and alleviating the semantic shift problem. To extract AD-related knowledge concepts, we refer to prior research and adopt the GPT-4 model, using a question-and-answer approach to generate feature descriptions related to AD and CN. To better transfer CLIP's knowledge to the AD diagnosis domain, we introduce learnable vectors as supplementary prompts. These learnable prompts, $\boldsymbol{p}_{cls}$, are concatenated with the prompt text, $\boldsymbol{v}_{cls}$, forming a new input [23], which is then fed into the CLIP text encoder to generate text feature embeddings for the two categories. Next, an adapter is used to map these embeddings to the visual semantic space, enabling the conversion from text space to visual space. Finally, we introduce a progressive text aggregation method to better learn and enhance the text semantic information.

Specifically, we introduce a progressive text aggregation method to better capture and enhance the semantic information from the text. As shown in Figure 3, we set up two learnable prototype features $\boldsymbol{P}_{cls}$, which use prototype aggregation attention to separately learn the information from the AD and CN category features. These prototype features are then fused with the text embeddings, enriching the semantic information and mitigating the significant semantic gap between the two types of features:

$$\boldsymbol{P}^{l}_{cls} = Norm\left(\text{softmax}\left(\frac{\boldsymbol{Q}^{p}_{cls}\boldsymbol{K}^{v}_{cls}}{\sqrt{d}}\right)\boldsymbol{V}^{v}_{cls}\right) + \boldsymbol{P}_{cls}, (cls = ad, cn) \tag{8}$$

$$\tilde{\boldsymbol{T}}^{l}_{cls} = \boldsymbol{T}_{cls} + \boldsymbol{P}^{l}_{cls} \tag{9}$$

Subsequently, the text aggregation attention mechanism is used to aggregate the discriminative information from the AD and CN classes, and a feed-forward

network (FFN) layer is employed to update the prototypes, generating two text prototype features with stronger discriminative ability:

$$\boldsymbol{T}_{cls}^{l} = Norm\left(\text{softmax}\left(\frac{\boldsymbol{Q}_{cls}^{t}\boldsymbol{K}_{cls}^{v}}{\sqrt{d}}\right)\boldsymbol{V}_{cls}^{v}\right) + \tilde{\boldsymbol{T}}_{cls}^{l}, \boldsymbol{T}_{cls}^{p} = FFN\left(\boldsymbol{T}_{cls}^{l}\right) \quad (10)$$

We perform three iterative operations to progressively aggregate the visual representations, resulting in a discriminative prototype feature embedding $\boldsymbol{T}_{cls}^{l}$. Subsequently, the semantic similarity between the aggregated slice visual embedding $\boldsymbol{F}$ and $\boldsymbol{T}_{cls}^{l}$ is computed to obtain the probabilities for different categories:

$$\hat{y}_{cls}^{f} = \frac{\exp\left(\cos\left(\boldsymbol{F}, \boldsymbol{T}_{cls}^{p}\right)\right)}{\sum_{i=1}^{C}\exp\left(\cos\left(\boldsymbol{F}, \boldsymbol{T}_{i}^{p}\right)\right)} \quad (11)$$

Subsequently, $\mathcal{L}_f = \mathcal{L}_{bce}\left(\hat{y}_{cls}^{f}, y\right)$ is used as a constraint for fine-grained classification. The final result is obtained by combining the coarse classification results with the fine-grained matching results. The total loss used for model training is:

$$\mathcal{L}_{oveall} = \mathcal{L}_f + \mathcal{L}_c \quad (12)$$

3 Experiments and Results

Datasets and Evaluation metrics To assess the performance of our proposed algorithm, we used the "ADNI1: Complete 1Yr 1.5T" dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) benchmark. This dataset consists of 639 participants, each of whom underwent baseline screening and follow-up MRI scans at 6 and 12 months, acquired with a 1.5 Tesla MRI scanner. Participants with MCI were further categorized into progressive MCI (pMCI) or stable MCI (sMCI) groups, depending on whether they converted to Alzheimer's disease within 36 months. The dataset includes 586 cognitively normal (CN), 474 AD, 162 pMCI, and 154 sMCI cases. The MRI images were registered to the MNI152 template space using the SyN algorithm in ANTs, followed by standard preprocessing steps: (1) bias field correction with the N4ITK method, (2) affine registration, and (3) skull stripping. We evaluated the algorithm on two binary classification tasks - AD vs CN and sMCI vs pMCI — using five-fold cross-validation. Performance was measured using four metrics: accuracy (ACC), specificity (SPE), sensitivity (SEN), and Matthews correlation coefficient (MCC). Notably, MCC is considered a more balanced and informative metric for binary classification, especially in the presence of class imbalance.

Implementation Details Our method was implemented using PyTorch and executed on an Intel(R) Xeon(R) CPU and NVIDIA GeForce RTX 4090 GPU. For a fair comparison, all methods in our experiments, including the proposed PPAL, were built on the VGG16 backbone. To enhance model generalization and robustness, we applied a series of random transformations to the MRI images during training. The model was trained using the Adam optimizer with an initial learning rate of 1e-4 and a decay rate of 0.0001. We set the batch size to 4 and trained for a maximum of 100 epochs, employing an early stopping strategy to prevent overfitting.

3.1 Comparison to state-of-the-art methods

Table 1. Results of 5-fold cross-validation on the ADNI dataset for AD diagnosis and MCI conversion prediction tasks. The gray area represents the results under zero-shot classification.

Networks	AD vs CN				sMCI vs pMCI			
	ACC	SPE	SEN	MCC	ACC	SPE	SEN	MCC
Attention Transformer [1]	82.60	**91.40**	71.70	65.10	62.30	66.50	48.73	23.80
AwareNet [15]	83.32	87.5	77.80	65.90	48.41	**77.40**	25.80	3.90
Majority Voting	80.40	89.70	68.90	60.50	61.40	60.10	62.90	22.90
CLIP [12]	78.44	75.31	80.98	56.36	63.72	58.90	68.83	27.83
ViLa-MIL Low [6]	79.19	66.53	**89.94**	58.16	63.09	56.44	70.13	26.79
CoOP [23]	82.86	79.29	84.72	64.13	62.46	60.74	64.29	25.02
AXIAL [10]	83.22	75.73	89.3	66.06	64.67	50.30	**79.87**	31.49
Ours	**85.38**	82.42	87.78	**70.39**	**67.19**	65.64	68.83	**34.47**

AD vs CN We compared our model with current state-of-the-art models based on 2D slice attention, including Attention Transformer [1], AwareNet [15], Majority Voting, AXIAL [10], and contrastive learning-based methods such as CLIP [12], ViLa-MIL Low [6] and CoOP [23]. As shown in Table 1, our method achieved the best results in the AD vs CN task, with ACC and MCC reaching 85.38% and 70.39%, respectively, outperforming AXIAL by 2.16% and 4.33%, demonstrating the effectiveness of our method in AD diagnosis. Additionally, with SPE at 82.42% and SEN at 87.78%, our method demonstrates strong diagnostic capability for both positive and negative samples.

sMCI vs pMCI In the MCI conversion task, diagnosing classification is more challenging due to the similar MRI characteristics between sMCI and pMCI. Specifically, to assess the model's zero-shot capability, we directly applied the weights trained on the AD vs. CN classification task to the MCI conversion prediction task without any additional fine-tuning. The results, presented in Table 1, show that the ACC and MCC reached 67.19% and 34.47%, respectively, outperforming both training-based and untrained methods. These findings highlight the potential of our approach in zero-shot transfer learning.

Interpretability analysis To enhance the interpretability of the model, we utilized explainable artificial intelligence (XAI) methods to visualize slice weight scores and key regions of interest, as shown in Figure 4. The visualization results indicate that our method can focus on critical slice information and primarily attend to key regions such as the hippocampus, parahippocampal gyrus, and amygdala. Table 2 presents the top five regions the model focuses on, which are consistent with the visualization results and clinical observations. Additionally, these regions align with the keywords specified in the text prompts, suggesting that the text prompts effectively guide the model to focus on regions relevant to Alzheimer's disease diagnosis. The attention map in Figure 5 further illustrates the model's ability to attend to important AD-related regions.

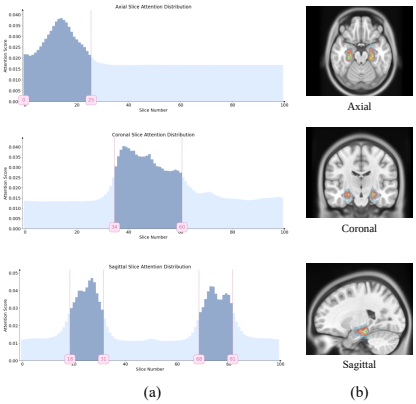

Fig. 4. (a) Slice attention weight distribution across three planes. (b) Visualization of model attention on ROI .

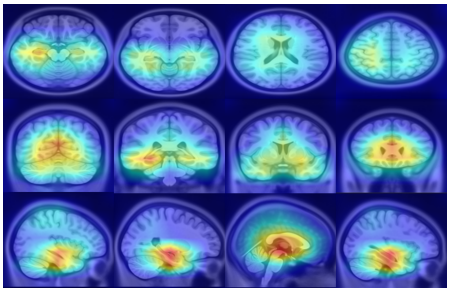

Fig. 5. GradCAM++ visualizations of four randomly selected slices per plane for our model.

Table 2. Top 5 ROI. V_r and P_r represent the overlap volume and overlap percentage of a specific region, respectively.

Brain Area	V_r	P_r
Hippocampus - right	1571	0.3348
Hippocampus - left	1536	0.3272
Parahippocampal - left	730	0.2696
Parahippocampal - right	532	0.1965
Amygdala - left	332	0.2012

Table 3. The ablation experiments conducted on TAV and SSC.

Methods	AD vs CN			
	ACC	SPE	SEN	MCC
Baseline	82.84	**84.52**	81.49	65.70
+TAV	83.60	78.87	87.43	66.75
+SSC	84.44	79.70	88.29	68.47
+SSC(w/o SA)	84.16	77.62	**89.47**	67.94
PPAL(w/o $\mathcal{L}_f$)	82.01	76.57	86.42	63.50
Ours	**85.38**	82.42	87.78	**70.39**

3.2 Ablation Study

We conducted an ablation study to evaluate the contribution of each component in our model, as shown in Table 3. Using the original VGG16 model as the baseline, our method achieved significant improvements of 2.54% and 4.69% in ACC and MCC, respectively. We progressively incorporated the SSC and TAV modules, where adding either module individually improved performance, and the best results were obtained when both modules were combined, further validating their effectiveness. Notably, in the absence of slice attention, the ACC and MCC of "Baseline+SSC" declined, highlighting the critical role of slice attention in the SSC module. Additionally, using $\mathcal{L}_f$ for coarse classification aids in aggregating visual features through text embeddings.

4 Conclusion

This paper proposes a method for Alzheimer's diagnosis that aggregates visual features using external knowledge, referred to as PPAL. This method effectively

transfers textual prior knowledge into medical image analysis. By utilizing a LLM to acquire diagnostic knowledge, it employs a progressive aggregation strategy to gradually integrate visual information related to the prior knowledge, achieving more refined knowledge embedding. To more efficiently extract diagnostic information, we adopt a slice selection and compression method that focuses on the most critical slice information and fuses them to obtain effective diagnostic insights. Experimental results demonstrate that our method can effectively select important slices, focus on key region information, and significantly improve the interpretability of the model.

Acknowledgments. This work was supported in part by the Guangzhou Science and Technology Department (Grant No. 2024D03J0013) and the Taihu Lake Innovation Fund for the School of Future Technology of South China University of Technology(Grant No. 2024B105611005).

Disclosure of Interests. The authors declare no competing interests.

References

1. Altay, F., et al.: Preclinical stage alzheimer's disease detection using magnetic resonance image scans. In: Proc. AAAI Conf. Artif. Intell. vol. 35, pp. 15088–15097 (2021)
2. Alzheimer's Association: 2019 alzheimer's disease facts and figures. Alzheimer's Dementia **15**(3), 321–387 (2019)
3. Brookmeyer, R., Johnson, E., Ziegler-Graham, K., Arrighi, H.M.: Forecasting the global burden of alzheimer's disease. Alzheimer's Dementia **3**(3), 186–191 (2007)
4. Fang, X., et al.: Aligning medical images with general knowledge from large language models. In: Int. Conf. Med. Image Comput. Comput.-Assist. Interv. pp. 57–67. Springer (2024)
5. Feng, X., et al.: A deep learning mri approach outperforms other biomarkers of prodromal alzheimer's disease. Alzheimer's Res. & Therapy **14**, 45 (2022)
6. Guo, Z., et al.: Texts as images in prompt tuning for multi-label image recognition. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 2808–2817 (2023)
7. Jin, D., et al.: Attention-based 3d convolutional network for alzheimer's disease diagnosis and biomarkers exploration. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI). pp. 1047–1051. IEEE (2019)
8. Li, Y., et al.: Image clustering with external guidance. In: Proc. Int. Conf. Mach. Learn. pp. 27890–27902 (2024)
9. Lin, Y., et al.: CLIP is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 15305–15314 (2023)
10. Lozupone, G., et al.: AXIAL: Attention-based explainability for interpretable alzheimer's localized diagnosis using 2d cnns on 3d mri brain scans. arXiv preprint arXiv:2407.02418 (2024)
11. Lu, M.Y., et al.: Visual language pretrained multiple instance zero-shot transfer for histopathology images. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 19764–19775 (2023)

12. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Proc. Int. Conf. Mach. Learn. pp. 8748–8763 (2021)
13. Rao, Y., et al.: DenseCLIP: Language-guided dense prediction with context-aware prompting. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 18082–18091 (2022)
14. Shi, J., et al.: ViLa-MIL: Dual-scale vision-language multiple instance learning for whole slide image classification. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 11248–11258 (2024)
15. Wang, C., et al.: Joint learning framework of cross-modal synthesis and diagnosis for alzheimer's disease by mining underlying shared modality information. Med. Image Anal. **91**, 103032 (2024)
16. Wang, R., et al.: Integrating human brain proteomes with genome-wide association data implicates new proteins in alzheimer's disease pathogenesis. Nat. Genet. **53**(2), 143–146 (2021)
17. Wang, T., Dai, Q.: A patch distribution-based active learning method for multiple instance alzheimer's disease diagnosis. Pattern Recognit. **150**, 110341 (2024)
18. Winblad, B., et al.: Defeating alzheimer's disease and other dementias: a priority for european science and society. Lancet Neurol. **15**(5), 455–532 (2016)
19. Wu, P., et al.: VADCLIP: Adapting vision-language models for weakly supervised video anomaly detection. In: Proc. AAAI Conf. Artif. Intell. vol. 38, pp. 6074–6082 (2024)
20. Wu, Y., et al.: An attention-based 3d cnn with multi-scale integration block for alzheimer's disease classification. IEEE J. Biomed. Health Inform. **26**, 5665–5673 (2022)
21. Yu, T., et al.: Task residual for tuning vision-language models. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 10899–10909 (2023)
22. Zhang, S., et al.: Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915 (2023)
23. Zhou, K., et al.: Conditional prompt learning for vision-language models. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 16816–16825 (2022)
24. Zhu, W., Sun, L., Huang, J., et al.: Dual attention multi-instance deep learning for alzheimer's disease diagnosis with structural mri. IEEE Trans. Med. Imaging **40**(9), 2354–2366 (2021)
25. Zhu, Y., Ma, J., Yuan, C., Zhu, X.: Interpretable learning based dynamic graph convolutional networks for alzheimer's disease analysis. Inf. Fusion **77**, 53–61 (2022)