

同济大学第十届“卓越杯”暨第十九届“挑战杯”
全国大学生课外学术科技作品竞赛校内选拔赛
作品申报书

作品名称：大模型辅助的前列腺超声影像癌症诊断系统

学院全称：计算机科学与技术学院

申报者（申报者代表）姓名：朱俊泽

类别：

- ☐ 自然科学类学术论文
- ☐ 哲学社会科学类社会调查报告和学术论文
- ☐ 科技发明制作 A 类
- ☒ 科技发明制作 B 类

说 明

1. 申报者应在认真阅读此说明各项内容后按要求详细填写。
2. 申报者在填写申报作品情况时，个人项目请填写 A1—1 和 A2 表，集体项目请填写 A1—2 和 A2 表。作品类别（自然科学类学术论文、哲学社会科学类社会调查报告、科技发明制作）分别填写 B1、B2 或 B3 表。所有申报者可根据情况填写 C 表。
3. 表内项目填写时一律用钢笔或打印，字迹要端正、清楚，此申报书可复制。
4. 学术论文、社会调查报告及所附的有关材料必须是中文（若是外文，请附中文本），请以 4 号楷体打印在 A4 纸上，附于申报书后，学术论文字数在 8000 字以内，社会调查报告字数在 15000 字以内（文章版面尺寸 14.5×22cm）。作品正文格式可参见附件 2：《同济大学“卓越杯”大学生课外学术科技作品竞赛论文模板》。申报书和作品正文按照学位论文的要求一起装订成册。
5. 申报书交至各学院团委，各学院收齐后交至校团委。

联系人：施老师、刘老师、许老师、臧同学

电话：65981406、65982406、18222009008

Email: tongji_tzb@163.com

A1-1. 申报者情况（个人项目）

说明：必须由申报者本人按要求填写，申报者情况栏内必须填写个人作品的第一作者（承担申报作品 60%以上的工作者），合作者必须是学生且不得超过 2 人；凡有合作者的个人作品须按学历最高的作者划分本科生作品或硕士研究生作品。

申报者情况	姓名	朱俊泽		性别	男
	出生年月	2005.7.31		学校	同济大学
	学院	计算机科学与技术学院		专业	数据科学与大数据技术
	年级	大二		学制	四年
	当前学历	高中		入学年月	2023.9
	身份证号	450202200507310010			
	作品名称	大模型辅助的前列腺超声影像癌症诊断系统			
	毕业论文题目	无			
	手机号码	18007721992		电子邮箱	2715978522@qq.com
合作者情况	姓名	性别	当前学历	学校	身份证号
	王雪宸	男	高中	同济大学	310112200411057811
	高晗博	男	高中	同济大学	140106200603220178
资格认定	院系负责人或导师意见	本作品是否为课外学术科技或社会实践活动成果。 <input checked="" type="checkbox"/> 是 <input type="checkbox"/> 否 负责人签名：倪张凯 2025 年 2 月 24 日			

A2. 指导教师情况

说明：1. 必须由申报者本人和指导教师充分沟通后填写；
2. 本表中的学院审核意见视为对指导教师指导情况的确认。

第一位指导教师情况	姓名	倪张凯	性别	男
	工作单位	同济大学	院系/部门	计算机科学与技术学院
	职务		职称	副教授
	研究方向	计算机视觉，图像处理，人工智能，生物医学影像		
	手机号码	13646019080	电子邮箱	zkni@tongji.edu.cn
第二位指导教师情况	姓名		性别	
	工作单位		院系/部门	
	职务		职称	
	研究方向			
	手机号码		电子邮箱	
第三位指导教师情况	姓名		性别	
	工作单位		院系/部门	
	职务		职称	
	研究方向			
	手机号码		电子邮箱	
学院审核意见		<div>(盖章) 年 月 日</div>		

B3. 申报作品情况（科技发明制作）

- 说明：1. 必须由申报者本人填写；
2. 本表必须附有研究报告，并提供图表、曲线、试验数据、原理构图、外观图（照片），也可附鉴定证书和应用证书；
3. 组别请按照作品发明点或创新点所在类别选择一项勾选；
4. 学科领域按照作品所选组别中的相应类别选择一项勾选。

作品全称	大模型辅助的前列腺超声影像癌症诊断系统
组别	<input type="radio"/> 机械与控制 <input checked="" type="checkbox"/> 信息技术 <input type="radio"/> 数理 <input type="radio"/> 生命科学 <input type="radio"/> 能源化工
学科领域	机械与控制： <input type="radio"/> 机械 <input type="radio"/> 仪器仪表 <input type="radio"/> 自动化控制 <input type="radio"/> 工程 <input type="radio"/> 交通 <input type="radio"/> 建筑 信息技术： <input checked="" type="checkbox"/> 计算机 <input type="radio"/> 电信 <input type="radio"/> 通讯 <input type="radio"/> 电子 数理： <input type="radio"/> 数学 <input type="radio"/> 物理 <input type="radio"/> 地球与空间科学 生命科学： <input type="radio"/> 生物 <input type="radio"/> 农学 <input type="radio"/> 药学 <input type="radio"/> 医学 <input type="radio"/> 健康 <input type="radio"/> 卫生 <input type="radio"/> 食品 能源化工： <input type="radio"/> 能源 <input type="radio"/> 材料 <input type="radio"/> 石油 <input type="radio"/> 化学 <input type="radio"/> 化工 <input type="radio"/> 生态 <input type="radio"/> 环保
作品设计、发明的目的和基本思路，创新点，技术关键和主要技术指标	<p>针对男性高发病前列腺癌症，本专利提出了一种改进的检测方法，结合医学分割大模型 MedSAM 2 的先验知识，对超声视频进行分割，协助优化对病灶区域的检测与标注。具体而言，对于包含病灶信息的超声视频，将其划分为 K 个等长片段 C_i ($i \in [1,K]$)，然后将这些片段逐一输入到 MedSAM 2 网络，获取每个片段的最优病灶区域分割结果。基于每个片段中的病灶区域，模型将其作为伪标签，并与其他相关区域进行相似度计算，从而修正视频级别的粗粒度标签，生成更加细粒度的片段级标签。这一过程矫正了由于扫描流程等因素导致的无病灶片段对最终分期诊断的影响，提高了肿瘤分期结果的准确性</p> <p>在得到修正后的片段级标签后，这些标签被输入到图像-文本预训练模型中，进行多模态学习。通过计算超声视频中的图像特征和医学文本描述之间的相似度，模型能</p>

	<p>够更好地理解图像内容与医学信息之间的潜在关联，从而实现分期和图像之间的对齐。</p> <p>该方法有效提升了肿瘤诊断精度，显著提高了模型对前列腺肿瘤的分期检测能力。最终，本方法可以更细致准确的进行肿瘤分期诊断，具有较高的临床应用价值，能够在超声视频的肿瘤检测和分期中发挥重要作用。</p>
作品的科学性先进性（必须说明与现有技术相比、该作品是否具有突出的实质性技术特点和显著进步。请提供技术性分析说明和参考文献资料）	<p>为了实现图像文本的对齐，从而更好的指导监督学习实现对前列腺肿瘤分期诊断是 AI for Medical 的一个重要课题，在这个过程中，数据样本的精细程度间接影响模型分期的准确程度，现有的方法要达到能够利用与实际诊断场景的程度，还存在以下问题：</p> <p>（1）使用微超声检测的检测实现较多^[1]，微超声能检测更小的解剖结构或者病变^[2]，但是出于设备经济效益考虑问题，超声检测是一个较为折中的检测方案。</p> <p>（2）现有针对传统超声数据做 AI for Medical 的工作利用的模型忽略了超声扫描数据的特征：有大量无关信息^{[3][4]}。</p> <p>本专利创新优越性</p> <p>本发明提出一种基于大模型先验知识利用超声影像数据的前列腺肿瘤分期预测方法，其创新性总结如下：</p> <p>（1）利用传统的超声数据检测，兼顾了检测的开销和检测的准确性。</p> <p>（2）无关数据压缩机制：本发明提出了一个基于大模型的前处理机制，在数据端压缩了前列腺超声影像扫描中存在大量无关信息对模型的影响。</p> <p>（3）筛选粗粒度标签机制：对于主要数据的粗标签，本发明提出了一个置信度机制，在大模型的分割下，初步筛选粗粒度标签下的超声视频，在时序上重点关注病灶区域。</p> <p>[1] Ditunno F, Franco A, Manfredi C, et al. Novel non-MRI imaging techniques for primary diagnosis of prostate cancer: micro-ultrasound, contrast-enhanced ultrasound, elastography, multiparametric ultrasound, and PSMA PET/CT[J]. Prostate cancer and prostatic diseases, 2024, 27(1): 29-36.</p> <p>[2] Basso Dias A, Ghai S. Micro-ultrasound: current role in prostate cancer diagnosis and future possibilities[J]. Cancers, 2023, 15(4): 1280.</p> <p>[3] Sun Y K, Zhou B Y, Miao Y, et al. Three-dimensional convolutional neural network model to identify clinically significant prostate cancer in transrectal ultrasound videos: a prospective, multi-institutional, diagnostic study[J]. EclinicalMedicine, 2023, 60.</p> <p>[4] Dias A B, O ’ Brien C, Correas J M, et al. Multiparametric ultrasound and micro-ultrasound in prostate cancer: a comprehensive review[J]. The British Journal of</p>

	Radiology, 2022, 95(1131): 20210633.
作品在何时、何地、何种机构举行的评审、鉴定、评比、展示等活动中获奖及鉴定结果	2024 年 12 月：计算机学院创新创业项目优秀项目
作品所处阶段	<input checked="" type="checkbox"/> 实验室阶段 <input type="checkbox"/> 中试阶段 <input type="checkbox"/> 生产阶段 <input type="checkbox"/> 其他_____
技术转让方式	接受多种形式的合作，包括技术开发、技术转让以及共建研发试验基地等形式。投资方可以投资入股，我方负责进行技术研发，以确保本专利能够顺利投入商业应用
作品可展示的形式	<input type="checkbox"/> 实物、产品 <input checked="" type="checkbox"/> 模型 <input type="checkbox"/> 图纸 <input type="checkbox"/> 磁盘 <input checked="" type="checkbox"/> 现场演示 <input checked="" type="checkbox"/> 图片 <input checked="" type="checkbox"/> 录像 <input type="checkbox"/> 样品

使用说明及该作品的技术特点和优势，提供该作品的适应范围及推广前景的技术性说明及市场分析和经济效益预测	<p>技术特点：利用大模型的丰富基础知识作为先验，不需要消费大量的医生资源。筛选压缩了无关信息对模型训练的影响，针对超声信息获取的特点压缩了没有病灶部位的片段。</p> <p>优势：利用传统超声影像做 Ai 检测，照顾了检测的成本，同时保证了检测的准确度，针对数据特征解决问题。</p> <p>推广前景：复旦大学附属肿瘤医院、上海市质子重离子医院、浙江大学医学院附属第一医院、上海交通大学医学院附属瑞金医院北院、上海市第一人民医院、上海市第六人民医院、上海市第九人民医院等拥有肿瘤科的医院辅助检测需要。</p>
专利申报情况	<div><div><input type="checkbox"/>已提出专利申请</div><div>申报号_____</div><div>申报日期： 年 月 日</div></div> <div><input type="checkbox"/>已获专利权批准</div> <div>批准号_____</div> <div>批准日期： 年 月 日</div> <div><input checked="" type="checkbox"/>未提出专利申请</div>

C. 当前国内外同类课题研究水平概述

说明：1. 申报者可根据作品类别和情况填写；
2. 填写此栏有助于评审。

任务：

利用传统的前列腺超声影像做前列腺癌症检测

同类工作：

目前国内外同类工作主要分为两种：利用超声影像的工作、利用微超声影像的工作

利用微超声影像的工作：

在利用前列腺微超声影像做前列腺癌症检测方面，和本工作的主要区别在于数据，微超声（micro-ultrasound）和传统超声（ultrasound）获取的方式、经济成本、获取到的信息都有很大区别，如下表展示

特点	超声影像	微超声影像
使用频率	1-20 MHz	20 MHz 及以上
分辨率	1-2 毫米	数十微米（μm）
设备要求	一般超声设备	高精度、昂贵设备，需要专业操作
适用范围	临床常规检查	研究和专业医疗领域

微超声影像从前列腺中提取的信息会比传统超声影像更加细致，同时更加要求设备和专业医生。因此利用微超声影响完成前列腺癌症检测的工作和本专利并非真正意义上的同类工作

利用超声影像的工作：

目前利用超声影像的同期工作，主要采用 3D 卷积网络来捕捉图片和视频的时序信息，这其中有一个问题会被忽视，那就是超声影像的获取方式和数据特点，通过医生的描述，超声影像的获取通过临床医生检查获取。通过非侵入的探头在患者部位附近扫描，在扫描过程的前后，会出现大片无关部分（没有扫描到病灶部位的部分），如果没有专业临床医生进行有效部位的筛选，这些无关部分干扰了模型的学习

综述：

Iqbal 等人研究了深度学习与非深度学习方法在前列腺癌检测中的应用，发现使用 LSTM 架构和 ResNet-101 网络能够有效提取前列腺 MRI 图像中的特征信息，且相比传统的手工方法或非深度学习方法（如支持向量机（Support Vector Machine, SVM）和 K-最近邻（K Nearest Neighbor, KNN）），深度学习方法在性能上有显著优势。具体而言，深度学习网络通过自动化学习图像中的复杂模式，能够有效提高检测的准确性和灵敏度，从而远超传统方法的表现。Sudhir 等人

则进一步探索了全片图像（Whole-Slide Image）在前列腺癌检测中的应用，他们将全片图像作为输入，结合深度学习网络进行训练，在二分类任务中成功将检测结果的正确性提升至 97.9%。这一研究表明，深度学习技术能够充分利用全片图像中的信息，提高病变区域的准确定位和判定能力。

随后，Hosseinzadeh 等人设计了一个结合前列腺癌先验知识的 CNN 网络，通过将 U-net 结构与分割图相结合，成功提高了 MRI 图像上的病灶检测性能。该方法通过深度学习模型自动学习前列腺癌的空间特征和形态变化，从而更精准地识别肿瘤区域。Sherif 等人则采用了监督学习方法训练 U-net，并结合 AH-Net 网络的优势，从而在前列腺癌的病灶检测与分割任务中取得了较为显著的进展。这一研究通过有效的网络设计与训练，提升了前列腺癌图像分割精度。

近年来，注意力机制的广泛应用为前列腺癌检测的精度提升带来了新的机遇。Mahdi 等人将注意力模块集成到深度学习网络中，设计了自监督特征提取和感兴趣区域（Region of Interest, ROI）裁剪与增强模块，以充分挖掘前列腺癌病灶区域的信息。这种基于注意力机制的网络结构能够更好地聚焦于病变区域，提取具有纹理和结构信息的特征，进而实现更高效的检测。Li 等人则通过结合注意力机制和多尺度信息，在 MRI 图像上实现了高效的自动化肿瘤检测和分割。通过多尺度特征的融合，网络能够在不同分辨率下检测到前列腺癌的不同阶段和类型，从而进一步提高了检测的准确性。

尽管近年来这些深度学习方法在 MRI 图像上取得了显著的进展，但大多数研究仍集中在 MRI 数据集上，针对经直肠超声（Transrectal Ultrasonography, TRUS）图像的高置信度检测网络的研究相对较少。经直肠超声作为一种广泛应用于临床的检查手段，其图像数据的特征与 MRI 图像有所不同，现有的深度学习模型在该领域的应用仍面临诸多挑战。因此，针对 TRUS 图像进行深度学习检测网络的研究具有重要的现实意义，尤其是在低资源地区，经济性和可操作性较强的经直肠超声图像的高效分析能够为前列腺癌的早期诊断提供重要支持。这一领域的进一步探索，不仅能够弥补 MRI 数据集的局限性，也能够推动前列腺癌早期筛查和诊断技术的发展。

正文部分会附上图

D.作品正文（格式参照附件 2）

同濟大學

大模型辅助的前列腺超声影像癌症诊断系统

同濟大學

TONGJI UNIVERSITY

同济大学“卓越杯”
大学生课外学术科技作品竞赛

论文题目：大模型辅助的前列腺超声影像癌症诊断系统

学院全称：计算机科学与技术学院

专业全称：数据科学与大数据技术

申报团队：朱俊泽

指导教师：倪张凯

申报类别：☒个人项目 ☐集体项目

大模型辅助的前列腺超声影像癌症诊断系统

摘要

前列腺癌是全球男性中第二常见的癌症类型，也是导致癌症相关死亡的主要原因之一。对于前列腺癌的检测有多种方法，医疗影像技术包含磁共振成像（Magnetic Resonance Imaging, MRI）、计算机断层扫描（Computed Tomography, CT）、X光成像以及微超声和超声成像技术等。这些技术提供了宝贵的视觉数据，帮助医生诊断和评估疾病。MRI 作为前列腺癌诊断的金标准具有价格高昂，设备普及率低等不利因素，相比之下，价格低廉设备成熟的超声影像技术可以作为一种值得进一步探索的替代方案，辅助医生进行高质量的前列腺癌诊断。

本发明提出了一种基于超声影像的前列腺癌症分期系统。根据超声影像的获取特点，同时考虑到医生实际诊断的流程，我们发现获取的超声图像中存在大量与诊断无关的视频片段，这些片段对于最终的前列腺癌症分期是有害的，因此如何有效压缩无关信息，提升对有效信息的关注，是提升模型学习和诊断能力的关键。基于此，我们探索了大模型辅助的前列腺癌症分期系统，借助大模型丰富的先验知识对无关信息进行压缩，同时充分考虑对有效信息的提取，从而提高模型对癌症分期的准确性，更好的辅助医生进行诊疗判断。

关键词：超声影像，前列腺癌症，MedSAM2 医学分割模型，CLIP 图像文本对齐

1	引 言	1
1.1	前列腺癌症检测意义	1
1.2	前列腺癌症检测方法	1
1.2.1	MRI 方法	1
1.2.2	CT 法	1
1.2.3	X 光成像法	1
1.3	预训练医学大模型	2
1.4	同期工作综述	2
1.5	本文所作的工作	3
2	理论部分	4
2.1	数据预处理	4
2.2	计算关键帧和相似度，压缩无关帧	4
2.3	图像文本对齐	5
3	实验部分	8
3.1	实验环境	8
3.1.1	数据集	8
3.1.2	计算资源	8
3.2	实验设定	8
3.2.1	采样方法	8
3.2.2	大模型分割	10
4	结论和展望	11
4.1	结论	11
4.2	展望	11
	参考文献	12

1 引言

1.1 前列腺癌症检测意义

前列腺癌是全球男性中第二常见的癌症类型，也是导致癌症相关死亡的主要原因之一。根据 2020 年数据，全球新发前列腺癌病例约为 141 万例，死亡病例约为 38 万例^[1]。其发病率在不同地区差异显著，高收入国家（如澳大利亚、北美、西欧等）发病率较高，而亚洲、北非等地区则相对较低。在中国，2020 年新发前列腺癌病例约为 12 万例，死亡病例约为 5 万例，位居男性癌症新发病例的第六位。此外，尽管在一些高收入国家，前列腺癌的发病率和死亡率趋于稳定或下降，但是在东欧和部分亚洲国家（包括中国），发病率仍在上升。前列腺癌的严重程度受地理区域、检测方法以及肿瘤生物学特性等因素影响，因此，制定符合当地实际的筛查与早诊早治策略至关重要。在《中国前列腺癌筛查与早诊早治指南（2022 年）》^[2]中，强调了提高筛查效果和规范性。尤其是在低资源地区，低成本且高效的诊断方法尤为关键。

1.2 前列腺癌症检测方法

1.2.1 MRI 方法

磁共振成像（Magnetic Resonance Imaging, MRI）作为医疗影像技术中最常用的影像学工具，MRI 具有较高的特异性，可用于前列腺癌的预测，但其敏感性较低，且受制于价格、禁忌症及设备普及等因素，难以普及到广泛的临床应用中。在临床检查中一般不会使用 MRI 方法进行检测，基于我们任务希望的普测性质，MRI 难以适应广泛的临床普测，和我们主要任务适配度不高。

1.2.2 CT 法

计算机断层扫描（Computed Tomography, CT）是一种非侵入性检查方法，不需要进行切开手术，因此对患者的伤害较小，且检查过程通常较快。CT 扫描可以提供前列腺及其周围组织的高分辨率图像。它有助于评估肿瘤的大小、形态及其与周围组织的关系，尤其是判断癌症是否扩散到其他器官或区域^[3]。但是 CT 扫描涉及 X 射线，因此会给患者带来辐射风险。对于需要多次检查的患者，辐射累计可能增加患癌症的风险。CT 的软组织对比度较低，这使得它在检测前列腺癌及其周围组织的微小变化时，较难区分健康和病变组织。

1.2.3 X 光成像法

X 光（X-ray）是一种广泛用于医学影像学的技术，虽然它在前列腺癌的诊断中使用较少，但在某些情况下仍然可以作为辅助检查工具。与 CT、MRI 等影像检查方法相比，X 光检查的成本较低，检查时间也较短，患者负担较轻。但由于 X 光技术主要适用于密度较大的组织（如骨骼）^[4]，它对前列腺的可视化较差。前列腺的形态和病变很难通过常规的 X 光影像清晰呈现。

1.2.4 微超声影像和超声影像法

作为另一种常见的影像诊断工具，超声成像在前列腺癌检测中也有所应用。与 MRI 相

比，超声成像具有成本低、适用人群广和对人体危害小的优势，更适用于低成本且高效的前列腺癌检测。但前列腺癌多发、散灶的特征使得单张影像的诊断价值降低，包绕整个前列腺腺体及周围组织、器官的超声视频弥补了单张影像诊断的不足。

在使用微超声(Micro-Ultrasound)还是使用传统超声(Ultrasound)，我们考虑到微超声的使用价格，设备要求和医生资源的消耗，在希望获得临床普测效果的前提下，选择使用相对价格合适、临床设备要求较低、医生资源要求较少的传统超声影像^[5]。在使用传统的超声影像设备获取超声影像时，由于对专业医生的要求较低，初级护理人员 and 初级医生对检测的操作会使得超声影像数据有相当长度的无关片段，也就是没有前列腺内容的片段，克服这个数据问题能够极大改善诊断系统。

1.3 预训练医学大模型

预训练的医学分割大模型（如基于深度学习的卷积神经网络，CNN、SAM）通常是为了提高在医学影像分析中的分割任务效果而设计的。这些模型通过大量医学影像数据进行训练，学习不同医学影像中的组织、器官、病变等区域的特征。预训练大模型在医学领域，由于数据的标注往往成本较高且有限，预训练模型常用迁移学习的方法，先在通用的大型数据集上进行训练，然后再在特定的医学影像数据上进行微调甚至零样本测试。这样可以有效地减少标注数据的需求，同时提升模型的效果。

在大量超声影像部位预训练的医学分割大模型 MedSAM2^[6]作为用广泛超声影像数据预训练的大模型，能够捕获超声影像特征，具有丰富的先验知识，通过采用 SAM 2 框架对 2D 和 3D 医疗数据进行统一分割，将分割任务统一视为视频物体追踪任务，从而实现对 2D 和 3D 数据的统一处理。该模型提出了一种新的自排序记忆库机制，该机制基于置信度和差异性动态选择信息嵌入，而不考虑时间顺序。由于其在多个数据集上的优秀泛化能力和高分割准确性，Medical SAM 2 被用于本专利中的视频片段病灶分割检测。

1.4 同期工作综述

Iqbal 等人^[7]研究了深度学习与非深度学习方法在前列腺癌检测中的应用，发现使用 LSTM 架构和 ResNet-101 网络能够有效提取前列腺 MRI 图像中的特征信息，且相比传统的手工方法或非深度学习方法（如支持向量机（Support Vector Machine, SVM）和 K-最近邻（K Nearest Neighbor, KNN）），深度学习方法在性能上有显著优势。具体而言，深度学习网络通过自动化学习图像中的复杂模式，能够有效提高检测的准确性和灵敏度，从而远超过传统方法的表现。Sudhir 等人^[8]则进一步探索了全片图像（Whole-Slide Image）在前列腺癌检测中的应用，他们将全片图像作为输入，结合深度学习网络进行训练，在二分类任务中成功将检测结果的正确性提升至 97.9%。这一研究表明，深度学习技术能够充分利用全片图像中的信息，提高病变区域的准确定位和判定能力。

随后，Hosseinzadeh 等人^[9]设计了一个结合前列腺癌先验知识的 CNN 网络，通过将 U-net 结构与分割图相结合，成功提高了 MRI 图像上的病灶检测性能。该方法通过深度学习模型自动学习前列腺癌的空间特征和形态变化，从而更精准地识别肿瘤区域。Sherif 等人^[10]则采用了监督学习方法训练 U-net，并结合 AH-Net 网络的优势，从而在前列腺癌的病灶检测与分割任务中取得了较为显著的进展。这一研究通过有效的网络设计与训练，提升了前列腺癌图像分割精度。

近年来，注意力机制的广泛应用为前列腺癌检测的精度提升带来了新的机遇。Mahdi

等人^[12]将注意力模块集成到深度学习网络中，设计了自监督特征提取和感兴趣区域（Region of Interest, ROI）裁剪与增强模块，以充分挖掘前列腺癌病灶区域的信息。这种基于注意力机制的网络结构能够更好地聚焦于病变区域，提取具有纹理和结构信息的特征，进而实现更高效的检测。Li 等人则通过结合注意力机制和多尺度信息，在 MRI 图像上实现了高效的自动化肿瘤检测和分割。通过多尺度特征的融合，网络能够在不同分辨率下检测到前列腺癌的不同阶段和类型，从而进一步提高了检测的准确性。

尽管近年来这些深度学习方法^[13-17]在 MRI 图像上取得了显著的进展，但大多数研究仍集中在 MRI 数据集上，针对经直肠超声（Transrectal Ultrasonography, TRUS）图像的高置信度检测网络的研究相对较少。经直肠超声作为一种广泛应用于临床的检查手段，其图像数据的特征与 MRI 图像有所不同，现有的深度学习模型在该领域的应用仍面临诸多挑战。因此，针对 TRUS 图像进行深度学习检测网络的研究具有重要的现实意义，尤其是在低资源地区，经济性和可操作性较强的经直肠超声图像的高效分析能够为前列腺癌的早期诊断提供重要支持。这一领域的进一步探索，不仅能够弥补 MRI 数据集的局限性，也能够推动前列腺癌早期筛查和诊断技术的发展。

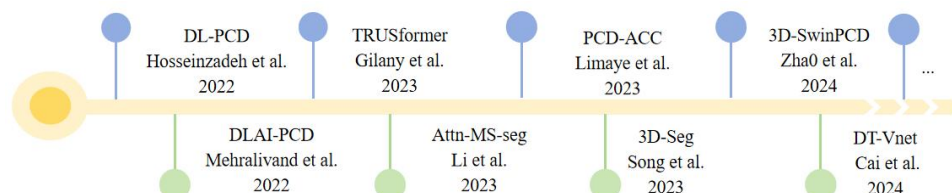


图 1-1 前列腺癌症诊断技术发展图

1.5 本文所作的工作

本专利将基础医疗大模型的丰富先验知识加入数据筛选过程，针对数据本身存在的大量非关键片段、无关片段，我们进行特定处理来压缩无关信息。之后进行 CLIP（Contrastive Language-Image Pre-Training）图像语言对比学习，我们做了如下工作：

- （1）选取了用相似数据形态预训练的医疗大模型作为先验知识
- （2）设计了选取关键帧和相似度的计算方法，通过统计方法丢弃无关片段，压缩的无关信息。
- （3）设计了针对视频的图像语言对比学习的词元和提示设计

2 理论部分

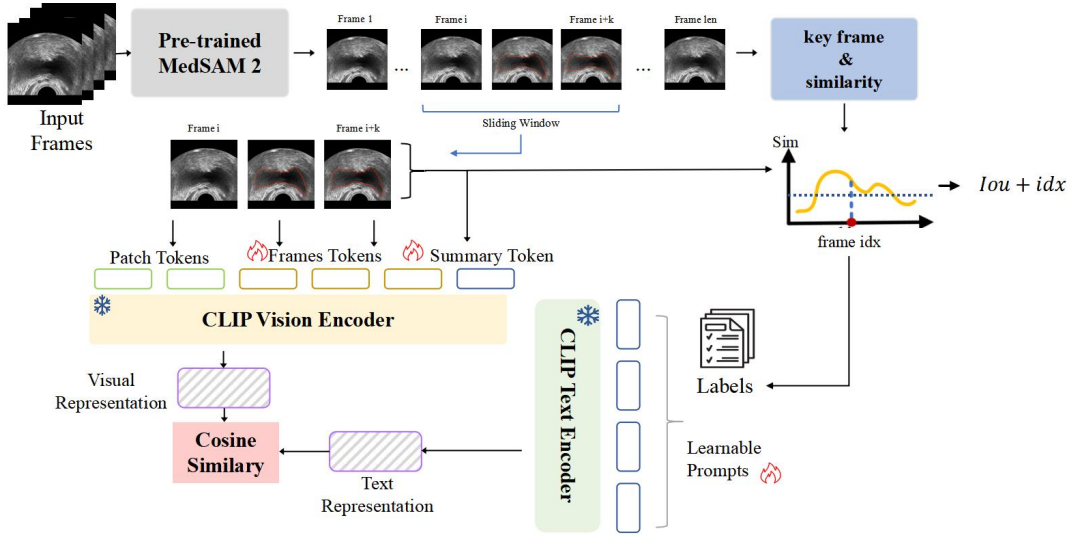


图 2-1 本专利的架构流程图

2.1 数据预处理

针对有 T 帧的视频 V ，我们采用步长为 $stride$ 的滑动窗口采样，滑动窗口长度为 num_frames ，这样就能取样出 $(T - num_frames) // stride$ 个视频的片段，此时经过预训练的 MedSAM2 分割模型，获得一个含有 Masked Region 的分割图像 $\{Seg^1, Seg^i \dots Seg^T\}_{i \in (1, T)}$ 。

2.2 计算关键帧和相似度，压缩无关帧

从数据的角度出发，超声影像扫描获得的数据存在极强连续性，即有效内容（含病灶内容）相对集中，无效内容（不含病灶内容）也是连续成片段的。因此针对帧之间距离和病灶分割内容我们使用聚类方法进行聚合。对第 i 帧，我们定义局部密度 ρ_i 为

$$\rho_i = \sum_j e^{-(d_{ij})^2}$$

其中 d_{ij} 为第 i 帧对第 j 帧的距离，定义为

$$d_{ij} = (1 - IoU_{ij}) + \frac{Idx_{ij}}{T}$$

其中 IoU_{ij} 为两帧之间的分割内容的重叠率 $IOU(Seg_i, Seg_j) = \frac{|Seg_i \cap Seg_j|}{|Seg_i \cup Seg_j|}$ ， T 为整个视频的帧数， Idx_{ij} 为第 i 帧和第 j 帧之间帧索引差的绝对值 $Idx_{ij} = |Idx_i - Idx_j|$

再计算每个数据点到相对密度 ρ_i 更高点的最小距离 δ_i

$$\delta_i = \min_{j: \rho_j \geq \rho_i} d_{ij}$$

根据每一帧的最小距离 δ_i 选择高密度中心，计算每个点的密度乘积 $\gamma_i = \delta_i * \rho_i$ 可获得密度乘积的阈值 $\gamma_{threshold}$ ，受到^[11]的启发我们设计了通过平均值和方差设置的阈值来筛选出相对的关键帧

$$\gamma_{threshold} = mean(\gamma) + std(\gamma)$$

筛选得到的密度中心 $\{\gamma_1, \gamma_2 \dots \gamma_n\}$ ，在密度中心被确定之后进行对密度中心对应的分割图片，通过分割内容最大的高密度中心确定唯一关键帧 F_{key_frame}

随后计算每一帧和关键帧 F_{key_frame} 的相似度 Sim_i

$$Sim_i = \frac{1}{2} (1 - \frac{Idx_i}{T} + IoU_i)$$

其中 Idx_i 是每第 $i \in (1, T)$ 帧和关键帧之间帧索引差的绝对值 $Idx_i = |Idx_i - Idx_{key_frame}|$ ， IoU_i 是第 i 帧和关键帧的分割内容重叠率 $IoU_i = \frac{|Seg_i \cap Seg_{key_frame}|}{|Seg_i \cup Seg_{key_frame}|}$ 。这样得到了整个视频片段的

相似度 $Sim = \{Sim_1, Sim_i \dots Sim_T\}$ ，结合之前分割出来的片段 $(T-num_frames)//stride$ 个视频的

片段。采样片段的集合 $\{t_1^1, t_1^2 \dots t_1^8\}_{N=(T-num_frames)//stride}^{i \in (1, N)}$ 我们有用帧索引对应的相似度集合

$\{Sim_1^1, Sim_1^2 \dots Sim_1^8\}_{N=(T-num_frames)//stride}^{i \in (1, N)}$ 首先计算相似度的阈值 $Sim_{threshold}$ ，其定义如下：

$$Sim_{threshold} = mean(Sim) + std(Sim)$$

也是受到了统计方法^[11]的启发我们设计了这样的阈值去筛选非关键信息，本着医学图像处理的原则，如果一个视频片段中有一帧或者更多帧相似度高于这个阈值，我们就保留这一段数据，反之这个数据片段就会被丢弃，不加入使用。这是由前列腺超声影像数据形式决定的，这样的超声数据采样时一开始和结束的时候有大量无关帧，没有病灶内容，通过前处理压缩掉这些干扰信息。

2.3 图像文本对齐

2.3.1 视频编码器（Video Encoder）

对于每一个分割出来的视频片段 $\{t_1^1, t_1^2 \dots t_1^8\}_{N=(T-num_frames)//stride}^{i \in (1, N)}$ ，根据 VIT 的要求，对每一帧 $T \in \mathbb{R}^{H \times W \times 3}$ 打成 $P \times P \times 3$ 的 N 个包，其中 $N=H \times W/P^2$ ，这些向量被扁平化成一组向量 $\{X_{t,i}\}_{i=1}^N$ ，对每一帧使用一个线性层 $P^{emb} \in \mathbb{R}^{3 \times p \times p \times D}$ 输出一个 D 维的向量。加入了一个额外的分类词元 cls token，最后每一帧输出的向量 $z_t^{(0)}$ 为

$$z_t^{(0)} = [X_{cls}, P^{emb}X_{t,1}, \dots, P^{emb}X_{t,N}] + e$$

其中 e 代表位置编码和时序编码，随后这样的向量被输入到视频编码器中，对于第 i 层的视频编码器来说有

$$z_t^{(i)} = f_{\theta_v}^{(i)}(z_t^{(i-1)})$$

最后的视觉特征第 v 帧图像的视觉特征 v_t 通过最后一层编码器之后，经过一个线性层 $P_{out} \in R^{D \times D'}$ 作为输出，将之前插入的分类词元 X_{cls} 提出，其中 l_{last} 代表最后一层编码器，视频的特征 v_t 被表示为

$$v_t = P_{out}^T z_{t,0}^{(l_{last})}$$

2.3.2 总结词元 (Summary Tokens)

我们对整段视频的特征进行一个总结，叫做 Summary Token，对于第 i 层的 Summary Token $S_t^{(i)}$ 我们将每一帧图像在第 $i-1$ 层视频编码器中的 cls Token 取出为 $Z_0^{(i-1)} = \{z_{1,0}^{(i-1)}, z_{2,0}^{(i-1)} \dots z_{T,0}^{(i-1)}\}$ ，通过一个线性投影 P_{sum} ，再自执行一次 LN (Layer-Norm) 层正则化后 MHSA (Multi-Head Self-Attention) 多头自注意力再加上自身，即

$$Z_0^{(i-1)} = P_{sum}^T Z_0^{(i-1)}$$

$$S^i = MHSA(LN(Z_0^{(i-1)})) + Z_0^{(i-1)}$$

2.3.3 全局提示词元 (Global Prompt Tokens)

为了让模型获得学习数据分布的能力，随机初始化一队可学习的向量 $G^{(i)} = \{g_1^{(i)}, g_2^{(i)} \dots g_T^{(i)}\}$

2.3.4 局部提示词元 (Local Prompt Tokens)

帧等级的提示 Token $L^{(i)} = \{l_1^{(i)}, l_2^{(i)} \dots l_T^{(i)}\}$ 也是随机初始化的可学习向量，帧等级 Local Prompt Tokens 利用了 cls token

$$l_t^{(i)} = l_t^{(i)} + z_{t,0}^{(i-1)}$$

2.3.5 最终帧输出

增加完了这些 Tokens，我们在最后一层 encoder 中，将 $S^{(last)}$, $G^{(last)}$, $L^{(last)}$ 添加到 $z_t^{(last-1)}$ 中计算 $z_t^{(last)}$ ，其中 FSA 是预训练的自注意力机制

$$[z_t^{(last)}, S^{(last)}, G^{(last)}, L^{(last)}] = FSA(LN([z_t^{(last)}, S^{(last)}, G^{(last)}, L^{(last)}]))$$

然后将添加的 $S^{(last)}$, $G^{(last)}$, $L^{(last)}$ 去除之后，对 $z_t^{(last)}$ 计算一个前馈神经网络 FFN

$$z_t^{(last)} = FFN(LN(z_t^{(last)})) + z_t^{(last)}$$

这样获得的帧表示用来计算帧视频输出

$$V_t = P_{out}^T z_t^{(l_{last})}$$

最后池化为视频表征

$$v = \text{AvgPool}(\{V_1, V_2 \dots V_t\})$$

2.3.6 文本编码器和提示

对于文本的 Encoder 编码器，我们使用预训练的 BERT 模型，模型一共有 12 层，每一层都由 MHSA 多头自注意力和 FFN 前馈神经网络组成，设 C_i ($i=1 \dots 12$) 为第 i 层文本编码器的

输出，那么有

$$C_i = FFN(MHSA(C_{i-1}))$$

最后在 $i=12$ 时，获得文本表征 $C = FFN(C_{12})$ ，采用提示学习（Prompt Learning）的方法作为文本输入，而不是手工设计的特征比如“这是一个{label}的视频”

2.3.7 对比学习目标

而言之。我们输入视频 V 和文本 C ，经过文本编码器和视频编码器后获得视频表征和文本表征如下

$$\begin{aligned} v &= f_{\theta_v}(V|S^{(last)}, G^{(last)}, L^{(last)}) \\ c &= f_{\theta_t}(C) \end{aligned}$$

对我们提取的视频表征 v 和文本表征 c ，定义余弦损失 $L_{cos}(v, c) = \frac{\langle v, c \rangle}{\|v\| \|c\|}$ ，我们使正确的 v ， c 对最大化 L_{cos} ，并且最小化其他错误的 v, c 对。

3 实验部分

3.1 实验环境

3.1.1 数据集

训练集、验证集：上海第十人民医院提供的数据集总共 800 余例

多中心测试集：宁波市第二人民医院提供数据 40 例左右、浙江大学医学院附属第一医院 50 例左右、复旦大学附属中山医院提供数据 60 例左右。

所有数据类别相对均衡，无病（Noca）案例 250 例左右，癌症初期（T0 期）案例 260 左右，癌症晚期（T1 期）案例 280 例左右。具体可如下表 3-1 所见。

3.1.2 计算资源

将训练批次大小（Batch Size, BS）设置为 64，整个过程中将学习率（Learning Rate, LR）设置为 $4e-4$ ，在两块 Tesla V100 显卡上训练。每一轮训练轮次设定为 50 个轮次，一共训练 2 天。

表 3-1 数据分布

数据	上海第十人民医院 (训练)	上海第十人民医院 (验证)	多中心医院
病例数量	640	160	150 左右
无病（Noca）数量	200	50	50 左右
癌症初期（T0 期） 数量	230	58	50 左右
癌症晚期（T1 期） 数量	210	52	50 左右

3.2 实验设定

3.2.1 采样方法

首先我们考虑了 Num_frames 间隔的采样方法，即针对 T 帧视频，采样出 $T/\text{Num_frames}$ 的片段输入后续的视频图片文本对齐，这样的采样方法导致数据的过稀疏，从而导致训练不收敛。

后采用滑动窗口的采样方法，即对于 T 帧视频，如果步长为 stride 为 s，那么会采样处 $(T - \text{Num_frames}) // s$ 的片段，在设计 stride 的时候我们考虑了数据在不同 stride 下的重复采样的冗余程度，分别设定了不同的 stride，经过我们的消融实验，发现 stride 为 2 时候实验效果最好。

在前处理的具体方式上，针对医生对数据的描述，我们打算采用两种方法：出于无关信息在本质上相当于无病灶部分，应该归于无病，所以将无关信息帧的 label 从 T0 期癌症和 T1 期癌症修改为良性标签；出于原本的良性标签的视频也存在一些无关区域，那么我们将无关信息帧片段直接丢弃。从实验结果来看，将无关信息帧直接进行丢弃效果更好。效果图

表如下：

表 3-2 实验结果

方法	Precision	Recall
Num_frames 间隔采样	不收敛	不收敛
滑动窗口采样	90.43%	90.03%
前处理（改 label）+滑动窗口采样	85.77%	82.24%
前处理（丢弃）+滑动窗口采样	91.55%	91.12%

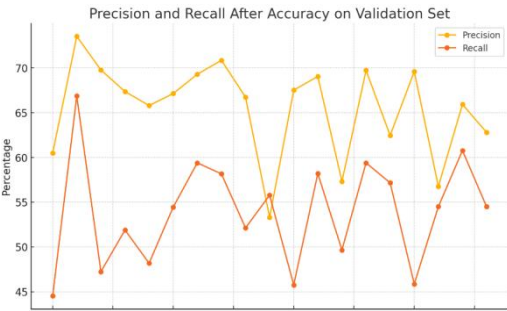


图 3-1 Num_frames 间隔采样精确率召回率

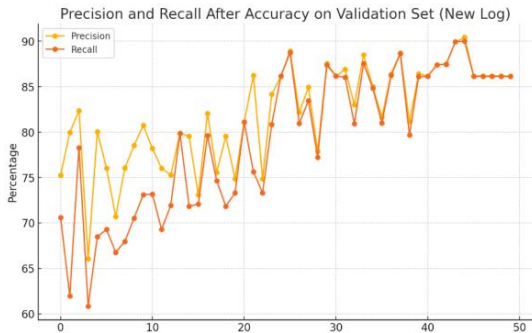


图 3-2 滑动窗口采样精确率召回率

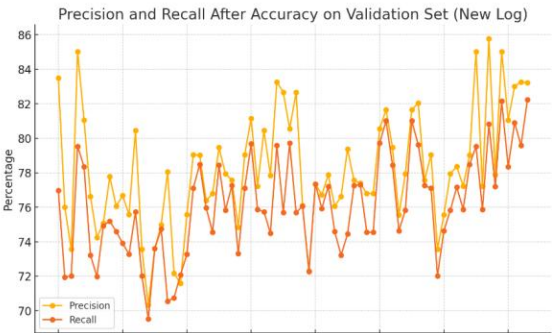


图 3-3 前处理（改 label）+滑动窗口采样精确率召回率

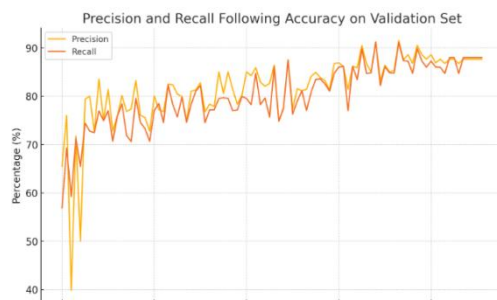


图 3-4 前处理（丢弃）+滑动窗口 采样精确率召回率

3.2.2 大模型分割

首先我们部署了预训练好的 MedSAM2 模型，它在有临床医生的提示的眼睛，大脑，腹部，肺部，胰腺等数据集上预训练，涵盖了 78 个数据集。使用 MedSAM2 工作提供的预训练模型进行分割，下面是分割可视化例子。由于 MedSAM2 的官方输入方式是 224×224 分辨率的图片，而我们的超声数据都是 512×512 的数据，我们对结果的采样直接采用上采样将分割内容恢复到原图的分辨率。

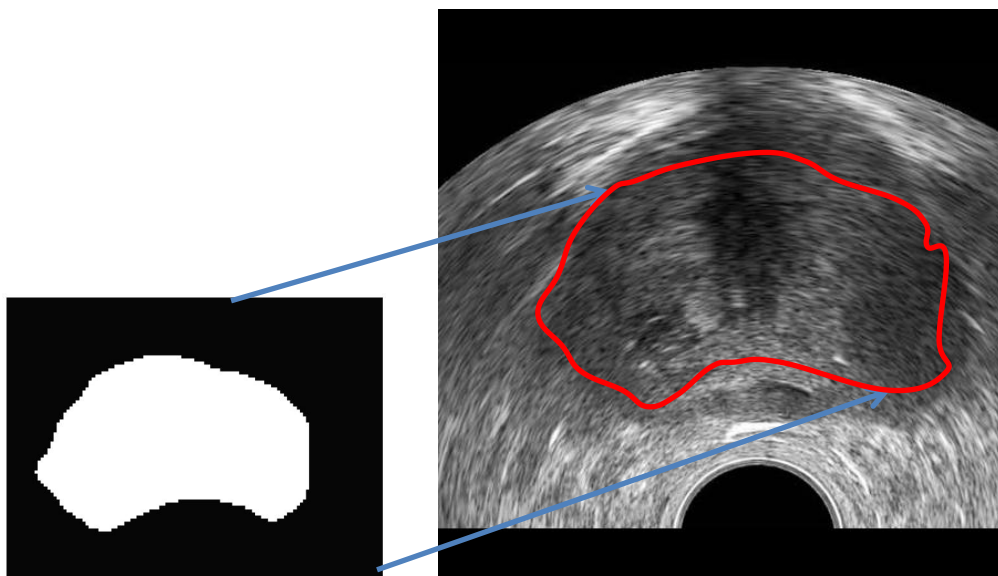


图 3-5 分割的可视化

4 结论和展望

4.1 结论

(1) 我们利用了成本较低，临床要求较少，对医生资源消耗较少的传统超声数据进行前列腺癌症的检测，提出了本发明作为临床普查的工具。

(2) 关于超声数据固有的关键数据稀疏、无关片段过多的特点，我们利用大模型的丰富基础知识作为先验检测而不是医生手动标注，对分割内容利用了机器学习的聚类方法选出了关键帧，通过相似度来筛选出无关帧无关信息。

(3) 关于采样方式，我们从数据密度入手，最后确定了步长为 2，窗口长度为 8 的滑动窗口采样方法。

在前列腺癌症诊断这个三分期任务上和上海第十人民医院的相关专家沟通，医院方面持积极态度，基本满足临床普测的任务要求。

4.2 展望

(1) 针对这样的训练范式，我们希望将本工作的训练方式扩展到其他同类任务中，即对有相同问题的数据--同样有大量无关内容的数据进行压缩。

(2) 在多中心验证时推理在一张 Tesla V100 中的推理时间在三分分钟左右，关于大模型的推理时间和能提供的先验信息可以再优化。

(3) 在应用中本专利希望能够实现小成本的临床普测，因此日后工作中要从加速推理方面入手。

(4) 在人机交互方面，我们希望后续针对大模型的输出添加一个下游任务：和诊断医生交互。PMC-LLaMA^[18]是一个针对医学领域知识训练的大型语言模型，通过提取和训练 4800 万篇医疗领域专业文献和超过 3 万本医学领域教科书中的文本，构建了一个包含 202 兆分词的医学问答、推理和对话数据集。该模型在医学领域的知识检索和文本对话中表现出色。在本项目中，我们通过在前列腺癌超声视频数据集上微调该模型，使其能够针对特定任务进行优化。通过将修正后的类标签与文本提取的特征一同输入该大模型，模型输出具有专业知识和置信度的初步诊断报告。最终，联合模型给出的诊断结果将共同提供给医生，辅助简化诊断流程并提高诊断准确性。借助上述大模型的相关技术，可以实现多种辅助功能，如生成初步诊断报告、进行图像与文本的知识对齐，或利用额外的分割图进行病灶定位指导。这些方法能够显著提高前列腺肿瘤超声视频分期检测的准确性，帮助更精确地评估病灶的特征和分期，从而更有效地支持医生在诊断决策过程中做出更具依据的判断，提高分期诊断的精度。

参考文献

- [1] Sung H, Ferlay J, Siegel R L, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA: a cancer journal for clinicians, 2021, 71(3): 209-249.
- [2] 赫捷, 陈万青, 李霓, 等. 中国前列腺癌筛查与早诊早治指南 (2022, 北京)[J]. 中华肿瘤杂志, 2022, 44(1): 29-53.
- [3] Liu, L., & Jiang, H. (2017). "CT and MRI in staging of prostate cancer." Cancer Imaging, 17(1), 24.
- [4] González, H., & Márquez, M. (2017). "Bone imaging in prostate cancer: The role of X-ray, CT, and MRI." European Journal of Nuclear Medicine and Molecular Imaging, 44(1), 13-22.
- [5] Dias A B, O'Brien C, Correias J M, et al. Multiparametric ultrasound and micro-ultrasound in prostate cancer: a comprehensive review[J]. The British Journal of Radiology, 2022, 95(1131): 20210633.
- [6] Zhu J, Qi Y, Wu J. Medical sam 2: Segment medical images as video via segment anything model 2[J]. arXiv preprint arXiv:2408.00874, 2024.
- [7] Iqbal S, Siddiqui G F, Rehman A, et al. Prostate cancer detection using deep learning and traditional techniques[J]. IEEE Access, 2021, 9: 27085-27100.
- [8] Perincheri S, Levi A W, Celli R, et al. An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy[J]. Modern Pathology, 2021, 34(8): 1588-1595.
- [9] Hosseinzadeh M, Saha A, Brand P, et al. Deep learning-assisted prostate cancer detection on bi-parametric MRI: minimum training data size requirements and effect of prior knowledge[J]. European Radiology, 2022: 1-11.
- [10] Mehralivand S, Yang D, Harmon S A, et al. Deep learning-based artificial intelligence for prostate cancer detection at biparametric MRI[J]. Abdominal Radiology, 2022, 47(4): 1425-1434
- [11] Sadiq B O, Muhammad B, Abdullahi M N, et al. Keyframe extraction techniques: A review[J]. ELEKTRIKA-Journal of Electrical Engineering, 2020, 19(3): 54-60.
- [12] Gilany M, Wilson P, Perera-Ortega A, et al. TRUSformer: Improving prostate cancer detection from micro-ultrasound using attention and self-supervision[J]. International Journal of Computer Assisted Radiology and Surgery, 2023, 18(7): 1193-1200
- [13] Li Y, Huang M, Zhang Y, et al. A dual attention-guided 3D convolution network for automatic segmentation of prostate and tumor[J]. Biomedical Signal Processing and Control, 2023, 85: 104755
- [14] Limaye S, Chowdhury S, Rohatgi N, et al. Accurate prostate cancer detection based on enrichment and characterization of prostate cancer specific circulating tumor cells[J]. Cancer Medicine, 2023, 12(8): 9116-9127
- [15] Song E, Long J, Ma G, et al. Prostate lesion segmentation based on a 3D

end-to-end convolution neural network with deep multi-scale attention[J]. Magnetic Resonance Imaging, 2023, 99: 98-109

[16]Li Y, Wynne J, Wang J, et al. Cross-shaped windows transformer with self-supervised pretraining for clinically significant prostate cancer detection in bi-parametric MRI[J]. Medical Physics, 2023

[17]Lu X, Liu X, Xiao Z, et al. Self-supervised dual-head attentional bootstrap learning network for prostate cancer screening in transrectal ultrasound images[J]. Computers in Biology and Medicine, 2023, 165: 107337

[18]Wu C, Lin W, Zhang X, et al. PMC-LLaMA: toward building open-source language models for medicine[J]. Journal of the American Medical Informatics Association, 2024: ocae045