基于文本引导下的多模态医学图像分析算法

樊 琳[1,2,3,4],龚 勋[1,2,3,4*],郑岑洋[1,2,3,4]

(1. 西南交通大学计算机与人工智能学院,四川成都 611756;2. 可持续城市交通智能化教育部工程研究中心,四川成都 611756;
3. 综合交通大数据应用技术国家工程实验室,四川成都 611756;
4. 四川省制造业产业链协同与信息化支撑技术重点实验室,四川成都 611756)

摘 要: 结合胃镜超声和白光内镜可以更准确地识别胃肠道间质瘤. 但是现有的多模态方法往往仅关注于图像特征,忽略了诊断文本信息中所包含的语义信息对于精确理解和诊断医学图像的重要性. 为此,本文提出一种新的基于文本引导下的多模态医学图像分析算法框架(Text-guided Multi-modal Medical image analysis framework,TMM-Net). TMM-Net使用多阶段的诊断文本来引导模型学习,以提取图像中的关键诊断信息特征,然后通过交叉模态注意力机制促进多模态特征之间的交互. 值得注意的是,TMM-Net通过预测病变属性来模拟临床诊断过程,从而增强了可解释性. 验证实验在两个中心包含10 025个模态数据对的数据集上进行. 结果表明,该方法相比目前最优的GISTs诊断方法精度提升7.7%,同时获得了最高的(Area Under the Curve,AUC)值:0.927,其可解释性可以更好地适合临床需求.

关键词: 多模态融合;模型可解释性;图像-文本匹配;胃肠道间质瘤;胃镜超声;白光内镜

基金项目: 国家自然科学基金(No.62376231);四川省重点研发项目(No.2023YFG0267);四川省卫生健康委员会科技项目(No.23LCYJ022)

中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112(2024)07-2341-15

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20231135

A Multi-Modal Medical Image Analysis Algorithm Based on Text Guidance

FAN Lin[1,2,3,4], GONG Xun[1,2,3,4*], ZHENG Cen-yang[1,2,3,4]

(1. *School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, Sichuan* 611756, *China*;
2. *Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, Chengdu, Sichuan* 611756, *China*;
3. *National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Chengdu, Sichuan* 611756, *China*;
4. *Manufacturing Industry Chains Collaboration and Information Support Technology Key Laboratory of Sichuan Province, Chengdu, Sichuan* 611756, *China*)

Abstract: Combining gastroscopy ultrasound and white light endoscopy can improve the accuracy of identifying gastrointestinal stromal tumors (GISTs). However, existing multi-modal methods often focus solely on image features and overlook the semantic relevance contained in diagnostic textual information, which is crucial for precise understanding and diagnosis of medical images. To address this issue, we propose a novel text-guided multi-modal medical image analysis framework (TMM-Net). TMM-Net extracts key diagnostic information features from images through a multi-stage guided model of diagnostic text, and then promotes the interaction of multi-modal features through cross-modal attention mechanisms. Notably, TMM-Net simulates the clinical diagnostic process by predicting lesion attributes, enhancing interpretability. Validation experiments were conducted on a dataset consisting of 10 025 modality data pairs from two centers. The results show that the proposed method achieves a 7.7% improvement in accuracy compared to the current state-of-the-art GISTs diagnostic method, with the highest AUC (Area Under the Curve) value of 0.927, and its interpretability may better suit clinical needs.

Key words: multi-modal fusion; model interpretability; image-text matching; gastrointestinal stromal tumor; gastroscopic ultrasound; white light endoscopy

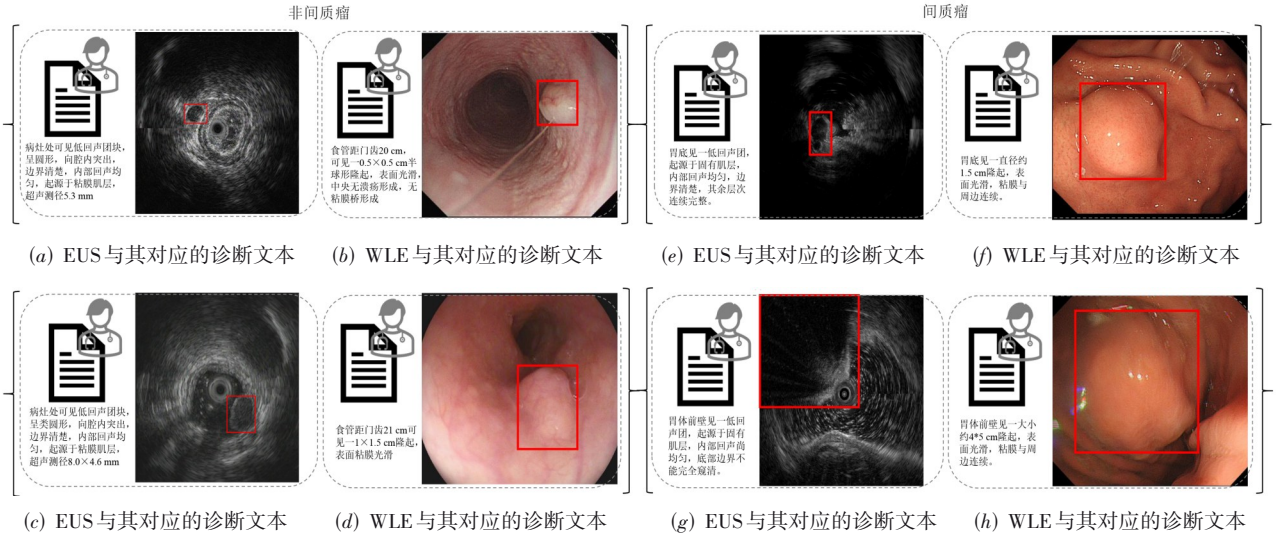
Foundation Item(s): National Natural Science Foundation of China (No.62376231); Sichuan Science and Technology Program (No.2023YFG0267); Science and Technology Project of Sichuan Provincial Health Commission (No.23LCYJ022)

收稿日期:2023-12-05;修回日期:2024-04-20;责任编辑:王天慧

*通讯作者:龚勋

1 引言

胃肠道间质瘤(GastroIntestinal Stromal Tumors,GISTs)属于消化道黏膜下肿瘤(SubEpithelial Lesions,SELs)中的一种,起源于消化道黏膜上皮以下组织.随着大众健康意识增强及消化内镜诊疗技术普及,SELs的检出率明显提高[1].尽管大部分SELs为良性病变,但其中胃肠道间质瘤都具有恶性潜能(约占10%~30%)[2].然而,GISTs的病理类型和生物学行为多样性增加了其诊治的难度,不确定或错误的GISTs决策将会导致不必要的病变切除、反复的内镜随访或患者的情绪困扰,造成极大的医疗负担.所以,GISTs的准确分化具有重要临床意义.内镜超声(Endoscopic UltraSono-graphy,EUS)联合白光内镜(White Light Endoscope,WLE)是目前SELs最优最简的诊断模式,2023年最新内镜诊治指南[3]建议根据EUS和WLE联合诊断进行GISTs的恶性风险评估.但是,GISTs与平滑肌瘤等具有重叠的成像特征,内镜医师很难精准诊断其病理性质(如图1所示,其中一对数据用大括号表示;红色的矩形表示病变的大小和位置;与图像对应的诊断信息列在图像左侧),而且诊断结果会受到主观性的影响,取决于临床医生的经验和技能.根据病变性质不同,医师诊断下GISTs的EUS与病理的符合率仅63%,WLE下SELs诊断特异性仅为29%[4].



(a) EUS与其对应的诊断文本 (b) WLE与其对应的诊断文本 (e) EUS与其对应的诊断文本 (f) WLE与其对应的诊断文本

(c) EUS与其对应的诊断文本 (d) WLE与其对应的诊断文本 (g) EUS与其对应的诊断文本 (h) WLE与其对应的诊断文本

图1 多模态信息对示例

基于深度学习算法的计算机诊断辅助系统(Computer Aided Diagnostic system,CAD)被开发出来解决医学图像分类问题,并且已被初步应用于食道癌[5]、胃癌[6]、结直肠息肉[7]和胰腺疾病[8]的EUS诊断领域,包括鉴别SELs中GISTs与非GISTs的研究[9~15].虽然CAD在GISTs诊断中取得一定的进展[16,17],但现有相关研究仍存在诸多局限性.(1)以EUS的单模态病理图像数据为主,对GISTs的识别准确性并不符合临床实践的要求.实际上,联合EUS和WLE能提供强有力的互补诊断信息,将会更有助于GISTs的诊断.(2)现有医学多模态方法大多采用将不同模态图像信息直接融合[18,19],融合空间具有不稳定性,难以拓展,甚至可能导致多模态特征的混淆和丢失.因此近年来一些基于注意力机制融合的方法[20,21]已被提出,但仍存在一些缺点,例如无法明确区分关键和不相关信息的融合,而且难以解释和理解模型在决策过程中的具体依据.(3)医学图像数据通常是高维度、高复杂度的数据,大多多模态方法仅依靠图像本身的视觉特征进行分析和理解,存在一定的局限性.(4)基于深度学习的多模态CAD模型缺乏可解释性.

受人类对图像学习过程的启发:通过解释自然语言中的可知特征来帮助理解图像,而不是直接从图像中学习未知特征,我们提出了一种利用医学图像诊断报告中的语义信息来引导多模态图像特征学习及融合的框架,目的是学习多模态融合下不同模态数据的强特征表示.诊断报告中包含了医生对相关影像的描述信息等,这些信息可以帮助模型更好地理解和分析医学图像.例如,在肿瘤诊断中,医生可以通过诊断文本中的病灶位置、大小、形态特征等信息来诊断病变类型和严重程度,而这些信息可以用来指导模型学习对应的图像特征.这种做法对医学多模态研究有三个方面的好处.(1)规范诊断报告比培训临床医师来规范图像的质量更容易,这有助于形成通用SELs语料库,通过使用语料库中的诊断信息来引导更多模态或更多中心的图像特征学习.相比于现有的固定模态和中心的方法,提出的TMM-Net有在多中心和多模态条件下拥有更好的拓展性和通用性.(2)诊断报告和图像之间存在特征

相关性,通过诊断报告的指导,可以将图像特征的不同形式和诊断信息中固定的描述联系起来,从而可以在不同模态图像上实现灵活的迁移,保证了图像特征提取的精确度,同时增加了不同模态图像特征融合的稳定性.(3)利用常见的诊断报告中的专业知识来指导和规范深度网络的训练,可以从医生的角度提供更可靠和准确的自动诊断.并且通过引导模式训练好的模型可以只接受图像输入,即文本和图像在训练和推理时取消了相互依赖嵌入.在实际应用中,这种模型更符合现实世界的需求,因为测试图像可能很难或根本无法获得相应的文本说明,特别是在医学图像领域.

本文提出一种多阶段文本引导下的多模态图像数据特征匹配融合框架,可以通过专业诊断报告的引导从不同模态医学图像中学习稳定的、可解释的特征表达并融合诊断.诊断报告是医师根据相关图像提取的核心可辨特征(如图1所示),使用规范性的文字来表达[22],并且在多中心、多模态环境下具有不变性,利用文本引导可便于模型在不同中心和模式之间轻松扩展.该框架包括三个阶段,分别为特征级别的引导阶段、原型级别的引导阶段和交叉模态注意力特征融合阶段.特征级别的引导阶段实现了诊断报告特征和图像特征在表示空间上的对齐,建立文本-图像对之间的对应关系.原型级别的引导阶段通过诊断文本特征与不同模态分支上的图像特征原型级别之间的匹配,实现文本特征引导下精确的图像诊断关键特征的学习,同时建立图像特征原型预测属性分支,来模拟临床过程.交叉模态注意力特征融合阶段通过WLE特征与EUS特征相互学习.论文的主要贡献如下:

(1)建立端到端的诊断报告、EUS和WLE图像的多模态联合训练模型.所提出的多模态网络明显优于GISTs诊断中单独使用任何单一信息源的方法,以及现有多模态框架.

(2)提出一种多阶段的诊断信息引导图像关键特征学习的多模态融合策略.该策略可通过自然语言信息引导网络对EUS和WLE图像关键诊断特征的学习,达到图像特征之间的模态对齐与领域对齐.并且该模型可拓展多中心数据.

(3)提出一种可解释性的CAD系统.该模型通过临床诊断文本来指导网络的训练,并且通过病变属性的预测来模拟临床诊断过程.根据这些信息,临床医生可以更好的理解AI CAD的推理,并利用其预测来协助诊断.

(4)在一个包含228个病例3 903个多模态信息对和一个包含58个病例6 122个多模态信息对的具有挑战性的多中心数据集上进行了广泛的实验,证明了所提出的框架的通用性和优越性.

2　相关工作

2.1　用于医学数据分析的多模态方法研究

在医学诊断中,不同模态的数据都具有独特的信息含义,且每种模态均能反映病情的不同特征.根据模态融合的层次,模态融合方式可分为输入级融合与特征级融合.输入级特征融合在多模态特征融合中属于最常见的融合方法.Li等人[23]提出MV-RAN融合多视角超声心动图,完成超声心动图序列分割与全心周期分析.Xu等人[24]提出多模态序列化学习框架,用于对模态缺失的阿尔兹海默症数据进行病情诊断.Li等人[18]通过(Coupled Neural P,CNP)系统进行多模态图像融合,利用两个CNP系统控制低频(NonsubSampled Shearlet Transform,NSST)系数融合.输入级特征融合模态融合难度较大,同时模型很难在两种以上的模态之间实现转移学习.而特征级融合可以更好地保留模态特定的信息,同时也可以更好地处理不同态之间的异质性差异.Chen等人[19]使用跨模态融合算法(Cross-Modal-Based Fusion,CMBF)捕获单一模态内以及多模态之间的特征信息,让模型学习到不同模态之间的交叉信息.Kaur等人[25]提出多目标差分进化融合技术,实现更优的模态特征融合效果.Zhang等人[26]通过文本特征对图像特征提取器进行引导,有效提取视觉特征,但其未考虑多模态视觉特征之间的有效融合.Zhou等人[27]提出密集连接高分辨率网络(Densely-connected High-Resolution Network,DHRNet)与Transformer相结合的无监督多模态医学图像融合方法,实现全局特征与局部特征之间的融合.

2.2　医学数据多中心分布差异研究

医学模型训练中经常面临着样本数量不足以及样本缺失的问题.为了缓解这些问题对模型训练的影响,通常会从多个中心收集医学数据.由于不同中心之间的数据分布差异[28-31],现有的一些研究通常通过多中心数据提高模型的泛化能力.Rajagopal等人[32]将联邦学习框架用于前列腺癌检测算法的跨中心异构数据训练中,使得模型的泛化能力有明显的提高.Vesal等人[33]提出有监督领域自适应技术和知识蒸馏损失解决多中心迁移学习和微调方法的局限性,在多中心前列腺数据集上拥有良好的鲁棒性.Starmans等人[34]提出多中心放射组学模型,在多中心前列腺癌检测上精确度高于单中心模型.Karani等人[35]在大脑、心脏和前列腺三个解剖结构的多中心数据集中提出Test-time adaptable改进措施能够提高网络在多中心数据集上的鲁棒性.Wang等人[36]提出周期一致性的跨域医学图像分割模型(Cycle-consistent Cross-domain Medical Image Segmentation,CyCMIS),将域之间的关系描述成多对多

映射而非一对一映射. Li 等人[37]在多中心(Late Gadolinium Enhanced Magnetic Resonance Imaging, LGE MRI)图像中,使用直方图匹配方式提高模型领域泛化能力.

2.3 多中心多模态结合

单一的研究中心和研究方法可能存在样本偏差、研究结果不可复制等问题,为提高研究结果可靠性,现有一些研究尝试将多模态与多中心的方法进行结合. Wu 等人[38]提出多模态多中心结合模型,利用多中心多模态 COVID-19 患者数据进行模型训练. Liu 等人[39]提出无监督多域自适应和空间神经注意结构的对称全卷积网络(Symmetric Full Convolutional Neural Network with the Unsupervised Multi-Domain Adaptation and a Spatial Neural Attention structure UMDA-SNA-SFCNN),实现跨模态跨域的图像分割. Tomar 等人[40]在跨模态跨中心数据中提出图像到图像的转换方法,实现了跨模态跨中心的图像分割. 尽管多中心多模态研究在医学领域中具有重要的实践意义和科学价值,但是目前这类研究的数量相对较少,可能是由于多中心多模态研究需要涉及多个独立的研究中心和多种不同的研究方法,需要更高的研究成本和更复杂的算法框架.

3 方法

本文提出的 TMM-Net 如图 2 所示,该框架包括三个阶段:特征级别的引导阶段、原型级别的引导阶段和交叉模态注意力特征融合阶段. 整个框架采用端到端的训练方式,不同模态原型级别引导阶段接受各自特征级别引导阶段输出特征作为输入,交叉模态注意力特征融合阶段接受两个模态的原型级别引导阶段输出特征并进行最终的融合诊断. 图像上方展示了三个阶段的训练顺序以及方式,下方展示了三个阶段的详细实现方式. 该体系结构的详细设计在第 3.1~3.3 节中展现.

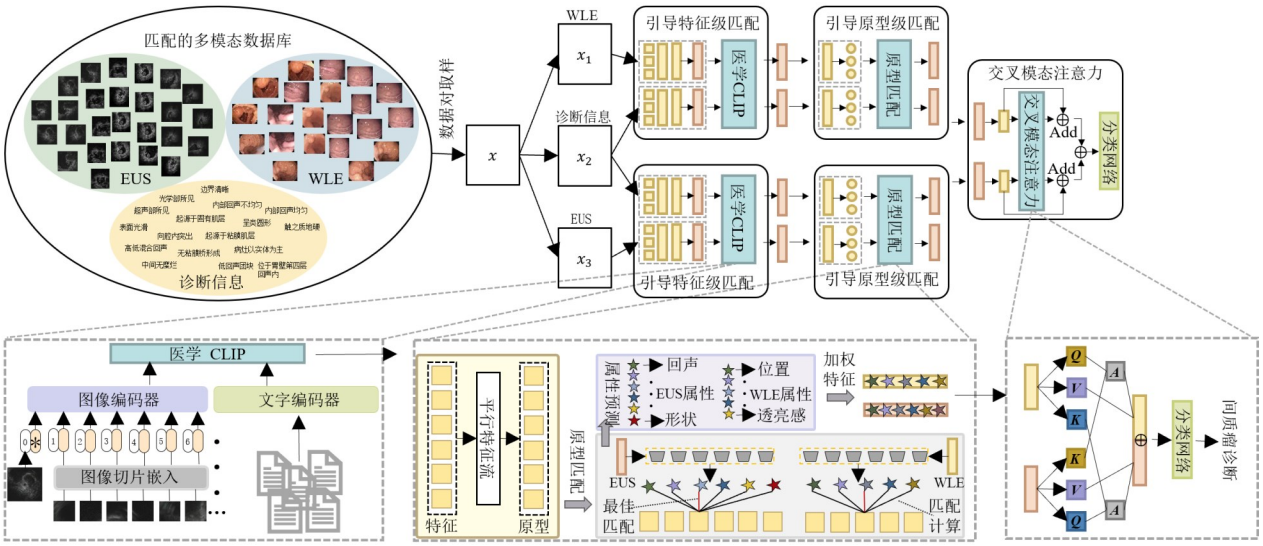


图 2 基于文本引导下的多模态医学图像分析算法框架.

3.1 阶段一:诊断文本引导模态对齐

使用诊断文本来指导模型学习图像特征是一项具有挑战性的任务,其中特征之间的模态对齐是关键问题之一. 第一阶段的主要目标是学习图像和诊断文本之间的对应关系,即将它们表示为相同的向量空间并对齐,从而实现跨模态的信息匹配和转换. Zhang 等人[41]和 Radford 等人[42]已经介绍了从与图像配对的文本中学习视觉表征的监督方法的潜力. 与他们使用广泛的自然语言不同,本文提出的方法使用特定领域的医学诊断信息,并首次研究了使用自然语言来监督引导多模态、多中心医学图像进行特征对齐的可能性. 接下来,我们将详细介绍使用特征匹配来进行引导图像特征学习的具体方法.

第一阶段中最基本的匹配网络,类似于(Contrastive Language-Image Pre-training, CLIP)[42]. 给定一批多模态数据对 N,分为 N_1:(EUS 图像 $\boldsymbol{I}_{\mathrm{EUS}}$,EUS 诊断文本 $\boldsymbol{T}_{\mathrm{EUS}}$)和 N_2:(WLE 图像 $\boldsymbol{I}_{\mathrm{WLE}}$,WLE 诊断文本 $\boldsymbol{T}_{\mathrm{WLE}}$),TMM-Net 的匹配框架被训练来分别预测一批数据对中的 $N_1(\boldsymbol{I}_{\mathrm{EUS}},\boldsymbol{T}_{\mathrm{EUS}})$内部配对和 $N_2(\boldsymbol{I}_{\mathrm{WLE}},\boldsymbol{T}_{\mathrm{WLE}})$内部配对. TMM-Net 通过联合训练两对图像编码器和文本编码器来学习两个多模态的嵌入空间,并通过文本特征的领域一致性来实现两个空间的联合. $\boldsymbol{I}_{\mathrm{EUS}}$ 和 $\boldsymbol{I}_{\mathrm{WLE}}$ 图像首先由各自的特征提取网络 $\mathrm{FI}_{\mathrm{EUS}}(\cdot)$ 和 $\mathrm{FI}_{\mathrm{WLE}}(\cdot)$ 进行处理,图像特征提取器采用 ViT(Vision Transformer)[43],已经在 ImageNet 上进行预训练. $\boldsymbol{T}_{\mathrm{EUS}}$ 和 $\boldsymbol{T}_{\mathrm{WLE}}$ 通过一个编码器 $\mathrm{FT}(\cdot)$ 映射文本特征到多模态嵌入空间,文本编码器由

Transformer[44]进行建模. 诊断信息都使用SELs语料库进行编码,并进行等长处理,最后使用空间注意力文本mask以规避错误的注意力计算. 算法1描述了具体流程. 其中,$L_{\text{EUS-Match}}$和$L_{\text{WLE-Match}}$损失中的计算单元cross_entropy_loss定义为

$$L_{\text{EUS-Match}}=-\frac{1}{n}\sum_{i=1}^{N}\left[\begin{array}{c} l\log\left(p\left(m|\boldsymbol{M}_{\text{EUS}}\right)\right) \\ +(1-l)\log\left(1-p\left(m|\boldsymbol{M}_{\text{EUS}}\right)\right)\end{array}\right] \tag{1}$$

$$L_{\text{WLE-Match}}=-\frac{1}{n}\sum_{i=1}^{N}\left[\begin{array}{c} l\log\left(p\left(m|\boldsymbol{M}_{\text{WLE}}\right)\right) \\ +(1-l)\log\left(1-p\left(m|\boldsymbol{M}_{\text{WLE}}\right)\right)\end{array}\right] \tag{2}$$

其中,$\boldsymbol{M}$代表两个模态数据各自的匹配矩阵;l是真实标签,代表出匹配矩阵中一对匹配数据;$p(m|\boldsymbol{M})$代表对应的多模态特征是否匹配的概率得分;$L_{\text{EUS-Match}}$和$L_{\text{WLE-Match}}$被用来促进EUS和WLE图像特征与SELs诊断文本特征进行匹配,从而达到特征的模态对齐和分布对齐.

算法1　特征级别文本引导下图像特征提取算法

输入: 多模态数据对$(\boldsymbol{I}_{\text{EUS}},\boldsymbol{T}_{\text{EUS}}),(\boldsymbol{I}_{\text{WLE}},\boldsymbol{T}_{\text{WLE}})$

输出: 匹配完成后特征$\text{FI}_{\text{EUS}}(\boldsymbol{I}_{\text{EUS}}),\text{FI}_{\text{WLE}}(\boldsymbol{I}_{\text{WLE}})$

1.根据SELs语料库对$\boldsymbol{T}_{\text{EUS}},\boldsymbol{T}_{\text{WLE}}$进行编码处理,获取编码后序列$\boldsymbol{T}_{\text{EUS-coding}},\boldsymbol{T}_{\text{WLE-coding}}$

2.将$\boldsymbol{T}_{\text{EUS-coding}},\boldsymbol{T}_{\text{WLE-coding}}$送入FT,提取特征

3.将原始$\boldsymbol{I}_{\text{EUS}}$和$\boldsymbol{I}_{\text{WLE}}$送入$\text{FI}_{\text{EUS}}$和$\text{FI}_{\text{WLE}}$提取特征

4.将$\text{FI}_{\text{EUS}}(\boldsymbol{I}_{\text{EUS}})$与$\text{FT}(\boldsymbol{T}_{\text{EUS}})$特征进行匹配,使用损失$L_{\text{EUS-Match}}$,匹配伪代码如下:

$$\text{FT}\left(\boldsymbol{T}_{\text{EUS-coding}}\right)=l_2\_\text{normalize}\left(\text{FT}\left(\boldsymbol{T}_{\text{EUS-coding}}\right)\right)$$

$$\text{FI}_{\text{EUS}}\left(\boldsymbol{I}_{\text{EUS}}\right)=l_2\_\text{normalize}\left(\text{FI}_{\text{EUS}}\left(\boldsymbol{I}_{\text{EUS}}\right)\right)$$

$$\boldsymbol{M}_{\text{EUS}}=\text{np.dot}\left(\text{FI}_{\text{EUS}}\left(\boldsymbol{I}_{\text{EUS}}\right),\text{FT}\left(\boldsymbol{T}_{\text{EUS-coding}}\right).T\right)$$

$$\text{loss}_{\text{EUS-match-0}}=\text{cross\_entropy\_loss}\left(\boldsymbol{M}_{\text{EUS}},\text{np.arange}\left(\text{batch}_{\text{size}}\right),\text{axis}=0\right)$$

$$\text{loss}_{\text{EUS-match-1}}=\text{cross\_entropy\_loss}\left(\boldsymbol{M}_{\text{EUS}},\text{np.arange}\left(\text{batch}_{\text{size}}\right),\text{axis}=1\right)$$

$$L_{\text{EUS-Match}}=(\text{loss}_{\text{EUS-match-0}}+\text{loss}_{\text{EUS-match-1}})/2$$

5.将$\text{FI}_{\text{WLE}}(\boldsymbol{I}_{\text{WLE}})$与$\text{FT}_{\text{WLE}}(\boldsymbol{T}_{\text{WLE}})$特征进行匹配,使用损失$L_{\text{WLE-Match}}$,匹配伪代码与上述伪代码对称,其中,

$L_{\text{WLE-Match}}=(\text{loss}_{\text{WLE-match-0}}+\text{loss}_{\text{WLE-match-1}})/2$

6.联合$L_{\text{EUS-Match}},L_{\text{WLE-Match}}$损失进行更新

7.得到匹配后特征$\text{FI}_{\text{EUS}}(\boldsymbol{I}_{\text{EUS}}),\text{FI}_{\text{WLE}}(\boldsymbol{I}_{\text{WLE}})$

3.2　阶段二:诊断文本引导特征原型匹配

EUS与WLE图像中通常包含大量信息,其中包含与病变诊断和病理学相关的诊断关键信息,也包含噪声与伪影. 提取图像中的关键诊断特征不仅可以减小噪声等对模型性能的影响,同时还能提升模型对GISTs的诊断能力. 但是关键诊断特征的提取仍然面临以下难点:(1)医学图像的采集方式、设备和处理过程可能存在差异,同时图像本身也比较复杂,难以保证数据的质量和可比性;(2)现有流行方法获得的原型具有抽象性,可能导致可解释性差,缺乏区分性;(3)关键诊断特征需要具有较好的转移性,即面对同病种的不同图像都能够提取出相同属性的关键特征. 为解决以上问题,本研究提出了一种方法,即通过引导图像特征原型与诊断文本特征原型之间的匹配来强化网络对于图像中诊断关键信息的提取,然后通过约束图像特征原型所表达的属性语义信息来对其进行解耦,并加强可解释性. 下面将详细介绍第二阶段的方法.

3.2.1　原型特征提取及匹配

不同于现有原型学习方法在整个图像空间中维护一个原型特征,我们以病例为单位构建文本原型特征来引导同病例下图像特征的学习,因为一个病例下的图像共享一个诊断文本,这样不仅可以实现针对性的引导,还可以增加原型特征之间的差异性,从而增强模型的泛化能力,避免转移性的下降. 图像与文本的原型特征提取网络均由几个平行的流组成,对应于每个待预测的病变属性,对EUS图像来说包含起源层次、形态、回声、回声异质性、生长方式、边界六个属性特征$(p_{\text{E1}},\cdots,p_{\text{E6}})$,相应的诊断文本特征原型为$(p_{\text{TE1}},\cdots,p_{\text{TE6}})$;对WLE图像来说包含肿瘤的部位、形态、表面粘膜、色泽、透亮感五个特征$(p_{\text{W1}},\cdots,p_{\text{W5}})$,相应的诊断文本特征原型为$(p_{\text{TW1}},\cdots,p_{\text{TW5}})$. 每个流由两层完全连接层(FC)来学习原型特定特征. 第二部分为通过诊断文本原型特征与图像特征之间的匹配来引导学习特定的属性信息,其中原型匹配损失定义如下:

$$L_{\text{proto-Match}}=-\frac{1}{n}\sum_{i=1}^{N}\left[\begin{array}{c} l\log\left(p\left(m|\boldsymbol{M}_{\text{p}}\right)\right) \\ +(1-l)\log\left(p\left(m|\boldsymbol{M}_{\text{p}}\right)\right)\end{array}\right] \tag{3}$$

其中,$\boldsymbol{M}_{\text{p}}$代表图像-文本对的原型匹配矩阵;EUS的匹配矩阵大小为6×6,由$(p_{\text{E1}},\cdots,p_{\text{E6}})\times(p_{\text{TE1}},\cdots,p_{\text{TE6}})^{\text{T}}$构成,相应的WLE的匹配矩阵大小为$5\times5$,由$(p_{\text{W1}},\cdots,p_{\text{W5}})\times(p_{\text{TW1}},\cdots,p_{\text{TW5}})^{\text{T}}$构成;$l$是真实标签,代表出匹配矩阵中一对匹配数据;$p(m|\boldsymbol{M}_{\text{p}})$代表对应的多模态特征是否匹配的概率得分. 根据诊断信息的引导,模型在一定程度上已具备提取图像中关键诊断特征的能力.

3.2.2　基于原型特征的属性预测

人类可以将他们的注意力转移到与特定任务最相关的图像区域. 例如,对于相同的EUS图像,医生通过观察病变边缘以及周围的组织层来判断其来源层次,同时更关注病变的内部区域以检测回声异质性. 受此启发,我们提出使用获取的图像原型特征来预测属性,然后将具有语义信息以及位置信息的原型特征分别加权集中到原始图像特征上,指导模型重点考虑诊断关

键特征以进行最终的预测. 通过这种方式,TMM-Net明确地模仿了诊断过程中的临床工作流程. 在这过程中,除了病变属性标注,我们没有使用任何的额外的标注信息,这加强了模型的可解释性.

图像的每个原型流对应一个诊断属性,对于每个属性分类任务,我们使用一个多分类问题来描述. $\boldsymbol{I}_{\mathrm{EUS}}$ 中第 i 张图像的属性真实标注为 $\mathrm{Attr}_{\mathrm{E}}=[y_{i1},y_{i2},y_{i3},y_{i4},y_{i5},y_{ic}]$,其中 $c=6$,对于 $\boldsymbol{I}_{\mathrm{WLE}}$,$c=5$,每个属性的类别数并不相同. 添加一个sigmoid层来规范化预测的输出 $P_{\mathrm{attr}}=(c|\boldsymbol{I})$:

$$\tilde{P}_{\mathrm{attr}}(c|\boldsymbol{I})=\frac{1}{1+\exp(-P_{\mathrm{attr}}(c|\boldsymbol{I}))} \tag{4}$$

其中,$\boldsymbol{I}$ 是输入图像,包括EUS和WLE;$\tilde{P}_{\mathrm{attr}}=(c|\boldsymbol{I})$ 表示 $\boldsymbol{I}$ 属于第 c 类的概率得分;对EUS图像,$c\in\{1,2,\cdots,6\}$,对WLE图像,$c\in\{1,2,\cdots,5\}$. 我们使用交叉熵损失来来优化属性预测的分支的参数:

$$L_{\mathrm{attr}}=-\frac{1}{N}\sum_{i=1}^{N}\Big(\sum_{c=1}^{C}y_{ic}\log(\tilde{P}_{\mathrm{attr}}(ic|\boldsymbol{I}))\Big) \tag{5}$$

其中,y_{ic} 是第 c 类的真实标签.

3.3 阶段三:交叉模态注意力机制下的特征融合

EUS和WLE图像在反映病变特征方面存在明显差异. EUS图像通常反映深层组织结构的变化,如病变的位置、大小、深度和侵犯周围组织的程度等. 而WLE图像主要反映表层病变的特征,如黏膜炎症、溃疡等. 因此,将EUS图像和WLE图像特征融合共同用于病变诊断可以充分利用它们各自的优势,提高病变诊断的准确性和可靠性. 但在某些情况下它们反应出的病变特征也会有重叠的地方,如肿瘤大小等. 为了避免信息冗余和相互干扰对特征融合的影响,我们引入了交叉模态注意力融合机制,结构如图3所示.

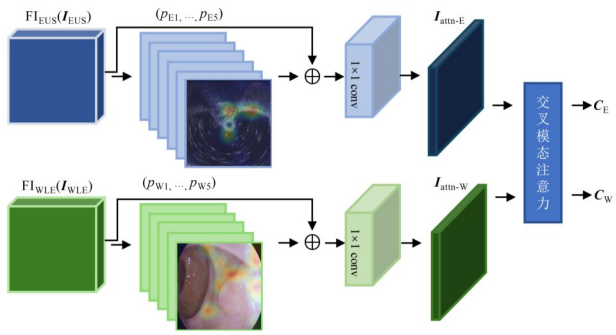


图3 交叉模态注意力融合机制

多个原型特征通过深度拼接的方式加权原始图像特征,加权后的特征再送入交叉模态注意力网络中,交叉模态注意力网络有两个分支,用于计算EUS与WLE图像特征之间的交互表示,突出与任务相关的关键诊断特征,再融合后进行最后的预测. 这样模型可以更强调相关特征,同时避免冗余或无关信息的干扰. 形式上,EUS或者WLE加权后的特征图计算为

$$\boldsymbol{I}_{\mathrm{attn}}=\mathrm{Conv}_{1\times1}\Big(\mathrm{concat}(\mathrm{FI}(\boldsymbol{I}),P_1,\cdots,P_c)\in\mathbb{R}^{\left(C_I+C_{P_1}+\cdots+C_{P_c}\right)\times H\times W}\Big) \tag{6}$$

其中,concat表示拼接操作;$\mathrm{Conv}_{1\times1}$ 表示 1×1 的卷积操作;$\boldsymbol{I}_{\mathrm{attn}}$ 表示加权后的特征,它的通道数为 $C_{\mathrm{I}}+C_{P_1}+\cdots+C_{P_c}$. 具体来说,将图像特征 $\mathrm{FI}(\boldsymbol{I})$ 和图像原型特征 P 沿着通道维拼接起来,进行卷积后得到加权后的图像特征 $\boldsymbol{I}_{\mathrm{attn}}$. 然后进行交叉模态注意力的融合学习.

将EUS以及WLE的加权后特征分别定义为 $\boldsymbol{I}_{\mathrm{attn\text{-}E}}\in\mathbb{R}^{h_1\times w_1\times c_1}$,$\boldsymbol{I}_{\mathrm{attn\text{-}W}}\in\mathbb{R}^{h_2\times w_2\times c_2}$,其中 h,w 分别表示图像的高度和宽度,c 表示图像中每个像素点的向量维度. 对 $\boldsymbol{I}_{\mathrm{attn\text{-}E}}$ 以及 $\boldsymbol{I}_{\mathrm{attn\text{-}W}}$ 分别进行线性变换,其中,$\boldsymbol{I}_{\mathrm{attn\text{-}E}}$ 变换后得到三个矩阵为

$$\boldsymbol{Q}_1\in\mathbb{R}^{h_1\times w_1\times d_k},\boldsymbol{K}_1\in\mathbb{R}^{h_1\times w_1\times d_k},\boldsymbol{V}_1\in\mathbb{R}^{h_1\times w_1\times d_v}$$

$\boldsymbol{I}_{\mathrm{attn\text{-}W}}$ 变换后得到三个矩阵为

$$\boldsymbol{Q}_2\in\mathbb{R}^{h_2\times w_2\times d_k},\boldsymbol{K}_2\in\mathbb{R}^{h_2\times w_2\times d_k},\boldsymbol{V}_2\in\mathbb{R}^{h_2\times w_2\times d_v}$$

其中,d_k 和 d_v 分别表示查询、键和值的向量维度. 这些矩阵是通过将 $\boldsymbol{I}_{\mathrm{attn\text{-}E}}$ 以及 $\boldsymbol{I}_{\mathrm{attn\text{-}W}}$ 分别与两组权重矩阵相乘得到的. 在 $\boldsymbol{I}_{\mathrm{attn\text{-}E}}$ 上计算的交叉模态注意力特征表示为 $\boldsymbol{C}_{\mathrm{E}}\in\mathbb{R}^{h_1\times w_1\times d_v}$,即将 $\boldsymbol{I}_{\mathrm{attn\text{-}W}}$ 中的信息融合到 $\boldsymbol{I}_{\mathrm{attn\text{-}E}}$ 中. 具体来说,对于 $\boldsymbol{I}_{\mathrm{attn\text{-}W}}$ 中的第 i 个像素点,将其对应的查询向量 $\boldsymbol{q}_i^1$ 与 $\boldsymbol{I}_{\mathrm{attn\text{-}W}}$ 中的所有键向量 $\boldsymbol{K}_2$ 进行计算得到注意力权重,然后再与所有的值向量 $\boldsymbol{V}_2$ 进行加权求和,得到新的表示 c_i^1,即 $c_i^1=\sum_{j=1}^{h_2w_2}\alpha_{ij}^{12}V_{2j}$,其中,$\alpha_{ij}^{12}$ 表示 $\boldsymbol{I}_{\mathrm{attn\text{-}E}}$ 中的第 i 个像素点对 $\boldsymbol{I}_{\mathrm{attn\text{-}W}}$ 中的第 j 个像素点的注意力权重. 最后,将所有的 c_i^1 重构成特征图的形式,得到,$\boldsymbol{C}_{\mathrm{E}}\in\mathbb{R}^{h_1\times w_1\times d_v}$. 同理,对于WLE,其交叉模态注意力特征表示为

$$\boldsymbol{C}_{\mathrm{W}}\in\mathbb{R}^{h_2\times w_2\times d_v}$$

最后得到互相关后的融合特征为 $\boldsymbol{F}_{\mathrm{fusion}}$,使用融合后的特征进行分类,分类损失定义如下:

$$L_{\mathrm{class}}=-\frac{1}{n}\sum_{i=1}^{N}\begin{bmatrix}y_i\log\left(p\left(y_i|\boldsymbol{F}_{\mathrm{fusion}}\right)\right)\\+\left(1-y_i\right)\log\left(1-p\left(y_i|\boldsymbol{F}_{\mathrm{fusion}}\right)\right)\end{bmatrix} \tag{7}$$

其中,y_i 是标签值,$p(y_i|\boldsymbol{F}_{\mathrm{fusion}})$ 是融合特征 $\boldsymbol{F}_{\mathrm{fusion}}$ 属 y_i 标签值的概率.

我们的目标是构建一个多模态模型,该模型由图像和语言两个路径组成,分别提取视觉和文本表征. 这两个路径相互作用,可以提高它们各自的功能. 同时,我们引入一种匹配机制来控制视觉特征和文本特征之间的耦合程度,以避免对文本表征的过度依赖. 这也更加符合实际应用情况,因为并非总是能够获取到诊断

文本. 通过匹配机制,我们有望从图像中提取出与诊断语义相关的视觉特征,从而提高图像理解的能力,并加强特征融合效果.

4　数据和实验设置

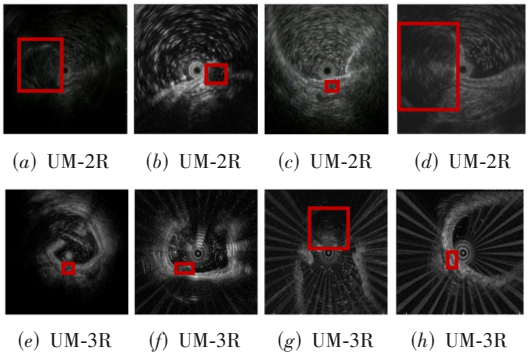
4.1　实验数据

胃肠道内镜超声相关数据从两个医院获得. 数据来自两个不同中心,都是通过专业内镜医师在对不同患者进行内镜检查时进行采集,所有的影像都是真实临床环境中采集的. 其中中心一包括228例患者数据,中心二包含58例患者数据. 为了充分利用已有数据中的图像,我们将每个病例中的EUS和WLE进行配对,配对后中心一包含3 903个多模态数据对,中心二包含6 122个多模态数据对. 采集到的图像中只包含单个肿瘤,分类标签是从相应的病理报告中获得的. 在单中心实验中,图像被随机打乱并分成训练集(70%)和测试集(30%),多中心实验中以中心一数据作为源领域,中心二数据作为目标领域. 这一胃肠道内镜超声数据集包含了从多个内镜超声机器中所获得的内镜超声图像,中心一数据是从12 MHz的奥林巴斯UM-2R获得的,中心二数据是从20 MHz的奥林巴斯UM-3R中获得的. 不同设备生成的内镜超声图像其对比度、分辨率、噪声和扫描窗口的大小等均存在不同,图4描述了一些典型的例子,图4(a)~(d)是从中心一获得,图4(e)~(h)是从中心二获得. 这些图像具有不同的对比度、强度分布、噪声. 与此同时,病变的大小和位置变化很大(用红色的方框表示). 需要指出的是,收集该数据集是为了接近真实世界的临床环境,而大多数现有的医疗数据集都是在一个中心/医院中使用某些类型的机器收集的. 如果一个算法能够在这样一个具有挑战性的数据集上表现良好,那么它很可能在实际应用中也能表现出类似的性能.

诊断信息根据专业医师对观察到的内镜超声图像描述中获得. SELs语料库构建方式为:从所有病例的诊断信息中搜集光学所见和超声所见的信息,并将这些信息按每一个字进行分割,将分割后的所有字按照出现的次数进行排序,并存入语料库中,语料库为一个字典,每一个字为字典中的键其值为每个字按照其出现次数排序之后得到的位次. 在自然语言处理领域中,文本通常被表示为高维向量的形式,其中每个维度代表文本中的一个特征或属性,这些特征的组合和权重决定了文本的含义和表达方式. 因此,SELs描述文本可以被视为一个在高维空间中聚合的特征向量,代表了消化道粘膜下肿瘤的特征和相关信息.

4.2　实现细节

EUS的WLE图像存储格式均为BMP格式,按照RGB格式读取图像. 为了保持图像原始的宽高比,将图像按照



(a) UM-2R　(b) UM-2R　(c) UM-2R　(d) UM-2R

(e) UM-3R　(f) UM-3R　(g) UM-3R　(h) UM-3R

图4　不同中心设备收集的图像的例子

比例缩小到短边为224像素,并使用224×224大小的正方形进行中心裁剪,并将图像中每个像素点归一化. 使用经过ImageNet预训练的Vision Transformer(ViT)以及处理语言的Transformer作为基线,ViT预训练后出色的图像处理性能对于小数据量下的医学图像任务可能会有所帮助. 在训练中,我们通过EUS和WLE图像之间1对1配比来扩充了多模态数据对的数量. 该网络被训练100个epoch,batch大小为32. 初始学习率为0.001,moment为0.9,在10个epoch后学习率减小10倍,该网络使用随机梯度下降(Stochastic Gradient Descent,SGD)进行优化. 整个实验都在Pytorch框架实现,使用Geforce GTX 4090 GPU进行模型训练和评估.

4.3　实验设置

对比算法:为了证明所提出的框架的有效性,我们使用我们的数据集实现了一些最先进的方法来进行公平的比较. 首先,我们选择了一些基于深度学习的SELs肿瘤诊断的最新的相关工作[45~50]. 其中包括单模态最新研究方法[46,47,51,52]以及多模态最新研究方法[45,48],为了保证实验结果的公正性和可比性,我们采用了论文中所报道的性能评估指标. 值得注意的是,提出的方法使用诊断文字信息来引导图像关键特征的学习,从而网络可自动关注到图像中感兴趣的位置(Region Of Interest,ROI),而有些方法[48,52]需要使用人工标注mask、剪裁或其他模型对ROI进行定位或预处理的额外阶段. 其次,为了研究一个基础的CNN模型针对胃镜超声图像数据的特征提取能力,我们实现了一些经典的分类方法,如ResNet50[49]、VGG-16[50]和DenseNet201[53],通过在这三个经典网络上进行多模态融合实验,可比较提出的方法的一些组件在特征提取与融合方面的优势,结果在表1中展示.

因为TMM-Net中图像和诊断文本的路径具有可解耦性,我们还进一步将提出的诊断文本引导路线融合到这三个网络中,以展示其对多模态基线网络的性能提升能力. 以ResNet50为例,具体做法为在使用ResNet50提取EUS与WLE图像特征过程中,按照

TMM-Net的设计,分为三个阶段.第一阶段使用特征级别的诊断文本引导,第二阶段进行原型级别的诊断文本引导,最后直接融合多模态特征进行最终预测.同时,我们还比较了基线网络、提升后的基线网络以及TMM-Net在第二阶段对于属性预测的能力,结果在表2中展示.

表1 对比算法的性能比较

方法	SEN	SPE	Acc	AUC	Interp	data modal	R-s	M-m	M-c
ResNet[49]	0.408	0.862	0.723	0.539	No	EUS	No	No	Yes
	0.376	0.822	0.715	0.486	No	WLE	No	No	Yes
	0.387	0.813	0.683	0.556	No	EUS+WLE	No	Yes	Yes
VGG16[50]	0.530	0.707	0.667	0.448	No	EUS	No	No	Yes
	0.387	0.790	0.700	0.470	No	WLE	No	No	Yes
	0.413	0.785	0.679	0.527	No	EUS+WLE	No	Yes	Yes
DenseNet201[53]	0.539	0.719	0.669	0.420	No	EUS	No	No	Yes
	0.264	0.797	0.698	0.584	No	WLE	No	No	Yes
	0.334	0.751	0.677	0.580	No	EUS+WLE	No	Yes	Yes
Voice-Assisted[45]	0.74	—	0.76	—	No	EUS+Voice	No	Yes	Yes
MMP-AI[48]	0.836	0.833	0.835	0.896	No	EUS+WLE	Yes	Yes	Yes
TMM-Net	**0.846**	**0.824**	**0.896**	**0.904**	**Yes**	**EUS+WLE+DI-Text**	**No**	**Yes**	**Yes**
EUS-CNN[51]	0.956	0.821	0.912	0.923	No	EUS	No	No	No
CNN-CAD[52]	0.830	0.755	0.792	—	No	EUS	Yes	No	No
CNN-based[46]	0.920	0.643	0.869	—	No	EUS	No	No	No
DIA-based[47]	0.865	0.759	0.835	—	No	EUS	No	No	No
TMM-Net	**0.943**	**0.994**	**0.989**	**0.927**	**Yes**	**EUS+WLE+DI-Text**	**No**	**Yes**	**No**

表2 提出的诊断文本引导框架对CNN网络性能的提升比较

方法	Data modal	G-Acc	P-Acc
ResNet50[49]	EUS+WLE	0.683	—
ResNet50+文本引导1	EUS+WLE +DI-Text	0.724	—
ResNet50+文本引导1,2	EUS+WLE +DI-Text	0.748	0.774
VGG16[50]	EUS+WLE	0.679	—
VGG16+文本引导1	EUS+WLE +DI-Text	0.719	—
VGG16+文本引导1,2	EUS+WLE +DI-Text	0.739	0.763
DenseNet201[53]	EUS+WLE	0.677	—
DenseNet201+文本引导1	EUS+WLE +DI-Text	0.722	—
DenseNet201+文本引导1,2	EUS+WLE +DI-Text	0.752	0.760
TMM-Net	**EUS+WLE +DI-Text**	**0.841**	**0.838**

消融实验:为了评估提出的TMM-Net框架中不同组件的有效性,我们进行了以下消融实验:(1)特征提取器ViT+单中心EUS图像(baseline1);(2)特征提取器ViT+单中心WLE图像(baseline2);(3)特征提取器+单中心EUS图像+EUS诊断信息特征级别引导(baseline1+EUS阶段1);(4)特征提取器ViT+单中心WLE图像+WLE诊断信息特征级别引导(baseline2+WLE阶段1);(5)特征提取器+单中心EUS图像+单中心WLE图像+直接特征融合(baseline1+baseline2+直接多模态);(6)特征提取器+单中心EUS图像+单中心WLE图像+特征级别引导(baseline1+baseline2+阶段1+直接多模态);(7)特征提取器+单中心EUS图像+单中心WLE图像+特征级别引导+原型级别引导(baseline1+baseline2+阶段1+阶段2+直接多模态);(8)特征提取器+单中心EUS图像+单中心WLE图像+特征级别引导+原型级别引导+交叉模态注意力(提出的TMM-Net).我们还在每个消融实验中都加入了多中心数据的直接验证精度,目的是为了验证诊断文本的多中心一致性特点对多中心数据集验证精度的提升作用.结果在表3中展示.

交叉验证:为了进一步估计所提出的模型是否可以推广到整个数据集,我们进行了一个五倍交叉验证实验.数据集被随机划分为五个子集.每个交叉验证模型都使用实验细节中描述的相同的超参数和数据增强方法进行训练.

表3　消融实验研究结果

方法	SEN	SPE	Acc	AUC	Data modal	Multi modal	Multi center
baseline1	0.864	0.987	0.936	0.817	EUS	No	No
baseline1	0.842	0.753	0.801	0.786	EUS	No	Yes
baseline2	0.901	0.966	0.917	0.801	WLE	No	No
baseline2	0.804	0.819	0816	0.745	WLE	No	Yes
baseline1+EUS阶段1	0.968	0.996	0.977	0.863	EUS+DI-Text	Yes	No
baseline1+EUS阶段1	0.836	0.833	0.828	0.789	EUS+DI-Text	Yes	Yes
baseline2+WLE阶段1	0.958	0.992	0.964	0.833	WLE+DI-Text	Yes	No
baseline2+WLE阶段1	0.867	0.721	0.822	0.751	WLE+DI-Text	Yes	Yes
baseline1+baseline2+直接多模态	0.935	0.828	0.871	0.802	EUS+WLE	Yes	No
baseline1+baseline2+直接多模态	0.861	0.753	0.793	0.728	EUS+WLE	Yes	Yes
baseline1+baseline2+阶段1+直接多模态	0.943	0.994	0.953	0.881	EUS+WLE+DI-Text	Yes	No
baseline1+baseline2+阶段1+直接多模态	0.843	0.764	0.831	0.864	EUS+WLE+DI-Text	Yes	Yes
baseline1+baseline2+阶段1+阶段2+直接多模态	0.916	0.964	0.983	0.913	EUS+WLE+DI-Text	Yes	No
baseline1+baseline2+阶段1+阶段2+直接多模态	0.839	0.825	0.841	0.894	EUS+WLE+DI-Text	Yes	Yes
TMM-Net	**0.943**	**0.994**	**0.989**	**0.927**	**EUS+WLE+DI-Text**	**Yes**	**No**
TMM-Net	**0.846**	**0.824**	**0.896**	**0.904**	**EUS+WLE+DI-Text**	**Yes**	**Yes**

4.4　评价指标

为了定量评估所提出的MAA-Net模型,我们使用了准确性(Acc)、精确性(Prec)、召回率(Rec)和受试者工作特征曲线下的面积(AUC)作为性能指标,它们被定义为

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP}} \tag{8}$$

$$\mathrm{Prec}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{9}$$

$$\mathrm{Rec}=\frac{\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}} \tag{10}$$

$$\mathrm{AUC}=\int_0^1 t_{\mathrm{pr}}(f_{\mathrm{pr}})\mathrm{d}f_{\mathrm{pr}}=P(x_1>x_0) \tag{11}$$

其中,TP、FN、TN、FP分别为真阳性、假阴性、真阴性、假阳性病例数;t_{pr}为真阳性率;f_{pr}为假阳性率;x_0和x_1分别为阴性实例和阳性实例的置信值.

5　实验结果与分析

5.1　对比算法

将所提出的TMM-Net与其他对比方法之间的性能比较的实验结果见表1所示."Interp"列表示该方法是否对诊断流程进行了解释性的展示,"R-s"为"ROI-stage"的缩写,表示该方法是否需要手工切割ROI或使用模型进行ROI的分割,来辅助准确关注到病变区域."M-m"为"Multi-modal"的缩写,代表是否使用多模态信息."M-c"为"Multi-center"的缩写,代表是否使用多中心数据.结果表明,提出的模型在不需要额外的ROI操作干预的情况下,在单中心数据集上获得了最高的精度(Acc:0.989,AUC:0.927).而且,在模型第二阶段,我们通过原型特征提供了图像属性预测的信息,并利用该信息对图像特征进行加权,从而在第三阶段使网络更加关注诊断所需的关键特征,从而提高了模型的最终诊断性能.通过提供属性预测概率,我们可以帮助医生更好地理解模型的诊断过程,增强模型的可解释性.图5展示了不同属性的特征注意力热力图,其中热力图被重新调整到相同的大小,并叠加到原始图像上.这些结果表明,不同属性的预测过程突出了不同的局部区域,从而更好地指导网络全面关注不同的诊断关键特征.尽管面临有挑战性的多中心数据的无监督训练,我们的模型也取得了多中心数据下的最高精度(Acc:0.896,AUC:0.904).这证明了所提出的TMM-Net在诊断信息的引导下可以同时处理来自多中心的EUS和WLE图像,学习到模态互补的区域和上下文信息来处理病变的位置、大小和外观的变化,且这些信息因为通过向相同诊断空间的映射而具有领域一致性.

表1还显示出现在流行的计算机视觉模型,如ResNet50、VGG-16、Densenet201[49,50,53],他们在处理多中心、多模态数据情况时,跟TMM-Net相比取得了相对较差的性能.特别是在无监督多中心数据测试下,性能下降明显,这可能是由于这些模型没有处理多模态、多中心数据情况下具有高变化特征的病变图像的能力.同时这也说明了该数据集具有多模态及多中心数据特征差异大的特点.而我们模型中的诊断文本引导方法不仅实现了特征的模态对齐(如图6所示),同时,诊断文本的领域一致性和模态交叉融合辅助实现了多中心数据的领域对齐(无监督多中心实验精度达到最高的0.896),从而在这种情况下体现出更好的性能.

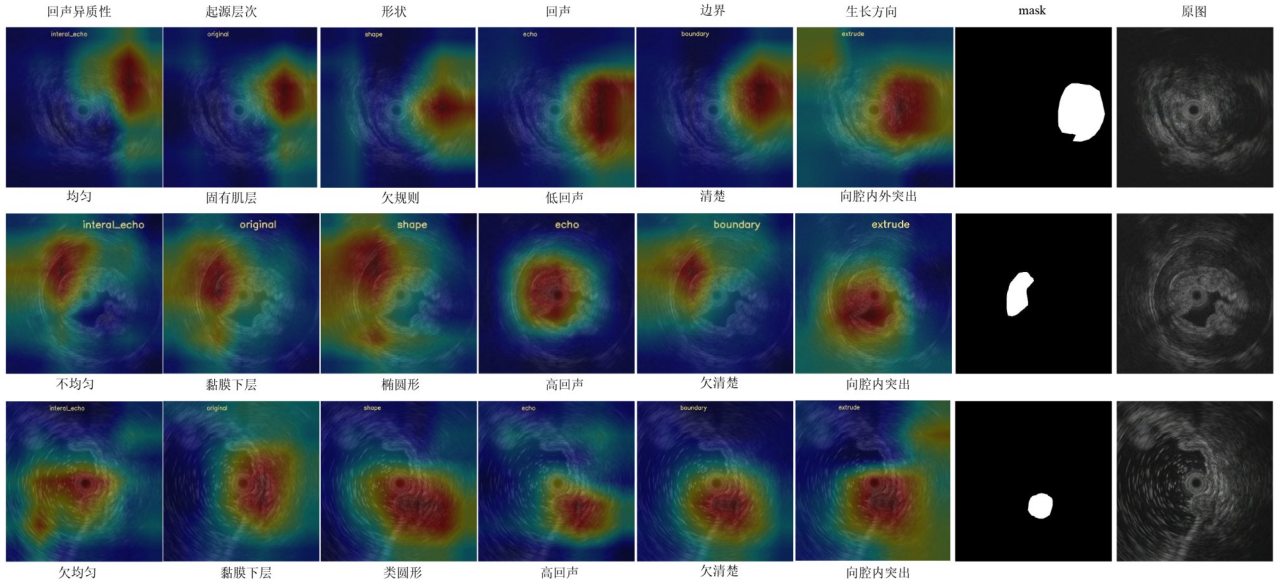


图5 EUS图像的属性预测关注热力图可视化例子

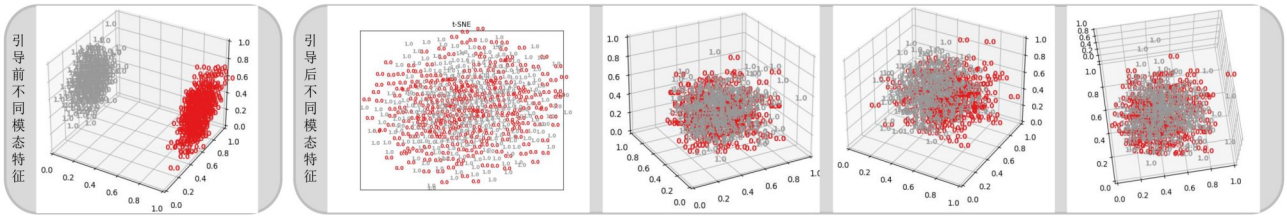


图6 EUS及WLE图像原始特征与诊断文本引导下融合后特征的降维可视化图

在表1中需要注意的是,“ROI-stage”模型在额外的ROI检测中可能会出现错误的结果,从而导致图像的错误裁剪和错误定位信息区域,分类网络不能对那些甚至可能不包含目标病变的图像进行准确的预测. CNN-CAD[51]通过手工剪裁ROI来避免了这类错误,但是可能因为病变的周围组织可能包含有用的背景信息,如来源层次等,导致了其精度的下降. 这一发现也验证了我们的模型设计,通过诊断文本给出的关键信息来引导图像对这些关键特征的学习,从而根据可判断病变性质的关键特征来自动引导对ROI的灵活关注. 如图7所示,我们提供了网络在最终预测时对整体图像的关注度可视化热力图,其中网络关注的病变区域与真实病变区域密切相关,且包含对周围组织信息的关注,符合临床医师对关注区域的选择. 同时提出的模型在第二阶段根据文本诊断信息特征原型的引导,还可针对病变的诊断属性进行预测,图5是对不同属性预测时的关注度热力图.

将提出的诊断文本引导框架组合到CNN网络中,与原始网络性能的提升比较结果在表2中显示. “文本引导1”代表加入第一阶段引导,“文本引导1,2代表加入第一以及第二阶段的引导”,“G-Acc”代表病变性质的分类精度,“P-Acc”代表属性的分类平均精度. 实验结果表明随着诊断文本引导的加入,三个CNN网络多模态跨领域识别精度都大幅上升,说明诊断文本引导提升了CNN对图像特征的提取能力,使其更关注与诊断相关的关键特征. 第一阶段的加入对于多中心精度提升更加显著,我们认识为诊断文本的领域一致性使其在引导图像特征的学习对齐中也对不同领域的图像特征进行了对齐. 同时,在表2中还显示了第二阶段中有关属性预测的平均结果与TMM-Net的比较,为了公平,TMM-Net只包含了阶段一和阶段二,然后将多模态特征直接融合. 实验结果表明,我们提出的TMM-Net在对GIST进行预测和对属性进行预测方面均取得了最佳结果. 这也从侧面证明了我们选择使用预训练后的ViT作为图像特征提取器的合理性和有效性.

5.2 消融实验

消融实验结果展示在表3中. 前四行baseline是EUS以及WLE的单中心以及多中心实验,结果表明特征提取器对单中心测试精度准确率较高,但是在无监督多中心下测试精度明显降低,这也说明了该多中心数据集数据领域之间差异大导致模型适配难度大,这也是现目前大多CAD面临的问题. 第5~8行是base-

line+特征匹配的结果,也是分别在EUS和WLE图像完成了单中心和多中心的测试,其中使用图像对应的诊断文本信息进行特征匹配,来引导图像特征的学习.结果表明加入诊断文本信息的引导后,图像分类任务的精度提升明显,在两中模态的图像上都有体现,并且还克服了一定程度的多中心领域偏移.

第9行和第10行展示了EUS和WLE特征直接融合后在单中心以及多中心任务下的结果.其多模态结果低于WLE和EUS任何单模态结果(例如:Acc:0.871<0.936),我们认为这是因为EUS与WLE的特征之间的模态差异明显,特征融合后受到互相的干扰,导致精度下降.但是,加上诊断信息引导的特征级别的多模态特征匹配后(第11行和第12行),模型的精度达到0.953,该结果直观反映出第一阶段诊断诊断文本特征级别的引导对于多模态特征对齐的有效性.同时,在第12行中反映了多中心测试的精度结果,表明诊断文本的多中心一致性对跨域任务也有一定的性能提升作用.第13行和第14行验证了阶段二诊断文本原型引导对最终预测精度的帮助,经过阶段二的优化后,多中心预测精度的提升效果似乎低于单中心预测.这可能是因为阶段二的主要目的是实现图像与诊断文本原型之间的准确匹配,以选择图像中的关键诊断特征来优化属性预测的精度.因此,网络更加关注提高该分布下的图像-文本对之间的匹配.最后两行展示了TMM-Net在单中心和多中心任务下的表现.结果表明,TMM-Net在不管在单中心或多中心任务中都表现出色,而交叉模态注意力似乎对提高多中心预测性能有较大的帮助.这是因为交叉模态注意力成功地融合了EUS和WLE特征,进一步扩大了特征空间,使其更加适用于多中心的情况.

图7显示了EUS和WLE在使用文本引导前后的网络关注,也就是涉及表2中第1、3、15行的数据.结果表明,在经过诊断文本引导的图像特征选择与融合后,网络对EUS图像和WLE图像的关注精度显著提高,并且还能够灵活关注到病变周围组织的层次结构,这更符合临床医生的视角.这也说明了诊断信息文本确实对图像关键特征的学习起到了引导作用,并且在特征一致性下的多模态特征融合也增强了网络识别病变的能力,从而实现了更高的分类精度.我们还可视化了消融实验每个层次的AUC曲线,以便更直接的观察各层次的模型效果及其变化趋势,如图8所示.

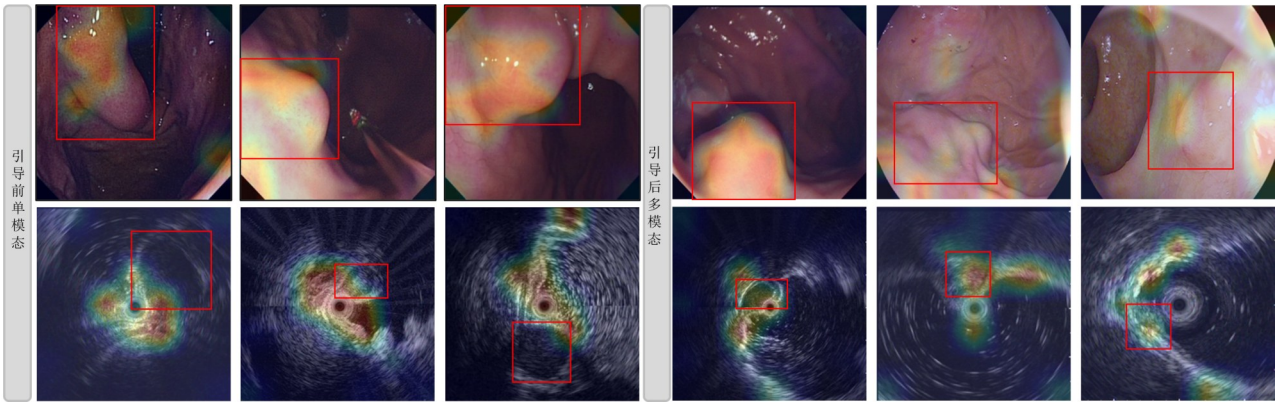


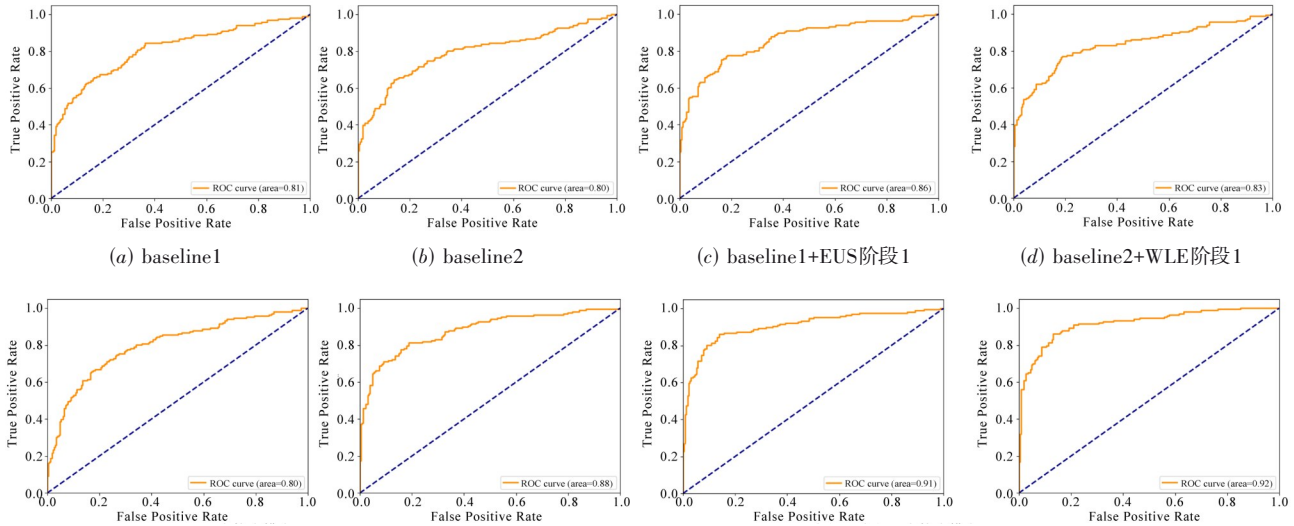
图7　特征匹配融合前后网络对EUS及WLE图的注意力热力图的可视化

5.3　交叉验证实验

为了进一步研究该模型在这个具有挑战性的数多模态据集上是否表现一致,我们在两个中心数据集上都单独进行了五倍交叉验证实验.表4显示了使用不同交叉验证分段数据训练和测试的TMM-Net模型的平均精度.结果表明,该模型对数据集的所有分区都是有效的.

5.4　可视化实验

为了更好地理解诊断的推理,同时也为了从视觉角度直观比较诊断文本引导下的特征匹配融合效果,我们提供了为识别WLE和EUS图片中的病变而生成的注意力热力图的可视化图(如图7所示),以及第二阶段中属性预测关注热力图可视化图(如图5所示).图7提供了随机6个不同病例的比较例子,红色矩形表示病变的大小和位置.其中,左边是3个示例在匹配前进行单模态训练时的网络注意力可视化图,可以明显看出关于WLE和EUS的注意力范围比较广泛,特别是EUS图像的注意力都集中在图像中间,但是对病变位置的关注却出现偏差.比如1个EUS图像,网络只在最中心进行了广泛关注,并没有找到病变的实际位置,这种情况在其他两个EUS图中也出现了.图7中右边三个示例展示了在多模态特征融合后,也就是第三阶段实现最终预测后的网络对整体图像的关注可视化图.结果表明,不管是EUS还是WLE图像,网络都实现了更精准的关注,而不是广泛的.特别是在EUS图像



(*a*) baseline1 (*b*) baseline2 (*c*) baseline1+EUS阶段1 (*d*) baseline2+WLE阶段1

(*e*) baseline1+baseline2+直接多模态 (*f*) baseline1+baseline2+阶段1+直接多模态 (*g*) baseline1+baseline2+阶段1+阶段2+直接多模态 (*h*) TMM-Net

图8 消融实验的病变性质预测的ROC曲线

表4 五倍交叉验证实验的平均结果

数据	设置	SEN	SPE	Acc	AUC
中心一	5-Fold	0.939	0.991	0.984	0.925
中心二	5-Fold	0.967	0.989	0.975	0.937

中，不仅每个病变的位置都被准确关注，而且还灵活关注到了病变周围组织，这也符合了临床诊断的关注点：医生在看EUS图像时，需要沿着病变位置两边查看组织层次来判断肿瘤来源．我们猜测，对于WLE图像中肿瘤形状、大小和大致位置的关注也帮助了对EUS的关注，因为左边引导前单模态关注图中表现出虽然网络对WLE图像中的病变关注较准确，但是在EUS图中却频频失利，正因为他们之间并没有进行多模态特征的交叉融合，这正体现了诊断文字引导的关键特征学习及模态交叉注意力机制的有效性．

为了进一步验证我们的方法，我们使用T-sne实现了原始WLE和EUS特征与最终融合后特征的降维可视化，如图8所示．结果表明，WLE图像特征与EUS图像特征之间存在明显的模态不匹配问题，在三维空间中它们的特征也呈现无交集的情况．EUS图像和WLE图像是从不同视角对病变区域进行成像，它们反映的病变信息也不同．经过SELs语料库中诊断信息进行引导匹配融合后，从多视角结果可以看出EUS和WLE特征已经实现了模态对齐．从消融实验中精度的明显变化中也可以看出这一点(表3)．

6 总结

本文提出了一种新的多阶段文本引导下的多模态图像特征提取框架，用于胃肠道间质瘤的诊断．该框架的设计是受到临床指南和诊断期间的人类行为的启发．该多阶段框架从特征级别、原型级别两个阶段，使用诊断文本特征分级引导模型对图像中关键诊断信息的学习，并使用一个属性预测分支来学习重要的临床属性，最后阶段采取模态交叉融合注意力机制来融合学习到的多模态特征完成病变性质的诊断．提出的文本引导框架还使用诊断文本的领域一致性来约束多模态、多中心图像特征的对齐．验证实验表明，该框架在胃肠道间质瘤分类方面优于其他最先进的方法，在多中心数据测试中也取得了最高精度，并提供了病变属性以及性质的注意力可视化图来解释其推理．

参考文献

[1] PARK E Y, KIM G H. Diagnosis of gastric subepithelial tumors using endoscopic ultrasonography or abdominopelvic computed tomography: Which is better? [J]. Clinical Endoscopy, 2019, 52(6): 519-520.

[2] PALLIO S, CRINÒ S F, MAIDA M, et al. Endoscopic ultrasound advanced techniques for diagnosis of gastrointestinal stromal tumours[J]. Cancers, 2023, 15(4): 1285.

[3] 中华医学会消化内镜分会NOTES、外科学组, 中国医师协会内镜医师分会消化内镜专业委员会, 中华医学会外科学分会胃肠外科学组．中国消化道黏膜下肿瘤内镜诊治专家共识(2023版)[J]. 中国实用外科杂志, 2023, 43

(3): 241-251.

[4] KHAN S, ZHANG R, FANG W L, et al. Reliability of endoscopic ultrasound using miniprobes and grayscale histogram analysis in diagnosing upper gastrointestinal subepithelial lesions[J]. Gastroenterology Research and Practice, 2020, 2020: 6591341.

[5] HORIE Y, YOSHIO T, AOYAMA K, et al. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks[J]. Gastrointestinal Endoscopy, 2019, 89(1): 25-32.

[6] HIRASAWA T, AOYAMA K, TANIMOTO T, et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images[J]. Gastric Cancer, 2018, 21(4): 653-660.

[7] BYRNE M F, CHAPADOS N, SOUDAN F, et al. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model[J]. Gut, 2019, 68(1): 94-100.

[8] KUWAHARA T, HARA K, MIZUNO N, et al. Usefulness of deep learning analysis for the diagnosis of malignancy in intraductal papillary mucinous neoplasms of the pancreas[J]. Clinical and Translational Gastroenterology, 2019, 10(5): 1-8.

[9] HIRAI K, KUWAHARA T, FURUKAWA K, et al. Artificial intelligence-based diagnosis of upper gastrointestinal subepithelial lesions on endoscopic ultrasonography images[J]. Gastric Cancer: Official Journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association, 2022, 25(2): 382-391.

[10] KIM Y H, KIM G H, KIM K B, et al. Application of a convolutional neural network in the diagnosis of gastric mesenchymal tumors on endoscopic ultrasonography images[J]. Journal of Clinical Medicine, 2020, 9(10): 3162.

[11] MINODA Y, IHARA E, KOMORI K, et al. Efficacy of endoscopic ultrasound with artificial intelligence for the diagnosis of gastrointestinal stromal tumors[J]. Journal of Gastroenterology, 2020, 55(12): 1119-1126.

[12] OH C K, KIM T, CHO Y K, et al. Convolutional neural network-based object detection model to identify gastrointestinal stromal tumors in endoscopic ultrasound images[J]. Journal of Gastroenterology and Hepatology, 2021, 36(12): 3387-3394.

[13] SEVEN G, SILAHTAROGLU G, SEVEN O O, et al. Differentiating gastrointestinal stromal tumors from leiomyomas using a neural network trained on endoscopic ultrasonography images[J]. Digestive Diseases (Basel, Switzerland), 2022, 40(4): 427-435.

[14] TANAKA H, KAMATA K, ISHIHARA R, et al. Value of artificial intelligence with novel tumor tracking technology in the diagnosis of gastric submucosal tumors by contrast-enhanced harmonic endoscopic ultrasonography[J]. Journal of Gastroenterology and Hepatology, 2022, 37(5): 841-846.

[15] YANG X T, WANG H, DONG Q, et al. An artificial intelligence system for distinguishing between gastrointestinal stromal tumors and leiomyomas using endoscopic ultrasonography[J]. Endoscopy, 2022, 54(3): 251-261.

[16] 张淑军, 彭中, 李辉. SAU-Net: 基于U-Net和自注意力机制的医学图像分割方法[J]. 电子学报, 2022, 50(10): 2433-2442.
ZHANG S J, PENG Z, LI H. SAU-Net: Medical image segmentation method based on U-Net and self-attention[J]. Acta Electronica Sinica, 2022, 50(10): 2433-2442. (in Chinese)

[17] 刘少鹏, 赵慧民, 洪佳明, 等. 面向医学图像生成的鲁棒条件生成对抗网络[J]. 电子学报, 2023, 51(2): 427-437.
LIU S P, ZHAO H M, HONG J M, et al. Medical image synthesis using robust conditional GAN[J]. Acta Electronica Sinica, 2023, 51(2): 427-437. (in Chinese)

[18] LI B, PENG H, LUO X H, et al. Medical image fusion method based on coupled neural P systems in nonsubsampled shearlet transform domain[J]. International Journal of Neural Systems, 2021, 31(1): 2050050.

[19] CHEN X, LU Y, WANG Y H, et al. CMBF: Cross-modal-based fusion recommendation algorithm[J]. Sensors (Basel, Switzerland), 2021, 21(16): 5275.

[20] WU P S, WANG Z D, ZHENG B X, et al. AGGN: Attention-based glioma grading network with multi-scale feature extraction and multi-modal information fusion[J]. Computers in Biology and Medicine, 2023, 152: 106457.

[21] LU S Y, LIU M Z, YIN L R, et al. The multi-modal fusion in visual question answering: A review of attention mechanisms[J]. PeerJ. Computer Science, 2023, 9: e1400.

[22] 金震东, 刘枫. 浅谈超声内镜的诊断标准及操作规范[J]. 临床消化病杂志, 2006, 18(3): 132-134.
JIN Z D, LIU F. Discussion on diagnostic criteria and operating specifications of ultrasonic endoscope[J]. Chinese Journal of Clinical Gastroenterology, 2006, 18(3): 132-134. (in Chinese)

[23] LI M, WANG C J, ZHANG H Y, et al. MV-RAN: Multi-view recurrent aggregation network for echocardiograph-

ic sequences segmentation and full cardiac cycle analysis[J]. Computers in Biology and Medicine, 2020, 120: 103728.

[24] XU L, WU H, HE C M, et al. Multi-modal sequence learning for Alzheimer's disease progression prediction with incomplete variable-length longitudinal data[J]. Medical Image Analysis, 2022, 82: 102643.

[25] KAUR M, SINGH D. Multi-modality medical image fusion technique using multi-objective differential evolution based deep neural networks[J]. Journal of Ambient Intelligence and Humanized Computing, 2021, 12(2): 2483-2493.

[26] ZHANG Z Z, CHEN P J, SHI X S, et al. Text-guided neural network training for image recognition in natural scenes and medicine[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(5): 1733-1745.

[27] ZHOU Q, YE S Z, WEN M W, et al. Multi-modal medical image fusion based on densely-connected high-resolution CNN and hybrid transformer[J]. Neural Computing and Applications, 2022, 34(24): 21741-21761.

[28] ORLHAC F, BOUGHDAD S, PHILIPPE C, et al. A postreconstruction harmonization method for multicenter radiomic studies in PET[J]. Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine, 2018, 59(8): 1321-1328.

[29] HERNANDEZ PETZSCHE M R, DE LA ROSA E, HANNING U, et al. ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset[J]. Scientific Data, 2022, 9(1): 762.

[30] DE STEFANO N, BATTAGLINI M, PARETO D, et al. MAGNIMS recommendations for harmonization of MRI data in MS multicenter studies[J]. NeuroImage Clinical, 2022, 34: 102972.

[31] BORDIN V, BERTANI I, MATTIOLI I, et al. Integrating large-scale neuroimaging research datasets: Harmonisation of white matter hyperintensity measurements across Whitehall and UK Biobank datasets[J]. NeuroImage, 2021, 237: 118189.

[32] RAJAGOPAL A, REDEKOP E, KEMISETTI A, et al. Federated learning with research prototypes: Application to multi-center MRI-based detection of prostate cancer with diverse histopathology[J]. Academic Radiology, 2023, 30(4): 644-657.

[33] VESAL S, GAYO I, BHATTACHARYA I, et al. Domain generalization for prostate segmentation in transrectal ultrasound images: A multi-center study[J]. Medical Image Analysis, 2022, 82: 102620.

[34] STARMANS M P A, ARIF M, et al. A multi-center, multi-vendor study to evaluate the generalizability of a radiomics model for classifying prostate cancer: High grade vs. low grade[J]. Diagnostics (Basel, Switzerland), 2021, 11(2): 369.

[35] KARANI N, ERDIL E, CHAITANYA K, et al. Test-time adaptable neural networks for robust medical image segmentation[J]. Medical Image Analysis, 2021, 68: 101907.

[36] WANG R Z, ZHENG G Y. CyCMIS: Cycle-consistent cross-domain medical image segmentation via diverse image augmentation[J]. Medical Image Analysis, 2022, 76: 102328.

[37] LI L, ZIMMER V A, SCHNABEL J A, et al. AtrialGeneral: Domain generalization for left atrial segmentation of multi-center LGE MRIs[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2021: 557-566.

[38] WU J T Y, DE LA HOZ M Á A, KUO P C, et al. Developing and validating multi-modal models for mortality prediction in COVID-19 patients: A multi-center retrospective study[J]. Journal of Digital Imaging, 2022, 35(6): 1514-1529.

[39] LIU J P, LIU H, GONG S B, et al. Automated cardiac segmentation of cross-modal medical images using unsupervised multi-domain adaptation and spatial neural attention structure[J]. Medical Image Analysis, 2021, 72: 102135.

[40] TOMAR D, LORTKIPANIDZE M, VRAY G, et al. Self-attentive spatial adaptive normalization for cross-modality domain adaptation[J]. IEEE Transactions on Medical Imaging, 2021, 40(10): 2926-2938.

[41] ZHANG Y H, JIANG H, MIURA Y, et al. Contrastive learning of medical visual representations from paired images and text[EB/OL]. (2020-10-02)[2023-12-05]. http://arxiv.org/abs/2010.00747.

[42] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[EB/OL]. (2021-02-26)[2023-12-05]. http://arxiv.org/abs/2103.00020.

[43] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2020-10-22)[2023-12-05]. http://arxiv.org/abs/2010.11929.

[44] VASWANI A, SHAZEER N, PARMAR N, et al. Atten-

tion is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[45] BONMATI E, HU Y P, GRIMWOOD A, et al. Voice-assisted image labeling for endoscopic ultrasound classification using neural networks[J]. IEEE Transactions on Medical Imaging, 2022, 41(6): 1311-1319.

[46] SEVEN G, SILAHTAROGLU G, SEVEN O O, et al. Differentiating gastrointestinal stromal tumors from leiomyomas using a neural network trained on endoscopic ultrasonography images[J]. Digestive Diseases (Basel, Switzerland), 2022, 40(4): 427-435.

[47] LEE M W, KIM G H, KIM K B, et al. Digital image analysis-based scoring system for endoscopic ultrasonography is useful in predicting gastrointestinal stromal tumors[J]. Gastric Cancer: Official Journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association, 2019, 22(5): 980-987.

[48] ZHU C, HUA Y F, ZHANG M, et al. A multimodal multipath artificial intelligence system for diagnosing gastric protruded lesions on endoscopy and endoscopic ultrasonography images[J]. Clinical and Translational Gastroenterology, 2023, 14(10): e00551.

[49] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[50] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2023-12-05]. http://arxiv.org/abs/1409.1556.

[51] OH C K, KIM T, CHO Y K, et al. Convolutional neural network-based object detection model to identify gastrointestinal stromal tumors in endoscopic ultrasound images[J]. Journal of Gastroenterology and Hepatology, 2021, 36(12): 3387-3394.

[52] KIM Y H, KIM G H, KIM K B, et al. Application of a convolutional neural network in the diagnosis of gastric mesenchymal tumors on endoscopic ultrasonography images[J]. Journal of Clinical Medicine, 2020, 9(10): 3162.

[53] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 2261-2269.

作者简介

樊　琳　女,博士研究生. 主要研究方向为计算机视觉及医学图像分析.
E-mail: linfan@my.swjtu.edu.cn

龚　勋　男,博士,教授,博士生导师. 主要研究方向为计算机视觉、人工智能和医学图像分析.
E-mail: xgong@home.swjtu.edu.cn

郑岑洋　男,硕士研究生. 主要研究方向为计算机视觉及医学图像分析.
E-mail: Z_C_Y@my.swjtu.edu.cn