面向肺部肿瘤分类的跨模态Light-3Dformer模型

周 涛[1,2],牛玉霞[1,2*],叶鑫宇[1,2],刘 隆[1,2],陆惠玲[3]
(1. 北方民族大学计算机科学与工程学院,宁夏银川 750021;2. 北方民族大学图像图形智能处理国家民委重点实验室,宁夏银川 750021;3. 宁夏医科大学医学信息与工程学院,宁夏银川 750004)

摘 要: 基于深度学习的三维多模态正电子发射型断层扫描/计算机断层扫描(Positron Emission Tomography/Computed Tomography,PET/CT)肺部肿瘤识别是一个重要的研究方向.肺部肿瘤病灶的空间形状不规则、与周围组织边界模糊,导致模型难以充分提取肿瘤特征,且模型在三维任务中需要较高的计算复杂度.针对上述问题,本文提出一种跨模态Light-3Dformer的三维肺部肿瘤识别模型.本文的主要创新工作有以下几个方面.首先,采用主、辅网络结构,其中主干网络提取PET/CT图像特征,辅助网络提取PET图像和CT图像特征,并采用轻量化跨模态协同注意力实现多模态特征增强和交互式学习.其次,设计Light-3Dformer模块,在该模块中,将Transformer的2次矩阵乘法操作更新为全局注意力机制Lightformer的线性元素乘法操作;设计级联Lightformer结构,其输出特征图和最初的输入特征图融合,通过并行和融合更多的深浅层特征,可以实现轻量化和提取丰富的梯度信息;设计无参数的注意力,该机制能从通道、空间和断层3个方面增强肺部肿瘤特征提取能力.再次,设计轻量化跨模态协同注意力模块(Light Cross-modal Collaborative Attention Module,LCCAM),该模块能充分学习三维多模态影像的跨模态优势信息,对深浅层特征进行交互式学习.最后,进行消融实验和对比实验,在自建的肺部肿瘤三维多模态数据集中,本文模型在计算量和运行时间最优的前提下,准确率和曲线下面积(Area Under the Curve,AUC)值分别达到90.19%和89.81%,与3D-SwinTransformer-S模型相比,参数量降低117倍,计算量降低400倍.实验结果表明:本文模型能更好地提取肺部肿瘤病灶的多模态信息,这为深度学习三维模型轻量化和多模态交互提供了新思路.

关键词: 肺部肿瘤;多模态图像;Transformer;Light-3Dformer;轻量化跨模态协同注意力
基金项目: 国家自然科学基金(No.62062003);宁夏自然科学基金(No.2023AAC03293)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2025)03-0951-11
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240642

Cross-Modal Light-3Dformer Model for Lung Tumor Classification

ZHOU Tao[1,2], NIU Yu-xia[1,2*], YE Xin-yu[1,2], LIU Long[1,2], LU Hui-ling[3]
(1. *School of Computer Science and Engineering*, *North Minzu University*, *Yinchuan*, *Ningxia* 750021, *China*;
2. *Laboratory of Image & Graphics Intelligent Processing of State Ethnic Affairs Commission*, *North Minzu University*, *Yinchuan*, *Ningxia* 750021, *China*; 3. *School of medical information & Engineering*, *Ningxia Medical University*, *Yinchuan*, *Ningxia* 750004, *China*)

Abstract: Recognition of 3D multimodal positron emission tomography/computed tomography (PET/CT) lung tumor using deep learning is an important research area. In medical images of lung tumors, the spatial shape of lesions is irregular and the boundary between the lesions and the surrounding tissues is blurred, which makes it difficult for the model to fully extract tumor features, and the computational complexity of the model is higher in three-dimensional tasks. To solve the above problems, a cross-modal Light-3Dformer 3D lung tumor recognition model is proposed in this paper. The main contributions of this paper are as follows. Firstly, the backbone network extracts PET/CT image features, and the auxiliary network extracts PET image features and CT image features. Multi-modal feature enhancement and interactive learning are realized by lightweight cross-modal collaborative attention. Secondly, Light-3Dformer module are designed. In this module, Updating the 2 times matrix multiplication operation of Transformer to the linear element multiplication operation of Lightformer; The cascade Lightformer structure is designed, the output feature map of the cascade Lightformer structure and the initial input feature map are fused, through parallel and deep and shallow feature fusion, lightweight and rich gradient information can be realized; Designing with parameter less attention, this structure can enhance the ability of lung tumor feature

收稿日期:2024-07-09;修回日期:2025-02-27;责任编辑:朱梅玉
*通讯作者:牛玉霞

extraction from three aspects: channel, space, and tomography image. Thirdly, lightweight cross-modal collaborative attention module (LCCAM) is designed, which can fully learn the cross-modal advantage information of 3D multi-modal images and carry out interactive learning of deep and shallow features. Finally, ablation experiments and comparative experiments. In the self-built 3D multi-modal data set of lung tumor, the accuracy and area under the curve (AUC) values of the model are 90.19% and 89.81%, respectively, under the premise of optimal computation and running time. Comparing with the 3D-SwinTransformer-S model, the computation quantity is reduced by 117 times, and the calculation quantity is reduced by 400 times. The experimental results show that the model can better extract multi-modal information of lung tumor lesions, which provides a new idea for lightweight and multi-modal interaction of deep learning 3D models.

Key words: lung tumor; multimodal images; Transformer; Light-3Dformer; light cross-modal collaborative attention

Foundation Item(s): National Natural Science Foundation of China (No.62062003); Natural Science Foundation of Ningxia Province (No.2023AAC03293)

1 引言

癌症被认为是世界上死亡和发病的关键原因,文献[1]指出预计2030年之前将有2 700万新癌症病例.早期识别和治疗肺部癌症可以降低死亡率,器官病灶的影像结构复杂,不同成像方式提供的诊断价值各异和信息多样,计算机断层扫描(Computed Tomography,CT)图像的密度分辨率较高和解剖关系明确,正电子发射型断层扫描(Positron Emission Tomography,PET)图像中,肿瘤由于更快的新陈代谢而具有较高对比度,图像分辨率较低导致肿瘤边缘模糊和强度不明显,单模态成像技术难以精准识别肺部肿瘤.肺部肿瘤的分析与诊断是一个复杂的过程,大量的阅片工作使得漏诊、误诊率高.因此,研发计算机辅助诊断系统以快速识别肺部肿瘤,从而有效提升诊断效率,是至关重要的.

深度学习模型通过学习影像数据获得优良特征表达,可以协助医生对疾病做出更精准诊断.文献[2]提出低剂量CT肺结节分类的自监督方法,利用邻近特征构建监督信息进行去噪,并通过联合去噪和分类获得了较高性能;文献[3]提出用于CT肺部肿瘤识别的双路径卷积神经网络(Convolutional Neural Network,CNN),分别学习局部和全局特征,通过判别相关分析进行集成获得了较好性能;文献[4]将三维肺部肿瘤CT转换为二维断层切片,输入到CNN中获得了99.8%准确率的识别性能;文献[5]提出二维三维级联的策略识别肺部肿瘤,然后进行精准的分割和分类,在肺部图像数据库联盟(Lung Image Database Consortium,LIDC)数据库中获得了90.01%的敏感度;文献[6]提出22层卷积结构的肺癌自动诊断模型,学习CT和医疗物联网数据的潜在特征,实现了91.6%的准确率和4级肺癌分类;文献[7]针对图像病变区域形状复杂、大小不一,与周围组织的边界模糊,提出新冠肺炎胸部X-ray图像识别模型M^3 Res-Transformer,在胸部X-ray图像识别任务中有效地提升了识别精度.

医学图像分类识别模型的轻量化设计是一个重要研究方向,其中针对肺部肿瘤疾病设计适合移动设备高效和易于部署的轻量级模型十分必要,基于CNN的轻量化通常使用分组操作、轻量级卷积操作、缩减模型层数和通道数等;文献[8]提出轻量级三维AlexNet肺部肿瘤分类模型,在LUNA16上实现了97.17%的准确率,并利用梯度类激活进行可视化;文献[9]提出双分支并行的轻量SqueezeNet,在LUNA16数据集上获得93.2%的二维影像准确率和94.3%的三维影像准确率;文献[10]提出三维多尺度双路径网络,降低了模型的复杂性,同时考虑三维CT的特性,获得了较高准确性;文献[11]提出基于重参的增强双流网络,提高识别小病变能力的同时减少了参数量,在性能和成本之间实现了很好的权衡.

深度学习在三维肺部影像上应用广泛,文献[12]提出三维放射学肺癌分类CT模型,获得了84%曲线下面积(Area Under the Curve,AUC)的较高性能;文献[13]提出多分辨率三维模型,学习肺部CT图像中不同分辨率的三维体积数据,获得87%的准确率;文献[14]提出三维肿瘤追踪方法,双编码器学习CT图像特征进行分割和分类,获得92%的召回率;文献[15]提出多视图卷积递归网络,利用CT肺部肿瘤的形状、大小和跨断层变化,获得了更好的泛化和稳健学习能力;Vision Transformer[16]广泛应用于智能医学图像处理领域,文献[17]提出三维Transformer识别肺部肿瘤,将三维CT图像转换为向量特征进行预测;文献[18]提出向量对局部和全局信息编码的MobileViT,将卷积中的局部处理方式替换为全局处理,模型兼具CNN和Transformer的特性,用少量参数和简单的训练方法,就可实现轻量高效的模型.

深度学习在PET/CT多模态肺部影像上应用广泛,文献[19]利用CNN获得PET与CT空间变化的融合图,量化每个不同模态特征的重要性以获得多模态互补表示,获得了99.29%的识别准确度;文献[20]利用集成CNN学习肺部肿瘤CT、PET、PET/CT多模态图像,以更少耗时获得了较高的识别准确率;文献[21]利用PET/

CT 肿瘤图像在标准剂量和低剂量下获得 95.9% 和 91.5% 的敏感度;文献[22]采用残差网络(Residual Network,ResNet)学习 PET 和 CT 特征,支持向量机学习临床数据,获得 82%AUC 结果;文献[23]提出跨模式三维网络预测肺癌,学习三维肿瘤块、临床特征和病理标签,在 401 个病例中取得了 92.6%AUC 的较优性能;文献[24]利用预训练的三维 VGG19 对 PET/CT 图像进行良恶性肿瘤识别,在 2 557 个病灶中实现了 88% 的准确率.

上述研究表明,面向三维影像数据的 CNN 和 Transformer 模型在肺部肿瘤识别中应用广泛,结合多模态图像的信息可以获得更准确的疾病信息. 在实际的三维肺部肿瘤识别任务中,三维影像信息更丰富和复杂,且肺部肿瘤存在形状不规则、差异性大等特点,导致模型的特征提取不充分. 而基于 CNN 或 Transformer 的方法通常通过更大数据集来提高性能,或者是引入更复杂的网络结构,这些策略都不同程度地增加了模型的时间复杂度和空间复杂度,导致需要的计算和存储资源都增加,限制了模型在复杂场景中的应用. 为此,本文提出一种基于跨模态 Light-3Dformer 的三维 PET/CT 肺部肿瘤分类识别模型,该方法的创新点主要有以下方面:

(1)为充分挖掘三种模态的特征,提高模型学习跨模态医学图像互补信息的能力,设计主辅结构的跨模态 Light-3Dformer 模型,主干网络提取 PET/CT 图像特征,辅助网络提取 PET 图像和 CT 图像特征,网络有 4 个阶段,每个阶段提取不同尺寸的特征,并在不同尺寸特征之间采用跨模态注意力对多个模态特征进行交互式学习.

(2)针对肺部肿瘤特征提取不充分和三维模型计算复杂度高的问题,设计轻量化的特征提取模块 Light-3Dformer,采用 Lightformer 结构和深浅层融合方式,采用线性元素乘法,高效提取多模态肺部肿瘤的深浅层特征;利用无参数的注意力机制,从通道、空间和断层 3 个方面对深浅层融合后的特征进行增强.

为充分学习三维多模态影像的跨模态优势信息,设计轻量化跨模态协同注意力模块(Light Cross-modal Collaborative Attention Module,LCCAM),对不同模态的深浅层特征进行交互式学习,利用 Lightformer 学习跨模态图像的语义相关性和建模多模态特征的远距离相互依赖关系,$\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 由不同模态特征产生,交互式增强跨模态肺部肿瘤特征,使模型可学习到更关键的功能和解剖信息,以获得更优的识别能力.

2　Light-3Dformer 方法模型

三维跨模态 Light-3Dformer 模型在四个阶段的不同分辨率特征图中,将 PET 图像和 CT 图像特征拼接合并后,与 PET/CT 图像特征通过轻量化跨模态协同注意力进行交互式互补,从三种模态图像中同时学习病灶特征,其中交互增强后的 PET 图像和 CT 图像还需按通道数切分,跨模态 Light-3Dformer 结构如图 1 所示. 其中特征提取模块 Light-3Dformer,在保持轻量化同时充分提取 PET、CT 和 PET/CT 信息,每阶段之间采用轻量化跨模态协同注意力进行交互式学习,增强肺癌特征,后经过三维全局池化层和分类层进行肺癌识别.

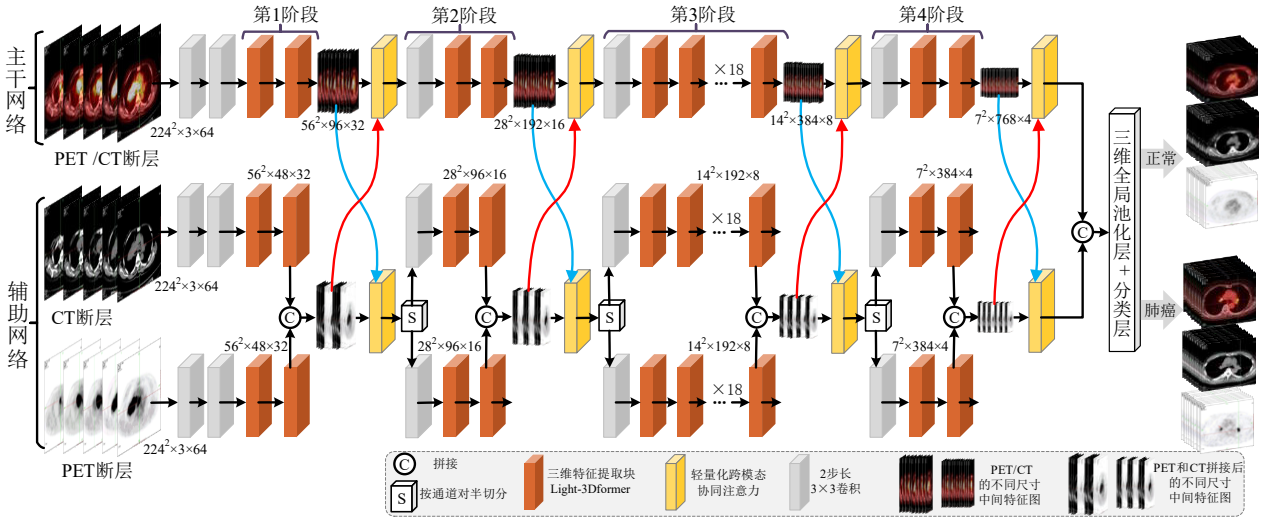

图 1　跨模态 Light-3Dformer 整体结构图

2.1　Light-3Dformer

CNN 通过局部相邻像素点之间的联系提取丰富的局部特征,在图像特征提取方面表现出较好的优势,但忽略了对全局上下文特征建立远程依赖关系的重要性. Transformer 中的自注意力机制能够建模图像全局信息,采用更有效的空间信息编码方法获得了更优秀的表现,但对局部细节信息的关注能力不足,在足够大的数据集上进行训练,局部信息关注问题能得

到缓解，大小有限的肺部肿瘤数据集会导致性能受限，并且在处理高分辨率图像时，Transformer导致的计算成本非常大. 三维模型相比于其他网络架构需要更多的资源消耗，在三维PET/CT多模态影像上直接采用Transformer，不仅性能有限而且资源消耗过多. 此外，在肺部肿瘤识别任务中，三维影像信息更丰富和复杂，而肺部肿瘤存在形状不规则、差异性大等特点，因此，不合理的设计会导致特征提取不充分的问题.

为此，本文设计Light-3Dformer模块，其特征提取采用了级联Lightformer结构，级联Lightformer结构的输出特征图和最初的输入特征图融合，其注意力部分利用无参数的注意力，从通道、空间、断层这3个方面对深浅层融合后的特征进行增强，结构如图2所示. Light-3Dformer通过并行更多的梯度流分支，在保证轻量化的同时获得更丰富的梯度信息，进而实现更高的效率和更优异的性能. 其中Lightformer结构引入了一种新颖的高效全局注意机制，该机制有效地用线性元素乘法取代了二次矩阵乘法运算，$\boldsymbol{K}$与$\boldsymbol{V}$交互可以用线性层代替，而不牺牲任何精度，Lightformer消除了矩阵乘法运算需要的较大资源消耗，显著降低了模型的计算复杂性；由于只使用$\boldsymbol{Q}$与$\boldsymbol{K}$交互，Lightformer复杂度$O(T\times D)$，与三维断层数T和序列数量D具有线性复杂性，相比于Transformer复杂度$O(H\times W\times T\times D)$，从二次幂的复杂度下降到线性复杂度，且无需考虑特征图的空间尺寸，因此，该全局注意力设计可用于所有阶段，在不同尺寸中均可高效地捕获上下文信息.

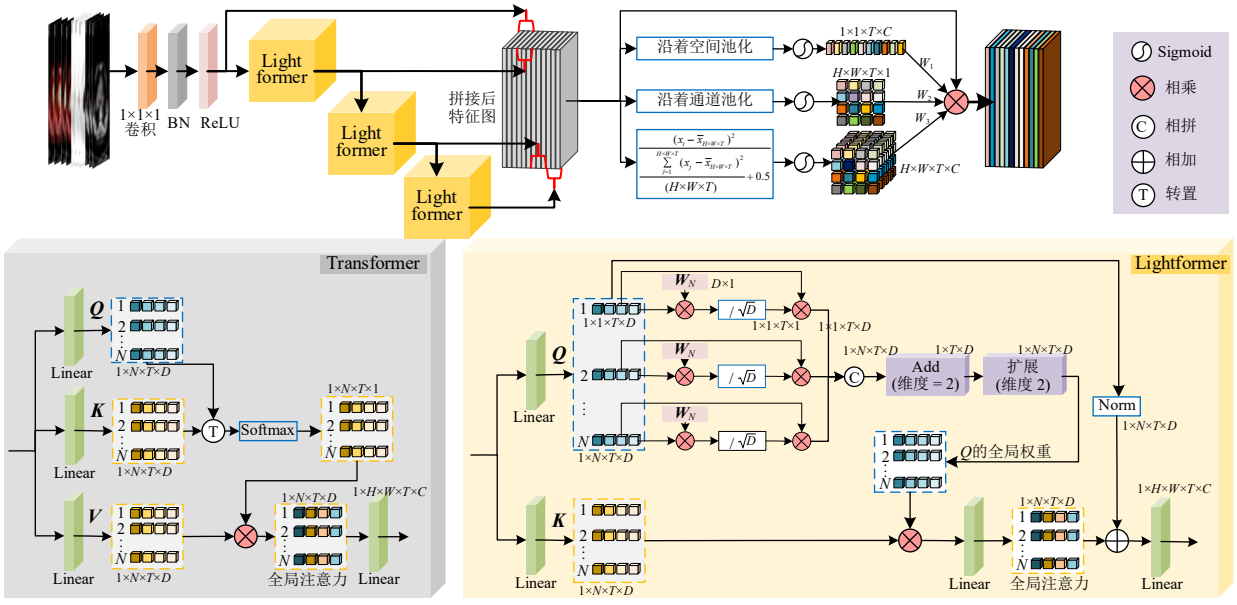

图2 Light-3Dformer结构

Transformer通过线性层Linear获得的查询($\boldsymbol{Q}$)、键值($\boldsymbol{K}$)和值($\boldsymbol{V}$)，$\boldsymbol{Q}$和$\boldsymbol{K}$进行计算，生成注意力权重并应用于输入特征的值$\boldsymbol{V}$，D为$\boldsymbol{Q}$的维度，T为矩阵转置，Transformer输出特征图X_{Trans}如式(1)所示：

$$X_{\text{Trans}}=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{D}}\right)\cdot\boldsymbol{V} \tag{1}$$

不同于Transformer的矩阵乘法，本文采用高效全局注意机制Lightformer，使用线性元素乘法和有效的相加注意力来编码全局上下文. 通过元素乘法来利用序列之间的成对交互，而不是使用点积运算来捕获全局上下文；$\boldsymbol{K}$与$\boldsymbol{V}$交互可以在不牺牲性能的情况下被删除，并且只需通过线性变换有效地编码查询键交互，以及使用更适于视觉任务的相加方式进行合并，就足以学习序列之间的关系. 通过全局$\boldsymbol{Q}$与$\boldsymbol{K}$交互来学习全局上下文信息，然后进行线性转换来计算全局上下文感知的注意力矩阵，该矩阵$\boldsymbol{W}_{\boldsymbol{QK}}$如式(2)所示：

$$\boldsymbol{W}_{\boldsymbol{QK}}=\text{Linear}\left(\boldsymbol{K}\times\left(\text{E}\left(\text{Add}\left(\sum_{i=1}^{n}Q_i\cdot\frac{W_i}{\sqrt{D}}\right)\right)\right)\right) \tag{2}$$

其中，Add为将第2个维度的值全部相加；E为将第2个维度的值进行填充. 首先，Lightformer使用元素乘法对全局查询$\boldsymbol{Q}$和和键$\boldsymbol{K}$矩阵之间的交互进行编码，获得的全局上下文感知的注意力矩阵与Transformer中的注意矩阵相似，两者均可以从每个序列中捕获信息，但Lightformer可以更灵活地学习到输入序列中的相关性，且计算成本较低，仅与序列长度呈线性复杂度. 其次，通过有效的相加注意力来编码全局上下文，对输入$\boldsymbol{Q}$进行归一化(Norm)以保证梯度流的稳定. 最后，采用线性层对$\boldsymbol{Q}$与$\boldsymbol{K}$进行交互，以学习序列的隐藏表示，

Lightformer输出特征图X_{Light}如式(3)所示:

$$X_{\text{Light}}=\text{Linear}\left(\boldsymbol{W}_{\boldsymbol{QK}}+\text{Norm}(\boldsymbol{Q})\right) \tag{3}$$

如图2所示,在Light-3Dformer模块中,有三个级联的Lightformer结构,这三个级联Lightformer结构的输出特征图和最初的输入特征图融合,通过并行和融合更多的深浅层特征,可以实现轻量化和提取丰富的梯度信息. Light-3Dformer特征提取部分的输出X_{Light3D}如式(4)所示:

$$X_{\text{Light3D}}=X\copyright\left(f_{\text{Light}}(X)\right)\copyright\left(f^{2}_{\text{Light}}(X)\right)\copyright\left(f^{3}_{\text{Light}}(X)\right) \tag{4}$$

其中,©为沿通道维度拼接.

三维肺部肿瘤数据不同于点云类型的三维数据,而是一段连续且厚度较小的断层数据组成,断层之间病灶的相对位置是固定的,如本文采用的PET/CT断层影像,尽管扫描区域为整个肺部,但由于肺部肿瘤形状不规则和所占区域偏小,会导致较多断层均为无关的背景信息. 为充分考虑轻量化需求和肺部肿瘤特点,本文在Light-3Dformer的注意力部分设计了一种三分支并行结构的三维无参数注意力,在通道和空间注意力分支中,采用无参数的池化操作,分别对通道和空间维度进行增强,同时,对断层维度也进行了增强,在第3分支中,计算特征值与特征平均值之间的差值,差值越大,表明该特征包含的信息量越多. 根据每个特征的信息量,赋予其不同的重要性,即实现逐像素权重.

通道注意力分支沿着通道对空间维度进行平均计算,以此学习每个通道和断层数中重要的肺部肿瘤特征,增强通道维度的特征表征能力. 空间注意力分支沿着空间维度对通道维度进行平均计算,以此学习每个空间位置中重要的肺部肿瘤特征,实现对空间和断层数维度中肺癌特征的增强. 第3条逐像素注意力分支不引入额外参数,仅需少量计算量就可提高模型的特征表达能力,计算每个特征值与当前通道特征平均值$\bar{x}_{H\times W\times T}$之间的差值,为避免部分特征与平均值之间差异不明显,逐像素注意力分支中使用平方的加权方式可产生更大的特征差异化,获得更多的易区分特征,也易于网络的优化训练,最终逐像素计算生成权重W_i,如式(5)所示:

$$W_i=(x_i-\bar{x}_{H\times W\times T})^2\Bigg/\left(\frac{\sum_{j=1}^{H\times W\times T}(x_j-\bar{x}_{H\times W\times T})^2}{H\times W\times T}+0.5\right) \tag{5}$$

其中,W_i为对每个特征依次进行计算;x_i为输入特征图的当前通道上第i个空间特征值;H、W和T分别为输入特征图的高度、宽度和断层数. 三维无参数注意力生成的通道、空间和逐像素权重,加权到原始特征图中以增强肿瘤特征和抑制背景信息. 三维无参数注意力被添加至每个Light-3Dformer末端,其可以在保持轻量化和较高计算效率的前提下,提升网络的信息捕获能力和识别性能,使网络更好地聚焦于肿瘤特征.

Light-3Dformer通过控制浅层和深层的梯度路径,使网络能够高效地学习到有效特征,具有较强的鲁棒性,Lightformer所采用的高效全局注意机制,将模型复杂度与空间尺寸的相关性解耦,且由二次型有效地降低到线性型,可以较好地兼顾模型性能和计算负担,充分学习三维多模态肺部肿瘤的丰富特征. Light-3Dformer结合无参数三分支注意力,无参数增加和极少的计算量消耗,便可更好地捕获到形状不规则、差异性大的肺部肿瘤.

2.2　轻量化跨模态协同注意力模块

成像机理不同的三维多模态图像之间存在很多不一致的信息,尽管通过Light-3Dformer中的无参数注意力,可以对各模态特征通道、空间和断层进行学习和重新校准,但在多模态肺部肿瘤识别中使用不合理的跨模态特征融合,会导致PET、CT和PET/CT多模态图像中包含肿瘤信息的关键特征难以有效提取和增强,从而导致识别精度较低. 此外,深层语义信息可以辅助不同模态的浅层定位信息,需要考虑不同模态深层特征和浅层特征之间信息传递的必要性. 为获取多模态丰富的细节和语义信息,并实现识别精度和识别速度的折中,本文设计LCCAM,对不同模态的深浅层特征进行交互和学习,具体结构如图3所示,其中,利用Lightformer充分学习跨模态图像的语义相关性,学习多模态特征的远距离相互依赖关系并进行重新校准,交互式增强多模态图像中肿瘤特征.

多模态融合包括像素级融合、特征级融合和决策级融合,像素级融合是将数据类型转化一致后进行拼接,特征级融合是将多模态数据经过各自的特征提取后进行拼接,决策级融合是将特征提取后会融合全部分支,由于本文识别对象为肺部肿瘤,决策级融合操作与特征级融合相同. 而像素级融合和特征级融合只对能对多模态数据融合一次,无法对关键特征进行捕获和增强,为此,本文现将PET和CT特征图拼接为PET-CT特征图,设计如图3所示的4种融合方式,如图3(a)跨模态融合注意力所示,在同一尺寸多模态特征图中,利用PET-CT对PET/CT进行增强. 如图3(b)跨模态协同注意力所示,在同一尺寸多模态特征图中,利用PET-CT和PET/CT特征图进行交互式的协同增强. 其中,Transformer用于编码PET-CT特征和PET/CT特征,主分支将特征映射为$\boldsymbol{Q}$,再与另一分支映射的$\boldsymbol{K}$和$\boldsymbol{V}$进行全局特征学习,通过在不同模态特征图内学习和交互,实现对三维多模态图像中病灶信息进行增强.

如图3(c)轻量化跨模态融合注意力所示,替换跨模态融合注意力中的Transformer为高效全局注意机制

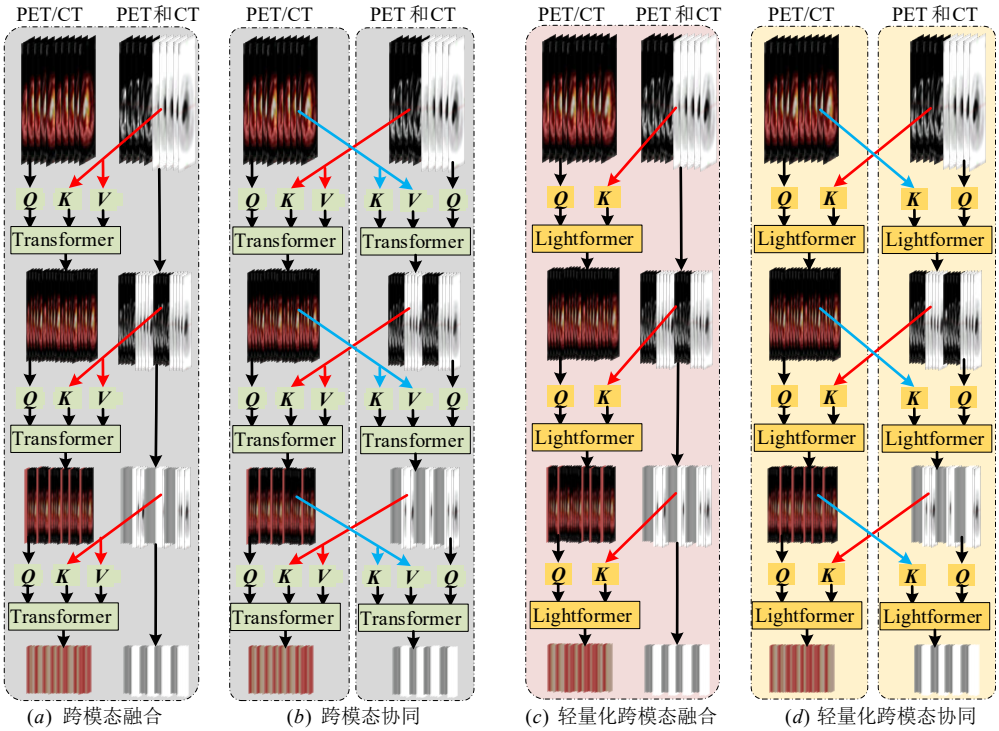

图3 轻量化跨模态协同注意力模块

Lightformer，其中，PET/CT特征被映射为$\boldsymbol{K}$，$\boldsymbol{K}$与PET/CT特征映射的$\boldsymbol{Q}$之间通过Lightformer进行增强，使用线性元素乘法和有效的相加注意力来编码跨模态全局上下文，融合跨模态特征并增强肺部肿瘤信息. 如图3(d)轻量化跨模态协同注意力所示，对PET-CT和PET/CT特征图进行高效的交互式协同增强，其使用仅与序列长度呈线性复杂度的Lightformer对输入序列中的相关性进行灵活学习，主分支将特征映射为$\boldsymbol{Q}$，再与另一分支映射的$\boldsymbol{K}$使用较低计算成本进行全局特征学习，然后通过有效的相加注意力来编码跨模态全局上下文，实现对不同模态特征图序列内的学习和交互，从而使模型可以高效地增强和学习三维跨模态肺部肿瘤病灶信息.

3 实验和讨论

3.1 实验数据集和评价指标

本文实验数据集选用的是宁夏某三甲医院在2014年1月至2021年7月期间收集的733例正常和845例肺部肿瘤临床患者，在Discovery MI仪器中进行肺部及躯干部图像采集，获取已配准的PET、CT和PET/CT三维肺部肿瘤图像. 同时合并由1 176例正常和419例肺部肿瘤患者的Data Science Bowl 2017数据集，每例大约有102~289张肺部断层切片. 按6∶2∶2比例分成训练集、验证集和测试集进行实验，本次实验为内嵌Ubuntu18.04 LTS子系统的64位Windows11专业版系统，128 GB内存，搭载2块Intel Xeon E5-2696v3的36核CPU处理器，并使用4块TITAN Xp显卡加速图像处理，采用自适应矩估计(Adaptive moment estimation Weight，AdamW)优化器进行优化，采用0.01的初始学习率和每10周期0.9的衰减策略，设置权重衰减值为1×10^{-5}，训练周期为300，训练批处理大小为24.

本文使用分类常用评价指标，根据模型预测结果分类错误和正确的个数，得到真正类(True Positive，TP)、假正类(False Positive，FP)、假负类(False Negative，FN)、真负类(True Negative，TN). 准确率(Accuracy)为全部类预测正确的比例，精确率(Precision)为正类且模型预测正确占所有正类的比例，召回率(Recall)为模型所预测出的正类占所有正类的比例，F_1分数(F_1)如式(6)所示：

$$F_1=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{6}$$

将受试者工作特征(Receiver Operating Characteristic，ROC)曲线以敏感度即真正类率(True Positive Rate，TPR)为纵轴、假正类率(False Positive Rate，FPR)为横轴进行绘制，TPR值等于召回率，FPR、特异度(True Negative Rate，TNR)如式(7)、式(8)所示：

$$\text{FPR}=\frac{\text{FP}}{\text{FP}+\text{TN}} \tag{7}$$

$$\text{TNR}=\frac{\text{TN}}{\text{FP}+\text{TN}} \tag{8}$$

ROC曲线下面积定义为AUC，越靠近左上角AUC值越大，表示模型的排序和分类性能会越好，评价指标

均是值越大表示模型越好.

3.2　消融实验与分析

为了评估模型每个模块的有效性,本文在3.1节所述肺部肿瘤PET/CT多模态三维数据集上进行消融实验. 使用3D-Mobile-ViT作基础模型进行8组实验,在这8组实验中,实验1到实验3采用的是单模态模型,实验4到实验8采用的是多模态模型. 实验1到实验3是在3D-Mobile-ViT模型基础上逐渐加入Lightformer、深浅层融合以及无参数的注意力3部分,实验1在PET/CT影像上使用Lightformer,实验2对Lightformer进行深浅层融合方式,实验3在实验2的基础上使用无参数注意力;实验4在实验3的基础上使用三种多模态图像进行特征级融合,实验5在改进特征级融合为跨模态融合注意力,实验6使用跨模态协同注意力,实验7使用轻量化跨模态融合注意力,实验8使用轻量化跨模态协同注意力. 实验结果对比如表1所示,主要从参数量、计算量、准确率、AUC值、召回率、F_1分数、精确率以及训练时间8个指标上对实验进行分析.

表1　消融实验结果对比

实验	模态	模型	参数量	计算量	准确率/%	AUC值/%	召回率/%	F_1分数/%	精确率/%	训练时间/s
—	单模态	3D-Mobile-ViT[15]	2.03 M	1.08 G	84.23	83.47	79.77	80.12	80.47	32 457
1		+Lightformer	1.20 M	523.39 M	84.93	84.21	80.65	81.01	81.36	30 236
2		+深浅层融合方式	402.38 K	144.35 M	85.51	84.54	79.77	81.44	83.18	27 991
3		+无参数的注意力	402.38 K	144.42 M	86.68	86.06	82.99	83.24	83.48	28 038
4	多模态	特征级融合三模态	403.17 K	144.69 M	87.50	86.94	84.16	84.29	84.41	28 093
5		跨模态融合注意力	453.06 K	196.37 M	88.20	87.77	85.63	85.26	84.88	29 054
6		跨模态协同注意力	502.95 K	248.05 M	89.37	88.94	86.80	86.68	86.55	29 308
7		轻量化跨模态融合注意力	410.14 K	157.89 M	89.02	88.59	86.51	86.26	86.01	28 511
8		轻量化跨模态协同注意力	417.09 K	170.55 M	90.19	89.81	87.98	87.72	90.19	28 749

实验1、实验3和实验8的热力图如图4所示,选取3个患者的部分断层数据进行分析,3个患者的断层数分别为3个、2个和2个,CT图像肿瘤和正常组织密度差异不明显,PET图像中肿瘤区域代谢旺盛,呈高亮,因此多模态图像可以更好地识别和定位病灶. 图4中伪彩被用来表示网络对图像不同区域的关注程度,红色程度越深表示网络对这个区域的关注度越高,反之,蓝色表示网络对该区域关注度越低.

图4第1行为三维PET/CT肺部肿瘤图像,已对部分病灶区域进行了标注,如表1所示,实验1到实验3为Light-3Dformer各组件的消融数据. 与3D-Mobile-ViT相比,实验1参数量和计算量下降40.88%和52.67%,准确率提升0.83%,表明Lightformer将模型复杂度由二次型有效地降低到线性型,可以较好地兼顾模型性能和计算负担. 实验2的参数量下降80.64%且训练时间缩短13.75%,指标整体上小幅提升,表面深浅层特征融合方式可以进一步轻量化,同时,获得更丰富的梯度信息. 其中,召回率降低表示肺癌病症的聚焦能力不足,从图4第3行的热力图也可看出模型关注区域较大,对病灶区域的关注存在较大误差且容易关注到非肺部区域. 实验3相比于实验2,无需增加额外参数,仅需部分计算量,所需额外计算时间较少,准确率和AUC值提升1.36%和1.79%,表明这种无参数的注意力在保证轻量化同时可提升模型识别肺癌的鲁棒性,从通道、空间和断层数维度,以及利用特征值与平均值的差计算信息量生成逐像素权重,使模型可以学习到更多的可区分特征,从图4第4行的热力图也可以看出模型可以较好地关注肺部肿瘤区域,但对于大小不明显和与器官特征相近的肺部病灶区分较难,如第2和8列关注区域均存在明显误差.

如表1所示,实验4到实验8为轻量化跨模态协同注意力的消融数据. 实验4相比于实验3的准确率和AUC值提升0.94%和1.02%,PET、CT和PET/CT多模态三维图像进行跨模态语义信息的特征互补,三模态进行特征级融合可以较好地增强模型对病灶的聚焦能力. 实验5跨模态融合注意力通过Transformer建模多模态特征的远距离相互依赖关系,将PET、CT对PET/CT进行增强和融合,使得模型更易识别出肺部肿瘤;实验6跨模态协同注意力是在实验5的基础上,使用Transformer将PET/CT对PET、CT进行增强和交互,准确率和AUC值进一步提升1.32%和1.33%,表明充分利用三维多模态图像的功能和解剖信息,可对包含肿瘤信息的特征进行有效增强.

实验7轻量化跨模态融合注意力是在实验5的基础上将Transformer替换为高效全局注意机制Lightformer,计算量降低19.59%的前提下提升0.92%的准确率,表明使用线性元素乘法和有效的相加注意力来编码跨模态全局上下文,既可避免二次矩阵乘法运算的复杂度较高问题,又可以增强肺部肿瘤信息和融合跨模态特征. 实验8轻量化跨模态协同注意力是在实验6的基础

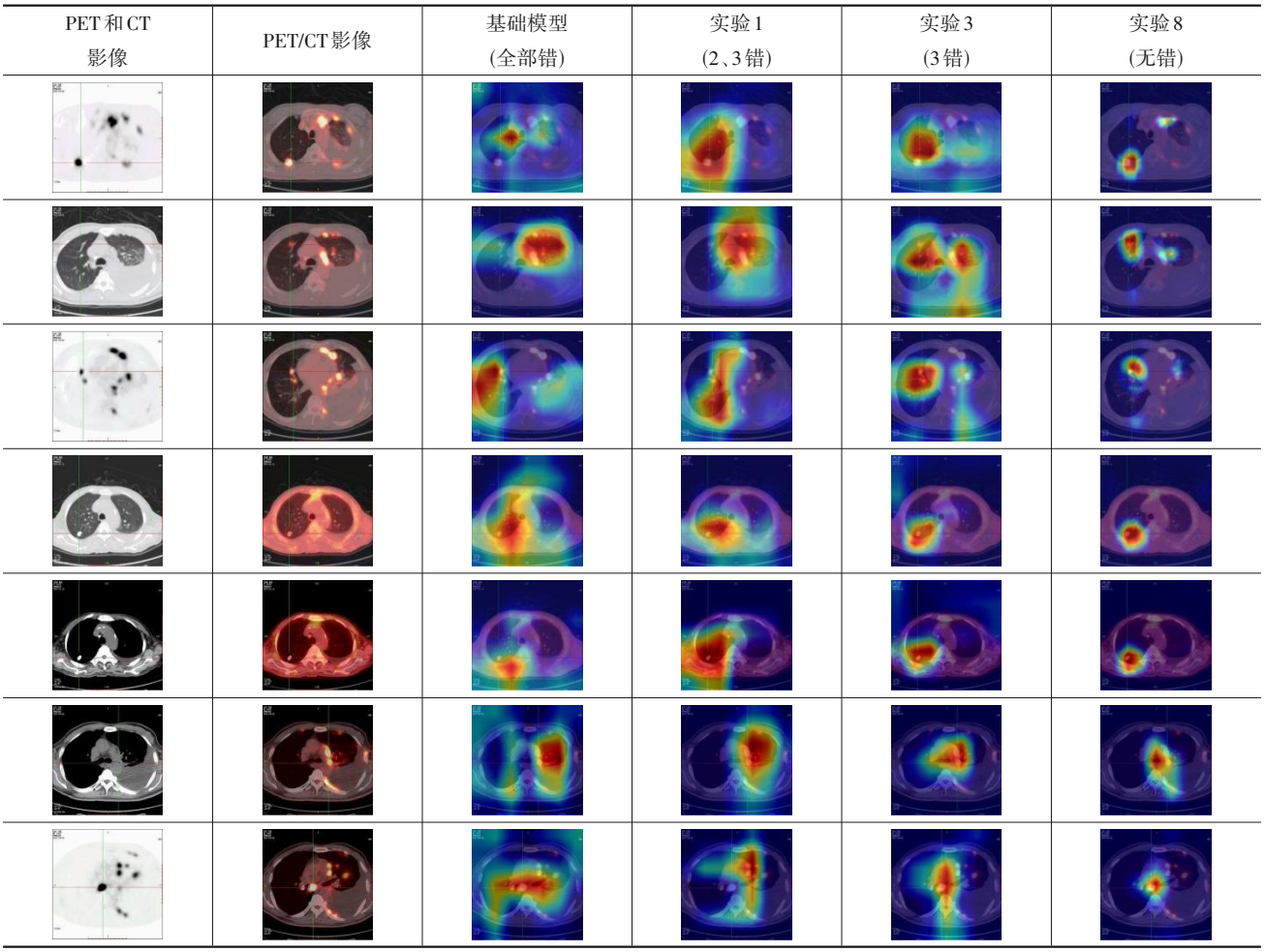

图4 各模型在肺部肿瘤影像上的热力图

上使用Lightformer，降低计算量的同时准确率和AUC值提升0.91%和0.97%，实现模型对不同模态深浅层特征之间进行高效学习和有效交互，从图4第5行的热力图可看出，模型充分利用多模态交互信息对病灶进行定位和识别，对多模态图像的语义相关性进行充分挖掘，可有效提高模型识别能力和检测精度. 实验8相比于3D-Mobile-ViT，参数量降低4.98倍，计算量降低6.48倍，准确率和AUC值提升7.07%和7.59%.

3.3 对比实验与分析

在肺部肿瘤PET/CT多模态三维数据集上，本文将提出的模型Light-3Dformer与2个CNN三维模型、3个Transformer三维模型，以及5个CNN结合Transformer的三维轻量化模型进行了对比实验，各个模型的具体分类结果如表2所示，实验结果表明：本文模型具有最小的参数量、计算量和运行效率，并获得较高的效率和较好的识别能力，对肺部肿瘤的识别能力显著.

相比于2个纯CNN三维模型，本文模型Light-3Dformer较参数量最小的3D-EfficientNet-b3，运行效率快了16.11%. 相比于3个纯Transformer三维模型，本文模型计算效率大幅提高同时获得性能提升，较3D-PoolFormer-S24模型训练时间缩短了23.30%，5项指标（准确率、AUC值、召回率、F_1分数、精确率）提升的平均值约为6.58%；较用于视频识别的3D-BEVT模型，5项指标分别提高了4.76%、4.89%、5.64%、6.09%和9.88%. 本文模型相比于5个CNN结合Transformer三维轻量化模型，其参数量、计算量和训练时间在肺部肿瘤识别方面均具有明显优势；3D-Mobile-ViT将Transformer向CNN嵌入，为尽可能降低计算量和提高推理速度，每阶段仅使用一次，而本文模型采用仅1/6计算量，精确率提升12.07%；与深度可分离全局注意力轻量化模型3D-EdgeNeXt-S相比，本文模型仅以超1/12的计算量实现了5.62%的准确率提升和6.42%的AUC值提升.

本文模型Light-3Dformer相比于性能较好的CNN结合Transformer模型，以较高计算效率对肺部肿瘤进行更精准的识别，在肺部肿瘤PET/CT多模态三维数据集上具有较高的识别精度和较好的分类能力. 3D-CVT-13结合CNN移位、缩放和失真不变性的特性和Trans-

表 2　不同模型在肺部肿瘤 PET/CT 多模态三维数据集上的实验结果

对比模型	参数量	计算量	准确率/%	AUC 值/%	召回率/%	F_1 分数/%	精确率/%	训练时间/s
3D-ResNet50[22]	46.203 M	39.981 G	81.07	79.47	71.55	75.08	78.96	36 256
3D-EfficientNet-b3[20]	1.624 M	205.719 M	82.59	80.53	70.38	76.31	83.33	34 271
3D-SwinTransformer-S[17]	48.751 M	68.210 G	83.64	83.09	80.35	72.25	76.64	38 589
3D-PoolFormer-S24[25]	31.687 M	32.020 G	85.75	85.43	83.87	82.42	81.02	37 486
3D-BEVT[17]	86.628 M	121.158 G	86.09	85.62	83.28	82.68	82.08	41 718
3D-Mobile-ViT[18]	2.028 M	1.082 G	84.23	83.47	79.77	80.12	80.47	32 457
3D-EdgeNeXt-S[26]	6.461 M	2.024 G	85.39	84.39	79.47	81.26	83.13	34 012
3D-CVT-13[27]	21.430 M	26.657 G	84.69	84.01	80.65	80.76	80.88	36 422
3D-CMT-S[27]	27.278 M	23.048 G	86.09	85.38	81.82	82.42	83.04	35 874
3D-NextVit-S[28]	19.900 M	34.610 G	86.92	86.20	82.69	83.43	84.18	34 891
Light-3Dformer	417.090 K	170.550 M	90.19	89.81	87.98	87.72	90.19	28 749

former 动态注意力、全局上下文和更好的泛化性能，与 3D-CVT-13 相比，本文模型的训练时间缩短了 21.06%，5 项指标提升的平均值约为 8.51%；3D-CMT-S 在全局特征提取中引入卷积操作进行细粒度特征提取，并采用模块层次化堆叠以提高性能和节省计算开销，与之相比，本文模型的训练时间进一步缩短了 19.86%，5 项指标提升的平均值约为 6.5%；此外，在工业部署场景中设计的 CNN-Transformer 混合架构 3D-NextVit-S 上，本文模型的训练时间缩短了 17.60%，并且在 5 项指标上分别提高了 3.76%、4.19%、6.39%、5.14% 和 7.13%.

图 5 为 11 种模型的 ROC 曲线，Light-3Dformer 曲线整体位于左上角，其分类性能良好，具有更好的鲁棒性，能较好地学习肺部肿瘤病灶信息. 图 6 为 11 种模型的 PR 曲线，Light-3Dformer 的曲线面积最大，可以看出本文模型性能明显最优.

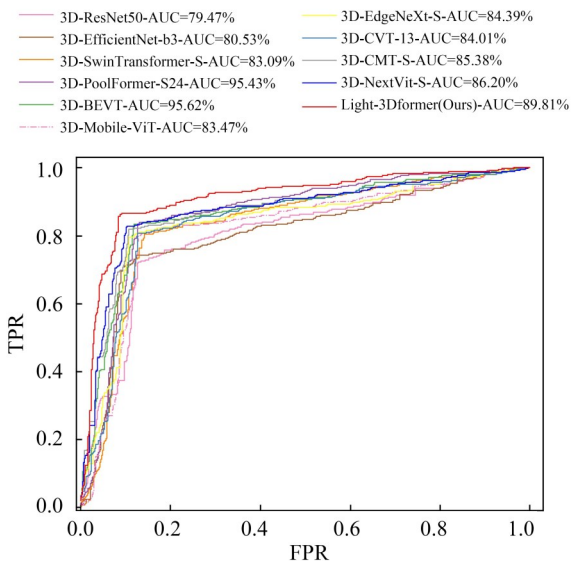

图 5　各模型的 ROC 曲线

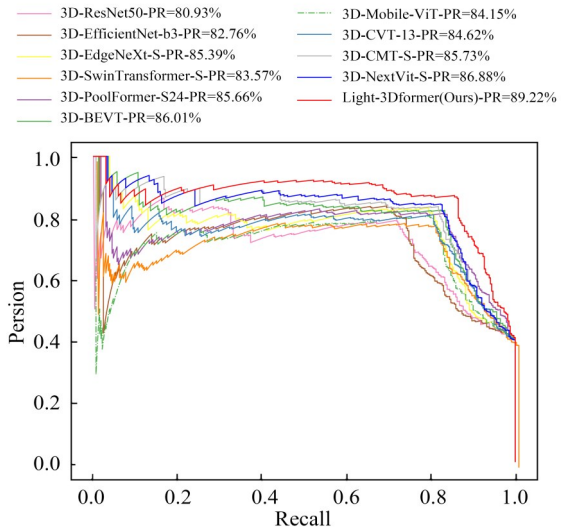

图 6　各模型的 PR 曲线

4　结论

为解决特征提取不充分和模型轻量化程度不足问题，本文提出一种用于肺部肿瘤识别的三维跨模态 Light-3Dformer 模型，主辅助网络学习具有功能和解剖信息的 PET、CT 和 PET/CT 三维多模态肺部肿瘤特征，设计 Light-3Dformer 实现轻量化的同时获得更丰富的梯度信息，其中，全局注意机制 Lightformer 将模型复杂度与空间尺寸的相关性解耦，可以较好地兼顾模型性能和计算代价，充分学习三维多模态肺部肿瘤的丰富特征，设计轻量化跨模态协同注意力机制，通过交互式增强方式，提升多模态图像中肿瘤特征的表征能力. 在肺部肿瘤 PET/CT 多模态三维数据集进行实验，结果表明：本文模型获得了 417.09 K 和 170.55 M 的最高效率，准确率为 90.19%、AUC 值为 89.81%、召回率为 87.98%、F_1 分数为 87.72% 和精确率为 90.19% 的最优识别性能. 热力图可视化技术表明：本文模型具有较好的鲁棒性，该模型对

三维多模态肺部肿瘤分类识别具有积极的意义,为深度学习三维模型轻量化和多模态交互提供了新思路.

参考文献

[1] SHI Z X, LIN J L, WU Y F, et al. Burden of cancer and changing cancer spectrum among older adults in China: Trends and projections to 2030[J]. Cancer Epidemiology, 2022, 76: 102068.

[2] LEI Y M, ZHANG J P, SHAN H M. Strided self-supervised low-dose CT denoising for lung nodule classification[J]. Phenomics, 2021, 1(6): 257-268.

[3] SORI W J, FENG J, GODANA A W, et al. DFD-Net: Lung cancer detection from denoised CT scan image using deep learning[J]. Frontiers of Computer Science, 2020, 15(2): 152701.

[4] PANDYA M, JARDOSH S, THAKKAR A. An efficient IISH-2D DCNN-based lung nodule classification using CT scan images[J]. International Journal of Modeling, Simulation, and Scientific Computing, 2023, 14(1): 2243005.

[5] DUTANDE P, BAID U, TALBAR S. LNCDS: A 2D-3D cascaded CNN approach for lung nodule classification, detection and segmentation[J]. Biomedical Signal Processing and Control, 2021, 67: 102527.

[6] FARUQUI N, YOUSUF M ABU, WHAIDUZZAMAN M, et al. LungNet: A hybrid deep-CNN model for lung cancer diagnosis using CT and wearable sensor-based medical IoT data[J]. Computers in Biology and Medicine, 2021, 139: 104961.

[7] 周涛, 刘赟璨, 侯森宝, 等. M^3 Res-Transformer: 新冠肺炎胸部 X-ray 图像识别模型[J]. 电子学报, 2024, 52(2): 589-601.
ZHOU T, LIU Y C, HOU S B, et al. M^3 res-transformer: Chest X-ray image recognition model of COVID-19[J]. Acta Electronica Sinica, 2024, 52(2): 589-601. (in Chinese)

[8] NEAL JOSHUA E S, BHATTACHARYYA D, CHAKKRAVARTHY M, et al. 3D CNN with visual insights for early detection of lung cancer using gradient-weighted class activation[J]. Journal of Healthcare Engineering, 2021, 2021: 6695518.

[9] TSIVGOULIS M, PAPASTERGIOU T, MEGALOOIKONOMOU V. An improved SqueezeNet model for the diagnosis of lung cancer in CT scans[J]. Machine Learning with Applications, 2022, 10: 100399.

[10] ZHANG H W, ZHANG W, WANG S S, et al. Deep 3D multi-scale dual path network for automatic lung nodule classification[J]. International Journal of Biomedical Engineering and Technology, 2022, 39(2): 149.

[11] ZHOU T, LIU F Z, YE X Y, et al. RNE-DSNet: A re-parameterization neighborhood enhancement-based dual-stream network for CT image recognition[J]. Engineering Science and Technology, an International Journal, 2024, 56: 101760.

[12] GUO Y X, SONG Q, JIANG M M, et al. Histological subtypes classification of lung cancers on CT images using 3D deep learning and radiomics[J]. Academic Radiology, 2021, 28(9): 258-266.

[13] FU Y, XUE P, ZHAO P, et al. 3D multi-resolution deep learning model for diagnosis of multiple pathological types on pulmonary nodules[J]. International Journal of Imaging Systems and Technology, 2022, 32(1): 74-87.

[14] KRONEMEIJER P S, GAVVES E, SONKE J J, et al. Tumor tracking in 4D CT images for adaptive radiotherapy[C]//Medical Imaging 2022: Image Processing. San Diego: SPIE, 2022: 54-61.

[15] NAEEM ABID M M, ZIA T, GHAFOOR M, et al. Multi-view convolutional recurrent neural networks for lung cancer nodule identification[J]. Neurocomputing, 2021, 453: 299-311.

[16] ZHOU T, NIU Y X, LU H L, et al. Vision transformer: To discover the "four secrets" of image patches[J]. Information Fusion, 2024, 105: 102248.

[17] NIU C, WANG G. Unsupervised contrastive learning based transformer for lung nodule detection[J]. Physics in Medicine and Biology, 2022, 67(20). DOI:10.1088/1361-10.1088/6560/ac92ba.

[18] MEHTA S, RASTEGARI M. MobileVIT: Light-weight, general-purpose, and mobile-friendly vision transformer[EB/OL]. (2022-03-04)[2024-07-09]. https://arxiv.org/abs/2110.02178v2.

[19] KUMAR A, FULHAM M, FENG D G, et al. Co-learning feature fusion maps from PET-CT images of lung cancer[J]. IEEE Transactions on Medical Imaging, 2020, 39(1): 204-217.

[20] SHI H Y, ZHANG N D, WU X Q, et al. Multimodal lung tumor image recognition algorithm based on integrated convolutional neural network[J]. Concurrency and Computation: Practice and Experience, 2020, 32(21): e4965.

[21] SCHWYZER M, FERRARO D A, MUEHLEMATTER U J, et al. Automated detection of lung cancer at ultralow dose PET/CT by deep neural networks-Initial results[J]. Lung Cancer, 2018, 126: 170-173.

[22] CHEN S, HAN X J, TIAN G W, et al. Using stacked deep learning models based on PET/CT images and clini-

cal data to predict EGFR mutations in lung cancer[J]. Frontiers in Medicine, 2022(9): 1041034.

[23] ZHAO X Y, WANG X, XIA W, et al. A cross-modal 3D deep learning for accurate lymph node metastasis prediction in clinical stage T1 lung adenocarcinoma[J]. Lung Cancer, 2020, 145: 10-17.

[24] BRADSHAW T, PERK T, CHEN S, et al. Deep learning for the detection of benign and malignant pulmonary nodules in non-screening chest CT scans[J]. Communications Medicine, 2018(1): 327.

[25] YU W H, LUO M, ZHOU P, et al. Metaformer is actually what you need for vision[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 10809-10819.

[26] MAAZ M, SHAKER A, CHOLAKKAL H, et al. Edgenext: Efficiently amalgamated cnn-transformer architecture for mobile vision applications[C]// In European Conference On Computer Vision. Israel: ECCV Workshops, 2023: 3-20.

[27] GUO J Y, HAN K, WU H, et al. CMT: Convolutional neural networks meet vision transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 12165-12175.

[28] LI J S, XIA X, LI W, et al. Next-VIT: Next generation vision transformer for efficient deployment in realistic industrial scenarios[EB/OL]. (2022-08-16)[2024-07-09]. https://arxiv.org/abs/2207.05501v4.

作者简介

周　涛　男,1977年出生于宁夏回族自治区吴忠市. 现为北方民族大学计算机科学与工程学院教授. 主要研究方向为医学图像分析处理、深度学习、模式识别.
E-mail: zhoutaonxmu@126.com.

牛玉霞　女,2000年出生于山西省长治市. 现为北方民族大学计算机科学与工程学院硕士研究生. 主要研究方向为智能医学图像分析处理.
E-mail: nyx2607133584@163.com.

叶鑫宇　男,1997年出生于湖北省天门市. 曾为北方民族大学计算机科学与工程学院硕士研究生. 主要研究方向为智能医学图像分析处理.
E-mail: 3303626778@qq.com.

刘　隆　男,2000年出生于安徽省安庆市. 现为北方民族大学计算机科学与工程学院硕士研究生. 主要研究方向为智能医学图像分析处理.
E-mail: liulong5254@163.com.

陆惠玲　女,1976年出生于河北省保定市. 现为宁夏医科大学医学信息与工程学院教授. 主要研究方向为医学图像分析处理、深度学习和模式识别等.
E-mail: lu_huiling@163.com