Computers in Biology and Medicine 163 (2023) 107110

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

End-to-end automatic pathology localization for Alzheimer's disease diagnosis using structural MRI

Gongpeng Cao [a], Manli Zhang [a], Yiping Wang [a], Jing Zhang [a], Ying Han [b], Xin Xu [c], Jinguo Huang [a,*], Guixia Kang [a,*]

[a] Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, No. 10 Xitucheng Road, Haidian District, Beijing, 100876, China
[b] Department of Neurology, Xuanwu Hospital of Capital Medical University, No. 45 Changchun Street, Xicheng District, Beijing, 100053, China
[c] Department of Neurosurgery, Chinese PLA General Hospital, No. 28 Fuxing Road, Haidian District, Beijing, 100853, China

ARTICLE INFO

Keywords:
Alzheimer's disease diagnosis
Pathology localization
End-to-end joint learning
Location coordinate prediction
Differentiable patch-cropping

ABSTRACT

Structural magnetic resonance imaging (sMRI) is an essential part of the clinical assessment of patients at risk of Alzheimer dementia. One key challenge in sMRI-based computer-aided dementia diagnosis is to localize local pathological regions for discriminative feature learning. Existing solutions predominantly depend on generating saliency maps for pathology localization and handle the localization task independently of the dementia diagnosis task, leading to a complex multi-stage training pipeline that is hard to optimize with weakly-supervised sMRI-level annotations. In this work, we aim to simplify the pathology localization task and construct an end-to-end automatic localization framework (AutoLoc) for Alzheimer's disease diagnosis. To this end, we first present an efficient pathology localization paradigm that directly predicts the coordinate of the most disease-related region in each sMRI slice. Then, we approximate the non-differentiable patch-cropping operation with the bilinear interpolation technique, which eliminates the barrier to gradient backpropagation and thus enables the joint optimization of localization and diagnosis tasks. Extensive experiments on commonly used ADNI and AIBL datasets demonstrate the superiority of our method. Especially, we achieve 93.38% and 81.12% accuracy on Alzheimer's disease classification and mild cognitive impairment conversion prediction tasks, respectively. Several important brain regions, such as rostral hippocampus and globus pallidus, are identified to be highly associated with Alzheimer's disease.

1. Introduction

Alzheimer's disease (AD), characterized by the irreversible loss of neurons and progressive impairment of cognitive functions [1], is one of the most prevalent neurodegenerative disorders among the elderly, accounting for 60% to 80% of the dementia cases [2]. Although there is still no effective treatment to reverse AD, accurate diagnosis and timely intervention at the prodromal stage (mild cognitive impairment, MCI) can effectively delay the onset of the disease [3]. Structural magnetic resonance imaging (sMRI) has been widely used for the early detection of AD as it can non-invasively capture AD-induced anatomical changes that emerge earlier than clinical symptoms [4,5]. However, diagnosis relying on clinicians to visually examine sMRI slice by slice is time-consuming and suffers from unavoidable subjective factors, leading to relatively imprecise diagnosis results. Therefore, it is of practical significance to develop reliable sMRI-based computer-aided diagnosis (CAD) systems to help clinicians identify subjects with suspected AD in time.

The key to sMRI-based computational dementia diagnosis is discriminative representation learning, as AD-induced brain atrophy is subtle and only occurs in a few local regions [6,7], leaving it challenging to extract discriminative feature representations from sMRI for accurate AD diagnosis. Motivated by the advances in deep learning, extensive studies have been focusing on using powerful deep neural networks (DNNs) as sMRI feature extractors to learn discriminative disease representations, including two-dimensional convolutional neural networks (2D CNNs), three-dimensional (3D) CNNs, and Transformers. 2D-CNN approaches [8–14] usually transfer ImageNet [15] pre-trained networks (e.g., ResNet [16]) to classify sMRI slices, and then aggregate all slice-level predictions to produce the subject-level prediction. 3D-CNN approaches apply 3D convolutions to the whole-brain sMRI [17–22] or some empirically predetermined anatomical regions [23,24], and directly perform subject-level prediction. Because of the additional kernel dimension, such methods tend to customize a shallow architecture to avoid heavy training parameters [25]. More recently, with the

* Corresponding authors.
E-mail addresses: hjg@bupt.edu.cn (J. Huang), gxkang@bupt.edu.cn (G. Kang).

https://doi.org/10.1016/j.compbiomed.2023.107110
Received 25 January 2023; Received in revised form 18 May 2023; Accepted 30 May 2023
Available online 2 June 2023

popularity of vision Transformers [26], several studies [27–29] also explore re-adapting Transformer architecture [30] to fit sMRI slices or the whole-brain sMRI for AD-related diagnosis tasks. While these studies have reported impressive diagnosis accuracy, further improvement is hampered due to neglecting the localization of dementia-related regions. This is because feature representations from the high layers of DNNs prefer to respond to the global semantics of the entire image [31], making it hard to extract discriminative representations without any guidance of pathological locations. On the other hand, the inability to localize pathological regions also makes the diagnosis decisions of these deep models opaque, which limits their deployment in clinical applications [32,33].

To mitigate this issue, a number of works propose weakly-supervised localization algorithms, which generate a dense saliency map with the same size as the original 3D sMRI to detect pathological locations. Typically, early solutions [7,34–38] combine local morphology features and statistical test models to create a p-value map, where locations with group-level differences are considered as pathological landmarks. Despite effective, such methods ignore individual heterogeneity and are sensitive to the choice of features. To overcome these problems, recent studies [39–41] use patch-based supervised strategies to train a fully convolutional network (FCN)-based pathology detector for generating subject-specific disease probability maps, which indicate the disease risk of each location. [42–44] employ visual explanation analysis techniques, such as class activation mapping (CAM) [45], gradient-weighted CAM (Grad-CAM) [46], and counterfactual reasoning [47], to derive disease attention maps from a pre-trained deep diagnosis model, revealing the contribution of each location to the disease. After obtaining the saliency maps, these methods either crop local sMRI patches from the identified pathological regions to train a patch-level diagnosis network [7,34–39,41,43], or directly utilize these saliency maps to guide a whole-brain level diagnosis network's focus on local pathological regions via the spatial attention mechanism [40,42,44]. By this way, discriminative local features are extracted for boosting the diagnosis accuracy, and the identified pathological regions can also interpret the diagnosis decisions, enabling such approaches to be explainable.

Although existing saliency map-based localization methods have shown positive effects, there still exist some limitations. First, generating dense saliency maps involves predicting the salient score for each voxel in 3D sMRI, which prefers the strong supervision of voxel-level fine-grained annotations. In practice, however, only sMRI-level category labels are available for weakly-supervised guidance, causing the optimization of the localization task to become difficult. To reduce the optimization difficulty, most methods choose to generate low-resolution saliency maps and then upsample them to the original sMRI size, which often results in rough and inaccurate localization results. Recent study [44] has attempted to directly generate high-resolution saliency maps using a generative adversarial network (GAN). But training GANs is notoriously challenging, they have to design multiple complex losses to ensure the training stability, resulting in extremely slow convergence speed. Second, limited by the unlearnable or post-hoc properties of localization algorithms, existing methods usually learn localization and diagnosis tasks separately rather than in an end-to-end unified framework. For instance, statistical test-based methods rely on unlearnable group comparison to generate saliency maps; FCN detector-based or visual explanation-based methods require post-hoc analysis on a pre-trained diagnosis model to derive saliency maps. This introduces a complicated multi-stage training pipeline, and may also result in sub-optimal results due to the potential inconsistency of the learning objectives of the two tasks. Therefore, how to simplify the localization task to construct an end-to-end joint localization-diagnosis framework still remains an unsolved problem.

To advance current localization-based diagnosis methods, we propose an *end-to-end* ***automatic*** *pathology* ***localization*** *framework for AD diagnosis* (**AutoLoc**), where the learning of localization and diagnosis is target-consistent and mutually reinforcing. Instead of generating saliency maps, our framework formulates the localization problem as predicting the coordinate of the most disease-relevant region to the diagnosis task in each sMRI slice, which naturally does not require voxel-level supervision, making it much easier to solve the localization task in a weakly-supervised fashion. To ensure that the localization task can be optimized together with the diagnosis task through standard backpropagation, we approximate the non-differentiable patch-cropping operation with the bilinear interpolation technique, which establishes the fully differentiable characteristic of AutoLoc framework, thereby supporting one-stage end-to-end training. More specifically, as illustrated in Fig. 1, our AutoLoc first uses a context-aware localizer network, taking as input the global slice information extracted from the sMRI slice sequence with a light-weight DNN, to decide the optimal top-left coordinate of the pathological patch for each slice. Then, the localized patches are interpolated from original slices instead of being cropped out, and sent into a high-capacity DNN to extract discriminative features for AD diagnosis. We conduct comprehensive experiments on ADNI and AIBL datasets, and the experimental results demonstrate the good performance of our AutoLoc framework on multiple AD-related diagnosis tasks, including AD classification and MCI conversion prediction.

In summary, we make the following contributions:

- An end-to-end deep learning framework is designed for joint pathology localization and AD diagnosis using sMRI.
- Coordinate-based localization paradigm is developed to detect AD-related pathological regions, which simplifies the localization task and makes it much easier to optimize under weak supervision.
- Interpolation-based patch cropping is introduced to allow gradient backpropagation between localization and diagnosis tasks, which supports the one-stage end-to-end training of our framework.
- Comprehensive experiments are conducted to demonstrate the superiority and generalizability of our approach. Detailed pathological analysis reveals some important brain regions that are highly associated with AD.

The rest of the paper is organized as follows. Section 2 introduces the proposed method. Section 3 shows the experimental setup and results analysis, including comparisons with baselines and state-of-the-art methods, pathology region analysis and ablation studies. Section 4 concludes the work.

2. Method

2.1. Network architecture

Overview. Fig. 1 illustrates an overview of our approach. Given a sMRI slice sequence along the selected axis (we default to selecting the sagittal axis as we observe that it performs better than coronal and axial axes, as shown in the Table 5), AutoLoc first quickly skims each slice at a lower resolution using a light-weight slice network f_S, aiming to obtain global slice features. Then, these global features are fed into a context-aware localizer network $\mathcal{G}$ to predict pathological region coordinates. Afterwards, pathological patches located at these locations are cropped from original slices via bilinear interpolation and then passed into a high-capacity patch network f_P for extracting detailed local features. Finally, a classifier f_C aggregates the global and local information of all slices to make AD-related decisions. Actually, the pipeline of our framework is similar to the clinicians' sMRI assessment process that they usually first skim each slice to find suspected pathological regions (corresponding to the procedures of slice and localizer networks), and then carefully analyze these local regions for an accurate diagnosis (corresponding to the procedures of patch and

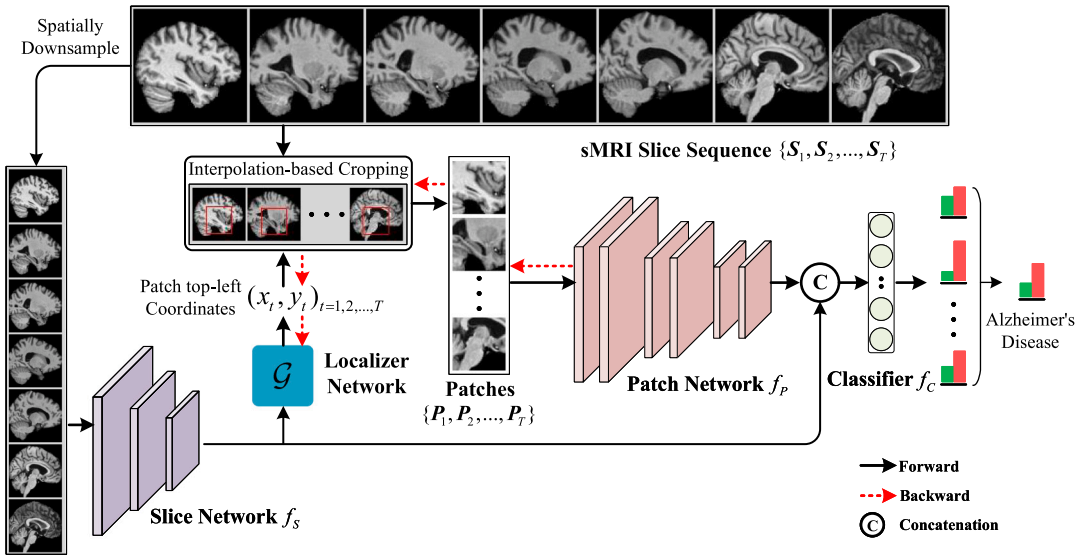

Fig. 1. Overview of AutoLoc framework. Given a sMRI slice sequence, our framework first feeds it into a light-weight slice network f_S to take a quick skim through all slices and extract global slice features. Then a simple localizer network $\mathcal{G}$ is followed to predict the optimal location coordinate of the pathological patch within each slice based on the derived global information. Afterwards, the localized pathological patches are cropped from original slices and sent into a high-capability patch network f_P for extracting disease-related local features, and a classifier f_C is used to obtain the diagnosis results by aggregating the global and local information of all slices. We use the bilinear interpolation technique to approximate the non-differentiable patch-cropping operation, building a fully end-to-end trainable framework for joint pathology localization and Alzheimer's disease diagnosis. See texts for more details.

classifier networks). In the following, we describe each component in detail.

Slice network f_S and patch network f_P are two 2D backbone networks that extract complementary deep features from the input. Slice network f_S aims to take a quick glance at each slice, providing the localizer with basic global information to decide which region is the most disease-related under the given diagnosis task. Therefore, a low-capacity network (e.g., MobileNet [48]) is applied to instantiate f_S. Patch network f_P, conversely, is designed to fully mine discriminative local features from localized pathological patches. Hence, we adopt a high-capacity network (e.g., ResNet) for f_P. It is worth noting that f_P is also computationally efficient as it only deals with local patches with small size.

Formally, given T sagittal slices $\{\boldsymbol{S}_1, \boldsymbol{S}_2, \dots, \boldsymbol{S}_T\}$ with spatial size $H \times W$ from a sMRI scan, we first downsample these slices to a lower spatial resolution $\hat{H} \times \hat{W}$ for efficiency and forward them to f_S to obtain global slice-level feature vectors $\boldsymbol{z}_t^S$:

$$\boldsymbol{z}_t^S = f_S(\text{Down}(\boldsymbol{S}_t)), \quad t = 1, 2, \dots, T, \tag{1}$$

where t is the slice index and $\text{Down}(\cdot)$ denotes the spatial downsampling operation. On the contrary, f_P receives pathological patches $\{\boldsymbol{P}_1, \boldsymbol{P}_2, \dots, \boldsymbol{P}_T\}$ of size $D \times D$, which are cropped from $\{\boldsymbol{S}_1, \boldsymbol{S}_2, \dots, \boldsymbol{S}_T\}$ respectively, and generates fine-gained local feature vectors $\boldsymbol{z}_t^P$:

$$\boldsymbol{z}_t^P = f_P(\boldsymbol{P}_t), \quad t = 1, 2, \dots, T. \tag{2}$$

Localizer network $\mathcal{G}$ takes the global features $\boldsymbol{z}_t^S$ from slice network f_S as input, and directly predicts the location coordinates of salient pathological patches within sMRI slices. Such a design is feasible as previous studies [49–51] have shown that deep networks are capable of learning excellent representations for object localization even with weakly-supervised labels (i.e., the sMRI-level category labels in this work). Fig. 2 illustrates the architecture of our localizer network. Considering the context relationship across slices is necessary for accurate location prediction, we compose $\mathcal{G}$ with a one-layer long short-term memory (LSTM) module, one fully-connected layer parameterized by $\mathbf{W}_h$ and a sigmoid function σ. In this way, the information of previous slices is first accumulated to the hidden states of LSTM:

$$\boldsymbol{h}_t, \boldsymbol{c}_t = \text{LSTM}\left(\boldsymbol{z}_t^S, \boldsymbol{h}_{t-1}, \boldsymbol{c}_{t-1}\right), \tag{3}$$

where $\boldsymbol{h}_t$ and $\boldsymbol{c}_t$ are the hidden state and cell content of LSTM at slice step t. Then, $\boldsymbol{h}_t$ is mapped to a binary tuple by a non-linear transformation:

$$(u_t, v_t) = \sigma(\mathbf{W}_h \boldsymbol{h}_t), \tag{4}$$

where $u_t, v_t \in [0, 1]$. We treat (u_t, v_t) as the normalized top-left coordinate of pathological patch $\boldsymbol{P}_t$ in original slice $\boldsymbol{S}_t$. Assume that the top-left corner of $\boldsymbol{S}_t$ is the origin of the pixel coordinate system, where the x-axis and y-axis are defined from left-to-right and top-to-bottom, respectively. We can convert (u_t, v_t) to pixel coordinate (x_t, y_t) based on the given slice size $H \times W$ and patch size $D \times D$:

$$\begin{aligned} x_t &= (W - D)u_t, \\ y_t &= (H - D)v_t. \end{aligned} \tag{5}$$

Once the locations of pathological regions are hypothesized, corresponding local patches can be cropped from original slices and sent into patch network f_P to extract detailed disease-specific features for accurate AD diagnosis. However, the cropping operation is non-differentiable, causing the direct optimization issue that the gradient from f_P cannot be back-propagated to localizer network $\mathcal{G}$. This means that the network parameters of $\mathcal{G}$ cannot be updated through the standard backpropagation algorithm during training. To tackle this problem, we adopt bilinear interpolation to achieve approximate patch cropping, which fills the gap between f_P and $\mathcal{G}$, and elegantly makes AutoLoc an end-to-end learning framework. We defer the details of this technique to Section 2.2.

Classifier f_C is a common fully-connected layer that integrates the information from all slices and outputs the subject-level diagnosis result of the sMRI. To be specific, we concatenate the global slice feature vector $\boldsymbol{z}_t^S$ and the local patch feature vector $\boldsymbol{z}_t^P$ as a comprehensive representation of slice $\boldsymbol{S}_t$, then pass it to classifier f_C to generate slice-wise prediction, and finally average the predictions of all slices as the subject-level prediction:

$$\boldsymbol{p} = \text{Softmax}\left(\frac{1}{T}\sum_{t=1}^{T} f_C(\text{Concat}\left(\boldsymbol{z}_t^S, \boldsymbol{z}_t^P\right))\right), \tag{6}$$

where $\boldsymbol{p} \in \mathbb{R}^C$ refers to the probability scores for each class and C denotes the number of categories. $\text{Concat}(\cdot)$ and $\text{Softmax}(\cdot)$ represent the

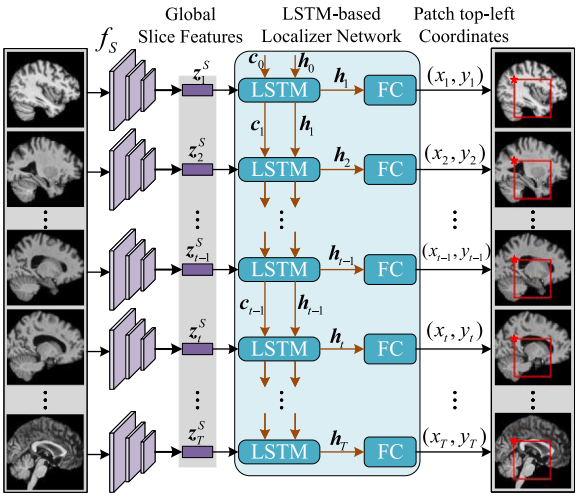

Fig. 2. Illustration of the localizer network. Taking the global slice features $\boldsymbol{z}_t^S$ as input, localizer network first models the context relationship among slices with a LSTM module, and then transforms the hidden states $\boldsymbol{h}_t$ to the top-left coordinates of the pathological patches using a non-linear mapping (consisting of a fully-connected layer (FC) and a sigmoid function).

concatenation operation and softmax function, respectively. Notably, such a design directly reuses the global features from slice network f_S to facilitate the diagnosis task. This differs from previous region localization-based diagnosis approaches, which require an additional branch [43] or build attention-based whole-brain classification networks [40,44] to extract global information, increasing the complexity and computation overhead of the pipeline.

2.2. Bilinear interpolation-based patch cropping

To remove the non-differentiable restriction of the cropping operation, we utilize the bilinear interpolation trick [49,51,52] to resample the localized pathological patch $\boldsymbol{P}_t$ from corresponding input slice $\boldsymbol{S}_t$. This derives a analytical representation between the pixel values of $\boldsymbol{P}_t$ and the predicted patch coordinate (x_t, y_t), enabling the gradient to be back-propagated to the localizer network. In the following, we first describe the coordinate parameterization of pixels within $\boldsymbol{P}_t$, then introduce the forward calculation and backward propagation of bilinear interpolation.

Coordinate Parameterization. According to Eqs. (4) and (5), localizer network $\mathcal{G}$ produces the top-left coordinate (x_t, y_t) of $\boldsymbol{P}_t$ in $\boldsymbol{S}_t$. Based on (x_t, y_t), the coordinates of other pixels in $\boldsymbol{P}_t$ can be easily formulated by adding a fixed offset constant:

$$(x_{t[ij]}, y_{t[ij]}) = (x_t, y_t) + \boldsymbol{c}_{ij}, \tag{7}$$

here, $(x_{t[ij]}, y_{t[ij]})$ represents the coordinate of the pixel point in ith row and jth column of $\boldsymbol{P}_t$, especially, $(x_{t[00]}, y_{t[00]})$ refers to the top-left coordinate (x_t, y_t). $\boldsymbol{c}_{ij} \in \{0, 1, \dots, D-1\}^2$ is the offset vector from (x_t, y_t) to $(x_{t[ij]}, y_{t[ij]})$, which reflects their relative position relationship. Establishing such a coordinate relationship between (x_t, y_t) and $(x_{t[ij]}, y_{t[ij]})$ is necessary for aggregating gradient information from other pixels to (x_t, y_t) in backpropagation.

Interpolation Calculation. Considering that $(x_{t[ij]}, y_{t[ij]})$ is a continuous real-valued coordinate, it is practically impossible to directly get corresponding pixel in $\boldsymbol{S}_t$. Instead, as shown in Fig. 3, we use the nearest four pixels around $(x_{t[ij]}, y_{t[ij]})$ in $\boldsymbol{S}_t$ to estimate its pixel value via bilinear interpolation, which is given by:

$$\begin{aligned}\boldsymbol{P}_{t[ij]} = \sum_{\alpha,\beta=0}^{1} &\left(1 - \left|\lfloor x_{t[ij]} \rfloor + \alpha - x_{t[ij]}\right|\right) \\ &\left(1 - \left|\lfloor y_{t[ij]} \rfloor + \beta - y_{t[ij]}\right|\right) \boldsymbol{S}_{t[mn]},\end{aligned} \tag{8}$$

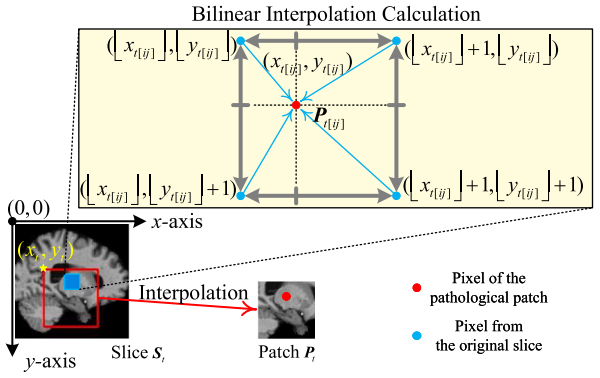

Fig. 3. Illustration of the bilinear interpolation-based patch-cropping operation. We use bilinear interpolation to compute the pixel values $\boldsymbol{P}_{t[ij]}$ of patch $\boldsymbol{P}_t$ from the nearest four pixels in slice $\boldsymbol{S}_t$, which constructs a differentiable analytical representation for the patch-cropping operation.

where $m = \lfloor y_{t[ij]} \rfloor + \beta$, $n = \lfloor x_{t[ij]} \rfloor + \alpha$. $\boldsymbol{P}_{t[ij]}$ represents the value of the pixel in ith row and jth column of $\boldsymbol{P}_t$, and similarly for $\boldsymbol{S}_{t[mn]}$. $|\cdot|$ and $\lfloor\cdot\rfloor$ denote absolute-value and rounding-down operations, respectively. In this way, we can generate the localized pathological patch $\boldsymbol{P}_t$ by traversing all possible i, j in Eq. (8).

Gradient Backpropagation. To allow backpropagation of the given loss $\mathcal{L}$ through the interpolation mechanism, we have to define the derivatives of $\mathcal{L}$ to (x_t, y_t), i.e., $\partial\mathcal{L}/\partial x_t$ and $\partial\mathcal{L}/\partial y_t$. Since they have similar forms, we take $\partial\mathcal{L}/\partial x_t$ as an example and calculate it by the chain rule, which is given by:

$$\frac{\partial \mathcal{L}}{\partial x_t} = \sum_{i,j} \frac{\partial \mathcal{L}}{\partial \boldsymbol{P}_{t[ij]}} \frac{\partial \boldsymbol{P}_{t[ij]}}{\partial x_t}, \tag{9}$$

here, $\partial\mathcal{L}/\partial\boldsymbol{P}_{t[ij]}$ represents the gradient of the interpolated pixel value $\boldsymbol{P}_{t[ij]}$ with respect to training loss $\mathcal{L}$, which can be directly derived from slice network f_P in backpropagation. $\partial\boldsymbol{P}_{t[ij]}/\partial x_t$ indicates the backward propagation of bilinear interpolation, and can be solved with Eqs. (7) and (8) as following:

$$\begin{aligned}\frac{\partial \boldsymbol{P}_{t[ij]}}{\partial x_t} &= \frac{\partial \boldsymbol{P}_{t[ij]}}{\partial x_{t[ij]}} \\ &= \sum_{\alpha,\beta=0}^{1} (-1)^{\alpha+1} \left(1 - \left|\lfloor y_{t[ij]} \rfloor + \beta - y_{t[ij]}\right|\right) \boldsymbol{S}_{t[mn]}.\end{aligned} \tag{10}$$

By doing this, we successfully obtain the derivatives $\partial\mathcal{L}/\partial x_t$ and $\partial\mathcal{L}/\partial y_t$, such that the loss gradient can flow back to patch coordinates x_t and y_t, and therefore back to localizer network $\mathcal{G}$ given that (x_t, y_t) is the output of $\mathcal{G}$. This enables the parameters of $\mathcal{G}$ to be updated synchronously with other network parameters during backpropagation, which means that AutoLoc framework is fully end-to-end trainable.

2.3. Loss functions

Supervision Loss. Since only sMRI-level category labels are available and there are no location annotations (e.g., bounding boxes of pathological regions), the pathology localization task actually is weakly supervised by the classification signal. More specifically, we train AutoLoc to minimize the standard cross-entropy loss between the predictions of classifier f_C and ground truth over the training set $\mathcal{D}_{train}$:

$$\mathcal{L}_{cls} = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{y})\sim\mathcal{D}_{train}} \left[-\boldsymbol{y} \log(\mathcal{F}(\boldsymbol{X}; \boldsymbol{\Theta}_1))\right], \tag{11}$$

where $\boldsymbol{\Theta}_1 = \{\boldsymbol{\theta}_{f_S}, \boldsymbol{\theta}_{\mathcal{G}}, \boldsymbol{\theta}_{f_P}, \boldsymbol{\theta}_{f_C}\}$ and $\boldsymbol{\theta}_{name}$ represents the learnable network parameters. $(\boldsymbol{X}, \boldsymbol{y})$ is the training sMRI sample ($\boldsymbol{X} = \{\boldsymbol{S}_1, \boldsymbol{S}_2, \dots, \boldsymbol{S}_T\}$) with associated one-hot encoded category label vector.

To better guide the optimization of the two backbone networks (slice network f_S and patch network f_P), we further employ *deep*

supervision on AutoLoc. In specific, we use two fully-connected layers (parameterized by $\mathbf{W}_s$ and $\mathbf{W}_p$, respectively) to transform the outputs of f_S and f_P to logits, which are normalized to probability scores using a softmax function. Then, we directly compute the slice-wise and patch-wise classification loss by assigning the sMRI-level label to corresponding slices and patches:

$$\mathcal{L}_{dsup} = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{y})\sim\mathcal{D}_{train}}\left\{\sum_{t=1}^{T}\left[-\boldsymbol{y}\log(\mathcal{F}(\boldsymbol{S}_t;\boldsymbol{\Theta}_2)) - \boldsymbol{y}\log(\mathcal{F}(\boldsymbol{S}_t;\boldsymbol{\Theta}_3))\right]\right\}, \quad (12)$$

where $\boldsymbol{\Theta}_2=\{\theta_{f_S},\mathbf{W}_s\}$ and $\boldsymbol{\Theta}_3=\{\theta_{f_S},\theta_{\mathcal{G}},\theta_{f_P},\mathbf{W}_p\}$. By introducing such deep supervision, we can effectively facilitate the slice-level representation learning of f_S and f_P, and alleviate the gradient vanishing problem. The overall supervision loss to optimize our AutoLoc can be defined as:

$$\mathcal{L}_{sup} = \mathcal{L}_{cls} + \mathcal{L}_{dsup}. \quad (13)$$

Regularization. Notably, patch network f_P can only observe the patches specified by localizer network $\mathcal{G}$, which potentially makes f_P suffer from overfitting and even forces $\mathcal{G}$ to converge to sub-optimal solutions. Inspired by the training regularization strategy in [50,51], we adopt *random patch augmentation* to increase the input diversity of f_P. Specifically, in addition to the pathological patches selected by $\mathcal{G}$, we also randomly crop patches from the input slices according to a uniform distribution, and then feed them into f_P and f_C to get an additional supervision loss $\mathcal{L}_{rand}$ like Eq. (13). Thus, the final loss for training our AutoLoc framework is defined as:

$$\mathcal{L} = \frac{1}{2}(\mathcal{L}_{sup} + \mathcal{L}_{rand}). \quad (14)$$

3. Experiments

3.1. Experimental setup

Datasets and Pre-processing. Our experiments are mainly based on Alzheimer's Disease Neuroimaging Initiative database (ADNI, available at https://adni.loni.usc.edu/), where a total of 1336 T1-weighted baseline sMRI scans are collected. There are 419 participants with normal cognition (NC), 397 participants with AD dementia, 268 participants with stable MCI (sMCI) and 252 participants with progressive MCI (pMCI). If a MCI participant progressed to AD within 36 months from the baseline point, we label that participant as pMCI; otherwise, as sMCI. The detailed demographic information of these subjects is shown in Table 1. To harmonize all sMRI scans, we sequentially conduct bias-field correction, skull stripping and linear registration to the standard MNI-152 template using FSL software library [53]. The processed image has a size of 181 × 218 × 181 voxels with the spatial resolution of $1 \times 1 \times 1$ mm^3. We further exclude the uninformative slices in the margin, resulting in the final image size 96 × 168 × 160. All 3D images are normalized independently by subtracting the mean and dividing by the standard deviation (based on nonzero voxels).

Implementation Details. Considering the information redundancy between adjacent slices, it is unnecessary to process all slices of a sMRI. Therefore, we uniformly divide the sMRI into T=16 segments along the sagittal direction, and then randomly sample one slice from each segment as input during training. In the inference phase, we select the middle slice of each segment as input by default, which is sufficient to achieve good diagnosis performance. When exhaustive pathology analysis is required (as in Section 3.4), all slices of the sMRI will be fed into the trained model at the cost of huge computation. Besides, on-the-fly data augmentation strategies are deployed to avoid overfitting. Specifically, at the training stage, we augment input slices by random multi-scale cropping [54] followed with 224 × 224 ($H \times W$) resizing. As for the inference stage, we scale the short side of input slices to 256 while maintaining the aspect ratio, then center-crop them to 224 × 224 ($H \times W$). By default, we downsample the input slices to 96 × 96 ($\hat{H} \times \hat{W}$) as the input to slice network f_S, and set the size of cropped patches to 128 × 128 ($D \times D$).

Table 1
Demographic information of the studied subjects including Dataset, Group Type, Gender, Age and Mini-Mental State Examination (MMSE).

Dataset	Group Type	Gender (male/female)	Age (mean ± s.d.)	MMSE (mean ± s.d.)
ADNI	NC	185/234	72.96 ± 6.16	29.15 ± 1.08
	AD	220/177	75.11 ± 7.79	23.15 ± 2.28
	sMCI	160/108	72.40 ± 7.13	28.02 ± 1.68
	pMCI	142/110	74.08 ± 7.11	26.73 ± 1.79
AIBL	NC	19/29	70.69 ± 7.32	29.33 ± 0.97
	AD	25/30	73.75 ± 8.03	19.85 ± 5.51

To verify the versatility of AutoLoc framework, we implement it with two different DNN architectures: CNNs and vision Transformers. Specifically, we use MobileNetV2-75 (MN75) and ResNet-18 (RN18) as slice network f_S and patch network f_P in the CNN architecture, while deploying MobileVitV2-50 [55] (MV50) and PoolFormer-S12 [56] (PF12) instead in the Transformer architecture. If not specified, the hidden size of the LSTM module in localizer network $\mathcal{G}$ is set to 512, and three dropout layers with ratio of 0.5 are placed before the fully-connected classification layers of f_S, f_P and classifier f_C, respectively. We consider different optimization strategies for these two DNN architectures. CNN-based AutoLoc is trained using SGD optimizer with weight decay of $1e$-4 and momentum of 0.9. The initial learning rate is set to 0.001 for the localizer network and 0.01 for other components. As to the Transformer-based AutoLoc, we adopt AdamW optimizer with weight decay of 0.05 and initial learning rate of $1e$-4. In addition, we always initialize DNN backbones with available ImageNet pre-trained weights and train AutoLoc end-to-end for 100 epochs with a cosine decay learning rate scheduler.

All the experiments are conducted on a computing workstation equipped with 2 Intel Xeon CPU E5-2620 v4 (8-core 2.1 GHz), and an NVIDIA Quadro RTX6000 graphic card. Python (version 3.8) is used for software development. Deep learning models are developed using PyTorch (version 1.10), and the batch size for training is set to 16.

Data Splitting and Evaluation Metrics. We adopt the five-fold cross-validation protocol to evaluate our proposed method. That is, ADNI dataset is randomly and uniformly divided into five folds; three of them are used to train the model, and the remaining two folds are used for validation and testing, respectively. The metrics including accuracy (ACC), sensitivity (SEN), specificity (SPE) and the area under receiver operating characteristic curve (AUC) are calculated to quantify the classification performance. All metrics are reported as a mean across five folds of cross-validation along with the standard deviation.

3.2. Comparison with baselines

Baselines. We compare our approach with multiple 2D DNN baselines to verify the design of AutoLoc framework, including FullReso, GlobalFeat, LocalFeat and RandLoc. FullReso method directly applies a high-capacity DNN to original *full-resolution* (224 × 224) slices for AD diagnosis, consuming heavy computation. Concretely, in the CNN architecture, FullReso is instantiated with ResNet-18/34 networks, termed as $\text{FullReso}_{\text{RN18}}$ and $\text{FullReso}_{\text{RN34}}$ respectively; in the Transformer architecture, FullReso is instantiated using PoolFormer-S12/24 networks, termed as $\text{FullReso}_{\text{PF12}}$ and $\text{FullReso}_{\text{PF24}}$ respectively. GlobalFeat only uses the light-weight slice network of AutoLoc to extract coarse *global features* from low-resolution (96 × 96) slices for AD diagnosis. LocalFeat uses our AutoLoc framework but stops the global information before the classifier, making only *local* pathological *features* contribute to disease recognition. For RandLoc, we use AutoLoc framework, but discard the localizer network and *randomly* sample patch *locations* from a uniform distribution instead.

Table 2
Results (mean ± s.d.) of AutoLoc and baseline methods on ADNI test set for AD classification (AD *vs.* NC) and MCI conversion prediction (pMCI *vs.* sMCI) tasks.

Method	ACC (%)	SEN (%)	SPE (%)	AUC (%)
AD *vs.* NC, CNN-Architecture[a]				
FullReso$_{RN18}$	91.66 ± 2.15	90.67 ± 4.66	92.60 ± 1.58	96.22 ± 1.30
FullReso$_{RN34}$	91.79 ± 1.02	90.94 ± 3.33	92.61 ± 1.89	96.33 ± 1.40
GlobalFeat	88.85 ± 1.41	85.87 ± 3.85	91.66 ± 3.18	94.25 ± 1.77
LocalFeat	92.52 ± 2.33	90.93 ± 4.56	**94.03** ± 0.75	96.32 ± 1.88
RandLoc	90.73 ± 1.27	89.46 ± 3.23	91.94 ± 1.47	95.71 ± 1.33
AutoLoc	**93.38** ± 1.63	**92.94** ± 3.37	93.80 ± 0.89	**96.70** ± 1.66
pMCI *vs.* sMCI, CNN-Architecture[a]				
FullReso$_{RN18}$	78.83 ± 2.48	80.11 ± 6.43	77.64 ± 3.32	85.92 ± 1.49
FullReso$_{RN34}$	79.99 ± 2.25	83.31 ± 4.73	76.87 ± 4.20	86.35 ± 3.65
GlobalFeat	75.78 ± 2.51	77.72 ± 5.72	73.97 ± 6.24	83.09 ± 2.97
LocalFeat	79.80 ± 2.16	80.94 ± 1.17	**78.73** ± 4.53	86.27 ± 1.81
RandLoc	77.35 ± 2.56	78.70 ± 5.33	76.09 ± 5.91	85.37 ± 1.78
AutoLoc	**81.12** ± 2.12	**83.69** ± 4.18	78.71 ± 2.66	**87.34** ± 1.28
AD *vs.* NC, Transformer-Architecture[b]				
FullReso$_{PF12}$	91.30 ± 1.01	90.67 ± 4.59	91.89 ± 2.53	96.30 ± 1.55
FullReso$_{PF24}$	92.28 ± 3.11	91.18 ± 4.17	**93.32** ± 2.77	96.51 ± 1.20
GlobalFeat	87.50 ± 2.45	90.43 ± 1.88	84.73 ± 4.20	93.76 ± 1.88
LocalFeat	92.15 ± 1.46	**93.19** ± 4.51	91.18 ± 3.05	96.45 ± 1.48
RandLoc	90.64 ± 1.52	90.62 ± 5.50	90.66 ± 3.86	95.94 ± 1.47
AutoLoc	**92.40** ± 1.90	91.94 ± 5.28	92.85 ± 3.10	**96.70** ± 1.17
pMCI *vs.* sMCI, Transformer-Architecture[b]				
FullReso$_{PF12}$	78.25 ± 2.17	80.54 ± 4.33	76.09 ± 4.48	86.31 ± 1.05
FullReso$_{PF24}$	79.41 ± 2.56	82.91 ± 3.36	76.13 ± 6.51	87.05 ± 2.44
GlobalFeat	75.18 ± 1.83	78.49 ± 6.99	72.08 ± 4.97	82.72 ± 1.31
LocalFeat	79.24 ± 1.69	**84.12** ± 2.20	74.64 ± 4.42	86.44 ± 1.15
RandLoc	77.67 ± 1.98	82.18 ± 5.65	73.45 ± 7.39	85.40 ± 2.35
AutoLoc	**80.38** ± 2.33	83.31 ± 2.83	**77.64** ± 3.91	**87.15** ± 0.96

[a]In the CNN architecture, we implement AutoLoc with MobileNetV2-75 and ResNet-18, FullReso$_{RN18}$ with ResNet-18 (RN18), and FullReso$_{RN34}$ with ResNet-34 (RN34).
[b]In the Transformer architecture, we implement AutoLoc with MobileVitV2-50 and PoolFormer-S12, FullReso$_{PF12}$ with PoolFormer-S12 (PF12), and FullReso$_{PF24}$ with PoolFormer-S24 (PF24).

Classification Performance. Table 2 shows the classification results of different methods with CNN and Transformer architectures. We can see that AutoLoc outperforms RandLoc by a large margin on both two diagnosis tasks, which demonstrates the effectiveness of our localizer network in identifying salient pathological locations. Taking CNN architecture as an example (as well as in the rest of this section), "Random Localizer" only reaches 90.73% ACC and 95.71% AUC on AD classification, whereas AutoLoc using learned pathological locations can reach 93.38% and 96.70% respectively. Similarly, in the MCI conversion prediction task, our AutoLoc also achieves higher ACC (81.12% *vs.* 77.35%) and AUC (87.34% *vs.* 85.37%) than RandLoc.

With no surprises, FullResos can serve as strong baselines thanks to the full-resolution input and their powerful feature extraction capability. But it is noticeable that AutoLoc's classification performance is consistently better than FullReso baselines. For example, using the same ResNet-18 backbone, our AutoLoc outperforms FullReso$_{RN18}$ by 1.72% ACC and 0.48% AUC for AD classification, and 2.29% ACC and 1.42% AUC for MCI conversion prediction. Meanwhile, when compared to FullReso$_{RN34}$ equipped with the deeper backbone ResNet-34, our method still shows superiority over it. This implies that there may exist misleading noise regions in original slices, which interfere with FullResos' decisions and degrade their performance. However, AutoLoc can learn to locate the most informative local regions to avoid such interference for better diagnosis.

To further verify the above conjecture, we investigate the classification performance of global features (GlobalFeat) and local features (LocalFeat) individually. As shown in Table 2, the performance of

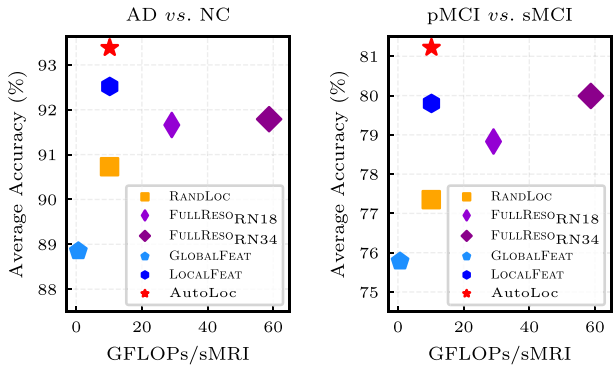

Fig. 4. Trade-off between computation overhead (measured by GFLOPs) and diagnosis accuracy for different models on ADNI test set.

GlobalFeat is not satisfactory, but LocalFeat surpasses FullReso$_{RN18}$ and achieves similar results to FullReso$_{RN34}$. This confirms that discriminative local representations are a critical factor affecting AD-related diagnosis performance, which is consistent with our conjecture. In addition, we can see that LocalFeat does not perform as well as AutoLoc due to the lack of global information. Therefore, it is equally essential to take into account global information in the design of sMRI-based CAD systems if we pursue higher classification performance.

Notably, above conclusions can also be drawn from using Transformer architecture, which shows that our framework is model-agnostic. Overall, the performance of Transformer architecture is slightly inferior to that of CNN architecture (we thus select the CNN architecture as the basis for follow-up analyzes). The possible reason is that vision Transformers are more hungry for training data and cannot generalize well when trained on small amounts of labeled data, as observed in [26]. This may be mitigated by leveraging self-supervised learning methods such as multi-view contrast [57], and we leave it as an interesting future work.

Efficiency *vs.* Accuracy. Fig. 4 presents the efficiency and accuracy trade-off, where we report the GFLOPs (giga floating point operations) used per sMRI (containing 16 uniformly sampled slices) at the inference stage to reflect the computation overhead. We can see that localizer-based approaches (AutoLoc and LocalFeat) can achieve high accuracy with few GFLOPs, whereas it is hard for other localizer-free baselines to improve the diagnosis performance while maintaining a low computation cost. Particularly, AutoLoc outperforms FullReso$_{RN18}$ and FullReso$_{RN34}$, the leading competitors among localizer-free baselines, with 2.85× and 5.75× less computation overhead, respectively. The gain in diagnosis accuracy is mainly attributed to the accurate localization of pathological regions, and the significant savings in computation budget are attributed to the fact that the high-capacity backbone in localizer-based approaches only processes very small task-relevant patches (e.g., 128 × 128 in our default setting). Taking ResNet-18 backbone as an example, inferring a 128 × 128 image patch only requires about 32% computation cost of processing the original full-resolution 224 × 224 slice. Besides, compared to RandLoc, AutoLoc only adds as small as 0.059 GFLOPs (~0.6% relative) but leads to at least 2% accuracy improvement, which demonstrates the good cost effectiveness of our localizer network. Compared to LocalFeat, AutoLoc brings almost no additional computation overhead but yields at least 1% accuracy improvement, which implies that our framework provides a cheap but effective manner to reuse global information.

Visualization of Semantic Features. To more intuitively understand the feature extraction capabilities of FullReso, GlobalFeat, LocalFeat and AutoLoc, we adopt t-distributed stochastic neighbor embedding (t-SNE) algorithm [58] to visualize their semantic feature (derived from the final fully-connected classification layer) distributions on the test set. Fig. 5 illustrates the two-dimensional t-SNE results, where each

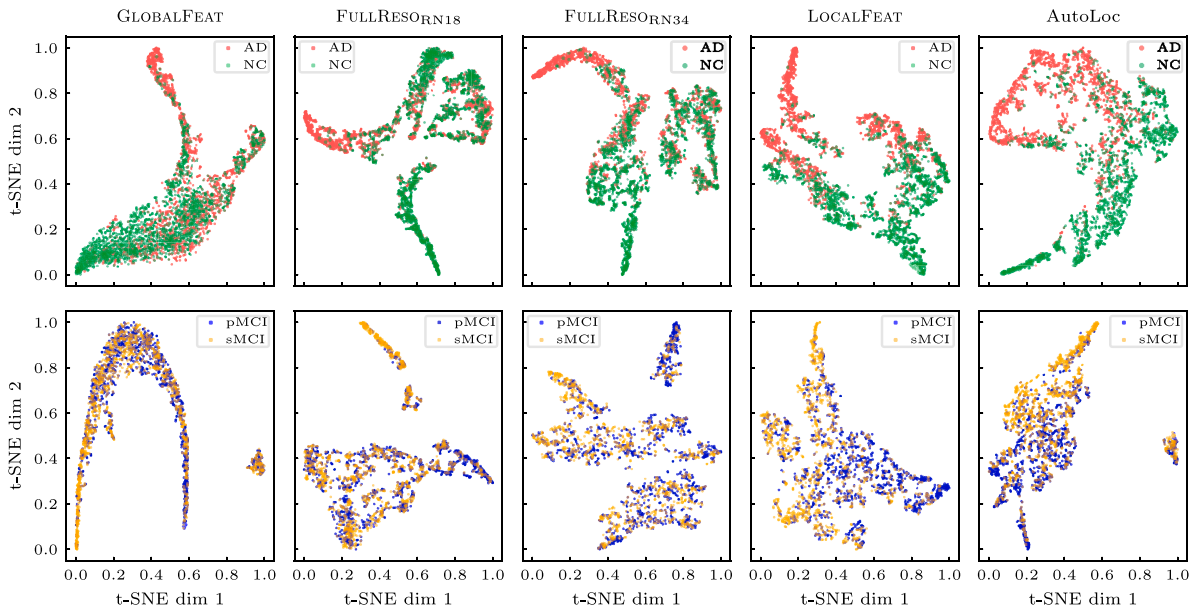

Fig. 5. Two-dimensional t-SNE visualization of the semantic features from different models (each column) on ADNI test set. The top row corresponds to AD *vs*. NC task, and the bottom row corresponds to pMCI *vs*. sMCI task.

Table 3
Results (mean ± s.d.) of AutoLoc and baseline methods on AIBL dataset for AD classification (AD *vs*. NC) task.

Method	ACC (%)	SEN (%)	SPE (%)	AUC (%)
CNN-Architecture[a]				
FullReso$_{RN18}$	89.32 ± 0.69	83.40 ± 1.31	**96.21** ± 0.96	96.37 ± 0.51
FullReso$_{RN34}$	90.49 ± 0.81	86.67 ± 1.82	94.96 ± 1.11	96.01 ± 0.75
GlobalFeat	85.44 ± 1.82	79.81 ± 3.46	92.04 ± 2.64	92.62 ± 1.25
LocalFeat	90.29 ± 1.37	85.92 ± 2.30	95.39 ± 0.85	97.03 ± 0.32
RandLoc	87.38 ± 0.99	84.37 ± 2.10	90.78 ± 1.50	94.23 ± 1.36
AutoLoc	**91.46** ± 1.06	**87.73** ± 2.31	95.79 ± 1.52	**97.11** ± 0.37
Transformer-Architecture[b]				
FullReso$_{PF12}$	87.96 ± 1.30	84.37 ± 2.86	92.16 ± 2.06	95.27 ± 0.79
FullReso$_{PF24}$	89.13 ± 1.99	84.77 ± 3.19	94.21 ± 1.60	96.40 ± 0.96
GlobalFeat	84.66 ± 1.06	82.18 ± 2.37	87.65 ± 2.71	93.73 ± 1.27
LocalFeat	89.32 ± 2.28	85.47 ± 3.47	93.79 ± 1.98	96.75 ± 0.59
RandLoc	85.44 ± 2.39	79.82 ± 5.45	92.32 ± 2.29	94.09 ± 0.68
AutoLoc	**90.29** ± 2.83	**86.21** ± 4.42	**95.02** ± 1.14	**97.21** ± 0.57

[a]In the CNN architecture, we implement AutoLoc with MobileNetV2-75 and ResNet-18, FullReso$_{RN18}$ with ResNet-18 (RN18), and FullReso$_{RN34}$ with ResNet-34 (RN34).
[b]In the Transformer architecture, we implement AutoLoc with MobileVitV2-50 and PoolFormer-S12, FullReso$_{PF12}$ with PoolFormer-S12 (PF12), and FullReso$_{PF24}$ with PoolFormer-S24 (PF24).

data point represents a slice. As expected, the features of the two categories in pMCI *vs*. sMCI are more closer compared to those in AD *vs*. NC. This corresponds to the fact that pMCI and sMCI patients all have fewer disease-related brain atrophy patterns, leaving it challenging to accurately distinguish them. It is clear that the feature distributions of AutoLoc and LocalFeat are more separable than other methods on both tasks, which verifies that the local regions localized by our method are highly class-discriminative. The features extracted from GlobalFeat present weak ability to separate different categories, which matches their poor classification results in Table 2. As to the full-resolution representations from FullReso$_{RN18}$ and FullReso$_{RN34}$, we can see that using a deeper network can improve the feature discriminability but also increases the computation cost significantly as shown in Fig. 4, suggesting sub-optimal results.

Generalizability Evaluation on AIBL Dataset. To further evaluate the generalization ability of our method in the cross-dataset scenario, we collect 103 baseline sMRI scans from Australian Imaging Biomarkers and Lifestyle Study of Ageing database (AIBL, available at https://aibl.csiro.au/), including 48 NC participants and 55 AD participants. The detailed demographic information is shown in Table 1. We directly use AIBL dataset as an additional test set to evaluate the models trained on ADNI dataset, without any further fine-tuning on AIBL dataset. As shown in Table 3, AutoLoc still maintains a high diagnosis performance in both CNN and Transformer architectures, and also achieves superior results compared to other baseline methods. This demonstrates the good generalization ability of our method.

3.3. Comparison with state-of-the-art methods

For a comprehensive comparison with recent studies on AD classification and MCI conversion prediction tasks, we summarize several state-of-the-art results reported in the published papers using sMRI data of ADNI database in Table 4. Although these results are not directly comparable due to their different experimental settings (e.g., data preprocessing steps, number of subjects, and division of the dataset), the simple and broad comparison here can still suggest some empirical observations.

As shown in Table 4, we can see that CNN-based AutoLoc outperforms previous 2D-CNN methods by a large margin and yields similar results to 3D-CNN methods. In general, 2D-CNN methods offer worse performance than 3D-CNN methods owing to the omission of inter-slice context information. While in our AutoLoc framework, we capture such information through the LSTM module in localizer network and exploits it to assist in pathology localization, elevating the discriminative power of slice-level representations. Compared with recent Transformer-based diagnosis methods, our Transformer-based AutoLoc also shows better performance, which demonstrates that our framework is model-agnostic and allows flexible deployment of various ImageNet backbone networks for AD diagnosis. In addition, AutoLoc achieves competitive performance compared with current saliency map-based localization methods, confirming the effectiveness of our framework for joint pathology localization and disease classification. In particular, AutoLoc is a fully differentiable framework and requires only one-stage end-to-end training, which is considerably much simpler and more efficient than other localization methods. For example, the recent leading work [44] formulates pathology localization as a counterfactual reasoning task and solves it in the GAN framework. Despite the

Table 4
Comparison with state-of-the-art studies using sMRI data of ADNI for AD classification (AD *vs.* NC) and MCI conversion prediction (pMCI *vs.* sMCI) tasks. The best and second-best results are respectively marked in bold and underlined fonts.

Reference	Subjects (AD/NC/pMCI/sMCI)	AD *vs.* NC				pMCI *vs.* sMCI			
		ACC(%)	SEN(%)	SPE(%)	AUC(%)	ACC(%)	SEN(%)	SPE(%)	AUC(%)
2D CNN-based diagnosis methods									
Pan et al. 2020 [9]	237/262/115/173	84.0	–	–	92.0	62.0	–	–	59.0
Ebrahimi et al. 2021 [10]	225/225/-/-	91.0	94.0	88.0	–	–	–	–	–
Kang et al. 2021 [11]	187/229/138/181	90.4	93.9	83.8	89.7	63.5	57.6	64.3	62.5
Zhang et al. 2022 [13]	139/159/-/-	90.0	92.8	87.5	–	–	–	–	–
3D CNN-based diagnosis methods									
Cui et al. 2019 [23]	192/223/165/231	92.3	90.6	93.7	**97.0**	75.0	73.3	76.2	79.7
Zhao et al. 2021 [20]	-/-/205/465	–	–	–	–	**83.0**	71.1	84.9	85.4
Wu et al. 2022 [21]	384/389/-/-	91.3	88.3	94.2	94.6	–	–	–	–
Li et al. 2022 [22]	299/330/-/-	93.2	95.0	89.8	92.4	–	–	–	–
Transformer-based diagnosis methods									
Kushol et al. 2022 [27]	159/229/-/-	88.2	**95.6**	77.4	–	–	–	–	–
Jang et al. 2022 [29]	1612/3174/-/-	93.2	–	–	96.3	–	–	–	–
Saliency map-based pathology localization methods									
Lian et al. 2020 [36]	358/429/205/465	90.3	82.4	**96.5**	95.1	80.9	52.6	85.4	78.1
Lian et al. 2022 [43]	358/429/205/465	91.9	88.7	94.5	96.5	82.7	57.9	**86.6**	79.3
Zhu et al. 2021 [37]	389/400/172/232	92.4	91.0	93.8	96.5	80.2	77.1	82.6	85.1
Zhang et al. 2021 [40]	327/416/-/-	92.0	90.3	93.1	96.2	–	–	–	–
Oh et al. 2023 [44]	359/431/251/497	**94.9**	–	–	–	77.0	–	–	–
Proposed coordinate prediction-based pathology localization framework									
AutoLoc(Transformer)	397/419/252/268	92.4	91.9	92.9	96.7	80.4	83.3	77.6	87.2
AutoLoc(CNN)		93.4	92.9	93.8	96.7	81.1	**83.7**	78.7	**87.3**

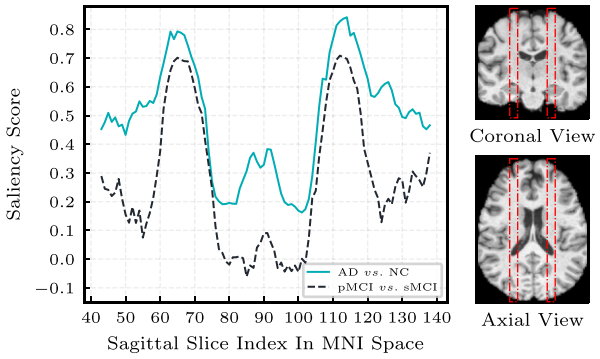

Fig. 6. Measured slice saliency distributions for AD *vs.* NC and pMCI *vs.* sMCI tasks by our AutoLoc on ADNI test set. The higher the saliency score of a slice, the more relevant it is to the diagnosis task. We mark out the top-20 salient slices in coronal and axial views using red dotted boxes.

excellent results, it is still limited to the post-hoc localization paradigm, leading to a multi-stage training pipeline and slow convergence.

3.4. Pathology region analysis

In this subsection, we present the discriminative pathological regions identified from the whole-brain sMRI using our AutoLoc framework. For this purpose, we first analyze the task-oriented discriminability of each slice and its pathology localization result, then integrate the pathological locations of the most discriminative (top-20) slices to obtain the disease-related brain regions. All analysis results are derived from the test set.

Slice Saliency Analysis. To quantitatively assess the ability of each slice to distinguish AD from NC, we calculate the difference of the average AD-class probability score between AD and NC groups to measure its saliency. Theoretically, the more discriminative a slice is, the more its average AD-class probability score converges to 1 for AD group and to 0 for NC group, hence the upper bound of the saliency score (difference) is 1 and conversely the lower bound is −1. Similarly for pMCI *vs.* sMCI. As shown in Fig. 6, the most salient slices mainly locate in the medial temporal lobe, where the atrophy of subregions (e.g., hippocampus) is an important diagnosis marker for at-risk AD subjects. Besides, the saliency distributions of the two tasks exhibit similar trends, but the saliency scores measured in pMCI *vs.* sMCI are consistently lower than those measured in AD *vs.* NC, which indicates that the dementia-induced structural changes in MCI brains are at an intermediate state between NC and AD. We can also observe that the saliency distributions show symmetry across the brain, which is in line with previous observations [59,60] that both right and left hemispheres are affected in the progression of AD.

Pathology Localization Analysis. We then analyze the localization results of our method by comparing with GradCAM algorithm, which is a post-hoc gradient-guided localization method and has been widely used in AD-related region localization. For GradCAM, we apply it to $\text{FullReso}_{\text{RN18}}$ to generate the saliency maps for pathology localization, since $\text{FullReso}_{\text{RN18}}$ uses the same high-capacity backbone as our AutoLoc. The average localization results of AD *vs.* NC and pMCI *vs.* sMCI are illustrated in Figs. 7(a) and 7(b), respectively. We can see that the identified pathological patches (red bounding boxes) using AutoLoc largely overlap with the highlighted areas in GradCAM-generated saliency maps, which verifies the effectiveness of our method for disease-related regions localization. In addition, it is intuitive to recognize more fine-grained structural atrophy within the localized discriminative patches by applying GradCAM to AutoLoc. As shown in Figs. 7(a) and 7(b), the saliency maps derived within our pre-identified pathological patches are more capable of capturing local and subtle abnormalities. For instance, AutoLoc-based GradCAM can attend to the ventricular regions in slices 75∼90, whose enlargement is an important biomarker in AD diagnosis [61]. This demonstrates the flexibility of our framework in locating pathological regions.

Identified Disease-related Brain Regions. We further mark out the major brain regions associated with AD by collecting the pathological areas from top-20 salient slices. As displayed in Fig. 8, the brain regions identified from AD classification and MCI conversion prediction are very similar and only partially different, which implies the high correlation of the two tasks. Specifically, they all include

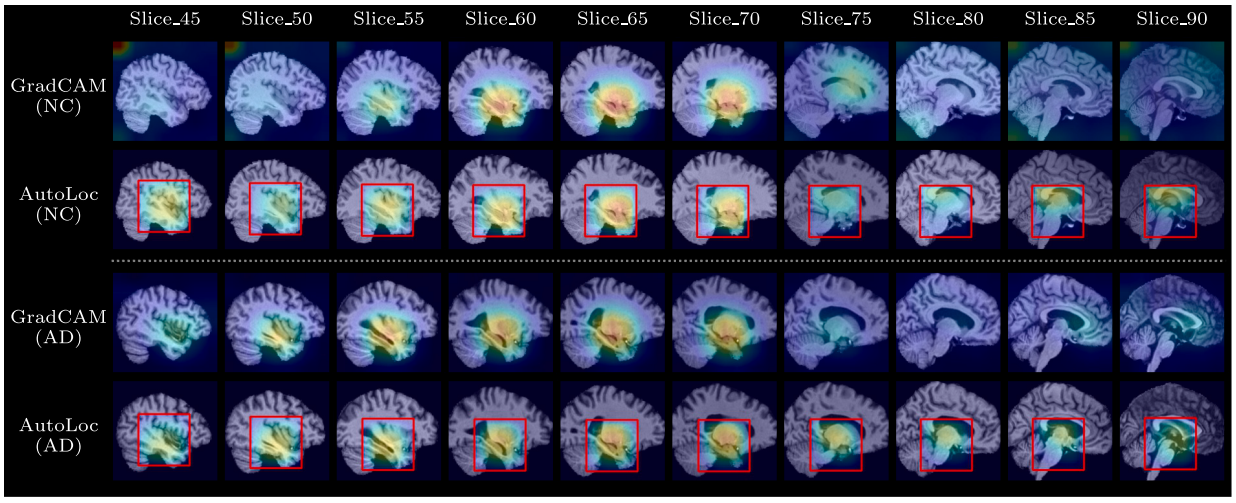

Fig. 7(a). Average pathology localization results of AutoLoc and GradCAM on ADNI test set for AD *vs*. NC task. The overlapped heatmap represents the localization results of FullReso$_{RN18}$-based GradCam, the red bounding box shows the location of the pathological patch (with the size of 128×128) predicted by our AutoLoc, and the highlighted areas within the red box are more fine-grained pathological structures identified by directly applying GradCAM to AutoLoc.

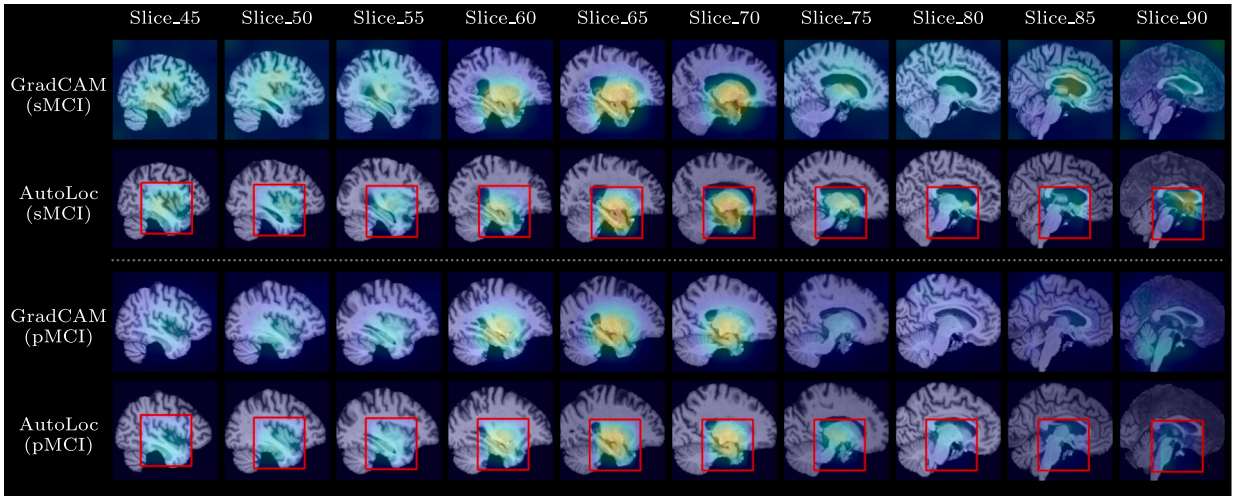

Fig. 7(b). Average pathology localization results of AutoLoc and GradCAM on ADNI test set for pMCI *vs*. sMCI task. The overlapped heatmap represents the localization results of FullReso$_{RN18}$-based GradCam, the red bounding box shows the location of the pathological patch (with the size of 128×128) predicted by our AutoLoc, and the highlighted areas within red box are more fine-grained pathological structures identified by directly applying GradCAM to AutoLoc.

rostral hippocampus, globus pallidus, amygdala, thalamus and area 35/36, i.e., parahippocampal gyrus. It is worth noting that the sMRI-based measurements of these brain regions have been reported to be effective for AD recognition in previous studies [9,37,43]. On the other hand, some of these regions (e.g., hippocampus and amygdala) play an essential role in episodic and spatial memory, which is closely correlated with the cognitive deficit symptoms of AD [62]. Besides, the dorsolateral putamen and entorhinal cortex are marked out for AD *vs*. NC and pMCI *vs*. sMCI respectively, which demonstrates that our method is able to adaptively localize task-oriented discriminative regions.

3.5. Ablation studies

To comprehensively evaluate our AutoLoc, we conduct in-depth ablation studies based on CNN architecture. The effect of each specific design of AutoLoc is investigated as follows.

Effectiveness of Loss Functions. We first explore the effectiveness of each component in our loss functions mentioned in Section 2.3, including classifier loss $\mathcal{L}_{cls}$, deep supervision loss $\mathcal{L}_{dsup}$ and random patch augmentation loss $\mathcal{L}_{rand}$. As shown in Table 6, training with $\mathcal{L}_{cls}$ alone leads to the worst classification performance (89.83% ACC and 95.25% AUC), which shows that the supervision signals from the final classifier is not enough to provide satisfactory optimization guidance. As expected, training by adding auxiliary supervision offers a marked rise in accuracy performance. Concretely, combining $\mathcal{L}_{cls}$ with $\mathcal{L}_{dsup}$ (2nd row) and $\mathcal{L}_{rand}$ (3rd row) yields 1.71% and 2.81% improvements in ACC, respectively. This is because $\mathcal{L}_{dsup}$ directly enhances the feature discriminability of the two backbone networks (slice network and patch network); $\mathcal{L}_{rand}$ prevents the patch network and localizer network from overfitting to local static information. Moreover, adding $\mathcal{L}_{rand}$ brings a more significant performance improvement than adding $\mathcal{L}_{dsup}$, which reveals that the coordinate prediction-based localization paradigm potentially suffers from non-negligible overfitting risk. Finally, training by adding both $\mathcal{L}_{dsup}$ and $\mathcal{L}_{rand}$ provides the best results, showing that these two strategies are mutually reinforcing in our framework.

Effectiveness of Localizer Network. In Table 7, we compare various choices of the localizer network, including random sampling (Random), multi-layer perceptron (MLP), LSTM and bidirectional LSTM

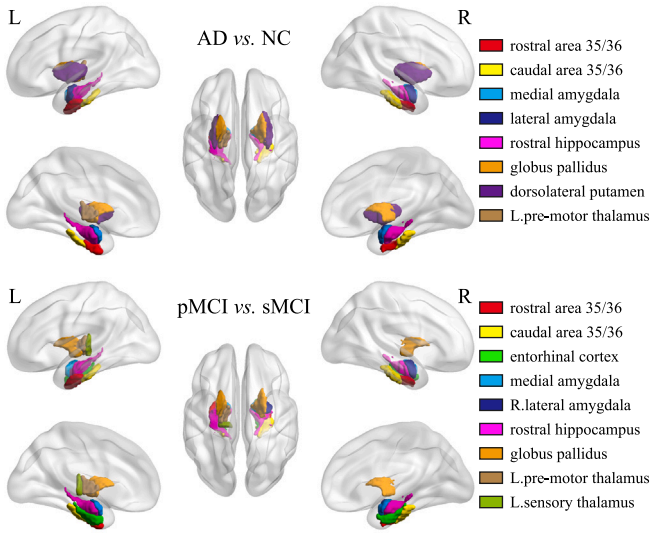

Fig. 8. Identified disease-related brain regions by our AutoLoc on AD *vs.* NC and pMCI *vs.* sMCI tasks.

Table 5
Results (mean ± s.d.) of AutoLoc using different slice axes on ADNI test set for AD classification task.

Axis	ACC (%)	SEN (%)	SPE (%)	AUC (%)
coronal view	90.68 ± 2.15	89.18 ± 3.86	92.12 ± 3.16	95.56 ± 1.32
axial view	91.91 ± 1.58	88.92 ± 3.01	**94.75** ± 1.21	95.85 ± 1.37
sagittal view	**93.38** ± 1.63	**92.94** ± 3.37	93.80 ± 0.89	**96.70** ± 1.66

Table 6
Ablation results (mean ± s.d.) of loss functions on ADNI test set for AD classification task.

$\mathcal{L}_{cls}$	$\mathcal{L}_{dsup}$	$\mathcal{L}_{rand}$	ACC (%)	SEN (%)	SPE (%)	AUC (%)
✓			89.83 ± 1.16	86.64 ± 1.75	92.85 ± 1.98	95.25 ± 1.51
✓	✓		91.54 ± 1.63	89.40 ± 4.16	93.56 ± 1.92	95.99 ± 1.67
✓		✓	92.64 ± 2.18	92.94 ± 3.82	92.36 ± 1.78	96.15 ± 1.75
✓	✓	✓	**93.38** ± 1.63	**92.94** ± 3.37	**93.80** ± 0.89	**96.70** ± 1.66

Table 7
Results (mean ± s.d.) of AutoLoc with different localizer networks on ADNI test set for AD classification task.

Localizer	ACC (%)	SEN (%)	SPE (%)	AUC (%)
Random	90.73 ± 1.27	89.46 ± 3.23	91.94 ± 1.47	95.71 ± 1.33
MLP	91.42 ± 0.42	90.92 ± 3.27	91.90 ± 2.94	95.90 ± 1.32
LSTM	**93.38** ± 1.63	**92.94** ± 3.37	**93.80** ± 0.89	**96.70** ± 1.66
BiLSTM	92.40 ± 2.23	91.94 ± 3.64	92.84 ± 1.09	96.31 ± 1.37

(BiLSTM). It is clear that all trainable localizers achieve considerably better results than random sampling, which verifies the effectiveness of our framework in optimizing the localization task. In addition, we observe that MLP localizer presents lower performance than LSTM localizer (ACC: 91.42% *vs.* 93.38%) and BiLSTM localizer (ACC: 91.42% *vs.* 92.40%). This can be attributed to the fact that MLP overlooks the exploitation of inter-slice context information, which reduces the receptive field and limits its performance. We also observe that BiLSTM localizer cannot further promote the diagnosis performance and even decreases the classification accuracy by around 1% compared to LSTM. One possible reason is that BiLSTM in our framework suffers from gradient vanishing problem, which makes optimization more difficult.

Influence of Cropped Patch Size. Table 8 summarizes the results of setting different pathological patch sizes for AutoLoc. One can see that

Table 8
Results (mean ± s.d.) of AutoLoc with different patch sizes ($D \times D$) on ADNI test set for AD classification (AD *vs.* NC) and MCI conversion prediction (pMCI *vs.* sMCI) tasks.

PatchSize	ACC (%)	SEN (%)	SPE (%)	AUC (%)
AD *vs.* NC				
96 × 96	91.04 ± 2.50	90.66 ± 5.93	91.41 ± 1.74	96.28 ± 1.57
128 × 128	**93.38** ± 1.63	**92.94** ± 3.37	**93.80** ± 0.89	**96.70** ± 1.66
160 × 160	92.90 ± 0.82	92.20 ± 2.27	93.56 ± 2.20	96.53 ± 1.69
pMCI *vs.* sMCI				
96 × 96	78.43 ± 4.31	82.88 ± 5.56	74.27 ± 6.50	84.97 ± 4.48
128 × 128	**81.12** ± 2.12	83.69 ± 4.18	**78.71** ± 2.66	**87.34** ± 1.28
160 × 160	79.59 ± 3.05	**84.08** ± 4.66	75.38 ± 2.39	87.20 ± 2.62

Table 9
Results (mean ± s.d.) of AutoLoc using different backbone capacities on ADNI test set for AD classification task.

Backbones ($f_S + f_P$)[a]	ACC (%)	SEN (%)	SPE (%)	AUC (%)
MV75 + MV75	90.07 ± 0.85	87.11 ± 4.74	92.85 ± 3.10	95.27 ± 1.63
RN18 + MV75	91.05 ± 1.28	88.89 ± 3.48	93.08 ± 1.38	96.29 ± 1.04
MV75 + RN18	**93.38** ± 1.63	**92.94** ± 3.37	93.80 ± 0.89	**96.70** ± 1.66
RN18 + RN18	92.89 ± 0.62	91.43 ± 2.60	**94.28** ± 2.30	96.40 ± 1.40

[a] f_S and f_P denote the slice network and the patch network in AutoLoc, respectively. MV75 and RN18 denote the low-capacity MobileNetV2-75 and the high-capacity ResNet-18, respectively.

the best results are obtained with the intermediate patch size of 128, too small or too large size will result in worse performance. For example, when compared to the patch size of 96 and 160, $D = 128$ improves the ACC by 2.34% and 0.48% on AD classification task, and leads to an ACC increase of 2.69% and 1.53% on MCI conversion prediction task. This is a reasonable phenomenon, as in these cases, smaller patches may fail to include complete pathological areas and miss some vital discriminative information, while larger patches potentially introduce noisy and misleading regions that might pollute the slice representation.

Influence of the Capacity of Backbone Networks. Slice network f_S and patch network f_P are two backbone networks in AutoLoc, responsible for extracting basic global slice information and detailed pathological patch information, respectively. Table 9 compares the performance of AutoLoc using different capacities of f_S and f_P. We can find that configuring a low-capacity network for f_P leads to a significant drop in diagnosis performance, for instance, the diagnosis accuracy of MV75 + MV75 and RN18 + MV75 is 3.31% and 2.33% lower than that of MV75 + RN18, respectively. This suggests that a low-capacity f_P is not sufficient to capture the subtle AD-induced atrophy that is crucial for accurate AD diagnosis, and a high-capacity network is better suited for f_P. Additionally, we observe that configuring two high-capacity networks (RN18 + RN18) for f_S and f_P in AutoLoc cannot further improve the diagnosis performance. Instead, it slightly decreases the diagnosis accuracy compared to MV75 + RN18, indicating the risk of overfitting. Therefore, configuring a low-capacity network for f_S and a high-capacity network for f_P can achieve a good balance between learning discriminative pathological representations and mitigating the risk of overfitting.

3.6. Limitations and future work

Although our proposed AutoLoc has achieved good performance in pathology localization and AD-related diagnosis tasks, its performance and practical value could be further improved in the future by carefully dealing with the following limitations or challenges. First, in our current implementation, the size of pathological patches is fixed and equivalent across sMRI slices. However, the structural changes caused by brain atrophy may present different scales in different slices. In the

future work, we could potentially make our framework more flexible by developing a multi-scale patch-size mechanism. Second, the random patch augmentation strategy currently used to regularize AutoLoc may introduce false positive samples, which can pollute the data representation and impair the diagnosis performance. Therefore, exploring more efficient and effective augmentation strategies to mitigate the overfitting risk of AutoLoc will be essential for further performance improvement. Third, we only consider the class-balanced data scenario in our current study. However, given that clinical data tends to be unbalanced, extending our approach to the more challenging class-unbalanced scenario will help to improve its generalizability. Finally, our current validation on ADNI and AIBL datasets is still limited. To facilitate the clinical translation, our next step is to collect more AD-related data to further validate our method. Moreover, it is interesting to adapt AutoLoc to other brain diseases, which will help the research community to establish a unified pathology localization and diagnosis framework.

4. Conclusion

In this paper, we have proposed an end-to-end automatic localization (AutoLoc) framework for joint pathology detection and AD diagnosis using sMRI, involving two key mechanisms: coordinate prediction-based localization and bilinear interpolation-based cropping. The former models the localization problem as predicting the coordinates of pathological regions instead of generating dense saliency maps as previous studies, which naturally avoids the need for voxel-level annotations and thus reduces the difficulty of optimizing the localization task under weak supervision. The latter approximates the non-differentiable patch-cropping operation with the bilinear interpolation trick, which ensures the gradient backpropagation between localization and diagnosis tasks, thereby enabling our framework to perform end-to-end joint training. We evaluated the validity and generalization of our framework on ADNI and AIBL datasets for AD classification and MCI conversion prediction. The experimental results demonstrated the effectiveness of our approach in localizing disease-related pathological regions, which leads to more discriminative data representations and superior classification performance than localization-free methods. When compared to saliency map-based localization approaches, our proposed localization paradigm also presented better or comparable diagnosis performance, while offering a considerably simpler and more efficient training pipeline. Moreover, we conducted detailed pathology localization analysis, and identified several important AD-associated brain regions, including rostral hippocampus, globus pallidus, amygdala, thalamus, parahippocampal gyrus, dorsolateral putamen and entorhinal cortex. These regions confirm previous findings and also provide new insights into the underlying mechanisms of AD pathology. Finally, in-depth ablation studies illustrated the effect of each specific design of AutoLoc, which we hope could provide new ideas for the research community to construct more accurate and efficient localization-diagnosis frameworks.

CRediT authorship contribution statement

Gongpeng Cao: Designed the research and experiments, Drafted the manuscript, Collected and processed the data. **Manli Zhang:** Designed the research and experiments, Drafted the manuscript, Collected and processed the data. **Yiping Wang:** Improved the ideas, Collected and processed the data. **Jing Zhang:** Improved the ideas, Revised the manuscript. **Ying Han:** Improved the ideas. **Xin Xu:** Improved the ideas. **Jinguo Huang:** Revised the manuscript. **Guixia Kang:** Revised the manuscript.

Declaration of competing interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by Fundamental Research Funds for the Central Universities (2020XD-A06-1), the State Key Program of the National Natural Science Foundation of China (82030037), the National Natural Science Foundation of China (62203063) and the China Postdoctoral Science Foundation (2022M710464).

All authors approved the version of the manuscript to be published.

References

[1] W. Jagust, Vulnerable neural systems and the borderland of brain aging and neurodegeneration, Neuron 77 (2) (2013) 219–234.
[2] 2022 Alzheimer's disease facts and figures, Alzheimer's Dementia 18 (4) (2022) 700–789, http://dx.doi.org/10.1002/alz.12638.
[3] R.C. Petersen, O. Lopez, M.J. Armstrong, T.S. Getchius, M. Ganguli, D. Gloss, G.S. Gronseth, D. Marson, T. Pringsheim, G.S. Day, M. Sager, J. Stevens, A. Rae-Grant, Practice guideline update summary: Mild cognitive impairment: Report of the guideline development, dissemination, and implementation subcommittee of the American academy of neurology, Neurology 90 (3) (2018) 126–135.
[4] W. Jagust, Imaging the evolution and pathophysiology of Alzheimer disease, Nat. Rev. Neurosci. 19 (11) (2018) 687–700.
[5] G. Lombardi, G. Crescioli, E. Cavedo, E. Lucenteforte, G. Casazza, A.-G. Bellatorre, C. Lista, G. Costantino, G. Frisoni, G. Virgili, G. Filippini, Structural magnetic resonance imaging for the early diagnosis of dementia due to Alzheimer's disease in people with mild cognitive impairment, Cochrane Database Syst. Rev. (3) (2020).
[6] T. Tong, R. Wolz, Q. Gao, R. Guerrero, J.V. Hajnal, D. Rueckert, Multiple instance learning for classification of dementia in brain MRI, Med. Image Anal. 18 (5) (2014) 808–818.
[7] J. Zhang, Y. Gao, Y. Gao, B.C. Munsell, D. Shen, Detecting anatomical landmarks for fast Alzheimer's disease diagnosis, IEEE Trans. Med. Imaging 35 (12) (2016) 2524–2533.
[8] M. Hon, N.M. Khan, Towards Alzheimer's disease classification through transfer learning, in: IEEE International Conference on Bioinformatics and Biomedicine, IEEE, 2017, pp. 1166–1169.
[9] D. Pan, A. Zeng, L. Jia, Y. Huang, T. Frizzell, X. Song, Early detection of Alzheimer's disease using magnetic resonance imaging: a novel approach combining convolutional neural networks and ensemble learning, Front. Neurosci. 14 (2020) 259.
[10] A. Ebrahimi, S. Luo, R. Chiong, Deep sequence modelling for Alzheimer's disease detection using MRI, Comput. Biol. Med. 134 (2021) 104537.
[11] 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.
[12] M. Tanveer, A.H. Rashid, M.A. Ganaie, M. Reza, I. Razzak, K.-L. Hua, Classification of Alzheimer's disease using ensemble of deep neural networks trained through transfer learning, IEEE J. Biomed. Health Inf. 26 (4) (2022) 1453–1463.
[13] Y. Zhang, Q. Teng, Y. Liu, Y. Liu, X. He, Diagnosis of Alzheimer's disease based on regional attention with sMRI gray matter slices, J. Neurosci. Methods 365 (2022) 109376.
[14] J.V. Shanmugam, B. Duraisamy, B.C. Simon, P. Bhaskaran, Alzheimer's disease classification using pre-trained deep networks, Biomed. Signal Process. Control 71 (2022) 103217.
[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.
[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 770–778.
[17] S. Basaia, F. Agosta, L. Wagner, E. Canu, G. Magnani, R. Santangelo, M. Filippi, Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks, NeuroImage Clin. 21 (2019) 101645.
[18] D. Jin, J. Xu, K. Zhao, F. Hu, Z. Yang, B. Liu, T. Jiang, Y. Liu, Attention-based 3D convolutional network for Alzheimer's disease diagnosis and biomarkers exploration, in: IEEE 16th International Symposium on Biomedical Imaging, IEEE, 2019, pp. 1047–1051.
[19] Z. Xia, G. Yue, Y. Xu, C. Feng, M. Yang, T. Wang, B. Lei, A novel end-to-end hybrid network for Alzheimer's disease detection using 3D CNN and 3D CLSTM, in: IEEE 17th International Symposium on Biomedical Imaging, IEEE, 2020, pp. 1–4.
[20] Y.-X. Zhao, Y.-M. Zhang, M. Song, C.-L. Liu, Region ensemble network for MCI conversion prediction with a relation regularized loss, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 185–194.
[21] 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. Health Inf. 26 (11) (2022) 5665–5673.

[22] J. Li, Y. Wei, C. Wang, Q. Hu, Y. Liu, L. Xu, 3-D CNN-based multichannel contrastive learning for Alzheimer's disease automatic diagnosis, IEEE Trans. Instrum. Meas. 71 (2022) 1–11.
[23] R. Cui, M. Liu, Hippocampus analysis by combination of 3-D DenseNet and shapes for Alzheimer's disease diagnosis, IEEE J. Biomed. Health Inf. 23 (5) (2019) 2099–2107.
[24] M. Liu, F. Li, H. Yan, K. Wang, Y. Ma, L. Shen, M. Xu, A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease, Neuroimage 208 (2020) 116459.
[25] S. Liu, C. Yadav, C. Fernandez-Granda, N. Razavian, On the design of convolutional neural networks for automatic detection of Alzheimer's disease, in: Proceedings of the Machine Learning for Health NeurIPS Workshop, PMLR, 2020, pp. 184–201.
[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2020.
[27] R. Kushol, A. Masoumzadeh, D. Huo, S. Kalra, Y.-H. Yang, Addformer: Alzheimer's disease detection from structural MRI using fusion transformer, in: IEEE 19th International Symposium on Biomedical Imaging, IEEE, 2022, pp. 1–5.
[28] J. Zhu, Y. Tan, R. Lin, J. Miao, X. Fan, Y. Zhu, P. Liang, J. Gong, H. He, Efficient self-attention mechanism and structural distilling model for Alzheimer's disease diagnosis, Comput. Biol. Med. 147 (2022) 105737.
[29] J. Jang, D. Hwang, M3T: Three-dimensional medical image classifier using multi-plane and multi-slice transformer, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2022, pp. 20686–20697.
[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[31] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer, 2014, pp. 818–833.
[32] B.H. van der Velden, H.J. Kuijf, K.G. Gilhuijs, M.A. Viergever, Explainable artificial intelligence (XAI) in deep learning-based medical image analysis, Med. Image Anal. 79 (2022) 102470.
[33] Y. Wang, Y. Yang, G. Cao, J. Guo, P. Wei, T. Feng, Y. Dai, J. Huang, G. Kang, G. Zhao, SEEG-Net: An explainable and deep learning-based cross-subject pathological activity detection method for drug-resistant epilepsy, Comput. Biol. Med. 148 (2022) 105703.
[34] 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.
[35] M. Liu, J. Zhang, E. Adeli, D. Shen, Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis, IEEE Trans. Bio-Med. Eng. 66 (5) (2019) 1195–1206.
[36] C. Lian, M. Liu, J. Zhang, D. Shen, Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI, IEEE Trans. Pattern Anal. Mach. Intell. 42 (4) (2020) 880–893.
[37] 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.
[38] M. Ashtari-Majlan, A. Seifi, M.M. Dehshibi, A multi-stream convolutional neural network for classification of progressive MCI in Alzheimer's disease using structural MRI images, IEEE J. Biomed. Health Inf. 26 (8) (2022) 3918–3926.
[39] S. Qiu, P.S. Joshi, M.I. Miller, C. Xue, X. Zhou, C. Karjadi, G.H. Chang, A.S. Joshi, B. Dwyer, S. Zhu, M. Kaku, Y. Zhou, Y.J. Alderazi, A. Swaminathan, S. Kedar, M.-H. Saint-Hilaire, S.H. Auerbach, J. Yuan, E.A. Sartor, R. Au, V.B. Kolachalama, Development and validation of an interpretable deep learning framework for Alzheimer's disease classification, Brain 143 (6) (2020) 1920–1933.
[40] Z. Zhang, L. Gao, G. Jin, L. Guo, Y. Yao, L. Dong, J. Han, THAN: task-driven hierarchical attention network for the diagnosis of mild cognitive impairment and Alzheimer's disease, Quant. Imaging Med. Surg. 11 (7) (2021) 3338–3354.
[41] B. Yan, Y. Li, L. Li, X. Yang, T.-q. Li, G. Yang, M. Jiang, Quantifying the impact of pyramid squeeze attention mechanism and filtering approaches on Alzheimer's disease classification, Comput. Biol. Med. 148 (2022) 105944.
[42] Q. Li, X. Xing, Y. Sun, B. Xiao, H. Wei, Q. Huo, M. Zhang, X.S. Zhou, Y. Zhan, Z. Xue, F. Shi, Novel iterative attention focusing strategy for joint pathology localization and prediction of MCI progression, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 307–315.
[43] C. Lian, M. Liu, Y. Pan, D. Shen, Attention-guided hybrid network for dementia diagnosis with structural MR images, IEEE Trans. Cybern. 52 (4) (2022) 1992–2003.
[44] K. Oh, J.S. Yoon, H.-I. Suk, Learn-explain-reinforce: Counterfactual reasoning and its guidance to reinforce an Alzheimer's disease diagnosis model, IEEE Trans. Pattern Anal. Mach. Intell. 45 (4) (2023) 4843–4857.
[45] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 2921–2929.
[46] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: IEEE International Conference on Computer Vision, IEEE, 2017, pp. 618–626.
[47] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harvard J. Law Technol. 31 (2017) 841.
[48] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 4510–4520.
[49] J. Fu, H. Zheng, T. Mei, Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 4476–4484.
[50] Y. Wang, K. Lv, R. Huang, S. Song, L. Yang, G. Huang, Glance and focus: a dynamic approach to reducing spatial redundancy in image classification, Adv. Neural Inf. Process. Syst. 33 (2020) 2432–2444.
[51] Y. Wang, Y. Yue, Y. Lin, H. Jiang, Z. Lai, V. Kulikov, N. Orlov, H. Shi, G. Huang, Adafocus v2: End-to-end training of spatial dynamic networks for video recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2022, pp. 20030–20040.
[52] M. Jaderberg, K. Simonyan, A. Zisserman, k. kavukcuoglu, Spatial transformer networks, Adv. Neural Inf. Process. Syst. 28 (2015).
[53] M. Jenkinson, C.F. Beckmann, T.E. Behrens, M.W. Woolrich, S.M. Smith, Fsl, Neuroimage 62 (2) (2012) 782–790.
[54] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, Temporal segment networks: Towards good practices for deep action recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 20–36.
[55] S. Mehta, M. Rastegari, Separable self-attention for mobile vision transformers, 2022.
[56] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, S. Yan, Metaformer is actually what you need for vision, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2022, pp. 10809–10819.
[57] G. Cao, Y. Wang, M. Zhang, J. Zhang, G. Kang, X. Xu, Multiview long-short spatial contrastive learning for 3D medical image analysis, in: IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2022, pp. 1226–1230.
[58] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008) 2579–2605.
[59] C.J. Jack, R. Petersen, Y. Xu, P. O'brien, G. Smith, R. Ivnik, B. Boeve, E. Tangalos, E. Kokmen, Rates of hippocampal atrophy correlate with change in clinical status in aging and AD, Neurology 55 (4) (2000) 484–490.
[60] P.M. Thompson, K.M. Hayashi, G. De Zubicaray, A.L. Janke, S.E. Rose, J. Semple, D. Herman, M.S. Hong, S.S. Dittmer, D.M. Doddrell, A.W. Toga, Dynamics of gray matter loss in Alzheimer's disease, J. Neurosci. 23 (3) (2003) 994–1005.
[61] S.M. Nestor, R. Rupsingh, M. Borrie, M. Smith, V. Accomazzi, J.L. Wells, J. Fogarty, R. Bartha, Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database, Brain 131 (9) (2008) 2443–2454.
[62] G.C. Schwindt, S.E. Black, Functional imaging studies of episodic memory in Alzheimer's disease: a quantitative meta-analysis, Neuroimage 45 (1) (2009) 181–190.