DOI:10.19651/j.cnki.emt.2416955

结合视图感知 CNN 和 Transformer 的阿尔茨海默病诊断研究*

吴慧东[1] 刘立程[1] 潘 丹[2]
(1.广东工业大学信息工程学院 广州 510006;2.广东技术师范大学电子与信息学院 广州 510665)

摘 要:为解决阿尔茨海默病(AD)患者大脑结构性核磁共振影像(sMRI)病变细微复杂和空间异质性分布引起的病症诊断准确率低的问题,提出了一种结合卷积神经网络(CNN)和 Transformer 优势的混合架构,用于 AD 病症诊断。首先,设计了多视图特征编码器,通过构造融合混合注意力机制的视图局部特征提取器分支,从 sMRI 的冠状面、矢状面和轴向面方向提取潜在互补信息,并通过多视图信息交互学习策略增强病灶区域的语义表征。其次,设计了级联式多尺度融合子网络,逐层融合多尺度特征图以生成更丰富判别信息。最后,利用 Transformer 编码器建模了全脑 sMRI 的全局特征表示。在阿尔茨海默病神经影像倡议(ADNI)数据集上的结果显示,本文方法在 AD 分类和轻度认知障碍(MCI)转化预测任务的准确率分别达到了 94.05%和 81.59%,优于多种现有方法。

关键词:阿尔兹海默病;结构性核磁共振成像;混合架构;多视图信息;多尺度特征
中图分类号:TN911.73 **文献标识码**:A **国家标准学科分类代码**:520.20

Combining view-aware CNN and Transformer for Alzheimer's disease diagnosis research

Wu Huidong[1] Liu Licheng[1] Pan Dan[2]
(1. School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China;
2. School of Electronics and Information, Guangdong Polytechnic Normal University, Guangzhou 510665, China)

Abstract: To address the low diagnostic accuracy of Alzheimer's disease (AD) caused by the subtle complexity and spatial heterogeneity of brain lesions in structural magnetic resonance imaging (sMRI) of AD patients, a hybrid architecture that combines the strengths of convolutional neural networks (CNN) and Transformers is proposed for the AD diagnosis. First, a multi-view feature encoder is designed, in which a view local feature extractor with integrated hybrid attention mechanisms is employed to extract complementary information from the coronal, sagittal, and axial views of sMRI. The semantic representation of lesion regions is further enhanced through a multi-view information interaction learning strategy. Second, a cascaded multi-scale fusion subnetwork is designed to progressively fuse multi-scale feature map information, enhancing discriminative ability. Finally, a Transformer encoder is used to model the global feature representation of full-brain sMRI. Results on the Alzheimer's disease neuroimaging initiative (ADNI) dataset show that the proposed method in this paper achieves classification accuracies of 94.05% for AD and 81.59% for mild cognitive impairment (MCI) conversion prediction, outperforming several existing methods.

Keywords: Alzheimer's disease; structural magnetic resonance imaging; hybrid architecture; multi-view feature; multi-scale feature

0 引 言

阿尔茨海默病(Alzheimer's disease, AD)是一种常见的慢性神经退行性疾病,主要表现为记忆力、感官和认知能力的逐渐衰退[1]。随着全球人口老龄化的加剧,世界卫生组织预计到 2050 年,AD 患者人数将达到 1.39 亿[2]。轻度认知障碍(mild cognitive impairment, MCI)被认为是 AD 的前驱阶段,MCI 患者根据在初次筛查后 36 个月内是

收稿日期:2024-09-22
*基金项目:国家自然科学基金项目(61976058)、广州市科技计划项目(202206010007)、广东省科技计划项目(2021B0101220006)

否转化为 AD，可分为进展型（progressive MCI，pMCI）和稳定型（stable MCI，sMCI）两类[3]。由于目前尚无有效的 AD 治疗方法，早期诊断并进行预防性干预对于阻止疾病进展具有重要意义。

近年来，许多研究广泛探索了机器学习和深度学习算法在大脑结构性核磁共振成像（structural magnetic resonance imaging，sMRI）的 AD 诊断应用，旨在为临床医生提供更可靠的计算机辅助诊断支持。基于机器学习的 AD 诊断方法通常需要借助专家知识从 sMRI 中手动提取感兴趣特征（如海马区体积和灰质密度）[4]，再用于训练支持向量机等分类器。然而，由于特征设计过程冗长耗时，且易受主观性影响，可能会限制模型性能。以卷积神经网络（convolutional neural network，CNN）为代表的深度学习架构通过端到端的方式，能够自动从 sMRI 中提取特征，并实现了优于临床专家的准确性[5]。然而，由于 CNN 的局部感受野限制了其在全脑 sMRI 中建模局部病变结构复杂关联的能力，使得全局信息的有效表征受到限制。为此，基于自注意力机制（self-attention，SA）在建模长程依赖关系方面的优势，Vision Transformer（VIT）[6]被应用于医学影像诊断研究[7]。为了利用卷积运算在提取病理局部细节特征（如纹理和边缘形态）方面的优势，已有研究开始将 CNN 与 Transformer 相结合，构建了基于混合架构的 AD 诊断方法[5,8-9]。

现有 AD 诊断的深度学习方法，通常仅从 sMRI 的单一视图方向（如冠状面）提取特征，然而，脑组织中的病变位置分布通常表现出复杂的空间异质性[10]，导致在不同视图中提取的特征信息存在差异性和互补性[11]。因此，提取隐藏在 sMRI 不同视图方向上的空间和语义信息，并融合生成判别信息更丰富的高级特征表示[12]，能够促进模型识别更细微复杂的病变信息，提高 AD 诊断的可靠性。

因此，为了充分利用 AD 患者全脑 sMRI 中病变区域之间的复杂关联关系及局部细节信息，本文提出了一种结合视图特征感知 3DCNN 和 Transformer 编码器的混合架构网络（multi-view feature perception network，MVFP-Net）。在卷积网络设计了一种多视图特征编码器（multi-view feature encoder，MVFE），其中的视图局部特征提取器（view-local feature extractor，VLFE）分支融合全脑病灶结构在冠状面、矢状面和轴状面视图方向的互补信息；而跨视图注意力（cross-view attention，CVA）分支促进跨视图信息交互，以增强模型对全脑 sMRI 的语义理解。此外，基于少量卷积和池化操作，设计了级联式多尺度融合子网络（cascaded multi-scale fusion subnetwork，CMSF），通过逐层级复用不同尺度的特征图，使卷积网络输出深层特征图具有更丰富的语义和空间信息。最后，通过 Transformer 编码器建模局部病变信息之间的复杂关联关系，以实现全脑 sMRI 从局部到全局的信息表征。

1　MVFP-Net 模型

1.1　模型整体架构

如图 1 所示，首先，形状为 1×112×112×112 的 sMRI 通过由两个步幅为 2 的卷积层构成的主干块，以学习浅层细粒度信息。其次，由 3 个连续的特征提取阶段来提取深层信息，阶段结构如图 1(a)所示。受 ConvNext 在医学影像分类的成功应用启发[13]，本研究将 ConvNext 残差块作为每个阶段的特征聚合单元（feature aggregation unit，FAU），利用大核感受野和通道缩放来增强复杂病变信息的聚合能力。将 FAU 的输出设为原视图特征图 $\boldsymbol{V}_0 \in \mathbb{R}^{C\times H\times W\times D}$，$C$、$H$、$W$ 和 D 分别表示通道、高度、宽度和深度。经过维度变换操作和分别获得视图特征图 $\boldsymbol{V}_1 \in \mathbb{R}^{C\times W\times D\times H}$ 和 $\boldsymbol{V}_2 \in \mathbb{R}^{C\times H\times D\times W}$。由于特征流尺度满足 $H=W=D=L$，故各视图特征图可表示为 $\boldsymbol{V}_i \in \mathbb{R}^{C\times L\times L\times L}$，$i\in[0,1,2]$。以标识不同视图方向。接着，进入 MVFE 中的 VLFE 和 CVA 分支，$\boldsymbol{V}_0$、$\boldsymbol{V}_1$ 和 $\boldsymbol{V}_2$ 分别进行视图信息的提取和交互后，再采用通道拼接和卷积运算进行融合。此外，在阶段之间由步幅为 2 的卷积层进行空间下采样，即第 j 阶段输出特征图 $\boldsymbol{F}_j \in \mathbb{R}^{C_j\times L_j\times L_j\times L_j}$，其中 $C_j=64\times 2^{j-1}$，$L_j=28/2^{j-1}$，$j\in[1,2,3]$。随后，将 $\boldsymbol{F}_1$、$\boldsymbol{F}_2$ 和 $\boldsymbol{F}_3$ 通过 CMSF 子网络实现多尺度特征融合，再将输出特征图经过嵌入映射重塑为视觉 Token 矩阵。最终，由 Transformer 编码器建模全脑局部病变信息的全局表示，将分类头向量经过全连接层实现病症预测。

1.2　多视图特征编码器

1）VLFE 模块

为了利用全脑病灶结构在不同视图方向上的互补信息，并生成更高级的语义表示，本研究在 MVFE 上设计了 VLFE 分支。如图 2 所示，构建了 3 个并行网络分别处理 3 个视图方向特征图。为了使网络关注 AD 相关的判别性区域，在每个并行网络上设计了一种轻量型混合注意力模块，以生成空间和通道维度的混合域注意力分数。即视图方向特征图 $\boldsymbol{V}_i$ 的空间域注意力图表示为 $\boldsymbol{W}_s(\boldsymbol{V}_i) \in \mathbb{R}^{1\times L\times L\times L}$：

$$W_s(V_i)=BN(Conv_1(BN(DWConv_3(BN(DWConv_3(Conv_1(V_i))))))) \tag{1}$$

其中，$Conv_k$ 和 $DWConv_k$ 分别为核大小为 k 的普通和深度卷积。考虑到计算效率，在通道域注意力分支采用了轻量级的 ECA 通道注意力机制[14]，通道域注意力图定义为 $\boldsymbol{W}_c(\boldsymbol{V}_i) \in \mathbb{R}^{C\times 1\times 1\times 1}$：

$$W_c(V_i)=Conv_{\kappa(C)}(Avgpool(V_i)) \tag{2}$$

$$\kappa(C)=\left|\frac{\log_2 C+1}{2}\right|_{odd} \tag{3}$$

其中，$\kappa(C)$ 为随通道动态调整的核大小，$|t|_{odd}$ 表示取接近 t 的奇整数。将空间和通道注意力图合并后，得到混合域注意力图 $\boldsymbol{W}_i(\boldsymbol{V}_i) \in \mathbb{R}^{C\times L\times L\times L}$：

$$W_i(V_i)=Sigmoid(W_s(V_i)\oplus W_c(V_i)) \tag{4}$$

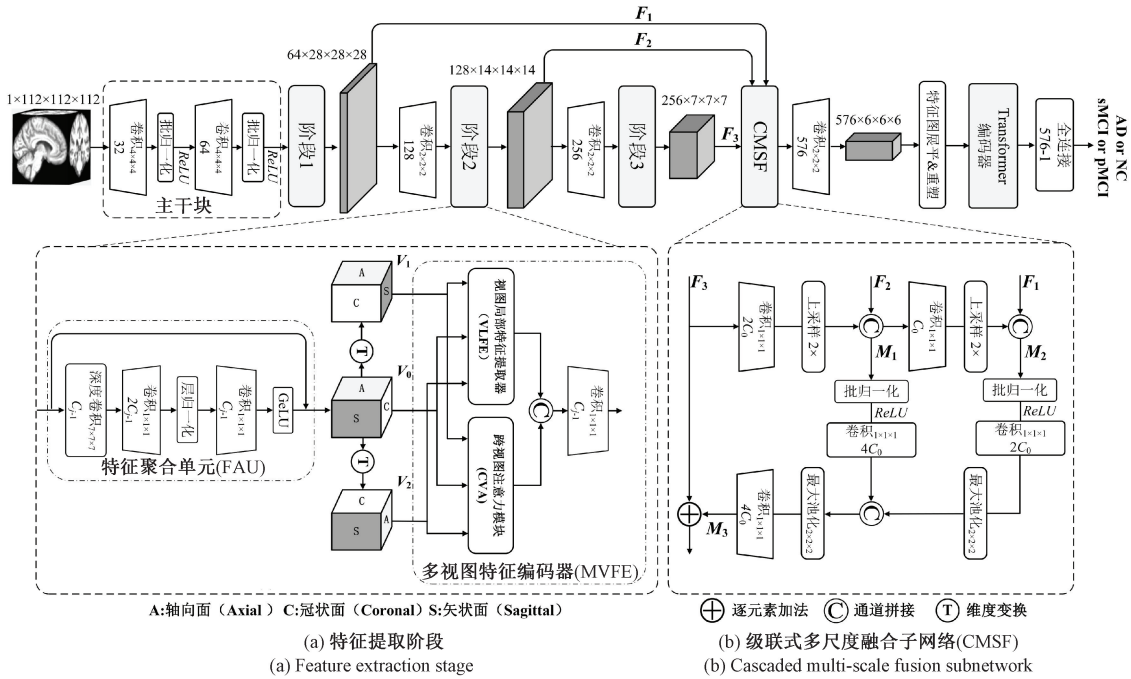


图 1 MVFP-Net 整体框架

Fig. 1 The MVFP-Net overall architecture

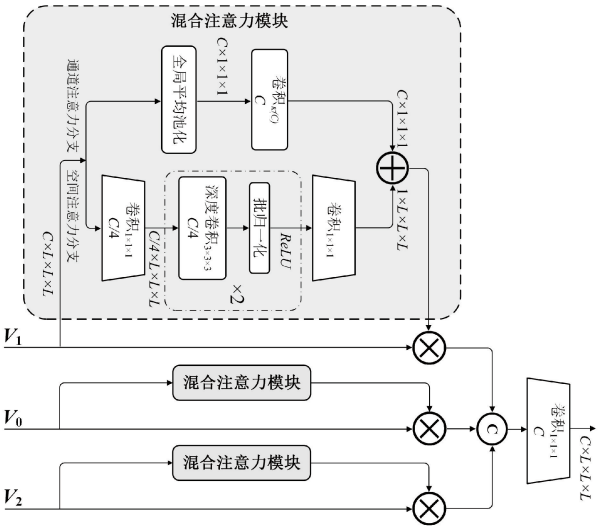


图 2 VLFE 模块结构

Fig. 2 Structure of the VLFE module

其中,*Sigmoid* 表示归一化函数。将 $\boldsymbol{W}_i(\boldsymbol{V}_i)$ 与 $\boldsymbol{V}_i$ 进行逐元素相乘,实现全脑局部区域重要性权重的动态调整。最终,3 个并行网络的输出沿通道拼接后,通过卷积运算融合视图信息。

2)CVA 模块

为了增强网络对空间结构语义信息的表征能力,本研究基于自注意力信息加权融合的跨视图交互学习策略,设计了 CVA 模块,结构如图 3 所示。首先,由核大小为 K_p 的平均池化层对每个视图特征图 $\boldsymbol{V}_i$ 进行空间下采样,输出的尺寸满足 $L'=L/K_p$。其次,通过输出通道数为 $3C$ 的卷积层,输每个视图的查询 $\boldsymbol{q}_i$、键 $\boldsymbol{k}_i$ 和值 $\boldsymbol{v}_i$ 向量矩阵。为了实现跨视图信息交互学习,将每个视图的 $\boldsymbol{q}_i$、$\boldsymbol{k}_i$ 和 $\boldsymbol{v}_i$ 进行空间域维度拼接后。再通过 SA 编码不同视图局部 Token 向量间的关联程度,该过程表示为:

$$CVA(q_i,k_i,v_i;M_{Atten})=SoftMax\left(\frac{\prod_{i=0}^{2}q_i\prod_{i=0}^{2}k_i^T}{\sqrt{C}}\right)\prod_{i=0}^{2}v_i \tag{5}$$

其中,运算符 $\prod$ 表示空间域拼接。$\boldsymbol{M}_{Atten}\in\mathbb{R}^{3(\boldsymbol{L}')^3\times\boldsymbol{C}}$ 表示经过自注意力加权融合后的 Token 向量矩阵。为优化梯度信息传递,进行了残差连接。接着,将 $\boldsymbol{M}_{Atten}$ 根据原空间域拼接顺序拆分为对应 3 个视图的 Token 向量矩阵,重塑还原为特征图形式后,输出特征图 $\boldsymbol{V}'_i\in\mathbb{R}^{C\times L'\times L'\times L'}$。最终,依次通过逐元素相加融合和倍率因子为 K_p 的三线性插值上采样,恢复为原输入特征图形状。

1.3 CMSF 子网络

由于与 AD 相关的病理变化呈现异质性分布,且病灶组织结构具有不同空间尺度。因此,为了实现病灶信息的层次性表征,设计了 CMFS 子网络。如图 1(b)所示,给定各个阶段的输出 $\boldsymbol{F}_1$、$\boldsymbol{F}_2$ 和 $\boldsymbol{F}_3$,首先,将 $\boldsymbol{F}_3\in\mathbb{R}^{256\times7\times7\times7}$ 通过卷积运算和三线性插值上采样,依次在通道和空间维度上对齐至 $\boldsymbol{F}_2$ 的特征空间,再沿通道维度拼接,得到多尺度层级特征图 $\boldsymbol{M}_1\in\mathbb{R}^{256\times14\times14\times14}$。类似地,将 $\boldsymbol{M}_1$ 和 $\boldsymbol{F}_1$ 合并得到层级特征图 $\boldsymbol{M}_2\in\mathbb{R}^{128\times28\times28\times28}$。其次,对 $\boldsymbol{M}_1$ 和 $\boldsymbol{M}_2$ 均

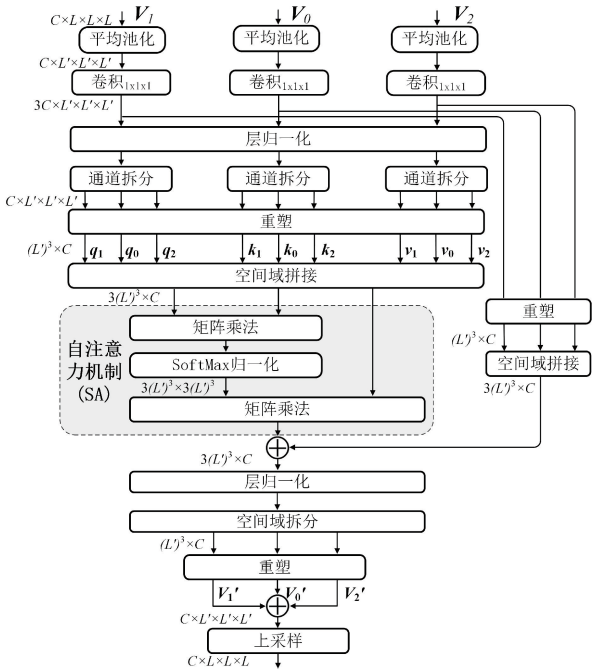


图 3　CVA 模块结构

Fig. 3　Structure of the CVA module

采用卷积运算来实现多层级特征图的通道和空间域交互，以促进信息深度融合。接着，将 $\boldsymbol{M}_2$ 通过池化操作对齐至 $\boldsymbol{M}_1$ 的特征空间后，依次经过通道拼接、空间池化下采样和卷积运算进一步融合，输出多层级特征图 $\boldsymbol{M}_3 \in \mathbb{R}^{256\times7\times7\times7}$。最终，融合 3 种尺度信息的 $\boldsymbol{M}_3$ 通过逐元素相加方式与 $\boldsymbol{F}_3$ 建立残差连接，输出多尺度信息表示。

1.4　Transformer 编码器

本研究利用 Transformer 块中的自注意力机制，加权融合卷积深层特征图中的局部信息，建立全脑 sMRI 的全局特征表示。具体来说，先将卷积网络输出特征图的通道维度映射为 d。经过特征图展平和重塑后，得到对应全脑空间 N 个局部区域的 N 个 d 维视觉 Token 向量。再与分类头向量拼接，并嵌入位置信息，得到 Token 向量矩阵 $\boldsymbol{Z} \in \mathbb{R}^{(N+1)\times d}$。如图 4 所示，Transformer 编码器由 l 个块组成，每个块主要由多头自注意力(multi-head self attention，MSA)层和多层感知器(multi-layer perceptron，MLP)构成。给定编码器的输入 Z。第 n 个 Transformer 块的输出可表示为 Z_n：

$$Z_n = MLP_n(LN(Z'_n)) + Z'_n, n \in [1,\cdots,l] \tag{6}$$

$$Z'_n = MSA_n(LN(Z_{n-1})) + Z_{n-1} \tag{7}$$

$$MSA_n = Concat(SA_n{}^1,\cdots,SA_n{}^h)W^o \tag{8}$$

$$SA_n^i = Soft\max(\frac{Q_n^i(K_n{}^i)^T}{\sqrt{d}})V_n^i, i \in [1,\cdots,h] \tag{9}$$

$$Q_n = Z_{n-1}W^Q, K_n = Z_{n-1}W^K, V_n = Z_{n-1}W^V \tag{10}$$

其中，i 表示第 i 个自注意力头，W^Q、W^K 和 W^V 对应生成查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 矩阵的权重参数。

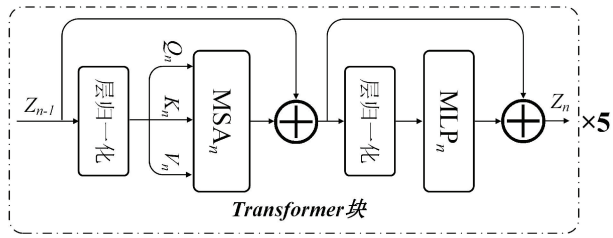


图 4　Transformer 编码器

Fig. 4　Transformer encoder

2　实　　验

2.1　数据集

本研究使用的数据集来源于阿尔茨海默病神经影像倡议(ADNI)公共数据库(https://adni.loni.usc.edu/)，共涵盖 1 504 名受试者的 4 796 张 1.5T/3T T1 加权 sMRI 影像。数据集相关信息如表 1 所示。为了避免影像特征空间未对齐以及采集噪声的影响，本研究采用标准化的影像预处理流程，包括结构自适应非局部均值去噪、非均匀强度归一化、空间配准和颅骨剥离。这些步骤均由 MATLAB 的统计参数映射工具箱(SPM12)(http://www.fil.ion.ucl.ac.uk/spm/software/spm)实现。

表 1　数据集人口统计学和临床相关信息

Table 1　Demographic and clinical information of the dataset

类别	数量(男/女)	年龄(mean±std)	MMSE(mean±std)	CDR(mean±std)
AD	197/167	75.35±7.85	23.28±2.26	0.77±0.27
NC	204/234	74.13±5.79	29.03±1.12	0.03±0.01
sMCI	273/191	73.35±7.72	27.59±1.86	0.49±0.06
pMCI	154/109	74.22±7.06	26.55±1.89	0.50±0.08

注：正常对照组(normal control，NC)，简易精神状态检查评分(MMSE)，临床痴呆评分(CDR)。

2.2　实验参数设定

本研究实验均在配备 NVIDIA RTX A5000 24 G GPU 的 Linux 工作站上完成，使用深度学习框架 PyTorch 1.8.1(Python 3.8.0)搭建模型。在实验前将数据集按 4∶1∶1比例随机分为训练、验证和测试子集，以独立测试集上的结果作为最终性能。为增加训练数据集多样性，在训练进程中加载影像时应用了包含沿 x、y 和 z 轴的随机旋转(±15°)、随机翻转和大小为 10 px×10 px×10 px 的随机遮掩，随机概率均设置为 0.5。

本研究分别评估了 AD 分类任务(AD vs. NC)和 MCI 转化预测任务(sMCI vs. pMCI)，采用了二进制标签平滑交叉熵损失函数优化参数。迭代次数设为 100，批次大小为 8，使用权重衰减为 1×10^{-4} 的 Adam 优化器，将学习率设为 5×10^{-5}。在模型超参数中，原始视图 $\boldsymbol{V}_0$ 对应为全脑

sMRI 的冠状面视图,$\boldsymbol{V}_1$ 和 $\boldsymbol{V}_2$ 分别对应矢状面和轴面视图特征图。3 个阶段中,CVA 的平均池化核 K_p 分别设置为 2、2 和 1。在 Transformer 编码器中,Transformer 块数量 l、嵌入维度 d 和 MSA 中的头数 h 分别设置为 5、576 和 12。

本研究使用了几种医学分类任务常见的性能指标,包含准确性(accuracy, ACC)、灵敏度(sensitivity, SEN)、特异性(specificity, SPE)和受试者工作特征曲线下面积(area under receiver operating characteristic curve, AUC)。对应的数学表达式如下:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$SEN = \frac{TP}{TP + FN} \tag{12}$$

$$SPE = \frac{TN}{TN + FP} \tag{13}$$

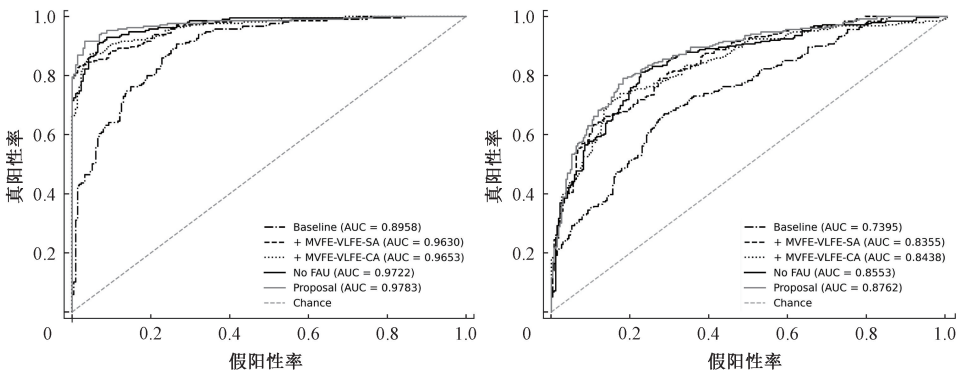
其中,TP、TN、FP 和 FN 分别表示预测真阳性(true positive)、真阴性(true negative)、假阳性(false positive)和假阴性(false negative)。本实验将 AD 和 pMCI 受试者设为正样本类别,NC 和 sMCI 受试者为负样本类别,因而 SEN 和 SPE 分别反映了模型识别患病和正常样本的性能,AUC 则反映了模型对病症类别识别的整体性能。

2.3 消融实验

1)MVFE 和 FAU 的有效性

本节通过一系列消融实验验证 MVFE 的有效性。具体来说,将移除 MVFE 后的模型作为基线(记作"Baseline"),与依次在 VLFE 中添加空间注意力(记作"+MVFE-VLFE-SA")、添加通道注意力(记作"+MVFE-VLFE-CA")以及添加 CVA 模块后的实验结果(Proposal)进行比较。此外,还通过移除 FAU 来验证其性能增益,记作"No FAU"。

消融实验结果如图 5 和表 2 中 ROC 曲线所示,与移除 MVFE 的基线模型相比,首先,引入混合注意力后的 VLFE 使网络在两个分类任务中均取得了约 10%的 ACC 提升,这表明通过动态调整通道和空间权重,使 VLFE 能够自适应地过滤冗余信息,并聚焦关键病变区域,实现性能提升。其次,添加 CVA 后进一步提高了多项指标值,反映了跨视图注意力加权融合能够促进了视图信息之间相互学习,增强了病灶结构的语义理解。最后,添加 FAU 使得 MCI 转化预测任务的所有指标项均获得 0.09%至 2.09%的提升,说明在 MVFE 之前设置 FAU 有效增强了细节信息的聚合能力,促进了病症识别能力。



(a) AD vs. NC (b) sMCI vs. pMCI

图 5 消融 MVFE 和 FAU 的 ROC 曲线

Fig. 5 ROC curves for ablation of MVFE and FAU

表 2 MVFE 和 FAU 的消融实验结果

Table 2 Ablation study results of the MVFE and FAU

方法	FAU	CVA	MVFE		AD vs. NC				sMCI vs. pMCI			
			VLFE-SA	VLFE-CA	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE
Baseline	√				89.58	82.07	89.20	74.88	73.95	69.33	67.48	71.05
+MVFE-VLFE-SA	√		√		96.30	91.43	83.10	94.23	83.55	77.12	73.57	83.33
+MVFE-VLFE-CA	√		√	√	96.53	92.59	92.94	93.62	84.38	79.26	76.45	**84.64**
No FAU		√	√	√	97.22	93.38	92.42	94.26	85.53	80.72	83.77	78.22
本文	√	√	√	√	**97.83**	**94.05**	**93.25**	**95.68**	**87.62**	**81.59**	**85.10**	78.31

为了进一步讨论 MVFE 的影响,本研究使用 Grad-CAM[15] 技术对 MVFP-Net 及移除 MVFE 后的简化版本进行热力图可视化。网络末端 Transformer 块的第一个 LN 层被设为可视化目标层,从 AD 和 pMCI 组受试者中分别随机选择 3 名受试者 sMRI,生成了热力图和对应的预测置信度分数 P。如图 6 所示,与简化版本相比,MVFP-Net 的热力值分布覆盖了更多 AD 相关脑区(如海马区、脑室附近)。此外,添加 MVFE 还能够提高模型类别预测置

信度。例如较为困难的 pMCI 受试者＃1 和＃3 的置信度分数出现明显提升，解释了 MCI 转化预测任务具有高灵敏度的原因。综上结果说明 MVFE 通过感知和利用 sMRI 不同视图方向的空间结构信息，能够增强模型对病灶区域的定位能力，从而增强模型性能。

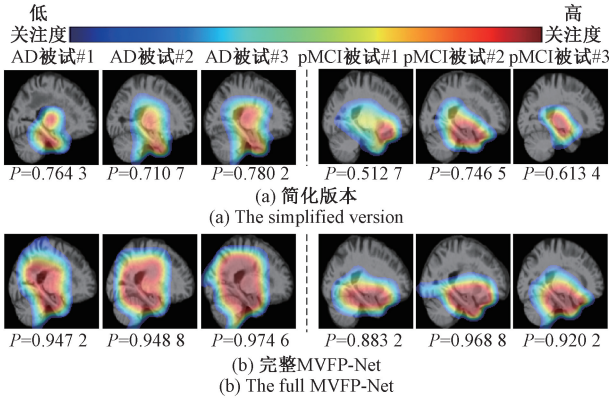


图 6　部分样本 sMRI 的热力图和置信度分数

Fig. 6　Heat maps and confidence scores for sMRI of selected samples

2）CMSF 子网络的有效性

本节通过消融实验以验证该子网络的有效性，对比了本模型和移除 CMSF 子网络后的性能表现。如图 7(a)和(b)所示，CMSF 子网络显著提高了两个任务的大多数性能指标。即使在更困难 MCI 转化预测任务中，MVFP-Net 依然表现出显著更好的性能，ACC、AUC 和 SEN 分别提高了 1.26%、1.44%和 4.90%。这些结果反映了 CMSF 子网络通过深度融合不同层级的多尺度空间信息，生成了更具语义丰富度的判别特征，从而为模型决策提供了更有效的支持。

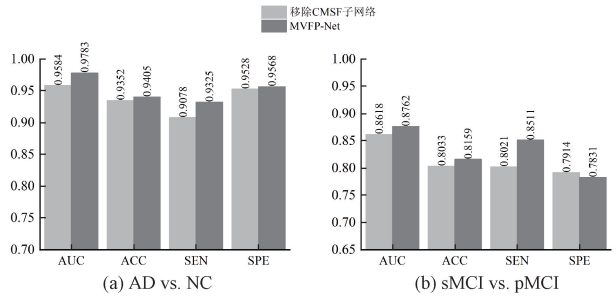


图 7　CMSF 子网络的消融实验结果

Fig. 7　Ablation results of the CMSF subnetwork

3）混合架构模型的有效性

本实验验证 CNN 和 Transformer 混合架构网络的有效性，以及 Transformer 编码器结构的优化设计。本实验以 3DVIT[6]的标准版本（深度为 12，嵌入维度为 576）作为纯视觉 Transformer 的对比模型，记作“3DVIT”。同时，将编码器替换为由 MLP 构成的分类器（隐藏节点为 512），构建基于纯 CNN 的对比模型，记作“MVFP-CNN”。

结果如表 3 和图 8 中 ROC 曲线所示，所提出的 MVFP-Net 在 AD 分类和 MCI 转化预测任务中的大多数性能指标上，明显优于纯 CNN 或纯 Transformer 架构的模型。这表明，MVFP-Net 有效结合了 CNN 在图像局部细节提取方面的优势和 Transformer 编码器在序列分析中的优势，将全脑 sMRI 中的局部表示建模为全局表示，进一步提升了分类性能。此外，本文在 AD 分类任务上通过消融编码器深度 l 和嵌入维度 d 对结构进行优化。如表 4 所示，根据结果确定了当 d 和 l 分别等于 576 和 5 时，模型具有最优性能。

表 3　与基于纯 Transformer 和纯 CNN 架构网络的对比结果

Table 3　Comparison results with pure Transformer and pure CNN architecture based networks　%

方法	Transformer 编码器	CNN	AD vs. NC				sMCI vs. pMCI			
			AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE
3DVIT	√		72.80	71.46	67.60	75.35	65.61	63.48	59.35	67.29
MVFP-CNN		√	95.42	90.22	86.97	94.43	81.98	75.59	67.89	**82.70**
本文方法	√	√	**97.83**	**94.05**	**93.25**	**95.68**	**87.62**	**81.59**	**85.10**	78.31

2.4　对比实验

1）与混合架构方法对比

为了突显本方法在类似方法中的优势，本实验复现了 3 种基于 CNN 和 Transformer 的混合架构方法，并进行了性能对比。复现结果均基于原论文网络框图或提供的官方代码，且均在本文数据集上评估。如表 5 所示，与其他 3 种方法中的最佳性能相比，所提出的 MVFP-Net 在 AD 分类和 MCI 转换预测任务中的 ACC 性能分别提高了 1.63%和 2.78%。尽管 sMCI 和 pMCI 患者的脑萎缩程度相似且低于 AD 患者[2]，所提出方法仍然在多项指标上反映出更优越的性能。这些方法中，文献[5]在 3DCNN 中引入了并行多尺度特征融合分支，以提取病变结构的多尺度信息，并将年龄信息嵌入特征图，利用 Transformer 编码器捕捉年龄关联性用以改善模型决策。然而，密集并行多尺度融合策略将影响特征融合效率，将年龄信息与特征图整合可能会影响局部细节的表达。文献[8]使用 3DCNN 学习全脑 sMRI 空间结构信息后，再获取特征图的 2D 切片微调预训练后的 2DCNN，最终由 Transformer 模块建模切

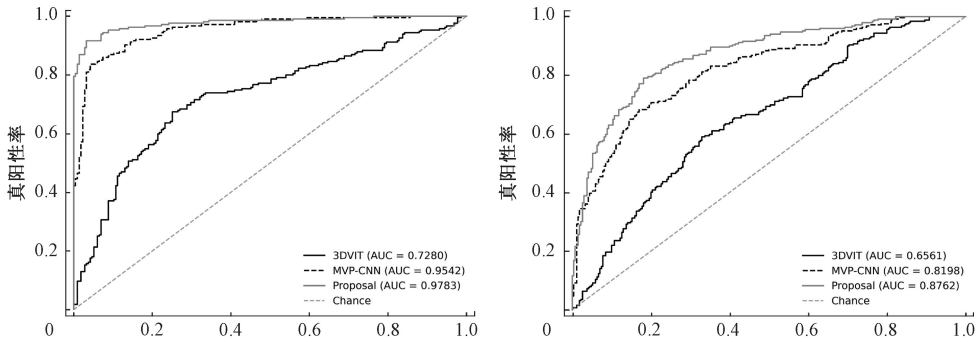


图 8 混合架构网络消融 ROC 曲线

Fig. 8 Hybrid architecture network ablation ROC curve

表 4 Transformer 编码器嵌入维度 d 和深度 l 的消融结果

Table 4 Ablation results for Transformer encoder embedding dimension d and depth l

固定 $d=576$		固定 $l=5$	
深度 l	ACC/%	嵌入维度 d	ACC/%
1	92.13	192	88.68
3	92.62	384	93.72
5	**94.05**	**576**	**94.05**
7	93.31	768	93.36
9	93.04	960	92.21

片间依赖关系。文献[9]则利用 VGG-16 提取 2D 切片特征,并采用窗口自注意编码器来建立切片图像内局部病变结构的空间语义联系。相比之下,本方法利用 3DCNN 捕获来自全脑 sMRI 中的空间病变信息,并由 Transformer 编码器将特征表示增强为全脑空间水平。此外,本文方法采用级联式逐层级复用的多尺度融合策略,在卷积主干上基于少量的卷积和池化操作,实现更深入、高效的特征融合。利用 sMRI 的多视图信息增强复杂病变新信息表征,在不引入额外监督信息的条件下实现了性能提升。

表 5 与基于 CNN 和 Transformer 的混合架构方法的对比结果

Table 5 Comparison results with hybrid architecture approach based on CNN and Transformer

对比	AD vs. NC/%				sMCI vs. pMCI/%			
方法	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE
文献	93.43	89.37	82.17	91.22	78.92	73.98	62.20	76.08
[5]	[94.21±1.01]	[91.19±0.79]	[83.19±1.56]	[92.89±1.09]	[80.23±1.11]	[75.31±1.04]	[62.13±1.25]	[78.33±1.53]
文献	96.03	92.42	92.05	91.97	83.72	78.81	81.42	73.08
[8]	[96.88±1.52]	[93.88±0.92]	[93.11±1.74]	[92.42±1.51]	[84.82±0.68]	[81.19±1.28]	[82.67±1.72]	[75.19±1.12]
文献	95.59	91.23	85.26	94.82	82.87	78.32	**85.66**	68.92
[9]	[95.81±1.56]	[92.30±1.24]	[87.72±1.62]	[95.38±0.75]	[84.09±1.20]	[79.21±0.96]	[87.32±1.32]	[70.92±2.02]
本文	**97.83**	**94.05**	**93.25**	**95.68**	**87.62**	**81.59**	85.10	**78.31**
	[98.17±1.15]	[95.10±1.08]	[94.38±1.21]	[95.89±0.85]	[88.49±1.25]	[82.72±0.98]	[85.79±1.43]	[80.49±0.93]

注:[mean±std]表示五折交叉验证的均值和标准差

2)与其他深度学习方法的对比

本节总结了几种在 ADNI 数据集进行 AD 诊断研究的其他类型方法。这些方法根据 sMRI 影像分析角度进行区分,主要包括两种 2D 切片级方法[16-17]、两种 3D 影像块级方法[18-19]和两种全脑影像级方法[11,20]。

这些方法所报告的结果如表 6 所示,与非全脑影像级的方法相比,本文通过结合 3D 卷积网络和自注意力机制的优势,利用了分布在全脑 sMRI 的病灶结构的语义信息和关联关系,在大部分指标项反映了其具有性能优势。在 MCI 转化预测任务的精度略微落后于方法[18],这可能与 3D 影像块级方法在获取鉴别性影像块过程中提供了先验知识有关。与全脑影像级的方法相比,在相似规模的数据集上,本文除了 MCI 转化预测任务的较低 SPE 指标外,在其他指标项表现出 1.23%到 10.6%的提升,反映了结合全脑多视图信息对模型性能提升的促进作用。本文通过模仿临床实践中多角度、多层次的 AD 患者 sMRI 影像分析方式,设计了多视图信息感知和逐阶段多尺度特征融合策略,使模型能够更精确识别患者大脑中复杂的病变结构,解释了在其他类型方法中具备的性能优势。

表 6　与其他使用 sMRI 的深度学习方法性能对比

Table 6　Performance comparison with other deep learning methods using sMRI

参考方法	类型	受试者分布	AD vs. NC/%				sMCI vs. pMCI/%			
		(AD/NC/sMCI/pMCI)	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE
文献[16]	2D 切片,2DCNN	187/229/181/138	89.72	90.36	93.94	83.78	62.50	63.49	57.56	64.29
文献[17]	2D 切片,2DCNN	397/419/268/252	96.70	93.40	92.9	93.80	87.30	81.10	83.70	78.70
文献[18]	3D 影像块,3DCNN	353/650/232/172	96.70	92.00	92.00	91.90	85.70	**81.90**	81.80	81.60
文献[19]	3D 影像块,3DCNN	389/400/232/172	96.50	92.40	91.00	93.80	85.10	80.20	77.10	**82.60**
文献[11]	全脑影像,3DCNN	326/413/470/242	95.00	91.10	88.80	91.40	78.90	80.10	52.00	85.60
文献[20]	全脑影像,3DCNN	408/773/445/269	96.60	91.60	89.20	93.50	81.40	80.20	74.50	82.10
本文	全脑影像,3DCNN+Transformer	335/441/465/263	**97.83**	**94.05**	**93.25**	**95.68**	**87.62**	81.59	**85.10**	78.31

3)模型计算效率对比

本节比较了两种主流纯卷积网络、纯视觉 VIT 和两种混合架构方法的参数量、浮点计算量(floating point operations, FLOPs)与对应的 ACC,以评估本文方法计算效率,结果如表 7 所示。本文方法的复杂度低于 3DVGG16、3DResNet18、3DVIT 以及文献[5],并取得了更高的 ACC。尽管文献[9]中的混合架构方法使用 2D 卷积提取浅层信息减少了参数量和 FLOPs,但本文方法利用了更多的全脑结构信息以支持模型决策,ACC 的显著优势说明了其具备更可靠的诊断性能。

表 7　模型计算效率对比

Table 7　Comparison of model computational efficiency

方法	参数量/MB	FLOPs/GB	ACC/%	
			AD vs. NC	sMCI vs. pMCI
3DVGG16	139.56	59.65	74.59	65.00
3DResNet18	55.25	39.66	84.23	68.84
3DVIT	88.20	38.42	71.46	63.48
文献[5]	34.69	41.98	89.37	73.98
文献[9]	**23.18**	**30.85**	91.23	78.32
本文	27.15	34.35	**94.05**	**81.59**

3　模型可解释性

本节通过生成平均类激活图来解释模型决策所依赖的 AD 相关生物标志物。由于浅层输出包含更多细节信息,能够揭示更多潜在生物标志物,将模型第二阶段的输出特征图用以计算类激活图。在测试集中随机选择了 50 名 AD 和 pMCI 受试者的 sMRI,经过类激活图计算、求和取平均后叠加在模板图像,获得平均激活图,并以由蓝至红的颜色映射反映病变程度。

如图 9 所示,通过比较 sMRI 模板图像(虚线圈标示了已有结论中的 AD 相关区域)、AD 和 pMCI 受试者的平均类激活图在冠状面、轴状面和矢状面的切片图像,发现了本网络决策所依赖的大脑区域包括颞叶、顶叶等皮质区域,以及海马区、杏仁核和海马旁回等皮层下结构。由于这些区域与记忆、认知和情绪功能密切相关,进而能够解释伴随 AD 病变中出现的临床症状,并与现有研究结果一致[5,8]。除了脑组织外,脑室结构附近的较高激活值分布,证明了脑萎缩引起的脑室扩大也是 AD 的重要标志物[19,21]。可视化结果直观反映了模型能够有效识别 AD 相关病变脑区,并获取丰富的判别信息为病症诊断提供更可靠的决策支持。

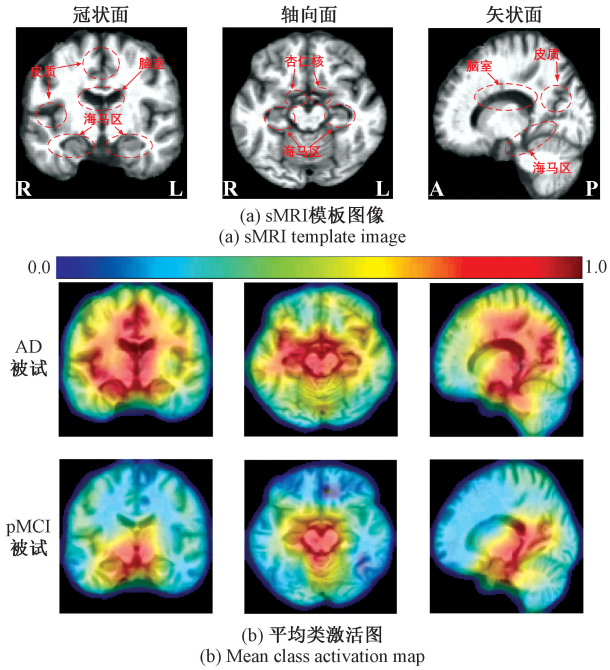


图 9　基于平均类激活图的模型可解释性结果

Fig. 9　Model interpretability results based on mean class activation maps

4　结　　论

本文提出了一种结合 CNN 和 Transformer 的深度学习网络,在 sMRI 上进行 AD 诊断研究。在卷积网络中,设

计了多视图特征编码器感知并融合来自 sMRI 的轴向、冠状和矢状视图方向的互补信息,并利用跨视图注意力加权融合策略来增强复杂病变信息的表征能力,此外,通过逐层级联方式进行融合多尺度特征,丰富了卷积深层特征图的空间和语义信息。在 Transformer 编码器建立了全脑病变信息的全局表示。实验结果表明,本文方法结合了混合架构的优势,通过引入多视图信息增强了分类性能。在 AD 病症诊断相关任务中的评估结果显示了其显著的性能优势。未来研究将致力于在更大规模的多中心异构数据集上优化模型的泛化能力,并通过引入稀疏注意力机制提升计算效率,重点聚焦关键脑区,以降低计算开销。

参考文献

[1] MASTERS C L, BATEMAN R, BLENNOW K, et al. Alzheimer's disease[J]. Nature Reviews Disease Primers, 2015, 1(1): 1-18.

[2] BETTER M A. Alzheimer's disease facts and figures[J]. Alzheimer's Dement, 2024, 20: 3708-3821.

[3] CUIJPERS Y, LENTE V H. Early diagnostics and Alzheimer's disease: Beyond 'cure' and 'care'[J]. Technological Forecasting and Social Change, 2015, 93: 54-67.

[4] ZHANG F, TIAN S J, CHEN S P, et al. Voxel-based morphometry: Improving the diagnosis of Alzheimer's disease based on an extreme learning machine method from the ADNI cohort [J]. Neuroscience, 2019, 414: 273-279.

[5] GAO X Y, CAI H J, LIU M H. A hybrid multi-scale attention convolution and aging transformer network for Alzheimer's disease diagnosis[J]. IEEE Journal of Biomedical and Health Informatics, 2023, 27(7): 3292-3301.

[6] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. ArXiv preprint arXiv:2010.11929, 2020.

[7] HOANG G M, KIM U H, KIM J G. Vision transformers for the prediction of mild cognitive impairment to Alzheimer's disease progression using mid-sagittal sMRI [J]. Frontiers in Aging Neuroscience, 2023, 15: 1102869.

[8] JANG J, HWANG D. M3T: Three-dimensional medical image classifier using multi-plane and multi-slice transformer [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 20718-20729.

[9] HU ZH T, LI Y Y, WANG ZH, et al. Conv-Swinformer: Integration of CNN and shift window attention for Alzheimer's disease classification[J]. Computers in Biology and Medicine, 2023, 164: 107304.

[10] 李志刚,胡德安,纪勇,等. 阿尔茨海默病多病程自适应筛查模型的构建[J]. 仪器仪表学报, 2023, 44(1): 182-189.
LI ZH G, HU D AN, JI Y, et al. Construction of a multiple disease courses adaptive screening model for Alzheimer's disease[J]. Chinese Journal of Scientific Instrument, 2023, 44(1): 182-189.

[11] CHEN L, QIAO H ZH, ZHU F. Alzheimer's disease diagnosis with brain structural MRI using multiview-slice attention and 3D convolution neural network[J]. Frontiers in Aging Neuro-science, 2022, 14: 871706.

[12] WANG Y Q, LI Z H, MEI J R, et al. SwinMM: Masked multi-view with swin transformers for 3D medical image segmentation [C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Cham: Springer Nature Switzerland, 2023: 486-496.

[13] YAO Y H, YANG J Y, SUN H J, et al. DeepGraFT: A novel semantic segmentation auxiliary ROI-based deep learning framework for effective fundus tessellation classification[J]. Computers in Biology and Medicine, 2024, 169: 107881.

[14] WANG Q L, WU B G, ZHU P F, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11534-11542.

[15] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]. IEEE International Conference on Computer Vision, 2017: 618-626.

[16] KANG W J, LIN L, ZHANG B W, et al. Multi-model and multi-slice ensemble learning architecture based on 2D convolutional neural networks for Alzheimer's disease diagnosis [J]. Computers in Biology and Medicine, 2021, 136: 104678.

[17] CAO G P, ZHANG M L, WANG Y P, et al. End-to-end automatic pathology localization for Alzheimer's disease diagnosis using structural MRI[J]. Computers in Biology and Medicine, 2023, 163: 107110.

[18] ZHANG X, HAN L X, HAN L H, et al. SMRI-PatchNet: A novel efficient explainable patch-based deep learning network for Alzheimer's disease diagnosis with Structural MRI [J]. IEEE Access, 2023, 11:108603-108616.

[19] ZHU W Y, SUN L, HUANG J SH, et al. Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI[J]. IEEE Transactions on Medical Imaging, 2021, 40(9): 2354-2366.

[20] HAN K F, HE M, YANG F, et al. Multi-task multi-level feature adversarial network for joint Alzheimer's disease diagnosis and atrophy localization using sMRI[J]. Physics in Medicine & Biology, 2022, 67(8): 085002.

[21] PAN D, ZENG AN, YANG B Y, et al. Deep learning for brain MRI confirms patterned pathological progression in Alzheimer's disease [J]. Advanced Science, 2023, 10(6): 2204717.

作者简介

吴慧东,硕士研究生,主要研究方向为深度学习在阿尔兹海默病诊断的应用。
E-mail:1505261567@qq.com

刘立程(通信作者),博士,副教授,硕士生导师,主要研究方向为信号处理、模式识别、人工智能在视频图像中的应用。
E-mail:celcliu@gdut.edu.cn

潘丹,博士,副教授,硕士生导师,主要研究方向为人工智能在医学图像处理的应用。
E-mail:pandan@gpnu.edu.cn