Original Article

THAN: task-driven hierarchical attention network for the diagnosis of mild cognitive impairment and Alzheimer's disease

Zhehao Zhang[1], Linlin Gao[1], Guang Jin[1], Lijun Guo[1], Yudong Yao[2], Li Dong[1], Jinming Han[3]; the Alzheimer's Disease NeuroImaging Initiative

[1]Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China; [2]Research Institute for Medical and Biological Engineering, Ningbo University, Ningbo, China; [3]Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

Contributions: (I) Conception and design: Z Zhang, L Gao; (II) Administrative support: L Gao; (III) Provision of study materials or patients: The Alzheimer's Disease NeuroImaging Initiative; (IV) Collection and assembly of data: Z Zhang; (V) Data analysis and interpretation: Z Zhang, L Gao; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Linlin Gao, PhD. Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315210, China. Email: gaolinlin@nbu.edu.cn.

Background: To assist doctors to diagnose mild cognitive impairment (MCI) and Alzheimer's disease (AD) early and accurately, convolutional neural networks based on structural magnetic resonance imaging (sMRI) images have been developed and shown excellent performance. However, they are still limited in their capacity in extracting discriminative features because of large sMRI image volumes yet small lesion regions and the small number of sMRI images.

Methods: We proposed a task-driven hierarchical attention network (THAN) taking advantage of the merits of patch-based and attention-based convolutional neural networks for MCI and AD diagnosis. THAN consists of an information sub-network and a hierarchical attention sub-network. In the information sub-network, an information map extractor, a patch-assistant module, and a mutual-boosting loss function are designed to generate a task-driven information map, which automatically highlights disease-related regions and their importance for final classification. In the hierarchical attention sub-network, a visual attention module and a semantic attention module are devised based on the information map to extract discriminative features for disease diagnosis.

Results: Extensive experiments were conducted for four classification tasks: MCI versus (*vs.*) normal controls (NC), AD *vs.* NC, AD *vs.* MCI, and AD *vs.* MCI *vs.* NC. Results demonstrated that THAN attained the accuracy of 81.6% for MCI *vs.* NC, 93.5% for AD *vs.* NC, 80.8% for AD *vs.* MCI, and 62.9% for AD *vs.* MCI *vs.* NC. It outperformed advanced attention-based and patch-based methods. Moreover, information maps generated by the information sub-network could highlight the potential biomarkers of MCI and AD, such as the hippocampus and ventricles. Furthermore, when the visual and semantic attention modules were combined, the performance of the four tasks was highly improved.

Conclusions: The information sub-network can automatically highlight the disease-related regions. The hierarchical attention sub-network can extract discriminative visual and semantic features. Through the two sub-networks, THAN fully exploits the visual and semantic features of disease-related regions and meanwhile considers global features of sMRI images, which finally facilitate the diagnosis of MCI and AD.

Keywords: Mild cognitive impairment (MCI); Alzheimer's disease (AD); information sub-network; hierarchical attention sub-network (HAS); structural magnetic resonance imaging (sMRI)

Submitted Jan 26, 2021. Accepted for publication Apr 26, 2021.
doi: 10.21037/qims-21-91
View this article at: http://dx.doi.org/10.21037/qims-21-91

Introduction

Alzheimer's disease (AD) is an irreversible and chronic neurodegenerative disease. It has been a leading cause of disability for the elderly over 65 years old (1). It has been reported that more than 33 million people worldwide were suffering from AD in 2018. The number is predicted to increase to 100 million by 2050. The cost of the disease was about 666 billion US dollars in 2018 and it is forecasted to double by 2030 (2). Mild cognitive impairment (MCI) is a prodromal stage of AD. Studies have shown that 20% of patients with MCI could deteriorate into AD within 4 years (3). Although there is no effective way to block the progression of MCI and AD (denoted as MCI/AD), some treatments have been developed to delay their progression. Therefore, it is becoming increasingly essential for the scientific community to develop effective methods to diagnose MCI/AD as early and accurately as possible.

Structural magnetic resonance imaging (sMRI) images are more useful for the early diagnosis of MCI/AD compared with the clinical assessment of cognitive impairment. This is because brain changes induced by AD have been proven to occur 10–15 years before symptom onset (4), and these brain changes can be non-invasively captured by sMRI images (5). Currently, sMRI images are extensively employed for computer-aided MCI/AD diagnosis based on machine learning methods (6-23). Among these methods, the convolutional neural network (CNN)-based methods (9-23) have demonstrated outstanding performance due to their excellent ability to extract task-driven features.

Existing CNN-based MCI/AD classification methods can be divided into four categories. They are regions-of-interest (ROI)-based methods, whole image-based methods, patch-based methods, and attention-based methods. The ROI-based methods first pre-segment disease-related regions according to the domain knowledge of AD experts. After that, different CNNs are designed to extract features from these regions and make final classification (9-11). However, the pre-segmented disease-related regions vary based on different experts and generally cannot cover all the lesion regions. Moreover, these methods require complex pre-processing steps. The whole image-based methods extract features from entire sMRI images directly (12). They require no expert knowledge and fully exploit the global features of an image. However, this type of method cannot accurately extract disease-related features, because the volume of each sMRI image is large while the lesion regions are small, leading to inaccurate classification. Overall, the ROI-based and whole image-based methods cannot automatically extract features from disease-related regions. The patch-based and attention-based methods relieve such a limitation to some extent.

The patch-based methods automatically select disease-related image patches first. After that, they extract features from these image patches and fuse these features to classify patients from normal controls (NCs). For example, image patches were first selected based on distinct anatomical landmarks in a data-driven manner (13-17). After that, the labels of image patches were assigned to the labels of their corresponding sMRI images. Further, in (13), multi-CNNs were trained for feature extraction and classification of these patches, and the majority voting strategy was used for whole image classification. In (14), an improved deep multi-instance learning network was designed to integrate patch-level features for final sMRI image classification. This network avoids the usage of patch-level labels. In (15,16), Liu *et al.* extended the deep multi-instance learning network on multi-modality data to improve the diagnosis of AD/MCI. Additionally, Lian *et al.* (17) further strengthened the correlation information of image patches by designing a hierarchical fully convolutional network (FCN) for MCI/AD diagnosis. Moreover, Qiu *et al.* (18) trained an FCN model to generate a participant-specific disease probability map based on randomly selected 3,000 image patches. They then selected high-risk voxels from the disease probability map, an implicit patch-selection strategy. Further, they trained a multilayer perceptron for AD diagnosis. These patch-based methods do not require expert knowledge. Moreover, they can better extract the visual features [i.e., low-level features such as luminance and edge (19), which are extracted from the shallow layers of CNNs] and semantic features [i.e., high-level features like objects (19), which are extracted from the deep layers of CNNs] of the disease-related image patches for classification. However, most of the methods completely neglect the remaining image patches, leading to dismissing of the global features of sMRI images.

The attention-based methods generate attention maps with different weights. By combining an attention map with the feature map from a certain layer, some features can be emphasized and meanwhile the other features are not neglected, either. Jin *et al.* (20) designed a weighted attention block inserted in 3D ResNet (24) to improve the ability in extracting features of disease-related regions. The weighted attention block was employed on a middle

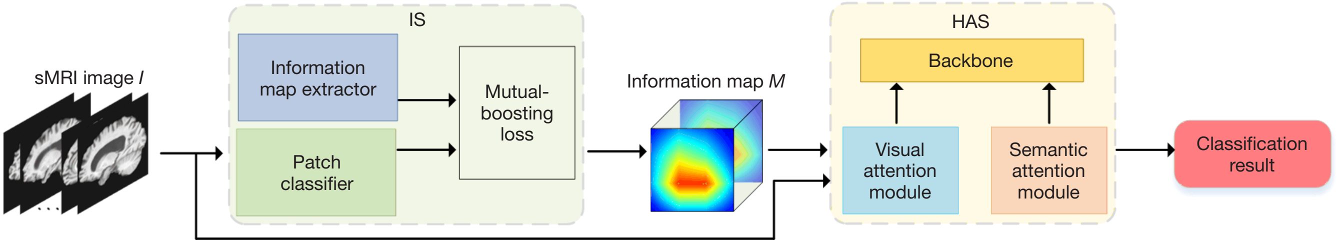

Figure 1 Structure of the task-driven hierarchical attention network (THAN). sMRI, structural magnetic resonance imaging; IS, information sub-network; HAS, hierarchical attention sub-network.

layer of 3D ResNet and the generation of the attention map has no direct constraint. Li *et al.* (21) proposed an iterative attention focusing strategy for the localization of pathological regions and for the classification between progressive and stable MCI. The iterative attention focusing strategy was utilized on a deep layer of each iterative sub-network. In addition, Lian *et al.* (22) constructed a dementia attention block to automatically identify subject-specific discriminative locations from whole sMRI images. This attention block was used on a deep layer of the network. Furthermore, Zhang *et al.* (23) developed a deep cross-modal attention network, which focused on learning the "deep relations" among different brain regions from diffusion tensor imaging and resting-state functional MRI. Overall, the attention-based methods require no expert knowledge, either. However, most attention maps are solely employed on the deep layers of networks for semantic feature extraction. This results in the visual features of disease-related regions not being well extracted and further affects the semantic feature extraction and final disease diagnosis. Even though a few attention-based methods utilize attention maps for visual feature extraction, these attention maps are generated without direct constraints. Thereby, they may be not closely related to disease regions.

To better extract features of disease-related regions and global features, we developed a task-driven hierarchical attention network (THAN) for MCI/AD diagnosis. THAN consists of an information sub-network (IS) and a hierarchical attention sub-network (HAS), as shown in *Figure 1*. IS can generate a task-driven information map that automatically highlights disease-related regions and their importance for final classification. It contains an information map extractor, a patch classifier, and a mutual-boosting loss function. The information extractor is assisted by the patch classifier to generate an effective information map. The mutual-boosting loss function aims to promote the performance of the information map extractor and the patch classifier. In HAS, a visual attention module and a semantic attention module are devised to successively enhance the extraction of discriminative features.

Methods

The proposed THAN consists of IS and HAS. IS aims to generate an effective information map. HAS makes use of the information map to produce hierarchical attention maps that guide the sub-network to extract discriminative visual and semantic features for final image classification. In this section, we first present IS and then discuss HAS. After that, the dataset and preprocessing steps are introduced. Moreover, evaluation metrics and implementation details are listed.

Information sub-network

IS, inspired by NTS-Net (25), is a mutual-boosting network, as shown in *Figure 2*. It is composed of two modules and a loss function. One module is an information map extractor and the other is a patch classifier. The goal of the information map extractor is to generate an effective information map with the assistance of the patch classifier. To be specific, the patch classifier first produces informative image patches based on the information map. It then tries to better classify these image patches and to assist the information map extractor to generate an effective information map. Moreover, the effective information map in turn improves the generation of informative image patches and the profound classification of these patches. The two modules boost each other in the above manner. Further, a mutual-boosting loss function is designed to combine the two modules and to promote IS to generate the final effective information map. The details of the

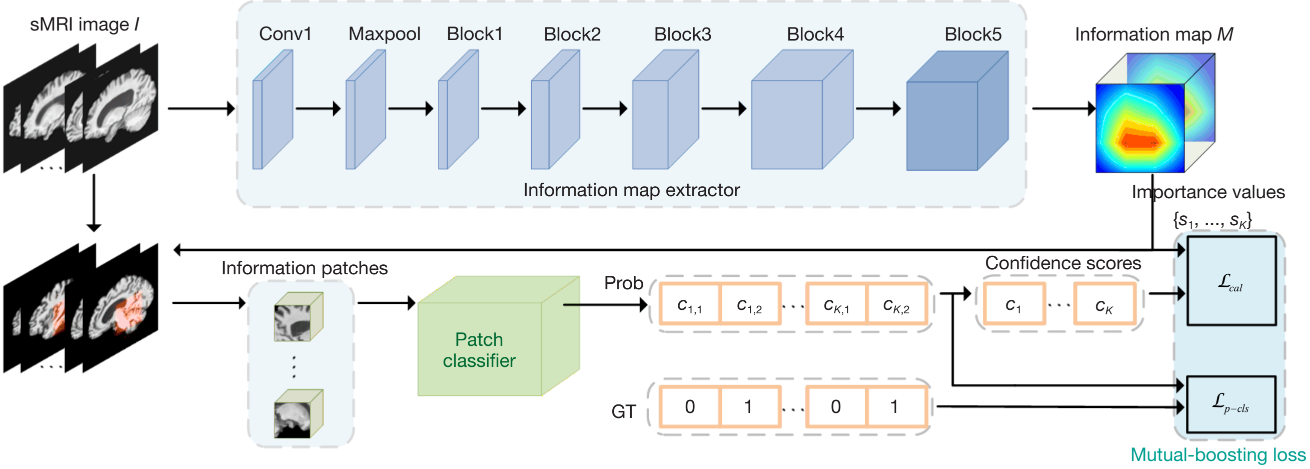

Figure 2 Framework of the information sub-network (IS). Ground truth (GT) is denoted in the one-hot format. sMRI, structural magnetic resonance imaging.

Table 1 Structures of networks. Global average pooling (GAP) is short for the global average pooling layer

Block names	Information map extractor *F*	3D ResNet10 (or 34)
Conv1	kernel size =7*7*7, channel =64, stride = (2, 2, 2)	
Max pooling	kernel size =3*3*3, stride = (2, 2, 2)	
Block1 (residual structure)	$\begin{bmatrix}3*3*3, 64\\ 3*3*3, 64\end{bmatrix}\times 2$	$\begin{bmatrix}3*3*3, 64\\ 3*3*3, 64\end{bmatrix}\times 1(3)$
Block2 (residual structure)	$\begin{bmatrix}3*3*3, 128\\ 3*3*3, 128\end{bmatrix}\times 2$	$\begin{bmatrix}3*3*3, 128\\ 3*3*3, 128\end{bmatrix}\times 1(4)$
Block3 (residual structure)	$\begin{bmatrix}3*3*3, 256\\ 3*3*3, 256\end{bmatrix}\times 2$	$\begin{bmatrix}3*3*3, 256\\ 3*3*3, 256\end{bmatrix}\times 1(6)$
Block4 (residual structure)	$\begin{bmatrix}3*3*3, 512\\ 3*3*3, 512\end{bmatrix}\times 2$	$\begin{bmatrix}3*3*3, 512\\ 3*3*3, 512\end{bmatrix}\times 1(3)$
Block5	$\begin{bmatrix}1*1*1, 256\\ 1*1*1, 128\\ 1*1*1, 1\end{bmatrix}\times 1$	–
	–	GAP and FC

FC, fully connected layer.

two modules and the mutual-boosting loss function are discussed below.

Information map extractor

Given an sMRI image *I*, an information map *M* is generated according to the information map extractor *F*, denoted as

$$M = F(I) \qquad [1]$$

where $M \in \mathrm{R}^{L*W*H}$, *L*, *W*, and *H* represent the length, width, and height of *M*, respectively.

The structure of *F* is shown in *Figure 2*. It is stacked by Conv1, Maxpool, Block1, Block2, Block3, Block4, and Block5. The layers from Conv1 to Block4 aim to extract discriminative features from *I*. They have the same structure as 3D ResNet18. Block5 is developed to gradually map the feature map with *N* channels to the information map *M* with only one channel. It is made up of multiple transition layers. The specific structure of *F* is displayed in the middle column of *Table 1*.

Patch classifier

An image patch in an original image can be downsampled to a value in a feature map through several layers and with certain scales. Conversely, each value in a feature map can be mapped to an image patch in the original image through the same layers yet with reversed scales. The mapping relations are shown in *Figure 3*. Since *M* is obtained by applying 32:1 spatial downsampling to *I* through *F*, each value $s_k \in M$ ($k \in \{1, 2, ..., L \times W \times H\}$) corresponds to an image patch p_k in *I*, where p_k is with the size of 32*32*32. We describe s_k as the importance value of p_k for final classification because of the semantic character of *M*.

Given a feature map with two channels, it becomes a 2D vector through the global average pooling (GAP) layer.

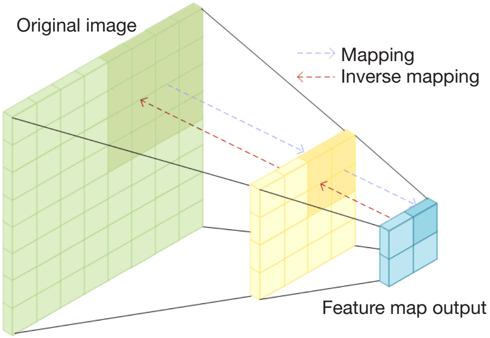

Figure 3 Mapping relations.

The 2D vector then produces a binary classification result after passed the Softmax layer. In this process, the highest value in the 2D vector determines the classification result. That is, the highest average value of the two channels in the feature map determines the classification result. This implies that the points with higher values in a feature map make more positive contributions to final classification. Based on this theory of deep convolutional neural networks (DCNNs), the points with higher values in a feature map are selected and highlighted for better classification. For *M*, the points with higher s_k make more contributions to final classification. That means that the higher s_k is, the more important p_k is.

K largest importance values are selected from *M*, denoted as $\{s_1, ..., s_k, ..., s_K\}$. Then the corresponding *K* image patches are obtained from *I*, denoted as $\{p_1, ..., p_k, ..., p_K\}$. Therefore, they are the *K* most informative patches in *I* and are most likely related to the category of *I*. Their labels are hence assigned to the label of *I*. We utilize one 3D ResNet10 to classify these *K* patches. The structure of 3D ResNet10 is shown in the last column of *Table 1*.

Mutual-boosting loss function

The mutual-boosting loss function consists of a patch-level classification loss function and a calibrating loss function.

The patch-level classification loss function is used to improve the classification performance of the *K* informative patches. It is realized by the widely used classification loss function, i.e., cross-entropy loss,

$$\mathcal{L}_{p-cls} = -\frac{1}{K}\sum\nolimits_{k=1}^{K}\sum\nolimits_{n=1}^{N} y_{k,n}\log\left(c_{k,n}\middle|p_k\right) \qquad [2]$$

where $y_{k,n}$ is the one-hot format of the ground truth of the patch p_k; $c_{k,n}$ represents the probability that p_k belongs to the *n*th category ($n \in \{1, 2, \dots, N\}$).

In the output of p_k, each element (i.e., $c_{k,n}$) indicates the probability that p_k belongs to the corresponding category. Specifically, $c_{k,n}$ is the probability that p_k belongs to the *n*th category. The category that has the highest probability is the true category of p_k. Therefore, we define the highest probability as the confidence score of p_k being truly classified, which is denoted as c_k,

$$c_k = Max\{c_{k,n}\} \qquad [3]$$

The calibrating loss function is used to keep the consistency between the importance values of informative patches and their confidence scores. It is formalized as follows,

$$\mathcal{L}_{cal} = -\frac{1}{K}\sum\nolimits_{k=1}^{N}\log(s_k - c_k)^2 \qquad [4]$$

where $s_k \in M$ represents the importance value of p_k and c_k denotes the confidence score of p_k. The loss function is the sum of the distances between each pair of s_k and c_k. It encourages the short distance between s_k and c_k and penalizes the long distance between them.

The mutual-boosting loss function is the joint of the above two loss functions, that is,

$$\mathcal{L}_{infor} = \mathcal{L}_{p-cls} + \mathcal{L}_{cal} \qquad [5]$$

It encourages the better classification of informative patches and the consistency between the importance values and the confidence scores of the informative patches. That means that the patch classifier better classifies the informative patches, which are extracted based on the information map. It meanwhile assists in improving the performance of the information map extractor to produce a more effective information map. The information map in turn boosts the generation of informative patches and the performance of the patch classifier. Therefore, the proposed IS is a mutual-boosting network.

Hierarchical attention sub-network

The task-driven information map *M* has the semantic information for MCI/AD diagnosis. It implies not only disease-related regions but also their importance for image classification. Therefore, we employ *M* to generate hierarchical attention maps to construct HAS. HAS is thereby guided to pay more attention to the visual and semantic features of disease-related regions, and also to

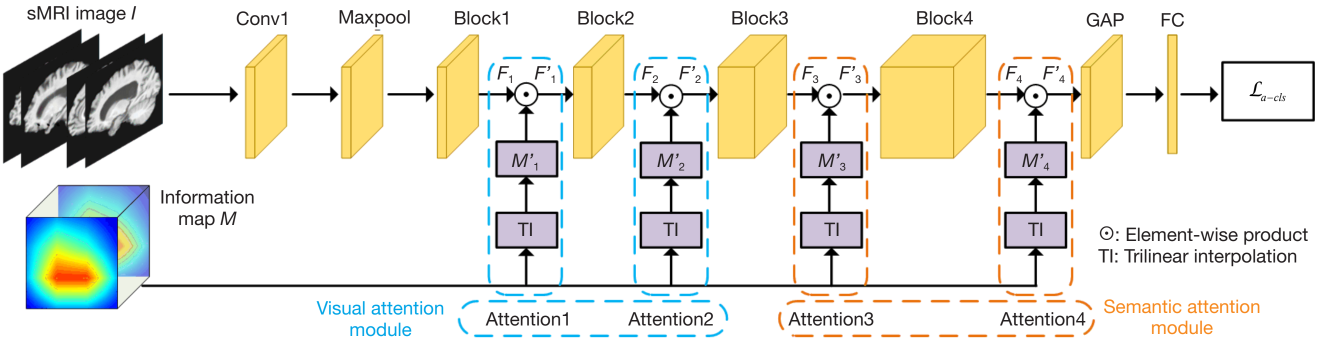

Figure 4 Framework of the hierarchical attention sub-network (HAS). sMRI, structural magnetic resonance imaging; GAP, global average pooling; FC, fully connected layer.

consider the features of the other regions. Therefore, HAS can well exploit the discriminative features of sMRI images for disease diagnosis.

The framework of HAS is depicted in *Figure 4*. Two attention modules are devised to focus on the discriminative feature extraction. They are the visual attention module and the semantic attention module. The visual attention module aims to extract discriminative low-level features, i.e., features from the shallow layers of HAS. It is comprised of two attention blocks, i.e., Attention1 and Attention2 in *Figure 4*. The semantic attention module is used to extract discriminative high-level features, i.e., features from the deep layers of HAS. It consists of another two attention blocks, i.e., Attention3 and Attention4 in *Figure 4*. Since the size of M is different from that of the feature maps in HAS, M cannot be directly used as an attention map. To solve this, each attention block produces an attention map M_i' based on M and a feature map F_i first. Then M_i' is combined with F_i to enhance the features of the disease-related regions in F_i. Each attention block can be formalized as

$$M_i' = TI(M, F_i) \quad [6]$$

$$F_i' = M_i' \odot F_i \quad [7]$$

where M and F_i are the input of each attention block; F_i' is the output; TI is the trilinear interpolation function, making M_i' the same size as F_i; $\odot$ is the element-wise multiplication to highlight important features in F_i. The backbone of HAS is 3D ResNet34. Its structure is shown in the last column in *Table 1*.

It is worth noting that the proposed HAS is superior to the general spatial attention network. This is because all the attention maps in HAS are generated based on the effective information map and they can correctly highlight the features of disease-related regions, no matter the visual features from the shallow layers or the semantic features from the deep layers. However, the general spatial attention network usually generates attention maps from the network itself. This causes the attention maps generated at shallow layers not closely coinciding with disease-related regions due to gradient vanishing. This further leads to the visual features not being well extracted. The non-discriminative visual features impact the extraction of semantic features, which blocks the good performance of the general spatial attention network.

The cross-entropy loss function is employed in HAS for the final classification of I,

$$\mathcal{L}_{a-cls} = \sum_{n=1}^{N} y_n \log\left(c_n \middle| I, M\right) \quad [8]$$

where y_n is the one-hot format of the true label of I and c_n represents the probability that I, together with M, belongs to the nth category.

It is noted that since the generation of hierarchical attention maps is directly dependent on M and M is a task-driven information map, we regard the generation of hierarchical attention maps as a task-driven operation. Therefore, we name the proposed network as a THAN. Moreover, even though there are studies about hierarchical attention networks, the specific frameworks of these hierarchical attention networks are different and they are designed based on their own tasks, which are not efficient for other tasks.

Table 2 Demographic and clinic information of the subjects from ADNI

Variable	MCI	AD	NC
Number	396	327	416
Sex (F/M)	144/252	156/171	213/203
Age	74.7±7.5	75.0±7.9	74.6±5.8
Education	15.7±3.0	15.1±3.0	16.3±2.7
MMSE	27.0±1.8	23.2±2.1	29.1±1.1

Age, education and MMSE are defined as mean ± standard deviation. MMSE, mini-mental state examination; ADNI, Alzheimer's Disease Neuroimaging Initiative; MCI, mild cognitive impairment; NC, normal control; AD, Alzheimer's disease.

Dataset and preprocessing

We utilized the public dataset of Alzheimer's Disease Neuroimaging Initiative (ADNI, http://adni.loni.usc.edu/) as our experimental data. ADNI was launched in 2003 as a public-private partnership. Its primary goal is to test whether serial magnetic resonance imaging, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. In this work, we downloaded the 1.5T T1-weighted sMRI images of 1,139 subjects at the baseline time point from ADNI, including 396 MCI, 327 AD, and 416 NC. *Table 2* shows the demographic and clinical information of the 1,139 subjects, where MMSE represents mini-mental state examination and std represents standard deviation.

All the sMRI images were pre-processed with a "Minimal" pipeline (26), i.e., skull stripping and affine registration (the other is an "Extensive" pipeline, including skull stripping and non-linear registration). The two pre-processing steps were realized with the tool of FMRIB Software Library 5.0 (https://fsl.fmrib.ox.ac.uk/). Affine registration was used to linearly align the sMRI images with the template of MNI152 (27,28) to remove global linear differences and also to resample the sMRI images into the spatial resolution of 1×1×1 mm^3. After this, each sMRI image contains the unnecessary background portion, which has no impact on image classification but increases computation time. Therefore, we removed the background portion of each sMRI image along the minimum vertical external matrix of the brain portion. After this operation, image sizes become different. They are 148±2, 182±2, and 153±3 (in the format of mean ± standard deviation). To keep the same size of these sMRI images, we scaled each sMRI image until the maximum side of each sMRI image is up to 128 by utilizing trilinear interpolation. Additionally, we padded the other two sides of each sMRI image to 128. Thereby, all the sMRI images have the size of 128*128*128. It is noted that we scale images into the size of 128*128*128 because 128 is the multiple of 32 (the size of image patches). Two hundred and fifty-six is not used because images with the size of 256*256*256 are far bigger than that of 128*128*128 and they require more graphics processing unit (GPU) memory and time to train and validate.

Evaluation metrics and implementation details

The proposed THAN is evaluated on four classification tasks: MCI against (*vs.*) NC, AD *vs.* NC, AD *vs.* MCI, and AD *vs.* MCI *vs.* NC. In the first two tasks, NC subjects are regarded as negative cases and AD or MCI subjects are regarded as positive ones. For AD *vs.* MCI, AD subjects are denoted as the positive cases and MCI subjects represent the negative ones. For AD *vs.* MCI *vs.* NC, one type of subject is regarded as positive cases and the remaining two types are as negative cases, which is referred to (29). According to (30), this task was conducted three times and the positive cases varied each time. The average of the three times was computed as the final performance of AD *vs.* MCI *vs.* NC. Six metrics are assessed. They are accuracy (ACC), specificity (SPE), sensitivity (SEN, also called Recall), precision (PRE), F1-score (F1), and area under the curve (AUC) of receiver operating characteristic.

For every task, we divide all the corresponding sMRI images into the training, validation, and test sets with the ratio of 7:2:1. It is noted that each type of sMRI image in every task also has the same ratio (i.e., 7:2:1) among the training, validation, and test sets, which can avoid data imbalance to some extent. For instance, for the task of MCI *vs.* NC, the number of sMRI images of MCI subjects and that of NC subjects have the ratio of 7:2:1 in the training, validation and test sets. The training and validation sets are used to train THAN. Specifically, for every task, we first trained IS with 80 epochs. The best trained IS were preserved when the training loss converged to 0 and the validation loss converged to less than 1. After that, an information map was generated for every sMRI image in the training and validation sets using the preserved best-trained IS. Every sMRI image and its information map formed a pair. Furthermore, pairs of sMRI images and information maps were used to train HAS with 80 epochs.

The best trained HAS was saved when the ACC of the validation set reached the largest value. The independent test set aims to evaluate the trained THAN, i.e., the saved best-trained IS and HAS. To demonstrate the robustness of THAN, we utilized 3-fold cross-validation for all the table-related experiments. Results are the mean of the 3 experiments. For figure-related experiments, we display the results of the certain one among the 3 times, since we do not find a proper way to fuse the experimental results of the 3 times. Moreover, all the methods, including THAN and the comparison methods, were implemented on PyTorch and they were trained and evaluated on an NVIDIA RTX 2080Ti GPU. During the training process, these methods utilized the optimizer of stochastic gradient descent with the momentum (31) of 0.9 and with the weight decay of 0.0001. The learning rates of these methods were increased to 0.001 using the warmup strategy (32) and then were lowered by a tenth every 30 epochs. Further, the batch sizes of these methods were set to 4. In addition, no data augmentation was used in these methods.

Results

Comparison results with two advanced attention modules

The proposed THAN was compared with two advanced attention modules. One is the squeeze-and-excitation (SE) module (33). It aims to pay attention to the relationships among channels of a feature map. The other is the convolutional block attention module (CBAM) (34). It is an extension of SE and utilizes both channel-wise and spatial-wise attention on a feature map. We kept the backbone of HAS unchanged and rewrote the SE and CBAM modules from 2D to 3D, respectively. We then replaced the visual and semantic attention modules (four attention blocks totally) in HAS with four 3D SE and CBAM modules, respectively. The two new networks are denoted as ResNet + SE and ResNet + CBAM. The comparison results of ResNet, ResNet + SE, ResNet + CBAM, and THAN are summarized in *Table 3*.

We can observe that (I) for MCI *vs.* NC, THAN gains the best values in terms of the five metrics among the four networks, and its SEN is just 0.7% lower than the SEN of ResNet + CBAM. Further, compared with ResNet (i.e., the baseline network), THAN and ResNet + CBAM achieve better values on the six metrics, and ResNet + SE obtains larger values on five metrics except the SEN value, which is the same with ResNet's SEN. (II) For AD *vs.* NC, THAN achieves the best values among the four networks in terms of the six metrics. However, both ResNet + SE and ResNet + CBAM are all behind ResNet for the six metrics. (III) For AD *vs.* MCI, THAN achieves the best performance for the five metrics except SEN among the four networks, with SEN being 0.1% behind ResNet + SE's. ResNet + CBAM is superior to ResNet except SPE and PRE, yet ResNet + SE does not have obvious advantages than ResNet. (IV) For AD *vs.* MCI *vs.* NC, THAN attains the best performance among the four networks in terms of the six metrics. The performance of ResNet + SE and ResNet + CBAM is similar and is better than that of ResNet. These results indicate that, overall, THAN is better than the two state-of-the-art attention modules and the baseline network for the four tasks; However, SE and CBAM cannot always promote the performance of the four tasks. This can be explained by the following reasons: since the SE and CBAM modules used in shallow layers do not have direct constraints, they are not well trained due to gradient vanishing. This causes the result that the visual features extracted using the SE and CBAM modules are not well extracted, which further degrades the extraction of semantic features and the performance of image classification. However, for THAN, the visual and semantic attention modules generate attention maps based on an effective information map, highlighting disease-related regions. Therefore, the attention maps in both shallow and deep layers can emphasize disease-related features. This benefits discriminative visual and semantic feature extraction and further facilitates image classification.

Comparison results with state-of-the-art methods

Regarding the four kinds of methods mentioned in Introduction, since their codes are not public, we realized one method in each kind and compared them with THAN. Specifically, for the attention-based methods, the approach in (20) was realized because the other attention-based methods are for different tasks (21,22) or use multi-modality data (23). The only whole image-based method (12) was realized and compared. For the patch-based methods, since the approach in (18) is the most recent study, it was compared. For the ROI-based method, we just realized the method in (9). The four comparison methods were conducted on our dataset and using the same 3-fold cross-validation. Their mean comparison results are summarized in *Table 4*.

It can be seen that (I) for MCI *vs.* NC, THAN outperforms these comparison methods in terms of ACC,

Table 3 Comparison results with two advanced attention-based modules (%)

Methods	MCI *vs.* NC						AD *vs.* NC						AD *vs.* MCI						AD *vs.* MCI *vs.* NC					
	ACC	SPE	SEN	PRE	F1	AUC	ACC	SPE	SEN	PRE	F1	AUC	ACC	SPE	SEN	PRE	F1	AUC	ACC	SPE	SEN	PRE	F1	AUC
ResNet	78.0	80.2	75.4	76.7	76.1	84.3	90.5	90.9	89.9	88.7	89.3	95.8	78.6	80.4	76.3	76.2	76.2	85.1	58.5	79.6	59.6	58.4	58.5	64.0
ResNet + SE	79.0	80.2	77.3	77.3	77.3	84.9	89.2	90.1	87.7	87.7	87.7	95.1	78.6	78.0	80.0	75.2	77.4	84.9	61.2	80.6	61.0	60.8	60.9	64.3
ResNet + CBAM	79.3	81.0	77.1	78.0	77.6	84.7	89.1	89.7	88.2	87.3	87.6	95.0	79.0	79.7	78.0	76.1	77.0	85.2	61.3	80.7	61.0	60.8	60.9	64.1
THAN	80.1	80.3	78.2	79.2	78.6	85.0	92.0	93.1	90.3	91.6	90.9	96.2	80.7	81.0	79.9	77.6	78.7	86.2	62.9	81.8	64.5	61.6	62.9	64.9

MCI, mild cognitive impairment; NC, normal control; AD, Alzheimer's disease; THAN, task-driven hierarchical attention network; CBAM, convolutional block attention module.

Table 4 Comparison results with state-of-the-art methods (%)

Methods	MCI *vs.* NC						AD *vs.* NC						AD *vs.* MCI						AD *vs.* MCI *vs.* NC					
	ACC	SPE	SEN	PRE	F1	AUC	ACC	SPE	SEN	PRE	F1	AUC	ACC	SPE	SEN	PRE	F1	AUC	ACC	SPE	SEN	PRE	F1	AUC
ROI	77.1	79.7	74.5	76.3	76.0	81.4	83.1	85.6	80.1	81.4	80.7	87.4	70.0	73.0	66.5	67.0	66.8	73.0	58.0	78.9	58.4	57.7	58.0	63.1
Patch	73.5	68.6	78.4	70.6	74.3	79.7	82.9	88.6	75.1	84.6	79.4	90.7	69.1	70.3	67.4	65.4	66.1	76.8	60.5	80.2	60.5	60.0	60.2	63.8
WI	77.7	79.5	75.9	77.5	76.6	86.3	88.1	90.2	85.2	87.6	86.4	94.3	72.4	71.4	73.5	68.1	70.6	78.6	60.9	80.2	63.6	59.8	61.6	63.9
Att	76.0	77.7	74.2	76.3	75.1	82.9	86.1	89.6	82.6	86.3	83.9	93.5	75.5	77.3	72.6	73.9	72.8	81.5	59.8	79.7	60.9	60.1	60.5	64.1
Our	80.1	80.3	78.2	79.2	78.6	85.0	92.0	93.1	90.3	91.6	90.9	96.2	80.7	81.0	79.9	77.6	78.7	86.2	62.9	81.8	64.5	61.6	62.9	64.9

MCI, mild cognitive impairment; NC, normal control; AD, Alzheimer's disease; ROI, region of interest; WI, the whole image-based method; Att, the attention-based method; ACC, accuracy; SPE, specificity; SEN, sensitivity; PRE, precision; F1, F1-score; AUC, area under the curve.

Table 5 Selection of the hyper-parameter K (%)

Number of patches	MCI *vs.* NC						AD *vs.* NC						AD *vs.* MCI						AD *vs.* MCI *vs.* NC					
	ACC	SPE	SEN	PRE	F1	AUC	ACC	SPE	SEN	PRE	F1	AUC	ACC	SPE	SEN	PRE	F1	AUC	ACC	SPE	SEN	PRE	F1	AUC
3	77.0	74.7	81.2	75.6	78.3	82.5	90.8	91.8	89.6	89.5	89.5	95.5	75.3	78.6	71.8	73.9	72.3	82.1	60.8	80.3	61.4	60.1	60.8	64.6
4	77.6	76.8	78.6	76.7	77.6	84.7	92.0	93.1	90.3	91.6	90.9	96.2	76.5	76.2	76.4	72.8	74.5	81.4	61.0	80.6	60.6	60.6	60.6	64.8
5	79.4	79.0	79.6	78.6	79.0	86.0	91.5	88.1	90.7	90.1	90.5	96.5	78.0	82.0	73.1	77.0	74.9	83.0	62.8	81.3	62.8	62.4	62.6	65.1
6	81.0	80.0	80.2	79.3	79.8	86.1	90.2	92.5	87.5	90.4	88.8	95.6	80.7	81.0	79.9	77.6	78.7	86.2	62.9	81.8	64.5	61.6	62.9	64.9
7	79.2	77.6	80.8	77.82	79.3	86.7	89.1	90.4	87.7	88.1	87.8	94.8	79.3	78.2	77.0	75.3	77.7	86.1	62.3	81.1	62.1	61.7	61.9	62.9
8	78.4	78.8	77.8	77.7	77.8	85.6	90.8	91.5	87.2	89.8	88.3	95.3	80.6	81.9	79.2	78.1	78.6	86.4	61.7	80.8	62.5	61.0	61.8	63.5
9	77.6	79.3	75.7	78.2	76.8	84.1	89.4	90.8	87.5	88.3	87.9	94.9	76.2	77.9	74.2	73.4	73.7	82.2	61.5	80.8	61.3	61.2	61.3	63.9
10	77.5	77.4	75.6	76.6	76.0	83.7	88.4	90.4	85.7	87.6	86.3	94.0	74.8	75.2	73.0	71.0	72.0	79.3	60.2	80.1	60.2	59.7	59.9	64.4
20	77.5	75.2	81.3	75.7	77.6	83.5	88.7	90.6	86.6	88.0	87.2	95.3	74.0	70.4	78.0	68.6	73.0	79.0	57.9	78.9	58.3	56.9	57.6	61.4
30	75.8	78.0	73.7	76.3	74.9	82.3	84.3	88.7	79.1	84.6	81.6	91.1	74.0	80.9	65.3	73.8	69.3	77.2	57.5	78.9	57.3	56.8	57.1	62.4

MCI, mild cognitive impairment; NC, normal control; AD, Alzheimer's disease; ACC, accuracy; SPE, specificity; SEN, sensitivity; PRE, precision; F1, F1-score; AUC, area under the curve.

SPE, PRE, F1, and AUC, and it just 0.2% falls behind the patch-based method of (18) and achieves excellent performance compared with the ROI-based methods. (II) For AD *vs.* NC, AD *vs.* MCI, and AD *vs.* MCI *vs.* NC, THAN achieves the most excellent performance when compared with these comparison methods. These results imply that THAN improves the performance of MCI/AD diagnosis and achieve better results when compared with the above methods.

Evaluation of the IS

We first show the selection of the hyper-parameter K in IS. After that, we visualize the information map generated by IS to show its effectiveness in highlighting disease-related regions.

The only hyper-parameter K in IS (i.e., the number of informative patches) was investigated based on the classification performance of THAN. We first varied K from 3 to 10 for each task and kept the other settings unchanged. We found that the performance of the four tasks increases first and then degrades. With this trend, we then observed the results when K is 20 and 30, and found that the performance of the four tasks continues falling. The trend that the larger number of patches does not have better performance is due to the following reasons: the selected informative patches are highly related to disease lesions and their corresponding regions in original sMRI images are assigned higher weights for image classification. However, not all the image patches contain lesions. With the increase of K, some of the selected informative patches do not contain disease lesions while they are assigned higher weights for image classification. The incorrect weights on these incorrect informative patches lead to the falling of classification performance. Therefore, there is the trend that the larger number of patches does not have better performance. All the results are shown in *Table 5*. We can see that (I) for MCI *vs.* NC, the four metrics of ACC, SPE, PRE, and F1 are the best when K =6; SEN and AUC achieve the highest points when K is 20 and 7, respectively. According to the overall performance, we set K =6 for MCI *vs.* NC. (II) For AD *vs.* NC, ACC, SPE, PRE, and F1 achieve the best points when K =4. The other two metrics are the highest when K =5. Therefore, we set K =4 for AD *vs.* NC. (III) For AD *vs.* MCI, ACC, SEN, and F1 obtain the largest values when K =6, SPE is the best when K =5, and PRE and AUC are the highest when K =8. Therefore, we set K =6 for this task. (IV) For AD *vs.* MCI *vs.* NC, the

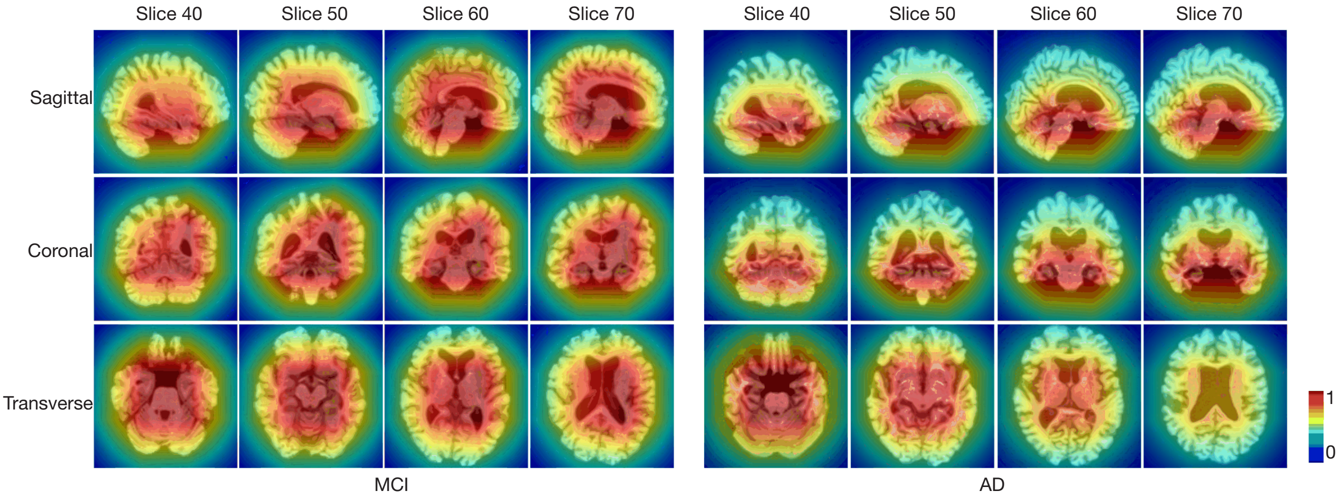

Figure 5 Information map visualization of an MCI (mild cognitive impairment) subject in MCI *vs*. NC (normal control) and an AD (Alzheimer's disease) subject in AD *vs*. NC.

values of ACC, SEN, and F1 are the largest when K =6, and SPE and AUC values are the best when K =5. According to the overall performance, we set K =6 for this task. Even though the optimal values of K for the four tasks are different, they are determined after the network of each task is well trained. That means we do not need to consider K in the test process.

To display the effectiveness of IS in localizing disease-related regions, we visualized the generated information maps of an MCI subject in MCI *vs*. NC and an AD subject in AD *vs*. NC, respectively. Specifically, we first transformed the generated 3D information map of a subject into the same size as the input sMRI image using Eq. [6]. We then selected the 40th, 50th, 60th, and 70th 2D maps from the coronal, sagittal, and transverse views of the transformed 3D information map, respectively. After that, we combined 12 maps with the 2D slices from the input sMRI image to produce the visualized 2D information maps. The results are displayed in *Figure 5*. The color indicates the importance of regions for classification. The redder the color is, the more important the whole region is for AD or MCI diagnosis. It can be seen that (I) for both AD and MCI subjects, the redder regions contain hippocampi and ventricles. This indicates that hippocampus and ventricle play important role in image classification, which is in accordance with the conclusion that hippocampus atrophy and ventricle enlargement are the biomarkers of MCI and AD (35). This observation implies that IS can effectively highlight disease-related regions. Nevertheless, even though hippocampi and ventricles are of importance for disease diagnosis, experts still observe other regions of images and make final diagnosis considering the features of the entire images in clinical trials. Moreover, since the pathogenesis of AD has not been discovered and there may be latent lesion regions related to MCI and AD, it is necessary to consider the features of the remaining regions except hippocampi and ventricles. The contributions of the features of other regions for image classification are decided based on the weights of these attention maps. (II) The redder regions of the MCI subject are larger than that of the AD subject. This may be because the MCI and NC subjects are harder to distinguish than the AD and NC subjects, and IS requires to focus on more regions to generate more discriminative features for MCI diagnosis. It is noted that some redder regions are in the background. This is caused by the mapping relations between image patches and the values of an information map. *Figure 3* displays the mapping relations. According to the mapping relations, a value in the information map corresponds to a 32*32*32 image patch in the original image. Therefore, some image patches, mapped from some values in the information map, contain both brain tissue and background. Further, for coarsely visualizing the importance of image patches, a value in the information map can be mapped into a 32*32*32 patch with values and these values on the patch are realized through the trilinear interpolation to the value in the information map. It is noted that the patch with values and the image patch forementioned are two different concepts. Due

Table 6 Ablation study of the visual and semantic attention modules in HAS (%)

Visual	Semantic	MCI *vs.* NC						AD *vs.* NC						AD *vs.* MCI						AD *vs.* MCI *vs.* NC					
attention	attention	ACC	SPE	SEN	PRE	F1	AUC	ACC	SPE	SEN	PRE	F1	AUC	ACC	SPE	SEN	PRE	F1	AUC	ACC	SPE	SEN	PRE	F1	AUC
×	×	78.0	80.2	75.4	76.7	76.1	84.3	90.5	90.9	89.9	88.7	89.3	95.8	78.6	80.4	76.3	76.2	76.2	85.1	58.5	79.6	59.6	58.4	58.5	64.0
√	×	78.4	77.6	79.4	76.6	77.9	85.0	90.6	92.3	88.5	90.1	89.2	96.5	79.1	80.6	77.3	76.6	76.8	85.2	59.8	79.9	60.1	59.4	59.8	64.6
×	√	78.7	78.4	78.9	77.2	78.0	84.4	90.6	92.1	88.5	90.4	89.4	95.9	78.9	81.0	76.0	76.7	76.4	84.7	59.9	80.1	59.4	59.4	59.4	63.3
√	√	80.1	80.3	78.2	79.2	78.6	85.0	92.0	93.1	90.3	91.6	90.9	96.2	80.7	81.0	79.9	77.6	78.7	86.2	62.9	81.8	64.5	61.6	62.9	64.9

HAS, hierarchical attention sub-network; MCI, mild cognitive impairment; NC, normal control; AD, Alzheimer's disease; ACC, accuracy; SPE, specificity; SEN, sensitivity; PRE, precision; F1, F1-score; AUC, area under the curve.

to the trilinear interpolation, when the forementioned image patch, containing both brain tissue and background, corresponds to a higher value in the information map, it will be combined with a patch with higher values, which indicate redder color. Therefore, the background portions in these image patches are also with redder color. This coarse visualization is a common phenomenon in current methods, such as in (36) and (37). However, this phenomenon does not affect the performance of THAN. Since information maps are used to generate attention maps in different layers of HAS, after multiple convolution layers, HAS can recognize the useless background portions.

Ablation study of the HAS

To validate the effectiveness of the combination of the visual and semantic attention modules in boosting the discriminative feature extraction for disease diagnosis, we conducted the ablation study on different combinations of the two attention modules with the backbone of HAS. The results are illustrated in *Table 6*.

For MCI *vs.* NC, we find that (I) the backbone + visual attention and the backbone + semantic attention are superior to the backbone in terms of ACC, SEN, F1, and AUC. (II) The backbone + visual attention module + semantic attention module (i.e., THAN) achieves the best performance referring to the six metrics when compared with the backbone and the backbone + semantic attention, and it obtains the highest values referring to the five matrices except SEN when compared with the backbone + visual attention. These results imply that both the visual attention module and the semantic attention module facilitate the extraction of discriminative features and further promote the performance of MCI *vs.* NC to some extent. Moreover, the performance of MCI *vs.* NC is best improved when the two attention modules are combined.

For AD *vs.* NC, the following results can be acquired. (I) Both the backbone + visual attention module and the backbone + semantic attention module are a little better than the backbone overall. (II) THAN achieves the best performance among the four methods. These results imply that both the visual attention module and the semantic attention module boost the performance of AD *vs.* NC. Furthermore, the performance of AD *vs.* NC is improved to the highest point when the two attention blocks are employed together.

For AD *vs.* MCI, we can see that (I) the backbone + visual attention module is better than the backbone in terms

Table 7 Execution time of THAN for the test sets (ms)

Execution time	MCI *vs.* NC	AD *vs.* NC	AD *vs.* MCI	AD *vs.* MCI *vs.* NC
Mean time	45.48	45.67	45.15	45.32
Max time	46.28	46.96	47.02	48.21
Min time	44.64	44.77	44.65	43.96

THAN, task-driven hierarchical attention network; MCI, mild cognitive impairment; NC, normal control; AD, Alzheimer's disease.

of ACC, SPE, PRE, F1, and AUC; the backbone + semantic attention module is superior to the backbone for ACC, SPE, PRE, and F1. (II) THAN achieves the best performance for all the matrices among the four methods. These results imply that the performance of AD *vs.* MCI is promoted by the visual attention module and the semantic attention module, respectively. Moreover, it is best improved when the two attention modules are combined.

For AD *vs.* MCI *vs.* NC, we observe that (I) the backbone + visual attention module is better than the backbone in terms of the six metrics; the backbone + semantic attention module is superior to the backbone for ACC, SPE, PRE, and F1. (II) THAN achieves the best performance for all the metrics when compared with the four methods. These results indicate that AD *vs.* MCI *vs.* NC is improved by the two attention modules, respectively, and it is improved to the best point when the two attention modules are combined.

Overall, the performance of the four tasks is promoted by the visual attention block and the semantic attention block, respectively, and it is highly improved by combining the visual and semantic attention modules into the backbone. This demonstrates that HAS can effectively extract both discriminative visual and semantic features for disease diagnosis.

Discussion

We first investigate the execution time of THAN. Moreover, we visualize the features extracted using THAN. Finally, we compare THAN with state-of-the-art methods.

Execution time of THAN

To discuss the efficiency of THAN, we described the execution time of THAN on the test sets for the four tasks, as shown in *Table 7*. The mean time indicates the average execution time of an sMRI image in a task. The max time indicates the maximum execution time of an sMRI image in a task. The min time indicates the minimum execution time of an sMRI image in a task. It can be seen that the mean time, the min time and even the max time of every sMRI image in the four tasks are far less than 1 s. This demonstrates that the proposed THAN is very efficient, even though it is made up of two sub-networks.

Visualization of the original images and their features extracted using THAN

To further demonstrate the ability of THAN in extracting discriminative features, t-distributed stochastic neighbor embedding (t-SNE) (38) was utilized to visualize the original images of the test set and the features of the test set extracted using THAN. t-SNE is a technique to visualize high-dimensional data by giving each datapoint a location in a 2D or 3D map (38). To be specific, it projects features from a high dimension to a low dimension and meanwhile attempts to preserve the local structures of these features. Since the dimensions of high-dimensional features have no specific meanings, the two dimensions of the projected 2D features also have no detailed meanings. Furthermore, since the 2D map is simpler than the 3D map, it is chosen to visualize each datapoint. It is noted that t-SNE cannot demonstrate the classification performance of THAN, which is determined by both the features extracted and the classifier used.

Specifically, each original image was reshaped into a one-dimensional feature vector first. T-SNE was then utilized to visualize these original images. For the features extracted using THAN, the trained IS was taken to produce information maps for the sMRI images. After that, the first to the GAP layers of the trained HAS were used to produce a 512-dimensional feature vector for each sMRI image. Finally, t-SNE was employed to visualize these 512-dimensional feature vectors of the test set. Results are shown in *Figure 6*. A red point represents the original image or its feature vector of an MCI subject, a yellow point the original image or its feature vector of an NC subject, and a blue point the original image or its feature vector of an AD subject.

We can find that (I) compared with original images, the features extracted using THAN has the ability to differentiate different categories. This indicates that the proposed THAN can learn discriminative features from sMRI images. (II) The two kinds of features in AD *vs.* NC

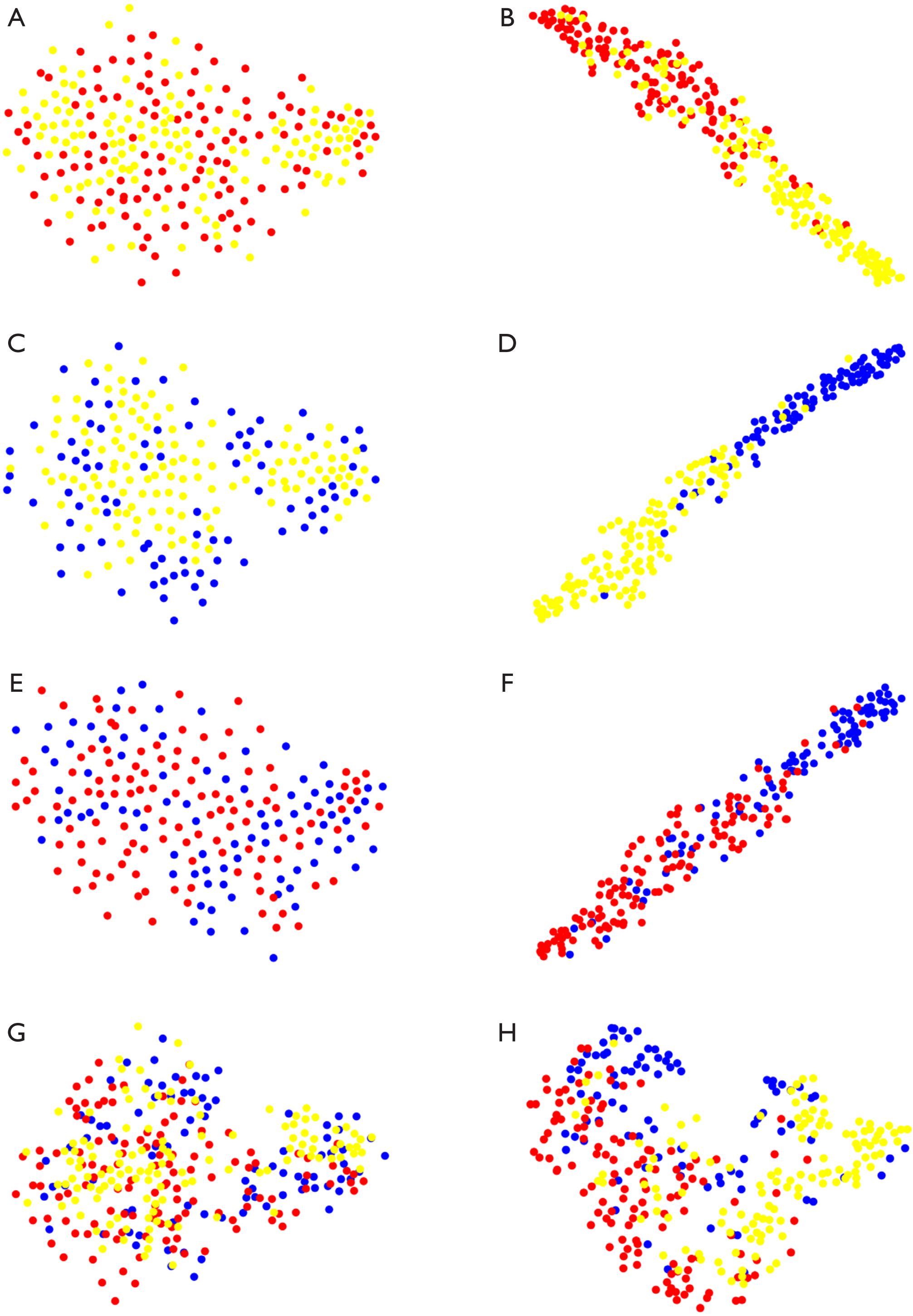

Figure 6 Visualization of the original images and their features of the test sets using t-SNE. The left four figures refer to the t-SNE results of the original images, and the right four figures refer to the t-SNE results of the features extracted using THAN. The four rows from the top to the bottom correspond to the tasks of MCI *vs.* NC, AD *vs.* NC, AD *vs.* MCI and AD *vs.* MCI *vs.* NC, respectively. MCI, mild cognitive impairment; NC, normal control; AD, Alzheimer's disease; t-SNE, t-distributed stochastic neighbor embedding.

are best distinguished compared with those in MCI *vs.* NC, AD *vs.* MCI, and AD *vs.* MCI *vs.* NC. This result matches the result in *Table 6*, that is, ACC of AD *vs.* NC is the best, followed by MCI *vs.* NC, by AD *vs.* MCI, and then by AD *vs.* MCI *vs.* NC. This indicates that AD *vs.* NC is the easiest task in the four tasks, and MCI *vs.* NC, AD *vs.* MCI, and AD *vs.* MCI *vs.* NC require more exploration, especially the task of AD *vs.* MCI *vs.* NC. Motivated by the observation, in the future, we will pay more attention to investigate the recognition between MCI and NC, between AD and MCI, and among AD, MCI, and AD.

Acknowledgments

Funding: This work was supported in part by the National Natural Science Foundation of China (grant No. 61901237), Natural Science Foundation of Zhejiang Province (grant No. LQ20F020013), Ningbo Natural Science Foundation (grant No. 2019A610102, 2019A610103, 2018A610057), and Open Progam of the State Key Lab of CAD&CG Zhejiang University (grant No. A1915). Data collection and sharing for this project was funded by the ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions such as AbbVie, Alzheimer's Association, and Araclon Biotech. A complete listing of ADNI investigators can be found online. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada.

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/qims-21-91). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Jia L, Quan M, Fu Y, Zhao T, Li Y, Wei C, Tang Y, Qin Q, Wang F, Qiao Y, Shi S, Wang Y, Du Y, Zhang J, Zhang J, Luo B, Qu Q, Zhou C, Gauthier S, Jia J. Dementia in China: epidemiology, clinical management, and research advances. Lancet Neurol 2020;19:81-92.
2. Patterson C. World Alzheimer report 2018. Available online: https://www.alz.co.uk/research/world-report-2018
3. Roberts R, Knopman DS. Classification and epidemiology of MCI. Clin Geriatr Med 2013;29:753-72.
4. Reiman EM, Jagust WJ. Brain imaging in the study of Alzheimer's disease. Neuroimage 2012;61:505-16.
5. Frisoni GB, Fox NC, Jack CR, Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer disease. Nat Rev Neurol 2010;6:67-77.
6. Brand L, Nichols K, Wang H, Shen L, Huang H. Joint multi-modal longitudinal regression and classification for alzheimer's disease prediction. IEEE Trans Med Imaging 2020;39:1845-55.
7. Suk HI, Lee SW, Shen D. Deep ensemble learning of sparse regression models for brain disease diagnosis. Med Image Anal 2017;37:101-13.
8. Klöppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, Fox NC, Jack CR Jr, Ashburner J, Frackowiak RSJ. Automatic classification of MR scans in Alzheimer's disease. Brain 2008;131:681-9.
9. Aderghal K, Benois-Pineau J, Afdel K. Classification of sMRI for Alzheimer's disease Diagnosis with CNN: Single Siamese Networks with 2D+? Approach and Fusion on ADNI. ICMR 2017: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval; June 6-9; Bucharest, Romania. New York: ACM, 2017:494-8.
10. Cui R, Liu M. Hippocampus Analysis by Combination of 3-D DenseNet and Shapes for Alzheimer's Disease Diagnosis. IEEE J Biomed Health Inform 2019;23:2099-107.
11. Xia Z, Yue G, Xu Y, Feng C, Yang M, Wang T, Lei B. A Novel End-to-End Hybrid Network for Alzheimer's Disease Detection Using 3D CNN and 3D CLSTM. ISBI 2020: 2020 IEEE 17th International Symposium on Biomedical Imaging; 2020 Apr 3-7; Iowa City, IA, USA.

New York: IEEE, 2020:1-4.

12. Korolev S, Safiullin A, Belyaev M, Dodonava Y. Residual and plain convolutional neural networks for 3D brain MRI classification. ISBI 2017: 2017 IEEE 14th International Symposium on Biomedical Imaging; 2017 Apr 18-21; Melbourne, Australia. New York: IEEE, 2017:835-8.
13. Liu M, Zhang J, Nie D, Yap P, Shen D. Anatomical landmark based deep feature representation for MR images in brain disease diagnosis. IEEE J Biomed Health Inform 2018;22:1476-85.
14. Liu M, Zhang J, Adeli E, Shen D. Landmark-based deep multi-instance learning for brain disease diagnosis. Med Image Anal 2018;43:157-68.
15. Liu M, Zhang J, Adeli E, Shen D. Deep multi-task multi-channel learning for joint classification and regression of brain status. MICCAI 2017: Proceedings of the 20th international conference on medical image computing and computer-assisted intervention; 2017 Sep 11-13; Quebec City, QC, Canada. Berlin: Springer, 2017:3-11.
16. Liu M, Zhang J, Adeli E, Shen D. Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis. IEEE Trans Biomed Eng 2019;66:1195-206.
17. Lian C, Liu M, Zhang J, Shen D. Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. IEEE Trans Pattern Anal Mach Intell 2020;42:880-93.
18. Qiu S, Joshi PS, Miller MI, Xue C, Zhou X, Karjadi C, Chang GH, Joshi AS, Dwyer B, Zhu S, Kaku M, Zhou Y, Alderazi YJ, Swaminathan A, Kedar S, Saint-Hilaire M, Auerbach SH, Yuan J, Sartor EA, Au R, Kolachalama VB. Development and validation of an interpretable deep learning framework for Alzheimer’s disease classification. Brain 2020;143:1920-33.
19. Long B, Yu CP, Konkle T. Mid-level visual features underlie the high-level categorical organization of the ventral stream. Proc Natl Acad Sci U S A 2018;115:E9015-24.
20. Jin D, Xu J, Zhao K, Hu F, Yang Z, Liu B, Jiang T, Liu Y. Attention-based 3D convolutional network for Alzheimer’s disease diagnosis and biomarkers exploration. ISBI 2019: 2019 IEEE 16th International Symposium on Biomedical Imaging; Apr 8-11; Venice, Italy. New York: IEEE, 2019:1047-51.
21. Li Q, Xing X, Sun Y, Xiao B, Wei H, Huo Q, Zhang M, Zhou X S, Zhan Y, Xue Z, Shi F. Novel iterative attention focusing strategy for joint pathology localization and prediction of MCI progression. MICCAI 2019: Proceedings of the 22nd international conference on medical image computing and computer-assisted intervention; 2019 Oct 13-17; Shenzhen, China. Berlin: Springer, 2019:307-15.
22. Lian C, Liu M, Wang L, Shen D. End-to-end dementia status prediction from brain MRI using multi-task weakly-supervised attention network. MICCAI 2019: Proceedings of the 22nd international conference on medical image computing and computer-assisted intervention; 2019 Oct 13-17; Shenzhen, China. Berlin: Springer, 2019:158-67.
23. Zhang L, Wang L, Zhu D. Jointly Analyzing Alzheimer's Disease Related Structure-Function Using Deep Cross-Model Attention Network. ISBI 2020: 2020 IEEE 17th International Symposium on Biomedical Imaging; 2020 Apr 3-7; Iowa City, IA, USA. New York: IEEE, 2020:563-7.
24. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. CVPR 2016: Proceedings of the IEEE conference on computer vision and pattern recognition; June 27-30; Las Vegas, NV, USA. New York: IEEE, 2019:770-8.
25. Yang Z, Luo T, Wang D, Hu Z, Gao J, Wang L. Learning to navigate for fine-grained classification. ECCV 2018: Proceedings of the European Conference on Computer Vision; Sep 8-14; Munich, Germany. Berlin: Springer, 2018:420-35.
26. Wen J, Thibeau-Sutre E, Diaz-Melo M, Samper-González J, Routier A, Bottani S, Dormont D, Durrleman S, Burgos N, Colliot O. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. Med Image Anal 2020;63:101694.
27. Fonov VS, Evans AC, Mckinstry RC, Almli CR, Collins DL. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage 2009;47:S102.
28. Fonov VS, Evans AC, Botteron KN, Almli CR, Mckinstry RC, Collins DL. Unbiased average age-appropriate atlases for pediatric studies. NeuroImage 2011;54:313-27.
29. Jayachandran A, Dhanasekaran R. Multi class brain tumor classification of MRI images using hybrid structure descriptor and fuzzy logic based RBF kernel SVM. Iranian Journal of Fuzzy Systems 2017;14:41-54.
30. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an Overview. arXiv:2008.05756 [preprint]. Available online: https://arxiv.org/abs/2008.05756
31. Qian N. On the momentum term in gradient descent learning algorithms. Neural Netw 1999;12:145-51.
32. Goyal P, Dollár P, Girshick R, Noordhuis P, Wesolowski

L, Kyrola A, Tulloch A, Jia Y, He K. Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv:1706.02677v2 [preprint]. 2018 [cited 2021 Jan 24]: [12 p.]. Available online: https://arxiv.org/abs/1706.02677

33. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. CVPR 2018: Proceedings of the IEEE conference on computer vision and pattern recognition; June 18-22; Salt Lake City, UT, USA. New York: IEEE, 2019:7132-41.

34. Woo S, Park J, Lee JY, Kweon IS. Cbam: Convolutional block attention module. ECCV 2018: Proceedings of the European conference on computer vision; 2018 Sep 8-14; Munich, Germany. Berlin: Springer, 2018:3-19.

35. Saied I, Arslan T, Chandran S, Smith C, Spires-Jones T, Pal S. Non-Invasive RF Technique for Detecting Different Stages of Alzheimer's Disease and Imaging Beta-Amyloid Plaques and Tau Tangles in the Brain. IEEE Trans Med Imaging 2020;39:4060-70.

36. Zhang X, Han L, Zhu W, Sun L, Zhang D. An Explainable 3D Residual Self-Attention Deep Neural Network for Joint Atrophy Localization and Alzheimer's Disease Diagnosis using Structural MRI. IEEE J Biomed Health Inform 2021. [Epub ahead of print]. doi: 10.1109/JBHI.2021.3066832.

37. Lian C, Liu M, Wang L, Shen D. Multi-Task Weakly-Supervised Attention Network for Dementia Status Estimation With Structural MRI. IEEE Trans Neural Netw Learn Syst 2021. [Epub ahead of print]. doi: 10.1109/TNNLS.2021.3055772.

38. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579-605.

Cite this article as: Zhang Z, Gao L, Jin G, Guo L, Yao Y, Dong L, Han J; the Alzheimer's Disease NeuroImaging Initiative. THAN: task-driven hierarchical attention network for the diagnosis of mild cognitive impairment and Alzheimer's disease. Quant Imaging Med Surg 2021;11(7):3338-3354. doi: 10.21037/qims-21-91