

智驱结肠镜：技术变革与前沿洞察

季葛鹏, 刘静怡, 徐鹏, Nick Barnes, Fahad Shahbaz Khan, Salman Khan, 范登平

摘要—结肠镜检查是目前最为灵敏的结直肠癌筛查方法之一。本研究旨在调查智驱结肠镜检查技术的前沿动态及其对多模态医疗应用的前瞻性影响, 为了实现这一目标, 本文立足于结肠镜场景感知的四大任务 (分类、检测、分割和视觉语言理解), 评估了当前以数据为中心和以模型为中心的研究现状。评估揭示出特定领域所面临的挑战, 并且强调了在结肠镜检查中深入开展多模态研究的必要性。为弥补领域空白, 我们构建了三项核心举措: 一个大规模的多模态指令微调数据集 ColonINST、一个针对结肠镜垂域的多模态语言模型 ColonGPT 以及一个多模态基准。为推动这一迅速发展领域的持续进步, 我们建立了一个公开网站以便获取最新的领域资讯: <https://github.com/ai4colonoscopy/IntelliScope>。

Index Terms—智驱结肠镜检查综述、息肉分割、多模态大语言模型、多模态基准、医疗人工智能。

1 引言

尽管结直肠癌 (CRC) 在高收入国家的发病率有所下降, 但在发展中国家的发病率仍呈现上升趋势, 其依然是全球确诊率排名第三的癌症 [2]。结肠镜检查是一种有效的 CRC 筛查方法, 借助配备摄像头的软管对结肠内部进行可视化检查。如图 (1-a) 所示, 在临床操作中还会使用圈套器、镊子及电灼设备等专用器械, 以切除癌前病变, 如锯齿状息肉和腺瘤性息肉。近期研究表明, 将人工智能 (AI) 应用于结肠镜检查相较于传统方法, 可将结直肠癌变的漏诊率降低约 50% [3]。这一成果也激发了我们对智驱结肠镜前沿技术的进一步探索。

结肠镜检查是一种内镜光学成像技术, 但由于结肠解剖结构复杂且多褶, 图像通常存在照明不均和视觉模式同质性高等问题, 这些特点与通用图像数据 (如 ImageNet [4]) 存在显著差异。此状况表明, 解读结肠镜检查数据亟需专门的方法。为此, 本文首先系统梳理了结肠镜检查中最新的智能技术, 评估当前研究现状, 识别该领域特有的挑战以及尚未充分探索的方向。我们发现当前结肠镜领域的多模态研究仍有大量空白。为填补领域空白, 本研究为该领域做出三方面的贡献, 如图 (1-b) 所示。

贡献. (a) 本文从数据驱动与模型驱动两个角度出发, 梳理了四项结肠镜场景感知任务 (参见图 2) 的最新研究进展。我们总结了自 2015 年以来发布的 63 个数据集及 137 种代表性深

- 通讯作者: 范登平 (dengpfan@gmail.com)。范登平现就职于南开国际先进研究院 (深圳福田), 以及中国天津南开大学计算机学院 VCIP 实验室。
- 季葛鹏和 Nick Barnes 在澳大利亚堪培拉的澳大利亚国立大学计算机学院工作。
- 刘静怡在日本横滨庆应义塾大学科技研究生院工作。
- 徐鹏在中国北京清华大学电子工程系工作。
- Fahad Shahbaz Khan 和 Salman Khan 在阿联酋阿布扎比的穆罕默德·本·扎耶德人工智能大学工作。
- 本文为 arXiv [1] 论文的中文翻译参考。本翻译初稿由李懿凝 (南开大学) 完成, 校稿由刘静怡、季葛鹏负责。

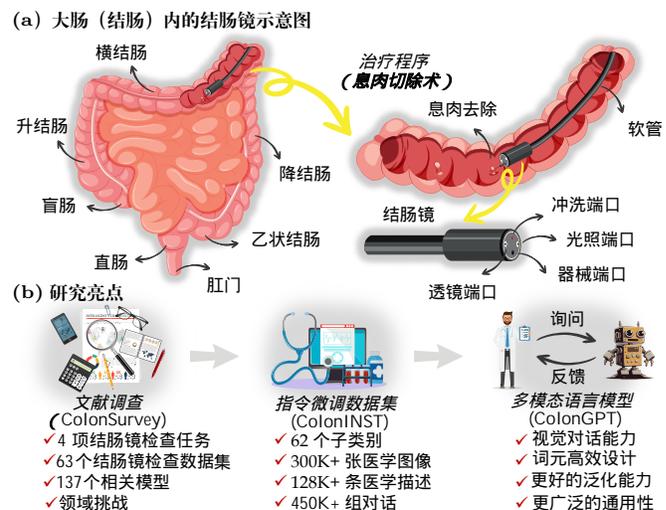


图 1. 结肠镜检查概述和研究亮点: 我们在图 (a) 中描述了消化道中大肠 (结肠) 的解剖结构、结肠镜检查过程中息肉切除流程以及结肠镜组件; 在图 (b) 中总结了本研究的三个核心亮点。

度学习方法的关键特征, 并进一步提炼了当前发展趋势与未来研究的潜在方向。(b) 本研究引入了 ColonINST, 一个专为多模态研究设计的开创性指令微调数据集, 旨在指导模型以交互方式执行用户驱动的任务。该数据集由 19 个公开数据源汇编而成, 涵盖 62 个子类别, 共包含 303,001 张结肠镜检查图像, 覆盖了结肠镜检查过程常见的各种临床场景。我们从两个方面扩展了这些图像样本: 首先, 利用多模态人工智能聊天机器人 GPT-4V [5] 自动生成了 128,620 条医学图像描述; 其次, 重构了 450,724 条适用于多模态任务人机对话数据。(c) 基于这些指令微调数据, 我们构建了多模态语言模型 ColonGPT, 可通过交互式对话为内镜医师提供辅助支持。为确保社区内普通用户能够复现实验结果, 本研究以资源友好的方式实现了 ColonGPT: 其采用了具有 0.4B 参数量的视觉编码器 SigLIP-SO [6] 和具有 1.3B 参数量的轻量级语言模型

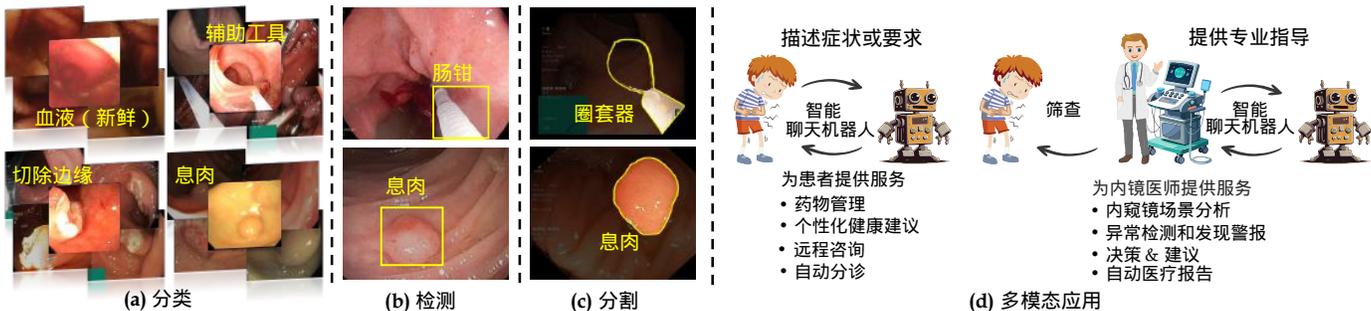


图 2. 从视觉到多模态视角的结肠镜场景感知. 在临床实践中, 纯视觉任务包括 (a) 分类、(b) 检测和 (c) 分割, 它们用于识别目标区域, 例如息肉或医疗器械. (d) 多模态应用通过执行符合临床需求的交互式、用户驱动任务来改进结肠镜检查流程. 借助聊天机器人, 系统不仅能够提供个性化建议, 自动化报告生成, 还能优化手术操作流程.

Phi1.5 [7] 进行开发. 与以往采用多层感知机恒等地处理视觉编码器中所有图元的视觉语言 (VL) 桥接方法 [8]–[10] 不同的是, 本研究提出了一种多粒度适配器, 可根据图元的重要性进行有选择性采样. 该策略在不影响性能的前提下, 将图元数量缩减到原来的约 34%, 使 ColonGPT 在新创建的多模态基准测试的三项任务中均位居前列. 尤为关键的是, ColonGPT 可以在两块英伟达 H200 显卡上于七小时内完成训练, 为后续相关研究的快速推进提供了良好基础.

研究范围. 本研究在在诸多方面区别于现有工作. 2020 年之前对传统方法 [11] 和深度学习方法 [12]–[15] 所展开的早期综述, 已无法满足当前技术发展的需求. 尽管近期有研究 [16] 探讨了结肠镜检查中的多种应用场景, 例如质量分析和异常检测, 但其缺乏系统性的定量分析. 此外, 现有的一些基准工作 [17]–[19] 主要聚焦于较为狭窄的子领域, 覆盖范围有限. 相比之下, 本文深入研究了与结肠镜场景感知中的四项关键任务, 并对其当前研究现状进行了全面评估, 旨在明确领域中的核心挑战与尚待探索的问题. 更重要的是, 我们不仅关于已有进展, 还致力于为多模态研究领域的未来发展奠定基础. 为此, 本文采取了三项举措: 构建多模态指令微调数据集、研发多模态语言模型以及建立多模态评测基准.

文章结构. 本文剩余部分的结构安排如下: 在 §2 章节中, 我们回顾了该领域的发展历程, 并分析了其面临的独特挑战; 在 §3 章节中, 系统梳理了 63 个与结肠镜检查相关的数据集; 随后在 §4 章节中, 调研并评估了 137 个代表性的深度学习模型; §5 章节介绍了我们为多模态研究提出的三项核心举措: 包括 ColonINST 数据集的构建、ColonGPT 模型的设计细节, 以及一个包含对比分析与消融实验的多模态评测基准; 最后在 §6 章节中总结全文并展望未来研究方向.

2 背景

2.1 起源和发展

结肠镜检查的发展历史中有两个关键的里程碑. 第一个里程碑出现在 1968 年, 当时胃肠外科医生 Hiromi Shinya 和 William Wolff 发现了结肠息肉与肠道肿瘤之间的关联, 但当时尚无合

适的设备对这些息肉进行检查. 1969 年, 他们发现了康宁公司 (Corning Incorporated) 制造的光纤材料, 并与奥林巴斯合作, 研制出了纤维结肠镜这一开创性的设备, 它不仅能够深入检查结肠, 还可通过钢丝圈对息肉进行切除. 第二个里程碑则是 1983 年电子结肠镜的问世 [80], 该技术使得结肠内部结构可在屏幕上可视化, 并可通过息肉切除套圈辅助进行息肉切除, 从而提高了病变检测率并减少了术中出血风险. 21 世纪迎来了人工智能时代, 计算机辅助诊断系统为结肠镜手术带来了更高的精度与可靠性 [81]. 本文重点探讨了智能技术对结肠镜检查 (一种内窥镜成像方式 [82]) 所带来的变革性影响; 其他相关的智能内窥镜技术, 如腹腔镜检查 [83], 在附录中进行了简要回顾.

2.2 内在特征和领域独特挑战

本文总结了结肠镜视觉任务所面临的五个独特挑战, 这些挑战主要源于结肠镜检查过程中复杂的操作方式及成像条件限制. (a) 非线性相机自运动. 由于操作空间受限, 结肠镜摄像头需要主动推进并以非线性、不规则的方式移动, 这种运动模式导致自运动补偿方法 [84] 难以适用, 并会引发运动模糊现象 [63]. (b) 医疗器械的干扰. 结肠镜检查过程中通常会用到诸如内窥镜、导丝和圈套器等医疗器械, 为了进行有效的分析, 算法需要具备正确区分这些器械与解剖结构的能力 [85]. (c) 受限的可视区域. 结肠内复杂的褶皱和盲点限制了结肠镜图像中的可见范围. 这就要求算法需具备从有限的视觉场景中提取有效信息的能力 [86]. (d) 光照不均匀. 结肠的黏膜表面常处于湿润状态并伴有反光, 导致成像中存在显著的光照变化, 包括漫反射、非朗伯反射和相互反射等复杂光学反射情况. 在这些条件下, 传统的基于光照建模的算法往往难以奏效 [87]. (e) 组织外观的可变性. 由于肠道运动、病变状态、个体差异以及医疗器械的干扰, 会导致黏膜表面在纹理和颜色上表现出显著变化. 此外, 良性息肉或其他病变通常具有较模糊或均匀边界 [88], 与周边组织融合度高, 进一步扩大了检测难度. 这些问题要求人工智能模型能够对固有的形态和颜色波动保持稳健响应的能力.

表 1

结肠镜检查数据集的统计信息. 各列内容包括: 图像数量 (#IMG) 和视频数量 (#VID)、分类标签 (Cls)、边界框 (Bbx)、分割掩码 (Seg)、文本描述 (Tx). 与结肠镜无关的类别 (如胃部、食管炎等) 用灰色标记.

数据集	发表出处	#IMG	#VID	Cls	Bbx	Seg	Tx	类别数量 (#C) → 类别名称	链接
CVC-ColonDB [20]	PR'12	300	15	-	-	-	-	#C1 → 息肉	Link
ETIS-Larib [21]	CARS'14	196	-	-	-	-	-	#C2 → 息肉、非息肉	Link
CVC-ClinicDB [22]	CMIG'15	612	31	-	-	-	-	#C1 → 息肉	Link
ASU-Mayo [23]	TMI'15	36,458	38	✓	✓	✓	✓	#C2 → 息肉、非息肉	Link
Ye 等人 [24]	MedIA'16	7,894	10	✓	✓	✓	✓	#C2 → 息肉、非实例	Link
Deeba 等人 [25]	IJCNN'16	100	-	✓	✓	✓	✓	#C2 → 出血、非出血	-
CU-ColonDB [26]	JBHI'16	1,930	-	✓	✓	✓	✓	#C3 → 增生性息肉、腺瘤性息肉、非息肉	-
ColonoscopicDS [27]	TMI'16	-	76	✓	✓	✓	✓	#C3 → 锯齿状腺瘤、增生性病变、腺瘤	Link
CVC-ClinicVideoDB [28]	MICCAI'17	10,924	18	✓	✓	✓	✓	#C2 → 息肉、非息肉	Link
Kvasir [29]	MMSys'17	8,000	-	✓	✓	✓	✓	#C8 → 盲肠、息肉、溃疡性结肠炎、染色并抬升的息肉、染色的切除边缘、乙线、幽门、食管炎	Link
Nerthus [30]	MMSys'17	5,525	21	✓	✓	✓	✓	#C4 → 波士顿肠道准备量表 (Boston-Bowel-Preparation-Scale) 0/1/2/3	Link
EndoSceneStill [31]	JHE'17	912	44	✓	✓	✓	✓	#C1 → 息肉	Link
KID1 [32]	EIO'17	137	-	✓	✓	✓	✓	#C10 → 血管扩张、溃疡、狭窄、绒毛水肿、结节性淋巴管扩张、乳腺囊肿、息肉、口腔、正常/无病变、管腔内出血	Link
KID2 [32]	EIO'17	2,371	47	✓	✓	✓	✓	#C4 → 血管异常、息肉样异常、炎症性异常、正常图像	Link
NBIPolyp-UCdb [33]	BSPC'19	86	11	✓	✓	✓	✓	#C2 → 腺瘤、增生性	Link
WLPolyp-UCdb [34]	EIO'19	3,040	42	✓	✓	✓	✓	#C2 → 息肉、正常黏膜	Link
ASEI [35]	MM'19	4,470	-	✓	✓	✓	✓	#C4 → 染色并抬升的息肉、染色的切除边缘、器械、息肉	Link
Cho 等人 [36]	PeerJ'19	328,927	112	✓	✓	✓	✓	#C1 → 盲肠	Link
EAD2019 [37]	arXiv'19	2,342	-	✓	✓	✓	✓	#C7 → 成像伪影、对比度、镜面反射、器械、气泡、运动模糊、饱和度	Link
Liu 等人 [38]	ISBI'20	14,317	18	✓	✓	✓	✓	#C2 → 息肉、非息肉	-
Kvasir-SEG [39]	MM'20	1,000	-	✓	✓	✓	✓	#C1 → 息肉	Link
PICCOLO [40]	AppSci'20	3,433	39	✓	✓	✓	✓	#C17 → 巴黎分类 (隆起型病变: 0-Ip/0-Ips/0-Is, 扁平隆起型病变: 0-IIa/0-IIa+c, 平坦型病变: 0-IIb)、尼斯分类 (1/2/3 型)、诊断 (腺瘤/腺瘤/增生)、组织学分层 (高级别异型增生/增生/浸润性腺瘤/低级别异型增生/无异型增生)	Link
EDD2020 [41]	arXiv'20	386	-	✓	✓	✓	✓	#C5 → 可疑区域、高级别异型增生、腺瘤、息肉、正常发育异常的巴雷特食管	Link
CAD-CAP [42]	EIO'20	25,124	1,686	✓	✓	✓	✓	#C4 → 血管病变、新鲜血液、溃疡性炎症病变、正常图像	-
ACP-ColonDB ₅₃₀ [43]	NPJDM'20	221,976	-	✓	✓	✓	✓	#C13 → 腺瘤性息肉、增生性息肉、其他息肉、出血、回盲瓣、器械、伪影、正常结肠结构、气泡、结肠内部背景、粪便、管腔、结肠外部背景	-
HyperKvasir [44]	SData'20	110,079	374	✓	✓	✓	✓	#C23 → 盲肠、直肠反转、波士顿肠道准备量表评分 0-1/2-3、溃疡性结肠炎 1/2/3/0-1/1-2/2-3 级、息肉、染色并抬升的息肉、染色的切除边缘、痔疮、巴雷特食管、回肠末端、乙线、食管炎 A 级、食管炎 B-D 级、幽门、胃反转、短节段巴雷特食管、嵌塞的粪便	Link
WCE-Polyp [45]	TMI'20	541	-	-	-	-	-	#C1 → 息肉	Link
EAD2020 [46]	MedIA'21	2,531	-	✓	✓	✓	✓	#C8 → 镜面反射、气泡、饱和、对比度、血液、器械、模糊、成像伪影	Link
BKAI-Small [47]	ISVC'21	1,200	-	✓	✓	✓	✓	#C3 → 非肿瘤性息肉、肿瘤性息肉、背景	Link
BKAI-Large [47]	ISVC'21	7,466	-	✓	✓	✓	✓	#C4 → 非肿瘤性息肉、肿瘤性息肉、未定义息肉、背景	Link
CPC-Paired [48]	MICCAI'21	681	-	✓	✓	✓	✓	#C2 → 增生性息肉、腺瘤	Link
LDPolyVideo [49]	MICCAI'21	901,666	263	✓	✓	✓	✓	#C2 → 息肉、非息肉	Link
Celik 等人 [50]	MICCAI'21	2,224	-	✓	✓	✓	✓	#C2 → 息肉、巴雷特食管	Link
Kvasir-Instrument [51]	MM'21	590	-	✓	✓	✓	✓	#C1 → 胃肠道手术工具 (例如: 圈套器、球囊和活检钳)	Link
CP-CHILD [52]	BMCMI'21	9,500	-	✓	✓	✓	✓	#C2 → 结肠息肉、正常或其他病理图像	Link
CROHN-IP1 [53]	EIO'21	3,498	-	✓	✓	✓	✓	#C7 → 红斑、水肿、阿弗他溃疡、溃疡 (3-10 毫米, 大于 10 毫米)、狭窄、非病理性	Link
C-E Crohn's Disease [54]	FMOLB'21	467	164	✓	✓	✓	✓	#C1 → 克罗恩病病变	-
SUN-database [55]	GIE'21	159,232	113	✓	✓	✓	✓	#C7 → 增生性息肉、低级别腺瘤、高级别腺瘤、传统锯齿状腺瘤、浸润性癌、无蒂锯齿状病变、阴性	Link
Kvasir-Sessile [56]	JBHI'21	196	-	-	-	-	-	#C1 → 息肉 (<10mm)	Link
Kvasir-Capsule [57]	SData'21	4,741,504	117	✓	✓	✓	✓	#C14 → 息肉、回盲瓣、淋巴管扩张、红斑、血管扩张、异物、糜烂、溃疡、血液 (新鲜)、血液 (含铁血黄素)、正常清洁黏膜、黏膜视图减少、幽门、十二指肠大乳头	Link
KUMC [58]	PONE'21	37,899	155	✓	✓	✓	✓	#C2 → 增生性息肉、腺瘤性息肉	Link
ERS* [59]	arXiv'22	1,354,667	1,520	✓	✓	✓	✓	#C27 → 溃疡性结肠炎 (活动期 / 静止期)、狭窄 (术后 / 炎症性 / 恶性)、息肉、黑变病、憩室病、瘘管、克罗恩病 (活动期 / 静止期)、脂肪瘤、直肠炎、痔疮、黏膜下肿瘤、孤立性溃疡、不明原因出血、回肠炎、憩室炎、结肠炎: 缺血性、结肠癌、血管发育异常、直肠溃疡、异物、息肉病综合征、术后表现、寄生虫	Link
Tian 等人 [60]	MICCAI'22	807,069	253	✓	✓	✓	✓	#C2 → 息肉、非息肉	Link
WCE-CCDD [61]	BSPC'22	6,000	-	✓	✓	✓	✓	#C4 → 溃疡、息肉、正常、食管炎	Link
PolypGen2.0 [62]	ISBI'22	3,446	46	✓	✓	✓	✓	#C2 → 锯齿状、腺瘤	Link
SUN-SEG [63]	MIR'22	159,232	1,013	✓	✓	✓	✓	#C7 → 增生性息肉、低级别腺瘤、高级别腺瘤、传统锯齿状腺瘤、浸润性癌、无蒂锯齿状病变、阴性	Link
SinGAN-Seg [64]	PONE'22	10,000	-	-	-	-	-	#C1 → 息肉	Link
ENDOTEST [65]	SJG'22	253,754	58	✓	✓	✓	✓	#C2 → 息肉、非息肉	Link
MEDVQA-GI [66]	CLEF'23	3,949	-	✓	✓	✓	✓	#C2 → 息肉、手术设备	Link
GastroVision [67]	ICMLW'23	8,000	-	✓	✓	✓	✓	#C27 → 辅助工具、血管扩张、管腔内血液、盲肠、结肠憩室、切除边缘、结肠癌、染色并抬升的息肉、红斑、溃疡、染色切除边缘、直肠倒转、大肠黏膜炎症、切除的息肉、结肠息肉、回盲瓣、大肠正常黏膜和血管模式、食管炎、巴雷特食管、十二指肠球部、食管静脉曲张、胃息肉、胃食管交界正常乙线、正常食管、正常胃、幽门、小肠末端回肠	Link
W-Polyp [68]	CVPR'23	1,450	-	-	-	-	-	#C1 → 息肉	Link
LIMUC [69]	IBD'23	11,276	-	✓	✓	✓	✓	#C4 → MAYO 内镜评分 (MES) 0/1/2/3	Link
PS-NBI2K [70]	JBHI'23	2,000	-	✓	✓	✓	✓	#C1 → 息肉	Link
PolypGen [17]	SData'23	8,037	23	✓	✓	✓	✓	#C2 → 息肉、阴性	Link
MedFMC** [71]	SData'23	22,349	-	✓	✓	✓	✓	#C5 → 溃疡、糜烂、息肉、肿瘤和非实例	Link
GB-WCE Dataset [72]	MD'23	226	-	✓	✓	✓	✓	#C2 → 出血或病变、正常	Link
REAL-Colon [73]	SData'24	2,757,723	60	✓	✓	✓	✓	#C2 → 息肉、阴性	Link
Xu 等人 [74]	TMI'24	251	-	✓	✓	✓	✓	#C4 → 梅奥内镜评分 (MES) 0/1/2/3	Link
Kvasir-VQA [75]	MMw'24	6,500	-	✓	✓	✓	✓	#C5 → 息肉、溃疡性结肠炎、器械、正常、食管炎	Link
CapsuleVision2024 [76]	CVIP'24	58,124	-	✓	✓	✓	✓	#C10 → 血管扩张、出血、糜烂、红斑、异物、淋巴管扩张、息肉、溃疡、蠕虫、正常	Link
COLON [77]	arXiv'24	~430,000	30	✓	✓	✓	✓	#C3 → 腺瘤、增生性、非病理病例	-
WCEBleedGen [78]	arXiv'24	2,618	-	✓	✓	✓	✓	#C2 → 出血、非出血	Link
PolypDB [79]	arXiv'24	3,934	-	✓	✓	✓	✓	#C1 → 息肉 (多种成像方式和多个医疗中心)	Link

* 注意 - ERS 数据集 [59] 总共包括 99 个注释类别. 为简洁起见, 此处仅列出 ERS 中与结肠相关的 27 个类别.

** 注意 - MedFMC 数据集 [71] 由五种模态的 23349 张医学图像组成. 此表仅列举了与内镜模态特定的类别.

3 重新审视结肠镜检查数据

3.1 结肠镜检查的医疗数据

表 1 展示了本研究对 63 个数据集的调查情况, 这些数据集包含与结肠镜场景感知四项任务相关的基本统计信息. 本研究

通过诸如“colonoscopy dataset”和“gastrointestinal disease dataset”等关键词进行检索, 并筛选出包含人体结肠相关的图像或视频的数据集. 特别是一些数据集还包含其他器官的图像, 比如数据集 [29] 中的幽门, 以及数据集 [67] 中的胃的图像. 接下来, 本文将根据不同的任务目标对这些选定的数据集

进行回顾分析。

- **分类数据集**已被广泛应用于多种目的，包括图像层面 [29], [32], [41], [44], [53], [59], [61], [67], [69], [74], [76] 或者视频层面 [55], [57], [63] 的结肠疾病分类、息肉识别 [28], [34], [38], [40], [42], [49], [52]、细粒度息肉分类 [26], [27], [48]、出血情况判断 [25], [72], [78]、异常识别 [60]、盲肠识别 [36] 以及术前评估 [30] 等任务。

- **检测数据集**为感兴趣的目标提供分类和定位标签，涵盖包括结肠疾病 [55], [58], [63]、辅助器械 [35], [43], [43]、息肉 [24], [49], [60], [65], [73]、内窥镜伪影 [37], [46] 以及胃肠道疾病 [44], [57]。此外，各类竞赛的组织通过建立数据收集和模型评估的共享平台，加速了结肠镜检查领域的发展，极大地推动了肠道疾病检测 [41] 和息肉检测 [39], [62], [70] 等研究方向的进步。

- **分割数据集**在结肠镜检查研究中的数据来源于两个方面。第一类数据主要基于真实图像，主要用于对诸如息肉 [17], [31], [33], [34], [45], [50]、克罗恩病 ([54] 以及手术辅助工具 [51] 等单一目标进行像素级分割。还有一些数据集，比如 BKAI-Small/Large [47]，为肿瘤性和非肿瘤性息肉提供实例级掩码。另外一部分数据集来自组织良好的竞赛，例如息肉分割数据集 [20]–[23], [62]、胃肠道疾病分割数据集 [41] 或者是针对现有数据库的扩展：提供逐像素掩码（如针对息肉的 [39], [56]、针对结直肠疾病的 [63]）或者提供涂鸦标签（如针对息肉的 [60], [68]）。第二类数据则来源于合成数据，比如 SinGAN-Seg [64] 利用生成模型合成息肉分割图像。

- **视觉语言 (VL) 数据集**至今仍相对稀缺，目前仅有两个已知专为该目的构建的数据集。MEDVQA-GI [66] 是首个涵盖三项视觉语言任务的数据集，包括视觉问答、视觉问题生成以及视觉定位问答。Kvasir-VQA [75] 则从现有的数据集 [44], [51] 中收集了 6,500 个问答对，用于支持胃肠道诊断相关任务，例如图像描述与视觉问答等。

3.2 讨论

基于对上述 63 个数据集的回顾，本研究提出了一些以数据为中心的观察，这些观察结果有望为未来研究提供启发。

- **数据粒度**仍有提升空间，以更好地支持病情理解与疗效评估。(a) 超过四分之一包含息肉的数据集提供了细粒度分类标签，但往往缺乏足够的细节。例如，BKAI-Small/-Large [47] 提供了肿瘤性与非肿瘤性息肉的实例级注释；ColonoscopicDS [27] 在视频层面对病变进行了增生性、锯齿状和腺瘤性分类；SUN-database [55] 提供了细粒度标签，包括息肉的尺寸（高度与宽度）、形态（有蒂、无蒂、扁平）及其解剖位置（如直肠、乙状结肠）。此外，多个以数据为核心的研究方向仍有待深入探索，例如：病变的时序演化、标签粒度的进一步提升、病变严重程度的分级标注，以及实例级别的目标注释等。(b) 另一个被普遍忽视的问题是标签的正交性假设，即默认各类别彼此独立。当前研究很少探讨类间潜在的关联性，例如炎症性肠病常伴随糜烂症状、克罗恩病与瘘管并发症同存、结直肠

癌常伴随出血表现。未来的研究应更系统地考虑因果关系 [89] 与共病情况 [90]，以更有效地建模这些潜在关联。

- **数据多样性**对于构建公平且可靠的模型而言至关重要。有三个方面值得关注。(a) 首先，由于罕见病例数量稀少且需要专业知识进行标注，当前针对罕见结直肠疾病的数据集仍较为有限。比如，据估计在美国每十万人中约有 58 至 241 人患有克罗恩病 [91]，但迄今为止只有三个数据集 [53], [54], [59] 涉及该疾病。这种不均衡分布导致数据驱动的模式在常见病例中的性能往往优于在罕见或新发病例中的表现。因此，增加对罕见胃肠道疾病的关注，有望提升模型处理长尾问题 [92] 与开放词汇任务 [93] 的能力。(b) 其次，结肠镜检查中的多模态研究尚处于早期阶段，可用于分析的数据仍较为有限 [66], [75]。因此，收集多样化的患者信息（如年龄、性别、饮食习惯）并结合专家解读（如临床报告、用药建议）将有助于推动个性化、低副作用的结肠镜检查实践 [94]。

- **数据不一致性**问题主要源于两个方面。(a) 一方面，由于临床经验、专业水平以及对细微差异的感知存在差异，专家间的解读结果往往不一致，从而引入主观判断与标注的不确定性。例如，ColonoscopicDS [27] 为每个样本提供了多位专家和初学者的原始诊断标签，这反映出标注者之间的潜在一致性与差异性。此外，SUN-SEG [63] 公布了其标注流程中被弃用的分割掩码，突显出在确定一致息肉边界方面所面临的挑战。(b) 另一方面，现有结肠镜数据集通常服务于特定研究目的，因而可能忽视其他目标或将其归类为背景。例如，Nerthus [29] 关注肠道准备质量评估，却未涉及息肉等病变信息；GastroVision [67] 中部分类别并非互斥，例如某个病例既可能被标注为“辅助工具”，也可归为“管腔内有血”。此外，Kvasir-Instrument [51] 仅对医疗器械进行分割标注，忽略了息肉等其他目标；而 Kvasir-SEG [39] 虽提供息肉标注，却未涵盖器械等对象。以上问题提示我们应进一步探索如何在部分标签学习 [95]、带噪声学习 [96] 或缺失标签学习 [97] 策略下实现鲁棒的模型训练。

4 重新审视结肠镜模型

本节系统回顾了自 2015 年以来发表于顶级会议与期刊的总共 137 篇结肠镜场景感知相关的深度学习研究成果。首先，我们梳理了该领域广泛关注的三大任务，包括 18 个分类模型 (§4.1 章节)、24 个检测模型 (§4.2 章节) 以及 86 个分割模型 (§4.3 章节)。如图 3 所示，这些模型在架构设计上可归纳为五类子结构。最后，我们在 §4.4 章节中补充讨论了九个与视觉语言 (VL) 相关的模型，展示了多模态方法在结肠镜场景中的最新探索。

4.1 分类模型

- **输入阶段**。表 2 汇总了用于每个深度模型的训练和测试数据。我们注意到，许多用于结肠镜分类模型 [107], [110], [113]–[115], [117], [119] 均基于内部数据构建，这在一定程度上限制

表 2

结肠镜检查分类模型概述. 数据集: CU = CU-ColonDB [26], CDS = ColonoscopicDS [27], Private = private data, HK=HyperKvasir [44], KC=Kvasir-Capsule [57]. 主干网络: CaffeNet [98], D-121 = DenseNet121 [99], R-12/-18/-50/-101 = ResNet12/18/50/101 [100], ViT-S16/ViT-B16 [101], MobV2 = MobileNetV2 [102], R50-Att = 带有注意力模块的 ResNet50 [103], C3D [104], Inc-v3 = Inceptionv3 [105], I3D [106]. “Customised” 意味着针对当前任务修改的基础网络, 或者是一个独立于基础网络选择的模型. 头部网络: 由全连接 (FC) 层和支持向量机 (SVM) 层实现的分层器, 或者使用 ℓ^2 范数来衡量输入和输出之间的差异. 架构: 如图 3 中所示的架构. 学习方式: 学习策略, 如全监督 (FS)、半监督 (SS)、无监督 (US) 和弱监督 (WS). 为简单起见, 除非另有说明, 以下表格使用一致的缩写.

	模型	出版物	核心设计	训练数据集	测试数据集	主干网络	架构	头部网络	学习方式	链接
图像处理模型	Zhang 等人 [26]	JBHI'16	领域迁移学习	CU, CDS	CU, CDS	CaffeNet	BF#1	SVM	FS	-
	RIIS-DenseNet [107]	MICCAI'18	旋转不变性、相似性约束	Private	Private	D-121	SF	FC	FS	-
	FSAD-Net [108]	MICCAI'20	交互信息最大化	Private	Private	D-121	BF#2	FC	US	Link
	Gammulle 等人 [109]	MICCAI'20	关系映射	Kvasir [29], Nerthus [30]	Kvasir [29], Nerthus [30]	R-50	MF#1	FC	FS	-
	ADGAN [38]	ISBI'20	双重对抗学习	Liu 等人 [38]	Liu 等人 [38]	Customised	BF#2	ℓ^2	US	-
	Carneiro 等人 [110]	MedIA'20	模型不确定性 & 校准	Private	Private	D-121	SF	FC	FS	-
	SSL-WCE [111]	MedIA'20	适应性聚集注意力	CAD-CAP [42]	CAD-CAP [42]	D-121	BF#2	FC	SS	Link
	PolypsAlign [48]	MICCAI'21	师生对齐	CPC-Paired [48]	CPC-Paired [48]	R-50	BF#2	FC	FS	Link
	CPC-Trans [112]	MICCAI'22	跨模态的表示方式一致性	CPC-Paired [48]	CPC-Paired [48]	ViT-S16	BF#2	FC	FS	Link
	FFCNet [113]	MICCAI'22	频域学习	Private	Private	R-18	SF	FC	FS	Link
DLGNet [114]	MedIA'23	高斯混合模型	Private	Private	R-18	BF#2	FC	FS	Link	
Yue 等人 [115]	TIM'23	阶级失衡损失	Private, HK	Private, HK	MobV2	SF	FC	FS	Link	
DAFON [116]	ESWA'24	小样本开放集学习	Kvasir-Capsule [57]	Kvasir-Capsule [57]	R-12	BF#2	FC	FS	-	
SSL-CPCD [74]	TMI'24	复合前置任务 - 类别判别	LIMUC [69]	Private, LIMUC [69]	R50-Att	BF#2	FC	FS	Link	
视频处理模型	BseNet [117]	MICCAI'18	无监督深度估计、LSTM [118]	Private	Private	C3D	SF	FC	FS	-
	Byrne 等人 [119]	Gut'19	实时评估系统	Private	Private	Inc-v3	SF	FC	FS	-
	Tamhane 等人 [120]	MICCAI'22	以视觉 transformer 为基础	Private	Private	ViT-B16	SF	FC	FS	-
	Tian 等人 [60]	MICCAI'22	多实例学习	WVAD [60]	WVAD [60]	I3D	SF	FC	WS	Link

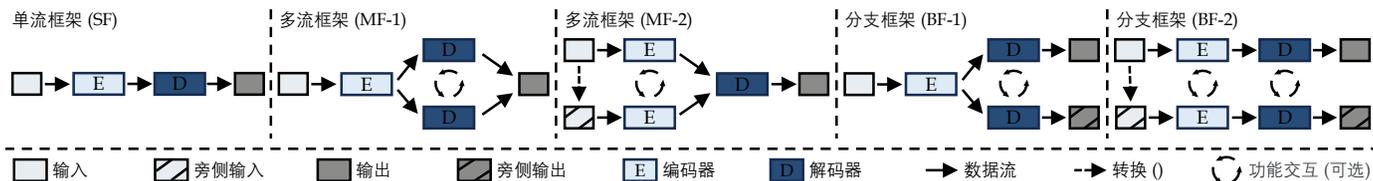


图 3. 基于深度的架构廊. 单流框架 (SF) 的特点是具有单个输入和输出, 且数据流是顺序的. 多流框架预测单个输出, 但涉及并行处理流, 这些并行处理流或在解码阶段 (MF#1), 或在编码阶段 (MF#2). 分支框架是对多流框架的扩展, 可从单个输入 (BF#1) 或多个输入 (BF#2) 产生多个输出. 这些旁侧输出通常会受到额外监督信号的监督, 例如边界线索.

了领域认可基准测试的形成. 这个问题源于各个研究所聚焦的分类目标存在差异, 例如一些工作致力于区分白光和窄带成像图像对中的息肉 [48], [112], 另一些则关注于从图像 [107] 或视频序列 [60] 中识别息肉. 此外, 也有研究聚焦于息肉尺寸的评估 [117]、结肠疾病的自动识别 [113] 或结肠镜检查中的解剖学标志识别 [120].

• **处理阶段.** 本文从两个关键属性出发探讨数据流管理策略: (a) 主干网络: 早期的模型 [107], [109] 通常会采用在 ImageNet 数据集 [4] 上预训练好的卷积主干网络 (例如 ResNet [100] 和 DenseNet [99]), 而近期研究则探索了更多的替代方案, 比如文献 [112] 中使用了视觉 Transformer, 文献 [115] 中使用了轻量级网络结构. 另一种策略是 SSL-CPCD [74], 它包括对模型进行预训练以生成特定领域的表征, 然后将其泛化到各种下游感知任务. (b) 架构: 在结构上呈现出多样性, 如图 3 所示. 一个基本的思路是使用单流框架 (SF), 该框架基于置信度校准 [110] 或三维卷积网络 [60], [117] 按顺序处理视觉特征. 此外, Gammulle 等人 [109] 提出了一种双解码流策略, 用于分层特征编码, 文研究将其归类为 MF#1. 第三, 为增强预测的可靠性, 部分模型采用分支架构进行多目标学习, 典型策略包括并行特征流融合 [26]、类间高斯损失 [114] 以及从全局到

局部的一致性优化 [112].

• **输出阶段.** (a) 预测头: 早期的模型 [26] 使用两个支持向量机 (SVM) 层来实现三类息肉的分类. 而现代方法通常采用全连接层作为最终的分层器, 因其结构简单且灵活. 一个特例是 ADGAN [38], 该方法采用生成对抗网络结构, 通过计算输入和输出图像之间的 ℓ^2 范数差值来识别异常. (b) 监督策略: 目前大多数模型仍依赖于全监督学习, 即在带有预先标注类别的数据上进行训练. 然而, 为提高数据利用效率, 部分研究开始探索更高效的学习范式, 包括半监督学习 [111]、弱监督学习 [60] 以及无监督学习 [38], [108] 等.

• **评论.** 本文从三方面总结当前分类模型的研究现状. (a) 像状态空间模型 [150] 这样新颖的视觉主干网络仍未得到充分探索. 此外, 将视觉语言模型 (如 CLIP [151]) 引入分类任务, 可能带来意想不到的结果. (b) 多类别分类的基准测试仍未得到充分研究. Kvasir 系列数据集 [29], [44], [57] 为进一步的研究提供了有价值的资源基础. 本文将在 §5.1 章节中进一步探讨这些公共数据集在协同作用方面的潜力. (c) 在结肠镜检查领域已经出现了一些新的任务设置. 例如, Tian 等人 [60] 从分布外的视角识别结肠镜检查视频中的异常帧. DAFON [116] 在少样本学习框架内解决了一个开放集分类问题.

表 3

结肠镜检查检测模型概述. 数据集: C6=CVC-ClinicDB [22], ETIS=ETIS-Larib [21], ASEI [35], C3=CVC-ColonDB [20], KUMC [58], LDPV=LDPolyVideo [49], SUN=SUN-database [55], PL=PICCOLO [40], KID=KID1&2 [32], CDS [27], KSe=Kvasir-Sessile [56], ASU=ASU-Mayo [23], CDB=CVC-ClinicVideoDB [28], ES=EndoSceneStill [31], CU [26], ACP=ACP-ColonDB530 [43]. **主干网络:** R-34/-50/-101 [100], CDN-53=CSPDarkNet53 [121], DN-53=DarkNet-53 [122], EffDet-D0=EfficientDet-D0 [123], AlexNet [124], V-16=VGG16 [125], R-50v2= ResNet50V2 [126]. **workflow (WF):** 一阶段 (OS) 或两阶段 (TS) workflow. **NMS:** 非极大值抑制. **EC:** 边缘线索.

	模型	出版物	核心设计	训练数据集	测试数据集	主干网络	WF 架构	NMS	EC	Sup	链接
图像处理模型	Yang 等人 [127]	TIM'20	并行检测 & 分割	Private, C6, ETIS	Private, C6, ETIS	R-50	TS BF#1			FS	-
	ConsolidatedPolypDA [128]	MedIA'21	高斯傅里叶域自适应	C6	ETIS, ASEI	R-101	TS BF#2			US	Link
	MDeNetplus [129]	MedIA'21	二维高斯形状预测	C6	C3, ETIS	R-34	OS MF#1	✓		FS	-
	FedInI [130]	MICCAI'22	联邦学习、结构因果模型	KUMC	KUMC	R-101	TS BF#2	✓		FS	Link
	Pacal 等人 [131]	CIBM'22	改进的 YOLOv3 [122]/v4 [121]	SUN, PL	SUN, PL, ETIS	CDN-53/DN-53	OS BF#1	✓		FS	Link
	SMPT++ [132]	TMI'22	无源域自适应	Private, C6, ETIS, ASEI, KID	Private, C6, ETIS, ASEI, KID	R-101	OS BF#1	✓		US	Link
	FRCNN-AA-CIF [133]	CIBM'23	注意力模块 & 上下文增强	Private	Private	R-101	TS BF#1	✓		FS	-
	Haugland 等人 [134]	MI'23	模态信息融合	Private, PL, CDS	PL, KUMC	EffDet-D0	OS BF#2	✓		FS	-
	SCAN++ [135]	TMM'23	增强语义条件自适应	C6, ASEI	C6, ASEI	R-101	OS BF#2			FS, US	Link
	TFCNet [136]	CIBM'24	细粒度特征补偿	C6, KUMC, LDPV	C6, KUMC, LDPV, KSe	CDN-53	OS BF#1			FS	-
DUT [137]	TNNLS'24	解耦无偏见教师	C6, ASEI, Private	ASEI, Private	R-101	OS BF#2			US	Link	
视频处理模型	Tajbaksh 等人 [138]	IPMI'15	补丁描述符 & 边缘分类	Private	Private	AlexNet	TS BF#1	✓		FS	-
	Tajbaksh 等人 [23]	TMI'15	扩展到 [138]	C3	C3, ASU	AlexNet	TS BF#1	✓		FS	-
	Yu 等人 [139]	JBHI'16	在线与离线集成	ASU	ASU	Customised	OS MF#2			FS	-
	Mo 等人 [140]	ICPR'18	基于 Faster R-CNN [141] 构建	CDB	C6, C3, CDB, ES	V-16	TS BF#1	✓		FS	-
	Qadir 等人 [142]	JBHI'19	时间依赖性	ASU, C6	ASU, CDB	V-16	TS BF#1	✓		FS	-
	AIPDT [143]	MICCAI'20	并行检测与跟踪	Private, CDB	CDB	DN-53, AlexNet	OS BF#2			FS	-
	AI-doscopist [43]	NPJDM'20	时空融合	C6, ETIS, C3, ASU, CU, ACP	C6, ETIS, C3, ASU, CU, ACP	R-50	OS BF#2	✓		FS	-
	STFT [144]	MICCAI'21	时空特征变换	ASU, CDB	ASU, CDB	R-50	OS BF#2			FS	Link
	Yu 等人 [145]	AIM'22	实例跟踪头 (即插即用)	Private, C6, CDB	Private, CDB, ETIS	V-16	OS BF#2	✓		FS	-
	EMSEN [146]	TII'22	可解释的多尺度 Shapley 解释	CDS	CDS	Customised	OS BF#2			FS	-
YONA [147]	MICCAI'23	特征对齐 & 对比学习	SUN, CDB, LDPV	SUN, CDB, LDPV	R-50	TS BF#2	✓		FS	Link	
Intrator 等人 [148]	MICCAI'23	自监督息肉再识别	Private	Private	R-50v2	OS MF#2			US	-	
V2I-DETR [149]	arXiv'24	视图到图像的知识提炼	SUN	SUN	R-50	OS BF#2			FS	-	

4.2 检测模型

• **输入阶段.** 检测模型通过对图像中的目标进行分类与边界框定位, 为手术干预和规划提供支持. 这类模型关注的目标类型多样, 常见任务包括: 在图像 [127]–[129], [131]–[135], [137]/视频 [23], [138]–[140], [142]–[149] 中识别息肉, 以及定位多种发现, 例如出血、息肉和辅助器械 [43].

• **处理阶段.** 对于分析过程, 有三个关键的配置项. (a) 主干网络: 网络初始化有两种常见策略. 第一类策略 [43], [127]–[129], [132], [133], [135], [137], [144], [147]–[149] 依赖于在 ImageNet 数据集 [4] 上预训练的 ResNet 系列网络 [100]. 第二类策略基于训练良好的目标检测器, 例如文献 [131], [143] 使用了 DarkNet 系列 [121], [122], 而文献 [134] 采用了 EfficientDet-D0 [123]. (b) 工作流程: 检测模型通常建立在通用网络架构之上. 在表 3 的“工作流程 (WF)”一栏中, 我们根据模型的处理阶段对其进行了分类. 两阶段工作流程将检测解耦为候选区域生成和目标分类两个阶段, 例如基于 Faster R-CNN [141] 的模型 (相关研究见文献 [128], [133], [140], [142]). 一阶段模型更注重推理速度与结构简洁性, 通常通过单次前向传播完成检测操作. 例如, 一些研究 [43], [131], [143] 基于 YOLO 框架 [121], [122], 而 Yu 等人 [145] 使用了 SSD 框架 [152]. (c) 架构: 检测模型负责预测目标的类别和空间坐标, 通常在分支框架 (BF#1/BF#2) 中实现, 如图 3 所示. 有两个特例 [129], [139] 对多流框架的设计进行了调整, 先生成像素级的注意力区域, 然后将其转换为边界框.

• **输出阶段.** (a) 后处理技术被用于消除重复的预测结果并选择最相关的目标, 其中非极大值抑制 (NMS) 是一种被广泛采用的方法 [43], [131]–[134], [140], [142], [145], [147]. (b) 辅助信息的引入可以提升预测的可靠性, 例如在 [23], [129], [138]

中使用的边缘线索为目标检测提供几何模式. (c) 监督策略目前以全监督学习策略为主, 常见形式包括区域级标签 [127], [131], [133], [134], [140], [142]–[146], [149] 和逐像素标签 [23], [129], [138], [139], 以及引入边界框辅助的对比学习 [147]. 另有一些模型 [128], [132], [135], [137] 探索了无监督域适应技术, 以实现跨结肠镜设备间的息肉检测迁移能力.

• **评论.** 我们对上述回顾提出以下几点观察结论 (a) 首先, 大多数模型主要聚焦于检测息肉, 而其他结肠镜检查结果受到的关注相对较少. 本研究鼓励进一步探索公开的多目标数据集 [35] 或多中心数据集 [79]. (b) 其次, 学习策略尚未得到充分探索. 通用的检测模型, 尤其是采用弱监督方式的模型 [256] 为该领域提供了有价值的参考, 因为它们对医学专家详细标注的需求较少, 所以可能更具可行性, 且具有更高的成本效益. (c) 再次, 除了经典的基于卷积的检测框架之外, 近年来新兴的方法也为该领域带来了新的发展机遇, 如基于 Transformer 架构的 DETR [257] 以及基于扩散的 DiffusionDet [258]. 此外, 探索跨任务协同作用同样具有前景, 已有三个基于视频的模型 [143], [145], [148] 证明了在统一息肉检测和跟踪任务方面的有效性. (d) 最后, 尽管如 SUN-database [55] (>158K 个样本) 和 LDPolyVideo [49] (>900K 个样本) 等数据集的规模相对较大, 但该领域仍然缺乏统一的、标准化的评估基准.

4.3 分割模型

与上述两个研究主题相比, 分割研究已相当成熟, 可参见表 4.

• **输入阶段.** 大多数分割模型主要专注于单个目标 (即息肉), 并普遍采用二值分割范式. 在模型的开发和比较过程中, 研究者通常遵循已被广泛认可的 PraNet 测试基准 [176]. 一个例外是 AFP-Mask 模型 [197], 其提供了一个无锚点的框架来分割

表 4

结肠镜检查分割模型概述. 数据集: C6 [22], ES [31], KS=Kvasir-SEG [39], C3 [20], ETIS [21], HK [44], ASU [23], CDB [28], BKAI=BKAI-Small [47], KSe [56], GI=GIANA [153], SUN-S [63], PG=PolypGen [70], K-I=Kvasir-Instrument [51]. 主干网络: ResUNet [154], R-18/-34/-50/-101 [100], R-50v2 [126], R2-50=Res2Net-50 [155], V-16 [125], DeiT [156], Eff-B4=EfficientNet-B4 [157], DLV3+=DeepLab V3+ [158], PB2/3/5=PVTv2-B2/-B3/-B5 [159], CvT [160], MiT-B2 [161], CMLP=CycleMLP [162], P-DETR= Point DETR [163], D-121=DenseNet121 [99], MN=MSCAN [164], Swin-T [165], SAM [166], SAM2 [167], ViT-B16 [101], DLV2=DeepLabV2 [168], HR-W48=HRNet-W48 [169], CN-T=ConvNeXt-Tiny [170], SFB3=SegFormer-B3 [161], M2Former=Mask2Former [171]. Edge-sensitive 通过显式 (EX) 使用边缘图作为监督/输入并隐式探索边缘感知表示 (IM#1) 或边缘感知不确定性 (IM#2) 进行分析.

模型	出版物	核心设计	训练数据集	测试数据集	主干网络	架构	边缘	学习方式	链接
Yuan 等人 [172]	JBHI'17	弱自下而上 & 强自上而下显著性	Private	C6	Customised	BF#1	-	US	-
SFA [173]	MICCAI'19	区域 & 边界约束	ES	ES	Customised	BF#1	EX	FS	Link
ResUNet++ [174]	ISM'19	增强型深部残差 UNet [154]	C6, KS	C6, KS	ResUNet	MF#1	-	FS	Link
ACSNet [175]	MICCAI'20	自适应上下文选择	ES, KS	ES, KS	R-34	BF#1	-	FS	Link
PraNet [176]	MICCAI'20	反向注意力, 并行分解器	C6, KS	C6, ES, KS, C3, ETIS	R2-50	BF#1	EX, IM#1	FS	Link
UI-CNN [177]	MedIA'20	蒙特卡罗引导反向传播	ES	ES	V-16	MF#1	IM#2	FS	Link
ThresholdNet [45]	TMI'20	置信度引导的流形混合增强	ES	ES, WCE-Polyp [45]	R-101	BF#2	EX, IM#1	FS	Link
SCR-Net [178]	AAAI'21	语义校准 & 精炼	KS	KS	Customised	MF#1	-	FS	Link
BI-GCN [179]	BMVC'21	边界感知图卷积	C6, KS	C6, ES, KS, C3, ETIS	R2-50	BF#1	EX	FS	Link
FDSEMI [180]	ICCV'21	合作性 & 对抗性学习	C6, KS	C6, KS	Customised	BF#2	IM#1	SS	Link
CCBANet [181]	MICCAI'21	级联上下文 & 平衡注意力	C6, ES, KS	C6, ES, KS	R-34	BF#1	IM#1	FS	Link
CCD [182]	MICCAI'21	约束对比分布学习	HK, Liu 等人 [38]	HK, Liu 等人 [38]	R-18	BF#2	-	US	Link
HRENet [183]	MICCAI'21	困难区域增强	C6, KS	C6, KS, C3	R-34	BF#1	EX	FS	Link
LOD-Net [184]	MICCAI'21	面向可学习性的衍生工具	C6, KS	C6, ES, KS, C3, ETIS	R-50	MF#2	EX	FS	Link
MSNet [185]	MICCAI'21	多尺度减法网络	C6, KS	C6, ES, KS, C3, ETIS	R2-50	MF#1	EX, IM#1	FS	Link
SANet [186]	MICCAI'21	颜色交换 & 概率校正	C6, KS	C6, ES, KS, C3, ETIS	R2-50	MF#2	-	FS	Link
TransFuse [187]	MICCAI'21	transformers 和 CNNs 融合	C6, KS	C6, ES, KS, C3, ETIS	R-50v2, ViT-B16	BF#2	EX	FS	Link
EndoUDA [50]	MICCAI'21	领域自适应, 变化的自编码器训练	Celik 等人 [50]	Celik 等人 [50]	Eff-B4	BF#2	EX	FS, US	Link
UACANet [188]	MM'21	不确定性增强的上下文注意力机制	C6, KS	C6, ES, KS, C3, ETIS	R2-50	BF#1	EX, IM#2	FS	Link
Jha 等人 [56]	JBHI'21	具有随机掩码 & 测试时间增强的 ResUNet++ [174]	C6, C3, ETIS, KS, ASU, CDB	C6, C3, ETIS, KS, ASU, CDB	ResUNet	MF#1	-	FS	Link
MPA-DA [189]	JBHI'21	互原自适应网络	ES, KS	ETIS	R-101	BF#2	IM#2	FS, US	Link
DW-HieraSeg [190]	MedIA'21	分层分割, 动态加权	ES	ES	DLV3+	BF#1	-	FS	Link
ICGNet [191]	IJCAI'22	基于上下文的反向轮廓引导	ES, KS	ES, C3, KS	R-34	BF#1	EX	FS	-
BoxPolyp [192]	MICCAI'22	带边界框标签的分割	C6, KS	C6, ES, KS, C3, ETIS	R2-50, PB2	BF#1	-	WS	Link
LDNet [193]	MICCAI'22	动态内核生成 & 更新	Private, C6, KS	Private, C6, KS, C3, ETIS	R2-50	BF#1	-	FS	Link
PPFormer [194]	MICCAI'22	息肉引导的自注意力机制, 局部-全局策略	C6, KS	C6, ES, KS, C3, ETIS	V-16, CvT	BF#1	EX	FS	-
SSFormer [195]	MICCAI'22	逐步局部 & 全局特征聚合	C6, KS	C6, C3, ETIS, KS	MiT-B2	MF#1	-	FS	Link
TRFR-Net [196]	MICCAI'22	任务相关特征补充	C3, ETIS, KS	C3, ETIS, KS	R-34	BF#2	-	FS, US	Link
AFP-Mask [197]	JBHI'22	无目标实例分割	Private, GI	Private, C6, ETIS	Customised	BF#1	-	FS	-
BCNet [198]	JBHI'22	跨层融合, 双边边界提取	KS	C6, ES, KS	R2-50	BF#1	EX	FS	-
BSCA-Net [199]	PR'22	切片片上上下文注意力机制	C6, KS	C6, ES, KS, C3, ETIS	R2-50	BF#1	EX	FS	Link
Polyp-Mixer [200]	TCSVT'22	基于上下文的感知神经网络	C6, KS	C6, KS, C3, ETIS	CMLP	BF#1	-	SS	Link
ACL-Net [201]	AAAI'23	亲和对比学习	C6, KS	C6, ES, KS, C3, ETIS	R-50	BF#2	-	FS	Link
WS-DefSegNet [68]	CVPRW'23	可变形 transformer, 稀疏前景损失	W-Polyp [68]	C6, KS, C3, ETIS, ES	R2-50	BF#1	-	WS, SS	Link
WeakPolyp [202]	MICCAI'23	掩码到边界框转换, 尺度一致性	SUN-S, Private	SUN-S, Private	PB2	BF#2	-	WS	Link
PEtNet [203]	MICCAI'23	高斯概率引导语义融合	C6, KS	C6, ES, KS, C3, ETIS	PB2	BF#1	-	FS	Link
s ² ME [204]	MICCAI'23	空谱互教, 集成学习	SUN-S	C6, KS, SUN-S, PG	Customised	BF#1	-	WS	Link
Su 等人 [205]	MICCAI'23	特征传播 & 聚合	C6, KS	C6, ES, KS, C3, ETIS	PB2	BF#1	EX	FS	Link
Polyp-PVT [206]	AIR'23	改进的金字塔视觉 transformer	Private	C6, ES, KS, C3, ETIS	PB2	BF#1	EX	FS	Link
RPA-Net [207]	IPMI'23	由粗到细的自监督	Private	C6, ETIS, KS	R-101	BF#2	-	FS, US	-
FEcNet [208]	JBHI'23	反馈增强门控网络	C6, KS	C6, ES, KS, C3, ETIS	R2-50	BF#1	EX	FS	-
BS-Loss [209]	JBHI'23	带位置约束的边界敏感损失	K-I	K-I	Customised	MF#1	EX	FS	Link
Point SEGRNet [210]	MedIA'23	多点对称一致性	C6, ETIS	C6, ETIS	P-DETR	BF#2	-	FS, WS, SS	-
DGNet [211]	MIR'23	深度梯度学习	C6, KS	C3, ETIS	Eff-B4	BF#1	EX	FS	Link
CFA-Net [212]	PR'23	跨层特征融合, 边界聚合	C6, KS	C6, ES, KS, C3, ETIS	R2-50	BF#1	EX	FS	Link
ColnNet [213]	TMI'23	统计注意力, 异常边界近似	C6, KS	C6, ES, KS, C3, ETIS	D-121	BF#1	EX	FS	-
FANet [214]	TNNLS'23	反馈注意力, 行程长度编码	C6, KS	C6, KS	Customised	MF#2	-	FS	Link
MCANet [215]	arXiv'23	多尺度跨轴注意力机制	C6, KS	C6, ES, KS, C3, ETIS	MN	MF#1	-	FS	Link
PolypNet [216]	AAAI'24	边界敏感注意力机制	C6, KS	C6, KS, C3, ETIS, ES	Swin-T	MF#1	IM#1	FS	Link
EMCAD [217]	CVPR'24	高效多尺度卷积注意力解码器	C6, C3, ETIS, KS, BKAI	C6, C3, ETIS, KS, BKAI	PB2	BF#1	-	FS	Link
Schön 等人 [218]	CVPR'24	SAM [166], 测试时自适应	K-I, CDB, KS	K-I, CDB, KS	SAM	BF#2	EX	WS	-
MH-pFLID [219]	ICML'24	联合学习, 注入-蒸馏范式	Private	Private	Customised	BF#2	-	FS	Link
ASPS [220]	MICCAI'24	SAM [166], 不确定性引导的正则化	C6, KS	C6, ES, KS, C3, ETIS	SAM, MN	BF#2	IM#2	FS	Link
Polyp-Mamba [221]	MICCAI'24	视觉状态空间模型, 语义关系挖掘	C6, KS	C6, ES, KS, C3, ETIS	VMamba [222]	MF#1	EX	FS	-
QueryNet [223]	MICCAI'24	分割 & 检测的统一框架	C6, KS	C6, ES, KS, C3, ETIS	M2Former	MF#2	-	FS	Link
LSSNet [224]	MICCAI'24	局部 & 浅层特征补充	C6, KS	C6, ES, KS, C3, ETIS	PB2	BF#1	EX	FS	Link
BSBP-RWKV [225]	MM'24	各向异性扩散, RWKV [226]	KS	KS	Customised	BF#1	EX	FS	-
CFATransUNet [227]	CIBM'24	通道级融合注意力机制与 Transformer	C6, KS	C6, KS	PB3	BF#1	-	FS	Link
PolypMixNet [228]	CIBM'24	一致性正则化, 软伪标签	C3, C6, KS, ETIS	C3, C6, KS, ETIS	R-34	BF#1	-	SS	Link
RGI-Aug [229]	JBHI'24	随机全局光照增强	C3, C6, ETIS, KS	C3, C6, ETIS, KS	Customised	BF#2	-	FS	Link
EMTS-Net [230]	JBHI'24	多任务协同网络	C6, KS	C6, ES, KS, C3, ETIS	R2-50	BF#1	-	FS	-
MSDE-Net [231]	JBHI'24	多尺度双编码网络	K-I	K-I	R-34, DeiT	MF#2	-	FS	-
Polyp-OD [232]	MIR'24	分布于建模之外, 潜在的标准化	SUN-S	SUN-S, C6, C3, ETIS, KS	ViT-B16	MF#2	-	US	Link
MedSAM [233]	NComs'24	SAM [166], 跨器官/模态调谐	Hybrid datasets	Hybrid datasets	SAM	MF#2	-	WS	Link
FOBS [234]	TCSVT'24	多级边界增强框架	KS, ES	KS, ES, ETIS, C3	DLV3+	BF#2	EX, IM#1	FS	Link
DCL-PS [235]	TMI'24	领域互动对比学习, 自训练	ETIS, HK, ES, KS	ES, KS	DLV2	BF#2	-	FS, US	Link
Gao 等人 [236]	TMI'24	语境学习, 双重相似性检验	C3	C3	SAM	BF#2	-	WS	Link
SliceMamba [237]	arXiv'24	视觉状态空间模型, 双向切片扫描	C6, KS	C6, KS	Customised	MF#1	-	FS	-
ProMamba [238]	arXiv'24	视觉状态空间模型, 快速分割	C6, KS	C6, KS, C3, ETIS, ES, BKAI	Vim [150]	MF#2	-	WS	-
SAM2-UNet [239]	arXiv'24	SAM2 [167], 基于适配器的调谐	C6, KS	C6, ES, KS, C3, ETIS	SAM2	BF#1	EX	FS	Link
U-KAN [240]	AAAI'25	U 型 Kolmogorov-Arnold 网络 [241]	C6	C6	Customised	MF#1	-	FS	Link
Puyal 等人 [242]	MICCAI'20	基于混合的时间相关性 2D/3D CNNs	Private, KS	Private	R-101	MF#2	-	FS	-
PNS-Net [243]	MICCAI'21	归一化自注意力, 渐进式学习	C6, C3, ASU, KS	C6, C3	R2-50	BF#2	-	FS	Link
SSTAN [244]	MICCAI'22	时-空注意力机制	LDPolyVideo [49]	LDPolyVideo [49]	ResUNet	BF#2	-	SS	Link
TCCNet [245]	IJCAI'22	时间一致性, 无上下文损失	C6, C3	C6, C3, ETIS	R2-50	BF#2	EX, IM#1	SS	Link
Puyal 等人 [246]	MedIA'22	使用最佳设置扩展 [242]	Private, KS	Private, SUN	R-101	BF#2	-	FS	-
PNS+ [63]	MIR'22	通过从全局到局部的学习逐步扩展 [243]	SUN-S	SUN-S	R2-50	BF#2	-	FS	Link
EUVPs [247]	AAAI'24	跨尺度区域链接, 横向尺度对齐	SUN-S, C6	SUN-S, C6	HR-W48	BF#2	-	FS	Link
LGRNet [248]	MICCAI'24	循环邻域传播, 希尔伯特选择性扫描	C6, C3, SUN-S	C6, C3, SUN-S	R2-50	BF#2	-	FS	Link
SALI [249]	MICCAI'24	短期对齐, 长期交互模块	SUN-S	SUN-S	PB5	BF#2	EX	FS	Link
Diff-VPS [250]	MICCAI'24	扩散模型, 对抗性时间推理	SUN-S	SUN-S	SFB3	BF#2	-	FS, US	Link
FlowICBNet [251]	CIBM'24	迭代反馈单元, 帧过滤 & 选择	SUN-S	SUN-S	R2-50	BF#2	-	FS	Link
Drag&Drop [252]	MIR'24	弱监督时间标注器	SUN-S	SUN-S	-	BF#2	-	WS	Link
SSTFV [253]	arXiv'24	自监督编码器, 子分支机制	SUN-S	SUN-S, CDB	R2-50	BF#2	-	US, FS	-
Vivim [254]	arXiv'24	视频状态空间模型, 时-空选择性扫描	KS, ASU, C3, C6	C3, C6	Customised	BF#2	EX	FS	Link
MAST [255]	arXiv'24	孪生 transformer, 混合注意力模块	SUN-S	SUN-S	PB2	BF#2	-	FS	Link

息肉实例. 此外, 近期的一些研究工作 [209], [218], [231] 也专注于在手术过程中对手术器械进行分割.

• **处理阶段.** (a) 主干网络: 针对分割任务的视觉编码器已得到了广泛研究. 一种常见的选择是使用在 ImageNet 数据

集 [4] 上经过预训练的通用主干网络, 例如使用卷积神经网络 (CNN) [176], [215], 视觉 Transformer [206], [216], CNN 与 Transformer 的混合网络 [187], [194], [224], 多层感知机 [200], 状态空间模型 [221], [237], [238], [254], 接受权重键值

表 5

图像息肉分割模型的比较. 使用两个测试集上的结构度量 (S [259]) 和 Dice 系数 (D) 的平均得分 (%) 对模型进行评估, 箱线图显示了它们在测试用例中的一致性和可变性的分布. 每个型号的名称都以灰色 标记.

模型	Kvasir-SEG (100 张测试图片) [39]				CVC-ClinicDB (62 张测试图片) [22]			
	S [# 排名]	0.0 → 0.5 → 1.0	D [# 排名]	0.0 → 0.5 → 1.0	S [# 排名]	0.0 → 0.5 → 1.0	D [# 排名]	0.0 → 0.5 → 1.0
Polyp-PVT [206]	92.51 [#1]		91.74 [#2]		95.00 [#2]		93.68 [#1]	
CFA-Net [212]	92.40 [#2]		91.47 [#4]		95.07 [#1]		93.25 [#2]	
MSNet [185]	92.31 [#3]		90.23 [#7]		94.68 [#3]		91.48 [#6]	
BoxPolyp [192]	92.30 [#4]		91.84 [#1]		93.70 [#6]		91.81 [#4]	
SSFormer [195]	92.21 [#5]		91.71 [#3]		92.87 [#9]		90.60 [#7]	
UACANet [188]	91.67 [#6]		91.21 [#5]		94.30 [#4]		92.63 [#3]	
PraNet [176]	91.50 [#7]		89.82 [#8]		93.68 [#7]		89.90 [#9]	
SANet [186]	91.45 [#8]		90.41 [#6]		93.98 [#5]		91.57 [#5]	
DGNet [211]	90.98 [#9]		89.72 [#9]		93.39 [#8]		90.44 [#8]	
MCANet [215]	90.25 [#10]		89.55 [#10]		91.79 [#10]		89.70 [#10]	
Polyper [216]	90.08 [#11]		89.12 [#11]		91.29 [#11]		88.63 [#11]	
UNet++ [260]	86.21 [#12]		82.08 [#12]		87.33 [#13]		79.42 [#13]	
UNet [261]	85.76 [#13]		81.83 [#13]		89.00 [#12]		82.25 [#12]	
SFA [173]	78.14 [#14]		72.31 [#14]		79.33 [#14]		70.06 [#14]	

(RWKV) [225] 以及 Kolmogorov-Arnold 网络 [225]. 另一种常用策略是直接采用训练良好的感知模型, 例如在研究 [210] 中使用了 Point DETR 模型 [163], 以及在研究 [190], [234], [235] 中应用了 DeepLab 系列模型 [158], [168]. 最近, 研究重心逐渐向可提示 (promptable) 架构转移. 第一类方法是基于基础模型展开, 包括使用位置提示对 SAM 模型 [166] 进行微调 [233], 在测试阶段对 SAM 进行适配 [218], 利用卷积神经网络与 Transformer 的混合网络 [220], 或者将可训练的适配器层融入到 SAM2 的编码器中 [239]. 另一类方法侧重于通过上下文学习实现对未见场景的适应 [236]. (b) 架构设计: 该领域倾向于采用编解码器设计, 因其在分层感知特征方面表现出色. 当前模型通常选择多流或分支框架, 如图 3 所示. 已有大量研究探索了该框架的拓展改进方式, 包括融入残差连接流 [56], [174])、探究跨任务协同机制 [223]、提供额外的边缘线索引导 [211]、使用模型集成策略 [187], [194], [220]、计算隐层统计量 [232]、通过三维卷积 [242], [246]) 或自注意力模块 [63], [243], [249] 探索时空关系, 以及采用师生学习范式 [68], [210]. (c) 边缘敏感分析: 几何模式有助于增强模型区分前景物体和背景的能力. 当前的研究主要沿两条路径展开. 第一种是显式使用从图像梯度中得到的边缘图, 既可直接用于监督 [173], [183], [211], 也可作为辅助输入 [50]. 部分研究还在损失函数中强调对边缘敏感机制的计算, 例如边界加权策略 [194], [239] 和自定义方式 [184], [209]. 第二种是嵌入了对边缘敏感表示 (例如使用反向注意力机制 [176]、形态学算子 [216] 以及减法算子 [45]) 或者通过不确定性量化来建模边缘感知 [177], [188]. 此外, 也有一些方法采用混合策略, 例如在 MSNet 模型 [185] 中同时集成了减法算子与边缘敏感损失项.

• **输出阶段.** 大多数模型以全监督的方式进行训练, 并在不同的解码阶段引入深度监督机制, 如文献 [176], [206] 中所示. 近来, 模型开发逐渐转向更加数据高效的训练策略. 例如, 通过

弱监督的掩码到边界框转换的方法 [202]; 无监督技术则包括基于对比学习 [182]、分布外建模 [232] 以及伪标签监督 [192], [244]. 此外, 混合监督策略也得到了探索, 例如模型 [50], [189], [196], [207] 通常在源域中进行全监督训练, 然后以无监督学习的方式适配到目标域. Point SEGTR [210] 结合了三种不同类型的监督信号以提升模型性能. 部分师生网络也会采用混合监督机制, 如 Ren 等人 [68] 对教师模型采用弱监督信号训练, 而学生模型则通过半监督方式优化.

• **评论.** 为反映该领域的最新进展, 本文在两个主流测试数据集上系统评估了 14 个开源图像分割模型, 详情参见表 5. 所有模型均在 Fan 等人发布的基准测试 [176] 上进行训练. 实验结果表明, 当前的学习主流的学习策略尚未被充分探索, 例如, 弱监督模型 BoxPolyp 在 Kvasir-SEG 数据集上取得了最高的 D 得分 (91.84%), 显示出弱监督策略的潜力. 此外, 尽管部分模型在平均性能指标上表现优越, 其预测稳定性却存在不足. 例如, 在 Kvasir-SEG 上获得最高 S 得分的 Polyp-PVT 模型, 其结果的四分位距明显宽于 SSFormer 等其他模型, 表明其不同图像上的性能波动更大. 基于上述分析, 本文指出了若干个值得关注的未来研究方向. (a) 首先, 现有的黄金基准 [176] 包含的图像样本不到 1500 个, 且仅涵盖单一类别 (息肉). 一般来说, 扩大数据规模和增加数据多样性, 可能是提升模型鲁棒性和泛化能力的自然路径. 这一需求正推动着相关创新, 例如用于减轻专家标注负担的半自动标注工具 [166], [167], 以及通过扩散技术 [262] 和自回归生成技术 [263] 合成高保真度的内容. (b) 除此之外, 无限制扩增数据规模并不可持续. 开发数据高效策略 [95]–[97], 即使用更少或更弱的标签, 对于该领域的普通用户来说更具成本效益. (c) 最后, 为医生提供程序性支持很重要, 包括异常检测、导航规划、风险评估和干预建议等. 在这方面可借鉴其他领域的先进工作 [264].

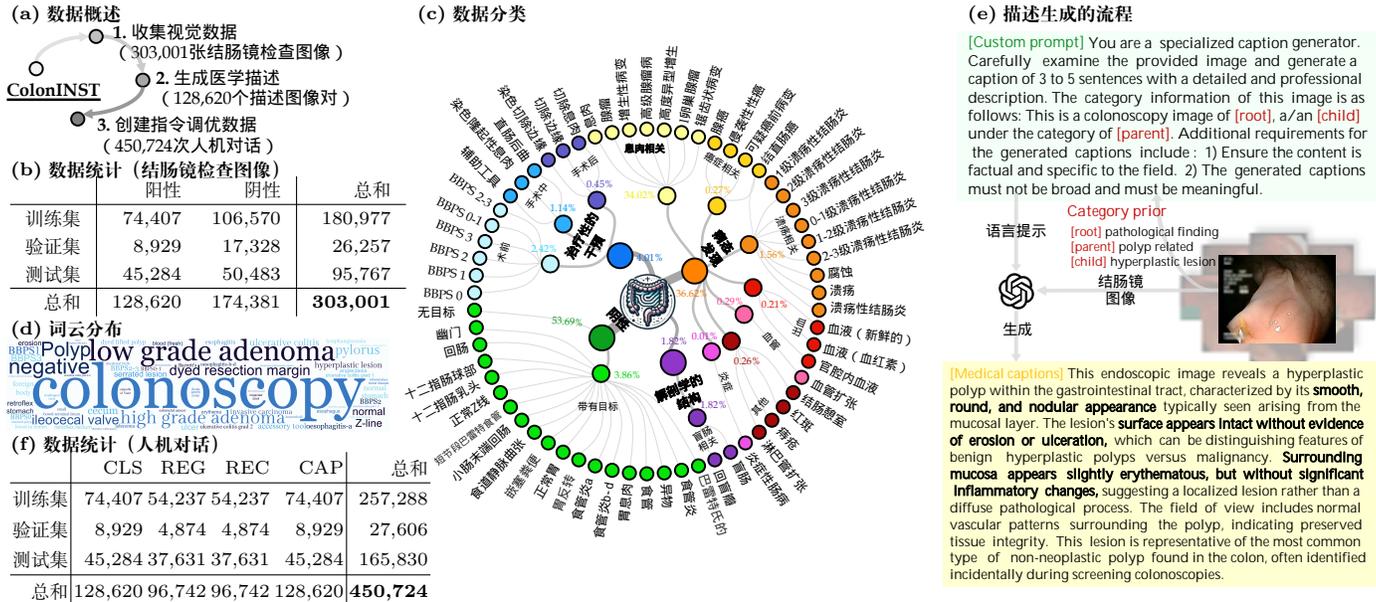


图 4. ColonINST 的细节. (a) 创建用于多模态研究的指令微调数据集的三个步骤. (b) 指定用于训练、验证和测试的结肠镜检查图像数量. (c) 三级类别的数据分类. (d) 按名称大小分类的词云. (e) 使用 GPT-4V 的视觉语言提示模式 [5] 的描述生成流程. (f) 为四个任务创建的人机对话数量.

4.4 视觉语言模型

与上述三个主题相比,多模态研究的参考文献相较有限. 现有的大多数方法采用判别式模型,旨在学习多模态输入之间的决策边界. 已有研究证明该类模型在指称分割任务中的有效性,例如通过将病变属性(如息肉的大小和数量)的文本注意力融入到 U 型网络 [265]、扩散模型 [266] 或者混合网络 [267] 中. 也有其他研究 [268], [269] 基于训练良好的 GLIP 模型 [270] 构建了用于息肉检测的提示工程流程. 此外, SAM 模型是一种可提示架构,进一步推动了近期的研究趋势. 一个典型策略是使用从图像中获取的对象坐标来提示 SAM. 此外, SAM 能够在视觉语言 (VL) 设置下运行,通过图像-文本激活图 [271] 或零样本目标检测模型 [272], [273] 获取位置提示. 在 MEDVQA-GI 竞赛 [66] 中,大多数解决方案都是基于判别的方法,将视觉语言任务建模为分类映射问题,即为图像-文本对分配预定义的标签. 另一种思路是基于生成的解决方案 [274],通过对预训练的 BLIP-2 模型 [275] 进行适配以生成预测结果.

• **评论.** 结肠镜检查的视觉语言 (VL) 研究滞后可能有两个原因. (a) 以数据为中心的问题. 缺乏结构化且高质量的图像-文本,已成为制约研究进展的重要因素. 未来可以从已有的研究思路中获取启发. 首先,可从社交媒体 [276] 和科学文献 [277] 中爬取未标注的图像-文本数据,可用于构建特定领域的基础模型. 其次,借助 GPT-4V [5] 等语言模型可以生成多样的专业描述,为扩展数据的知识空间提供了一种经济且可扩展的解决方案. (b) 以模型为中心的问题. 尽管通用多模态语言模型 (MLM [9], [278]) 已取得了显著进展,但当前结肠镜检查场景中的视觉语言技术仍相对滞后. 这些技术仅使用解码器,将多个任务(如检测、图像描述)整合至一个自回归框架(下

一个标记预测)中. 该类模型具备处理不同长度输入和输出文本的灵活性,且无需为不同任务单独设计特定的预测头.

5 推进用于结肠镜检查的多模态学习

近期,多模态语言模型 (MLMs) 在利用语言能力处理多模态信号方面展现出巨大潜力,尤其是在“感知和解析”视觉信息方面. 指令微调 [279] 在引导多模态语言模型执行符合用户偏好的特定领域任务中起到了关键作用. 本节介绍了推动多模态研究的三项关键举措: 首先,如何构建一个大规模的指令微调数据集 ColonINST (请参见 §5.1 章节); 其次,如何训练一个面向结肠镜检查场景的多模态语言模型 ColonGPT (请参见 §5.2 章节). 最后,我们为结肠镜检查对话任务提供了首个多模态基准测试 (请参见 §5.3 章节),并围绕 ColonGPT 开展了消融实验分析 (请参见 §5.4 章节),最后分享了从实证观察中所获得的经验 (请参见 §5.5 章节).

5.1 已建立的指令微调数据集: ColonINST

图 (4-a) 展示了本研究中用于构建指令微调数据的半自动化工作流程. 该流程包含三个步骤: 首先,从公共数据集中收集结肠镜检查图像; 其次,结合特定类别的提示,与多模态人工智能聊天机器人 GPT-4V [5] 进行交互,为所收集的阳性病例生成医学图像文本描述; 最后,我们对上述数据进行了重组成为指令微调样本对,使得模型能够以对话交互的形式执行四类结肠镜检查任务.

• **数据收集.** 如图 (4-b) 所示, ColonINST 数据集共包含 303,001 张结肠镜检查图像,涵盖 128,620 例阳性病例与 174,381 例阴性病例,分别来自 19 个不同的数据源. 为确保数据完整性并避免数据泄露,我们制定了一套系统的管理规

表 6

指令微调数据集 ColonINST 的详细信息. 对于每项任务, 本文提供了五个人工指令模板、用于组织人机对话的数据源和一个人机对话示例.

任务	指令模板	数据源	人机对话示例
分类	<ol style="list-style-type: none"> 1. Categorize the object. 2. Determine the object's category. 3. Identify the category of the object. 4. Classify the object's category. 5. Assign the object to its corresponding category. 	19 个数据源 → SUN-database [55], CVC-ColonDB [20], PolypGen [70], ETIS [21], KUMC [58], Kvasir [29], PSNB12K [17], GastroVision [67], EDD2020 [41], Kvasir-Capsule [57], CP-CHILD [52], BKAI-Small [47], PICCOLO [40], WCE-CCDD [61], CPC-Paired [48], HyperKvasir [44], Nerthus [30], Kvasi-Instrument [51], CVC-ClinicDB [22]	 <p>Human: "Assign the object to its corresponding category" Answer: "polyp"</p>
回归	<ol style="list-style-type: none"> 1. What category does {object coordinates} belong to? 2. Can you tell me the category of {object coordinates}? 3. Could you provide the category for {object coordinates}? 4. Please specify the category of {object coordinates}. 5. What is the category for {coordinates}? 	11 个数据源 → SUN-database [55], CVC-ColonDB [20], EDD2020 [41], ETIS [21], KUMC [58], Kvasir [29], PSNB12K [17], PolypGen [70], CVC-ClinicDB [22], Kvasir-Capsule [57], Kvasi-Instrument [51]	 <p>Human: "Could you provide the category for {<147><317><665><770>}?" Answer: "high grade adenoma"</p>
检索	<ol style="list-style-type: none"> 1. Where is the location of {object category}? 2. Could you give the position of {object category}? 3. Where is {category} located? 4. Could you specify the location of {object category}? 5. Please specify the coordinates of {object category}. 	11 个数据源 → SUN-database [55], CVC-ColonDB [20], EDD2020 [41], ETIS [21], KUMC [58], Kvasir [29], PSNB12K [17], PolypGen [70], CVC-ClinicDB [22], Kvasir-Capsule [57], Kvasi-Instrument [51]	 <p>Human: "Where is adenomatous located?" Answer: "{<128><406><216><496>}"</p>
描述	<ol style="list-style-type: none"> 1. Describe what you see in the image. 2. Interpret what the image shows. 3. Detail the visual elements in the image. 4. Explain the image's visuals thoroughly. 5. Offer a thorough explanation of the image 	19 个数据源 → SUN-database [55], CVC-ColonDB [20], PolypGen [70], ETIS [21], KUMC [58], Kvasir [29], PSNB12K [17], GastroVision [67], EDD2020 [41], Kvasir-Capsule [57], CP-CHILD [52], BKAI-Small [47], PICCOLO [40], WCE-CCDD [61], CPC-Paired [48], HyperKvasir [44], Nerthus [30], Kvasi-Instrument [51], CVC-ClinicDB [22]	 <p>Human: "Detail the visual elements in the image." Answer: "The image displays a medical endoscopic view ..."</p>

则用于数据划分. 对于已有预定义划分的数据集 (如 KUMC [58]、PICCOLO [40]、WCE-CCDD [61]、BKAI-Small [47]、CP-CHILD [52]、Kvasir-Instrument [51] 以及 PS-NBI2K [17]), 本研究严格遵循其原始的划分方案; 而对于未提供标准划分的数据集, 我们依据现有研究所采用的通用划分准则进行处理. 例如, CVC-ClinicDB [22]、CVC-ColonDB [20]、ETIS-Larib [21] 以及 Kvasir 中的息肉类图像 [29] 依据 Fan 等人 [176] 所提供的标准进行划分; SUN-database [55] 中的阳性样本则参考 Ji 等人 [63] 使用的标准. 其余数据集 (包括 HyperKvasir [44]、Kvasir-Capsule [57]、CPC-Paired [48]、Nerthus [30]、GastroVision [67]、EDD2020 [41]、PolypGen [70]、SUN 数据库中的阴性样本 [55], 以及 Kvasir 中除息肉类别以外的样本 [29]) 则统一按照 60%/10%/30% 的比例划分为训练、验证和测试集. 我们数据集最终的图像划分比例约为 59.73% (训练) / 8.77% (验证) / 31.61% (测试). 如图 (4-c) 所示, ColonINST 中所有图像被组织为一个三级标签层次结构, 涵盖 4 个根类别、13 个父类别以及 62 个子类别. 具体而言, 阳性根类别包括三类: 结肠疾病的病理表现 (110,970 例)、与结肠相关的解剖结构 (5,511 例) 以及与结肠镜检查相关的治疗干预措施 (12,139 例), 涵盖了术前、术中与术后阶段. 另一方面, 不在本研究关注范围内的目标 (例如在结肠镜检查过程中未出现的胃、食道、正常齿状线或者胃息肉) 及无目标图像 (如正常黏膜与结肠背景) 被统一归为阴性根类别 (174,381 例). 我们特意保留这些阴性样本, 以期其在未来胃肠道相关研究中发挥潜在价值. 图 (4-d) 展示了 ColonINST 所涵盖类别名称的词云分布.

• **描述生成.** 行为学研究表明, 语言能够影响视觉感知的高级 (如识别) 与低级 (如辨别) 过程 [280]. 这促使我们通过生成图像描述来增强阳性病例的语义丰富度, 例如为 128,620 张展示不同结肠镜检查结果的图像提供文本说明. 一个直接的方法是将类别标签嵌入基本模板中, 类似于 Radford 等人 [151] 所提出的 “一张 [类别] 的照片”. 然而, 这类简化句式所传递的信息有限, 往往难以实现理想的多模态对齐. 如图 (4-e) 所

示, 我们设计了一套流程, 用于生成更具描述性的医学图像文本. 具体而言, 我们针对每张结肠镜图像构建了基于类别条件的自定义提示, 将其作为先验知识输入多模态聊天模型 GPT-4V [5]. 相较于简单的模板句式, 生成的描述具备以下三方面优势: 首先, 这些描述文本更加具体地说明了目标的独特特征, 且提供了物体表面的细节, 例如 “表面光滑、呈圆形且有结节状外观”; 其次, 基于类别先验生成的文本能够区分良性与恶性病变, 例如将病灶描述为 “表面完整、无糜烂或溃疡迹象”; 最后, 生成的图像描述不仅关注病灶本身, 还涵盖其周围环境, 例如 “黏膜略显红斑, 但无明显炎症变化”, 从而提供了更全面、准确的医学解读.

• **组织微调数据.** 在最后一步, 本研究将所有阳性病例统一转化为单轮对话形式, 即 “图像 + 人类指令 → 模型回复”. 如图 (4-f) 所示, 研究从多种图像、标签和描述来源中整理出共计 450,724 个人机对话数据, 涵盖了四类不同的任务. 具体而言, 分类任务 (CLS) 要求模型为一张结肠镜图像分配一个类别标签. 利用定位标签, 我们引入了两个相关任务: 指称表达生成任务 (REG), 即对指定图像区域进行分类; 以及指称表理解任务 (REC), 即根据给定的类别定位目标物体. 此外, 还引入了图像描述生成任务 (CAP), 以 GPT-4V 生成的描述作为回复内容. 为了提升对话的多样性, 本研究为每项任务设计了五个问题模板, 并从中随机选择一个来构建人机指令对. 具体模板请参见表 6.

5.2 提出的多模态语言模型: ColonGPT

• **动机.** 如图 5 所示, 本研究评估了三种多模态人工智能聊天机器人 (LLaVA-v1.5 [9]、LLaVA-Med-v1.5 [281] 和 GPT-4V [5]) 在零样本文本生成能力. 实验结果表明, 这些模型往往给出不准确或模糊的回复, 难以适配特定的结肠镜检查任务. 基于这一观察, 我们开发了面向结肠镜检查场景的多模态语言模型 ColonGPT, 使其能够依据用户指令将图像准确分类为 “息肉” 等具体类别, 从而实现更精确、个性化的应用.

• **概述.** 本研究旨在验证语言模型 (LMs) 在医学光学成像领域中解析视觉与文本信号的能力. 为此, 我们提出了一个名为

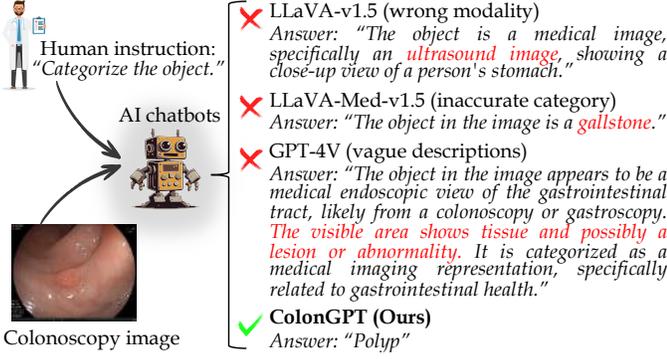


图 5. 结肠镜图像分类的响应比较. 我们将三个人工智能聊天机器人的零样本语言响应与我们的多模态模型 ColonGPT 的响应进行了评估对比.

ColonGPT 的基线模型, 用于根据人类指令执行结肠镜检查相关任务. 如图 6 左侧所示, 本研究提出一个名为 ColonGPT 的基线模型, 其遵循多模态语言模型 (MLM) 的通用框架设计 [9], 包含了四个关键组件. (a) 首先, 语言分词器将人类指令 X_q 编码为词元序列 $T_q \in \mathbb{R}^{N_q \times D}$, 其中词元长度为 N_q , 嵌入特征维度为 D . (b) 其次, 视觉编码器 (通常基于 Transformer 架构) 将一张尺寸为高 H 宽 W 的结肠镜检查图像 $X_v \in \mathbb{R}^{H \times W \times 3}$ 编码为视觉嵌入特征 $E_v \in \mathbb{R}^{\frac{HW}{P} \times d}$. 这里, P 表示图像分块的大小, d 表示图元的嵌入维度. (c) 随后, 多模态适配器将视觉嵌入 E_v 转换为 N_v 个视觉词元 $T_v \in \mathbb{R}^{N_v \times D}$, 以匹配语言词元 T_q 的嵌入维度 D . (d) 最后, 语言模型接收拼接后的图元 T_v 和文本词元 T_q 作为输入. 并利用概率链式法则, 以自回归的方式生成一个长度为 L 的响应序列 Y , 这个过程可以写为 $p(Y) = \prod_{i=1}^L p(y_i | T_v, T_q, Y_{<i})$, 其中 $Y_{<i} = [y_1, y_2, \dots, y_{i-1}]$ 是索引在 i 之前的已预测语言词元序列.

• **多粒度多模态适配器.** 先前的研究工作 [8]–[10] 通常采用由三个线性层和高斯误差线性单元 (GELU) 激活函数组成的多层感知机 (MLP) 结构, 作为多模态适配器. 然而, 这类方法在处理全部图元时会引入显著冗余, 因为并非所有图元在医学任务中的重要性相同; 此外, 考虑到该方法的计算复杂度与词元数量呈二次复杂度关系, 进一步加剧了计算负担. 为应对上述挑战, 本文提出了一种改进的多模态适配器, 在两个线性变换之间引入了多粒度池化机制. 如图 6 右侧所示, 首先通过一个线性层和 GELU 激活函数, 将输入的视觉嵌入 $E_v \in \mathbb{R}^{\frac{HW}{P} \times d}$ 从维度 d 映射至 D , 随后重塑为空间格式 $F_v \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P} \times D}$. 为在尽量减少图元数量的同时保持模型性能, 本文在池化阶段设计了三项改进策略, 并在 §5.4 章节中对其有效性进行了系统验证. (a) 多粒度视角. 本研究在多模态适配器中引入了一组核大小为 s_1, \dots, s_N 的自适应平均池化操作, 用于提取多粒度视觉特征. 该自适应机制能够灵活处理不同长度的输入序列, 其中每个池化特征的维度为 $\mathbb{R}^{s_n \times s_n \times D}$, 对应核大小为 s_n 的池化操作. (b) 位置编码. 受文献 [282] 启发, 为增强池化特征中的空间感知能力, 本文在

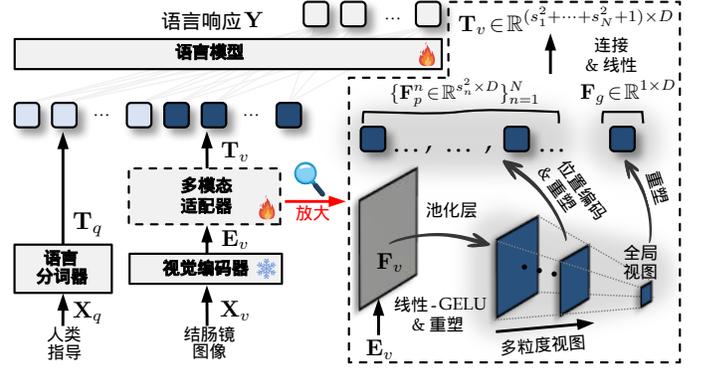


图 6. 多模态语言模型 ColonGPT 的细节.

每个池化特征上使用了具有合适大小的零值填充的二维卷积操作. 默认配置下, 在输入特征边缘添加一像素大小的零填充边界. (c) 全局视角. 此外, 为获取图像的全局视角, 本文在特征 F_v 上施加了核大小为 1 的全局平均池化操作, 得到形状为 $\mathbb{R}^{1 \times 1 \times D}$ 的全局视角向量. 随后, 将所有池化特征重塑为展平向量形式 $\{F_p^n \in \mathbb{R}^{s_n^2 \times D}\}_{n=1}^N$ 及 $F_g \in \mathbb{R}^{1 \times D}$. 最后, 本文将所有展平向量连接起来, 并通过一个线性变换生成最终的图像词元序列 $T_v \in \mathbb{R}^{N_v \times D}$, 其中 $N_v = s_1^2 + \dots + s_N^2 + 1$ 表示图元总数.

• **模型实现.** 为提高普通用户的可重复性, 我们以资源友好的方式实现了 ColonGPT, 能够与现有的视觉编码器和语言模型无缝集成. 具体而言, 该模型采用 SigLIP-SO (参数量为 0.4B) [6] 作为视觉编码器, 其输入图像分辨率为 $H = W = 384$, 图像块大小为 $P = 14$, 特征嵌入维度为 $d = 1152$. 在该配置下, 生成的视觉嵌入 E_v 具有形状 $\mathbb{R}^{729 \times 1152}$, 其中图元数量为 $729 = \lfloor \frac{384}{14} \rfloor^2$. 同时, 本文使用 Phi-1.5 (参数量为 1.3B) [7] 作为语言分词器与语言模型, 其嵌入维度为 $D = 2048$. 为进一步降低计算开销, 池化操作的核大小被设置为 $\{s_1, s_2\} = \{14, 7\}$, 使图元数量从原始的 729 降至 246, 减少幅度达 66.26%. 这种轻量化设计显著缩短了训练时间, 使得完整训练过程可在七小时内完成, 从而加速了概念验证与迭代开发.

• **训练方案.** 本研究基于 PyTorch 框架实现模型, 并使用两块 NVIDIA H200 显卡进行加速训练. 我们采用 AdamW 优化器, 初始学习率设为 2×10^{-3} , 并结合余弦退火策略进行动态调整. (a) 预对齐阶段: 首先在约 83K 个图像-描述对上训练多粒度多模态适配器, 以实现视觉编码器与语言模型之间的预对齐, 从而使模型具备理解图元的能力. 在该阶段, 视觉编码器与语言模型均保持冻结状态, 仅训练多模态适配器 (学习率设为 2×10^{-4}). (b) 监督微调阶段: 随后在约 202K 个图像-文本对上进行微调, 使模型能够适应结肠镜检查中的分类 (CLS)、指称表达生成 (REG) 与指称表达理解 (REC) 任务. 为了提升效率, 本文对语言模型使用低秩自适应 (LoRA) 策略 [283], 其中秩为 $r = 128$, 缩放因子为 $\alpha = 256$. 在该阶段,

表 7

三个对话任务的多模态基准. “LoRA” 是指使用低阶自适应 [283] 进行微调, “EXT” 表示在额外的一般数据上使用预训练的权重. 本研究在 ColonINST 比较了验证集中已见样本和测试集中未知样本的结果. 符号 \uparrow 表示得分越高, 表现越好.

模型	视觉编码器 (输入形状/链接)	语言模型 (模型大小/链接)	No.	LoRA	EXT	CLS 任务 ($\mathcal{A} \uparrow$)		REG 任务 ($\mathcal{A} \uparrow$)		REC 任务 (IoU \uparrow)	
						seen	unseen	seen	unseen	seen	unseen
MiniGPT-v2 [8]	EVA-G/14 (448px/link)	LLaMA2 (7B/link)	#A ₁	✓		91.49%	77.93%	94.69%	72.05%	23.45%	15.36%
			#A ₂	✓	✓	90.00%	76.82%	87.65%	70.23%	27.97%	31.13%
LLaVA-v1 [278]	CLIP-L/14 (224px/link)	Vicuna-v1.3 (7B/link)	#B ₁	✓		87.86%	72.08%	84.55%	68.11%	20.05%	12.72%
			#B ₂	✓	✓	89.61%	42.17%	86.87%	46.85%	21.81%	3.24%
LLaVA-v1.5 [9]	CLIP-L/14 (336px/link)	Vicuna-v1.5 (7B/link)	#C ₁	✓		92.97%	79.10%	98.58%	70.38%	55.72%	34.32%
			#C ₂	✓	✓	93.33%	80.89%	99.32%	72.88%	61.97%	42.31%
Bunny-v1.0-3B [10]	SigLIP-SO (384px/link)	Phi2 (2.7B/link)	#D ₁	✓		91.16%	75.50%	96.61%	69.45%	46.24%	31.24%
			#D ₂	✓	✓	92.47%	79.50%	96.02%	75.08%	54.00%	41.48%
MGM-2B [284]	CLIP-L/14 (336px/link) & ConvNeXt-L (768px/link)	Gemma (2B/link)	#E ₁	✓		92.97%	78.99%	98.17%	69.81%	39.78%	16.00%
			#E ₂		✓	93.24%	78.69%	98.75%	74.30%	57.25%	25.23%
MobileVLM-1.7B [285]	CLIP-L/14 (336px/link)	MobileLLaMA (1.4B/link)	#F ₁		✓	93.02%	78.75%	97.78%	73.14%	47.30%	31.46%
			#F ₂	✓	✓	93.64%	80.44%	97.87%	78.03%	51.36%	34.80%
LLaVA-Med-v1.0 [281]	CLIP-L/14 (224px/link)	LLaMA1 (7B/link)	#G ₁			93.52%	78.04%	97.74%	75.07%	41.60%	24.89%
			#G ₂		✓	93.84%	77.38%	97.35%	75.25%	39.43%	20.85%
LLaVA-Med-v1.5 [281]	CLIP-L/14 (224px/link)	Mistral-v0.2 (7B/link)	#H ₁	✓		93.62%	79.24%	99.30%	73.05%	64.69%	41.97%
			#H ₂	✓	✓	87.22%	66.51%	90.40%	70.00%	13.39%	12.95%
ColonGPT (Ours)	SigLIP-SO (384px/link)	Phi1.5 (1.3B/link)	-	✓		94.06%	83.24%	99.96%	80.18%	85.74%	56.24%

适配器和语言模型的学习率分别为 2×10^{-3} 和 2×10^{-4} . 训练过程共进行两个阶段, 每阶段分别训练三轮. 每块 GPU 的批大小为 16, 梯度每两步累积一次回传. 训练耗时方面, 预对齐阶段约 1.4 小时, 微调阶段约 5.3 小时.

5.3 多模态基准

• **对比模型.** 为构建该领域广泛认可的多模态基准, 本研究选取了八个主流的多模态语言模型 (MLMs) 作为对比方法, 其中包括六个通用模型和两个医学领域专用模型. 如表 7 所示, 每个对比模型均在两种训练设定下参与评估, 分别对应于应用低秩自适应 (LoRA) [283] 或基于额外预训练数据的初始化知识. 所有对比模型均在 ColonINST 所提供的完整训练集与验证集上进行了重新训练.

• **评估协议.** 本研究在多模态基准测试中针对三类对话任务进行了定量评估. 对于两项分类相关任务, 即分类任务 (CLS) 与指称表达生成任务 (REG), 我们采用准确率指标 (\mathcal{A}) 来衡量正确预测的类别数占总预测数的比例. 对于指称表达理解任务 (REC), 我们采用交并比 (IoU) 指标评估模型的定位精度. 此外, 针对图像描述生成任务 (CAP), 由于其语言输出具有一定主观性, 本研究从医学准确性的角度出发, 定性分析模型是否准确识别图像中的解剖结构、类别名称或相关临床信息描述.

• **学习能力.** 首先对每个模型进行开卷测试, 以定量评估每个模型在多大程度上能有效地将训练阶段的视觉和语言模式内化为自身知识. 具体来说, 我们在这些模型见过的训练样本上进行评估, 也就是对应 ColonINST 中的验证集. 如表 7 中 “seen” 列所示, 我们在 CLS 任务上取得 $\mathcal{A} = 94.06\%$ 、REG 任务上取得 $\mathcal{A} = 99.96\%$ 以及 REC 任务上取得 $\text{IoU} = 85.74\%$, 均实现了最高分数. 这表明 ColonGPT 模型具有更强的学习能力, 能够准确地对图像进行分类, 并理解与特定视觉区域相关的指称表达.

• **泛化能力.** 我们进一步进行了闭卷测试, 以评估各模型将所学知识推广到未知样本的能力, 即 ColonINST 中的测试样本. 如表 7 中 “unseen” 列所示, 结果一致表明在所有三项任务中, ColonGPT 在未知样本上表现更优. 回想一下, 在见过的样本场景中, 指称表达生成任务 (REG) 上 ColonGPT 模型表现略优于 LLaVA-v1.5. 然而, 在未见样本场景中, 这一差距进一步被扩大, ColonGPT 达到 80.18% 的准确率, 甚至比 7B 参数规模的 LLaVA-v1.5 (#C₂) 还高出了 7.3%. 此外, 与医学领域的定制模型相比, ColonGPT 作为专用于结肠镜检查的对话助手, 在多模态任务中展现出显著潜力, 尤其是在泛化到未见过数据方面. 例如, 在 CLS 任务的未知场景中, ColonGPT 模型的准确率比 #G₁ 和 #H₁ 分别高出 5.2% 和 4.0%.

• **定性分析.** 图 7 展示了我们模型在四个对话任务中的三种多模态能力. (a) 理解能力: 在分类任务 (CLS) 中, 模型能够识别细微的视觉特征, 在视觉相似的图像中区分 “高级别异型增生” 与 “腺瘤”. 在指称表达生成任务 (REG) 中, 模型能正确地将给定坐标处的复杂视觉特征转化为精确的医学术语. (b) 定位能力: 这需要 ColonGPT 能理解语言查询, 并在复杂的结肠环境中定位视觉目标. 指称表达理解任务 (REC) 的输出展示了 ColonGPT 能使用边界框精确地定位指定表达. (c) 描述能力: 需要将视觉信息综合为连贯的、与临床相关的文本. 模型提供了带蒂息肉的描述, 详细说明了其形态、背景特征及潜在的临床意义. 此外, 当图像中存在器械时, ColonGPT 还能描述治疗过程, 例如: “图像中可见用于切除息肉的圈套器”.

5.4 诊断实验

• **视觉编码器.** 诊断实验始于一个问题——什么样的视觉表征更适用于多模态结肠镜数据? 本研究基于多种大规模预训练策略, 构建了四类图像表征方式: 监督学习 (ConvNeXtV2 [286], ViT [101])、重建学习 (MAE [287]), 以及使用纯视觉对比学习 (DINOv2 [288]) 或使用视觉-语言 (VL) 对比学习

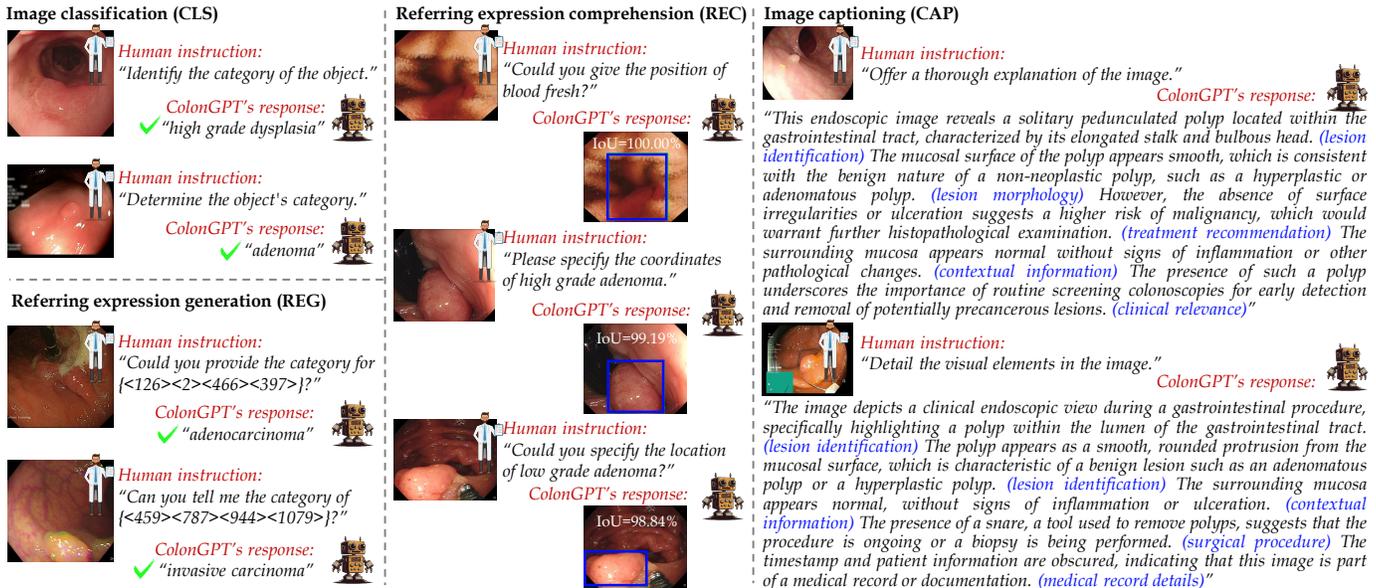


图 7. ColonGPT 模型的多模态能力展示. ColonGPT 模型可以通过对话交互执行各种多模态结肠镜检查任务, 包括基于理解 (CLS、REG)、定位 (REC) 和描述 (CAP) 的任务.

(CLIP [289], SigLIP [6]). 如表 (VIII-a) 所示, 所有的编码器都使用了来自 Huggingface 的预训练权重. 为了确保一致性, 本研究将 MAE 和 DINOv2 默认的位置嵌入从 224 像素手动插值到 384 像素 (标记为 *), 而其余模型保持默认输入不变. 观察结果表明, 使用视觉-语言 (VL) 数据的对比学习编码器的表现优于其他策略. 这说明在预训练期间通过弱文本对视觉表征进行初步对齐, 有助于将视觉嵌入更有效地映射至语言空间. 至于其他单模态编码器, 监督学习方法 (ConvNeXtV2、ViT) 和重建方法 (MAE) 都给出了令人满意的反馈. 然而, 纯视觉的对比学习模型 (DINOv2) 表现不佳¹, 这表明其视觉表征可能难以与语言空间对齐.

• **多粒度多模态适配器.** 它是连接视觉和语言模态的关键组件, 通过减少视觉词元来降低计算开销. 正如表 (VIII-b) 中所详述的, 本研究从三个方面分析其有效性. (a) 如何设定池化核大小? 作为参考, 本文借鉴了文献 [8], [9] 中提出的多模态适配器, 构建了 ColonGPT 的初始基线版本. 该适配器采用纯多层感知机 (MLP) 架构, 用于对所有输入词元进行均等处理. 然后, 我们进一步设计了五种变体, 通过逐步减小池化核的大小来压缩视觉词元, 其配置分别为: 16, 8, 1, 14, 7, 1, 12, 6, 1, 10, 5, 1 和 8, 4, 1. 综合考虑模型性能与计算成本, 最终选择了 14, 7, 1 作为最优配置, 这将视觉词元数量从 100% (729 个) 压缩至 33.74% (246 个), 同时在这三项多模态对话任务中仍保持了优异的表现. 例如, 在分类任务 (CLS) 与指称表达生成任务 (REG) 中, 性能分别提升了 5.14% 和 4.91%. (b) 全局上下文是否必要? 本研究从默认设置 {14, 7, 1} 的多

1. 值得注意的是, 我们进行了多次实验尝试, 通过调整输入尺寸、模型规模和学习率等超参数, 但对于多模态结肠镜检查任务而言, 均未产生令人满意的结果.

粒度适配器中移除全局视角特征, 得到一个设置为 {14, 7} 的对照变体. 实验结果显示性能下降了, 这表明在视觉嵌入环节, 对全局上下文信息进行有效捕获, 是实现更优成果的必要前提. (c) 位置编码是否重要? 如表 (VIII-b) 最后一行所示, 去除位置编码的模型在指称表达生成任务 (REG) 中性能显著下降, 从 80.18% 降至 76.35%. 这表明视觉序列的相对位置信息对于定位任务至关重要.

• **微调策略.** 最后, 本文探究了“如何在多模态结肠镜检查数据上有效地微调我们的模型”. 如表 (VIII-c) 所示, 本研究针对语言模型 Phi-1.5 构建了多个微调版本. 其中涵盖七种不同的低秩自适应 (LoRA) 配置, 以及一个全量微调的参考版本. 在配置为 $r/\alpha = 128/256$ 的 LoRA 变体中观察到了最佳性能. 进一步来看, 全量微调的版本在三项任务上的表现并不理想. 这表明在数据规模有限的情况下, 增加可微调参数并不会带来性能提升.

5.5 经验教训

本研究是对结肠镜检查领域中多模态指令微调技术的初步探索. 研究将多模态与多任务范式整合于一个因果语言模型之中, 该模型有两个特点: 一是在语言空间内解读视觉内容, 二是在基于下一个词元预测的框架下处理各种视觉任务. 最终, 我们从实验中得出了一些经验教训, 希望为未来多模态研究和发展提供指导性建议.

• **拥抱数据稀缺问题.** 一般来说, 多模态语言模型 (MLMs) [9], [278] 会在海量数据上进行两阶段训练, 例如使用约 55.8 万个样本进行多模态对齐, 随后使用约 66.5 万个指令微调样本以确保符合人类需求. 本研究采用了单阶段策略, 在相对较少的约 285K 个指令的训练数据上直接对 ColonGPT 进行微调. 在数据有限的结肠镜检查场景中, 这一策略似乎是有效的. 本

表 8

ColonGPT 中三个核心组件的诊断研究。“*”：对位置嵌入进行插值以获得更高的分辨率，特别是从 224px 到 384px。默认配置为灰色背景。

(a) 视觉编码器的不同呈现					(b) 多粒度多模态适配器					(c) 微调策略					
视觉编码器	输入/链接	CLS	REG	REC	词元 (比率)	CLS	REG	REC	策略	r	α	CLS	REG	REC	
ConvNeXtV2-L	384px/ link	81.51%	76.41%	53.09%	MLP 基线	729 (100.00%)	78.10%	75.27%	56.29%	全量微调	-	-	70.03%	63.07%	13.17%
ViT-L	384px/ link	83.99%	78.79%	56.64%	{16, 8, 1}	321 (44.03%)	79.84%	76.13%	53.15%	LoRA	8 16	80.51%	77.61%	51.61%	
MAE-L*	384px/ link	83.30%	79.51%	53.59%	{14, 7, 1}	246 (33.74%)	83.24%	80.18%	56.24%	LoRA	16 32	81.95%	77.87%	52.41%	
MAE-L	224px/ link	82.85%	79.19%	57.57%	{14, 7}	245 (33.61%)	80.67%	78.86%	58.05%	LoRA	32 64	82.63%	78.76%	52.11%	
DINOv2-L	224px/ link	22.48%	10.09%	5.65%	{12, 6, 1}	181 (24.83%)	79.46%	75.25%	52.22%	LoRA	64 128	82.27%	79.72%	52.20%	
CLIP-L	336px/ link	77.05%	71.92%	51.68%	{10, 5, 1}	126 (17.28%)	81.42%	76.73%	52.04%	LoRA	128 256	83.24%	80.18%	56.24%	
SigLIP-SO	384px/ link	83.24%	80.18%	56.24%	{8, 4, 1}	81 (11.11%)	81.55%	77.60%	53.39%	LoRA	256 512	81.75%	77.89%	54.03%	
					w/o Pos. Enc.	246 (33.74%)	80.56%	76.35%	55.77%	LoRA	512 1024	79.39%	74.54%	55.52%	

表 9

失败案例。CLS 任务的准确率在 ColonINST 基准中的五个长尾类别上进行评估，每个类别包含的样本数均少于 30 例。最后一列计算所有选定类别的平均得分。

Model	adenocarcinoma	suspicious precancerous lesion	inflammatory bowel disease	colon diverticula	resection margin	平均准确率
LLaVA-v1.5 (No. #C2)	0.00%	14.29%	0.00%	0.00%	50.00%	14.04%
ColonGPT (Ours)	57.14%	14.29%	22.22%	22.22%	37.50%	28.07%

研究提出了两种可行的方法来弥补这个以数据为核心的问题。(a) 扩大数据规模是提高特定领域表征能力的一种直接方法。一种具有成本效益的方法是，当公共数据源耗尽时考虑使用合成数据 [290]。(b) 多样化的人机对话形式有助于有效训练适用于结肠镜检查的人工智能专家模型。这一过程不仅可以通过先进的人工智能聊天机器人扩展问答对，还能够组织更多可执行的任务，例如将掩码转换为多边形以实现精细分割 [291] 或对视频分析中的多帧相关性进行建模 [292]。(c) 我们观察到当前模型在长尾案例上表现出一定的局限性。尽管它们在基准测试 (Tab. 7) 上取得了出色的整体性能，但在稀有类别中的有效性却显著下降 (Tab. 9)。这种差异表明，在处理长尾数据分布时仍存在挑战，因为由于训练样本不足，模型可能对多数类别产生过拟合。一个有前景的方向是将智能化框架 [?] 与显式推理过程相结合。

• **效率至上**。如前所述，我们在使用更少训练数据的情况下，依然优于其他竞争模型，取得了更佳的性能。这一成果得益于我们对 ColonGPT 的高效构建方式。(a) 结肠镜检查数据本身存在大量冗余信息。正如 §2.2 章节中所讨论的，大多数黏膜表面具有高度相似性，且良性病变常常呈现出与周围环境相似的伪装特征。为减少这种冗余，本研究提出了一种多粒度的多模态适配器，能够在不显著影响模型性能的前提下，对词元进行有选择性的采样。进一步优化方向可借鉴以往在词元压缩方面的经验 [293]。(b) 如表 7 所示，轻量级的 Phi.1.5 模型 [7] 展现出令人惊讶的效率，其性能甚至超过了一些参数规模达到 70 亿的对比模型。这一现象表明，规模更大的模型可能更依赖于更庞大数量的结肠镜检查数据。因此，未来的研究应更关注于提升参数效率，而非一味依赖大规模的计算资源，尤其是在医学领域中。一个颇具前景的方向是引入无编码器的解决方案 [294]，以简化多模态语言模型 (MLM) 的结构，从而更直接地实现对视觉像素的理解与解析。

• **改善空间感知能力**。本研究发现，模型根据语言描述准确确定

位目标的能力仍然有限。这一点在表 7 中指称表达理解任务 (REC) 的结果中体现得尤为明显：在未见样本上测试时，大多数模型的交并比 (IoU) 得分均低于 50%。为突破该性能瓶颈，本文提出两条可能的研究路径。(a) 首先，在构建 ColonGPT 时，我们采用了通用领域的预训练视觉编码器和语言模型。然而，这一策略面临两个核心挑战：一是通用数据与医学光学数据之间存在显著分布差异；二是视觉模态与语言模态之间仍存在语义鸿沟。对此，文献 [295] 提出，在进行指令微调前，先对多模态空间进行预训练和对齐，是缓解上述问题的一个具有前景的方向。(b) 其次，因果语言模型基于下一个词元预测的机制，在处理需要精确输出的任务（如算术计算）时存在错误累积问题 [296]。在 REC 任务中，语言模型并不直接负责精确预测目标位置坐标。因此，我们建议下一代框架应实现视觉与语言模块的功能解耦，使其各司其职。例如，文献 [297] 提出的并行结构即是一种有前景的方案，能够在同时生成语言描述的同时预测目标的分割掩码。

6 总结

本文聚焦于智驱结肠镜检查技术的前沿进展调查，及其在多模态领域中蕴含的广泛应用潜力。文章围绕两条主线展开论述：首先，我们系统调研了结肠镜检查中的四类场景感知任务，梳理并总结了当前的关键挑战与研究空白；其次，调研结果表明，结肠镜检查相关的多模态研究仍处于起步阶段，尚未得到充分挖掘。基于此，本文作出以下三项主要贡献：构建了大规模多模态指令微调数据集 ColonINST；开发了面向结肠镜检查场景的多模态语言模型 ColonGPT；三是建立了系统的多模态基准评测体系。

参考文献

- [1] G.-P. Ji, J. Liu, P. Xu, N. Barnes, F. S. Khan, S. Khan, and D.-P. Fan, “Frontiers in intelligent colonoscopy,” *arXiv preprint arXiv:2410.17241*, 2024.

- [2] C. Eng, T. Yoshino, E. Ruíz-García, N. Mostafa, C. G. Cann, B. O'Brian, A. Benny, R. O. Perez, and C. Cremolini, "Colorectal cancer," *The Lancet*, vol. 394, no. 10207, pp. 1467–1480, 2024.
- [3] M. B. Wallace, P. Sharma, P. Bhandari, J. East, G. Antonelli, R. Lorenzetti, M. Vieth, I. Speranza, M. Spadaccini, M. Desai *et al.*, "Impact of artificial intelligence on miss rate of colorectal neoplasia," *Gastro*, vol. 163, no. 1, pp. 295–304, 2022.
- [4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE CVPR*, 2009.
- [5] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [6] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *IEEE ICCV*, 2023.
- [7] Y. Li, S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, and Y. T. Lee, "Textbooks are all you need ii: phi-1.5 technical report," *arXiv preprint arXiv:2309.05463*, 2023.
- [8] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.
- [9] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *IEEE CVPR*, 2024.
- [10] M. He, Y. Liu, B. Wu, J. Yuan, Y. Wang, T. Huang, and B. Zhao, "Efficient multimodal learning from data-centric perspective," *arXiv preprint arXiv:2402.11530*, 2024.
- [11] V. S. Prasath, "Polyp detection and segmentation from video capsule endoscopy: A review," *J. Imaging*, vol. 3, no. 1, p. 1, 2016.
- [12] B. Taha, N. Werghe, and J. Dias, "Automatic polyp detection in endoscopy videos: A survey," in *IEEE IASTED*, 2017.
- [13] L. F. Sanchez-Peralta, L. Bote-Curiel, A. Picon, F. M. Sanchez-Margallo, and J. B. Pagador, "Deep learning to find colorectal polyps in colonoscopy: A systematic literature review," *AIIM*, vol. 108, p. 101923, 2020.
- [14] I. Pacal, D. Karaboga, A. Basturk, B. Akay, and U. Nalbantoglu, "A comprehensive review of deep learning in colon cancer," *CIBM*, vol. 126, p. 104003, 2020.
- [15] B. Münzer, K. Schoeffmann, and L. Böszörményi, "Content-based processing and analysis of endoscopic images and videos: A survey," *MTAP*, vol. 77, pp. 1323–1362, 2018.
- [16] M. Taghiakbari, Y. Mori, and D. von Renteln, "Artificial intelligence-assisted colonoscopy: A review of current state of practice and research," *WJG*, vol. 27, no. 47, p. 8103, 2021.
- [17] G. Yue, G. Zhuo, S. Li, T. Zhou, J. Du, W. Yan, J. Hou, W. Liu, and T. Wang, "Benchmarking polyp segmentation methods in narrow-band imaging colonoscopy images," *IEEE JBHI*, vol. 27, no. 7, pp. 3360–3371, 2023.
- [18] Z. Wu, F. Lv, C. Chen, A. Hao, and S. Li, "Colorectal polyp segmentation in the deep learning era: A comprehensive survey," *arXiv preprint arXiv:2401.11734*, 2024.
- [19] J. Mei, T. Zhou, K. Huang, Y. Zhang, Y. Zhou, Y. Wu, and H. Fu, "A survey on deep learning for polyp segmentation: Techniques, challenges and future trends," *VInt*, vol. 3, no. 1, p. 1, 2025.
- [20] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *PR*, vol. 45, no. 9, pp. 3166–3182, 2012.
- [21] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *CARS*, vol. 9, no. 2, pp. 283–293, 2014.
- [22] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilarino, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *CMIG*, vol. 43, pp. 99–111, 2015.
- [23] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE TMI*, vol. 35, no. 2, pp. 630–644, 2015.
- [24] M. Ye, S. Giannarou, A. Meining, and G.-Z. Yang, "Online tracking and retargeting with applications to optical biopsy in gastrointestinal endoscopic examinations," *MedIA*, vol. 30, pp. 144–157, 2016.
- [25] F. Deeba, F. M. Bui, and K. A. Wahid, "Automated growcut for segmentation of endoscopic images," in *IJCNN*, 2016.
- [26] R. Zhang, Y. Zheng, T. W. C. Mak, R. Yu, S. H. Wong, J. Y. Lau, and C. C. Poon, "Automatic detection and classification of colorectal polyps by transferring low-level cnn features from nonmedical domain," *IEEE JBHI*, vol. 21, no. 1, pp. 41–47, 2016.
- [27] P. Mesejo, D. Pizarro, A. Abergel, O. Rouquette, S. Beorchia, L. Poincloux, and A. Bartoli, "Computer-aided classification of gastrointestinal lesions in regular colonoscopy," *IEEE TMI*, vol. 35, no. 9, pp. 2051–2063, 2016.
- [28] Q. Angermann, J. Bernal, C. Sánchez-Montes, M. Hammami, G. Fernández-Esparrach, X. Dray, O. Romain, F. J. Sánchez, and A. Histace, "Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis," in *MICCAI-W*, 2017.
- [29] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt *et al.*, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *ACM MMSys*, 2017.
- [30] K. Pogorelov, K. R. Randel, T. de Lange, S. L. Eskeland, C. Griwodz, D. Johansen, C. Spampinato, M. Taschwer, M. Lux, P. T. Schmidt *et al.*, "Nerthus: A bowel preparation quality video dataset," in *ACM MMSys*, 2017.
- [31] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdal, and A. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *JHE*, vol. 2017, no. 1, p. 4037190, 2017.
- [32] A. Koulaouzidis, D. K. Iakovidis, D. E. Yung, E. Rondonotti, U. Kopylov, J. N. Plevris, E. Toth, A. Eliakim, G. W. Johanson, W. Marlicz *et al.*, "Kid project: an internet-based digital video atlas of capsule endoscopy for research purposes," *EIO*, vol. 5, no. 06, pp. E477–E483, 2017.
- [33] I. N. Figueiredo, L. Pinto, P. N. Figueiredo, and R. Tsai, "Unsupervised segmentation of colonic polyps in narrow-band imaging data based on manifold representation of images and wasserstein distance," *BSPC*, vol. 53, p. 101577, 2019.
- [34] P. N. Figueiredo, I. N. Figueiredo, L. Pinto, S. Kumar, Y.-H. R. Tsai, and A. V. Mamonov, "Polyp detection with computer-aided diagnosis in white light colonoscopy: comparison of three different methods," *EIO*, vol. 7, no. 02, pp. E209–E215, 2019.

- [35] T.-H. Hoang, H.-D. Nguyen, V.-A. Nguyen, T.-A. Nguyen, V.-T. Nguyen, and M.-T. Tran, "Enhancing endoscopic image classification with symptom localization and data augmentation," in *ACM MM*, 2019.
- [36] M. Cho, J. H. Kim, K. S. Hong, J. S. Kim, H.-J. Kong, and S. Kim, "Identification of cecum time-location in a colonoscopy video by deep learning analysis of colonoscope movement," *PeerJ*, vol. 7, p. e7256, 2019.
- [37] S. Ali, F. Zhou, C. Daul, B. Braden, A. Bailey, S. Realdon, J. East, G. Wagnieres, V. Loschenov, E. Grisan *et al.*, "Endoscopy artifact detection (ead 2019) challenge dataset," *arXiv preprint arXiv:1905.03209*, 2019.
- [38] Y. Liu, Y. Tian, G. Maicas, L. Z. C. T. Pu, R. Singh, J. W. Verjans, and G. Carneiro, "Photoshopping colonoscopy video frames," in *IEEE ISBI*, 2020.
- [39] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. d. Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *MMM*, 2020.
- [40] L. F. Sánchez-Peralta, J. B. Pagador, A. Picón, Á. J. Calderón, F. Polo, N. Andracka, R. Bilbao, B. Glover, C. L. Saratxaga, and F. M. Sánchez-Margallo, "Piccolo white-light and narrow-band imaging colonoscopic dataset: a performance comparative of models and datasets," *ApplSci*, vol. 10, no. 23, p. 8501, 2020.
- [41] S. Ali, N. Ghatwary, B. Braden, D. Lamarque, A. Bailey, S. Realdon, R. Cannizzaro, J. Rittscher, C. Daul, and J. East, "Endoscopy disease detection challenge 2020," *arXiv preprint arXiv:2003.03376*, 2020.
- [42] R. Leenhardt, C. Li, J.-P. Le Mouel, G. Rahmi, J. C. Saurin, F. Cholet, A. Boureille, X. Amiot, M. Delvaux, C. Duburque *et al.*, "Cad-cap: a 25,000-image database serving the development of artificial intelligence for capsule endoscopy," *EIO*, vol. 8, no. 03, pp. E415–E420, 2020.
- [43] C. C. Poon, Y. Jiang, R. Zhang, W. W. Lo, M. S. Cheung, R. Yu, Y. Zheng, J. C. Wong, Q. Liu, S. H. Wong *et al.*, "Ai-doscopist: a real-time deep-learning-based algorithm for localising polyps in colonoscopy videos with edge computing devices," *NPJDM*, vol. 3, no. 1, p. 73, 2020.
- [44] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen *et al.*, "Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *SData*, vol. 7, no. 1, p. 283, 2020.
- [45] X. Guo, C. Yang, Y. Liu, and Y. Yuan, "Learn to threshold: Thresholdnet with confidence-guided manifold mixup for polyp segmentation," *IEEE TMI*, vol. 40, no. 4, pp. 1134–1146, 2020.
- [46] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. B. Guo, B. Matuszewski *et al.*, "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy," *MedIA*, vol. 70, p. 102002, 2021.
- [47] P. Ngoc Lan, N. S. An, D. V. Hang, D. V. Long, T. Q. Trung, N. T. Thuy, and D. V. Sang, "Neounet: Towards accurate colon polyp segmentation and neoplasm detection," in *ISVC*, 2021.
- [48] Q. Wang, H. Che, W. Ding, L. Xiang, G. Li, Z. Li, and S. Cui, "Colorectal polyp classification from white-light colonoscopy images via domain alignment," in *MICCAI*, 2021.
- [49] Y. Ma, X. Chen, K. Cheng, Y. Li, and B. Sun, "Ldpolypvideo benchmark: a large-scale colonoscopy video dataset of diverse polyps," in *MICCAI*, 2021.
- [50] N. Celik, S. Ali, S. Gupta, B. Braden, and J. Rittscher, "Endouda: a modality independent segmentation approach for endoscopy imaging," in *MICCAI*, 2021.
- [51] D. Jha, S. Ali, K. Emanuelsen, S. A. Hicks, V. Thambawita, E. Garcia-Ceja, M. A. Riegler, T. de Lange, P. T. Schmidt, H. D. Johansen *et al.*, "Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy," in *MMM*, 2021.
- [52] W. Wang, J. Tian, C. Zhang, Y. Luo, X. Wang, and J. Li, "An improved deep learning approach and its applications on colonic polyp images detection," *BMCMI*, vol. 20, pp. 1–14, 2020.
- [53] A. de Maissin, R. Vallée, M. Flamant, M. Fondain-Bossiere, C. Le Berre, A. Coutrot, N. Normand, H. Mouchère, S. Coudol, C. Trang *et al.*, "Multi-expert annotation of crohn's disease images of the small bowel for automatic detection using a convolutional recurrent attention neural network," *EIO*, vol. 9, no. 07, pp. E1136–E1144, 2021.
- [54] Z. Kong, M. He, Q. Luo, X. Huang, P. Wei, Y. Cheng, L. Chen, Y. Liang, Y. Lu, X. Li *et al.*, "Multi-task classification and segmentation for explicable capsule endoscopy diagnostics," *FMOLB*, vol. 8, p. 614277, 2021.
- [55] M. Misawa, S.-e. Kudo, Y. Mori, K. Hotta, K. Ohtsuka, T. Matsuda, S. Saito, T. Kudo, T. Baba, F. Ishida *et al.*, "Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video)," *GIE*, vol. 93, no. 4, pp. 960–967, 2021.
- [56] D. Jha, P. H. Smedsrud, D. Johansen, T. de Lange, H. D. Johansen, P. Halvorsen, and M. A. Riegler, "A comprehensive study on colorectal polyp segmentation with resnet++, conditional random field and test-time augmentation," *IEEE JBHI*, vol. 25, no. 6, pp. 2029–2040, 2021.
- [57] P. H. Smedsrud, V. Thambawita, S. A. Hicks, H. Gjestang, O. O. Nedrejord, E. Næss, H. Borgli, D. Jha, T. J. D. Berstad, S. L. Eskeland *et al.*, "Kvasir-capsule, a video capsule endoscopy dataset," *SData*, vol. 8, no. 1, p. 142, 2021.
- [58] K. Li, M. I. Fathan, K. Patel, T. Zhang, C. Zhong, A. Bansal, A. Rastogi, J. S. Wang, and G. Wang, "Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations," *PONE*, vol. 16, no. 8, p. e0255809, 2021.
- [59] J. Cychnerski, T. Dziubich, and A. Brzeski, "Ers: a novel comprehensive endoscopy image dataset for machine learning, compliant with the mst 3.0 specification," *arXiv preprint arXiv:2201.08746*, 2022.
- [60] Y. Tian, G. Pang, F. Liu, Y. Liu, C. Wang, Y. Chen, J. Verjans, and G. Carneiro, "Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection," in *MICCAI*, 2022.
- [61] F. J. P. Montalbo, "Diagnosing gastrointestinal diseases from endoscopy images through a multi-fused cnn with auxiliary layers, alpha dropouts, and a fusion residual block," *BSPC*, vol. 76, p. 103683, 2022.
- [62] S. Ali and N. Ghatwary, "Endoscopic computer vision challenges 2.0," 2022. [Online]. Available: <https://endocv2022.grand-challenge.org/>
- [63] G.-P. Ji, G. Xiao, Y.-C. Chou, D.-P. Fan, K. Zhao, G. Chen, and L. Van Gool, "Video polyp segmentation: A deep learning perspective," *MIR*, vol. 19, no. 6, pp. 531–549, 2022.
- [64] V. Thambawita, P. Salehi, S. A. Sheshkal, S. A. Hicks, H. L. Hammer, S. Parasa, T. d. Lange, P. Halvorsen, and M. A. Riegler, "Singan-seg: Synthetic training data generation

- for medical image segmentation,” *PONE*, vol. 17, no. 5, p. e0267976, 2022.
- [65] D. Fitting, A. Krenzer, J. Troya, M. Banck, B. Sudarevic, M. Brand, W. Böck, W. G. Zoller, T. Rösch, F. Puppe *et al.*, “A video based benchmark data set (endotest) to evaluate computer-aided polyp detection systems,” *SJG*, vol. 57, no. 11, pp. 1397–1403, 2022.
- [66] S. Hicks, A. Storås, P. Halvorsen, T. de Lange, M. Riegler, and V. Thambawita, “Overview of imageclef medical 2023—medical visual question answering for gastrointestinal tract,” in *CLEF (Working notes)*, 2023.
- [67] D. Jha, V. Sharma, N. Dasu, N. K. Tomar, S. Hicks, M. Bhuyan, P. K. Das, M. A. Riegler, P. Halvorsen, T. de Lange *et al.*, “Gastrovision: A multi-class endoscopy image dataset for computer aided gastrointestinal disease detection,” in *ICML-W*, 2023.
- [68] G. Ren, M. Lazarou, J. Yuan, and T. Stathaki, “Towards automated polyp segmentation using weakly-and semi-supervised learning and deformable transformers,” in *IEEE CVPR-W*, 2023.
- [69] G. Polat, H. T. Kani, I. Ergenc, Y. Ozen Alahdab, A. Temizel, and O. Atug, “Improving the computer-aided estimation of ulcerative colitis severity according to mayo endoscopic score by using regression-based deep learning,” *IBD*, vol. 29, no. 9, pp. 1431–1439, 2023.
- [70] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O. E. Salem, D. Lamarque, C. Daul, M. A. Riegler, K. V. Anonsen *et al.*, “A multi-centre polyp detection and segmentation dataset for generalisability assessment,” *SData*, vol. 10, no. 1, p. 75, 2023.
- [71] D. Wang, X. Wang, L. Wang, M. Li, Q. Da, X. Liu, X. Gao, J. Shen, J. He, T. Shen *et al.*, “A real-world dataset and benchmark for foundation model adaptation in medical image classification,” *SData*, vol. 10, no. 1, p. 574, 2023.
- [72] H. Khan, Ali; Malik, “Gastrointestinal bleeding wce images dataset,” 2023, doi: [10.17632/8pbj274w.1](https://doi.org/10.17632/8pbj274w.1).
- [73] C. Biffi, G. Antonelli, S. Bernhofer, C. Hassan, D. Hirata, M. Iwatate, A. Maieron, P. Salvagnini, and A. Cherubini, “Real-colon: A dataset for developing real-world ai applications in colonoscopy,” *SData*, vol. 11, no. 1, p. 539, 2024.
- [74] Z. Xu, J. Rittscher, and S. Ali, “Ssl-cpcd: Self-supervised learning with composite pretext-class discrimination for improved generalisability in endoscopic image analysis,” *IEEE TMI*, 2024, doi: [10.1109/TMI.2024.3411933](https://doi.org/10.1109/TMI.2024.3411933).
- [75] S. Gautam, A. Storås, C. Midoglu, S. A. Hicks, V. Thambawita, P. Halvorsen, and M. A. Riegler, “Kvasir-vqa: A text-image pair gi tract dataset,” in *ACM MM-W*, 2024.
- [76] P. Handa, A. Mahbod, F. Schwarzahans, R. Woitek, N. Goel, D. Chhabra, S. Jha, M. Dhir, D. Gunjan, J. Kakarla *et al.*, “Capsule vision 2024 challenge: Multi-class abnormality classification for video capsule endoscopy,” in *CVIP*, 2024.
- [77] L. Ruiz, F. Sierra-Jerez, J. Ruiz, and F. Martinez, “Colon: The largest colonoscopy long sequence public database,” *arXiv preprint arXiv:2403.00663*, 2024.
- [78] P. Handa, M. Dhir, A. Mahbod, F. Schwarzahans, R. Woitek, N. Goel, and D. Gunjan, “Wcebleedgen: A wireless capsule endoscopy dataset and its benchmarking for automatic bleeding classification, detection, and segmentation,” *arXiv preprint arXiv:2408.12466*, 2024.
- [79] D. Jha, N. K. Tomar, V. Sharma, Q.-H. Trinh, K. Biswas, H. Pan, R. K. Jha, G. Durak, A. Hann, J. Varkey *et al.*, “Polypdb: A curated multi-center dataset for development of ai algorithms in colonoscopy,” *arXiv preprint arXiv:2409.00045*, 2024.
- [80] M. Sivak, “Gastrointestinal endoscopy: past and future,” *Gut*, vol. 55, no. 8, pp. 1061–1064, 2006.
- [81] T. M. Berzin and E. J. Topol, “Adding artificial intelligence to gastrointestinal endoscopy,” *The Lancet*, vol. 395, no. 10223, p. 485, 2020.
- [82] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, “Wireless capsule endoscopy,” *Nature*, vol. 405, no. 6785, pp. 417–417, 2000.
- [83] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, “Endonet: a deep architecture for recognition tasks on laparoscopic videos,” *IEEE TMI*, vol. 36, no. 1, pp. 86–97, 2016.
- [84] S. Shao, Z. Pei, W. Chen, W. Zhu, X. Wu, D. Sun, and B. Zhang, “Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue,” *MedIA*, vol. 77, p. 102338, 2022.
- [85] J. C. Á. Cerón, G. O. Ruiz, L. Chang, and S. Ali, “Real-time instance segmentation of surgical instruments using attention and multi-scale feature fusion,” *MedIA*, vol. 81, p. 102569, 2022.
- [86] Y. Blau, D. Freedman, V. Dashinsky, R. Goldenberg, and E. Rivlin, “Unsupervised 3d shape coverage estimation with applications to colonoscopy,” in *IEEE ICCV-W*, 2021.
- [87] Y. Zhang, S. Wang, R. Ma, S. K. McGill, J. G. Rosenman, and S. M. Pizer, “Lighting enhancement aids reconstruction of colonoscopic surfaces,” in *IPMI*, 2021.
- [88] D.-P. Fan, G.-P. Ji, P. Xu, M.-M. Cheng, C. Sakaridis, and L. Van Gool, “Advances in deep concealed scene understanding,” *VInt*, vol. 1, no. 1, p. 16, 2023.
- [89] W. M. de Vos and E. A. de Vos, “Role of the intestinal microbiome in health and disease: from correlation to causation,” *Nutr. Rev.*, vol. 70, no. suppl_1, pp. S45–S56, 2012.
- [90] Y. Li and P. Agarwal, “A pathway-based view of human diseases and disease relationships,” *PONE*, vol. 4, no. 2, p. e4346, 2009.
- [91] B. Veauthier and J. R. Hornecker, “Crohn’s disease: diagnosis and management,” *AFP*, vol. 98, no. 11, pp. 661–669, 2018.
- [92] L. Yang, H. Jiang, Q. Song, and J. Guo, “A survey on long-tailed visual recognition,” *IJCV*, vol. 130, no. 7, pp. 1837–1872, 2022.
- [93] J. Wu, X. Li, S. X. H. Yuan, H. Ding, Y. Yang, X. Li, J. Zhang, Y. Tong, X. Jiang, B. Ghanem *et al.*, “Towards open vocabulary learning: A survey,” *IEEE TPAMI*, vol. 46, no. 7, pp. 5092–5113, 2024.
- [94] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, “Large language models encode clinical knowledge,” *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.
- [95] J. Zhang, Y. Xie, Y. Xia, and C. Shen, “Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets,” in *IEEE CVPR*, 2021.
- [96] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, “Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis,” *MedIA*, vol. 65, p. 101759, 2020.
- [97] H.-F. Yu, P. Jain, P. Kar, and I. Dhillon, “Large-scale multi-label learning with missing labels,” in *ICML*, 2014.
- [98] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Gir-

- shick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in *ACM MM*, 2014.
- [99] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *IEEE CVPR*, 2017.
- [100] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE CVPR*, 2016.
- [101] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *ICLR*, 2021.
- [102] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *IEEE CVPR*, 2018.
- [103] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *ECCV*, 2018.
- [104] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *IEEE ICCV*, 2015.
- [105] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *IEEE CVPR*, 2016.
- [106] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *IEEE CVPR*, 2017.
- [107] Y. Yuan, W. Qin, B. Ibragimov, B. Han, and L. Xing, “Riisensenet: rotation-invariant and image similarity constrained densely connected convolutional network for polyp detection,” in *MICCAI*, 2018.
- [108] Y. Tian, G. Maicas, L. Z. C. T. Pu, R. Singh, J. W. Verjans, and G. Carneiro, “Few-shot anomaly detection for polyp frames from colonoscopy,” in *MICCAI*, 2020.
- [109] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, “Two-stream deep feature modelling for automated video endoscopy data analysis,” in *MICCAI*, 2020.
- [110] G. Carneiro, L. Z. C. T. Pu, R. Singh, and A. Burt, “Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy,” *MedIA*, vol. 62, p. 101653, 2020.
- [111] X. Guo and Y. Yuan, “Semi-supervised wce image classification with adaptive aggregated attention,” *MedIA*, vol. 64, p. 101733, 2020.
- [112] W. Ma, Y. Zhu, R. Zhang, J. Yang, Y. Hu, Z. Li, and L. Xiang, “Toward clinically assisted colorectal polyp recognition via structured cross-modal representation consistency,” in *MICCAI*, 2022.
- [113] K.-N. Wang, Y. He, S. Zhuang, J. Miao, X. He, P. Zhou, G. Yang, G.-Q. Zhou, and S. Li, “Ffcnet: Fourier transform-based frequency learning and complex convolutional network for colon disease classification,” in *MICCAI*, 2022.
- [114] K.-N. Wang, S. Zhuang, Q.-Y. Ran, P. Zhou, J. Hua, G.-Q. Zhou, and X. He, “Dlgnnet: A dual-branch lesion-aware network with the supervised gaussian mixture model for colon lesions classification in colonoscopy images,” *MedIA*, vol. 87, p. 102832, 2023.
- [115] G. Yue, P. Wei, Y. Liu, Y. Luo, J. Du, and T. Wang, “Automated endoscopic image classification via deep neural network with class imbalance loss,” *IEEE TIM*, vol. 72, pp. 1–11, 2023.
- [116] Y. Luo, X. Guo, L. Liu, and Y. Yuan, “Dynamic attribute-guided few-shot open-set network for medical image diagnosis,” *ESWA*, vol. 251, p. 124098, 2024.
- [117] H. Itoh, H. R. Roth, L. Lu, M. Oda, M. Misawa, Y. Mori, S.-e. Kudo, and K. Mori, “Towards automated colonoscopy diagnosis: binary polyp size estimation via unsupervised depth learning,” in *MICCAI*, 2018.
- [118] J. Schmidhuber, S. Hochreiter *et al.*, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [119] M. F. Byrne, N. Chapados, F. Soudan, C. Oertel, M. L. Pérez, R. Kelly, N. Iqbal, F. Chandelier, and D. K. Rex, “Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model,” *Gut*, vol. 68, no. 1, pp. 94–100, 2019.
- [120] A. Tamhane, T. Mida, E. Posner, and M. Bouhnik, “Colonoscopy landmark detection using vision transformers,” in *MICCAI-W*, 2022.
- [121] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “Yolov4: Optimal speed and accuracy of object detection,” *arXiv preprint arXiv:2004.10934*, 2020.
- [122] J. Redmon, “Yolov3: An incremental improvement,” *arXiv preprint arXiv:1804.02767*, 2018.
- [123] M. Tan, R. Pang, and Q. V. Le, “Efficientdet: Scalable and efficient object detection,” in *IEEE CVPR*, 2020.
- [124] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, “Siamrpn++: Evolution of siamese visual tracking with very deep networks,” in *IEEE CVPR*, 2019.
- [125] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *ICLR*, 2015.
- [126] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in *ECCV*, 2016.
- [127] X. Yang, Q. Wei, C. Zhang, K. Zhou, L. Kong, and W. Jiang, “Colon polyp detection and segmentation based on improved mrcnn,” *IEEE TIM*, vol. 70, pp. 1–10, 2020.
- [128] X. Liu, X. Guo, Y. Liu, and Y. Yuan, “Consolidated domain adaptive detection and localization framework for cross-device colonoscopic images,” *MedIA*, vol. 71, p. 102052, 2021.
- [129] H. A. Qadir, Y. Shin, J. Solhusvik, J. Bergsland, L. Aabakken, and I. Balasingham, “Toward real-time polyp detection using fully cnns for 2d gaussian shapes prediction,” *MedIA*, vol. 68, p. 101897, 2021.
- [130] X. Liu, W. Li, and Y. Yuan, “Intervention & interaction federated abnormality detection with noisy clients,” in *MICCAI*, 2022.
- [131] I. Pacal, A. Karaman, D. Karaboga, B. Akay, A. Basturk, U. Nalbantoglu, and S. Coskun, “An efficient real-time colonic polyp detection with yolo algorithms trained by using negative samples and large datasets,” *CIBM*, vol. 141, p. 105031, 2022.
- [132] X. Liu and Y. Yuan, “A source-free domain adaptive polyp detection framework with style diversification flow,” *IEEE TMI*, vol. 41, no. 7, pp. 1897–1908, 2022.
- [133] R. Gong, S. He, T. Tian, J. Chen, Y. Hao, and C. Qiao, “Frcnn-aa-cif: An automatic detection model of colon polyps based on attention awareness and context information fusion,” *CIBM*, vol. 158, p. 106787, 2023.
- [134] M. R. Haugland, H. A. Qadir, and I. Balasingham, “Deep learning for improved polyp detection from synthetic narrow-band imaging,” in *SPIE Med. Imaging*, 2023.
- [135] W. Li, X. Liu, and Y. Yuan, “Scan++: Enhanced semantic conditioned adaptation for domain adaptive object detection,” *IEEE TMM*, vol. 25, pp. 7051–7061, 2023.

- [136] X. Pan, Y. Mu, C. Ma, and Q. He, “Tfcnet: A texture-aware and fine-grained feature compensated polyp detection network,” *CIBM*, vol. 171, p. 108144, 2024.
- [137] X. Liu, W. Li, and Y. Yuan, “Decoupled unbiased teacher for source-free domain adaptive medical object detection,” *IEEE TNNLS*, vol. 35, no. 6, pp. 7287–7298, 2024.
- [138] Tajbaksh, Nima and Gurudu, Suryakanth R and Liang, Jianming, “A comprehensive computer-aided polyp detection system for colonoscopy videos,” in *IPMI*, 2015.
- [139] L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng, “Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos,” *IEEE JBHI*, vol. 21, no. 1, pp. 65–75, 2016.
- [140] X. Mo, K. Tao, Q. Wang, and G. Wang, “An efficient approach for polyps detection in endoscopic videos based on faster r-cnn,” in *IEEE ICPR*, 2018.
- [141] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *NeurIPS*, 2015.
- [142] H. A. Qadir, I. Balasingham, J. Solhusvik, J. Bergsland, L. Aabakken, and Y. Shin, “Improving automatic polyp detection using cnn by exploiting temporal dependency in colonoscopy video,” *IEEE JBHI*, vol. 24, no. 1, pp. 180–193, 2019.
- [143] Z. Zhang, H. Shang, H. Zheng, X. Wang, J. Wang, Z. Sun, J. Huang, and J. Yao, “Asynchronous in parallel detection and tracking (aipdt): Real-time robust polyp detection,” in *MICCAI*, 2020.
- [144] L. Wu, Z. Hu, Y. Ji, P. Luo, and S. Zhang, “Multi-frame collaboration for effective endoscopic video polyp detection via spatial-temporal feature transformation,” in *MICCAI*, 2021.
- [145] T. Yu, N. Lin, X. Zhang, Y. Pan, H. Hu, W. Zheng, J. Liu, W. Hu, H. Duan, and J. Si, “An end-to-end tracking method for polyp detectors in colonoscopy videos,” *AIIM*, vol. 131, p. 102363, 2022.
- [146] D. Wang, X. Wang, S. Wang, and Y. Yin, “Explainable multi-task shapley explanation networks for real-time polyp diagnosis in videos,” *IEEE TII*, vol. 19, no. 6, pp. 7780–7789, 2022.
- [147] Y. Jiang, Z. Zhang, R. Zhang, G. Li, S. Cui, and Z. Li, “Yona: You only need one adjacent reference-frame for accurate and fast video polyp detection,” in *MICCAI*, 2023.
- [148] Y. Intrator, N. Aizenberg, A. Livne, E. Rivlin, and R. Goldenberg, “Self-supervised polyp re-identification in colonoscopy,” in *MICCAI*, 2023.
- [149] Y. Jiang, Z. Zhang, J. Wei, C.-M. Feng, G. Li, X. Wan, S. Cui, and Z. Li, “Let video teaches you more: Video-to-image knowledge distillation using detection transformer for medical video lesion detection,” in *BIBM*, 2024.
- [150] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, “Vision mamba: Efficient visual representation learning with bidirectional state space model,” in *ICML*, 2024.
- [151] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *ICML*, 2021.
- [152] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *ECCV*, 2016.
- [153] J. Bernal, N. Tajbaksh, F. J. Sanchez, B. J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham *et al.*, “Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge,” *IEEE TMI*, vol. 36, no. 6, pp. 1231–1249, 2017.
- [154] Z. Zhang, Q. Liu, and Y. Wang, “Road extraction by deep residual u-net,” *IEEE GRSL*, vol. 15, no. 5, pp. 749–753, 2018.
- [155] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, “Res2net: A new multi-scale backbone architecture,” *IEEE TPAMI*, vol. 43, no. 2, pp. 652–662, 2019.
- [156] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in *ICML*, 2021.
- [157] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *ICML*, 2019.
- [158] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *ECCV*, 2018.
- [159] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, “Pvt v2: Improved baselines with pyramid vision transformer,” *CVMJ*, vol. 8, no. 3, pp. 415–424, 2022.
- [160] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, “Cvt: Introducing convolutions to vision transformers,” in *IEEE ICCV*, 2021.
- [161] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “Segformer: Simple and efficient design for semantic segmentation with transformers,” in *NeurIPS*, 2021.
- [162] S. Chen, E. Xie, C. Ge, R. Chen, D. Liang, and P. Luo, “Cyclemlp: A mlp-like architecture for dense prediction,” in *ICLR*, 2022.
- [163] L. Chen, T. Yang, X. Zhang, W. Zhang, and J. Sun, “Points as queries: Weakly semi-supervised object detection by points,” in *IEEE CVPR*, 2021.
- [164] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, “Segnext: Rethinking convolutional attention design for semantic segmentation,” in *NeurIPS*, 2022.
- [165] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *IEEE ICCV*, 2021.
- [166] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, “Segment anything,” in *IEEE ICCV*, 2023.
- [167] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryal, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, “Sam 2: Segment anything in images and videos,” *arXiv preprint arXiv:2408.00714*, 2024.
- [168] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE TPAMI*, vol. 40, no. 4, pp. 834–848, 2017.
- [169] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, “Deep high-resolution representation learning for visual recognition,” *IEEE TPAMI*, vol. 43, no. 10, pp. 3349–3364, 2020.
- [170] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A convnet for the 2020s,” in *IEEE CVPR*, 2022.
- [171] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, “Masked-attention mask transformer for universal image segmentation,” in *IEEE CVPR*, 2022, pp. 1290–1299.

- [172] Y. Yuan, D. Li, and M. Q.-H. Meng, “Automatic polyp detection via a novel unified bottom-up and top-down saliency approach,” *IEEE JBHI*, vol. 22, no. 4, pp. 1250–1260, 2017.
- [173] Y. Fang, C. Chen, Y. Yuan, and K.-y. Tong, “Selective feature aggregation network with area-boundary constraints for polyp segmentation,” in *MICCAI*, 2019.
- [174] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, “Resunet++: An advanced architecture for medical image segmentation,” in *IEEE ISM*, 2019.
- [175] R. Zhang, G. Li, Z. Li, S. Cui, D. Qian, and Y. Yu, “Adaptive context selection for polyp segmentation,” in *MICCAI*, 2020.
- [176] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, “Pranet: Parallel reverse attention network for polyp segmentation,” in *MICCAI*, 2020.
- [177] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen, “Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps,” *MedIA*, vol. 60, p. 101619, 2020.
- [178] H. Wu, J. Zhong, W. Wang, Z. Wen, and J. Qin, “Precise yet efficient semantic calibration and refinement in convnets for real-time polyp segmentation from colonoscopy videos,” in *AAAI*, 2021.
- [179] Y. Meng, H. Zhang, D. Gao, Y. Zhao, X. Yang, X. Qian, X. Huang, and Y. Zheng, “Bi-gcn: Boundary-aware input-dependent graph convolution network for biomedical image segmentation,” in *BMVC*, 2021.
- [180] H. Wu, G. Chen, Z. Wen, and J. Qin, “Collaborative and adversarial learning of focused and dispersive representations for semi-supervised polyp segmentation,” in *IEEE ICCV*, 2021.
- [181] T.-C. Nguyen, T.-P. Nguyen, G.-H. Diep, A.-H. Tran-Dinh, T. V. Nguyen, and M.-T. Tran, “Ccbnet: cascading context and balancing attention for polyp segmentation,” in *MICCAI*, 2021.
- [182] Y. Tian, G. Pang, F. Liu, Y. Chen, S. H. Shin, J. W. Verjans, R. Singh, and G. Carneiro, “Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images,” in *MICCAI*, 2021.
- [183] Y. Shen, X. Jia, and M. Q.-H. Meng, “Hrenet: A hard region enhancement network for polyp segmentation,” in *MICCAI*, 2021.
- [184] M. Cheng, Z. Kong, G. Song, Y. Tian, Y. Liang, and J. Chen, “Learnable oriented-derivative network for polyp segmentation,” in *MICCAI*, 2021.
- [185] X. Zhao, L. Zhang, and H. Lu, “Automatic polyp segmentation via multi-scale subtraction network,” in *MICCAI*, 2021.
- [186] J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, and S. Cui, “Shallow attention network for polyp segmentation,” in *MICCAI*, 2021.
- [187] Y. Zhang, H. Liu, and Q. Hu, “Transfuse: Fusing transformers and cnns for medical image segmentation,” in *MICCAI*, 2021.
- [188] T. Kim, H. Lee, and D. Kim, “Uacnet: Uncertainty augmented context attention for polyp segmentation,” in *ACM MM*, 2021.
- [189] C. Yang, X. Guo, M. Zhu, B. Ibragimov, and Y. Yuan, “Mutual-prototype adaptation for cross-domain polyp segmentation,” *IEEE JBHI*, vol. 25, no. 10, pp. 3886–3897, 2021.
- [190] X. Guo, C. Yang, and Y. Yuan, “Dynamic-weighting hierarchical segmentation network for medical images,” *MedIA*, vol. 73, p. 102196, 2021.
- [191] X. Du, X. Xu, and K. Ma, “Icgnnet: Integration context-based reverse-contour guidance network for polyp segmentation,” in *IJCAI*, 2022.
- [192] J. Wei, Y. Hu, G. Li, S. Cui, S. Kevin Zhou, and Z. Li, “Boxpolyp: Boost generalized polyp segmentation using extra coarse bounding box annotations,” in *MICCAI*, 2022.
- [193] R. Zhang, P. Lai, X. Wan, D.-J. Fan, F. Gao, X.-J. Wu, and G. Li, “Lesion-aware dynamic kernel for polyp segmentation,” in *MICCAI*, 2022.
- [194] L. Cai, M. Wu, L. Chen, W. Bai, M. Yang, S. Lyu, and Q. Zhao, “Using guided self-attention with local information for polyp segmentation,” in *MICCAI*, 2022.
- [195] J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, and S. Song, “Stepwise feature fusion: Local guides global,” in *MICCAI*, 2022.
- [196] Y. Shen, Y. Lu, X. Jia, F. Bai, and M. Q.-H. Meng, “Task-relevant feature replenishment for cross-centre polyp segmentation,” in *MICCAI*, 2022.
- [197] D. Wang, S. Chen, Q. Chen, Y. Cao, B. Liu, X. Liu, and X. Sun, “Afp-mask: Anchor-free polyp instance segmentation in colonoscopy,” *IEEE JBHI*, vol. 26, no. 7, pp. 2995–3006, 2022.
- [198] G. Yue, W. Han, B. Jiang, T. Zhou, R. Cong, and T. Wang, “Boundary constraint network with cross layer feature integration for polyp segmentation,” *IEEE JBHI*, vol. 26, no. 8, pp. 4090–4099, 2022.
- [199] Y. Lin, J. Wu, G. Xiao, J. Guo, G. Chen, and J. Ma, “Bsca-net: Bit slicing context attention network for polyp segmentation,” *PR*, vol. 132, p. 108917, 2022.
- [200] J.-H. Shi, Q. Zhang, Y.-H. Tang, and Z.-Q. Zhang, “Polyp-mixer: An efficient context-aware mlp-based paradigm for polyp segmentation,” *IEEE TCSVT*, vol. 33, no. 1, pp. 30–42, 2022.
- [201] H. Wu, W. Xie, J. Lin, and X. Guo, “Acl-net: semi-supervised polyp segmentation via affinity contrastive learning,” in *AAAI*, 2023.
- [202] J. Wei, Y. Hu, S. Cui, S. K. Zhou, and Z. Li, “Weakpolyp: You only look bounding box for polyp segmentation,” in *MICCAI*, 2023.
- [203] T. Ling, C. Wu, H. Yu, T. Cai, D. Wang, Y. Zhou, M. Chen, and K. Ding, “Probabilistic modeling ensemble vision transformer improves complex polyp segmentation,” in *MICCAI*, 2023.
- [204] A. Wang, M. Xu, Y. Zhang, M. Islam, and H. Ren, “S²me: Spatial-spectral mutual teaching and ensemble learning for scribble-supervised polyp segmentation,” in *MICCAI*, 2023.
- [205] Y. Su, Y. Shen, J. Ye, J. He, and J. Cheng, “Revisiting feature propagation and aggregation in polyp segmentation,” in *MICCAI*, 2023.
- [206] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, “Polyp-pvt: Polyp segmentation with pyramid vision transformers,” *CAAI AIR*, vol. 2, p. 9150015, 2023.
- [207] J. Wang and C. Chen, “Unsupervised adaptation of polyp segmentation models via coarse-to-fine self-supervision,” in *IPMI*, 2023.
- [208] Q. Jin, H. Hou, G. Zhang, and Z. Li, “Fegnet: A feedback enhancement gate network for automatic polyp segmentation,” *IEEE JBHI*, vol. 27, no. 7, pp. 3420–3430, 2023.
- [209] J. Du, K. Guan, P. Liu, Y. Li, and T. Wang, “Boundary-sensitive loss function with location constraint for hard region segmentation,” *IEEE JBHI*, vol. 27, no. 2, pp. 992–1003, 2023.
- [210] Y. Shi, H. Wang, H. Ji, H. Liu, Y. Li, N. He, D. Wei, Y. Huang, Q. Dai, J. Wu *et al.*, “A deep weakly semi-supervised framework

- for endoscopic lesion segmentation,” *MedIA*, vol. 90, p. 102973, 2023.
- [211] G.-P. Ji, D.-P. Fan, Y.-C. Chou, D. Dai, A. Liniger, and L. Van Gool, “Deep gradient learning for efficient camouflaged object detection,” *MIR*, vol. 20, no. 1, pp. 92–108, 2023.
- [212] T. Zhou, Y. Zhou, K. He, C. Gong, J. Yang, H. Fu, and D. Shen, “Cross-level feature aggregation network for polyp segmentation,” *PR*, vol. 140, p. 109555, 2023.
- [213] S. Jain, R. Atale, A. Gupta, U. Mishra, A. Seal, A. Ojha, J. Kuncewicz, and O. Krejcar, “Coinnet: A convolution-involution network with a novel statistical attention for automatic polyp segmentation,” *IEEE TMI*, vol. 42, no. 12, pp. 3987–4000, 2023.
- [214] N. K. Tomar, D. Jha, M. A. Riegler, H. D. Johansen, D. Johansen, J. Rittscher, P. Halvorsen, and S. Ali, “Fanet: A feedback attention network for improved biomedical image segmentation,” *IEEE TNNLS*, vol. 34, no. 11, pp. 9375–9388, 2023.
- [215] H. Shao, Q. Zeng, Q. Hou, and J. Yang, “Mcanet: Medical image segmentation with multi-scale cross-axis attention,” *arXiv preprint arXiv:2312.08866*, 2023.
- [216] H. Shao, Y. Zhang, and Q. Hou, “Polyper: Boundary sensitive polyp segmentation,” in *AAAI*, 2024.
- [217] M. M. Rahman, M. Munir, and R. Marculescu, “Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation,” in *IEEE CVPR*, 2024.
- [218] R. Schön, J. Lorenz, K. Ludwig, and R. Lienhart, “Adapting the segment anything model during usage in novel situations,” in *IEEE CVPR*, 2024, pp. 3616–3626.
- [219] L. Xie, M. Lin, T. Luan, C. Li, Y. Fang, Q. Shen, and Z. Wu, “Mh-pflid: Model heterogeneous personalized federated learning via injection and distillation for medical data analysis,” in *ICML*, 2024.
- [220] H. Li, D. Zhang, J. Yao, L. Han, Z. Li, and J. Han, “Asps: Augmented segment anything model for polyp segmentation,” in *MICCAI*, 2024.
- [221] Z. Xu, F. Tang, Z. Chen, Z. Zhou, W. Wu, Y. Yang, Y. Liang, J. Jiang, X. Cai, and J. Su, “Polyp-Mamba: Polyp Segmentation with Visual Mamba,” in *MICCAI*, 2024.
- [222] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, “Vmamba: Visual state space model,” *NeurIPS*, 2024.
- [223] J. Chai, Z. Luo, J. Gao, L. Dai, Y. Lai, and S. Li, “QueryNet: A Unified Framework for Accurate Polyp Segmentation and Detection,” in *MICCAI*, 2024.
- [224] W. Wang, H. Sun, and X. Wang, “LSSNet: A Method for Colon Polyp Segmentation Based on Local Feature Supplementation and Shallow Feature Supplementation,” in *MICCAI*, 2024.
- [225] X. Zhou and T. Chen, “Bsbp-rwkv: Background suppression with boundary preservation for efficient medical image segmentation,” in *ACM MM*, 2024.
- [226] B. Peng, E. Alcaide, Q. G. Anthony, A. Albalak, S. Arcadinho, S. Biderman, H. Cao, X. Cheng, M. N. Chung, L. Derczynski, X. Du, M. Grella, K. K. GV, X. He, H. Hou, P. Kazienko, J. Kocon, J. Kong, B. Koptyra, H. Lau, J. Lin, K. S. I. Mantri, F. Mom, A. Saito, G. Song, X. Tang, J. S. Wind, S. Woźniak, Z. Zhang, Q. Zhou, J. Zhu, and R.-J. Zhu, “RWKV: Reinventing RNNs for the transformer era,” in *EMNLP*, 2023.
- [227] C. Wang, L. Wang, N. Wang, X. Wei, T. Feng, M. Wu, Q. Yao, and R. Zhang, “Cfatransunet: Channel-wise cross fusion attention and transformer for 2d medical image segmentation,” *CIBM*, vol. 168, p. 107803, 2024.
- [228] X. Jia, Y. Shen, J. Yang, R. Song, W. Zhang, M. Q.-H. Meng, J. C. Liao, and L. Xing, “Polypmixnet: Enhancing semi-supervised polyp segmentation with polyp-aware augmentation,” *CIBM*, vol. 170, p. 108006, 2024.
- [229] Z. Zhang, Y. Li, and B.-S. Shin, “Generalizable polyp segmentation via randomized global illumination augmentation,” *IEEE JBHI*, vol. 28, no. 4, pp. 2138–2151, 2024.
- [230] M. Wang, X. An, Z. Pei, N. Li, L. Zhang, G. Liu, and D. Ming, “An efficient multi-task synergetic network for polyp segmentation and classification,” *IEEE JBHI*, vol. 28, no. 3, pp. 1228–1239, 2024.
- [231] L. Yang, Y. Gu, G. Bian, and Y. Liu, “Msde-net: A multi-scale dual-encoding network for surgical instrument segmentation,” *IEEE JBHI*, vol. 28, no. 7, pp. 4072–4083, 2024.
- [232] G.-P. Ji, J. Zhang, D. Campbell, H. Xiong, and N. Barnes, “Rethinking polyp segmentation from an out-of-distribution perspective,” *MIR*, vol. 21, no. 4, pp. 631–639, 2024.
- [233] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, “Segment anything in medical images,” *NComms*, vol. 15, no. 1, p. 654, 2024.
- [234] Z. Liu, S. Zheng, X. Sun, Z. Zhu, Y. Zhao, X. Yang, and Y. Zhao, “The devil is in the boundary: Boundary-enhanced polyp segmentation,” *IEEE TCSVT*, vol. 34, no. 7, pp. 5414–5423, 2024.
- [235] Z. Lu, Y. Zhang, Y. Zhou, Y. Wu, and T. Zhou, “Domain-interactive contrastive learning and prototype-guided self-training for cross-domain polyp segmentation,” *IEEE TMI*, 2024.
- [236] J. Gao, Q. Lao, Q. Kang, P. Liu, C. Du, K. Li, and L. Zhang, “Boosting your context by dual similarity checkup for in-context learning medical image segmentation,” *IEEE TMI*, 2024.
- [237] C. Fan, H. Yu, L. Wang, Y. Huang, L. Wang, and X. Jia, “Slicemamba with neural architecture search for medical image segmentation,” *arXiv preprint arXiv:2407.08481*, 2024.
- [238] J. Xie, R. Liao, Z. Zhang, S. Yi, Y. Zhu, and G. Luo, “Promamba: Prompt-mamba for polyp segmentation,” *arXiv preprint arXiv:2403.13660*, 2024.
- [239] X. Xiong, Z. Wu, S. Tan, W. Li, F. Tang, Y. Chen, S. Li, J. Ma, and G. Li, “Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation,” *arXiv preprint arXiv:2408.08870*, 2024.
- [240] C. Li, X. Liu, W. Li, C. Wang, H. Liu, and Y. Yuan, “U-kan makes strong backbone for medical image segmentation and generation,” in *AAAI*, 2025.
- [241] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, “Kan: Kolmogorov-arnold networks,” *arXiv preprint arXiv:2404.19756*, 2024.
- [242] J. G.-B. Puyal, K. K. Bhatia, P. Brandao, O. F. Ahmad, D. Toth, R. Kader, L. Lovat, P. Mountney, and D. Stoyanov, “Endoscopic polyp segmentation using a hybrid 2d/3d cnn,” in *MICCAI*, 2020.
- [243] G.-P. Ji, Y.-C. Chou, D.-P. Fan, G. Chen, H. Fu, D. Jha, and L. Shao, “Progressively normalized self-attention network for video polyp segmentation,” in *MICCAI*, 2021.
- [244] X. Zhao, Z. Wu, S. Tan, D.-J. Fan, Z. Li, X. Wan, and G. Li, “Semi-supervised spatial temporal attention network for video polyp segmentation,” in *MICCAI*, 2022.

- [245] X. Li, J. Xu, Y. Zhang, R. Feng, R.-W. Zhao, T. Zhang, X. Lu, and S. Gao, “Tccnet: Temporally consistent context-free network for semi-supervised video polyp segmentation.” in *IJCAI*, 2022.
- [246] J. G.-B. Puyal, P. Brandao, O. F. Ahmad, K. K. Bhatia, D. Toth, R. Kader, L. Lovat, P. Mountney, and D. Stoyanov, “Polyp detection on video colonoscopy using a hybrid 2d/3d cnn,” *MedIA*, vol. 82, p. 102625, 2022.
- [247] Z. Fang, X. Guo, J. Lin, H. Wu, and J. Qin, “An embedding-unleashing video polyp segmentation framework via region linking and scale alignment,” in *AAAI*, 2024.
- [248] H. Xu, Y. Yang, A. I. Aviles-Rivero, G. Yang, J. Qin, and L. Zhu, “Lgrnet: Local-global reciprocal network for uterine fibroid segmentation in ultrasound videos,” in *MICCAI*, 2024.
- [249] Q. Hu, Z. Yi, Y. Zhou, F. Peng, M. Liu, Q. Li, and Z. Wang, “Sali: Short-term alignment and long-term interaction network for colonoscopy video polyp segmentation,” in *MICCAI*, 2024.
- [250] Y. Lu, Y. Yang, Z. Xing, Q. Wang, and L. Zhu, “Diff-vps: Video polyp segmentation via a multi-task diffusion network with adversarial temporal reasoning,” in *MICCAI*, 2024.
- [251] L. Wan, Z. Chen, Y. Xiao, J. Zhao, W. Feng, and H. Fu, “Iterative feedback-based models for image and video polyp segmentation,” *CIBM*, vol. 177, p. 108569, 2024.
- [252] Y.-C. Chou, B. Li, D.-P. Fan, A. Yuille, and Z. Zhou, “Acquiring weak annotations for tumor localization in temporal and volumetric data,” *MIR*, vol. 21, no. 2, pp. 318–330, 2024.
- [253] Z. Xu, J. Rittscher, and S. Ali, “Sstfb: Leveraging self-supervised pretext learning and temporal self-attention with feature branching for real-time video polyp segmentation,” *arXiv preprint arXiv:2406.10200*, 2024.
- [254] Y. Yang, Z. Xing, and L. Zhu, “Vivim: a video vision mamba for medical video object segmentation,” *arXiv preprint arXiv:2401.14168*, 2024.
- [255] G. Chen, J. Yang, X. Pu, G.-P. Ji, H. Xiong, Y. Pan, H. Cui, and Y. Xia, “Mast: Video polyp segmentation with a mixture-attention siamese transformer,” *arXiv preprint arXiv:2401.12439*, 2024.
- [256] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *IEEE CVPR*, 2016.
- [257] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *ECCV*, 2020.
- [258] S. Chen, P. Sun, Y. Song, and P. Luo, “Diffusiondet: Diffusion model for object detection,” in *IEEE ICCV*, 2023.
- [259] M.-M. Cheng and D.-P. Fan, “Structure-measure: A new way to evaluate foreground maps,” *IJCV*, vol. 129, pp. 2622–2638, 2021.
- [260] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: Redesigning skip connections to exploit multiscale features in image segmentation,” *IEEE TMI*, vol. 39, no. 6, pp. 1856–1867, 2019.
- [261] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *MICCAI*, 2015.
- [262] Q. Chen, X. Chen, H. Song, Z. Xiong, A. Yuille, C. Wei, and Z. Zhou, “Towards generalizable tumor synthesis,” in *IEEE CVPR*, 2024.
- [263] K. Tian, Y. Jiang, Z. Yuan, B. PENG, and L. Wang, “Visual autoregressive modeling: Scalable image generation via next-scale prediction,” in *NeurIPS*, 2024.
- [264] M. Hu, P. Xia, L. Wang, S. Yan, F. Tang, Z. Xu, Y. Luo, K. Song, J. Leitner, X. Cheng *et al.*, “Ophnet: A large-scale video benchmark for ophthalmic surgical workflow understanding,” in *ECCV*, 2024.
- [265] N. K. Tomar, D. Jha, U. Bagci, and S. Ali, “Tganet: Text-guided attention for improved polyp segmentation,” in *MICCAI*, 2022.
- [266] Y. Zhao, J. Li, L. Ren, and Z. Chen, “Dtan: Diffusion-based text attention network for medical image segmentation,” *CIBM*, vol. 168, p. 107728, 2024.
- [267] Y. Zhao, J. Li, and Z. Hua, “Tact: Text attention based cnn-transformer network for polyp segmentation,” *IJIST*, vol. 34, 2023.
- [268] Z. Qin, H. Yi, Q. Lao, and K. Li, “Medical image understanding with pretrained vision language models: A comprehensive study,” in *ICLR*, 2023.
- [269] M. Guo, H. Yi, Z. Qin, H. Wang, A. Men, and Q. Lao, “Multiple prompt fusion for zero-shot lesion detection using vision-language models,” in *MICCAI*, 2023.
- [270] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, “Grounded language-image pre-training,” in *IEEE CVPR*, 2022.
- [271] S. Wang, Y. Zhu, X. Luo, Z. Yang, Y. Zhang, P. Fu, M. Wang, Z. Song, Q. Li, P. Zhou *et al.*, “Knowledge extraction and distillation from large-scale image-text colonoscopy records leveraging large language and vision models,” *arXiv preprint arXiv:2310.11173*, 2023.
- [272] R. Biswas, “Polyp-sam++: Can a text guided sam perform better for polyp segmentation?” *arXiv preprint arXiv:2308.06623*, 2023.
- [273] Y. Zhao, Y. Zhou, Y. Zhang, Y. Wu, and T. Zhou, “TextPolyp: Point-supervised Polyp Segmentation with Text Cues ,” in *MICCAI*, 2024.
- [274] S. Wang, W. Zhou, Y. Yang, H. Huang, Z. Ye, T. Zhang, and D. Yang, “Adapting pre-trained visual and language models for medical image question answering,” in *CLEF (Working notes)*, 2023.
- [275] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *ICML*, 2023.
- [276] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou, “A visual-language foundation model for pathology image analysis using medical twitter,” *NM*, vol. 29, no. 9, pp. 2307–2316, 2023.
- [277] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic, “Galactica: A large language model for science,” *arXiv preprint arXiv:2211.09085*, 2022.
- [278] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” in *NeurIPS*, 2024.
- [279] C. Li, Y. Ge, D. Li, and Y. Shan, “Vision-language instruction tuning: A review and analysis,” *TMLR*, 2024.
- [280] G. Lupyan, R. A. Rahman, L. Boroditsky, and A. Clark, “Effects of language on visual perception,” *TICS*, vol. 24, no. 11, pp. 930–944, 2020.
- [281] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, “Llava-med: Training a

- large language-and-vision assistant for biomedicine in one day,” in *NeurIPS*, 2024.
- [282] M. A. Islam, S. Jia, and N. D. Bruce, “How much position information do convolutional neural networks encode?” in *ICLR*, 2020.
- [283] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, “Lora: Low-rank adaptation of large language models,” in *ICLR*, 2022.
- [284] Y. Li, Y. Zhang, C. Wang, Z. Zhong, Y. Chen, R. Chu, S. Liu, and J. Jia, “Mini-gemini: Mining the potential of multi-modality vision language models,” *arXiv preprint arXiv:2403.18814*, 2024.
- [285] X. Chu, L. Qiao, X. Lin, S. Xu, Y. Yang, Y. Hu, F. Wei, X. Zhang, B. Zhang, X. Wei *et al.*, “Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices,” *arXiv preprint arXiv:2312.16886*, 2023.
- [286] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, “Convnext v2: Co-designing and scaling convnets with masked autoencoders,” in *IEEE CVPR*, 2023.
- [287] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *IEEE CVPR*, 2023.
- [288] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, “Dinov2: Learning robust visual features without supervision,” *TMLR*, 2024.
- [289] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *ICML*, 2021.
- [290] P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu, L. Heim, and M. Hobbhahn, “Position: Will we run out of data? limits of llm scaling based on human-generated data,” in *ICML*, 2024.
- [291] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and L. Yuan, “Florence-2: Advancing a unified representation for a variety of vision tasks,” in *IEEE CVPR*, 2024.
- [292] D. Jiang, X. He, H. Zeng, C. Wei, M. Ku, Q. Liu, and W. Chen, “Mantis: Interleaved multi-image instruction tuning,” *TMLR*, 2024.
- [293] J. B. Haurum, S. Escalera, G. W. Taylor, and T. B. Moeslund, “Which tokens to use? investigating token reduction in vision transformers,” in *IEEE ICCV-W*, 2023.
- [294] H. Diao, Y. Cui, X. Li, Y. Wang, H. Lu, and X. Wang, “Unveiling encoder-free vision-language models,” in *NeurIPS*, 2024.
- [295] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, “Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks,” in *IEEE CVPR*, 2024.
- [296] G. Bachmann and V. Nagarajan, “The pitfalls of next-token prediction,” in *ICML*, 2024.
- [297] X. Huang, J. Wang, Y. Tang, Z. Zhang, H. Hu, J. Lu, L. Wang, and Z. Liu, “Segment and caption anything,” in *IEEE CVPR*, 2024.