Information Fusion 126 (2026) 103579

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

Interpretable multi-view fusion network via multi-view dual alignment and private bias filtering for Alzheimer's disease analysis

Jinghao Xu [a], Chenxi Yuan [b], Yi Jing [a], Huifang Shang [c,*], Xiaoshuang Shi [a,*], Xiaofeng Zhu [a]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China
[b] Department of Informatics, Ying Wu College of Computing, New Jersey Institute of Technology, Newark, NJ, 07102, USA
[c] Department of Neurology, Laboratory of Neurodegenerative Disorders, National Clinical Research Center for Geriatrics, West China Hospital, Sichuan University, Chengdu, 610041, China

ARTICLE INFO

Dataset link: https://github.com/nollexu/AFFNet

Keywords:
Alzheimer's disease
Attention
Multi-view fusion
sMRI

ABSTRACT

Structural magnetic resonance imaging (sMRI) combined with multi-view learning has been preliminarily explored in Alzheimer's disease (AD) analysis. However, existing methods usually face two key limitations: (i) they fail to fully exploit the inherent consistency of multiple views to design constraints for alignment and feature normalization; (ii) they lack effective mechanisms to separate discriminative information from view-specific noise. Hence, the fused representation may fail to effectively preserve cross-view complementary information, or even contain redundant noise, which often limits the performance of the final model. To address these challenges, we propose an innovative Alignment-Filtering-Fusion Network (AFFNet), which consists of four collaborative modules. Specifically, the multi-view feature extraction module integrates 3D and 2D convolutional networks to capture spatial structural information and extract multi-view features. The multi-view dual alignment module fully exploits the inherent supervision in multi-view data by introducing dual constraints of semantic and attention alignment, ensuring the regularization of complementary multi-view information while enhancing cross-view consistency. The private bias filtering module employs cross-view contrastive loss, orthogonal decomposition, and semantic regularization to identify and separate view-specific noise unrelated to the classification task, improving feature discriminability and laying the foundation for subsequent fusion. Finally, the multi-view fusion and classification module performs mean fusion on the aligned and filtered multi-view features to achieve complementary information integration for AD classification. Extensive experiments on widely used ADNI and AIBL datasets demonstrate that AFFNet significantly outperforms existing methods in AD classification accuracy and model interpretability. *All data list and source codes are available at:* https://github.com/nollexu/AFFNet.

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative disease and the most common cause of dementia [1]. It usually starts unknowingly and gradually progresses to symptoms, such as memory loss and cognitive dysfunction. There is no cure for AD, so early and accurate diagnosis is essential to delay the progression of the disease. Structural magnetic resonance imaging (sMRI) can reveal subtle changes in brain structure, especially atrophy of the hippocampus, providing an important biomarker for AD diagnosis [2], and thus it is considered as an effective tool for early detection of AD.

In recent years, deep learning technology has made significant progress in image recognition and classification tasks, and many methods have been proposed for early AD prediction using the 3D sMRI data, but existing research still has some key limitations. Single-view methods (including 2D-based models [3–6] and 3D-based models [7–12]) either sacrifice spatial integrity or suffer from high computational costs and risk of overfitting. More importantly, they often ignore the inherent multi-view consistency and complementary properties of 3D sMRI, which are critical to alleviating model overfitting and improving diagnostic performance. For instance, although imaging angles differ, the physical location and tissue characteristics of lesions remain consistent across views. However, due to differences in imaging orientation, MRIs from different directions may lead to variations in visual features, such as lesion shape or boundary clarity, making certain anatomical structures or lesion characteristics more prominent in specific views. Multi-view methods [13–21] aim to integrate information from different perspectives to enhance model performance. Unfortunately, they often do not adequately utilize the cross-view consistency constraints,

* Corresponding authors.
E-mail addresses: hfshang2002@126.com (H. Shang), xsshi2013@gmail.com (X. Shi).

https://doi.org/10.1016/j.inffus.2025.103579
Received 10 February 2025; Received in revised form 16 May 2025; Accepted 24 July 2025
Available online 5 August 2025

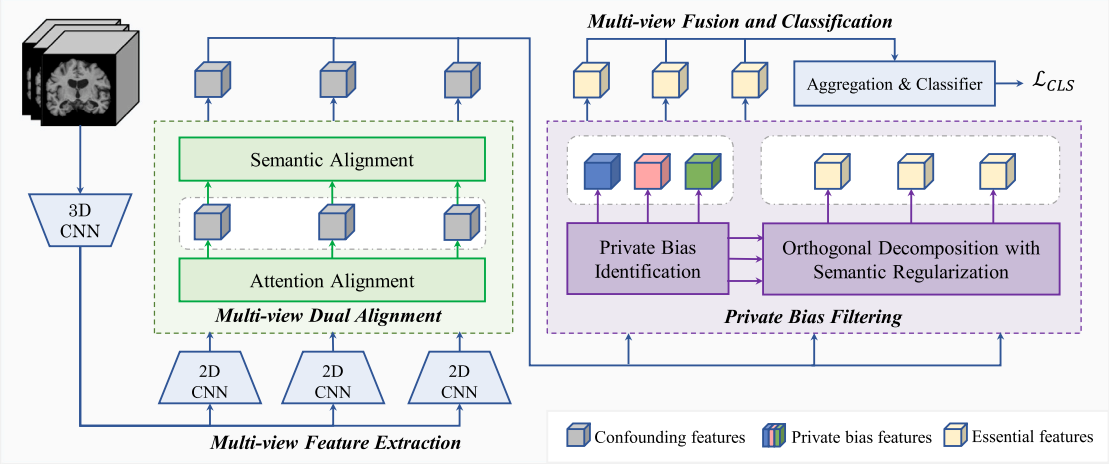

Fig. 1. The overall architecture of AFFNet.

which play a key role in aligning features and integrating complementary information effectively. Moreover, they lack effective mechanisms for separating discriminative information from view-specific noise. As a result, the fused representations often fail to fully preserve cross-view complementary information, may introduce redundant noise, or lead to overly smoothed features, ultimately limiting the performance of the final model.

To address aforementioned limitations, we propose an Align-Filter-Fusion Network (AFFNet), which aligns features and filters out private noise while effectively preserving and integrating cross-view complementary information. For clarity, we show the architecture of the proposed framework in Fig. 1. Here, confounding features refer to feature representations that contain both private bias information and essential features. Private bias features are view-specific and irrelevant to the classification task, while essential features are discriminative information retained after effective alignment and filtering. Specifically, AFFNet first extracts low-level structural features and builds multi-view representations using a 2D-3D hybrid backbone. Then, we design a dual alignment module to weakly align features from both the semantic and attention activation levels, thereby enhancing representation consistency and preserving complementary information. Next, to further disentangle view-specific noise, AFFNet introduces a private bias filtering module that identifies task-irrelevant features through cross-view contrastive loss, and removes them using orthogonal decomposition with semantic regularization, ultimately obtaining aligned and discriminative essential features. Finally, considering the semantic consistency of multi-view features, we average the essential features from all views in the aligned space for downstream classification. Extensive experiments on two benchmark datasets, ADNI and AIBL, demonstrate that AFFNet not only achieves state-of-the-art performance in AD diagnosis but also exhibits strong interpretability.

2. Main contributions

In summary, the major contributions of this paper are listed as follows:

- We propose a novel multi-view feature extraction and fusion framework for AD diagnosis, so as to fully leverage the multi-view features of 3D sMRI and meanwhile preserve the underlying spatial information.
- We design a multi-view dual alignment module, which combines semantic and attention alignments to simultaneously explore consistent multi-view features and normalize the complementary information of each view.
- We design a private bias filtering module based on an innovative cross-view contrastive loss and orthogonal decomposition, so as to separate view-specific and useless information and promote multi-view fusion.
- Extensive experiments demonstrate the superior classification and interpretation performance of the proposed framework over recent state-of-the-art AD diagnosis methods.

3. Related works

In recent years, many deep learning methods have been proposed to predict early AD through neuroimaging data. Based on the differences of feature processing strategies, these methods can be roughly divided into two categories: single-view and multi-view. We briefly review them in the following.

3.1. Single-view learning methods for AD diagnosis

Single-view learning methods can be further divided into 2D-based [3–6] and 3D-based methods [7–12]. 2D-based methods select slices from the original MRI as the input for model training according to some specific criteria. For instances, Kang et al. [5] train the model by sampling multiple selected slices at fixed intervals along the coronal plane of the sMRI as the model input. Hon and Khan [3] employ image entropy to select the most informative slices as input data, and combine transfer learning technologies for AD diagnosis.

3D-based methods typically utilize local patches at predefined locations or directly take the entire MRI image as the model input. For example, Liu et al. [7] and Zhu et al. [8] take patches at predefined locations as the input and combine multi-instance learning and attention mechanism to build AD diagnosis models. In another independent work, Han et al. [22] augments the data by aligning sMRI to multiple brain templates, and then applies the Siamese network to extract features for AD diagnosis. Recently, Xu et al. [11] localize stable discriminative regions from sMRI by using an innovative logits-constraint attention, and then they focus on exploring multi-scale features of these regions and employ multi-layer graph neural networks to optimize the feature representation to improve diagnostic performance.

2D-based methods are efficient but lose 3D spatial information, 3D-based ones make full use of spatial information but increase complexity and overfitting risk. Additionally, both of them neglect the inherent multi-view features of 3D MRI, which might improve the model robustness and performance. By contrast, our method combines the advantages of 3D and 2D feature extraction, fully explores and fuses the discriminative multi-view features contained in different slice directions using a novel multi-view learning framework.

3.2. Multi-view learning methods for AD diagnosis

Multi-view learning methods aim to integrate information from multiple perspectives to enhance model performance. In the context of AD diagnosis, existing multi-view approaches can be broadly categorized into 2D-based [13,14,19–21] and 2D–3D hybrid methods [15–18]. For example, Weng et al. [13] draw inspiration from the SlowFast network in video processing, and propose a 2D-based multi-view SlowFast network for AD diagnosis. Liu et al. [19] propose a multi-plane multi-scale feature-level fusion model (MPS-FFA), which combine clinical score information and introduce a feature similarity discriminator to improve the diagnostic performance. Zhang et al. [20] propose a multi-slice multi-view fusion network for AD diagnosis by combining lightweight convolution operations, so as to effectively enhance the prediction collaboration between views by introducing a label consistency mechanism. Chen et al. [16] utilize three independent 2D CNNs and one 3D CNN to extract features from different orientations of sMRI as well as from the overall 3D representation, concatenating these features for the final classification decision. Zhang et al. [18] design a similar architecture and introduce a slice-weighting module based on self-attention mechanisms, which can capture the contextual information between slices and generate attention weights to enhance representation learning. Jang and Hwang [17] integrate a 3D CNN with three pre-trained and frozen 2D sub-networks to extract multi-view representations, and subsequently feed the slice-level features from all views into a multi-layer Transformer [23] to further facilitate feature interaction and fusion.

Despite the achievements of existing methods, they often fail to fully exploit the consistency constraints across views, which are crucial for feature alignment and effective integration of complementary information. In addition, these methods lack an effective mechanism to distinguish discriminative information from view-specific noise, leading to the potential introduction of redundant noise or over-smoothing of features in the fused representation, ultimately affecting model performance. By contrast, we design a multi-view dual alignment module that introduces semantic and attentional consistency as additional constraints, enhancing the consistency of multi-view representations and the extraction of complementary information. Meanwhile, our method identifies and removes private bias information within each view before fusion, thereby significantly improving the quality of the fused representation and overall model performance.

4. Methodology

4.1. Overview of the proposed method

To effectively capture and fuse multi-view discriminative features from 3D sMRI for automatic AD diagnosis, we propose a deep multi-view learning framework, named AFFNet, as illustrated in Fig. 1. The proposed framework consists of four main stages:

- **Multi-view feature extraction**, which combines 3D and 2D convolutional layers to capture underlying spatial information and extract multi-view slice-level features;
- **Multi-view dual alignment**, which aggregates the multi-view slice-level features into multi-view confounding features, and enhances multi-view consistency at both the attention activation and feature semantic levels to ensure effective retention and regularization of complementary information;
- **Private bias filtering**, which identifies and separates view-specific biases that are uninformative for the classification task, thereby distilling the multi-view essential features;
- **Multi-view fusion and classification**, which integrates essential features for AD diagnosis.

For clarity, we describe each of these four stages in detail in the following subsections.

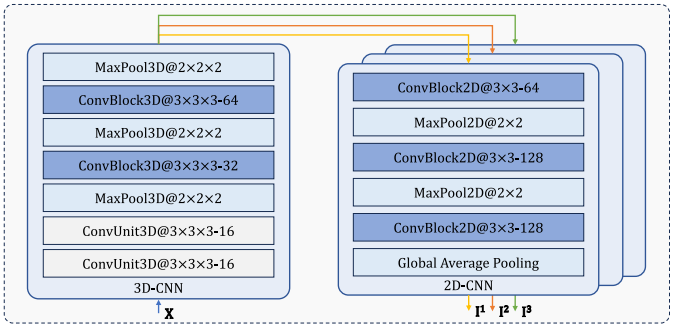

Fig. 2. The architecture of the multi-view feature extraction module. The kernel size (e.g., 3 × 3×3) and the number of output channels (e.g., 16) are represented as "3 × 3×3-16".

4.2. Multi-view feature extraction

Fig. 2 illustrates the overall architecture of our multi-view feature extraction module, which consists of a 3D CNN and three parallel 2D CNN branches with identical structures. In the figure, ConvUnitKD denotes a K-dimensional convolution (ConvKD) followed by a Batch Normalization layer and a ReLU activation function, while ConvBlockKD represents a BasicBlock [24], which takes $\mathbf{X}^l$ as the input and obtains the output:

$$\mathbf{X}^{l+1} = \mathrm{ReLU}(\mathrm{BN}(\mathrm{ConvKD}(\mathrm{ConvUnitKD}(\mathbf{X}^l))) + \mathbf{X}^l). \quad (1)$$

Note that we neglect the dimension matching operation of the skip connection in the Eq. (1) for simplicity.

In the 3D CNN part, the network includes three max-pooling layers and six convolutional layers (with each ConvBlock consisting of two convolutional layers). The numbers of output channels for these convolutional layers are 16, 16, 32, 32, 64, and 64, respectively. Each 2D CNN branch consists of two max-pooling layers, one global average pooling layer, and six convolutional layers, with output channels of 64, 64, 128, 128, 128, and 128, respectively. All convolution kernels have a size of 3, a stride of 1, and a padding of 1, thus not changing the spatial dimensions of the feature maps. All max-pooling layers use a kernel size of 2, stride of 2, and no padding, achieving a 2× spatial downsampling.

Given a batch of data $\mathbf{X} \in \mathbb{R}^{N\times 1\times H\times W\times D}$, where N is the batch size, H, W, and D represent the height, width and depth of each sMRI, respectively. We first obtain the 3D representation $\mathbf{I}_{3D} \in \mathbb{R}^{N\times 64\times \frac{H}{8}\times \frac{W}{8}\times \frac{D}{8}}$ by the first three 3D convolution modules. Then, we merge the batch dimension with an arbitrary spatial dimension to obtain the multi-view inputs $\mathbf{I}^1_{2D} \in \mathbb{R}^{(N\times \frac{H}{8})\times 64\times \frac{W}{8}\times \frac{D}{8}}$, $\mathbf{I}^2_{2D} \in \mathbb{R}^{(N\times \frac{W}{8})\times 64\times \frac{H}{8}\times \frac{D}{8}}$ and $\mathbf{I}^3_{2D} \in \mathbb{R}^{(N\times \frac{D}{8})\times 64\times \frac{H}{8}\times \frac{W}{8}}$. By utilizing three parallel 2D sub-networks and applying a global average pooling (GAP) operation on the spatial dimension, we obtain slice-level representations of multiple views $\mathbf{I}^1 \in \mathbb{R}^{N\times \frac{H}{8}\times 128}$, $\mathbf{I}^2 \in \mathbb{R}^{N\times \frac{W}{8}\times 128}$, and $\mathbf{I}^3 \in \mathbb{R}^{N\times \frac{D}{8}\times 128}$.

To simplify the introduction of subsequent modules, we rewrite the slice-level features of each view as $\{\mathbf{I}^m \in \mathbb{R}^{N\times S^m\times d}\}_{m=1}^{M}$, where S^m denotes the number of slices in the m-th view, d represents the number of feature dimensions, and M is the number of views.

4.3. Multi-view dual alignment

Multi-view alignment plays an important role in optimizing multi-view feature representation and promoting multi-view information fusion [25,26]. Existing alignment methods usually align different views through consistency properties, but in this process, specific information within the view is often eliminated, resulting in the loss of complementary information [27]. Thus, we develop a multi-view dual

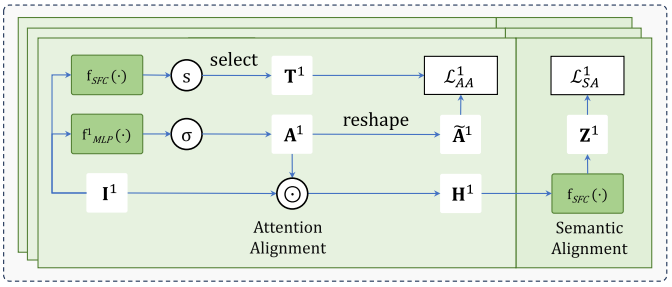

Fig. 3. Architecture of the multi-view dual alignment module.

alignment module that aims to promote multi-view feature consistency while preserving and normalizing complementary information. This module, as illustrated in Fig. 3, consists of two interdependent sub-modules: attention alignment and semantic alignment. It receives multi-view slice-level features $\{\mathbf{I}^m\}_{m=1}^M$ as input and outputs multi-view confounding features $\{\mathbf{H}^m \in \mathbb{R}^{N\times d}\}_{m=1}^M$.

The core objective of semantic alignment is to ensure semantic consistency across different views by unifying the decision space and encouraging similar decision behaviors among them. In contrast, attention alignment aims to activate consistent discriminative regions across views while suppressing irrelevant responses, thereby avoiding inconsistent or noisy attention patterns and enhancing the robustness of multi-view learning.

Specifically, given the multi-view slice-level features $\{\mathbf{I}^m\}_{m=1}^M$, we first feed them into the attention alignment module, in which the attention weights of each slice are calculated through M two-layer MLPs, and these weights are used to weight and aggregate the slice-level features to generate the multi-view confounding features $\{\mathbf{H}^m \in \mathbb{R}^{N\times d}\}_{m=1}^M$, i.e.,

$$\mathbf{A}^m = \sigma(\mathrm{f}_{MLP}^m(\mathbf{I}^m)), \tag{2}$$

$$\mathbf{H}^m = \frac{1}{S^m}\sum_{s=1}^{S^m}(\mathbf{I}^m \odot \mathbf{A}^m)_{:,s,:}, \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid activation function, $\mathrm{f}_{MLP}^m(\cdot)$ represents an independent two-layer MLP operation corresponding to the m-th view, $\mathbf{A}^m \in \mathbb{R}^{N\times S^m\times 1}$ denotes the slice-level attention of the m-th view, and $\odot$ represents the Hadamard product with broadcasting mechanism.

Based on $\{\mathbf{H}^m\}_{m=1}^M$, we introduce a semantic alignment constraint to ensure the semantic consistency between features from different views and regularize the feature representations of complementary information by unifying the decision space, it is:

$$\mathbf{Z}^m = \mathrm{f}_{SFC}(\mathbf{H}^m), \tag{4}$$

$$\mathcal{L}_{SA} = \sum_{m=1}^M \mathcal{L}_{SA}^m = -\frac{1}{N}\sum_{m=1}^M\sum_{n=1}^N\sum_{c=1}^C \mathbf{Y}_{n,c}\log(\mathrm{s}(\mathbf{Z}^m)_{n,c}), \tag{5}$$

where $\mathrm{f}_{SFC}(\cdot)$ denotes a globally shared fully connected layer, $\mathbf{Z}^m \in \mathbb{R}^{N\times C}$ is a matrix of logit vectors corresponding to $\mathbf{H}^m$, C is the total number of categories, $\mathrm{s}(\cdot)$ is the softmax function, and $\sum_{c=1}^C s(\mathbf{Z}^m)_{n,c} = 1$. $\mathbf{Y}_{n,c} \in \{0,1\}$ indicates whether the n-th sample belongs to class c, it is 1 if the sample belongs to the class, and 0 otherwise.

To further guide the attention distribution, we introduce an attention alignment loss that encourages consistent activation of semantically relevant slices while suppressing trivial regions:

$$\mathbf{T}^m = \mathrm{s}(\mathrm{f}_{SFC}(\mathbf{I}^m))_{:,:,\hat{y}_n}, \tag{6}$$

$$\mathcal{L}_{AA} = \sum_{m=1}^M \mathcal{L}_{AA}^m = \frac{1}{N}\sum_{m=1}^M\sum_{n=1}^N\sum_{s=1}^{S^m}\mathbf{T}_{n,s}^m \log\left(\frac{\mathbf{T}_{n,s}^m}{\tilde{\mathbf{A}}_{n,s}^m}\right), \tag{7}$$

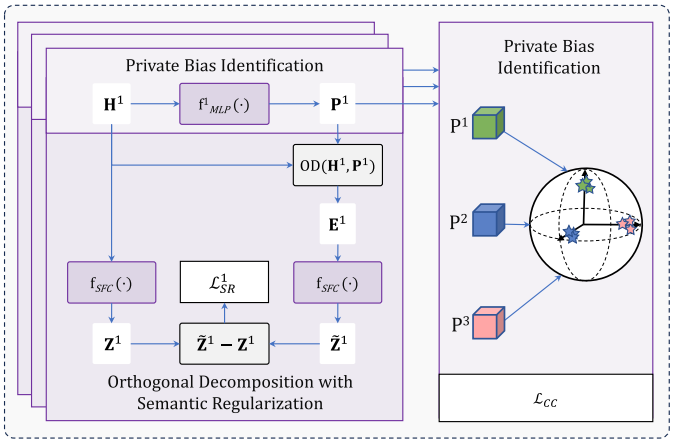

Fig. 4. Details of the private bias filtering module.

where $\hat{y}_n$ represents the index corresponding to the true class label of the n-th sample, $\mathbf{T}^m \in \mathbb{R}^{N\times S^m}$ represents the probability of all slices in the m-th view belonging to the true category of the sample, and $\tilde{\mathbf{A}}^m \in \mathbb{R}^{N\times S^m}$ is the result of reshaping $\mathbf{A}^m$.

Unlike traditional methods that achieve alignment by enforcing numerical similarity of cross-view features, our proposed dual-alignment module facilitates collaboration across views by promoting the alignment of desirable properties, such as semantic consistency and discriminative consistency. Under a unified decision space, the model can effectively organize and constrain complementary information from different views to better serve the classification task. Meanwhile, because dual alignment imposes only weak constraints, the complementary information is preserved rather than eliminated, laying a solid foundation for subsequent multi-view fusion.

4.4. Private bias filtering

Although multi-view alignment can effectively regularize feature representation, there might still contain useless private information inside a single view, which might be harmful for classification tasks and hinder the effectiveness of multi-view information fusion [28,29]. To this end, we construct a private bias filtering module to reduce the interference of irrelevant information, so as to optimize the feature representation of each view, as shown in Fig. 4. The module takes multi-view confounding features $\{\mathbf{H}^m\}_{m=1}^M$ as input and outputs multi-view essential features $\{\mathbf{E}^m \in \mathbb{R}^{N\times d}\}_{m=1}^M$.

First, we utilize M independent two-layer MLPs to extract the multi-view private bias information $\{\mathbf{P}^m \in \mathbb{R}^{N\times d}\}_{m=1}^M$ based on $\{\mathbf{H}^m\}_{m=1}^M$, i.e.,

$$\mathbf{P}^m = \mathrm{f}_{MLP}^m(\mathbf{H}^m). \tag{8}$$

Private bias refers to information that is specific to a particular view but irrelevant to the classification task. In other words, the private bias within the same view should be consistent, while private biases across different views should be uncorrelated. To model this, we introduce a cross-view contrastive loss, which encourages intra-view consistency and inter-view discrepancy in the private representations, thereby facilitating the identification of view-specific noise, i.e.,

$$\mathrm{d}(\mathbf{P}_i^m, \mathbf{P}_j^m) = \frac{\left\langle \mathbf{P}_i^m, \mathbf{P}_j^m \right\rangle}{\left\|\mathbf{P}_i^m\right\|\left\|\mathbf{P}_j^m\right\|}, \tag{9}$$

$$\mathcal{L}_{CC} = -\frac{1}{N}\sum_{n_1=1}^N\sum_{m_1=1}^M \log\frac{\sum_{n_2=1}^N e^{\mathrm{d}(\mathbf{P}_{n_1}^{m_1}\cdot\mathbf{P}_{n_2}^{m_1})}}{\sum_{n_3=1}^N\sum_{m_2=1}^M e^{|\mathrm{d}(\mathbf{P}_{n_1}^{m_1}\cdot\mathbf{P}_{n_3}^{m_2})|}}, \tag{10}$$

where $\langle\cdot,\cdot\rangle$ is dot product operator.

Since private bias is irrelevant to the classification task, it should be orthogonal to the essential (task-relevant) representation. To achieve this, we adopt an orthogonal decomposition strategy, projecting the confounding features $\{\mathbf{H}^m\}_{m=1}^M$ onto the subspace orthogonal to the private bias $\{\mathbf{P}^m\}_{m=1}^M$, thereby obtaining the essential features $\{\mathbf{E}^m \in \mathbb{R}^{N\times d}\}_{m=1}^M$,as follows:

$$\mathbf{E}^m = \mathrm{OD}(\mathbf{H}^m, \mathbf{P}^m) = \mathbf{H}^m - \frac{\mathbf{H}^m \cdot \mathbf{P}^{m\top}}{\mathbf{P}^m \cdot \mathbf{P}^{m\top}}\mathbf{P}^m, \tag{11}$$

where $\mathrm{OD}(\cdot,\cdot)$ denotes the orthogonal decomposition operation.

The essential features are expected to achieve comparable or superior classification performance to $\{\mathbf{H}^m\}_{m=1}^M$. To preserve task-discriminative semantics after decomposition, we introduce a semantic regularization loss during training. Specifically, we reuse the previous $\mathrm{f}_{SFC}(\cdot)$ and impose a classification loss on the logits difference between $\{\mathbf{E}^m\}_{m=1}^M$ and $\{\mathbf{H}^m\}_{m=1}^M$, formulated as:

$$\tilde{\mathbf{Z}}^m = \mathrm{f}_{SFC}(\mathbf{E}^m), \tag{12}$$

$$\mathcal{L}_{SR} = \sum_{m=1}^{M}\mathcal{L}_{SR}^m = -\frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{c=1}^{C}\mathbf{Y}_{n,c}\log(\mathrm{s}(\tilde{\mathbf{Z}}^{\mathbf{m}} - \mathbf{Z}^{\mathbf{m}})_{n,c}), \tag{13}$$

where $\tilde{\mathbf{Z}}^m \in \mathbb{R}^{N\times C}$ is a matrix of logit vectors corresponding to $\mathbf{E}^m$.

4.5. Multi-view fusion and optimization objective

Given that the essential features from different views $\{\mathbf{E}^m\}_{m=1}^M$ have been effectively aligned (i.e., exhibiting consistent semantic categories in a unified decision space) and the private biases removed, and thus, we adopt mean fusion—a simple and natural aggregation strategy—to retain complementary information across views while avoiding potential overfitting risks introduced by concatenation or more complex fusion methods. As the final step, we introduce a new fully connected layer $\mathrm{f}_{FC}(\cdot)$ to achieve the classification:

$$\mathcal{L}_{CLS} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}\mathbf{Y}_{n,c}\log(\mathrm{s}(\mathrm{f}_{FC}(\frac{1}{M}\sum_{m=1}^{M}\mathbf{E}^m))_{n,c}). \tag{14}$$

The overall optimization objective of the proposed framework is:

$$\mathcal{L} = \mathcal{L}_{CLS} + \mathcal{L}_{SA} + \mathcal{L}_{AA} + \alpha\mathcal{L}_{CC} + \beta\mathcal{L}_{SR}, \tag{15}$$

where α and β are two hyperparameters to balance the cross-view contrastive loss and semantic constraints. For clarity, we present the training procedure of our algorithm in Appendix A.

5. Experiments

5.1. Datasets

We evaluate the proposed framework on two widely-used datasets: the Alzheimer's Disease Neuroimaging Initiative[1](ADNI) and the Australian Imaging, Aging Biomarkers and Lifestyle Flagship Study[2] (AIBL). The **ADNI** dataset spans three phases, namely ADNI1, ADNI2, and ADNI3. We strictly adhere to the official ADNIMERGE.csv[3] to obtain baseline scans for all non-overlapping subjects. For MCI subjects, we include only those with at least 36 months of longitudinal follow-up data. Due to limitations in the preprocessing tools, we manually inspect and remove images that exhibit issues during preprocessing to ensure high-quality data integrity. In total, the final ADNI dataset used in our experiments is composed of 1,828 baseline sMRI scans. According to standard clinical diagnostic criteria, these subjects are categorized into three groups: Alzheimer's disease (AD), mild cognitive impairment (MCI), and normal control (NC). Specifically, the dataset includes 396 AD subjects, 584 MCI subjects, and 848 NC subjects. Furthermore, MCI participants are subdivided based on whether they progress to AD within 36 months after baseline assessment, resulting in 323 stable MCI (sMCI) and 261 progressive MCI (pMCI) participants.

For the **AIBL** dataset, we focus on AD and NC subjects. Following a similar approach to ADNI, we select all baseline data according to official guidelines and manually remove problematic samples. Then, we obtain a final dataset of 531 independent subjects' baseline sMRI scans, including 75 AD subjects and 456 NC subjects.

The demographic information of the subjects in the two datasets is summarized in Table 1. To evaluate the overall differences in continuous and categorical variables between different categories, we follow the method in the literature [30] and use one-way ANOVA and χ^2 test for statistical analysis, respectively, and summarize the obtained p-values in Table 1. Considering the significant differences in age and sex among subjects of different categories in the ADNI dataset, we employed 2D t-SNE to assess whether these factors introduced noticeable confounding effects in the imaging data. The results are provided in Appendix B.

5.2. Image preprocessing

Our sMRI preprocessing pipeline follows a series of normalization steps. First, we utilize SimpleITK library [31] to convert DCM format files to NIFTI format. Then, we apply the fslreorient2std tool to normalize the image orientation and ensure consistency with the standard template. Next, we adopt the robustfov tool to crop the sMRI, which effectively removes the non-target regions in the neck and lower head, laying a foundation for subsequent registration and skull dissection. Subsequently, we leverage the FLIRT tool [32] to register all sMRI scans to the Colin27 template [33] using rigid-body and affine transformations. This step aims to eliminate inter-subject global linear differences and ensure spatial alignment of anatomical structures, enabling the model to focus more on disease-related features instead of anatomical variability across individuals. Additionally, this process standardizes the image resolution (i.e., $1\times1\times1\mathrm{mm}^3$) and size (i.e., $191\times217\times191$), providing a consistent input for subsequent analyses. Finally, we apply the BET tool [34] for skull stripping, removing non-brain tissues, ensuring that the analysis is focused solely on brain regions and preventing irrelevant structures from introducing noise, which could negatively affect classification performance. The used tools, including fslreorient2std, robustfov, FLIRT and BET, are integrated in the FSL software package [35].

5.3. Comparison methods

We compare the proposed AFFNet with ten popular methods under the same experimental settings. The comparison methods are briefly introduced as follows:

- **VoxCNN** [36]: which extends VGGNet to a 3D version for AD diagnosis.
- **VoxResNet** [36]: which extends ResNet to a 3D version for AD diagnosis.
- **AttResNet** [37]: which implements a sparse attention through 3D convolution with ReLU function, and embeds it into 3D ResNet to provide interpretable analysis.
- **M3T** [17]:which integrates a 3D CNN with three pre-trained and frozen 2D sub-networks to extract multi-view representations, and uses a multi-layer Transformer to perform feature interaction and fusion.
- **M^{2}FAN** [9]: which proposes a multi-task multi-level feature adversarial network for AD diagnosis.
- **AMSNet** [10]: which designs a multi-scale fusion module based on dilated convolution and integrates it into the network to extract multi-scale information from sMRI.

[1] http://adni.loni.usc.edu
[2] https://aibl.csiro.au
[3] https://ida.loni.usc.edu/explore/jsp/search/search.jsp?project=ADNI#studyFiles

Table 1
Baseline demographic information of subjects included in two sets (i.e., ADNI and AIBL).

Dataset	Group Type & *p*-values	Sex (Male/Female)	Age (Mean±SD)	Education in years (Mean±SD)	MMSE (Mean±SD)
ADNI	AD	218/178	74.86 ±7.85	15.23 ±2.91	23.16 ±2.18
	pMCI	150/111	73.90 ±7.09	15.89 ±2.77	26.72 ±1.82
	sMCI	202/121	72.56 ±7.64	16.09 ±2.78	28.02 ±1.59
	NC	365/483	72.27 ±6.65	16.52 ±2.51	29.07 ±1.16
	p-value	6.700e−10	1.098e−08	2.107e−13	<1.0e−200
AIBL	AD	29/46	73.56 ± 7.52	–	20.31 ± 5.47
	NC	191/265	72.36 ± 6.12	–	28.71 ± 1.22
	p-value	6.906e−01	1.297e−01	–	<1.0e−200

- **AMSF** [18]: which employs three 2D CNNs and one 3D CNN to extract features from different orientations of sMRI images and their overall 3D representation, with the features concatenated for final classification.
- **MPS-FFA** [19]: which proposes a multi-plane multi-scale feature fusion model for AD diagnosis. It incorporates clinical score information and introduces a feature similarity discriminator to enhance diagnostic performance. To ensure a fair comparison, we remove the branch related to clinical scores, focusing solely on image-based feature extraction and fusion.
- **MMFNet** [20]: which proposes a multi-slice and multi-view fusion lightweight network for AD diagnosis, and introduces a label consistency constraint to promote prediction consistency across different views.
- **LA-GMF** [11]: which builds a two-branch network for AD diagnosis, with one branch mining the significant regions and the other one extracting multi-scale features.

5.4. Experimental setting

We conduct all our experiments using the PyTorch framework on a single GPU (i.e., NVIDIA GeForce 3090 24 GB). Our model is trained for a total of 100 epochs with a batch size of 4. During training, we adopt the Adam optimizer [38] and set the learning rate to 0.0001. To reduce the effect of background, we crop all experimental sMRIs to 160 × 192 × 160. Additionally, we augment the training images through translation and mirroring operations to increase the diversity of training data. Specifically, the translation operation randomly selects one of the six directions (up, down, front, back, left, right) to shift a voxel with equal probability, while the mirror operation transforms the left and right brain of the sMRI symmetrically with a probability of 0.5.

For the ADNI dataset, we design three tasks to evaluate the model's performance: ADNI-2CLS (AD-NC), ADNI-3CLS (AD-MCI-NC), and ADNI-4CLS (AD-pMCI-sMCI-NC), with increasing levels of difficulty. ADNI-3CLS builds upon the binary classification task by adding the distinction between MCI subtypes. ADNI-4CLS further increases complexity by differentiating between progressive MCI (pMCI) and stable MCI (sMCI), which is of significant clinical importance for early intervention. All experiments adopt a 5-fold cross-validation strategy, i.e., the dataset is randomly divided into 5 subsets, of which 4 subsets (80% of the subjects in the dataset) are used for model training, the remaining subset is used for testing, and the average of the five experiments is used as the final result. For the AIBL dataset, we adopt it as an independent test set to evaluate the model generalization ability on the popular AD-NC binary classification task (i.e., AIBL-2CLS).

We employ six key evaluation metrics—Area Under the Curve (AUC), F_1-score, Accuracy, Sensitivity, Specificity, and Precision—to comprehensively assess the model's performance during the testing phase. Furthermore, we perform statistical analysis by reporting 95% confidence intervals to ensure the reliability of the evaluation.

5.5. Classification results

Table 2 presents the classification performance of eleven methods on ADNI and AIBL datasets for the AD-NC binary classification task. As we can see, the proposed AFFNet achieves the best performance among all methods. Specifically, AFFNet improves the three major metrics AUC, F_1-score and Accuracy on the ADNI dataset by 1.3%, 2.4% and 1.4%, respectively, compared to the best competitors. On the AIBL dataset, AFFNet outperforms the best competitors by 0.8%, 2.0%, and 0.5% in terms of AUC, F_1-score, and Accuracy, respectively.

To further illustrate the strength of the proposed AFFNet, we show the classification results of the nine methods on the ADNI-3CLS task and the ADNI-4CLS tasks in Table 3. M^2FAN and MMFNet are excluded because they are specifically designed for binary tasks. As shown in Table 3, AFFNet also achieves the best results on these two more challenging tasks. Specifically, for the ADNI-3CLS task, the gain of AFFNet is 2.1%, 4.1%, and 4.5% over the best competitors in terms of AUC, F_1-score and Accuracy, respectively; For the ADNI-4CLS task, the gain of AFFNet is 2.6%, 4.2%, and 3.3% over the best competitors in terms of AUC, F_1-score and Accuracy, respectively.

Tables 2–3 illustrate the superior performance of the proposed framework, probably because: (i) Our framework fuses 3D and 2D convolutions, which enables it to effectively integrate spatial information while extracting multi-view features; (ii) By introducing multi-view dual alignment and private bias filtering, the model cannot only enhance the consistency of multiple views while retaining the complementary information of each view, but also effectively identify and exclude view-specific bias information, thereby further promoting the effective fusion of multi-view information.

It is worth noting that the two 2D-3D hybrid multi-view methods M3T and AMSF do not perform as expected in the experiments. We hypothesize that the main reason is that they only use the basic cross-entropy loss to guide the end-to-end training process, which fails to fully utilize the inherent constraints within the multi-view to align the multi-view representations. Additionally, M3T utilizes multiple layers of Transformer to facilitate interactions between multi-view features, but in scenarios with limited dataset size, this approach might lead to excessive feature smoothing. Meanwhile, AMSF directly concatenates the multi-view feature vectors with the global feature vector representing the entire 3D sMRI, which might lead to the curse of dimensionality and feature redundancy.

5.6. Ablation study

In this subsection, we conduct ablation experiments on AFFNet based on the AD-NC binary classification task to verify the effectiveness of various modules within AFFNet.

Ablation of Key Modules. We present the results of key modules ablation in Table 4. The baseline method "B" means using a single classification loss to guide multi-view feature extraction and fusion. MDA stands for multi-view dual alignment modules, including semantic alignment (SA) and attention alignment (AA). PBF denotes the private bias filtering module, which contains a cross-view contrastive loss (CC) and a semantic regularization loss (SR). As we can see from Table 4,

Table 2
The results of different methods for AD-NC binary classification using five-fold cross-validation.

Methods	**ADNI-2CLS**					
	AUC (95% CI)	F_1-score (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)
VoxCNN	0.944 (0.927, 0.960)	0.857 (0.832, 0.882)	0.912 (0.896, 0.929)	0.823 (0.795, 0.852)	0.954 (0.934, 0.974)	0.894 (0.853, 0.935)
VoxResNet	0.936 (0.919, 0.954)	0.853 (0.835, 0.871)	0.910 (0.897, 0.923)	0.818 (0.794, 0.843)	0.953 (0.928, 0.978)	0.892 (0.843, 0.942)
AttResNet	0.954 (0.940, 0.969)	0.873 (0.850, 0.896)	0.921 (0.904, 0.938)	0.848 (0.796, 0.901)	0.955 (0.916, 0.994)	0.904 (0.829, 0.978)
M3T	0.915 (0.897, 0.932)	0.810 (0.789, 0.832)	0.882 (0.867, 0.896)	0.793 (0.771, 0.815)	0.923 (0.903, 0.943)	0.829 (0.791, 0.867)
M^2FAN	0.953 (0.945, 0.960)	0.884 (0.864, 0.904)	0.929 (0.918, 0.941)	0.848 (0.819, 0.878)	**0.967 (0.954, 0.980)**	**0.924 (0.895, 0.952)**
AMSNet	0.958 (0.945, 0.960)	0.888 (0.864, 0.904)	0.930 (0.918, 0.941)	0.869 (0.819, 0.878)	0.959 (0.954, 0.980)	0.908 (0.895, 0.952)
AMSF	0.931 (0.920, 0.941)	0.841 (0.809, 0.874)	0.901 (0.885, 0.917)	0.828 (0.752, 0.904)	0.935 (0.911, 0.959)	0.858 (0.821, 0.895)
MPS-FFA	0.926 (0.912, 0.940)	0.840 (0.814, 0.866)	0.901 (0.887, 0.915)	0.816 (0.767, 0.864)	0.941 (0.920, 0.962)	0.867 (0.829, 0.905)
MMFNet	0.955 (0.939, 0.971)	0.877 (0.861, 0.894)	0.924 (0.914, 0.934)	0.859 (0.819, 0.898)	0.954 (0.934, 0.974)	0.899 (0.860, 0.937)
LA-GMF	0.953 (0.937, 0.969)	0.879 (0.840, 0.919)	0.926 (0.903, 0.949)	0.851 (0.778, 0.924)	0.961 (0.936, 0.986)	0.913 (0.865, 0.961)
AFFNet	**0.971 (0.959, 0.982)**	**0.912 (0.888, 0.936)**	**0.944 (0.928, 0.960)**	**0.912 (0.893, 0.930)**	0.960 (0.936, 0.982)	0.913 (0.868, 0.958)
Methods	**AIBL-2CLS**					
	AUC (95% CI)	F_1-score (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)
VoxCNN	0.930 (0.916, 0.944)	0.710 (0.686, 0.734)	0.905 (0.885, 0.924)	0.821 (0.744, 0.899)	0.918 (0.884, 0.953)	0.632 (0.554, 0.711)
VoxResNet	0.922 (0.899, 0.944)	0.654 (0.570, 0.739)	0.883 (0.835, 0.931)	0.763 (0.689, 0.836)	0.903 (0.844, 0.962)	0.581 (0.459, 0.704)
AttResNet	0.930 (0.914, 0.945)	0.702 (0.652, 0.751)	0.900 (0.867, 0.933)	0.816 (0.716, 0.916)	0.914 (0.861, 0.967)	0.628 (0.511, 0.745)
M3T	0.898 (0.871, 0.925)	0.602 (0.503, 0.701)	0.847 (0.778, 0.915)	0.784 (0.692, 0.876)	0.857 (0.766, 0.948)	0.506 (0.339, 0.673)
M^2FAN	0.926 (0.907, 0.945)	0.763 (0.708, 0.817)	0.927 (0.907, 0.946)	0.829 (0.772, 0.887)	0.943 (0.920, 0.965)	0.709 (0.617, 0.801)
AMSNet	0.948 (0.933, 0.963)	0.773 (0.736, 0.810)	0.930 (0.919, 0.941)	0.851 (0.791, 0.911)	0.943 (0.932, 0.953)	0.710 (0.670, 0.749)
AMSF	0.911 (0.924, 0.949)	0.647 (0.615, 0.680)	0.880 (0.846, 0.913)	0.773 (0.658, 0.889)	0.897 (0.840, 0.954)	0.572 (0.454, 0.690)
MPS-FFA	0.916 (0.907, 0.924)	0.663 (0.621, 0.704)	0.881 (0.856, 0.906)	0.821 (0.777, 0.866)	0.891 (0.857, 0.925)	0.558 (0.492, 0.625)
MMFNet	0.934 (0.923, 0.945)	0.710 (0.663, 0.757)	0.903 (0.879, 0.928)	0.829 (0.805, 0.854)	0.915 (0.883, 0.948)	0.624 (0.538, 0.711)
LA-GMF	0.937 (0.924, 0.949)	0.741 (0.711, 0.770)	0.922 (0.912, 0.933)	0.784 (0.757, 0.811)	0.945 (0.934, 0.956)	0.703 (0.660, 0.746)
AFFNet	**0.956 (0.947, 0.965)**	**0.793 (0.763, 0.822)**	**0.935 (0.919, 0.952)**	**0.869 (0.785, 0.954)**	**0.946 (0.915, 0.977)**	**0.737 (0.644, 0.830)**

Table 3
Five-fold cross-validation results of different methods on ADNI-3CLS and ADNI-4CLS tasks.

Methods	**ADNI-3CLS**					
	AUC (95% CI)	F_1-score (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)
VoxCNN	0.777 (0.759, 0.794)	0.609 (0.586, 0.632)	0.652 (0.631, 0.674)	0.614 (0.596, 0.632)	0.811 (0.800, 0.822)	0.630 (0.601, 0.659)
VoxResNet	0.773 (0.759, 0.787)	0.592 (0.570, 0.614)	0.634 (0.621, 0.647)	0.596 (0.565, 0.626)	0.801 (0.789, 0.812)	0.612 (0.608, 0.616)
AttResNet	0.793 (0.784, 0.802)	0.591 (0.535, 0.648)	0.642 (0.616, 0.669)	0.610 (0.575, 0.644)	0.810 (0.796, 0.823)	0.613 (0.567, 0.659)
M3T	0.740 (0.700, 0.780)	0.522 (0.450, 0.593)	0.580 (0.562, 0.598)	0.546 (0.509, 0.582)	0.776 (0.759, 0.794)	0.529 (0.411, 0.647)
AMSNet	0.812 (0.803, 0.822)	0.613 (0.597, 0.630)	0.655 (0.633, 0.677)	0.627 (0.609, 0.645)	0.816 (0.808, 0.824)	0.630 (0.579, 0.682)
AMSF	0.773 (0.757, 0.789)	0.579 (0.550, 0.608)	0.629 (0.615, 0.642)	0.584 (0.549, 0.618)	0.799 (0.787, 0.812)	0.599 (0.574, 0.623)
MPS-FFA	0.758 (0.739, 0.778)	0.582 (0.533, 0.632)	0.633 (0.596, 0.671)	0.602 (0.573, 0.632)	0.807 (0.790, 0.824)	0.592 (0.532, 0.652)
LA-GMF	0.791 (0.783, 0.799)	0.628 (0.595, 0.662)	0.661 (0.643, 0.680)	0.623 (0.587, 0.659)	0.820 (0.809, 0.830)	0.646 (0.609, 0.682)
AFFNet	**0.829 (0.823, 0.835)**	**0.655 (0.633, 0.677)**	**0.691 (0.678, 0.704)**	**0.655 (0.642, 0.668)**	**0.834 (0.826, 0.841)**	**0.675 (0.647, 0.703)**
Methods	**ADNI-4CLS**					
	AUC (95% CI)	F_1-score (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)
VoxCNN	0.766 (0.757, 0.775)	0.433 (0.408, 0.458)	0.622 (0.614, 0.630)	0.473 (0.458, 0.487)	0.852 (0.846, 0.858)	0.496 (0.479, 0.514)
VoxResNet	0.746 (0.716, 0.775)	0.414 (0.381, 0.447)	0.616 (0.599, 0.633)	0.460 (0.433, 0.487)	0.847 (0.833, 0.861)	0.521 (0.424, 0.619)
AttResNet	0.760 (0.735, 0.785)	0.380 (0.327, 0.432)	0.612 (0.603, 0.620)	0.442 (0.427, 0.457)	0.842 (0.837, 0.846)	0.403 (0.254, 0.552)
M3T	0.722 (0.681, 0.763)	0.331 (0.309, 0.353)	0.571 (0.543, 0.599)	0.405 (0.377, 0.432)	0.826 (0.811, 0.840)	0.286 (0.253, 0.319)
AMSNet	0.798 (0.776, 0.821)	0.443 (0.420, 0.465)	0.635 (0.618, 0.651)	0.483 (0.467, 0.500)	0.857 (0.849, 0.864)	0.481 (0.413, 0.548)
AMSF	0.743 (0.734, 0.752)	0.373 (0.336, 0.409)	0.606 (0.597, 0.615)	0.438 (0.419, 0.456)	0.842 (0.833, 0.850)	0.382 (0.253, 0.512)
MPS-FFA	0.758 (0.749, 0.768)	0.380 (0.351, 0.408)	0.607 (0.595, 0.620)	0.439 (0.424, 0.454)	0.841 (0.835, 0.846)	0.435 (0.324, 0.547)
LA-GMF	0.783 (0.771, 0.796)	0.473 (0.437, 0.510)	0.633 (0.622, 0.644)	0.493 (0.476, 0.510)	0.860 (0.854, 0.865)	0.516 (0.496, 0.536)
AFFNet	**0.819 (0.798, 0.839)**	**0.493 (0.412, 0.575)**	**0.656 (0.631, 0.681)**	**0.522 (0.467, 0.577)**	**0.869 (0.855, 0.884)**	**0.550 (0.482, 0.617)**

both the two modules MDA and PBF are beneficial to boosting the model performance. Specifically, in the ADNI-2CLS task, by implementing the semantic alignment strategy, the "B+SA" method achieves improvements of 0.8%, 0.9%, and 0.4% in AUC, F_1-score, and Accuracy compared to the "B" method, respectively. Additionally, the "B+MDA" method, which incorporates both semantic and attention alignment strategies, elevates the AUC to 96.7%, the F_1-score to 89.4%, and the Accuracy to 93.4%. This method also performs well in the AIBL-2CLS

Table 4
Ablation study on different key components inside AFFNet based on the AD-NC binary classification task.

B	MDA		PBF		ADNI-2CLS					
	SA	AA	CC	SR	AUC	F_1-score	Accuracy	Sensitivity	Specificity	Precision
✓	–	–	–	–	0.953	0.884	0.928	0.871	0.954	0.898
✓	✓	–	–	–	0.961	0.893	0.932	0.881	0.956	0.905
✓	✓	✓	–	–	0.967	0.894	0.934	0.874	0.962	0.916
✓	✓	✓	✓	–	0.967	0.902	0.939	0.886	**0.963**	**0.920**
✓	✓	✓	✓	✓	**0.971**	**0.912**	**0.944**	**0.912**	0.960	0.913
B	MDA		PBF		AIBL-2CLS					
	SA	AA	CC	SR	AUC	F_1-score	Accuracy	Sensitivity	Specificity	Precision
✓	–	–	–	–	0.949	0.782	0.930	**0.877**	0.939	0.708
✓	✓	–	–	–	0.947	0.769	0.928	0.845	0.941	0.714
✓	✓	✓	–	–	0.951	0.783	0.934	0.835	0.951	**0.747**
✓	✓	✓	✓	–	0.952	0.782	0.933	0.853	0.946	0.724
✓	✓	✓	✓	✓	**0.956**	**0.793**	**0.935**	0.869	**0.960**	0.737

Table 5
Ablation study on multi-view feature extraction and fusion based on the AD-NC binary classification task.

Model	ADNI-2CLS					
	AUC	F_1-score	Accuracy	Sensitivity	Specificity	Precision
Coronal plane	0.959	0.881	0.927	0.856	0.960	0.911
Sagittal plane	0.958	0.880	0.926	0.853	0.960	0.910
Axial plane	0.959	0.886	0.928	0.871	0.955	0.903
w/o 3D CNN	0.965	0.894	0.933	0.884	0.956	0.905
Concat	0.961	0.900	0.937	0.884	**0.962**	**0.918**
AFFNet	**0.971**	**0.912**	**0.944**	**0.912**	0.960	0.913
Model	AIBL-2CLS					
	AUC	F_1-score	Accuracy	Sensitivity	Specificity	Precision
Coronal plane	0.946	0.757	0.927	0.795	**0.949**	0.733
Sagittal plane	0.937	0.755	0.924	0.819	0.942	0.706
Axial plane	0.943	0.748	0.919	0.843	0.932	0.676
w/o 3D CNN	0.947	0.746	0.921	0.816	0.939	0.690
Concat	0.949	0.745	0.925	0.853	0.937	0.700
AFFNet	**0.956**	**0.793**	**0.935**	**0.869**	0.946	**0.737**

task, showing increases of 0.4%, 1.4%, and 0.6% in AUC, F_1-score, and Accuracy, respectively. When we further integrate the PBF module into the model, performance in the ADNI-2CLS task experiences a notable improvement, with the AUC increasing to 97.1%, the F_1-score to 91.2%, and Accuracy to 94.4%. Similarly, we observe improvements on the AIBL dataset, achieving an AUC of 95.6%, an F_1-score of 79.3%, and an Accuracy of 93.5%.

Ablation of Multi-view Feature Extraction and Fusion. To evaluate the importance of multi-view feature extraction and fusion in 3D sMRI analysis, we conduct a series of ablation studies based on the AD-NC binary classification task. These experiments include: models using only a single view (coronal, sagittal, or axial), a variant where the 3D CNN in AFFNet is replaced with three 2D CNNs of identical architecture (denoted as w/o 3D CNN), and a model where the mean aggregation in AFFNet is replaced with a simple concatenation (denoted as concat). The results are presented in Table 5. Compared to methods relying on a single view, AFFNet consistently outperforms across three key evaluation metrics (AUC, F_1-score, and ACC), underscoring the essential role of multi-view fusion in improving classification performance. The 3D CNN component also leads to performance gains, suggesting that capturing low-level spatial information benefits the classification task. Moreover, mean aggregation of the aligned and filtered features in the aligned space proves more effective than simple concatenation.

Furthermore, by combining the experimental results in Tables 4–5, we observe that although the effective fusion of multi-view information positively affects the classification decision, relying on only a single classification loss function to guide the fusion process of multi-view features might degrade the model performance. This finding validates our research motivation, indicating the need for more complex and refined mechanisms to optimize feature extraction and information integration during multi-view fusion.

5.7. Parameter analysis

In this subsection, we analyze the effect of two key hyperparameters α and β in the proposed framework. First, we search the optimal hyperparameter configurations in different classification tasks during the range of $[0.2, 0.4, 0.6, 0.8, 1.0]$. Then, we show the experimental results obtained by the control variable method under the selected optimal hyperparameter settings in Fig. 5, i.e., fixing the optimal value of α and adjusting the value of β, and vice versa. As we can see, on the ADNI-2CLS task, the optimal configurations are $\alpha = 0.4$ and $\beta = 0.4$; On the ADNI-3CLS task, the optimal configurations are $\alpha = 1.0$ and $\beta = 0.6$; On the ADNI-4CLS task, the optimal configurations are $\alpha = 0.8$ and $\beta = 0.6$. However, it is worth noting that our method is robust to changes of α and β, since different values of α and β do not affect the model performance significantly.

5.8. Interpretation experiments

In this section, we adopt both qualitative and quantitative methods to explore the potential of AFFNet in interpretability (i.e., locating discriminative regions).

Qualitative Evaluation. We visualize the attention maps of AFFNet and three comparison methods with attention mechanisms to demonstrate the ability of our method to locate discriminative regions across multiple tasks (ADNI-2CLS, ADNI-3CLS, and ADNI-4CLS). Specifically, for AFFNet, we extract slice-level attention from its three view branches (with shapes $S^1 \times 1 \times 1$, $1 \times S^2 \times 1$, and $1 \times 1 \times S^3$, respectively), and use a broadcasting mechanism and matrix multiplication to integrate these slices into 3D attention maps ($S^1 \times S^2 \times S^3$). The broadcasting mechanism and matrix multiplication in this process lead to the banded pattern in our attention map. Subsequently, the generated 3D attention

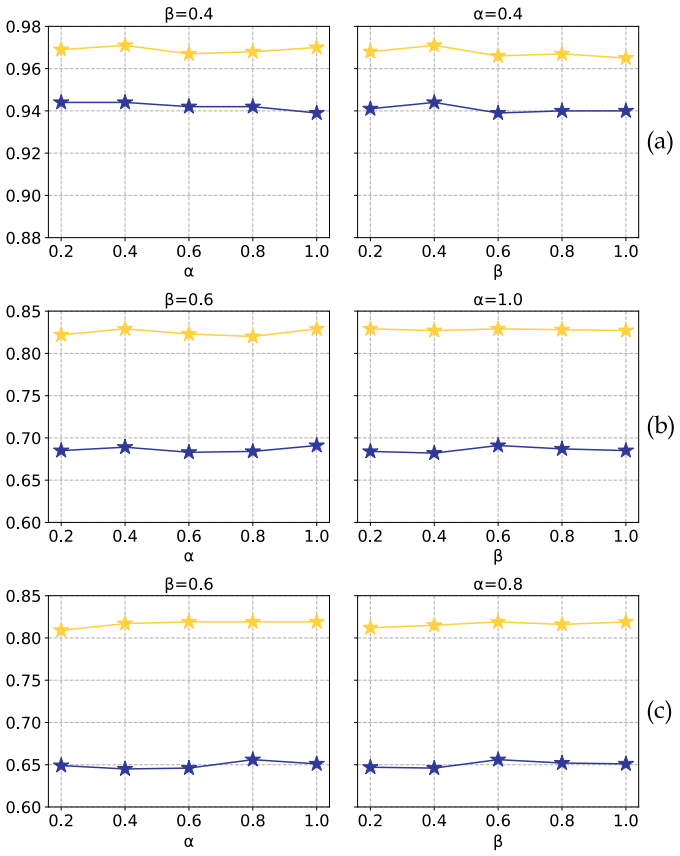

Fig. 5. Effect of varying α and β in Eq. (15) on model performance, where (a), (b), (c) denote the results on the ADNI-2CLS, ADNI-3CLS, and ADNI-4CLS tasks, respectively. ★ and ★ denote the Accuracy and AUC, respectively.

maps are upsampled and superimposed on the original images to provide intuitive visualization results. For the comparison models, we directly upsample the attention maps and overlay them on the original images to obtain visualizations. We show the attention maps of the four methods in Fig. 6, where the color gradient from blue to red indicates that the weight changes from low to high.

As we can see from Fig. 6, the attention map of AttResNet exhibits sparse characteristics, which might be caused by using ReLU as the activation function for the attention output. The attention map of AMSNet covers almost the entire foreground area, probably because dilated convolutions are used to fuse multi-scale information, leading to mixed local features and affecting the model's interpretability. Different from them, both LA-GMF and AFFNet demonstrate strong discriminative localization capabilities, effectively highlighting brain regions associated with AD, such as the hippocampus and parahippocampal gyrus. However, there are differences in their attention weights. The discrepancy might stem from the fact that LA-GMF aims to learn strong consistency between attention weights and category semantics, which leads to the attention map overfitting the category semantics. By contrast, AFFNet learns the distributional similarity between attention activations and slice-level true category prediction probabilities via its attention alignment module, and thus it offers more flexibility, enhancing AFFNet's performance in both localization and classification tasks.

Quantitative Evaluation. Because subtle differences in the same key brain regions between different subjects often constitute the key basis for disease diagnosis, it is crucial to locate discriminative regions stably. To systematically evaluate the consistency and reliability of various methods in identifying discriminative regions, we design a set of quantitative evaluation strategies, presented below. First, based on existing literature [39,40], we identify the brain regions most closely associated with AD. Their names and corresponding areas in the AAL template [41] are listed in Table 6.

Subsequently, for each method, we upsample the obtained attention maps to the original image size and segment them into multiple brain regions using the AAL template. We then calculate the average attention weight of each region as its score. The top 20 regions with the highest scores for each sample serve as the sample-level discriminative regions identified by the model. From all samples, we select the top 20 regions with the highest selection frequency as the population-level discriminative regions for each method. It is worth noting that we include M^{2}FAN in the comparison methods, and although it does not provide attention maps that can be visualized, it extracts patches from each brain region according to the AAL template and calculates attention scores, which is similar to our method. Finally, based on the brain regions closely associated with AD as reported in existing literature, and the population-level discriminative regions identified by the models, we calculate the frequency-weighted average precision AP_{fw} as the quantitative indicator of the positioning accuracy of different methods, it is:

$$AP_{fw} = \frac{\sum_{i=1}^{N_t} (\mathrm{P}(i) \times \mathrm{e}(i) \times \mathrm{f}(i))}{N_s}, \tag{16}$$

which is a variant of average precision (AP) to consider both the importance ranking of the brain regions mined by each method and the frequency of each brain region being selected, N_s and N_t represent the number of significant and total brain regions in the sequence, respectively. $\mathrm{e}(i) \in \{0, 1\}$ represents whether the i-th brain region is significant. $\mathrm{f}(i)$ is the frequency at which the i-th brain region is selected, and $\mathrm{P}(i)$ represents the precision of the i-th region containing significant brain regions, it is calculated by:

$$\mathrm{P}(i) = \frac{m_i}{i}, \tag{17}$$

where m_i is the number of searched significant brain regions in the first i brain regions.

We present the quantitative results in Table 7, which shows that the most stable brain areas identified by AttResNet are No. 59 (i.e. Parietal_Sup_L) and No. 60 (i.e. Parietal_Sup_R), while AMSNet tends to stably highlight No. 89 (i.e. Temporal_Inf_L) and No. 90 (i.e. Temporal_Inf_R) brain area. The association between these areas and AD diagnosis is not significant in the literature. Although M^{2}FAN can effectively locate brain area 39 (i.e. ParaHippocampal_L), its overall positioning accuracy is still poor. By contrast, LA-GMF can effectively and stably locate areas that are crucial for AD diagnosis, but its stability is still lower than the proposed AFFNet. For more clarity, we provide a visualization example of key brain regions based on the 027_S_0120 sample in Fig. 7, using the same slicing configuration as that in Fig. 6. Therefore, Figs. 6–7 and Table 7 all demonstrate the superiority of AFFNet in localization performance, indicating its significant advantages in identifying brain areas related to AD diagnosis.

6. Discussion

Extensive experiments indicate that the performance gains of AFFNet primarily stem from its effective integration of multi-view features. Unlike most existing methods that rely solely on the classification loss to extract discriminative information for AD diagnosis, AFFNet introduces cross-view consistency through a weak alignment strategy, implicitly regularizing complementary features across views (see Fig. C.1(a)–(b) in Appendix C for visualization). Additionally, its generalization performance on the AIBL dataset further demonstrates the effectiveness of the dual alignment strategy in mitigating overfitting. Furthermore, as shown in Table 4 and Fig. C.1(c) in Appendix C, the private bias filtering module enhances feature discriminability, underscoring the importance of explicitly addressing view-specific biases in multi-view learning—an aspect often overlooked by previous multi-view methods.

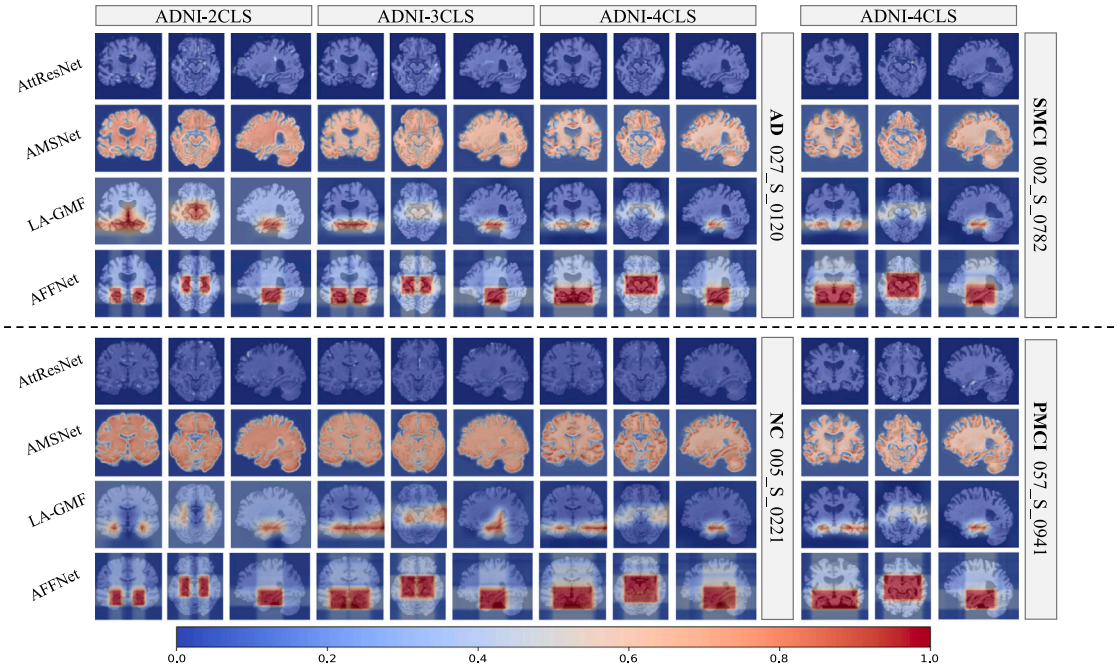

Fig. 6. Comparison of attention maps from different models across various task settings. AttResNet and AMSNet suffer from interpretability issues: AttResNet produces sparse maps, while AMSNet's maps overly cover the foreground. By contrast, LA-GMF and AFFNet effectively highlight key brain regions, with AFFNet showing superior robustness and stability in identifying discriminative areas.

Table 6
The corresponding names and numbers of important brain regions related to AD in the AAL template.

ROI	AAL name	AAL number
hippocampal, uncus	Hippocampus_L, Hippocampus_R	37/38
parahippocampal, entorhinal cortex	ParaHippocampal_L, ParaHippocampal_R	39/40
amygdale	Amygdala_L, Amygdala_R	41/42
fusiform	Fusiform_L, Fusiform_R	55/56
precuneus	Precuneus_L, Precuneus_R	67/68

Table 7
Comparison of population-level discriminant regions and the AP_{fw} identified by different models.

Methods	Selected ROIs and corresponding frequencies	AP_{fw}
AttResNet	59(0.65), 60(0.64), 24(0.55), **41**(0.53), 23(0.52), 69(0.47), 11(0.43), 19(0.43), 63(0.43), 57(0.42), 13(0.41), 22(0.41), 17(0.40), **42**(0.39), 26(0.39), 65(0.39), 83(0.37), 61(0.37), 4(0.34), 1(0.33)	0.019
M^{2}FAN	**39**(0.60), 87(0.59), 27(0.58), 26(0.58), 89(0.58), 25(0.57), 88(0.57), **41**(0.55), **42**(0.54), 84(0.53), 15(0.52), **40**(0.52), 28(0.52), 5(0.50), 21(0.47), 16(0.46), 83(0.46), 6(0.46), **38**(0.46), 75(0.45)	0.121
AMSNet	90(0.70), 89(0.69), 75(0.58), 76(0.58), 53(0.56), 114(0.56), 115(0.53), 54(0.52), **36**(0.52), 28(0.52), **56**(0.49), 73(0.48), 18(0.47), 74(0.45), **41**(0.42), 106(0.41), 17(0.40), 29(0.39), 5(0.39), **42**(0.39)	0.016
LA-GMF	**42**(0.99), **41**(0.99), **40**(0.99), **38**(0.98), **56**(0.96), **39**(0.86), **37**(0.83), **55**(0.83), 89(0.79), 83(0.76), 87(0.72), 84(0.71), 98(0.69), 108(0.68), 100(0.66), 96(0.65), 97(0.64), 90(0.60), 95(0.59), 107(0.53)	0.745
AFFNet	**37**(1.00), **38**(1.00), **39**(1.00), **40**(1.00), **41**(1.00), **42**(1.00), **56**(1.00), **55**(1.00), 96(0.99), 98(0.99), 97(0.98), 76(0.97), 75(0.97), 95(0.91), 108(0.87), 73(0.84), 107(0.80), 74(0.76), 100(0.71), 99(0.55)	**0.800**

While AFFNet has demonstrated promising results, its scalability to larger datasets remains to be fully explored. In future work, we plan to integrate multiple publicly available datasets to construct large-scale benchmarks, thereby enabling a more comprehensive evaluation of AFFNet's generalization ability. Additionally, AFFNet focuses solely on utilizing sMRI data, while accurate AD diagnosis typically benefits from multi-modal integration, such as combining clinical records, PET, and other imaging modalities. Considering the semantic differences between different modalities, in the future we plan to enhance the dual alignment module for multi-modal settings by integrating the confidence scores of specific modalities to guide cross-modal semantic alignment. Meanwhile, we will explore the applicability of the privacy bias filtering module in the field of multi-modal learning, i.e., whether it can effectively identify and suppress irrelevant signals specific to each modality, thereby facilitating robust multi-modal integration. Moreover, AFFNet still relies on lightweight CNNs for feature extraction, whereas recent large language models [42] and pretrained multi-modal frameworks [43,44] have demonstrated remarkable generalization abilities in medical-related tasks. Therefore, we plan to explore integrating these powerful models into our framework to further improve diagnostic performance and scalability.

7. Conclusions

In this paper, we propose a novel deep multi-view learning framework, AFFNet, which is designed to fully exploit the multi-view information embedded in 3D sMRI for boosting AD diagnosis. By integrating both 3D and 2D convolutions, the framework effectively preserves low-level spatial details and mitigates the loss of critical structural

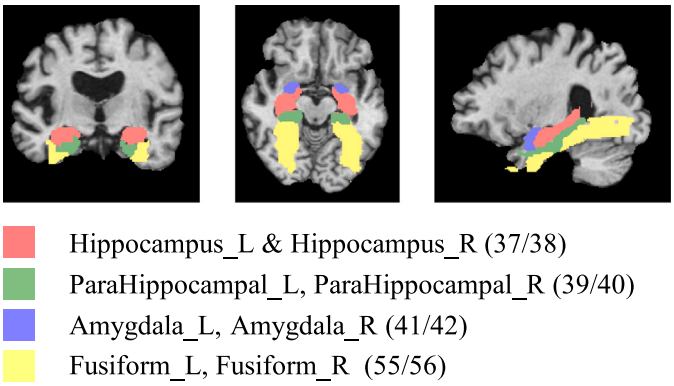

Fig. 7. Visualization of key brain regions of the sample 027_S_0120 under the same slice settings as Fig. 6.

information. Additionally, the proposed dual alignment module enhances feature consistency across views while retaining complementary information, thereby improving the quality of multi-view representations. Moreover, the private bias filtering module leverages a cross-view contrastive loss and the orthogonal decomposition with semantic regularization to identify and eliminate intra-view biases, further promoting effective feature fusion. Experimental results demonstrate that AFFNet achieves superior performance in AD classification and produces more discriminative and interpretable attention maps, providing clinicians with a more trustworthy basis for supporting AD diagnosis.

CRediT authorship contribution statement

Jinghao Xu: Writing – original draft, Validation, Investigation, Visualization, Methodology, Data curation. **Chenxi Yuan:** Validation, Visualization, Methodology. **Yi Jing:** Validation, Visualization. **Huifang Shang:** Validation, Methodology, Writing – review & editing, Supervision. **Xiaoshuang Shi:** Writing – review & editing, Supervision, Funding acquisition, Validation, Methodology. **Xiaofeng Zhu:** Supervision, Writing – review & editing, Funding acquisition.

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

This work was supported in part by the National Key Research and Development Program of China (No. 2022YFA1004100), and the Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China and West China Hospital (No. ZYGX2022YGRH009).

Appendix A. Algorithm flow

In this appendix, we provide a detailed description of the training procedure for the proposed algorithm to facilitate a better understanding, as shown in Algorithm A.1.

Algorithm A.1 AFFNet

Input: Training data D, the number of training epochs T.
Output: Well trained AFFNet.

for $\tau \in [1, T]$ **do**
  Random select a batch of samples $\{\mathbf{X}, \mathbf{Y}\}$ from D;
  $\mathbf{I}_{3D} \leftarrow \mathrm{f}_{3D}(\mathbf{X})$ ▷ Obtain 3D representation.
  $\{\mathbf{I}_{2D}^{m}\}_{m=1}^{M} \leftarrow \mathrm{reshape}(\mathbf{I}_{3D})$ ▷ Obtain 2D multi-view input features.
  $\{\mathbf{I}^{m}\}_{m=1}^{M} \leftarrow \{\mathrm{f}_{2D}^{m}(\mathbf{I}_{2D}^{m})\}_{m=1}^{M}$ ▷ Obtain multi-view slice-level features.
  $\{\mathbf{H}^{m}\}_{m=1}^{M} \leftarrow$ Eqs. (2)-(3) ▷ Obtain multi-view global confounding features.
  $\mathcal{L}_{SA} \leftarrow$ Eqs. (4)-(5) ▷ Perform semantic alignment.
  $\mathcal{L}_{AA} \leftarrow$ Eqs. (6)-(7) ▷ Perform attention alignment constraints.
  $\mathcal{L}_{CC}, \mathcal{L}_{SR}, \{\mathbf{E}^{m}\}_{m=1}^{M} \leftarrow$ Eqs. (8)-(13) ▷ Filter private bias.
  $\mathcal{L}_{CLS} \leftarrow$ Eq. (14) ▷ Fuse $\{\mathbf{E}^{m}\}_{m=1}^{M}$ and perform the final classification.
  $\mathcal{L} \leftarrow$ Eq. (15) ▷ The final objective function.
  Back-propagate $\mathcal{L}$ to update model parameters;
end for

Appendix B. Visualization of the ADNI dataset

In this appendix, we present the 2D t-SNE visualization results based on the ADNI dataset to assess the potential confounding effects of age and sex on imaging data. Specifically, we downsample the preprocessed sMRI scans by a factor of 8 and use the flattened features as the input to t-SNE. To evaluate the effect of age, we divide the age of subjects into four intervals: <60, [60, 70), [70, 80) and ≥80, and perform 2D embedding visualization with these age intervals as labels. For the evaluation of sex, we directly utilize the sex of the subject as a label for 2D embedding visualization. As we can see from Fig. B.1, the visualization of sMRI data shows that there is no obvious cluster, no matter whether they are grouped by age or sex. This indicates that the preprocessed imaging data does not have significant age- or sex-related bias.

Appendix C. t-SNE visualization of multi-view features

To more clearly demonstrate the working mechanism of our proposed module, we perform t-SNE visualization of the features extracted under different settings, as shown in Fig. C.1.

Specifically, Fig. C.1(a) and (b) present the visualizations of multi-view features before and after applying the dual alignment module, respectively. Before alignment, features from different views are intermingled in the embedding space with significant color overlaps, suggesting that complementary information across views is not effectively captured. Although a certain degree of inter-class separability exists, the intra-class compactness is relatively poor. After applying the dual alignment module, features from different views converge toward a shared clustering center, revealing a well-organized cross-view structure. Additionally, features within each view exhibit enhanced intra-view cohesion, and the intra-class compactness is substantially improved. These results clearly demonstrate that our alignment strategy introduces weak consistency across views while effectively preserving complementary information.

Fig. C.1(c) further shows the visualizations of private bias features and essential features extracted by the private bias filtering module. It can be observed that private bias information is successfully identified and disentangled. Meanwhile, the essential features exhibit strong intra-view cohesion and clear inter-class separability, indicating that the extracted essential representations not only retain complementary information but also enhance semantic discriminability.

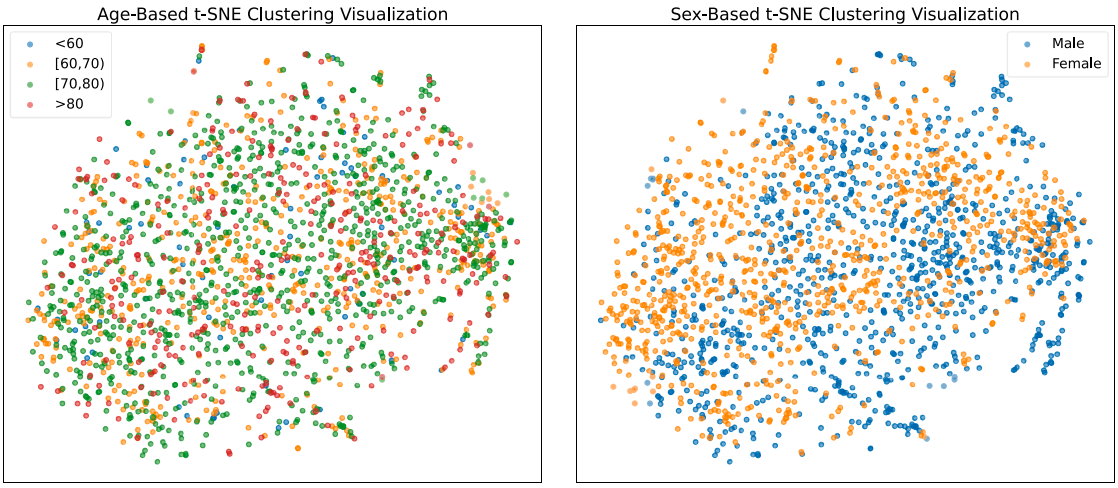

Fig. B.1. 2D *t*-SNE embedding visualization of preprocessed structured MRI data after 8x downsampling, categorized by age and sex.

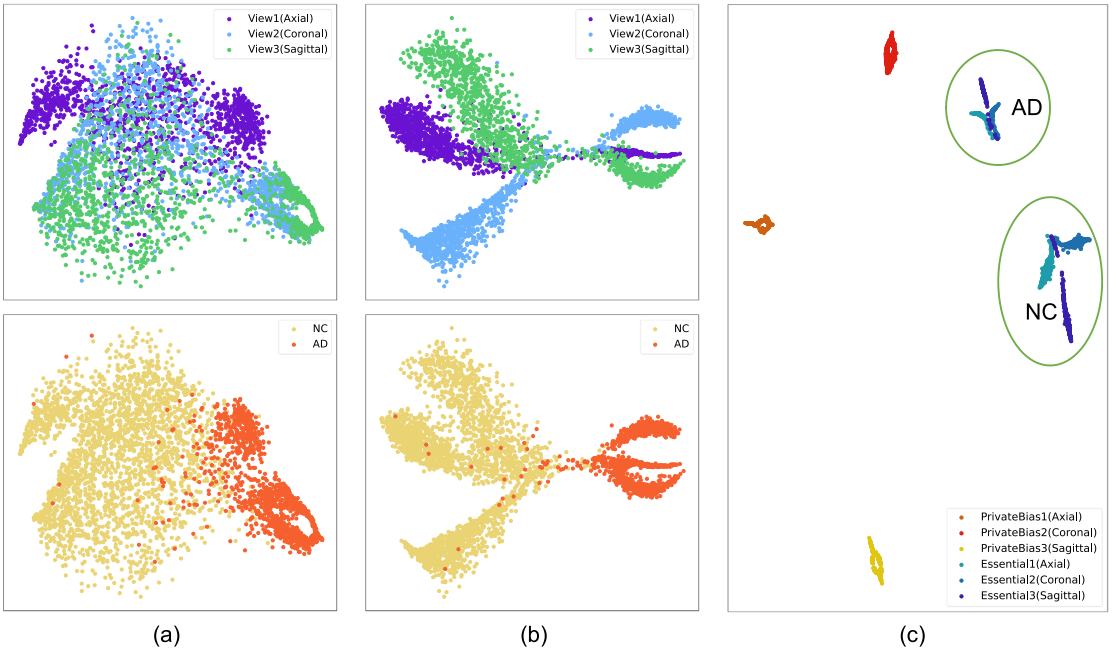

Fig. C.1. 2D *t*-SNE visualizations of learned features. (a) Visualization of multi-view features before dual alignment. (b) Visualization of multi-view features after dual alignment. (c) Visualization of private bias features and essential features extracted by the private bias filtering module.

Data availability

All data list and source codes are available at: https://github.com/nollexu/AFFNet.

References

[1] R. Matej, A. Tesar, R. Rusina, Alzheimer's disease and other neurodegenerative dementias in comorbidity: a clinical and neuropathological overview, Clin. Biochem. 73 (2019) 26–31.

[2] P. Vemuri, C.R. Jack, Role of structural MRI in alzheimer's disease, Alzheimer's Res. Ther. 2 (2010) 1–10.

[3] M. Hon, N.M. Khan, Towards alzheimer's disease classification through transfer learning, in: 2017 IEEE International Conference on Bioinformatics and Biomedicine, IEEE, 2017, pp. 1166–1169.

[4] R. Mendoza-Léon, J. Puentes, L.F. Uriza, M.H. Hoyos, Single-slice alzheimer's disease classification and disease regional analysis with supervised switching autoencoders, Comput. Biol. Med. 116 (2020) 103527.

[5] W. Kang, L. Lin, B. Zhang, X. Shen, S. Wu, Multi-model and multi-slice ensemble learning architecture based on 2d convolutional neural networks for alzheimer's disease diagnosis, Comput. Biol. Med. 136 (2021) 104678.

[6] Y. Ban, X. Zhang, H. Lao, Diagnosis of alzheimer's disease using structure highlighting key slice stacking and transfer learning, Med. Phys. 49 (9) (2022) 5855–5869.

[7] M. Liu, J. Zhang, E. Adeli, D. Shen, Landmark-based deep multi-instance learning for brain disease diagnosis, Med. Image Anal. 43 (2018) 157–168.

[8] W. Zhu, L. Sun, J. Huang, L. Han, D. Zhang, Dual attention multi-instance deep learning for alzheimer's disease diagnosis with structural MRI, IEEE Trans. Med. Imaging 40 (9) (2021) 2354–2366.

[9] K. Han, M. He, F. Yang, Y. Zhang, Multi-task multi-level feature adversarial network for joint alzheimer's disease diagnosis and atrophy localization using sMRI, Phys. Med. Biol. 67 (8) (2022) 085002.

[10] Y. Wu, Y. Zhou, W. Zeng, Q. Qian, M. Song, An attention-based 3D CNN with multi-scale integration block for alzheimer's disease classification, IEEE J. Biomed. Heal. Inform. 26 (11) (2022) 5665–5673.

[11] J. Xu, C. Yuan, X. Ma, H. Shang, X. Shi, X. Zhu, Interpretable medical deep framework by logits-constraint attention guiding graph-based multi-scale fusion for alzheimer's disease analysis, Pattern Recognit. 152 (2024) 110450.

[12] Q. Chen, Q. Fu, H. Bai, Y. Hong, Longformer: Longitudinal transformer for alzheimer's disease classification with structural MRIs, in: Proceedings of the

IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 3575–3584.
[13] Z. Weng, J. Meng, Z. Ding, J. Yuan, S3F: A multi-view slow-fast network for alzheimer's disease diagnosis, in: 2020 IEEE International Conference on Multimedia and Expo, IEEE, 2020, pp. 1–6.
[14] J.Y. Choi, B. Lee, Combining of multiple deep networks via ensemble generalization loss, based on MRI images, for alzheimer's disease classification, IEEE Signal Process. Lett. 27 (2020) 206–210.
[15] H. Qiao, L. Chen, F. Zhu, A fusion of multi-view 2D and 3D convolution neural network based MRI for alzheimer's disease diagnosis, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, IEEE, 2021, pp. 3317–3321.
[16] L. Chen, H. Qiao, F. Zhu, Alzheimer's disease diagnosis with brain structural mri using multiview-slice attention and 3D convolution neural network, Front. Aging Neurosci. 14 (2022) 871706.
[17] J. Jang, D. Hwang, M3T: Three-dimensional medical image classifier using multi-plane and multi-slice transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20718–20729.
[18] Y. Zhang, S. Peng, Z. Xue, G. Zhao, et al., AMSF: attention-based multi-view slice fusion for early diagnosis of alzheimer's disease, PeerJ Comput. Sci. 9 (2023) e1706.
[19] F. Liu, H. Wang, S.-N. Liang, Z. Jin, et al., MPS-FFA: A multiplane and multiscale feature fusion attention network for alzheimer's disease prediction with structural MRI, Comput. Biol. Med. 157 (2023) 106790.
[20] Q. Zhang, Y. Long, H. Cai, S. Yu, Y. Shi, X. Tan, A multi-slice attention fusion and multi-view personalized fusion lightweight network for alzheimer's disease diagnosis, BMC Med. Imaging 24 (1) (2024) 258.
[21] H. Huang, W. Pedrycz, K. Hirota, F. Yan, A multiview-slice feature fusion network for early diagnosis of alzheimer's disease with structural MRI images, Inf. Fusion (2025) 103010.
[22] K. Han, G. Li, Z. Fang, F. Yang, Multi-template meta-information regularized network for alzheimer's disease diagnosis using structural MRI, IEEE Trans. Med. Imaging 43 (5) (2024) 1664–1676, http://dx.doi.org/10.1109/TMI.2023.3344384.
[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, et al., Attention is all you need, in: Advances in Neural Information Processing Systems, vol. 30, 2017.
[24] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[25] C. Zhang, J. Cheng, Q. Tian, Multi-view image classification with visual, semantic and view consistency, IEEE Trans. Image Process. 29 (2019) 617–627.
[26] P. Ren, C. Li, H. Xu, Y. Zhu, et al., Viewco: Discovering text-supervised segmentation masks via multi-view semantic consistency, 2023, arXiv preprint arXiv:2302.10307.
[27] X. Jia, X.-Y. Jing, X. Zhu, S. Chen, et al., Semi-supervised multi-view deep discriminant representation learning, IEEE Trans. Pattern Anal. Mach. Intell. 43 (7) (2020) 2496–2509.
[28] J. Xu, H. Tang, Y. Ren, L. Peng, X. Zhu, L. He, Multi-level feature learning for contrastive multi-view clustering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16051–16060.
[29] J. Xu, Y. Ren, X. Shi, H.T. Shen, X. Zhu, UNTIE: Clustering analysis with disentanglement in multi-view information fusion, Inf. Fusion 100 (2023) 101937.
[30] S. Qiu, M.I. Miller, P.S. Joshi, J.C. Lee, et al., Multimodal deep learning for alzheimer's disease dementia assessment, Nat. Commun. 13 (1) (2022) 3404.
[31] Z. Yaniv, B.C. Lowekamp, H.J. Johnson, R. Beare, SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research, J. Digit. Imaging 31 (3) (2018) 290–303.
[32] M. Jenkinson, P. Bannister, M. Brady, S. Smith, Improved optimization for the robust and accurate linear registration and motion correction of brain images, Neuroimage 17 (2) (2002) 825–841.
[33] C.J. Holmes, R. Hoge, L. Collins, R. Woods, A.W. Toga, A.C. Evans, Enhancement of MR images using registration for signal averaging, J. Comput. Assist. Tomogr. 22 (2) (1998) 324–333.
[34] S.M. Smith, Fast robust automated brain extraction, Hum. Brain Mapp. 17 (3) (2002) 143–155.
[35] M. Jenkinson, C.F. Beckmann, T.E. Behrens, M.W. Woolrich, S.M. Smith, Fsl, Neuroimage 62 (2) (2012) 782–790.
[36] S. Korolev, A. Safiullin, M. Belyaev, Y. Dodonova, Residual and plain convolutional neural networks for 3D brain MRI classification, in: 2017 IEEE 14th International Symposium on Biomedical Imaging, IEEE, 2017, pp. 835–838.
[37] D. Jin, J. Xu, K. Zhao, F. Hu, et al., Attention-based 3D convolutional network for alzheimer's disease diagnosis and biomarkers exploration, in: 2019 IEEE 16Th International Symposium on Biomedical Imaging, IEEE, 2019, pp. 1047–1051.
[38] D.P. Kingma, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
[39] G. Karas, P. Scheltens, S. Rombouts, R. Van Schijndel, et al., Precuneus atrophy in early-onset alzheimer's disease: a morphometric structural MRI study, Neuroradiology 49 (2007) 967–976.
[40] C.J. Galton, K. Patterson, K. Graham, M.A. Lambon-Ralph, et al., Differing patterns of temporal atrophy in alzheimer's disease and semantic dementia, Neurology 57 (2) (2001) 216–225.
[41] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, et al., Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain, Neuroimage 15 (1) (2002) 273–289.
[42] A.J. Thirunavukarasu, D.S.J. Ting, K. Elangovan, L. Gutierrez, T.F. Tan, D.S.W. Ting, Large language models in medicine, Nature Med. 29 (8) (2023) 1930–1940.
[43] Z. Zhao, Y. Liu, H. Wu, Y. Li, et al., Clip in medical imaging: A comprehensive survey, 2023, arXiv preprint arXiv:2312.07353.
[44] H. Xiao, F. Zhou, X. Liu, T. Liu, et al., A comprehensive survey of large language models and multimodal large language models in medicine, 2024, arXiv preprint arXiv:2405.08603.