Computerized Medical Imaging and Graphics 126 (2025) 102664

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

CRAD: Cognitive Aware Feature Refinement with Missing Modalities for Early Alzheimer's Progression Prediction

Fei Liu [a], Shiuan-Ni Liang [a,b], Mohamed Hisham Jaward [c], Huey Fang Ong [d], Huabin Wang [e,*], for the Alzheimer's Disease Neuroimaging Initiative[1], the Australian Imaging Biomarkers and Lifestyle flagship study of ageing[2]

[a] Department of Electrical and Robotics Engineering, School of Engineering, Monash University Malaysia, Bandar Sunway, Malaysia
[b] Medical Engineering and Technology Hub, School of Engineering, Monash University Malaysia, Bandar Sunway, Malaysia
[c] School of Business, Arts, Social Sciences and Technology, University of Suffolk, Ipswich, United Kingdom
[d] School of Information Technology, Monash University Malaysia, Bandar Sunway, Malaysia
[e] School of Computer Science and Technology, Anhui University, Hefei, China

ARTICLE INFO

Keywords:
Alzheimer's disease diagnosis
Multimodal knowledge distillation
Confidence regularization
MRI

ABSTRACT

Accurate diagnosis and early prediction of Alzheimer's disease (AD) often require multiple neuroimageing modalities, but in many cases, only one or two modalities are available. This missing modality hinders the accuracy of diagnosis and is a critical challenge in clinical practice. Multimodal knowledge distillation (KD) offers a promising solution by aligning complete knowledge from multimodal data with that of partial modalities. However, current methods focus on aligning high-level features, which limit their effectiveness due to insufficient transfer of reliable knowledge. In this work, we propose a novel Consistency Refinement-driven Multi-level Self-Attention Distillation framework (CRAD) for Early Alzheimer's Progression Prediction, which enables the cross-modal transfer of more robust shallow knowledge with self-attention to refine features. We develop a multi-level distillation module to progressively distill cross-modal discriminating knowledge, enabling lightweight yet reliable knowledge transfer. Moreover, we design a novel self-attention distillation module (PF-CMAD) to transfer disease-relevant intermediate knowledge, which leverages feature self-similarity to capture cross-modal correlations without introducing trainable parameters, enabling interpretable and efficient distillation. We incorporate a consistency-evaluation-driven confidence regularization strategy within the distillation process. This strategy dynamically refines knowledge using adaptive distillation controllers that assess teacher confidence. Comprehensive experiments demonstrate that our method achieves superior accuracy and robust cross-dataset generalization performance using only MRI for AD diagnosis and early progression prediction. The code is available at https://github.com/LiuFei-AHU/CRAD.

1. Introduction

Alzheimer's disease (AD), a progressive neurodegenerative disorder, is clinically characterized by a gradual decline in cognitive functions that ultimately leads to complete dementia. Mild Cognitive Impairment (MCI) represents an intermediate stage between normal cognitive aging and AD, with approximately 50% of individuals diagnosed with MCI progressing to AD within five years (Petersen et al., 2018). Early and accurate diagnosis and prediction of AD are crucial in delaying the progression of dementia (Yiannopoulou and Papageorgiou, 2020; Bouts et al., 2019). Multimodal neuroimaging, such as magnetic resonance imaging (MRI) and positron emission tomography (PET), provides complementary insights for the early diagnosis of AD (Dubois et al., 2023; Rudroff et al., 2024). However, incomplete or missing modalities in clinical settings remain a significant barrier to reliable

* Corresponding author.
E-mail address: wanghuabin@ahu.edu.cn (H. Wang).

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

[2] Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organization (CSIRO) which was made available at the ADNI database (www.loni.usc.edu/ADNI). The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at www.aibl.csiro.au.

https://doi.org/10.1016/j.compmedimag.2025.102664
Received 23 April 2025; Received in revised form 5 November 2025; Accepted 6 November 2025
Available online 10 November 2025

diagnosis. Although multimodal knowledge distillation (KD) offers a promising solution by transferring knowledge from teacher models (trained on complete modalities) to student models (handling partial modalities) (Guan et al., 2021; Yang et al., 2023; Chen et al., 2023; Van Sonsbeek et al., 2021; Wang et al., 2023b; Song et al., 2023), existing methods suffer from two critical limitations: (1) insufficient utilization of discriminative intermediate features that encode early Alzheimer's disease (AD) biomarkers (e.g., hippocampus atrophy, changes in gray matter density) (Weber et al., 2021; Deng et al., 2024), and (2) vulnerability to imperfect supervision (the teacher model may not consistently provide reliable knowledge). Current knowledge distillation (KD) frameworks focus primarily on high-level features or soft labels (Yang et al., 2023; Van Sonsbeek et al., 2021; Wang et al., 2023b; Song et al., 2023), often neglecting the richness of intermediate features that capture localized structural abnormalities critical for early detection of Alzheimer's disease (AD) (Zhai et al., 2024; Hu et al., 2023; Deng et al., 2024). Recent studies (Zhai et al., 2024; Hu et al., 2023) have highlighted the importance of transferring robust intermediate features. These features provide richer information about the disease, enabling multi-granularity analysis and offering better robustness against noise and overfitting. In addition, they are more effective than high-level features in diagnosing early AD, as they have high spatial resolution and local sensitivity, which allows accurate detection of subtle structural changes (Deng et al., 2024). Specifically, early pathological markers of AD, such as hippocampus atrophy, reduced gray matter density, and localized brain abnormalities, are typically identified through intermediate features (Weber et al., 2021). Therefore, highlighting intermediate features can enhance the transfer of discriminating knowledge and improve the generalization performance of student models.

On the other hand, effectively screening task-relevant intermediate features remains challenging for traditional feature distillation. Recently, attention distillation methods (Wang et al., 2019, 2020a) have shown a promising ability to filter knowledge by learning the attentive semantic context of the teacher model. However, existing methods have not fully studied the attention transfer of hierarchical discriminating knowledge. In particular, attention distillation of hierarchical intermediate features can capture discriminating contextual information at different levels that can achieve multi-granularity knowledge transfer.

Meanwhile, to achieve more accurate knowledge transfer, knowledge refinement is critical for KD. Conventional knowledge refinement strategies, such as gated regularization (Yang et al., 2023, 2022b), rely on teacher certainty while ignoring scenarios where student models may outperform teachers. In this regard, we suggest introducing a dynamic refinement mechanism, a more flexible regularization considering the relative gap between the teacher and student models, balancing teacher confidence and student-teacher consistency. Moreover, a flexible paradigm that effectively combines consistency regularization with the distillation of hierarchical knowledge attention is crucial to robust knowledge transfer.

To address these issues, we propose a **C**onsistency **R**efinement-driven Multi-level Self-**A**ttention **D**istillation framework (CRAD), which introduces hierarchical multi-level distillation with dynamic knowledge refinement via consistency evaluation. First, we enhance the teacher model in two ways: by improving the disease awareness capability of its intermediate features through a cognitive awareness feature refinement module (CAFR), improving the quality of knowledge transfer, and further reducing redundancy by disentangling modality-specific knowledge via orthogonal disentanglement (Chen et al., 2023; Yang et al., 2022a). Then, unlike existing works focused on high-level feature alignment (Yang et al., 2023; Wang et al., 2019), we design a special cross-modal self-attention distillation module (PF-CMAD) to transfer disease-relevant intermediate knowledge. This module leverages feature similarity-based self-attention to capture cross-modal correlations without introducing trainable parameters (parameter-free), enabling interpretable and efficient distillation. In addition, by hierarchically aligning multi-level features from intermediate to high-level, we preserve spatial resolution and disease sensitivity crucial for detecting subtle early AD biomarkers (e.g., localized atrophy) (Weber et al., 2021; Deng et al., 2024) from local to global perspectives. Next, we develop a smooth distillation unit (SDU) within multi-level distillation, incorporating a consistency-aware confidence regularization strategy that dynamically refines knowledge by evaluating teacher certainty and student-teacher prediction divergence, ensuring accurate knowledge transfer. Compared to the existing knowledge refinement method (Yang et al., 2023) that only takes teacher models' certainty, we penalize inconsistent teacher–student outputs while promoting reliable cross-modal correlations.

This study aims to bridge a critical gap in early Alzheimer's diagnosis: accurately predicting disease progression when key neuroimaging modalities (e.g., PET) are missing. We propose CRAD, a framework that distills essential diagnostic knowledge from complete multimodal data into models using only partial inputs (e.g., MRI), while maintaining reliability and computational efficiency. Comprehensive experiments demonstrate that our method achieves state-of-the-art diagnostic accuracy and generalization on benchmark datasets. The main contributions are summarized as follows.

- We propose a novel hierarchical distillation framework that utilizes multilevel attention alignment and noise suppression to enable effective cross-modal knowledge transfer for accurate diagnosis and progression prediction of AD.
- A lightweight parameter-free attention distillation module is developed for efficient, robust attentive feature alignment. We utilize the self-similarity of features and incorporate a consistency-aware confidence regularization to minimize unreliable knowledge transfer by evaluating the confidence level and prediction consistency.
- Extensive experiments demonstrate that our method outperforms the state-of-the-art methods in the early diagnosis of AD, even using mono-modality, and visualization analysis reveals its potential to localize disease-specific biomarkers.

This paper is organized as follows. We introduce and review the related work in Section 2. The proposed method is then presented in detail in Section 3. The experimental results are presented and discussed in Section 4. We summarize this work in Section 5.

2. Related work

2.1. Disease diagnosis with multimodal data

Multimodal models have recently received substantial attention for their ability to harness complementary information from diverse data sources, significantly improving the performance of Alzheimer's Disease (AD) diagnosis (Shi et al., 2018, 2022; Qiu et al., 2022; Zhang et al., 2021; Qiu et al., 2024). For example, Shi et al. (2018) proposed a multimodal fragmented deep polynomial network (MM-SDPN) to integrate features extracted from neuroimaging data for the diagnosis of AD. Similarly, Shi et al. (2022) and Zhang et al. (2021) developed advanced feature selection strategies to identify and fuse the most discriminating information from multimodal data, allowing for more effective utilization of complementary features. Despite that, they are often plagued by computational complexity and susceptibility to interference from redundant information. Recent studies (Chen et al., 2023; Lu et al., 2020; Yang et al., 2022a; Wang et al., 2023a) suggest that focusing on modality-specific information can reduce interference from irrelevant or redundant information, thus enhancing the robustness of multimodal models. However, these methods often require significant computational and storage resources, rendering them impractical for resource-constrained real-world applications. Hence, we suggest

introducing an orthogonal loss in the training stage to disentangle modal-specific knowledge without requiring additional computation and memory costs at inference.

The lack of sufficient multimodal data is another key challenge. Liu et al. (2015) introduced a zero-masking strategy, replacing missing modalities with zero values to preserve model functionality. Subspace projection methods (Zhou et al., 2019b,a, 2020) extracted features from existing modalities and then averaged in a latent space to replace missing data. Alternatively, imputation methods aim to reconstruct missing modalities based on available data (Wang et al., 2024; Gao et al., 2021). In addition, feature-sensitive imputation techniques (Pan et al., 2022; Gao et al., 2023) emphasize extracting disease-related information from existing modalities during the imputation process. To address the challenge of incomplete multimodal longitudinal data, Xu et al. (2022) proposed a deep latent representation collaborated sequence learning framework that handles arbitrary modality-missing patterns and variable-length sequences through degradation networks and RNN-based progression modeling. Building upon this, Dao et al. (2024) introduced LMDP-Net with a variational autoencoder-based fusion module to handle modality uncertainty and an improved LSTM mechanism (IRLSTM) to optimize information flow in longitudinal data. However, these methods often suffer from high computational complexity, resulting in increased training and inference costs, and may introduce biased information, ultimately leading to suboptimal performance. Hence, transferring knowledge from complete modalities to primary modalities is a potential way to avoid the impact of missing modalities and reduce computational complexity, and knowledge distillation is a popular framework for efficiently transferring knowledge from complex teacher models to simple student models.

Recent advancements in multimodal integration have also emphasized the importance of capturing both shared and modality-specific information to improve diagnostic robustness. For instance, low-rank tensor fusion techniques and shared-specific feature modeling frameworks have been proposed to exploit complementary information while reducing redundancy across modalities (Wang et al., 2023a; Qiu et al., 2024). These methods aim to learn a common latent space where multimodal data can be effectively combined, even in the presence of missing or incomplete modalities. Moreover, several studies have begun to incorporate clinical metadata, such as cognitive scores, genetic markers, and demographic information, alongside neuroimaging data to create more holistic and clinically actionable models (Qiu et al., 2022; Wang et al., 2024; Wu et al., 2025). This trend toward integrative and clinically informed multimodal learning highlights a growing recognition that combining imaging with non-imaging data can significantly enhance early diagnosis and progression prediction in Alzheimer's disease, achieving more personalized and precise clinical applications.

2.2. Knowledge distillation

Knowledge distillation (KD) (Gou et al., 2021) is widely adopted to transfer knowledge from teacher models to student models (Yang et al., 2023; Wang et al., 2023b; Song et al., 2023). Yang et al. (2023) and Song et al. (2023) introduced knowledge distillation to improve the diagnostic performance of AD based on MRI.Van Sonsbeek et al. (2021) employed variational knowledge distillation to transfer disease-related knowledge from Electronic Health Records (EHR) to X-ray images.

However, traditional knowledge distillation methods often focus on knowledge transfer while overlooking the ability to discern the importance of knowledge. In contrast, attention distillation offers a more comprehensive approach by transferring the representational power of teacher models, particularly in capturing contextual dependencies. For example, Wang et al. (2020a) proposed a novel distillation framework that emphasizes the transfer of self-attention scores from the teacher model. By distilling attention maps, the student model can more effectively mimic the behavior of the teacher model. Wang et al. (2019) demonstrated significant improvements in transfer learning efficiency and performance through attention distillation. Nevertheless, the distillation of intermediate features has not been fully explored. In particular, due to the robust characteristics of intermediate features compared to high-level features and soft labels, the distillation of attention on intermediate features can transfer more shallow knowledge, improving the generalization performance of student models.

In addition, Yang et al. (2023) introduced a gated regularized distillation mechanism, enabling the student model to learn reliable knowledge from the teacher model. Similarly, (Yang et al., 2022b) refined the distillation through a confidence regularization distillation mechanism. However, the output of the teacher model may be inaccurate because of noise, leading to inconsistent knowledge transfer and thus affecting the distillation effect. We suggest dynamically applying confidence scores for flexible gated distillation to enhance accurate knowledge transfer by comparing the outputs between the teacher and student models. Moreover, although attention distillation (Wang et al., 2019, 2020a) and gated regularization (Yang et al., 2023) partially mitigate the issue of insufficient transfer of discriminating knowledge and reliable distillation, their static designs (e.g., fixed attention modules (Hu et al., 2018), error-based uncertainty (Yang et al., 2022b)) limit adaptability to dynamic missing-modality scenarios.

To this end, we propose a Consistency Refinement-driven Multi-level Self-Attention Distillation framework, which integrates confidence regularization and attention mechanisms to improve the distillation efficiency and the generalization performance of student models. In particular, we design a parameter-free attention module to align multiscale intermediate features, and then dual confidence regularization strategies ensure accurate knowledge transfer.

2.3. Challenges in clinical deployment

Beyond the technical limitations of existing multimodal and distillation methods, several broader challenges impede the widespread adoption of multimodal AI systems in clinical practice. A significant hurdle is the inherent heterogeneity of medical data, which varies in resolution, acquisition protocols, and quality across institutions. This variability can lead to domain shift, reducing model generalization when deployed in real-world settings (Ghifary et al., 2015). Furthermore, missing modalities are not merely a technical inconvenience but a systemic issue in healthcare, influenced by factors such as cost, patient compliance, and clinical guidelines (Haque et al., 2017). While imputation and distillation offer partial solutions, they often assume a static missingness pattern, which rarely holds in dynamic clinical environments. Finally, computational and infrastructural constraints in hospitals, such as limited GPU resources and data privacy requirements, favor lightweight, efficient models that can operate near real-time without compromising patient data security (Arbabshirani et al., 2018). These practical considerations underscore the need for robust, efficient, and interpretable multimodal learning frameworks that are not only accurate but also deployable in diverse clinical contexts.

3. Method

This section first presents an overview of our proposed framework. Subsequently, each specifically designed module for the proposed framework is introduced in detail.

3.1. Problem setting

Let $X = \{X_i\}_{i=1}^{N}$ represent the training data used in this study, $Y = \{Y_i\}_{i=1}^{N}$ are the corresponding diagnostic labels, where N is the number of data, and (X_i, Y_i) indicates the ith data and label, respectively. Each data contains multiple modalities, i.e., $X_i = \{X_{i,j}\}_{j=1}^{M_i}$, where M_i represents the number of available modalities in X_i. In particular, the modalities used in this study include MRI, PET, and Mini-Mental State Examination (MMSE). In particular, not all subjects have PET; most have only MRI. The objective is to predict the disease label Y based on the given X.

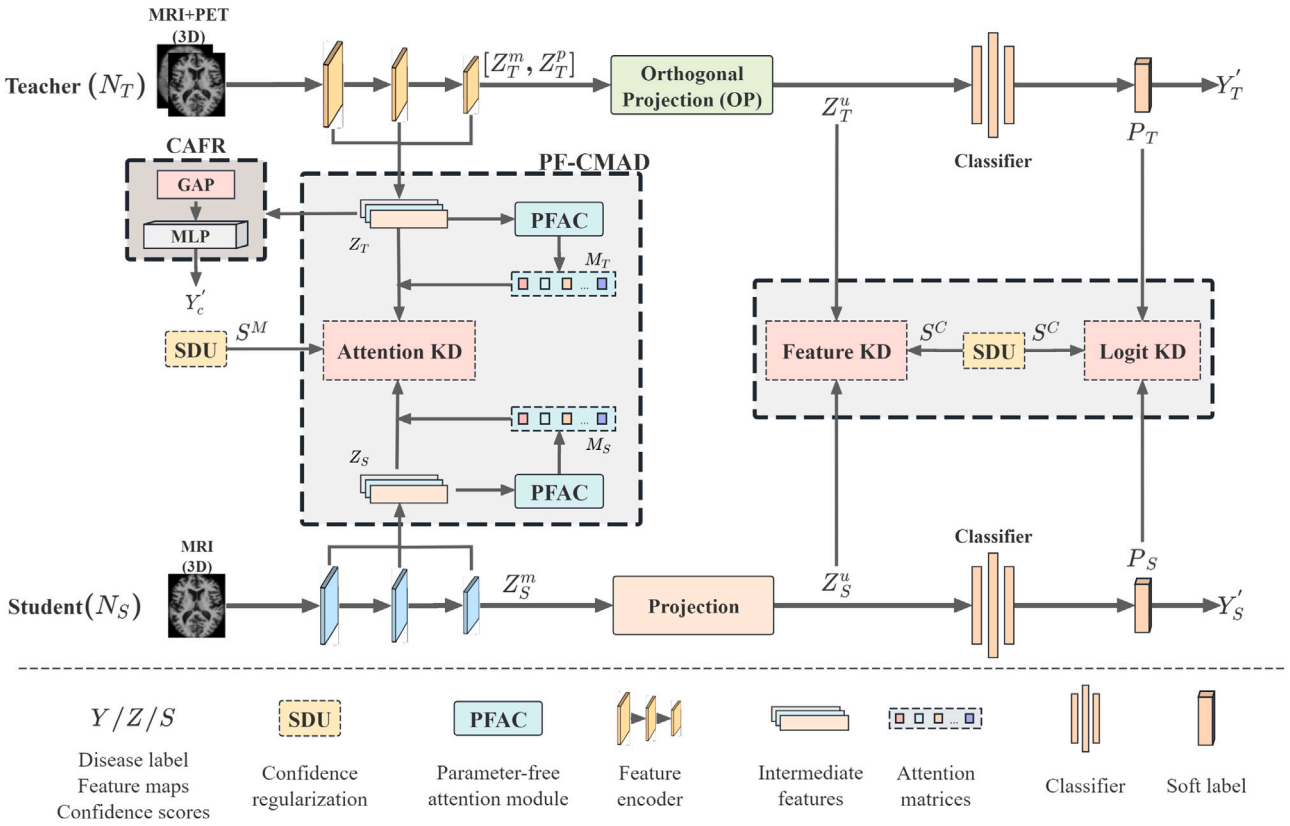

Fig. 1. Overview of the proposed Consistency Refinement-driven Multi-level Self-Attention Distillation framework. The $\mathcal{N}_T$ learns from paired MRI and PET, whereas $\mathcal{N}_S$ only learns from MRI. First, for $\mathcal{N}_T$, the cognitive awareness feature refinement (CAFR) improves the disease-aware ability of the intermediate features by predicting cognitive scores, and the orthogonal projection (OP) disentangles the modality-specific representation to reduce redundancy. Then, the cross-modal self-attention distillation module (PF-CMAD) aligns intermediate features between $\mathcal{N}_T$ and $\mathcal{N}_S$, achieving efficient shallow knowledge distillation. Moreover, the smooth distillation unit (SDU) employs the consistency-aware regularization strategy to refine the distillation on intermediate- and high-level features.

3.2. Overall architecture

The proposed CRAD framework is designed to improve AD diagnosis under missing modalities through multimodal knowledge distillation. As illustrated in Fig. 1, the teacher model $\mathcal{N}_T$ learns from multimodal data, while the student model $\mathcal{N}_S$ learns solely from MRI data. The encoder extracts shallow features from the input data X. These features are then projected into a latent space to capture high-level features. The classifier outputs the predicted disease label Y'. This process can be described as follows:

$$Y' = \text{Classifier}(\text{Projection}(\text{Encoder}(X))). \tag{1}$$

The teacher model aims to provide high-quality multimodal knowledge that will be transferred to the MRI-only student model. A cognitive awareness feature refinement module (CAFR) is integrated into the encoder of $\mathcal{N}_T$ to identify intermediate discriminating features by predicting the clinical cognitive score (i.e., MMSE), while a orthogonal disentanglement module (OP) is employed to the output features of encoder to reduce feature redundancy (see Section 3.2.1 for details). The teacher model is pre-trained based on paired MRI and PET images, and it is supervised by the ground truth, namely, the disease label. In contrast to the teacher model, the student model is a lightweight architecture with only a feature encoder and classifier. We enhance the student model by transferring knowledge from the teacher model, in addition to being supervised by the ground truth.

During knowledge distillation, we do not train the teacher model. Instead, we design a parameter-free cross-modal self-attention distillation module (PF-CMAD) to distill intermediate knowledge from $\mathcal{N}_T$ to $\mathcal{N}_S$ (see Section 3.3). In addition, in contrast to the traditional gated regularization described in Section 3.4, we propose a smooth distillation unit (SDU) to implement a consistency-aware confidence regularization strategy, improving the reliability of knowledge distillation (see Section 3.5). The details of these components are presented in the following subsections. We provide a train procedure in Appendix A.1 and a detailed data flow (input and output) in Appendix A.2.

3.2.1. Multimodal teacher model

For the teacher model $\mathcal{N}_T$, we use a 3D convolutional neural network (CNN) as a feature encoder to extract multimodal features. Although our proposed framework supports any 3D CNN, VGG (Simonyan and Andrew, 2015) is selected as the primary backbone because it achieves the best performance with moderate complexity (Table 9 illustrates the backbone comparison). First, an auxiliary cognitive awareness feature refinement module (see Fig. 1 (CAFR)) is designed to improve disease awareness of intermediate features $Z_T = [Z_T^1, Z_T^2, \ldots, Z_T^K]$ by predicting the MMSE score, where K is the number of layers of intermediate features. Let Z_T^k be the features of the kth layer. $Y_c' = h\left(\text{GAP}(Z_T^k; W^k)\right)$ is the predicted value of the true MMSE Y_c, where $h(\cdot)$ is used to calculate Y_c' from Z_T^k with learnable parameters W^k, while GAP$(\cdot)$ is the Global Average Pooling operation. The Mean Squared Error Loss $\mathcal{L}_{MSE}$ is applied to evaluate the prediction error between Y_c' and Y_c. As shown in Eq. (2), by minimizing $\mathcal{L}_{MSE}$, the $\mathcal{N}_T$

is encouraged to capture disease-related features.

$$\begin{aligned}\mathcal{L}_{MSE} &= \frac{1}{K}\sum_{k=1}^{K}\|Y_c' - Y_c\|_2^2 \\ &= \frac{1}{K}\sum_{k=1}^{K}\|h\left(\mathrm{GAP}\left(Z_T^k\right); W^k\right) - Y_c\|_2^2\end{aligned} \tag{2}$$

Meanwhile, inspired by Ranasinghe et al. (2021), we introduce orthogonal projection loss $\mathcal{L}_{ORT}$ to disentangle modality-specific features. Specifically, we perform an orthogonal decomposition among multimodal features $[Z^m, Z^p]$ of $\mathcal{N}_T$, where the Z_T^m and Z_T^p are the features extracted from the paired MRI and PET. $\mathcal{L}_{ORT}$ aims to refine the consistent distribution of intra-modal features while increasing the distance of inter-modal features:

$$\mathcal{L}_{ORT} = \sum_{(i,j)\in\{(m,p),(p,m)\}}\Big(\left(1 - s(Z_T^i, Z_T^i)\right) + d(Z_T^i, Z_T^j)\Big), \tag{3}$$

where s and d indicate the feature cosine similarity. We utilize $\mathcal{L}_{ORT}$ to constrain s to be close to 1, while d is close to 0.

Subsequently, the disentangled Z_T^m and Z_T^p are fused as Z_T^u and then input into the classifier to predict the disease label Y_T'. We use the Cross-Entropy function to calculate classification loss:

$$\mathcal{L}_{CE} = -\sum\left(Y\log Y_T' + (1-Y)\log(1-Y_T')\right). \tag{4}$$

Therefore, the optimization objective of the teacher model can be formulated as:

$$\mathcal{L}_{N_T} = \mathcal{L}_{CE} + \gamma_1\mathcal{L}_{MSE} + \gamma_2\mathcal{L}_{ORT}, \tag{5}$$

where γ_1 and γ_2 are weight factors to balance the contribution of $\mathcal{L}_{MSE}$ and $\mathcal{L}_{ORT}$.

3.2.2. Distillation between teacher and student models

Similarly to $\mathcal{N}_T$, we use a simple 3D convolutional neural network to extract the features Z_S^m from the MRI. The Z_S^m is then projected into the latent space to learn the high-level semantic features Z_S^u for the classifier to output the predicted disease label Y_S'.

Following previous studies (Guan et al., 2021; Chen et al., 2023; Van Sonsbeek et al., 2021), we align the high-level features between $\mathcal{N}_T$ and $\mathcal{N}_S$ by knowledge distillation, i.e., feature distillation (FD). This process can be described as:

$$\mathcal{L}_{KD}^U = \sum KL\left(\tilde{Z}_T^u \otimes \left(\tilde{Z}_T^u\right)^T \| \tilde{Z}_S^u \otimes \left(\tilde{Z}_S^u\right)^T\right), \tag{6}$$

where the $\otimes$ represents matrix multiplication and the KL is Kullback–Leibler divergence (Kullback and Leibler, 1951). In particular, Z_T^u and Z_S^u are normalized along the channel dimension, i.e., $\tilde{Z}_T^u = Z_T^u/(\max(\|Z_T^u\|_2, \epsilon))$, and ϵ is a small positive real number used to avoid division by zero.

Knowledge distillation is also applied to align the soft labels, i.e., soft label distillation (SD). The loss of distillation $\mathcal{L}_{KD}^P$ is calculated on the soft labels Y_T' and Y_S', defined in formula (7). Therefore, the distillation loss between $\mathcal{N}_S$ and $\mathcal{N}_T$ can be summarized as $\mathcal{L}_{KD} = \mathcal{L}_{KD}^U + \mathcal{L}_{KD}^P$.

$$\mathcal{L}_{KD}^P = \sum KL\left(Y_T', Y_S'\right) \tag{7}$$

Moreover, as shown in Fig. 2, we not only focus on high-level feature alignment (Yang et al., 2023; Wang et al., 2019), but also design a special cross-modal self-attention distillation module (PF-CMAD) to distill the attentive intermediate features (refer to Section 3.3 for details). This module leverages feature similarity-based self-attention to capture cross-modal correlations. In addition, we develop a smooth distillation unit (SDU) that employs a consistency-aware confidence regularization strategy to dynamically control the distillation process by evaluating prediction divergence between student and teacher models, ensuring reliable knowledge transfer. Please refer to Sections 3.4 and 3.5 for details of this confidence regularization strategy.

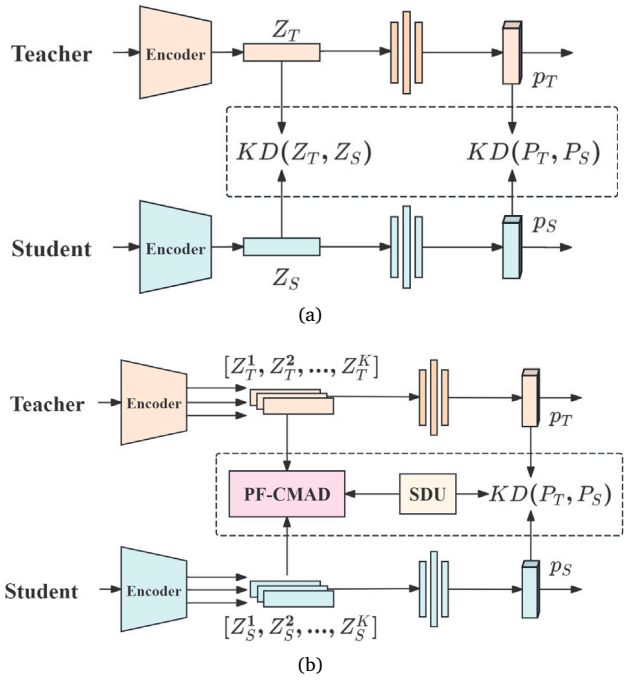

Fig. 2. (a) Previous knowledge distillation focuses on aligning high-level features. (b) Our proposed "Consistency Refinement-driven Multi-level Self-Attention Distillation" aligns attentive features (PF-CMAD) with consistency evaluation (SDU) within multiple layers of the student-teacher model.

For the student model $\mathcal{N}_S$, it is not only supervised by the teacher model's output, but also learns from the disease label Y. Thus, the objective function for optimizing $\mathcal{N}_S$ is defined as:

$$\mathcal{L}_{N_S} = \mathcal{L}_{CE} + \lambda\mathcal{L}_{KD}, \tag{8}$$

where λ is a weight factor to balance the contribution of $\mathcal{L}_{KD}$.

3.3. Parameter-free cross-modal self-attention distillation

Apart from distilling knowledge by aligning high-level features, we propose the cross-modal transfer of more robust shallow knowledge with self-attention and refinement. We develop a self-attention distillation module (PF-CMAD) to hierarchically transfer cross-modal discriminating knowledge and incorporate a consistency-driven confidence regularization strategy (SDU) to refine knowledge. As shown in Fig. 3, the PF-CMAD utilizes hierarchical attention distillation of intermediate features to capture contextual information at different levels for multi-granularity knowledge transfer. For each self-attention block, we design a simple yet effective parameter-free attention converter (PFAC) based on features' self-similarity to identify the feature's importance and then transfer discriminating shallow knowledge by fusing the intermediate attentive features, i.e., attention distillation (ATD). In contrast to traditional attention modules (e.g., SENet (Hu et al., 2018)) that calculate the attention weights with learnable parameters, our proposed PFAC adaptively infers attention maps without introducing trainable parameters. As shown in Fig. 4, the PFAC first normalizes the features along the channel dimension, then performs Global Average Pooling (GAP) and Global Max Pooling (GMP), followed by concatenation and reshaping to Z' with $c \times 1$ dimension, where c is the channel number. The attention matrix M is obtained by:

$$M = \mathrm{Sigmoid}(\mathrm{diag}(Z' \otimes (Z')^T)), \tag{9}$$

where $\mathrm{diag}(\cdot)$ takes the diagonal elements and $\mathrm{Sigmoid}(\cdot)$ is the activation function. Subsequently, the M is used to highlight the key

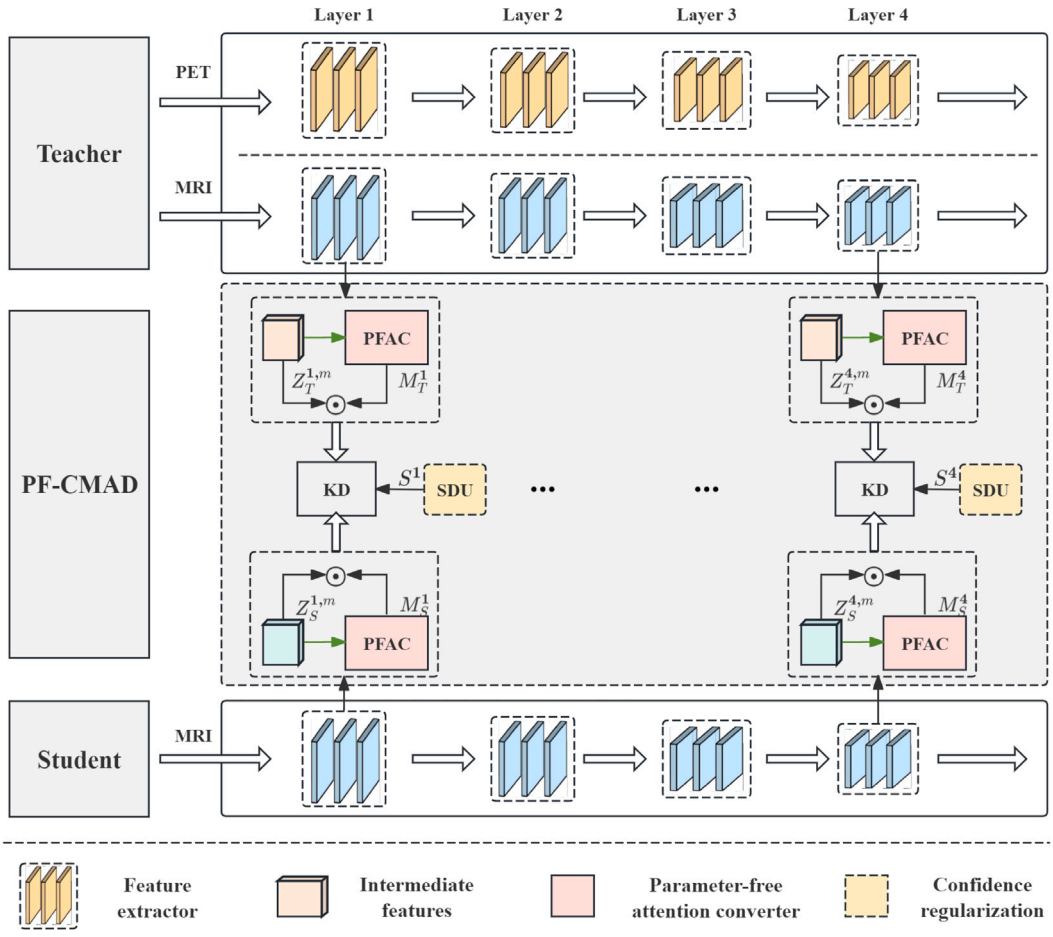

Fig. 3. Detailed design of the proposed Parameter-Free Cross-Modal Self-Attention Distillation (PF-CMAD).

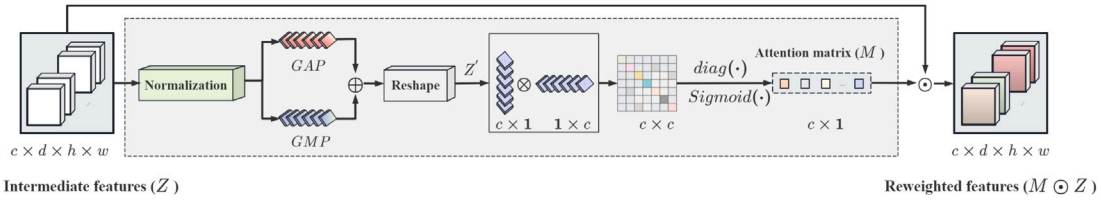

Fig. 4. Diagram of the proposed Parameter-Free Attention Converter (PFAC) module. The PFAC first normalizes the features Z along the channel dimension, then performs Global Average Pooling (GAP) and Global Max Pooling (GMP), followed by concatenation and reshaping to Z'. Then, the attention matrix M is used to screen discriminating features that are aligned between $\mathcal{N}_S$ and $\mathcal{N}_T$, as defined in Eq. (10).

intermediate features. Then we use $\mathcal{L}_{KD}^{F}$ to align attentive features between $\mathcal{N}_S$ and $\mathcal{N}_T$:

$$\mathcal{L}_{KD}^{F} = \frac{1}{K}\sum_{k=1}^{K} \| M_S^k Z_S^{k,m} - M_T^k Z_T^{k,m} \|_2^2, \tag{10}$$

where $[M_S^k, Z_S^{k,m}]$ and $[M_T^k, Z_T^{k,m}]$ are the attention matrix and feature maps (MRI) of kth intermediate layer of $\mathcal{N}_S$ and $\mathcal{N}_T$ respectively, and K is the number of layers.

3.4. Gated regularization distillation

Traditional gated regularization calculates the confidence score S^C, and applies it to distillation loss $\mathcal{L}_{KD}$, refining the knowledge of the teacher model. Usually, S^C is obtained by measuring the distance between the output of $\mathcal{N}_T$ and the ground truth. For example, S^C can be calculated by the Euclidean distance function dis(·), and then a clip(·) is employed to limit the upper bound of S^C. Subsequently, the distance is converted to a confidence score limited to [0, 1]. This can be formulated as:

$$S^C = 1 - \text{clip}\left(\text{dis}\left(Y_T', Y\right)\right), \tag{11}$$

where Y_T' and Y are the output of $\mathcal{N}_T$ and the ground truth, respectively.

However, the confidence score obtained by Eq. (11) ignores the relative errors between $\mathcal{N}_S$ and $\mathcal{N}_T$, a penalty should be applied to S^C if the prediction error of $\mathcal{N}_T$ is greater than that of $\mathcal{N}_S$, namely, learning from $\mathcal{N}_T$ should be softened. Intuitively, S^C can be constrained by an additional regularization term ψ, where $\psi = 0$ indicates $\text{dis}\left(Y_S{}', Y\right) \leq \text{dis}\left(Y_T', Y\right)$, otherwise $\psi = 1$. Setting the confidence score to 0 may result in suboptimal results. Therefore, we employ a higher penalty for larger errors and a lower penalty for smaller errors. The ψ is defined as:

$$\psi = \left(\frac{\text{dis}\left(Y_S', Y\right)}{\text{dis}\left(Y_T', Y\right) + \text{dis}\left(Y_S', Y\right) + \epsilon}\right)^2, \tag{12}$$

Table 1
Demographic information and data distribution of the studied subjects.

Dataset	Group	Modality		Sex		Age	MMSE
		Paired	MRI-only	Male	Female	(Mean ± Std)	(Mean ± Std)
ADNI	AD	183	89	146	126	75.03 ± 7.65	23.14 ± 2.09
	pMCI	86	171	149	108	73.78 ± 7.12	26.94 ± 1.81
	sMCI	131	400	305	226	72.42 ± 7.73	27.88 ± 1.73
	NC	232	121	175	178	74.97 ± 5.78	29.11 ± 1.09
AIBL	AD	–	74	30	44	73.35 ± 7.93	20.18 ± 5.44
	pMCI	–	11	7	4	74.90 ± 5.97	26.27 ± 1.60
	sMCI	–	69	33	36	75.36 ± 7.54	27.04 ± 2.13
	NC	–	85	30	55	75.52 ± 6.63	28.71 ± 1.35

where ϵ is a small positive real number used to avoid division by zero.

Then, the objective function for optimizing the $\mathcal{N}_S$ can be further reformulated as:

$$\mathcal{L}_{N_S} = \lambda \mathcal{L}_{CE} + (1-\lambda)\psi S^C \mathcal{L}_{KD}. \tag{13}$$

In addition, we design a flexible regularization strategy to ensure accurate knowledge transfer from a global perspective. Specifically, the confidence scores for regularization are calculated by predicting cognitive scores (from intermediate features) and disease labels (from global features), respectively. Then, these two types of confidence scores are used as independent regularization penalties to distill shallow and high-level knowledge, respectively (see Dual Smooth Distillation Units (SDU) in Section 3.5).

3.5. Smooth Distillation Units (SDU)

In contrast to the conventional knowledge refinement strategy (Yang et al., 2023) that only evaluates the teacher model's certainty based on soft probabilities, we propose the smooth distillation unit (SDU) to obtain flexible and reliable knowledge transfer. Specifically, different regularization terms are employed to refine knowledge at different levels. Moreover, we further evaluated the consistency of the output of the student-teacher model at different levels to avoid transferring noisy signals (see Section 3.4). Let $Z = \left[Z^1, Z^2, \ldots, Z^K\right]$ denote the multiscale intermediate features. Confidence score S^k of the kth layer is obtained by measuring the distance between the predicted MMSE score Y_c' and the ground truth Y_c:

$$S^k = 1 - \text{clip}\left(\| \left(Y_c' - Y_c\right) \|_2^2\right). \tag{14}$$

The confidence scores for all intermediate layers can be denoted as $S^M = \left[S^1, S^2, \ldots, S^K\right]$, then S^M is used to regularize the distillation on intermediate features.

Moreover, we apply dual confidence scores S^M and S^C, and extend the $\mathcal{L}_{KD}$ to further refine the knowledge transfer between $\mathcal{N}_S$ and $\mathcal{N}_T$:

$$\mathcal{L}_{KD} = S^M \mathcal{L}_{KD}^F + S^C\left(\mathcal{L}_{KD}^P + \mathcal{L}_{KD}^U\right), \tag{15}$$

where S^M and S^C (see Eq. (11)) are two confidence scores for distillation on shallow and high-level knowledge, respectively.

4. Results and discussion

In this section, we first introduce the data preprocessing and performance evaluation metrics. Then, we briefly review the competing methods and compare our proposed method with them on two AD-related tasks based on extensive analysis.

4.1. Experimental setting

4.1.1. Dataset

Our study utilizes two databases, including the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Australian Imaging, Biomarker and Lifestyle Flagship Study of Aging (AIBL). We acquired 1.5T/3T T1-weighted structural MRI scans and 18F-FDG Positron Emission Tomography scans from the ADNI database. Moreover, we retrieved 1.5T/3T T1-weighted MRI scans from the AIBL database. All of the scans are taken at their baseline/screening visits. The data are categorized into three groups: Alzheimer's disease (AD), Mild Cognitive Impairment (MCI), and Normal Control (NC), following their diagnosis labels. The MCI can be divided into progressive MCI (pMCI) and stable MCI (sMCI). The pMCI means MCI subjects would convert to AD within 36 months after the baseline visit, while sMCI means MCI subjects would remain stable after the baseline visit. The demographic details of the studied subjects are shown in Table 1. The subjects labeled 'AD' and 'NC' were selected for the AD diagnosis task. For predicting MCI conversion (pMCI v.s. sMCI), the subjects labeled 'MCI' at the baseline screen were selected. Note that subjects in this study were selected based on their diagnostic label, without considering other detailed criteria such as sex, age, slice thickness, or device manufacturer. The data from the ADNI database are used to train and test the models, while the data in AIBL are used only for testing the models' generalization.

We evaluated the CRAD framework on two tasks: diagnosing Alzheimer's disease (AD-NC classification) and predicting the conversion of Mild Cognitive Impairment (MCI) to AD (pMCI-sMCI classification) using five-fold cross-validation. The data was divided into five folds at the subject level, ensuring a balanced distribution of classes: AD, NC, sMCI, and pMCI. During each of the five training sessions, one fold was reserved for testing, and the remaining four folds were used for training. The ratio of training data to test data is 8:2, and 10% of the training data was randomly selected as a validation set, ensuring no overlap with the test set.

4.1.2. Preprocessing

Following common practice, we performed a preprocessing pipeline on the original images, including spatial registration and tissue segmentation. We performed registration to transform MRI and PET to the MNI152 template (Fonov et al., 2011) based on the Statistical Parametric Mapping and Computational Anatomy Toolbox (Gaser et al., 2024). Besides, the PET scans are aligned to the space of the corresponding MRI. After the preprocessing, MRI and PET are resized to $113 \times 113 \times 137$ voxels.

4.2. Implementation

The proposed CRAD framework is implemented with the PyTorch framework (Paszke et al., 2019) and trained with an NVIDIA GTX 3090 GPU for 300 epochs. We used Adam as the optimizer with a learning rate fixed to 0.0001 and a batch size of 8. The hyperparameter λ is set to 0.8 while γ_1 and γ_2 are set to 1.0 in our experiments. Following the common practice, we applied multiple metrics, including Accuracy (ACC), Sensitivity (SEN), Specificity (SPE), the weighted F1-score, and the area under the receiver operating characteristic curve (AUC), to evaluate the performance of the proposed method and competing methods.

4.3. Comparison with competing methods

We conducted comprehensive comparisons between the proposed CRAD framework and eight existing methods, including two baseline models and six state-of-the-art approaches: (1) a single-modal baseline (SM-BL) (Korolev et al., 2017) using only MRI; (2) a multimodal baseline (MM-BL) (Han et al., 2019) with both MRI and PET; (3) a multimodal disease-induced network (MDL-Net) (Qiu et al., 2024); (4) an imputation-based model (TPA-GAN) (Gao et al., 2021) that synthesizes PET from MRI; (5) a gated regularization knowledge distillation method (CReg-KD) (Yang et al., 2023); (6) an attentive feature distillation scheme (AFDS) (Wang et al., 2019); and (7) two recent cross-modal distillation techniques (DFTD (Chen et al., 2023) and IC-MKD (Kwak

Table 2
Performance comparison of different methods on the ADNI dataset. The best results are **highlighted**. The results are shown with mean and standard deviation (Mean ± Std) across five folds.

Methods	AD vs. NC					pMCI vs. sMCI				
	ACC↑	AUC↑	SEN↑	SPE↑	F1-score↑	ACC↑	AUC↑	SEN↑	SPE↑	F1-score↑
SM-BL (Korolev et al., 2017)	0.894 ± 0.04[a]	0.881 ± 0.05[a]	0.934 ± 0.04[a]	0.827 ± 0.11[a]	0.911 ± 0.04[a]	0.850 ± 0.04[a]	0.853 ± 0.03[a]	0.924 ± 0.10	0.781 ± 0.14[a]	0.854 ± 0.02[a]
MM-BL (Han et al., 2019)	0.915 ± 0.06[a]	0.914 ± 0.06[a]	0.934 ± 0.06	0.894 ± 0.09[a]	0.927 ± 0.05[a]	0.906 ± 0.04	0.906 ± 0.05	0.933 ± 0.06	0.880 ± 0.08	0.903 ± 0.05
MDL-Net (Qiu et al., 2024)	0.953 ± 0.04	0.956 ± 0.04	0.954 ± 0.06	0.958 ± 0.06[a]	0.957 ± 0.04	**0.911 ± 0.04**	0.906 ± 0.05	0.918 ± 0.05	0.893 ± 0.08	0.907 ± 0.04
TPA-GAN (Gao et al., 2021)	0.861 ± 0.02[a]	0.924 ± 0.01[a]	0.860 ± 0.02[a]	0.861 ± 0.05[a]	0.887 ± 0.01[a]	0.846 ± 0.01[a]	0.885 ± 0.01[a]	0.749 ± 0.03[a]	**0.912 ± 0.02**	0.782 ± 0.04[a]
CReg-KD (Yang et al., 2023)	0.934 ± 0.02[a]	0.937 ± 0.02[a]	0.935 ± 0.06	0.940 ± 0.06[a]	0.943 ± 0.02[a]	0.906 ± 0.03	0.913 ± 0.03	**0.956 ± 0.04**	0.871 ± 0.06	0.904 ± 0.03
AFDS (Wang et al., 2019)	0.945 ± 0.02[a]	0.930 ± 0.04[a]	**0.981 ± 0.03**	0.878 ± 0.09[a]	0.956 ± 0.01[a]	0.874 ± 0.03[a]	0.863 ± 0.03[a]	0.901 ± 0.08	0.824 ± 0.12	0.871 ± 0.03[a]
DFTD (Chen et al., 2023)	0.934 ± 0.05[a]	0.945 ± 0.04[a]	0.906 ± 0.09[a]	0.983 ± 0.03	0.941 ± 0.04[a]	0.872 ± 0.02[a]	0.875 ± 0.01[a]	0.858 ± 0.02[a]	0.892 ± 0.04[a]	0.884 ± 0.01[a]
IC-MKD (Kwak et al., 2025)	0.935 ± 0.03[a]	0.932 ± 0.02[a]	0.926 ± 0.06	0.937 ± 0.06[a]	0.944 ± 0.03[a]	0.853 ± 0.04[a]	0.847 ± 0.04[a]	0.816 ± 0.14[a]	0.878 ± 0.08[a]	0.848 ± 0.07[a]
CRAD	**0.959 ± 0.02**	**0.966 ± 0.01**	0.947 ± 0.04	**0.983 ± 0.03**	**0.965 ± 0.01**	**0.911 ± 0.04**	**0.919 ± 0.03**	0.937 ± 0.04	0.900 ± 0.08	**0.907 ± 0.04**

[a] Denotes that the performance improvements of our proposed method have statistical significance ($p < 0.05$) based on a paired t-test.

Table 3
Generalization performance comparison of different methods on the AIBL dataset. The best results are **highlighted**. The results are shown with mean and standard deviation (Mean ± Std) across five folds.

Methods	AD vs. NC					pMCI vs. sMCI				
	ACC↑	AUC↑	SEN↑	SPE↑	F1-score↑	ACC↑	AUC↑	SEN↑	SPE↑	F1-score↑
SM-BL (Korolev et al., 2017)	0.870 ± 0.01	0.874 ± 0.01	0.919 ± 0.05	**0.866 ± 0.05**	0.866 ± 0.01	0.804 ± 0.07[a]	0.860 ± 0.06[a]	0.937 ± 0.09[a]	0.782 ± 0.09[a]	0.685 ± 0.12[a]
MM-BL (Han et al., 2019)	0.812 ± 0.07[a]	0.822 ± 0.06	0.932 ± 0.05	0.711 ± 0.15[a]	0.823 ± 0.05[a]	0.800 ± 0.14[a]	0.860 ± 0.08[a]	0.955 ± 0.10[a]	0.765 ± 0.18	0.662 ± 0.12[a]
MDL-Net (Qiu et al., 2024)	0.826 ± 0.04[a]	0.834 ± 0.03[a]	0.926 ± 0.05	0.742 ± 0.11[a]	0.831 ± 0.02[a]	0.865 ± 0.10	0.865 ± 0.06[a]	0.866 ± 0.14	0.865 ± 0.14	0.724 ± 0.11
TPA-GAN (Gao et al., 2021)	0.851 ± 0.03[a]	**0.921 ± 0.01**	0.840 ± 0.06[a]	0.871 ± 0.06[a]	**0.877 ± 0.03**	0.834 ± 0.03[a]	0.889 ± 0.01[a]	0.749 ± 0.08[a]	**0.893 ± 0.04**	**0.783 ± 0.05**
CReg-KD (Yang et al., 2023)	0.807 ± 0.03[a]	0.811 ± 0.03[a]	0.862 ± 0.04[a]	0.760 ± 0.07[a]	0.806 ± 0.02[a]	0.767 ± 0.11[a]	0.805 ± 0.09[a]	0.866 ± 0.09[a]	0.745 ± 0.13[a]	0.596 ± 0.14[a]
AFDS (Wang et al., 2019)	0.780 ± 0.06[a]	0.786 ± 0.05[a]	0.874 ± 0.07[a]	0.698 ± 0.15[a]	0.789 ± 0.03[a]	0.849 ± 0.08	0.864 ± 0.06[a]	0.889 ± 0.08[a]	0.840 ± 0.10	0.698 ± 0.11
DFTD (Chen et al., 2023)	0.873 ± 0.01	0.874 ± 0.01[a]	0.885 ± 0.03[a]	0.864 ± 0.01	0.865 ± 0.01[a]	0.844 ± 0.07	0.847 ± 0.09	0.852 ± 0.13[a]	0.842 ± 0.06	0.672 ± 0.13[a]
IC-MKD (Kwak et al., 2025)	0.870 ± 0.04[a]	0.871 ± 0.02[a]	0.881 ± 0.03[a]	0.857 ± 0.04	0.861 ± 0.01[a]	0.789 ± 0.02[a]	0.813 ± 0.03[a]	0.852 ± 0.06[a]	0.775 ± 0.03[a]	0.598 ± 0.04[a]
CRAD	**0.874 ± 0.03**	0.879 ± 0.02	**0.938 ± 0.04**	0.820 ± 0.07	0.873 ± 0.02	**0.886 ± 0.06**	**0.921 ± 0.05**	**0.978 ± 0.05**	0.865 ± 0.07	0.768 ± 0.11

[a] Denotes that the performance improvements of our proposed method have statistical significance ($p < 0.05$) based on a paired t-test.

Table 4
Component ablation results of the teacher model on the AD diagnosis task.

	OP	CAFR	ACC↑	AUC↑	SEN↑	SPE↑	F1-score↑
1			0.881 ± 0.02	0.871 ± 0.04	0.909 ± 0.03	0.832 ± 0.11	0.899 ± 0.02
2	✓		0.941 ± 0.03	0.939 ± 0.03	0.930 ± 0.05	0.949 ± 0.05	0.946 ± 0.03
3		✓	0.934 ± 0.01	0.924 ± 0.02	0.966 ± 0.03	0.883 ± 0.06	0.944 ± 0.01
4	✓	✓	**0.965 ± 0.02**	**0.969 ± 0.02**	**0.980 ± 0.03**	**0.958 ± 0.06**	**0.964 ± 0.02**

et al., 2025)). All methods were trained and evaluated under the same dataset settings to ensure a fair comparison.

As summarized in Table 2, our CRAD method consistently achieves superior performance in both AD vs. NC classification and pMCI vs. sMCI prediction tasks. Specifically, CRAD attains the highest scores in ACC, AUC, and F1-Score, demonstrating its effectiveness and robustness. It is noteworthy that multimodal methods (Qiu et al., 2024; Han et al., 2019) generally outperform the single-modal baseline (Korolev et al., 2017), underscoring the benefit of integrating complementary information from multiple modalities. Furthermore, knowledge distillation-based approaches (Yang et al., 2023; Chen et al., 2023; Wang et al., 2019; Kwak et al., 2025) yield noticeably better results than the imputation-based TPA-GAN (Gao et al., 2021), affirming the advantage of distillation over synthesis in handling missing modalities.

To assess generalization capability, we evaluated all models on the AIBL dataset, which is not used for training. As shown in Table 3, although all methods exhibit performance degradation due to domain shift, CRAD maintains the highest accuracy and robustness, further validating its strong generalization across datasets. Among distillation techniques, AFDS (Wang et al., 2019) and CReg-KD (Yang et al., 2023) show competitive results in the ADNI dataset, while DFTD (Chen et al., 2023) and IC-MKD (Kwak et al., 2025) show more robust performance.

In summary, these results highlight that CRAD not only effectively integrates multimodal information but also achieves competing performance through our proposed cognitive-aware attention distillation mechanism, even in the presence of missing data.

4.4. Ablation study

In this subsection, we evaluated the effectiveness of different components and analyzed attention distillation, confidence regularization, orthogonal projection, and modality gaps in various settings.

4.4.1. Component ablation experiments

To validate the contribution of each proposed module within the CRAD framework, we conducted extensive ablation studies on both the teacher and student networks. All experiments were performed on the ADNI dataset for the AD vs. NC classification task.

The teacher model's ablation results are presented in Table 4. The baseline teacher (Row 1) achieves an ACC of 0.881. Adding the Orthogonal Projection (OP) module (Row 2) significantly improves performance (ACC: 0.941), confirming its effectiveness in disentangling multimodal features. Introducing the Cognitive Awareness Feature Refinement (CAFR) module (Row 3) also provides a substantial boost (ACC: 0.934), demonstrating that the auxiliary task of MMSE prediction successfully enhances the learning of disease-relevant features. The combination of both OP and CAFR modules (Row 4) yields the best teacher performance (ACC: 0.965), indicating that feature disentanglement and disease-aware refinement are complementary.

The student model's ablation results are presented in Table 5. The MRI-only baseline (Row 1) serves as the starting point. Adding standard soft-label and feature distillation (SD+FD) with the SDU unit (Row 2) provides a strong baseline. Incorporating our proposed PF-CMAD module (Rows 4, 5, 6) consistently improves performance over the SD+FD baseline, with the most significant gains in specificity (SPE), highlighting its strength in refining feature distillation. The SDU unit also shows a clear positive impact by comparing rows with and without it. The complete CRAD framework (Row 7), integrating SDU, SD, FD, and PF-CMAD, achieves the best performance across almost all metrics (ACC: 0.959, AUC: 0.966), validating the synergistic effect of all components.

We found that there is a subtle performance fluctuation when combining Feature Distillation (FD) or Attention Distillation (ATD) with Soft-Label Distillation (SD); this phenomenon arises from distillation target conflicts, which we resolve via our proposed Smooth Distillation

Table 5
Component ablation results of the student model on the AD diagnosis task. The "SD" and "FD" represent soft-label distillation and feature distillation.

	SDU	SD	FD	PF-CMAD	ACC↑	AUC↑	SEN↑	SPE↑	F1-score↑
1					0.889 ± 0.05	0.870 ± 0.06	0.942 ± 0.04	0.799 ± 0.15	0.908 ± 0.04
2	✓	✓			0.928 ± 0.05	0.926 ± 0.05	0.933 ± 0.07	0.918 ± 0.06	0.951 ± 0.04
3	✓	✓	✓		0.928 ± 0.04	0.936 ± 0.04	0.900 ± 0.07	0.971 ± 0.04	0.935 ± 0.04
4	✓	✓		✓	0.922 ± 0.05	0.914 ± 0.06	**0.956 ± 0.05**	0.872 ± 0.15	0.936 ± 0.03
5	✓		✓	✓	0.935 ± 0.05	0.933 ± 0.05	0.934 ± 0.06	0.932 ± 0.07	0.945 ± 0.05
6		✓	✓	✓	0.934 ± 0.03	0.931 ± 0.03	0.945 ± 0.04	0.918 ± 0.06	0.944 ± 0.03
7	✓	✓	✓	✓	**0.959 ± 0.02**	**0.966 ± 0.01**	0.947 ± 0.04	**0.983 ± 0.03**	**0.965 ± 0.01**

Unit (SDU). FD forces high-level feature alignment between the teacher (MRI+PET) and the student (MRI-only), ignoring distribution shifts. ATD transfers attention maps of multimodal features, which may cause incompatibility with the MRI-only student's feature space. In other words, the FD and ATD attempt to align distributionally incompatible features, introducing noise that degrades performance. We mitigate this conflict via dual confidence regularization (SDU), which evaluates the teacher's certainty for intermediate and high-level features. By adaptively reweighting distillation losses, SDU suppresses conflicting distillation (FD/ATD), preserving only reliable knowledge transfer.

In summary, the ablation studies show that each module contributes to performance gains, and their combination is essential for achieving the optimal result.

4.4.2. Influence of attention distillation

To further validate the design of our proposed Parameter-Free Cross-Modality Attention Distillation (PF-CMAD) module, which utilizes an attention converter (PFAC) to calculate attention maps, we conducted a comprehensive comparative analysis with multiple widely used attention mechanisms: a baseline method without attention (w/o), four conventional attention methods with learnable parameters (SENet (Hu et al., 2018), ECA (Wang et al., 2020b), CBAM (Woo et al., 2018), SA (Vaswani et al., 2017)), and three parameter-free attention methods (AFDS (Wang et al., 2019), SimAM (Yang et al., 2021), and PFAA (Körber, 2022). In addition, we evaluated several variants: PFAC-w/o (a pooling operation without normalization), PFAC-w (a pooling operation with traditional normalization), and PFAC-w+ (a pooling operation with normalization from the PFCA (Shi et al., 2023), namely, our full PF-CMAD). We aim to verify whether introducing learnable parameters leads to better performance in our knowledge distillation framework for medical image analysis. As shown in Table 6, all attention modules bring improvements over the baseline (first row). Among the learnable modules, SENet and CBAM achieve competitive performance. However, our parameter-free PF-CMAD (PFAC-w+) consistently achieves the best balance across metrics, especially in accuracy (ACC) and specificity (SPE). It outperforms all parameterized attention modules without adding any learnable parameters. This indicates that carefully designed parameter-free attention can effectively highlight critical cross-modal features while avoiding overfitting and enhancing generalization, proving especially suitable for clinical applications with limited and noisy data. PF-CMAD offers a superior alternative to learnable attention modules in the context of cross-modality distillation, by reducing model complexity while increasing robustness and performance.

4.4.3. Effectiveness of confidence regularization

To evaluate the effectiveness of the SDU, we compared it with the traditional gating regularization methods (Yang et al., 2023, 2022b). As shown in Table 7, employing gating regularization improves the overall performance by approximately 1%, suggesting that knowledge refinement is useful. In addition, if we directly set the ψ to zero when the student's output is more confident, the AUC further improves by about 2%, implying that soft regularization (Soften) is better than simple gating regularization because it considers the consistency of results. The proposed soft regularization (Soften+) by considering relative error achieves the best performance, which suggests that applying a soft regularization strategy can enhance the reliability of distillation.

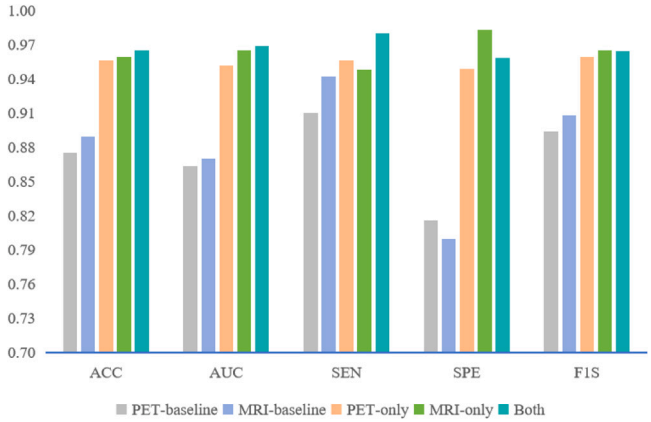

Fig. 5. Performance analysis under different modality inputs.

4.4.4. Impact of orthogonal projection

The orthogonal projection (OP) loss (Eq. (3)) aims to maximize intra-modal self-similarity by preserving critical information while minimizing inter-modal feature similarity to suppress redundancy. In contrast to Chen et al. (2023) and Kwak et al. (2025), our proposed orthogonal projection reduces computational cost and the risk of over-fitting on limited medical data because it has no extra parameters.

We evaluated various projection settings, such as orthogonal projection, shared projection, and fusion projection. As shown in Table 8, the orthogonal projection is better than the shared and fusion projections because it reduces redundant information. Combining orthogonal and shared projection on modality-specific and modality-shared features (Orthogonal+) achieves the best AUC.

4.4.5. Modality ablation analysis

A comprehensive ablation study was conducted to evaluate the performance of the proposed method under different modal inputs. As shown in Fig. 5, our teacher model achieved optimal performance when complete modalities (MRI+PET) were available (ACC: 0.965, AUC: 0.969), validating the effectiveness of multimodal information fusion. The proposed single-modal models (MRI-only and PET-only) significantly outperformed their corresponding baselines; the MRI-only model improved ACC from 0.889 to 0.959 and AUC from 0.870 to 0.965, while the PET-only model increased ACC from 0.875 to 0.956 and AUC from 0.863 to 0.952. These results confirm that the knowledge distillation framework successfully transferred knowledge from the multimodal teacher to the single-modal students, substantially enhancing the representation capability of individual modalities. Between the two single-modal variants, the MRI-only model slightly outperformed on most metrics, whereas the PET-only model showed marginally higher sensitivity, indicating the method's strong adaptability to different input modalities. Importantly, the performance of the single-modal models closely approximated that of the complete multimodal model, demonstrating that the proposed approach minimizes performance degradation while maximizing clinical practicality with limited resources.

Table 6
Performances of various attention modules.

Attention	ACC↑	AUC↑	SEN↑	SPE↑	F1-score↑
w/o	0.928 ± 0.04	0.936 ± 0.04	0.900 ± 0.07	0.971 ± 0.04	0.935 ± 0.04
SENet[a]	0.941 ± 0.03	0.944 ± 0.03	0.933 ± 0.05	0.955 ± 0.04	0.947 ± 0.02
ECA[a]	0.935 ± 0.04	0.942 ± 0.04	0.917 ± 0.12	0.967 ± 0.05	0.938 ± 0.04
CBAM[a]	0.935 ± 0.05	0.939 ± 0.05	0.917 ± 0.04	0.962 ± 0.05	0.943 ± 0.04
SA[a]	0.934 ± 0.03	0.931 ± 0.03	0.947 ± 0.06	0.915 ± 0.02	0.945 ± 0.03
AFDS	0.945 ± 0.02	0.930 ± 0.04	**0.981 ± 0.03**	0.878 ± 0.09	0.956 ± 0.01
SimAM	0.951 ± 0.02	0.939 ± 0.01	0.972 ± 0.04	0.906 ± 0.02	0.960 ± 0.02
PFAA	0.947 ± 0.03	0.953 ± 0.03	0.935 ± 0.04	0.971 ± 0.04	0.954 ± 0.02
PFAC-w/o	0.950 ± 0.02	0.950 ± 0.02	0.944 ± 0.05	0.955 ± 0.05	0.957 ± 0.02
PFAC-w	0.958 ± 0.02	0.962 ± 0.02	0.959 ± 0.03	0.964 ± 0.04	0.964 ± 0.01
PFAC-w+	**0.959 ± 0.02**	**0.966 ± 0.01**	0.947 ± 0.04	**0.983 ± 0.03**	**0.965 ± 0.01**

[a] Indicates that the attention modules have learnable parameters.

Table 7
Comparison of different confidence regularization methods.

Gating	ACC↑	AUC↑	SEN↑	SPE↑	F1-score↑
w/o	0.934 ± 0.03	0.931 ± 0.03	0.945 ± 0.04	0.918 ± 0.06	0.944 ± 0.03
Gating	0.942 ± 0.03	0.941 ± 0.03	0.945 ± 0.05	0.936 ± 0.05	0.951 ± 0.03
Soften	0.956 ± 0.02	0.964 ± 0.01	**0.949 ± 0.05**	0.978 ± 0.04	0.963 ± 0.01
Soften+	**0.959 ± 0.02**	**0.966 ± 0.01**	0.947 ± 0.04	**0.983 ± 0.03**	**0.965 ± 0.01**

Table 8
Comparison of different projection methods. "(w/o)" as a baseline represents a fusion projection that only concatenates features. "Ort" means Orthogonal.

Projection	ACC↑	AUC↑	SEN↑	SPE↑	F1-score↑
w/o	0.902 ± 0.06	0.887 ± 0.07	0.931 ± 0.05	0.843 ± 0.15	0.919 ± 0.05
Shared	0.928 ± 0.02	0.924 ± 0.02	0.931 ± 0.05	0.918 ± 0.06	0.936 ± 0.02
Ort	**0.934 ± 0.02**	0.926 ± 0.03	**0.953 ± 0.05**	0.898 ± 0.10	**0.943 ± 0.03**
Ort+	**0.934 ± 0.03**	**0.934 ± 0.03**	0.913 ± 0.04	**0.956 ± 0.05**	0.940 ± 0.03

Table 9
Performances of the proposed CRAD framework with different backbones.

Backbone	ACC↑	AUC↑	SEN↑	SPE↑	F1-score↑
VGG-16	**0.959 ± 0.02**	**0.966 ± 0.01**	0.947 ± 0.04	**0.983 ± 0.03**	**0.965 ± 0.01**
ResNet-18	0.950 ± 0.02	0.952 ± 0.02	**0.959 ± 0.03**	0.944 ± 0.04	0.957 ± 0.02
ResNet-50	0.933 ± 0.01	0.942 ± 0.01	0.923 ± 0.03	0.961 ± 0.05	0.945 ± 0.01
Densenet-121	0.944 ± 0.02	0.943 ± 0.02	0.945 ± 0.06	0.941 ± 0.06	0.953 ± 0.02

Table 10
Distillation comparison of different intermediate layers, where the layer numbers varying from 1 to 4 represent distillation from shallow to deep layers.

Layers	ACC	AUC	SEN	SPE	F1-score
1	0.926 ± 0.05	0.925 ± 0.05	0.934 ± 0.05	0.917 ± 0.07	0.938 ± 0.04
2	0.922 ± 0.04	0.928 ± 0.04	0.900 ± 0.07	0.956 ± 0.07	0.929 ± 0.04
3	0.908 ± 0.04	0.916 ± 0.04	0.892 ± 0.06	0.941 ± 0.06	0.920 ± 0.04
4	0.909 ± 0.04	0.911 ± 0.04	0.887 ± 0.05	0.936 ± 0.05	0.919 ± 0.04
3,4	0.915 ± 0.04	0.917 ± 0.05	0.925 ± 0.05	0.910 ± 0.07	0.927 ± 0.04
1,2	0.951 ± 0.03	0.955 ± 0.03	0.946 ± 0.05	0.964 ± 0.04	0.957 ± 0.03
All	0.959 ± 0.02	0.965 ± 0.01	0.948 ± 0.04	0.983 ± 0.03	0.965 ± 0.01

4.5. Impact of backbone architectures

To evaluate the influence of different backbones, we used 3D VGG (Simonyan and Andrew, 2015), 3D ResNet (He et al., 2016), and 3D Densenet (Huang et al., 2017) as encoders, respectively. As shown in Table 9, the performance of AD diagnosis only shows minor differences, indicating that the proposed CRAD is robust with different backbones. Interestingly, compared to the simple backbones (i.e., VGG (Simonyan and Andrew, 2015) and ResNet-18 (He et al., 2016)), the more complex backbones (i.e., ResNet-50 (He et al., 2016) and Densenet-121 (Huang et al., 2017)) seem to have a minor performance decrease. This may be due to the insufficient training data, which can lead to overfitting for larger models.

4.6. Layer selection for knowledge distillation

A key design consideration in our distillation framework is the selection of intermediate layers from which knowledge is transferred. Different layers in a deep network capture different levels of feature abstraction. Shallow layers typically retain structural and detailed information, while deeper layers encode semantic and high-level representations. Relying on a single layer may lead to incomplete knowledge transfer, limiting the student model's ability to mimic the teacher's full behavioral spectrum.

To determine the optimal layer combination, we conducted an extensive ablation study. We evaluated various layer configurations, including single-layer and multi-layer distillation settings. As shown in Table 10, while single-layer distillation (e.g., Layer 1 or 2) already provides competitive results, the best performance was achieved when features from all four layers were used jointly (ACC = 0.959, AUC = 0.965). This suggests that both low-level and high-level features offer complementary knowledge that collectively enhances the student's learning. Notably, the combination of early layers (1 and 2) also performed strongly, indicating the importance of shallow-level features for this task.

These results affirm that multi-layer feature integration is essential for effective knowledge distillation. Hence, in our proposed CRAD framework, we distill knowledge from all intermediate layers to maximize the student model's representational capacity and diagnostic accuracy.

Table 11
Hyperparameters sensitivity evaluation of the teacher and student models.

Parameter		ACC	AUC	SEN	SPE	F1-score
λ	0.2	0.942 ± 0.02	0.946 ± 0.02	0.928 ± 0.03	0.964 ± 0.04	0.948 ± 0.02
	0.4	0.944 ± 0.02	0.950 ± 0.02	0.949 ± 0.05	0.952 ± 0.04	0.953 ± 0.01
	0.5	0.955 ± 0.03	0.963 ± 0.02	0.948 ± 0.03	0.981 ± 0.01	0.966 ± 0.02
	0.6	0.945 ± 0.02	0.949 ± 0.02	0.943 ± 0.06	0.956 ± 0.08	0.949 ± 0.02
	0.8	0.959 ± 0.02	0.965 ± 0.01	0.948 ± 0.04	0.983 ± 0.03	0.965 ± 0.01
	1.0	0.959 ± 0.02	0.962 ± 0.01	0.960 ± 0.02	0.964 ± 0.02	0.964 ± 0.01
γ_1, γ_2	0.0,1.0	0.941 ± 0.03	0.939 ± 0.03	0.930 ± 0.05	0.949 ± 0.05	0.946 ± 0.03
	1.0,0.0	0.934 ± 0.01	0.924 ± 0.02	0.966 ± 0.03	0.883 ± 0.06	0.944 ± 0.01
	0.2,0.8	0.945 ± 0.02	0.944 ± 0.02	0.962 ± 0.03	0.927 ± 0.01	0.951 ± 0.01
	0.4,0.6	0.935 ± 0.03	0.935 ± 0.04	0.922 ± 0.03	0.949 ± 0.09	0.942 ± 0.02
	0.6,0.4	0.943 ± 0.03	0.949 ± 0.03	0.935 ± 0.05	0.964 ± 0.04	0.951 ± 0.03
	0.8,0.2	0.934 ± 0.03	0.927 ± 0.06	0.929 ± 0.03	0.926 ± 0.13	0.948 ± 0.02

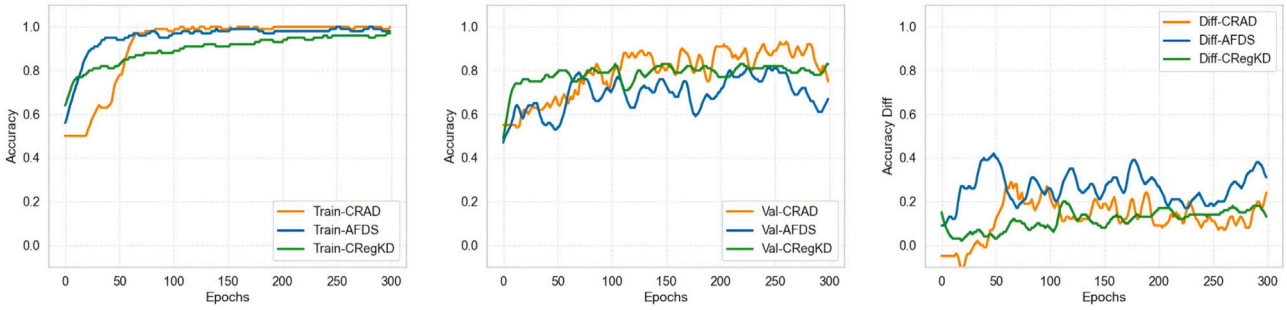

Fig. 6. Plots on the learning curves of different knowledge distillation models (left: train accuracy, middle: validation accuracy, right: differences of train and validation accuracy). The orange represents our model, the blue represents the attention distillation model (Wang et al., 2019), and the green represents the regularized distillation model (Yang et al., 2023). The difference curves between the training and validation accuracy are also displayed. Our models show faster convergence with less variation between the train and validation datasets.

Table 12
Evaluation of model complexity of different methods.

Methods	Params. (M)↓	Flops (G)↓	$T_{inf.}$ (ms) ↓
SM-BL (Korolev et al., 2017)	56.8	77.5	4.2
MM-BL (Han et al., 2019)	74.8	132.6	13.1
MDL-Net (Qiu et al., 2024)	2.8	27.4	8.9
TPA-GAN (Gao et al., 2021)	20.6	289.1	33.8
CReg-KD (Yang et al., 2023)	46.7	221.9	5.9
AFDS (Wang et al., 2019)	33.3	155.6	4.9
DFTD (Chen et al., 2023)	1.5	10.8	6.9
IC-MKD (Kwak et al., 2025)	8.3	8.7	3.9
CRAD	46.6	73.9	4.1

4.7. Influence of hyperparameter

To ensure the robustness and reproducibility of our proposed CRAD framework, we conducted a comprehensive sensitivity analysis on its key hyperparameters: the distillation loss coefficient λ and the loss weighting coefficients γ_1 and γ_2. The coefficient λ controls the relative importance of the knowledge distillation loss versus the task-specific classification loss. We tested λ across a wide range of values [0.2, 0.4, 0.5, 0.6, 0.8, 1.0]. As shown in Table 11, the model performance is robust across values from 0.5 to 1.0, with the optimal balance of accuracy (ACC), robustness (AUC), and specificity (SPE) achieved at $\lambda = 0.8$. This indicates that emphasizing knowledge transfer from the teacher is beneficial, but requires balancing with the student's own task learning. The weights γ_1 and γ_2 balance the contribution of feature-level refinement and disentanglement. Ablations show that using either loss alone ($\gamma_1 = 1.0, \gamma_2 = 0.0$ or $\gamma_1 = 0.0, \gamma_2 = 1.0$) leads to imbalanced performance, e.g., high SEN but low SPE, or vice versa. The optimal performance across most metrics was achieved with $\gamma_1 = 0.6$ and $\gamma_2 = 0.4$, confirming that both intermediate feature alignment and final output matching are essential for effective distillation. These results demonstrate that CRAD is stable under a reasonable range of hyperparameters, and our chosen values are well-justified to maximize generalization and performance.

4.8. Model performance evaluation

Our proposed model exhibits a trade-off in efficiency and complexity compared to competing methods, as detailed in Table 12. It achieves the lowest inference time of 4.2 ms per sample on one RTX 3090 GPU, owing to its optimized architecture that minimizes redundant computations. In addition, our model requires only 73.9G FLOPs and 46.6M parameters, fewer than most competing methods, underscoring its computational efficiency. These advantages make our model highly suitable for real-world applications with strict computational constraints.

While our method, CRAD, does not offer the lowest computational complexity, it is specifically designed to excel in generalization, robustness, and real-world applicability, which are critical in clinical settings and justify its complexity. For example, CRAD significantly reduces cross-dataset performance degradation (8.5% vs. 12.7% for MDL-Net (Qiu et al., 2024)), demonstrating stronger generalization capability. Moreover, unlike (Gao et al., 2021; Qiu et al., 2024; Han et al., 2019), which require both MRI and PET modalities during inference, CRAD's student model operates robustly using only MRI, improving practicality in environments where PET is scarce or unavailable. Thus, while CRAD introduces additional complexity, it offers essential advantages in terms of real-world usability, cross-domain stability, and resilience to missing data, making it a more suitable solution for clinical scenarios.

Fig. 6 shows the accuracies obtained for our proposed model (orange), the attention distillation model (Wang et al., 2019) (blue), and the regularized distillation model (Yang et al., 2023) (green), and all of them are averaged across all cross-validations. As Fig. 6 (left and middle) shows, the CRAD reaches a stable state after 100 epochs and achieves the highest accuracy. In addition, the accuracy gap between train and validation datasets suggests the stability of the model, as shown in Fig. 6 (right). Our model shows faster convergence and a smaller accuracy gap. This demonstrates its enhanced stability and efficiency during both training and validation.

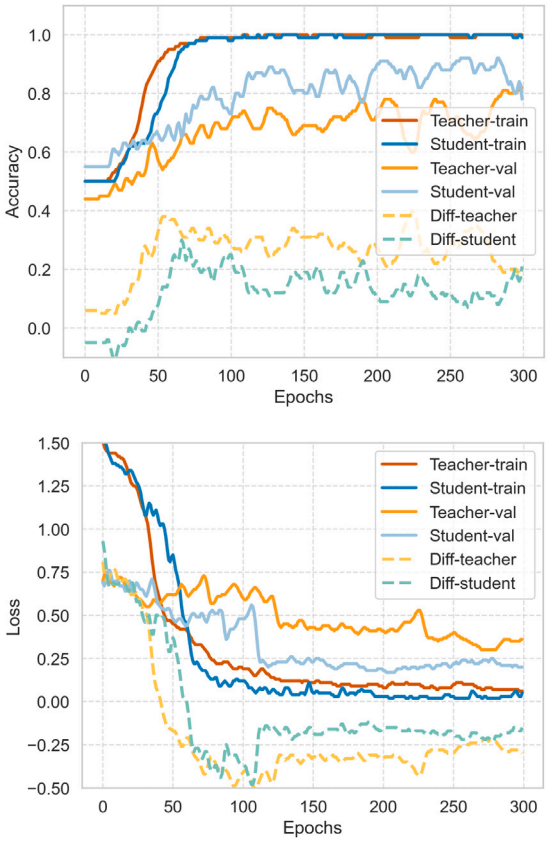

Fig. 7. Comparison of performance between teacher and student models. Top: Training accuracy, validation accuracy, and their difference curves highlight the superior performance of the student model in accuracy. Bottom: Training loss, validation loss, and their difference curves demonstrate the stability of the student model. Note that the difference curves of the student model are close to the zero axis, indicating better stability.

As illustrated in Fig. 7, the learning curves (solid lines) and difference curves (dashed lines) are plotted for the teacher and the student models of the proposed CRAD in terms of loss and accuracy. In the learning curves, the accuracy and loss of the teacher model are represented in orange, while those of the student model are depicted in blue. It can be observed that the student model converges faster than the teacher model, as evidenced by the training and validation accuracy curves after 100 epochs. This indicates that attention distillation facilitates faster learning of the student model from the teacher model. Furthermore, the smaller average accuracy discrepancies (student: 0.12 vs. teacher: 0.26) between training and validation data suggest the effectiveness of attention distillation in reducing overfitting, and the smaller average loss discrepancies (student: 0.06 vs. teacher: 0.19) between training and validation datasets demonstrate the enhanced stability of the student model during training.

4.9. Visualization and failure case analysis

We computed the mutual information between brain regions and diagnostic labels to identify brain regions critical for AD diagnosis. The spatial distribution of these regions is visualized in Fig. 8, while their original mutual information values with disease labels are detailed in Table 13. These brain regions are the left Hippocampus (lHIP), left Parahippocampal Gyrus (lPHG), right Hippocampus (rHIP), right Middle Occipital Gyrus (rMOG), left Inferior Temporal Gyrus (lITG), left Amygdala (lAMY), right Parahippocampal Gyrus (rPHG), right

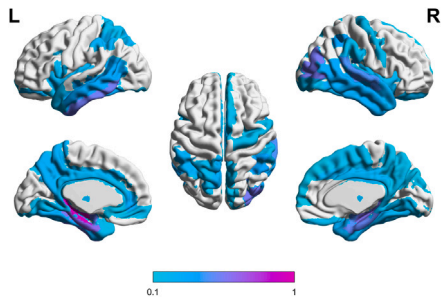

Fig. 8. Overview of the brain regions related to AD. Regions with purple and blue colors illustrate that they are more important for diagnosing AD. The visualization is drawn with the Brain-Net Viewer (Xia et al., 2013). Notably, the mutual information is normalized to 0–1.

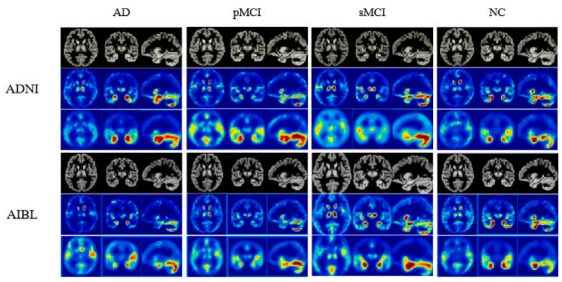

Fig. 9. Visualization of the feature maps, which are obtained by averaging all the data of the same groups. For each dataset, the first row is the original brain images, and the following rows represent shallow intermediate feature maps.

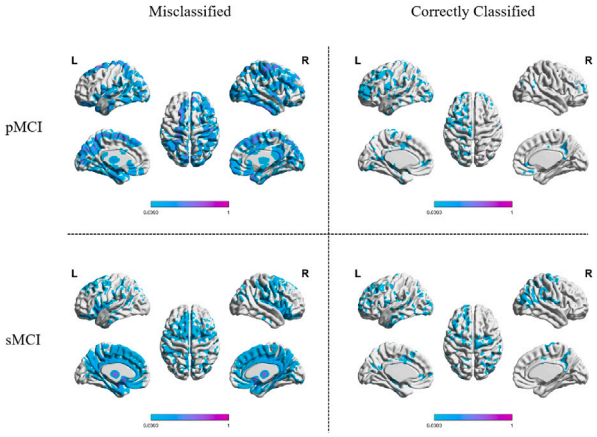

Fig. 10. Visualizations of case analysis. The left column shows cases of incorrect classification, and the right column shows cases of correct classification. The upper row is progressive MCI (pMCI), and the lower row is stable MCI (sMCI).

Amygdala (rAMY), right Superior Temporal Gyrus (rSTG), and right Inferior Parietal Gyrus (rIPG), encompassing the supramarginal and angular gyri, respectively. These regions are prominently associated with AD pathology, underscoring their diagnostic contribution.

As illustrated in Fig. 9, the visualized feature maps highlight regions of high importance, denoted by deeper color intensities. These regions exhibit strong correspondence with the critical brain areas identified in Fig. 8, further validating the disease-aware capability of our proposed method. Visualization analysis also reveals that most brain regions contribute minimally to AD diagnosis, indicating the need for the model

Table 13
Top 10 brain regions with the greatest mutual information with disease labels.

lHIP	lPHG	rHIP	rMOG	lITG	lAMY	rPHG	rAMY	rSTG	rIPG
0.1953	0.1203	0.1129	0.1079	0.1041	0.1012	0.0977	0.0901	0.0881	0.0740

to focus more intensively on extracting and leveraging discriminating features.

Despite the strong overall performance of CRAD, we observed that the majority of misclassifications occurred between pMCI and sMCI groups, a challenge in AD research due to the subtlety of early neurodegenerative changes (Petersen et al., 2018; Weber et al., 2021). These misclassifications may be due to atypical presentations or very early disease stages. To better understand these errors, we visualized feature distributions for misclassified samples. The cases are categorized into two groups (correctly classified and misclassified), and the average features captured by the proposed model are displayed (the features are scaled up to the original shape). As shown in Fig. 10, the up left (pMCI misclassified as sMCI) shows the model attended broadly to temporal and parietal regions, but failed to highlight the left Hippocampus strongly enough, the region for predicting progression, while the down left (sMCI misclassified as pMCI) indicates the model overemphasized the occipital region, which is less specific to AD pathology, while under-weighting atrophy in the Parahippocampal Gyrus. For correctly classified samples, the proposed model focuses on key subtle areas, such as the Middle Frontal Gyrus and Precuneus.

These visualizations suggest that while CRAD generally focuses on clinically relevant regions, it can sometimes be distracted by non-specific structural changes or fail to capture very subtle atrophy patterns. This may be due to the inherent heterogeneity within MCI subgroups. Future directions include integrating additional biomarkers and designing more sensitive feature extractors for early structural changes.

4.10. Discussion

This study proposes a Consistency Refinement-driven Multi-level Self-Attention Distillation framework to improve multimodal knowledge transfer for disease diagnosis and address the challenge of missing modalities in clinical practice. Compared with existing knowledge distillation methods, the CRAD has several unique advantages.

First, existing knowledge distillation methods (Yang et al., 2023; Van Sonsbeek et al., 2021; Song et al., 2023) ignore the ability to identify discriminating intermediate features. Our proposed CRAD can adaptively distill attentive features through a simple yet effective attention converter module, which can be viewed as a regularization term that refines knowledge. Moreover, we align multi-level features to distill cross-modal discriminating knowledge. Second, the existing knowledge refinement ignores the relative error between the output of the teacher and student models. We develop a consistency-driven confidence regularization that smooths distillation by introducing a dynamic regularization term ψ to balance knowledge transfer between the teacher and student models. We further incorporate this regularization within the hierarchical self-attention distillation process. Third, compared to existing feature disentanglement methods (Chen et al., 2023; Lu et al., 2020; Yang et al., 2022a; Wang et al., 2019, 2020a), the proposed orthogonal projection and parameter-free attention distillation are light designs without additional parameters, making them particularly suitable for resource-constrained real-world applications. Moreover, we focus more on transferring robust shallow knowledge, which contains subtle disease-related changes, than high-level knowledge.

Despite its strengths, the proposed method shows limitations. First, the teacher model can only be trained with paired multimodal data, which limits the amount of data it can learn. In addition, from the results of Table 3, the disparities among different domains contribute to suboptimal generalization performance. This indicates that it is essential to learn domain-invariant features. Thus, we will study a flexible multimodal distillation framework that transfers knowledge under random missing modalities. Meanwhile, we will bridge the domain gap and incorporate it within the distillation framework.

5. Conclusions

In this study, we introduced the Consistency Refinement-driven Multi-level Self-Attention Distillation framework, a novel approach designed to address the challenges of Alzheimer's disease (AD) diagnosis and Mild Cognitive Impairment (MCI) conversion prediction under incomplete multimodal data. The CRAD framework incorporates three key innovations: (1) the cross-modal attention distillation module (PF-CMAD), which leverages a parameter-free attention converter (PFAC) to distill attentive features efficiently; (2) the smooth distillation unit (SDU), which employs consistency-based confidence regularization to enhance the reliability and stability of the multi-level distillation process; and (3) cross-modal orthogonal projection (OP), which disentangles inter-modal features to reduce redundancy without introducing additional learnable parameters. Collectively, these components form a lightweight and efficient multimodal distillation framework that is highly adaptable to AD diagnosis under missing modalities.

Extensive experimental evaluations demonstrate that the proposed CRAD framework outperforms state-of-the-art methods in AD diagnosis-related tasks, achieving superior performance in handling incomplete multimodal data. Visualization experiments further validate the framework's ability to identify discriminating brain regions associated with AD, providing interpretable insights into its diagnostic capabilities. This study advances the field of multimodal knowledge distillation, providing a scalable and efficient solution for early and accurate diagnosis of AD, with potential applications in other medical imaging domains. In the future, we will explore cross-modal and domain-invariant distillation techniques to enhance the robustness of AD diagnosis. Additionally, we plan to fully leverage unpaired multimodal data in the distillation process, addressing a critical limitation in current multimodal learning paradigms. We will also validate CRAD in multi-center studies (e.g., more disease types) and optimize inference under resource-limited scenarios to support real-time diagnosis.

CRediT authorship contribution statement

Fei Liu: Writing – original draft, Visualization, Validation, Project administration, Methodology, Conceptualization. **Shiuan-Ni Liang:** Writing – review & editing, Supervision, Resources. **Mohamed Hisham Jaward:** Writing – review & editing, Supervision, Methodology. **Huey Fang Ong:** Writing – review & editing, Supervision. **Huabin Wang:** Writing – review & editing, Supervision, Resources, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, United States, the National Institute of Biomedical Imaging and Bioengineering, United States, and through generous contributions from the following: AbbVie,

Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare, United Kingdom; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Appendix

A.1. Training procedure

We present an algorithm block that summarizes the training procedure of the CRAD framework, as shown in Algorithm 1. The training process contains two main phases: teacher training and student distillation. The teacher training involves the CAFR module for cognitive score prediction and OP for feature disentanglement. The student phase uses knowledge distillation with PF-CMAD for attention-based feature alignment and SDU for confidence regularization.

Algorithm 1 Training Procedure of the Proposed CRAD framework

Input:
Multimodal dataset $X = X_i$, labels Y, cognitive scores Y_c
Teacher model $\mathcal{N}_T$, student model $\mathcal{N}_S$
Hyperparameters: $\gamma_1, \gamma_2, \lambda$
Output: Trained student model $\mathcal{N}_S$
Phase 1: Train Teacher Model $\mathcal{N}_T$
for each batch of paired MRI and PET data **do**
  Extract intermediate features $Z_T = \{Z_T^m, Z_T^p\}$
  Apply Cognitive Awareness Feature Refinement (CAFR):
    Predict MMSE score Y_c'
    Compute MSE loss: $\mathcal{L}_{MSE}$ Eq. (2)
  Apply Orthogonal Projection (OP)
    Disentangle MRI and PET features Z_T^m, Z_T^p
    Compute orthogonal loss: $\mathcal{L}_{ORT}$ Eq. (3)
  Predict Y_T' and compute classification loss: $\mathcal{L}_{CE}$ Eq. (4)
  Update $\mathcal{N}_T$ via backpropagation
end for
Phase 2: Distill Knowledge to Student Model $\mathcal{N}_S$
Freeze teacher model $\mathcal{N}_T$
for each batch of MRI-only data **do**
  Extract student features Z_S, predict Y_S'
  Extract teacher features Z_T, predict Y_T'
  Compute distillation losses Eq. (15)
  Apply Smooth Distillation Unit (SDU) Eq. (12)
  Update $\mathcal{N}_S$ via backpropagation
end for

A.2. Feature dimensions and alignment

To ensure clarity, we detail the dimensionality of feature representations at each stage of the proposed CRAD framework. The overall feature flow and dimensionality changes are summarized in Table A.1. The detailed description is presented as follows.

Intermediate Feature Extraction & Distillation: The feature encoder, based on a 3D CNN, extracts hierarchical representations from the input scans. The output of each convolutional block is a tensor (B, C, D, H, W), where C is the number of channels, and (D, H, W) are the spatial dimensions. To prepare these features for distillation, we reshape them into a 3D tensor (B, C, L), where $L = D \times H \times W$ is the number of spatial locations. For instance, the output of Layer 4 $(B, 512, 7, 8, 7)$ is reshaped to $(B, 512, 392)$. This standardized format allows for efficient computation of distillation losses between corresponding layers of the teacher and student models.

Projection, Fusion, and High-Level Alignment: In the teacher model, the modality-specific features Z_T^m and Z_T^p (both $(B, 512, 392)$) are processed by the Orthogonal Projection module. The fused feature Z_T^u is obtained by concatenating them along the channel dimension, resulting in a tensor of shape $(B, 1024, 392)$, which is then projected back to $(B, 512, 392)$ using a $1 \times 1 \times 1$convolution. The student model projects its MRI feature $Z_S^m(B, 512, 392)$ to the same latent space. The high-level feature distillation loss $\mathcal{L}_{KD}^U$ is computed directly between the teacher's Z_T^u and the student's projected feature, both of dimension $(B, 512, 392)$.

Final Prediction and Output Distillation: Following global average pooling (which reduces the features from $(B, 512, 392)$ to $(B, 512)$), the classifier predicts logits of shape (B, C), where C is the number of classes. The soft-label distillation loss $\mathcal{L}_{KD}^P$ aligns these output distributions between the teacher and student models.

A.3. Supplementary experiments

A.3.1. Ablation analysis of CAFR module

We have conducted ablation experiments to evaluate the effectiveness of the proposed CAFR module. As shown in Table A.2, the results demonstrate that CAFR not only enhances the intermediate features and overall performance of the teacher model but also, more importantly, leads to a stronger student model through improved knowledge distillation. The comparative results are summarized below. Specifically, we observe an increase in ACC from 0.881 to 0.934 and in AUC from 0.871 to 0.924. This confirms that CAFR effectively enhances the intermediate feature representations of the multimodal teacher, leading to a more powerful and robust model. When the student is trained with a CAFR-enhanced teacher, it achieves superior performance, particularly in ACC (0.923 vs. 0.918) and AUC (0.926 vs. 0.918), which are the primary indicators for classification tasks. This demonstrates that the improved feature quality of the teacher directly translates into more effective knowledge transfer, resulting in a better student model. These results validate that the CAFR module is integral to our framework. It successfully strengthens the teacher's feature representations, which in turn enables the distillation of a more accurate and reliable student model.

A.3.2. Ablation analysis of OP module

The OP loss is applied only during the teacher model's training phase when both MRI and PET modalities are fully available. It serves to enhance feature disentanglement and reduce redundancy between the two modalities, thereby improving the robustness and representational quality of the teacher model. Since the student model is trained to operate under missing-modality settings (e.g., MRI-only), it does not utilize the OP module directly. Instead, it benefits from the refined feature representations distilled from

Table A.1
Feature dimensions and knowledge distillation alignment. The teacher network has both MRI and PET branches (marked with "*"), while the student network has only the MRI branch. "B" indicates batch size.

Stage		Description	Input shape	Output shape	KD
Input data (X)		MRI scans PET scans	(B,1,113,137,113) (B,1,113,137,113) *		
Feature extractor	Layer 1	3D convolution	(B,1,113,137,113) (B,1,113,137,114) *	(B,64,57,69,57) (B,64,57,69,58) *	Attention KD
	Layer 2	3D convolution	(B,64,57,69,57) (B,64,57,69,57) *	(B,128,28,34,28) (B,128,28,34,28) *	Attention KD
	Layer 3	3D convolution	(B,128,28,34,28) (B,128,28,34,28) *	(B,256,14,17,14) (B,256,14,17,14) *	Attention KD
	Layer 4	3D convolution	(B,256,14,17,14) (B,256,14,17,14) *	(B,512,7,8,7) (B,512,7,8,7) *	Attention KD
Orthogonal projection (N_T)		Z_T^m Z_T^p	(B,512,7,8,7) (B,512,7,8,7) *	(B,512,392) (B,512,392) *	
Fusion (N_T)		Z_T^u	(B,1024,392)	(B,512,392)	
Projection (N_S)		Z_S^u	(B,512,392)	(B,512,392)	
Global average pool		Spatial pooling	(B,512,392)	(B,512)	Feature KD
Classifier		Fully-connected layer	(B,512)	(B,class_number)	Logit KD

Table A.2
Ablation analysis of the CAFR module for enhancing the teacher model and the distillation.

Model	CAFR	ACC	AUC	SEN	SPE	F1-score
Teacher	Yes	0.934 ± 0.01	0.924 ± 0.02	0.966 ± 0.03	0.883 ± 0.06	0.944 ± 0.01
	No	0.881 ± 0.02	0.871 ± 0.04	0.909 ± 0.03	0.832 ± 0.11	0.899 ± 0.02
Student	Yes	0.923 ± 0.02	0.926 ± 0.03	0.903 ± 0.04	0.949 ± 0.09	0.931 ± 0.01
	No	0.918 ± 0.02	0.918 ± 0.03	0.913 ± 0.04	0.923 ± 0.11	0.928 ± 0.01

Table A.3
Ablation analysis of the OP module for enhancing the teacher model and the distillation. For student models, the "baseline" indicates that it was only distilled from the teacher model trained without the OP module, while the "with OP" means it was distilled from the teacher model trained with the OP module.

Model	Ablation	ACC	AUC	SEN	SPE	F1-score
Teacher	Baseline	0.881 ± 0.02	0.871 ± 0.04	0.909 ± 0.03	0.832 ± 0.11	0.899 ± 0.02
	With OP	0.941 ± 0.03	0.939 ± 0.03	0.930 ± 0.05	0.949 ± 0.05	0.946 ± 0.03
Student	Baseline	0.889 ± 0.05	0.870 ± 0.06	0.942 ± 0.04	0.799 ± 0.15	0.908 ± 0.04
	With OP	0.902 ± 0.06	0.887 ± 0.07	0.931 ± 0.05	0.843 ± 0.15	0.919 ± 0.05

the OP-enhanced teacher. To explicitly demonstrate the contribution of the OP module, we conducted an ablation study comparing the performance of both the teacher and student models with and without the incorporation of OP during teacher training. The results are summarized in Table A.3. For student models, the "baseline" indicates that it was only distilled from the teacher model trained without the OP module, while the "with OP" means it was distilled from the teacher model trained with the OP module.

The results indicate that applying the OP module during teacher training yields significant improvements in the teacher's performance across all metrics. More importantly, this gain is effectively transferred to the student model during distillation, as evidenced by consistent performance improvements (e.g., ACC increased from 0.889 to 0.902) even when the student receives only MRI input. This confirms that the OP module plays an essential role in learning more robust and transferable multimodal representations in the teacher model, which in turn enhances the student's capability in missing-modality scenarios. Therefore, while the OP module is not applied during student inference, it is critical for strengthening the teacher's feature learning, which forms the foundation of an effective distillation process.

Data availability

Data will be made available on request.

References

Arbabshirani, M.R., Fornwalt, B.K., Mongelluzzo, G.J., Suever, J.D., Geise, B.D., Patel, A.A., Moore, G.J., 2018. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. NPJ Digit. Med. 1 (1), 9.

Bouts, M.J., van der Grond, J., Vernooij, M.W., Koini, M., Schouten, T.M., de Vos, F., Feis, R.A., Cremers, L.G., Lechner, A., Schmidt, R., et al., 2019. Detection of mild cognitive impairment in a community-dwelling population using quantitative, multiparametric MRI-based classification. Hum. Brain Mapp. 40 (9), 2711–2722.

Chen, Y., Pan, Y., Xia, Y., Yuan, Y., 2023. Disentangle first, then distill: A unified framework for missing modality imputation and Alzheimer's disease diagnosis. IEEE Trans. Med. Imaging 42 (12), 3566–3578.

Dao, D.-P., Yang, H.-J., Kim, J., Ho, N.-H., Initiative, A.D.N., et al., 2024. Longitudinal Alzheimer's disease progression prediction with modality uncertainty and optimization of information flow. IEEE J. Biomed. Health Informatics 29 (1), 259–272.

Deng, R., Cui, C., Remedios, L.W., Bao, S., Womick, R.M., Chiron, S., Li, J., Roland, J.T., Lau, K.S., Liu, Q., Wilson, K.T., Wang, Y., Coburn, L.A.,

Landman, B.A., Huo, Y., 2024. Cross-scale multi-instance learning for pathological image diagnosis. Med. Image Anal. 94, 103124.
Dubois, B., von Arnim, C.A., Burnie, N., Bozeat, S., Cummings, J., 2023. Biomarkers in Alzheimer's disease: role in early and differential diagnosis and recognition of atypical variants. Alzheimers Res. Ther. 15 (1), 175.
Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L., Group, B.D.C., 2011. Unbiased average age-appropriate atlases for pediatric studies. Neuroimage 54 (1), 313–327.
Gao, X., Shi, F., Shen, D., Liu, M., 2021. Task-induced pyramid and attention GAN for multimodal brain image imputation and classification in Alzheimer's disease. IEEE J. Biomed. Health Inf. 26 (1), 36–43.
Gao, X., Shi, F., Shen, D., Liu, M., 2023. Multimodal transformer network for incomplete image generation and diagnosis of Alzheimer's disease. Comput. Med. Imaging Graph. 110, 102303.
Gaser, C., Dahnke, R., Thompson, P.M., Kurth, F., Luders, E., the Alzheimer's Disease Neuroimaging Initiative, 2024. CAT: a computational anatomy toolbox for the analysis of structural MRI data. GigaScience 13, giae049.
Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D., 2015. Domain generalization for object recognition with multi-task autoencoders. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2551–2559.
Gou, J., Yu, B., Maybank, S.J., Tao, D., 2021. Knowledge distillation: A survey. Int. J. Comput. Vis. 129 (6), 1789–1819.
Guan, H., Wang, C., Tao, D., 2021. MRI-based Alzheimer's disease prediction via distilling the knowledge in multi-modal data. NeuroImage 244, 118586.
Han, K., Pan, H., Gao, R., Yu, J., Yang, B., 2019. Multimodal 3D convolutional neural networks for classification of brain disease using structural MR and FDG-PET images. In: Proc. ICPCSEE. Guilin, China, pp. 658–668.
Haque, A., Guo, M., Alahi, A., Yeung, S., Luo, Z., Rege, A., Jopling, J., Downing, L., Beninati, W., Singh, A., et al., 2017. Towards vision-based smart hospitals: a system for tracking and monitoring hand hygiene compliance. In: Machine Learning for Healthcare Conference. PMLR, pp. 75–87.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.. CVPR, Las Vegas, NV, USA, pp. 770–778.
Hu, Y., Huang, Y., Zhang, K., 2023. Multi-scale information distillation network for efficient image super-resolution. Knowl.-Based Syst. 275, 110718.
Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.. CVPR, Salt Lake City, UT, USA, pp. 7132–7141.
Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.. CVPR, Honolulu, HI, USA, pp. 2261–2269.
Körber, N., 2022. Parameter-free average attention improves convolutional neural network performance (almost) free of charge. arXiv preprint arXiv:2210.07828.
Korolev, S., Safiullin, A., Belyaev, M., Dodonova, Y., 2017. Residual and plain convolutional neural networks for 3D brain MRI classification. In: Proc. Int. Symp. Biomed. Imaging. ISBI, Melbourne, VIC, Australia, pp. 835–838.
Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22 (1), 79–86.
Kwak, M.G., Mao, L., Zheng, Z., Su, Y., Lure, F., Li, J., 2025. A cross-modal mutual knowledge distillation framework for Alzheimer's disease diagnosis: Addressing incomplete modalities. IEEE Trans. Autom. Sci. Eng. 22, 14218–14233.
Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., Feng, D., Fulham, M.J., 2015. Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Trans. Biomed. Eng. 62 (4), 1132–1140.
Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N., 2020. Cross-modality person re-identification with shared-specific feature transfer. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.. CVPR, Seattle, WA, USA, pp. 13379–13389.
Pan, Y., Liu, M., Xia, Y., Shen, D., 2022. Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data. IEEE Trans. Pattern Anal. Mach. Intell. 44 (10), 6839–6853.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. In: Proc. NeurIPS. Vol. 32, Vancouver, BC, Canada.
Petersen, R.C., Lopez, O., Armstrong, M.J., Getchius, T.S., Ganguli, M., Gloss, D., Gronseth, G.S., Marson, D., Pringsheim, T., Day, G.S., et al., 2018. Practice guideline update summary: Mild cognitive impairment. Neurology 90 (3), 126–135.
Qiu, S., Miller, M.I., Joshi, P.S., Lee, J.C., Xue, C., Ni, Y., Wang, Y., De Anda-Duran, I., Hwang, P.H., Cramer, J.A., et al., 2022. Multimodal deep learning for Alzheimer's disease dementia assessment. Nat. Commun. 13 (1), 3404.
Qiu, Z., Yang, P., Xiao, C., Wang, S., Xiao, X., Qin, J., Liu, C.-M., Wang, T., Lei, B., 2024. 3D multimodal fusion network with disease-induced joint learning for early Alzheimer's disease diagnosis. IEEE Trans. Med. Imaging 43 (9), 3161–3175.
Ranasinghe, K., Naseer, M., Hayat, M., Khan, S., Khan, F.S., 2021. Orthogonal projection loss. In: Proc. IEEE/CVF Int. Conf. Comput. Vis.. ICCV, pp. 12333–12343.
Rudroff, T., Rainio, O., Klén, R., 2024. AI for the prediction of early stages of Alzheimer's disease from neuroimaging biomarkers–A narrative review of a growing field. Neurol. Sci. 1–11.
Shi, Y., Yang, L., An, W., Zhen, X., Wang, L., 2023. Parameter-free channel attention for image classification and super-resolution. arXiv preprint arXiv:2303.11055.
Shi, J., Zheng, X., Li, Y., Zhang, Q., Ying, S., 2018. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. IEEE J. Biomed. Health Inf. 22 (1), 173–183.
Shi, Y., Zu, C., Hong, M., Zhou, L., Wang, L., Wu, X., Zhou, J., Zhang, D., Wang, Y., 2022. ASMFS: Adaptive-similarity-based multi-modality feature selection for classification of Alzheimer's disease. Pattern Recognit. 126, 108566.
Simonyan, K., Andrew, Z., 2015. Very deep convolutional networks for large-scale image recognition. In: Proc. Int. Conf. Learn. Represent.. ICLR, pp. 1–14.
Song, T., Cao, G., Xiong, X., Kang, G., 2023. SDATNet: Self-Distillation Adversarial Training Network for AD classification. In: Proc. IEEE Int. Conf. Bioinform. Biomed.. BIBM, Istanbul, Turkey, pp. 2671–2678.
Van Sonsbeek, T., Zhen, X., Worring, M., Shao, L., 2021. Variational knowledge distillation for disease classification in chest x-rays. In: Inform. Proc. Med. Imaging. IPMI, Denmark, pp. 334–345.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G., 2023a. Multi-modal learning with missing modality via shared-specific feature modelling. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.. CVPR, Vancouver, BC, Canada, pp. 15878–15887.
Wang, L., Dai, W., Jin, M., Ou, C., Li, X., 2023b. Fundus-enhanced disease-aware distillation model for retinal disease classification from OCT images. In: Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.. MICCAI, Vancouver, BC, Canada, pp. 639–648.
Wang, K., Gao, X., Zhao, Y., Li, X., Dou, D., Xu, C.-Z., 2019. Pay attention to features, transfer learn faster CNNs. In: Proc. Int. Conf. Learn. Represent.. ICLR, Addis Ababa, Ethiopia.
Wang, C., Piao, S., Huang, Z., Gao, Q., Zhang, J., Li, Y., Shan, H., 2024. Joint learning framework of cross-modal synthesis and diagnosis for Alzheimer's disease by mining underlying shared modality information. Med. Image Anal. 91, 103032.
Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M., 2020a. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. Adv. Neural Inf. Process. Syst. 33, 5776–5788.
Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q., 2020b. ECA-Net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11534–11542.
Weber, C.J., Carrillo, M.C., Jagust, W., Jack, Jr., C.R., Shaw, L.M., Trojanowski, J.Q., Saykin, A.J., Beckett, L.A., Sur, C., Rao, N.P., et al., 2021. The worldwide Alzheimer's disease neuroimaging initiative: ADNI-3 updates and global perspectives. Alzheimer's Dement.: Transl. Res. Clin. Interv. 7 (1), e12226.
Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 3–19.
Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W., 2025. Towards generalist foundation model for radiology by leveraging web-scale 2d&3D medical data. Nat. Commun. 16 (1), 7866.
Xia, M., Wang, J., He, Y., 2013. BrainNet Viewer: a network visualization tool for human brain connectomics. PloS One 8 (7), e68910.
Xu, L., Wu, H., He, C., Wang, J., Zhang, C., Nie, F., Chen, L., 2022. Multi-modal sequence learning for Alzheimer's disease progression prediction with incomplete variable-length longitudinal data. Med. Image Anal. 82, 102643.
Yang, Q., Guo, X., Chen, Z., Woo, P.Y.M., Yuan, Y., 2022a. D2-Net: Dual disentanglement network for brain tumor segmentation with missing modalities. IEEE Trans. Med. Imaging 41 (10), 2953–2964.
Yang, Y., Guo, X., Ye, C., Xiang, Y., Ma, T., 2023. CReg-KD: Model refinement via confidence regularized knowledge distillation for brain imaging. Med. Image Anal. 89, 102916.
Yang, Y., Xutao, G., Ye, C., Xiang, Y., Ma, T., 2022b. Regularizing Brain Age Prediction via Gated Knowledge Distillation. In: Proc. MIDL. Vol. 172, Zurich, Switzerland, pp. 1430–1443.
Yang, L., Zhang, R.-Y., Li, L., Xie, X., 2021. SimAM: A simple, parameter-free attention module for convolutional neural networks. In: Proc. ICML. Vol. 139, pp. 11863–11874.
Yiannopoulou, K.G., Papageorgiou, S.G., 2020. Current and future treatments in Alzheimer disease: An update. J. Cent. Nerv. Syst. Dis. 12, 1179573520907397.
Zhai, Z., Liang, J., Cheng, B., Zhao, L., Qian, J., 2024. Strengthening attention: knowledge distillation via cross-layer feature fusion for image classification. Int. J. Multimed. Inf. Retr. 13 (2), 23.
Zhang, Y., Wang, S., Xia, K., Jiang, Y., Qian, P., 2021. Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion. Inf. Fusion 66, 170–183.

Zhou, T., Liu, M., Fu, H., Wang, J., Shen, J., Shao, L., Shen, D., 2019a. Deep multi-modal latent representation learning for automated dementia diagnosis. In: Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.. MICCAI, Shenzhen, China, pp. 629–638.

Zhou, T., Liu, M., Thung, K.-H., Shen, D., 2019b. Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data. IEEE Trans. Med. Imaging 38 (10), 2411–2422.

Zhou, T., Thung, K.-H., Liu, M., Shi, F., Zhang, C., Shen, D., 2020. Multi-modal latent space inducing ensemble SVM classifier for early dementia diagnosis with neuroimaging data. Med. Image Anal. 60, 101630.