基于 SUN-SEG 数据集的息肉分割研究

摘要

结直肠癌是全球第二大致命癌症，且在其晚期阶段患者的存活率急剧下降。因此，结肠镜检查作为一种有效的筛查工具，能够及早发现病变息肉并进行干预，从而大幅提升存活率。近年来，深度学习特别是人工智能技术在结肠镜息肉检测中取得了显著进展，但仍面临注释数据不足与视频数据的动态复杂性挑战。为此，本文介绍了 SUN-SEG 数据集和用于息肉分割的最新视频息肉分割方法 SALI，该数据集是一种高质量的结肠镜视频息肉分割数据集，旨在推动视频息肉分割任务（VPS）的发展，而该方法基于短期对齐模块和长期交互模块的视频息肉分割方法，其实现了优异的效果。

结肠镜视频为直肠癌的息肉分割提供了更丰富的信息。然而，由于内窥镜的快速移动和近距离观察，使得现有的方法存在较大的空间不一致性和连续的低质量帧，从而产生有限的分割精度。在此背景下，通过增强相邻特征的一致性和重建可靠的息肉表示来实现稳健的视频息肉分割。为了实现这一目标，本文提出了 SALI 网络，它是短期对齐模块 (SAM) 和长期交互模块 (LIM) 的混合体。SAM 通过可变形卷积学习相邻帧的空间对齐特征，并进一步协调它们以捕获更稳定的短期息肉表示。在低质量帧的情况下，LIM 将历史的息肉表示存储为长期记忆库，并探索追溯关系以交互地为当前分割重建更可靠的息肉特征。结合 SAM 和 LIM，SALI 视频分割网络对空间变化和低视觉线索表现出很强的稳健性。

总之，本研究通过使用 SALI 视频分割网络在 SUN-SEG 数据集上进行息肉分割，旨在推动结肠镜视频息肉分割技术的发展，为临床提供更高效、准确的辅助诊断工具，以期降低结直肠癌的漏诊率，提升早期发现的效果，最终提高患者的存活率。

关键词：结直肠癌；息肉；SUN-SEG 数据集；SALI

1 引言

作为第二大致命癌症和第三大常见恶性肿瘤，结直肠癌估计每年导致数百万发病和死亡。结直肠癌患者在疾病第一期的存活率超过 95%，但到第四期和第五期时存活率急剧下降至 35% 以下 [3]。因此，通过结肠镜检查和乙状结肠镜检查等筛查技术及早诊断阳性结直肠癌病例对于提高存活率至关重要。为了预防，医生可以切除有转变为癌症风险的结肠息肉。然而，这一过程高度依赖于医生的经验，并且息肉漏诊率很高，高达 22% - 28% [28]。因此，结肠镜检查视频中的息肉自动分割对于提供及时提示和更丰富的诊断信息至关重要。

近些年来，人工智能技术被应用于医生在结肠镜检查期间自动检测候选病变息肉。然而，由于两个问题，开发具有令人满意的检测率的人工智能模型仍然具有挑战性：(a) 注释数据有限。深度学习模型通常需要具有密集注释标签的大规模视频数据集。此外，缺乏社区认可的基准来评估这些方法的实际性能。(b) 动态复杂性。结肠镜检查通常涉及不太理想的摄像机移

动采集条件，例如结肠息肉的多样性（例如边界对比度、形状、方向、拍摄角度）、内部伪影（例如水流、残留物）和成像质量下降（例如颜色失真、镜面反射）。为此，有作者提出了一个注释数据丰富且精准的息肉分割数据集 SUN-SEG [16]，以促进视频息肉分割的深度学习模型的开发，关于该数据集的具体介绍在4.3。

在过去的几年里，许多静态图像的息肉分割方法 [9, 10, 19, 45] 已经被提出，但是它们忽略了时间维度上的有价值的信息，并且在视频上表现出有限的性能。充分利用时间信息对每一帧的准确特征表示进行建模是视频分割的关键，但由于结肠镜检查独特的成像模式，这是非常具有挑战性的。首先，与大多数自然场景中摄像机是固定的而对象是移动的不同，在结肠镜视频中，摄像机是移动的，而对象 (息肉) 和背景 (正常组织) 是固定的，这导致息肉和正常组织之间的光流模式不明显 (参见图1(A))，从而使基于光流的视频分割方法 [27, 39, 40] 无效。其次，伴随着快速摄像机轨迹的近距离观察会在非常短的时间内甚至在两个相邻的帧之间引起显著的帧变化 (参见图1(B))。因此，现有的视频息肉分割方法 [15, 16, 21, 28] 依靠全局注意块直接聚集特征，会受到短期特征不稳定的影响。第三，复杂的光照环境导致大量的低质量剪辑 (参见图1(C))，这需要长范围的时间建模来捕捉可靠的帧，但现有方法仅考虑较窄的时间跨度 [11，2，18，12]，并且可能无法在充满低质量帧的时刻分割。

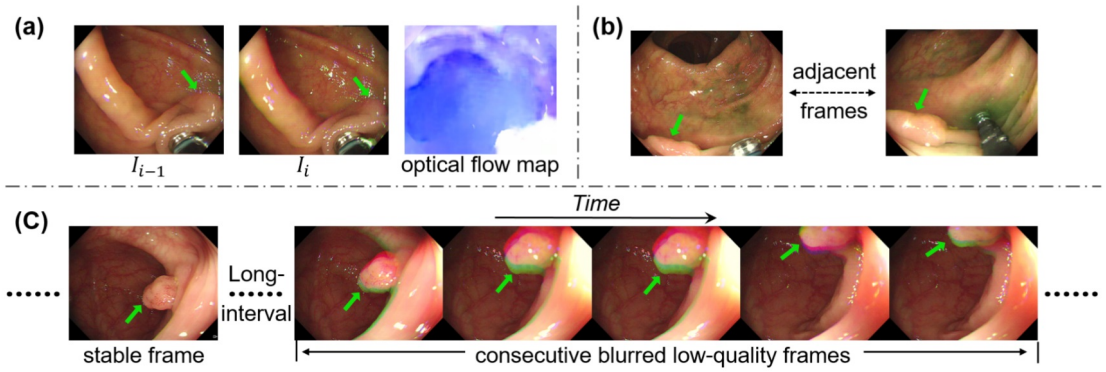

图 1. 息肉视频分割的三个挑战。(a) 光流动图（由 RAFT [32] 预测）无法显示任何物体运动信息。(b) 两个相邻帧之间存在显着差异。(c) 连续低质量帧的长序列。绿色箭头指向息肉。

为了解决上述挑战， [12] 提出了短期对齐和长期交互网络（SALI），它利用长期和短期感受野来建模时间连贯性。在短时间跨度内，我们设计了短期对齐模块（SAM），它首先通过可变形卷积对齐相邻帧的特征以减轻空间变化，然后探索对齐特征的相关性以聚合构建稳定的短期特征。在较长的时间跨度内， [12] 提出了长期交互模块（LIM），其中创建了一个记忆库来记住历史帧和预测。 [12] 通过一种新颖的掩蔽注意力模块在记忆库和短期查询之间建立时空相关性，以便即使面对质量引起的弱视觉线索也能以交互方式重建更可靠的息肉表征。

总而言之， [12] 的主要贡献如下：

- 提出了一种新型的视频息肉分割方法 SALI，该方法解决了现有方法对变异大的相邻帧和连续低质量帧的长序列的局限性。
- 提出了两个有效的新模块，称为短期对齐模块（SAM）和长期交互模块（LIM），分别增强时空特征的稳定性和可靠性。

- 大规模公共数据集的基准结果，即 SUN-SEG 证明，SALI 通过在四个测试子集上分别将 Dice 提高 2.1%、2.5%、4.1%、1.9%，实现了与其他最先进技术相比的卓越性能。

2 相关工作

早期的解决方案 [3, 8, 22, 23] 致力于通过人工挖掘图像特征的模式（例如颜色、形状、纹理和超像素）来识别结肠息肉。然而，由于这样设计的特征表示息肉的能力有限，以及息肉和硬模仿物之间的密切相似性，它们通常准确性较低 [38]。相比之下，数据驱动的人工智能技术可以以更好的学习能力应对这些具有挑战性的条件。本节主要重点关注跟踪最新的图像/视频息肉分割技术。

2.1 图像息肉分割

已经提出了几种从结肠镜检查图像中定位像素级息肉区域的方法，它们可以分为两大类。(a) 基于卷积神经网络 (CNN) 的方法。Brandao 等人 [4] 采用带有预训练模型的完全卷积网络 (FCN) 来分割息肉。后来，Akbari 等人 [2] 引入了一种改进的 FCN 来提高分割精度。受到 UNet [29] 在生物医学图像分割中取得巨大成功的启发，UNet++ [46] 和 ResUNet++ [14] 被用于息肉分割以提高性能。此外，PolypSeg [44]、ACS [41]、ColonSegNet [13] 和 SCR-Net [36] 探索了 UNetenhanced 架构在自适应学习语义上下文方面的有效性。之后新提出的方法中，SANet [19] 和 MSNet [43] 分别设计了浅层注意模块和减法单元，以实现精确高效的分割。此外，一些研究选择通过三种主流方式引入额外的约束: 施加明确的边界监督 [11, 17, 25, 31, 34]、引入隐式边界感知表示 [6, 10, 26]、探索模糊区域的不确定性 [18]。(b) 基于 Transformer 的方法。最近，Transformer [30] 因其强大的建模能力而越来越受欢迎。TransFuse [42] 将 Transformer 和 CNN 结合起来，称为并行分支方案，用于捕获全局依赖关系和低级空间细节。此外，还设计了一个 BiFusion 模块来融合来自两个分支的多层特征。Segtran [20] 提出了一种压缩注意力模块来规范自注意力，扩展模块学习多样化的表示。提出了一种位置编码方案来施加归纳连续性偏差。基于 PVT [35]，Dong 等人 [9] 引入了一个具有三个紧密组件的模型，即级联融合、伪装识别和相似性聚合模块。

2.2 视频息肉分割

尽管取得了进步，但现有的图像息肉分割方法存在固有的局限性，即忽视结肠镜检查视频中有价值的时间线索。因此，人们一直致力于将连续视频帧之间的时空特征结合起来。提出了混合 2/3D CNN 框架 [28] 来聚合时空相关性并实现更好的分割结果。然而，核大小限制了帧之间的空间相关性，限制了快速移动的息肉的准确分割。为了缓解上述问题，PNSNet [15] 引入了一个规范化的自我注意力 (NS) 模块来有效地学习具有邻居相关性的时空表示。在 [16] 中，作者深入研究了一种基于 NS 块的更有效的全球局部学习策略，该策略可以充分利用长期和短期时空依赖性。

3 本文方法

3.1 本文方法概述

图2显示了 SALI 的整体及其两个主要模块，即短期对齐模块（SAM）和长期互动模块（LIM）。在下面，将会详细介绍了每个模块并给出了实现细节。

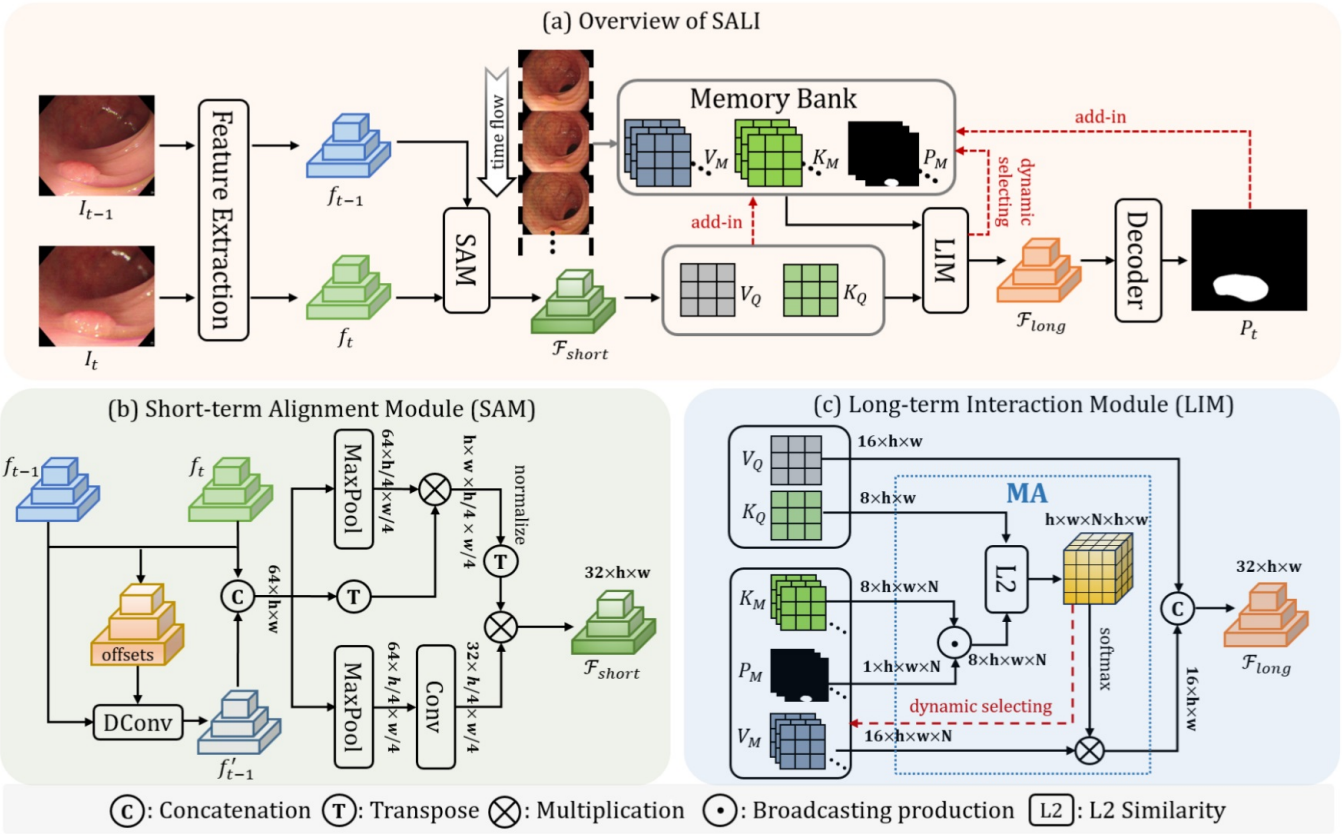

图 2. SALI 总体框架。(a)SALI 提出了短期对齐模块（SAM）和长期交互模块（LIM）两个模块来获取稳定可靠的时空特征。(b)SAM 首先对齐相邻特征，然后通过探索相关性构建稳定的短期特征。(c)LIM 利用掩蔽注意（MA）块将短期特征与记忆库中的长期视觉线索进行交互，以获得可靠的长期特征。

3.2 通过短期对齐模块进行稳定的特征学习

SALI 是一个在线视频分割网络，这意味着它“看不到”未来的帧。给定 t 时间点的当前帧 $I_t \in \mathbb{R}^{3\times H\times W}$，SALI 首先使用 SAM 来构建短期表示 $\mathcal{F}_{short}$，以前一帧 I_{t-1} 作为条件。具体来说，SAM 采用基于 transformer 的特征提取器 PVTv2 [35] 加上感受场块（RFB）[37] 来提取两个相邻帧的图像特征。对于每个帧，分别以缩减因子 8、16、32 获得三个特征地图，总共可以得到 $\{f_t^l, f_{t-1}^l\} \in \mathbb{R}^{32\times H/2^{l+2}\times W/2^{l+2}}, l = 1, 2, 3$。特征图在每个尺度 l 上独立处理，并且在最终解码之前不涉及跨尺度操作。因此，在下面省略了比例符号。

为了构建短期特征，我们需要聚合 $\{f_t, f_{t-1}\} \in \mathbb{R}^{32\times h\times w}$，如果特征地图处于 l 级，则 $h = \frac{H}{2^{l+2}}$ 和 $w = \frac{W}{2^{l+2}}$。然而，由于相机快速移动和近距离观察，它们之间会发生很大的空间变化，从而降低了聚集稳定性。有鉴于此，在聚合之前，SAM 首先将 f_{t-1} 与 f_t 对齐。简单的

方法是对帧执行配准，而变形场的估计非常困难，因为两个帧之间的视野通常不重叠。因此，SAM 直接倾斜对齐的特征，而不是显式估计变形场。

为此，SAM 采用可变形卷积层（DConv）[7]。具体来说，首先将 $\{f_t, f_{t-1}\}$ 级联并送入 1 x 1 卷积层 N_0 以估计核偏差，这反映了两个帧各自字段的差异。然后利用这些偏差创建 3 x 3 DConv，用于将 f_{t-1} 转换为 f'_{t-1}，其公式如下：

$$f'_{t-1} = \text{Dconv}\left(f_{t-1}, \mathcal{N}_0(\text{concat}[f_t, f_{t-1}])\right) \in \mathbb{R}^{32\times h\times w} \tag{1}$$

f'_{t-1} 预计将在端到端优化期间逐渐与 f_t 对齐。因此，我们可以通过级联来聚合它们，并使用自我注意力操作加上 3 x 3 卷积层 N_1 进一步协调它们，如图2（b）所示。最后，短期特征 $\mathcal{F}_{short}$ 可以计算如下：

$$\mathcal{F}_{short} = \text{Attention}(Q, K, V) \in \mathbb{R}^{32\times h\times w} \tag{2}$$

其中 $Q = \text{concat}[f_t, f'_{t-1}], K = \max\text{pool}(Q), V = \mathcal{N}_1(\max\text{pool}(Q))$，max-pooling 中的窗口大小设置为 4 x 4。

3.3 通过长期交互模块进行可靠的功能学习

从技术上讲，$\mathcal{F}_{short}$ 可以用于解码分割掩模，而由于复杂的照明环境，$\{I_t, I_{t-1}\}$ 可能包含低图像质量，从而使得解码不可靠。为了重建更可靠的代表，SALI 利用 LIM 来获得感知息肉代表长期记忆的能力。具体来说，LIM 首先利用 3 x 3 卷积层将短期特征 $\mathcal{F}_{short}$ 转换为一个键-值对，记为 $\{K \in \mathbb{R}^{8\times h\times w}, V \in \mathbb{R}^{8\times h\times w}\}$。随着视频流的进行，键-值对和最终的掩膜预测被不断添加到 N 长度的存储库中，产生历史键-值对的集合，即 $\{K_M, V_M\}$ 和掩膜集，即 $P_M \in \mathbb{R}^{1\times h\times w\times N}$。更清楚的是，如图2（c）所示，从当前短期特征推导出的键-值对被表示为 $\{K_Q, V_Q\}$。

我们的目标是通过使用存储在存储库中的之前看到的区分性息肉表示将 $\{K_Q, V_Q\}$ 重建为更可靠的长期特征 $\mathcal{F}_{long}$。为此，LIM 引入了一个掩膜注意力（MA）块，该块利用当前帧 K_Q 通过使用存储的知识库中的息肉区域来检索最相关的历史帧 K_M。然后，MA 块根据标准化相关性聚合检索到的帧的特征值。我们将这个过程称为长期交互，因为关键的想法是交互式地收集在大时间跨度内出现的那些可靠的视觉线索，以重建当前的可靠表示。

最后，将当前特征值与交互的特征值拼接得到 $\mathcal{F}_{long}$，公式如下：

$$\mathcal{F}_{long} = \text{concat}\left[V_Q, V_M \times \text{softmax}\left(\frac{\mathcal{S}(K_Q, K_M \odot P_M)}{\sqrt{d_k}}\right)\right] \in \mathbb{R}^{32\times h\times w} \tag{3}$$

其中 dk 是比例因子 [33]，S (·) 计算 L2 相似度 [5]，并且 ⊙ 表示广播机制操作。

为了平衡内存成本，将最大内存长度 N 设置为 35，并每 5 帧将键-值对添加到内存库中。如果存储库已满，则会调用动态选择策略，为新进入的元素腾出空间。具体来说，3.3等式中的 L2 相似性所指示的最不相关的历史帧将从键-值对集合被删除。

3.4 损失函数定义

对于分割，SALI 首先从所有三个尺度中获得 $\mathcal{F}_{long}$，然后利用部分解码器 [37] 聚合 $\mathcal{F}_{long}$ 以获得全局信息，最后使用全局信息通过反向注意力机制 [10] 预测当前时间的最终分割掩膜

P_t。然后，我们计算分割的地图 P_T 和地面真值掩蔽 Y_t 之间的交叉熵 (CE) 损失和交并比 (IoU) 损失的组合，以优化 SALI 模型，如下所示：

$$\mathcal{L} = \mathcal{L}_{CE}(P_t, Y_t) + \mathcal{L}_{IoU}(P_t, Y_t) \tag{4}$$

4 复现细节

4.1 与已有开源代码对比

利用 [12] 提供的的开源代码，进行复现，开源代码来自https://github.com/Scatteredrain/SALI。

4.2 实验环境搭建

遵循https://github.com/Scatteredrain/SALI/blob/main/README.md内的 Usage 下的操作步骤搭建即可。

4.3 数据集与评估指标

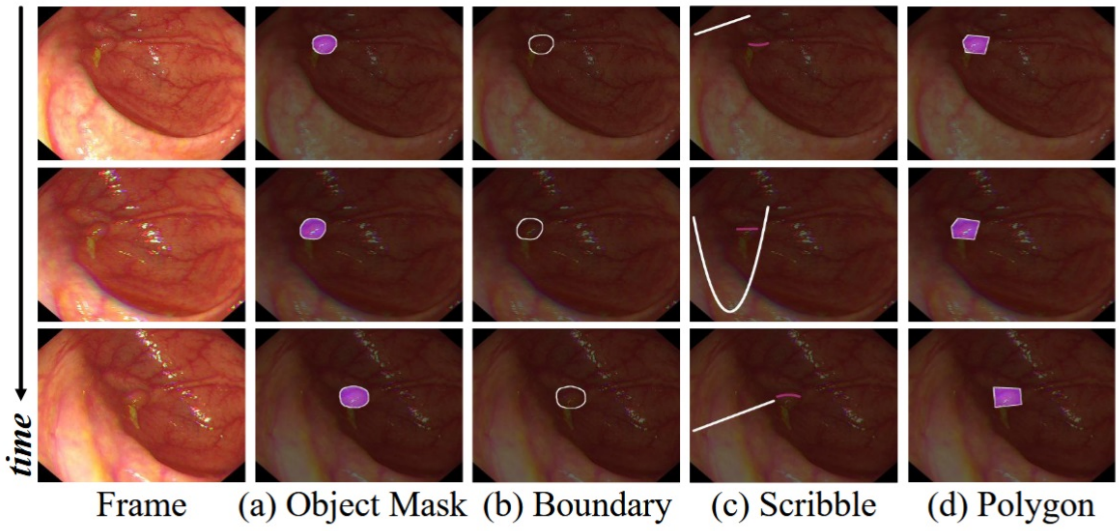

图 3. SUN-SEG 数据集中每个视频帧的多样化注释，包括对象掩膜 (a)、边界 (b) 和两个弱标签，即涂鸦 (c) 和多边形 (d)。

本文使用的 SUN-SEG 数据集是一个大规模的视频息肉分割数据集，由研究者从知名的 SUN 数据库 [24] 中扩展得来。该数据集包含 158690 帧来自结肠镜检查视频的帧，如图3所示，可以看到该数据集覆盖了多种类型的标注，包括属性、对象掩码、边界、涂鸦和多边形。这些丰富的标注不仅支持息肉的检测和定位，还有助于衍生任务的发展。SUN-SEG 数据集的构建考虑了实际结肠镜检查中的挑战性场景，如不同尺寸的息肉、不同的聚焦距离和速度、以及成像过程中的各种动态条件。数据集从 SUN 数据库的 113 个结肠镜视频中手动修剪出 378 个阳性和 728 个阴性视频片段，保持了它们的连续内在关系，每个片段大约持续 3 至 11 秒，

以 30fps 的实时帧率提供。此外，SUN-SEG 数据集还经过了严格的质量控制，按照 [10] 的工作流程，由经验丰富的标注者使用 Adobe Photoshop 提供各种标签，然后由结肠镜相关的研究人员重新验证这些初始标注的质量和正确性。SUN-SEG 数据集的引入，旨在推动结肠镜诊断、定位和衍生任务的发展，为 VPS 任务提供了一个高质量的密集标注基准。

为了更深入地了解模型性能，[12] 使用六个不同的指标来进行时间戳 s 处的预测 P_s 和地面真值 G_s 之间的模型评估，在这里简要介绍比较关键的前三种，包括：(a)DIce 系数 (Dice = $\frac{2\times|P_s\cap G_s|}{|P_s\cup G_s|}$)，测量预测掩膜和真实掩膜之间的相似性，并对假阳性/阴性预测进行惩罚。符号 $\cap$、$\cup$ 和 $|\cdot|$ 分别表示区域中的交集、并集和对应的像素数数量。(b) 像素敏感度 (Sen = $\frac{|P_s\cap G_s|}{|G_s|}$)，用于评估总体病变区域的真实阳性预测。由于结肠镜检查的目标是筛查息肉缺失率较低的息肉，因此患有息肉的人应该很有可能被识别出来。因此，可以通过采用灵敏度来惩罚假阴性预测，灵敏度是指该方法正确检测息肉的能力。(c)F-measure [1]($F_\beta = \frac{(1+\beta^2)\times\text{Prc}\times\text{Rcl}}{\beta^2\times(\text{Prc}+\text{Rcl})}$) 是由 β 加权的精确度和召回率的调和平均值，广泛用于通过结合精确度 (Prc $= \frac{|P_s\cap G_s|}{|P_s|}$) 和召回率 (Rcl $= \frac{|P_s\cap G_s|}{|G_s|}$) 进行更全面的评估。

5 实验结果分析

表 1. 复现结果对比

Method	SUN-SEG-Easy						SUN-SEG-Hard					
	S_α	E_ϕ^{mn}	F_β^w	F_β^{mn}	Sen	Dice	S_α	E_ϕ^{mn}	F_β^w	F_β^{mn}	Sen	Dice
Seen												
Claim	94.4	97.2	91.0	92.5	92.4	92.7	91.6	95.2	86.5	88.9	89.8	89.1
Mine	**93.6**	**96.5**	**89.6**	**91.5**	**90.6**	**91.5**	**90.2**	**93.9**	**84.4**	**86.8**	**86.9**	**86.8**
Unseen												
Claim	87.0	92.0	79.4	83.1	81.1	82.5	87.4	92.0	79.0	82.2	83.0	82.2
Mine	**85.8**	**90.1**	**77.8**	**81.7**	**78.1**	**81.2**	**87.2**	**90.9**	**79.2**	**82.8**	**81.0**	**82.6**

复现结果如表1所示，这里主要看 Dice 指标，可以看到我们的 Dice 指标在 Hard-Unseen 是比 [12] 的结果还要好的，但是在另外三个验证子集上是略低的，考虑到深度学习模型训练的随机性，这个大约 1% 是差距是可以接受的，因此可以得出结论，本次复现工作顺利完成。

6 总结与展望

在 [12] 的工作中，提出了 SALI，一种用于息肉视频分割的新型高效框架。SALI 提出了两个新模块，即 SAM 和 LIM，分别用于解决存在较大变化的相邻帧之间特征的稳定短期聚合和大量低质量帧上的可靠长期时间交互。在 SUN-SEG 数据集上与六种最先进 (SOTA) 视频分割方法进行全面而广泛的比较表明，SALI 的表现明显优于其他 SOTA 的方法。

但是作者并未对这个工作进行 fps 的测试，而视频分割框架中，fps 是一个十分关键的指标，这关系到该方法能否付诸实际，在日常生活中，电影电视都是 24fps 的，而经过粗糙的测

试，在 NVIDIA 3090 上的测试结果是只有 10fps，无法满足最低要求的 24fps，同时该文的记忆帧的设计并不是十分清晰易懂，未来可以从轻量化和明晰记忆帧的设计这两个方面对该工作进行优化。比如将 backbone 替换成参数量与计算量较小的或将注意力操作替换成状态空间模型 (SSM) 来进行轻量化的思路，将记忆库的更新策略进行优化和容量，现在每 5 帧添加并通过 L2 相似度作为选择策略，尝试将容量下降提高 fps 和设计更有效的记忆帧策略。

参考文献

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2009.

[2] Mojtaba Akbari, Majid Mohrekesh, Ebrahim Nasr-Esfahani, S.M. Reza Soroushmehr, Nader Karimi, Shadrokh Samavi, and Kayvan Najarian. Polyp segmentation in colonoscopy images using fully convolutional network. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, page 69–72. IEEE, July 2018.

[3] J. Bernal, J. Sánchez, and F. Vilariño. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, September 2012.

[4] Patrick Brandao, Evangelos Mazomenos, Gastone Ciuti, Renato Caliò, Federico Bianchi, Arianna Menciassi, Paolo Dario, Anastasios Koulaouzidis, Alberto Arezzo, and Danail Stoyanov. Fully convolutional neural networks for polyp segmentation in colonoscopy. In Samuel G. Armato and Nicholas A. Petrick, editors, *Medical Imaging 2017: Computer-Aided Diagnosis*. SPIE, March 2017.

[5] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation, 2021.

[6] Mengjun Cheng, Zishang Kong, Guoli Song, Yonghong Tian, Yongsheng Liang, and Jie Chen. *Learnable Oriented-Derivative Network for Polyp Segmentation*, page 720–730. Springer International Publishing, 2021.

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017.

[8] B.V. Dhandra, R. Hegadi, M. Hangarge, and V.S. Malemath. Analysis of abnormality in endoscopic images using combined hsi color space and watershed segmentation. In *18th International Conference on Pattern Recognition (ICPR' 06)*, page 695–698. IEEE, 2006.

[9] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *CAAI Artificial Intelligence Research*, page 9150015, December 2023.

[10] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020.

[11] Yuqi Fang, Cheng Chen, Yixuan Yuan, and Kai-yu Tong. *Selective Feature Aggregation Network with Area-Boundary Constraints for Polyp Segmentation*, page 302–310. Springer International Publishing, 2019.

[12] Qiang Hu, Zhenyu Yi, Ying Zhou, Fang Peng, Mei Liu, Qiang Li, and Zhiwei Wang. Sali: Short-term alignment and long-term interaction network for colonoscopy video polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 531–541. Springer, 2024.

[13] Debesh Jha, Sharib Ali, Nikhil Kumar Tomar, Havard D. Johansen, Dag Johansen, Jens Rittscher, Michael A. Riegler, and Pal Halvorsen. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. *IEEE Access*, 9:40496–40510, 2021.

[14] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Dag Johansen, Thomas De Lange, Pal Halvorsen, and Havard D. Johansen. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE, December 2019.

[15] Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Huazhu Fu, Debesh Jha, and Ling Shao. Progressively normalized self-attention network for video polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 142–152. Springer, 2021.

[16] Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research*, 19(6):531–549, 2022.

[17] Ge-Peng Ji, Lei Zhu, Mingchen Zhuge, and Keren Fu. Fast camouflaged object detection via edge-based reversible re-calibration network. *Pattern Recognition*, 123:108414, March 2022.

[18] Taehun Kim, Hyemin Lee, and Daijin Kim. Uacanet: Uncertainty augmented context attention for polyp segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM ' 21, page 2167–2175. ACM, October 2021.

[19] Go-Eun Lee, Jungchan Cho, and Sang-Il Choi. Sranet: Shallow and reverse attention network for colon polyp segmentation. February 2023.

[20] Shaohua Li, Xiuchao Sui, Xiangde Luo, Xinxing Xu, Yong Liu, and Rick Goh. Medical image segmentation using squeeze-and-expansion transformers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, IJCAI-2021, page 807–815. International Joint Conferences on Artificial Intelligence Organization, August 2021.

[21] Jianzhuang Lin, Wenzhong Yang, and Sixiang Tan. Combining transformer and reverse attention mechanism for polyp segmentation. In *Proceedings of the 4th International Conference on Biotechnology and Biomedicine*, page 125–137. SCITEPRESS - Science and Technology Publications, 2022.

[22] Omid Haji Maghsoudi. Superpixel based segmentation and classification of polyps in wireless capsule endoscopy. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, page 1–4. IEEE, December 2017.

[23] Alexander V. Mamonov, Isabel N. Figueiredo, Pedro N. Figueiredo, and Yen-Hsi Richard Tsai. Automated polyp detection in colon capsule endoscopy. *IEEE Transactions on Medical Imaging*, 33(7):1488–1502, July 2014.

[24] Masashi Misawa, Shin-ei Kudo, Yuichi Mori, Kinichi Hotta, Kazuo Ohtsuka, Takahisa Matsuda, Shoichi Saito, Toyoki Kudo, Toshiyuki Baba, Fumio Ishida, Hayato Itoh, Masahiro Oda, and Kensaku Mori. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointestinal Endoscopy*, 93(4):960–967.e3, April 2021.

[25] Balamurali Murugesan, Kaushik Sarveswaran, Sharath M Shankaranarayana, Keerthi Ram, Jayaraj Joseph, and Mohanasankar Sivaprakasam. Psi-net: Shape and boundary aware joint multi-task deep network for medical image segmentation. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, page 7223–7226. IEEE, July 2019.

[26] Tan-Cong Nguyen, Tien-Phat Nguyen, Gia-Han Diep, Anh-Huy Tran-Dinh, Tam V. Nguyen, and Minh-Triet Tran. *CCBANet: Cascading Context and Balancing Attention for Polyp Segmentation*, page 633–643. Springer International Publishing, 2021.

[27] Gensheng Pei, Fumin Shen, Yazhou Yao, Guo-Sen Xie, Zhenmin Tang, and Jinhui Tang. *Hierarchical Feature Alignment Network for Unsupervised Video Object Segmentation*, page 596–613. Springer Nature Switzerland, 2022.

[28] Juana González-Bueno Puyal, Kanwal K. Bhatia, Patrick Brandao, Omer F. Ahmad, Daniel Toth, Rawen Kader, Laurence Lovat, Peter Mountney, and Danail Stoyanov. *Endoscopic Polyp Segmentation Using a Hybrid 2D/3D CNN*, page 295–305. Springer International Publishing, 2020.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, page 234–241. Springer International Publishing, 2015.

[30] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *Medical Image Analysis*, 88:102802, August 2023.

[31] Yutian Shen, Xiao Jia, and Max Q.-H. Meng. *HRENet: A Hard Region Enhancement Network for Polyp Segmentation*, page 559–568. Springer International Publishing, 2021.

[32] Zachary Teed and Jia Deng. *RAFT: Recurrent All-Pairs Field Transforms for Optical Flow*, page 402–419. Springer International Publishing, 2020.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[34] Ruxin Wang, Shuyuan Chen, Chaojie Ji, Jianping Fan, and Ye Li. Boundary-aware context neural network for medical image segmentation. *Medical Image Analysis*, 78:102395, May 2022.

[35] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, March 2022.

[36] Huisi Wu, Jiafu Zhong, Wei Wang, Zhenkun Wen, and Jing Qin. Precise yet efficient semantic calibration and refinement in convnets for real-time polyp segmentation from colonoscopy videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):2916–2924, May 2021.

[37] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3902–3911. IEEE, June 2019.

[38] Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng Ann Heng. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. *IEEE Journal of Biomedical and Health Informatics*, 21(1):65–75, January 2017.

[39] Yichen Yuan, Yifan Wang, Lijun Wang, Xiaoqi Zhao, Huchuan Lu, Yu Wang, Weibo Su, and Lei Zhang. Isomer: Isomerous transformer for zero-shot video object segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 966–976. IEEE, October 2023.

[40] Kaihua Zhang, Zicheng Zhao, Dong Liu, Qingshan Liu, and Bo Liu. Deep transport network for unsupervised video object segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2021.

[41] Ruifei Zhang, Guanbin Li, Zhen Li, Shuguang Cui, Dahong Qian, and Yizhou Yu. *Adaptive Context Selection for Polyp Segmentation*, page 253–262. Springer International Publishing, 2020.

[42] Yundong Zhang, Huiye Liu, and Qiang Hu. *TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation*, page 14–24. Springer International Publishing, 2021.

[43] Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. *Automatic Polyp Segmentation via Multi-scale Subtraction Network*, page 120–130. Springer International Publishing, 2021.

[44] Jiafu Zhong, Wei Wang, Huisi Wu, Zhenkun Wen, and Jing Qin. *PolypSeg: An Efficient Context-Aware Network for Polyp Segmentation from Colonoscopy Videos*, page 285–294. Springer International Publishing, 2020.

[45] Tao Zhou, Yi Zhou, Kelei He, Chen Gong, Jian Yang, Huazhu Fu, and Dinggang Shen. Cross-level feature aggregation network for polyp segmentation. *Pattern Recognition*, 140:109555, August 2023.

[46] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, June 2020.