基于三维跨模态 ConvFormer 的肺部肿瘤识别

周涛[1,2], 叶鑫宇[1,2], 刘凤珍[1,2], 陆惠玲[3]

[1] (北方民族大学计算机科学与工程学院 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 银川 750021)
[3] (宁夏医科大学医学信息与工程学院 银川 750004)
(zhoutao@nmu.edu.cn)

摘 要: 针对三维医学影像因肺部肿瘤形状不规则、差异性大, 导致特征提取不充分和识别不准确的问题, 提出一种基于CNN和Transformer的三维跨模态肺部肿瘤识别模型3D-CConvFormer. 首先, 利用三分支网络学习三维PET, CT和 PET/CT 影像中病灶的特征; 其次, 设计全局特征与浅层局部特征融合的高效 ConvFormer 模块, 并利用自校正卷积对感受野进行有效扩展, 提高每个模态中对病灶信息的提取能力; 最后, 设计双分支不同分辨率的跨模态特征交互块, 利用 2 个全局注意力机制交叉学习不同模态、全局和局部信息, 交互式地增强跨模态特征提取能力. 实验采用的肺部肿瘤 3D 多模态数据集, 该数据集共有 3 173 例患者, 3D-CConvFormer 模型在参数量和运行时间较优的前提下, 获得了 89.25%的准确率和 88.74%的 AUC 值的最优性能, 为三维多模态肺部肿瘤疾病诊断提供可靠的计算机辅助.

关键词: 肺部肿瘤; ConvFormer; 跨模态特征交互; 三维 PET/CT 多模态影像
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024.20153

3D Cross-Modal ConvFormer for Lung Cancer Recognition

Zhou Tao[1,2], Ye Xinyu[1,2], Liu Fengzhen[1,2], and Lu Huiling[3]

[1] (*School of Computer Science and Engineering, Northern Minzu University, Yinchuan* 750021)
[2] (*The Key Laboratory of Image and Graphics Intelligent Processing of North Minzu University, Yinchuan* 750021)
[3] (*School of Medical Information and Engineering, Ningxia Medical University, Yinchuan* 750004)

Abstract: Due to the irregular shape and large difference of lung tumors in 3D medical images, the feature extraction of lesions is insufficient, and the recognition accuracy is not high, a 3D Cross-Modal ConvFormer is proposed. Firstly, three networks are utilized to learn the 3D PET, CT and PET/CT medical images. Secondly, a ConvFormer model is designed to fuse global and shallow local features, while self-correcting convolution expands the receptive field for better lesion extraction. Finally, a dual-branch cross-modal feature interaction block is designed to enhance cross-modal features and capture 3D multimodal details. This module uses two global attention mechanisms to improve the extraction of cross-modal and global-local information. The experiments use a 3D multimodal lung tumor dataset with 3173 patients. With optimized parameters and computation time, the 3D-CConvFormer achieves an accuracy of 89.25% and an AUC of 88.74%, providing reliable computer-aided diagnosis for 3D multimodal lung tumor.

Key words: lung cancer; ConvFormer; cross-modal feature interaction; 3D PET/CT multimodal images

收稿日期: 2023-02-22; 修回日期: 2023-10-18. **基金项目**: 国家自然科学基金(62062003); 宁夏自然科学基金(2022AAC03149, 2023AAC03293). **周涛**(1977—), 男, 教授, 博士生导师, 主要研究方向为计算机辅助诊断、模式识别; **叶鑫宇**(1999—), 男, 硕士研究生, 主要研究方向为医学图像处理、计算机辅助诊断; **刘凤珍**(1998—), 女, 硕士研究生, 主要研究方向为医学图像处理、计算机辅助诊断; **陆惠玲**(1976—), 女, 硕士, 副教授, 主要研究方向为图像图形智能处理.

癌症是全球第二大死亡原因, 其中肺部肿瘤在 2020 年造成了 221 万例死亡①. 肺部肿瘤通过早期检测有较高的治愈机会, 可以采用手术、放疗和化疗等降低死亡风险. 许多成像技术被用于肺部肿瘤的诊断和识别, 如计算机断层扫描(computed tomography, CT)、正电子发射断层扫描(positron emission computed tomography, PET)、磁共振成像和 X 光片等. 肺部肿瘤计算机辅助诊断深度学习模型[1]可辅助医生进行快速、准确的诊断. Zhang 等[2]利用连体卷积神经网络学习 CT 像素块之间内容距离, 多数投票规则识别肺部肿瘤. Hedner 等[3]设计低剂量 CT 肺部肿瘤识别网络, 在 1 179 名患者中识别到 68.00%良性和 97.00%恶性. 而实际肺部肿瘤临床影像往往为三维的肺部扫描切片影像.

基于三维肺部肿瘤影像的深度学习网络展现了其优越性. 肺部肿瘤尺寸小易导致类似结构被识别为肿瘤, Dutande 等[4]提出二维三维级联的策略, 对肿瘤体积立方体进行分类识别, 获得了较高精度. 由于肿瘤数据集规模较小且类别较少, Mei 等[5]收集了一份 9 个类别的肺结节数据集, 提出切片感知网络, 可捕获一个切片组的任何位置和任何通道之间的远程依赖关系, 有效地降低了假阳性率. 针对医学领域仅有少量注释图像可用的问题, Amyar 等[6]提出编码器对多个任务提取特征, 多任务多尺度框架可从肿瘤内部和周围区域学习丰富特征, 其在三维肺部肿瘤 PET 中患者治疗反应和生存率的预测结果优于单任务学习. 为了解决肿瘤个体差异大以及肿瘤与软组织视觉相似的问题, Wu 等[7]利用二维卷积神经网络(convolutional neural networks, CNN)提取矢状面、冠状面和轴状面的特征, 与三维 CNN 提取肿瘤的体积特征按概率进行加权聚合, 在肺结节分析 2016 (lung nodule analysis 2016, LUNA16)数据集上获得 90.08%的识别准确率. Fu 等[8]过滤掉不相关的切片并利用注意力模块学习三维 CT 肺部肿瘤图像, 获得了较高的识别准确率. Niu 等[9]转换三维 CT 图像为向量特征, 将其输入到 Transformer 中获得了较好肺部肿瘤识别性能. 三维 CNN 难以捕获到全局特征; 三维 Transformer 在所有特征上建立远程依赖关系, 所需参数量和计算量较大; 在肺部肿瘤识别任务中, 三维 CNN 和 Transformer 不仅需要采用合理结合方式, 而且还需要充分地学习多模态优势信息.

为了充分地学习三维多模态 PET/CT 图像中肿瘤的代谢和解剖信息, 许多深度学习模型被提出. Kao 等[10]指出, PET 检测肿瘤代谢活性, CT 检测人体解剖结构, PET/CT 组合可更好地识别肺部肿瘤. PET 和 CT 分别提取特征后进行集成, 尽管能够获得高质量的肿瘤识别结果, 但忽略了 PET/CT 集成的作用. Ming 等[11]通过融合 CT 和 PET 图像获得解剖和代谢信息, 较单模态识别准确率提升 6.00%, 同时提高了临床诊断效率. 肺部肿瘤检测 PET 中的高异常摄取比心脏生理摄取更重要, 为了考虑不同空间位置特征具有不同优先级和跨模态信息互补, Kumar 等[12]利用 CNN 获得三维 PET 与 CT 空间变化的融合图, 通过量化不同空间位置上每种模态特征的重要性, 将融合图与模态特定特征图相乘, 以获得多模态互补表示. 尽管获得了较优识别准确率, 但成像机理不同的图像之间存在很多不一致的信息, 直接相乘难以充分地突出多模态优势信息. 为了从不同成像模态中可靠地提取细粒度特征, Qin 等[13]利用空间注意力门控和通道注意力增强的三维密集网络, 有效地提取出 PET 和 CT 图像的细粒度特征, 可实现肺癌的无创诊断, 局部空间和通道注意力能够增强肺部肿瘤特征, 但忽略了肿瘤的全局信息. Zhao 等[14]提出跨模态三维网络预测肺部肿瘤, 并学习三维肿瘤块、临床和病理标签 3 种特征, 在 401 个病例中取得了 92.60%的曲线下面积(area under curve, AUC)值.

三维 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能够获得更准确的病理信息. 实际的多模态三维影像信息更丰富和复杂, 且肺部肿瘤具有形状不规则、差异性大等特点, 导致特征提取不充分和识别不准确的问题. 针对上述问题, 本文提出一种用于肺部肿瘤识别的三维跨模态 ConvFormer 模型——3D-CConvFormer 模型. 其通过 3 个分支分别学习三维 PET, CT 和 PET/CT 影像, 采用 CNN 与 Transformer 高效结合的特征提取模块 ConvFormer. 其中, 自校正注意力有效扩展了 CNN 感受野; 浅层全局特征被传递到深层进行全局和局部特征融合, 充分地学习三维多模态影像并提高每个模态中对病灶信息的提取能力; 设计的双分支跨模态特征交互模块, 利用 2 个全局注意力交换不同分辨率特定特征图, 交叉学习不同模态、全局和局部信息, 以实现全局与局部特征中肿瘤特征的有效增强, 充分地学习三维多模态图像中肺部肿瘤的代谢和解剖信息.

① https://www.who.int/news-room/fact-sheets/detail/cancer

1 本文方法

3D-CConvFormer 模型结构如图 1 所示. 3 分支网络分 4 个阶段对 PET, CT 和 PET/CT 提取信息, 其层数设计参考 NextVit-S[15], ConvFormer 学习多模态全局和局部信息, 跨模态特征交互模块增强病灶特征, 最后经分类层进行肺部肿瘤识别.

1.1 ConvFormer

CNN 通过局部相邻像素的联系提取局部特征, 但难以捕获全局特征. Transformer 对特征建立远程依赖关系, 但对局部细节关注不足. 尽管在足够大的数据集中, Transformer 对局部信息关注问题能得到缓解, 但临床肺部肿瘤数据集规模有限而导致其性能受限; 此外, 医学影像的分辨率较高也会导致模型参数量和计算量急剧上升. 三维模型比二维网络架构资源消耗更多, 为此, 本文在三维 PET/CT 多模态影像上设计 CNN 与 Transformer 高效结合的 ConvFormer 模块, 结构如图 2 所示. 其中, 浅层全局特征被传递到深层, 并融合全局特征与自校正卷积后的局部特征, 以充分地学习三维多模态影像的全局和局部特征, 进而提高每个模态中对病灶信息的提取能力.

如图 2 所示, CNN 自校正卷积利用卷积权值共享和卷积核间通信捕获关键特征, 低分辨率校准有效地扩展感受野, 避免复杂参数引入和烦琐的参数调整, 利用较低分辨率特征图 $\boldsymbol{X}_1$ 生成注意力权值, 指导原始空间中的特征学习, 使每个像素自适应地学习远距离上下文语义信息, 从而增强特征辨识性. 首先对 $\boldsymbol{X}_1$ 使用步长为 4 的 3×3×3 三维平均池化(Pool); 其次利用三维卷积($\mathrm{Conv}_{3\times3\times3}$)学习后进行三线性插值的上采样(Up), 输出特征图与 $\boldsymbol{X}_1$ 进行残差相加; 然后采用归一化指数函数(Softmax)

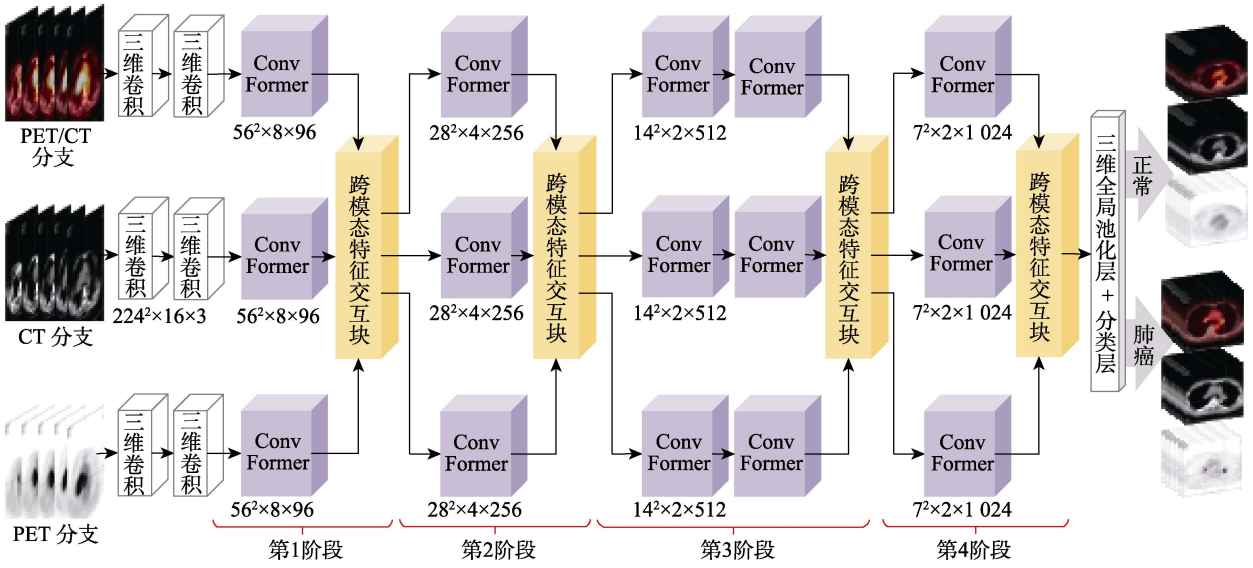


图 1 3D-CConvFormer 模型结构图

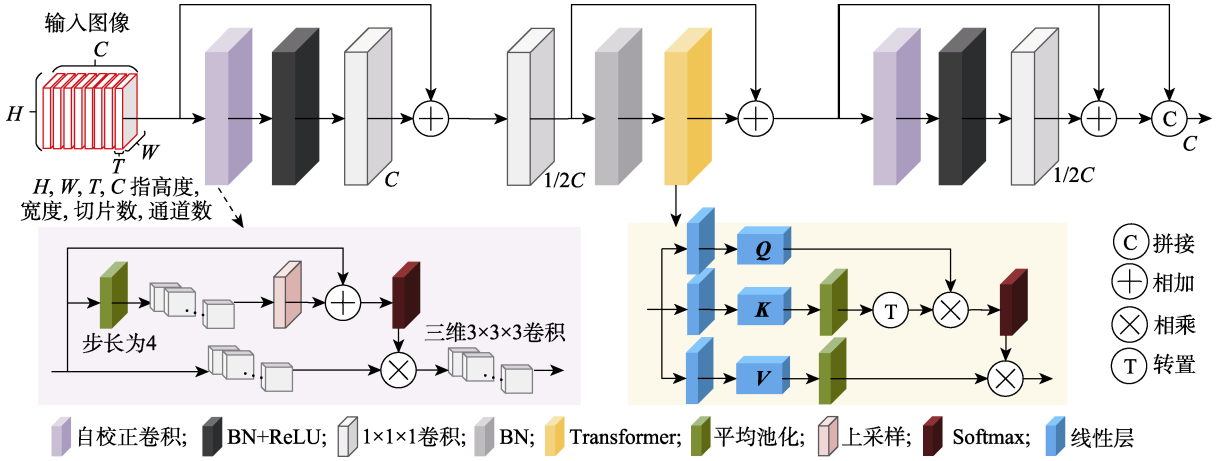


图 2 ConvFormer 的结构图

获得注意力权值W，加权到每个像素；最后再通过三维卷积得到$\boldsymbol{X}_2$，其中，

$$W=\mathrm{Softmax}\left(\mathrm{Up}\left(\mathrm{Conv}_{3\times3\times3}\left(\mathrm{Pool}(\boldsymbol{X}_1)\right)\right)+\boldsymbol{X}_1\right) \quad (1)$$

Transformer 部分利用线性层将特征图$\boldsymbol{X}_2$变换为$\boldsymbol{Q}$, $\boldsymbol{K}$和$\boldsymbol{V}$，利用平均池化压缩$\boldsymbol{K}$和$\boldsymbol{V}$的空间尺寸，以降低计算成本；然后利用缩放点积[9]计算全局注意图并应用于$\boldsymbol{V}$，输出特征图为

$$\boldsymbol{X}_3=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\times\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\times\boldsymbol{V} \quad (2)$$

其中，d_{k}是指$\boldsymbol{K}$的维度．批量归一化(BN)和残差能对全局特征更高效学习．输出特征图$\boldsymbol{X}_3$再次经过 CNN 部分得到特征图$\boldsymbol{X}_4$，对全局特征$\boldsymbol{X}_3$进行拼接复用，使模型在提高性能的同时尽可能地提高效率．ConvFormer 并行学习局部特征和全局特征，以较少的资源消耗学习更多可区分的特征，以提高语义判别能力，并缓解类别混淆，更好地捕获多模态肿瘤特征．

1.2　跨模态特征交互模块

成像机理不同的多模态图像之间存在很多不一致的信息，在肺部肿瘤识别中使用不合理的特征融合会导致准确率较低．此外，尽管通过自校正卷积和全局注意力机制对三模态特征学习和校准，但难以精准识别包含肿瘤信息的全局和局部特征．为此，本文设计双分支的跨模态特征交互模块，在不同分辨率内学习和交互，实现对全局与局部特征中病灶信息的增强，结构如图 3 所示．

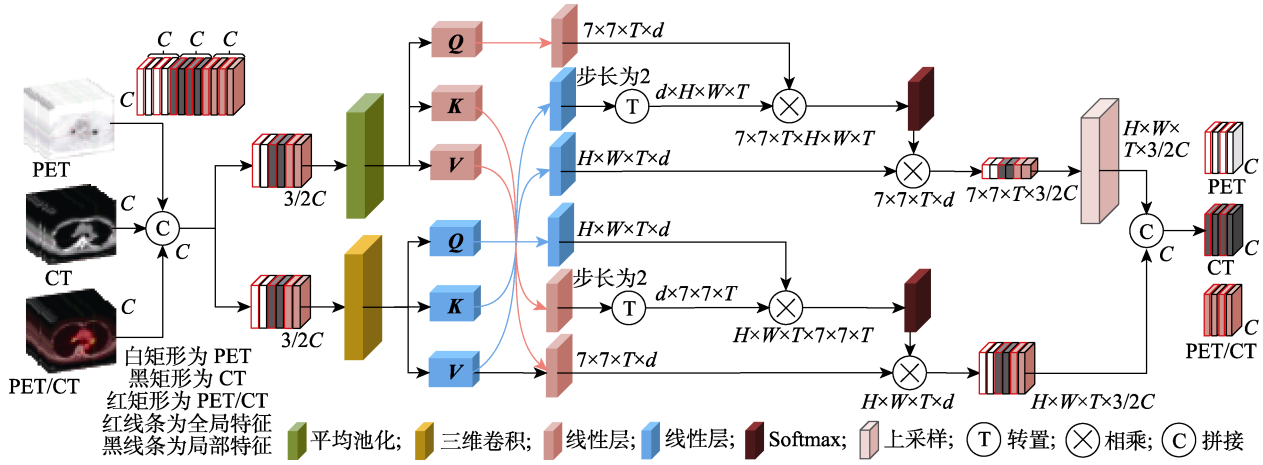


图 3　跨模态特征交互模块的结构图

如图 3 所示，按全局和局部均匀切分的一部分三模态特征图，通过三维平均池化将空间维度压缩到最小尺寸；另一部分经过三维卷积，其中低分辨率特征图变换为$\boldsymbol{Q}_{\mathrm{L}}$, $\boldsymbol{K}_{\mathrm{L}}$和$\boldsymbol{V}_{\mathrm{L}}$，高分辨率特征图变换为$\boldsymbol{Q}_{\mathrm{H}}$, $\boldsymbol{K}_{\mathrm{H}}$和$\boldsymbol{V}_{\mathrm{H}}$．全局特征关注区域更广和局部特征学习区域更集中，为避免直接拼接导致特征不完全对齐，通过交换不同分辨率特征生成的$\boldsymbol{K}_{\mathrm{L}}$与$\boldsymbol{V}_{\mathrm{L}}$以及$\boldsymbol{K}_{\mathrm{H}}$和$\boldsymbol{V}_{\mathrm{H}}$，然后进行全局计算，不同分辨率特征交互使得全局与局部特征对齐，实现病灶信息的增强，低分辨特征为

$$\boldsymbol{X}_{\mathrm{L}}=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{L}}\times\boldsymbol{K}_{\mathrm{H}}^{\mathrm{T}}}{\sqrt{d_{\mathrm{H}}}}\right)\times\boldsymbol{V}_{\mathrm{H}} \quad (3)$$

其中，d_{H}是指$\boldsymbol{K}_{\mathrm{H}}$的维度，高分辨特征为

$$\boldsymbol{X}_{\mathrm{H}}=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{H}}\times\boldsymbol{K}_{\mathrm{L}}^{\mathrm{T}}}{\sqrt{d_{\mathrm{L}}}}\right)\times\boldsymbol{V}_{\mathrm{L}} \quad (4)$$

其中，d_{L}是指$\boldsymbol{K}_{\mathrm{L}}$的维度．3 种模态生成的高分辨特征与低分辨特征之间进行交互，其中交叉的全局计算使得特征间信息传递是可学习的和动态的．

跨模态特征交互块利用低分辨特征充分地学习位置和语义这类深层信息，双路径并行处理不同分辨率以充分提取全局与局部优势特征．不同分辨率交互可建模复杂的跨尺度关系，以充分地利用多模态图像的语义相关性，自适应融合多模态特征中的全局和局部信息，融合 PET 中肿瘤的代谢信息和 CT 中肿瘤的解剖信息，进而提高模型分类性能和增强模型对肺部肿瘤病灶的聚焦能力．

2　实验和讨论

2.1　PET/CT 多模态三维数据集与评价指标

本文实验数据集选用从宁夏某三甲医院 2014 年 1 月~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[13]，每例大约有 102~289 张肺部切片.

数据集按 6∶2∶2 比例分成训练集、验证集和测试集，实验环境搭载 2 块 TITAN Volta 显卡，采用 Pytorch 搭建网络，自适应矩估计权重(adaptive moment estimation weight, AdamW)优化器进行优化和每 10 个周期 0.9 的衰减策略，设置初始学习率为 0.01，训练周期为 300，训练批处理大小为 8. 如图 4 所示, CT 中肿瘤和正常组织密度差异很难区分, PET 中肿瘤区域代谢旺盛呈高亮，因此多模态肺部肿瘤图像可以更好地识别和定位病灶.

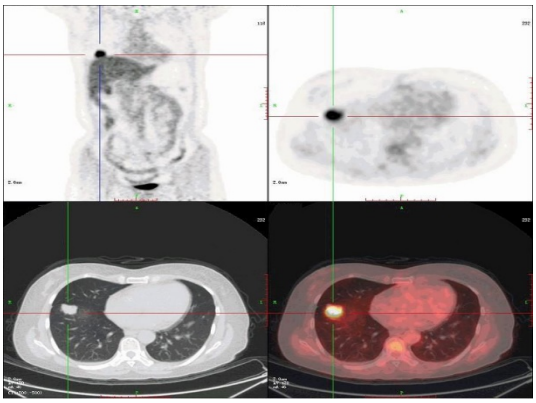


图 4 已配准的 PET, CT 和 PET/CT 图像

模型预测结果可划分为真正类 TP、假正类 FP、假负类 FN、真负类 TN. 准确率 A 为全部类预测正确的比例，精确率 P 为正类预测正确占所有正类的比例，召回率 R 为预测正类占所有正类的比例，F_1 分数表示为

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{5}$$

ROC 曲线(receiver operating characteristic curve)是以真正类率即召回率(true positive rate, TPR)为纵轴、假正类率(false positive rate, FPR)为横轴进行绘制. FPR 表示为

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{6}$$

ROC 越靠近左上角，AUC 值越大，表示模型排序与分类性能越好；其他指标值越大，表示模型越佳.

2.2 不同模态识别实验

通过不同模态相加测试多模态特征融合效果，在 ConvFormer 模型基础上进行 7 组实验，结果如表 1 所示. 实验 1~3 输入单模态图像，实验 4~6 输入 2 种多模态图像，实验 7 输入 3 种多模态图像.

表 1 不同模态识别的具体结果 %

实验	模态	A	AUC	R	F_1
1	CT	85.51	84.79	81.23	81.71
2	PET	85.98	85.28	81.82	82.30
3	PET/CT	86.33	85.67	82.40	82.77
4	CT+PET	86.57	85.71	81.52	82.86
5	CT+PET/CT	86.92	86.40	83.87	83.63
6	PET+PET/CT	87.27	86.69	83.87	83.99
7	PET+CT+PET/CT	87.73	87.23	84.75	84.63

实验 1~实验 3 中各项指标较低，因肺部肿瘤病灶错综复杂且与正常组织相连，仅使用单模态难以准确地识别. 实验 4 融合 CT 和 PET，较 PET/CT 单模态大部分指标均有小幅提升. 实验 5 和实验 6 将 PET/CT 分别与 CT 和 PET 融合，各项指标均获得提升，可以看出，学习多模态图像可以更好地捕获和识别肺部肿瘤. 实验 7 利用 3 种模态的优势互补，增强模型对病灶的聚焦能力，获得最优性能，A, AUC, R 和 F_1 这 4 项指标较 PET/CT 单模态分别提升了 1.62%, 1.82%, 2.85%和 2.24%.

2.3 消融实验

为评估各模块有效性，以 3D-Transformer 为基础模型，依次添加各模块，5 组实验结果如表 2 所示，实验 1、实验 3 和实验 5 中 5 个患者的热力图如图 5 所示. 第 1 行为不同模态的三维肺部肿瘤图像，第 5 行为已配准的三模态三维图像，红线与绿线交叉区域是标注出的病灶主要区域.

表 2 消融实验结果对比

模型	实验	替换添加	参数量	A/%	AUC/%	R/%	F_1/%	P/%	训练时间/s
3D-Transformer	1	3D-Transformer	13.738M	82.59	82.02	79.18	78.37	77.59	33 667
ConvFormer	2	ConvFormer	6.784M	84.81	84.06	80.35	80.83	81.31	28 236
	3	自校准卷积	9.422M	86.33	85.67	82.40	82.77	83.14	29 894
CConvFormer	4	三模态图像	9.422M	87.73	87.23	84.75	84.63	84.50	31 047
	5	跨模态特征交互模块	11.199M	89.25	88.74	86.22	86.47	86.73	32 153

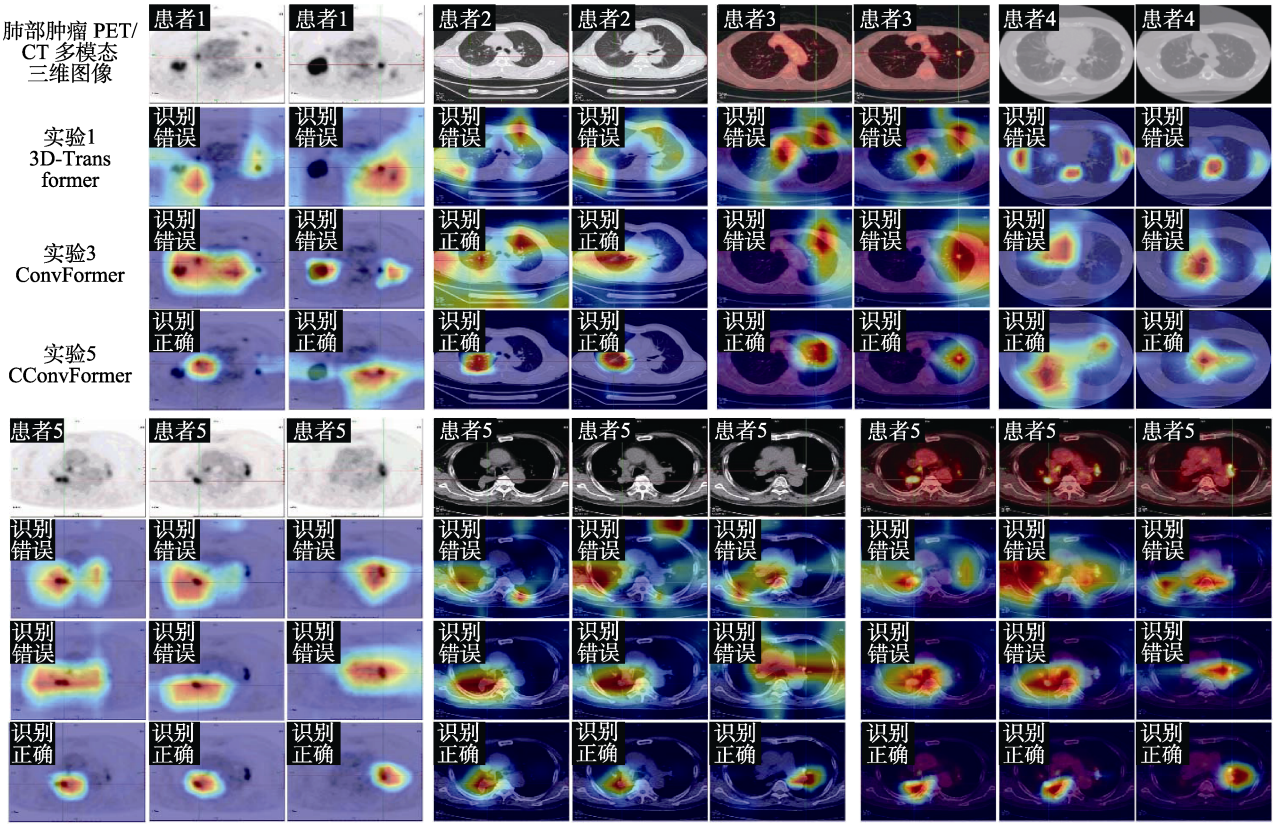


图 5　消融模型在肺部肿瘤影像上的热力图

虽然实验 1 第 2 行和第 6 行热力图关注区域较大，但对病灶区域的关注存在较大误差且关注到非肺部区域. 实验 2 的参数量减少 50.62%且训练时间缩短 16.13%，准确率提升 2.69%，ConvFormer 可提升性能和保持较高计算效率. 实验 3 准确率提升 1.79%，自校正注意力帮助模型学习更多可区分的肺部肿瘤特征，从图 5 第 3 行和第 7 行热力图看出，模型能较好关注肺部肿瘤区域，但对小病灶或与器官特征相近的病灶区分较难，如第 3 行第 7 列和第 7 行的关注区域存在明显误差. 实验 4 使用多模态数据后准确率提升 1.62%，多模态语义信息了增强模型对病灶的聚焦能力. 实验 5 准确率提升 1.73%，跨模态特征交互块通过不同分辨率特征交互，有效增强全局与局部肿瘤特征，从图 5 第 4 行和第 8 行的热力图看出，模型充分地学习了三维多模态图像中肿瘤的代谢和解剖信息，利用多模态特征互补，更精准地定位和识别病灶，有效地提高模型识别肺部肿瘤的能力.

2.4　对比实验

本文在肺部肿瘤 PET/CT 多模态三维数据集上，将 3D-CConvFormer 与 3 个 CNN 模型、3 个 Transformer 模型和 4 个 CNN 结合 Transforme 模型进行对比，结果如表 3 所示，性能较好的部分模型热力图如图 6 所示. 本文模型参数量仅低于 3D-EfficientNet-b3，训练时间最短；性能指标表现全面优于其他模型，表明其肺部肿瘤识别能力强且计算效率高.

与 3 个 CNN 模型相比，本文模型的训练时间较 3D-EfficientNet-b3 缩短了 6.18%，准确率较 3D-ConvNeXt-S 提升 5.24%. 与 3 个 Transformer 模型相比，本文模型计算效率大幅提高同时获得了更高的性能，训练时间较 3D-PoolFormer-S24 缩短 14.22%，5 项指标提升近 3.87%；较用于视频识别的 3D-BEVT 模型，5 项指标分别提升 3.67%，3.64%，3.53%，4.58%和 5.67%. 与 4 个 CNN 结合 Transformer 模型相比，本文模型在训练时间和肺部肿瘤识别能力上均具优势：训练时间较 3D-CVT-13 缩短 11.72%，性能提升 5.63%；较全局特征中引入卷积提取细粒度特征的 3D-CMT-S 训练时间缩短 11.57%，识别准确率提升 3.94%；较工业部署的 3D-NextVit-S 训练时间缩短 8.52%，5 项指标分别提高了 2.68%，2.95%，4.27%，3.64%和 3.03%. 从热力图可见，3D-CConvFormer 识别肺部肿瘤更精准，且具更高计算效率和鲁棒性.

图 7a 展示了 11 种模型的 ROC 曲线和 AUC 值，本文模型的曲线整体位于左上角，能有效学习

表 3 对比模型在肺部肿瘤 PET/CT 多模态三维数据集上的具体结果

对比模型	参数量	准确率/%	AUC/%	召回率/%	F_1/%	精确率/%	训练时间/s
3D-ResNet50[1]	46.203M	81.07	79.47	71.55	75.08	78.96	36 256
3D-EfficientNet-b3[7]	**1.624M**	82.59	80.53	70.38	76.31	83.33	34 271
3D-ConvNeXt-S[8]	54.954M	84.81	83.71	78.29	80.42	82.66	37 104
3D-SwinTransformer-S[16]	48.751M	83.64	83.09	80.35	72.25	76.64	38 589
3D-PoolFormer-S24[15]	31.687M	85.75	85.43	83.87	82.42	81.02	37 486
3D-BEVT[16]	86.628M	86.09	85.62	83.28	82.68	82.08	41 718
3D-CoaT-S[9]	24.385M	83.53	82.69	78.59	79.17	79.76	42 153
3D-CVT-13[17]	21.430M	84.69	84.01	80.65	80.76	80.88	36 422
3D-CMT-S[12]	27.278M	86.09	85.38	81.82	82.42	83.04	35 874
3D-NextVit-S[15]	19.900M	86.92	86.20	82.69	83.43	84.18	34 891
3D-CConvFormer	11.199M	**89.25**	**88.74**	**86.22**	**86.47**	**86.73**	**32 153**

注. 粗体表示最优值.

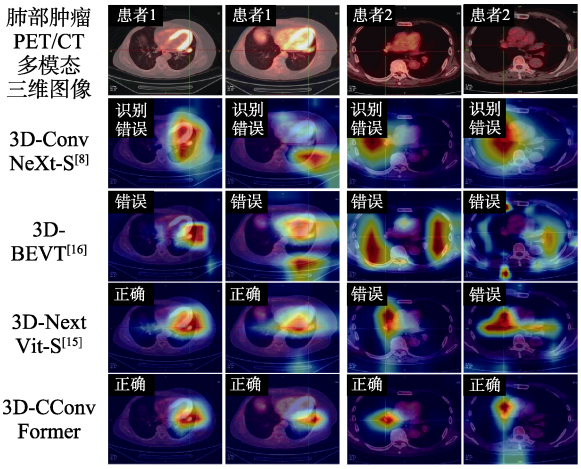
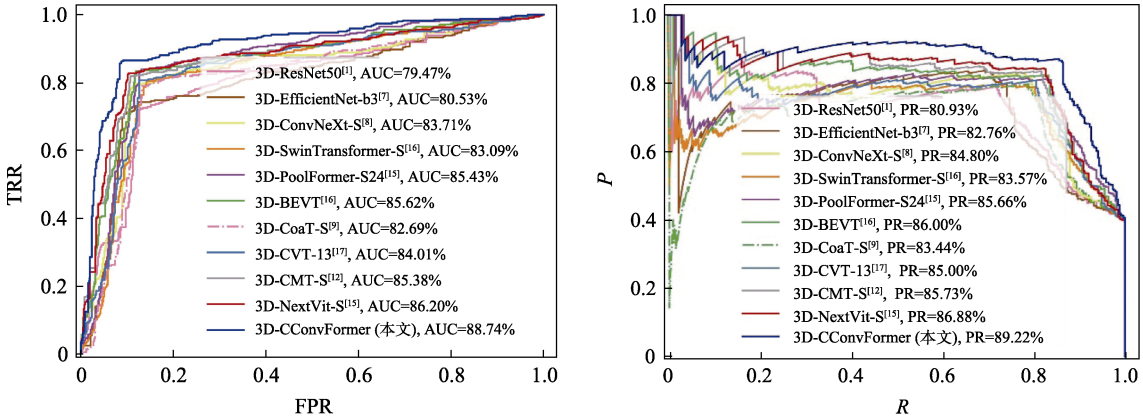


图 6 对比模型在肺部肿瘤影像上的热力图

和识别肺部肿瘤的全局与局部病灶信息. 图 7b 展示了 11 种模型的 PR(precision-recall)曲线, 本文模型的 PR 曲线下面积最大, 性能最优.

2.5 公共数据集对比实验

为了进一步验证 3D-CConvFormer 对三维肺部肿瘤识别的鲁棒性和泛化能力, 使用包括 888 个低剂量肺 CT 图像的公开三维肺部数据集 LUNA16[7], 其中 118 个肺部肿瘤标签由 4 位专家标注. 实验采用十折交叉验证, 结果如表 4 所示, AUC 曲线如图 8 所示, 可以看出, 本文模型的各项评价指标最高、AUC 曲线覆盖范围最大, 表明其对肺部肿瘤识别能力较强.



a. ROC曲线 b. PR曲线

图 7 11 种模型在肺部肿瘤 PET/CT 多模态三维数据集上的曲线对比

表 4 不同模型在 LUNA16 数据集上的具体结果

对比模型	A/%	AUC/%	R/%	F_1/%
3D-ConvNeXt-S[8]	94.32	84.43	70.83	77.27
3D-BEVT[16]	95.45	86.84	75.00	81.82
3D-NextVit-S[15]	97.16	93.09	87.50	89.36
3D-CConvFormer	**98.29**	**99.01**	**100.00**	**94.12**

注. 粗体表示最优值.

3 结 语

本文提出一种局部和全局特征同时提取的三维跨模态肺部肿瘤识别模型, ConvFormer 捕获各模态病灶信息, 跨模态特征交互模块通过不同分辨率特征的交互学习, 增强多模态肺部肿瘤特征.

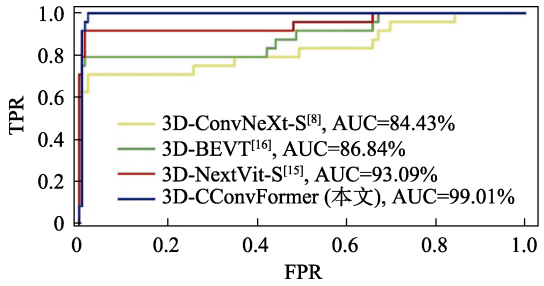


图 8　不同模型在 LUNA16 数据集上的 ROC 曲线

在 PET/CT 多模态三维数据集上的热力图和结果表明，本文模型以较高计算效率获得 89.25%的准确率和 88.74%的 AUC 值，充分地利用多模态的优势信息提升模型识别能力；在公开数据集 LUNA16 上，本文模型获得了较高准确率和良好鲁棒性.

在临床三维肺部影像上本文模型表现出强鲁棒性与泛化能力，具备较高准确率和较快计算速度，显著优于专家人工阅读切片的识别效率，模型有望辅助医生精准识别肺部肿瘤，其热力图与预测结果提高了解释性与临床可靠性. 受临床三维多模态数据配准困难和可训练样本限制，本文未能开展大规模临床验证，未来将在三维跨模态肺部肿瘤识别方面与医院进行更深入的合作和研究.

参考文献(References):

[1] Zhou Tao, Liu Yuncan, Lu Huiling, *et al*. ResNet and its application to medical image processing: research progress and challenges[J]. Journal of Electronics & Information Technology, 2022, 44(1): 149-167(in Chinese)
(周涛, 刘赟璨, 陆惠玲, 等. ResNet 及其在医学图像处理领域的应用：研究进展与挑战[J]. 电子与信息学报, 2022, 44(1): 149-167)

[2] Zhang K, Qi S L, Cai J M, *et al*. Content-based image retrieval with a convolutional siamese neural network: distinguishing lung cancer and tuberculosis in CT images[J]. Computers in Biology and Medicine, 2022, 140: 105096

[3] Hedner J A, Stenlof K, Ding Z, *et al*. Safety and efficacy of sulthiame in moderate to severe obstructive sleep apnea: a randomized placebo-controlled parallel-group trial[J]. European Respiratory Journal, 2021, 58: Article No.OA4383

[4] Dutande P, Baid U, Talbar S. LNCDS: A2D- 3D cascaded CNN approach for lung nodule classification, detection and segmentation[J]. Biomedical Signal Processing and Control, 2021, 67: Article No.102527

[5] Mei J, Cheng M M, Xu G, *et al*. SANet: a slice-aware network for pulmonary nodule detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(8): 4374-4387

[6] Amyar A, Modzelewski R, Vera P, *et al*. Multi-task multi-scale learning for outcome prediction in 3D PET images[J]. Computers in Biology and Medicine, 2022, 151: Article No.106208

[7] Wu Z, Ge R J, Shi G L, *et al*. MD-NDNet: a multi-dimensional convolutional neural network for false-positive reduction in pulmonary nodule detection[J]. Physics in Medicine & Biology, 2020, 65(23): Article No.235053

[8] Fu X H, Bi L, Kumar A, *et al*. An attention-enhanced cross-task network to analyse lung nodule attributes in CT images[J]. Pattern Recognition, 2022, 126: Article No.108576

[9] Niu C, Wang G. Unsupervised contrastive learning based transformer for lung nodule detection[J]. Physics in Medicine & Biology, 2022, 67(20): Article No.204001

[10] Kao Y S, Yang J. Deep learning-based auto-segmentation of lung tumor PET/CT scans: a systematic review[J]. Clinical and Translational Imaging, 2022, 10(2): 217-223

[11] Ming Y, Dong X Y, Zhao J H, *et al*. Deep learning-based multimodal image analysis for cervical cancer detection[J]. Methods, 2022, 205: 46-52

[12] 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

[13] Qin R X, Wang Z Z, Jiang L Y, *et al*. Fine-grained lung cancer classification from PET and CT images based on multidimensional attention mechanism[J]. Complexity, 2020, 2020: Article No.6153657

[14] 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

[15] Li J S, Xia X, Li W, *et al*. Next-ViT: next generation vision transformer for efficient deployment in realistic industrial scenarios[OL]. [2023-02-22]. https://arxiv.org/abs/2207.05501

[16] Wang R, Chen D D, Wu Z X, *et al*. BEVT: BERT pretraining of video transformers[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 14713-14723

[17] Wu H P, Xiao B, Codella N, *et al*. CvT: introducing convolutions to vision transformers[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 22-31