深度学习视角下的视频息肉分割

季葛鹏†[1], 肖国宝†[2], 周昱程†[3], 范登平✉[4], 赵凯[5], 陈耿[6] and Luc Van Gool[4]

[1]工程研究院, 澳大利亚国立大学, 堪培拉, 澳大利亚.
[2]计算机与控制工程学院, 闽江学院, 福州, 中国.
[3]计算机学院, 约翰霍普金斯大学, 巴尔的摩, 美国.
[4]计算机视觉实验室, 苏黎世联邦理工学院, 苏黎世, 瑞士.
[5]放射科学系, 加利福尼亚大学洛杉矶分校, 洛杉矶, 美国.
[6]计算机学院, 西北工业大学, 西安, 中国.

项目主页: https://github.com/GewelsJI/VPS.

Abstract

本文在深度学习时代下，呈现了首个关于视频息肉分割（VPS）技术的系统性研究。近年来，由于缺乏具有大规模细粒度分割标签的数据集，视频息肉分割研究领域的发展并不顺利。为解决上述问题，本文首次引入一个具有逐帧标注的高质量视频息肉分割数据集，名为SUN-SEG，其中包含了来自著名SUN数据集中的**158, 690**张结肠镜视频帧。并额外提供了不同类型的标签，即：属性标签、目标掩码、边缘标签、线标签和多边形标签。其次，本文设计了一个简单且高效的基线模型，名为PNS+，其包含全局编码器、局部编码器和归一化自注意力（NS）模块。全局编码器和局部编码器分别接收一个锚帧和多个连续帧作为输入，用以提取长期表征和短期表征，然后通过两个归一化自注意力模块渐进式地更新。实验结果表明PNS+模型取得了最佳的性能和实时的推理速度（170fps），使其成为视频息肉分割任务中颇具前景的解决方案。本文接着在SUN-SEG数据集上广泛地评测了**13**个经典的息肉/目标分割模型，并且提供了基于属性的评测结果。最后，本文讨论了领域内亟待解决的几个问题，并为视频息肉分割研究社区提出了若干潜在研究方向。

Keywords: 视频息肉分割、数据集、自注意力、结肠镜检查、腹腔

1 引言

结直肠癌（CRC）是全球第二大致命癌症和第三大常见的恶性肿瘤，据估计每年会在全球范围内造成数百万人发病和死亡。结直肠癌患者在第一阶段的生存概率超过95%，但在第四和第五阶段却大幅下降到35%以下[2]。因此，通过结肠镜、乙状结肠镜等筛查技术对阳性结直肠癌病例进行早期预诊，对于提高患者生存率具有重要意义。为达到预防目的，内科医师可以切除有癌变风险的结肠息肉。然而，这一过程高度依赖于医师的经验水平，且出现了较高的息肉漏诊率（即：22% ∼ 28%[3]）。

近年来，人工智能（AI）技术被医生用于执行结肠镜检查过程中进行病变息肉自动检测。然而，开发出具有令人满意的检测率的人工智能方案仍具有挑战性，其主要存在以下两个问题：

†代表具有同等贡献。✉代表通讯作者(dengpfan@gmail.com)。本文为MIR2022 [1]的中文翻译版，由周昱程、季葛鹏翻译，范登平、陈耿、赵凯校稿。

（a）有限的标注数据： 深度学习模型通常需要具有密集标注的大规模视频数据集。此外，研究社区内也缺乏一个广泛认可的评测基准用于评估对比方法的真实能力（例如：准确率和效率)。

（b）动态复杂性： 结肠镜检查通常涉及到不太理想的相机运动和图像采集条件，包括息肉的多样性（例如：边缘对比度、形状、方向、角度）、肠道杂物（例如：水流、残留物）和成像退化（例如：颜色失真、镜面反射）。为此，本文呈现了一个系统性的研究工作，用以推动深度学习模型在视频息肉分割（VPS）领域的发展。主要贡献如下:

- **视频息肉分割数据集：** 本文提出一个名为SUN-SEG的大规模视频息肉分割数据集，其包含了从SUN[4]中选取的158,690个视频帧。本文还提供了各类标签，包括：属性标签、目标掩码、边缘标签、线标签和多边形标签，用于进一步推动结肠镜诊断、定位及其衍生任务的发展。
- **视频息肉分割基线模型：** 本文设计了一个简单且高效的视频息肉分割基线模型，名为PNS+，其由一个全局编码器、一个局部编码器和两个归一化自注意力（NS）模块组成。全局编码器和局部编码器分别用于从锚帧和多个连续帧中提取长期和短期表征。归一化自注意力模块则用于在所提取特征之间耦合注意力线索时，动态地更新感受野。实验表明PNS+在具有挑战性的SUN-SEG数据集上取得了最佳性能。
- **视频息肉分割评测基准：** 为了对视频息肉分割发展提供更为全面的理解，本文进行了首个大规模基准评测，其包含了对13个（即：5个基于图像和8个基于视频）前沿的息肉分割/目标分割方法进行评测。根据评测基准的结果，本文观察到视频息肉分割任务尚未很好的解决。这为未来进一步的探索留下了很大的空间。

本工作的会议版本发表于[5]。在本扩展工作中，主要引入了如下三处贡献:

- 在第3节中, 本文提出了一个具有密集标注的高质量视频息肉分割数据集（SUN-SEG），其具有五个扩展的标签，即：属性标签、目标掩码、边缘标签、线标签和多边形标签。
- 本文基于文献[5]中所提出的归一化自注意力模块，设计了一个简单且高效的基线模型PNS+，以实现对长期和短期依赖的建模（请参见第4节）。
- 如第5节所示，本文在视频息肉分割任务上贡献了第一个大规模的评测基准。其包含了13个最新的息肉/目标分割对比模型。

2 相关工作

本节从以下两个方面回顾了计算机辅助息肉诊断的最新进展：结肠镜相关数据集（请参见第2.1节）和结肠镜相关方法（请参见第2.2节）。

2.1 结肠镜相关数据集

近年来，已有多个与人类结肠镜检查相关的数据集被收集。如表1所示，总结了19个主流数据集和本文的SUN-SEG数据集中的关键统计数据。根据任务定义将其划分为三个主要方向。

2.1.1 分类

有四个主流的数据集最初被用于识别胃肠道病变。ColonoscopicDS[8]收集了76个结肠镜视频，其包含三种类型的肠道病变，包括：增生性病变、锯齿状病变和腺瘤病变。Kvasir[14]包含了8个解剖学标志（包括：息肉、食管炎、溃疡性结肠炎、Z线、幽门、盲肠、染色息肉和染色切除边缘），每一种类型对应着1,000张图像。Hyper-Kvasir[16]进一步从374个结肠镜视频

表1 人类结肠镜检查的20个当前数据集统计。**#IMG** 代表图像数量。**#VID**代表视频数量。**DL**代表密集标注。**CLS**代表分类标签。**BBX**代表包围盒标签。**PM**代表像素级别标签。

数据集	发表年份	#IMG	#VID	DL	CLS	BBX	PM	网址
CVC-ColonDB[2]	2012	300	13				✓	Link
ETIS-Larib[6]	2014	196	34				✓	Link
CVC-ClinicDB[7]	2015	612	31				✓	Link
ColonoscopicDS[8]	2016	-	76		✓			Link
ASU-Mayo[9]	2016	36,458	38	✓			✓	Link
CVC-ClinicVideoDB[10]	2017	11,954	18	✓		✓		Link
CVC-EndoSceneStill[11]	2017	912	44				✓	Link
KID2[12, 13]	2017	2,371	47		✓		✓	Link
Kvasir[14]	2017	8,000	-		✓			Link
EDD2020[15]	2020	386	-		✓	✓	✓	Link
SUN-database[4]	2020	158,690	113	✓	✓	✓		Link
Hyper-Kvasir[16]	2020	110,079	374		✓	✓	✓	Link
Kvasir-SEG[17]	2020	1,000	-				✓	Link
PICCOLO[18]	2020	3,433	40		✓		✓	Link
Kvasir-Capsule[19]	2021	4,741,504	117	✓	✓	✓		Link
CP-CHILD-A[20]	2021	8,000	-		✓			Link
CP-CHILD-B[20]	2021	1,500	-		✓			Link
LDPolypVideo[21]	2021	40,266	160	✓	✓	✓		Link
KUMC[22]	2021	37,899	155	✓	✓	✓		Link
PolypGen[23]	2021	6,282	26		✓	✓	✓	Link
SUN-SEG (OUR)	2022	158,690	1,013	✓	✓	✓	✓	Link

中收集了110,079个样本，其中包含了三种类型的标签：23种不同病变发现的10,662个类别标签，以及1,000张分割掩码和检测包围盒标签。值得注意的是，Hyper-Kvasir中所有的分割掩码都是从Kvasir-SEG[16]中挑选而来的。近期，CP-CHILD-A/-B[20]记录了来自儿童的结肠镜数据，其包含用于分类任务的两个类别（即：结肠息肉和正常或其他病理图像）。

2.1.2 检测

目前有5个广泛应用的视频数据集主要用于检测任务。CVC-ClinicVideoDB[10]作为早期的视频数据集，包含了18个视频。其总帧数为11,954，其中的10,025帧至少包含一个息肉。SUN[4]作为最大的密集标注的视频息肉检测数据集。其包含了从99名患者采集的49,136个含病灶数据和对应的包围盒标签。近期，两个视频数据集（即：Kvasir-Capsule[19]和KUMC[22]）被应用于检测和分类任务。具体而言，前者提供了14种病变类别的47,238个包围盒标签，而后者提供了带有包围盒标签的37,899个视频帧。与上述的数据集不同，LDPolypVideo[21]包含了来自160个结肠镜视频的40,266帧及其对应的圆形标签。

2.1.3 分割

对于视频数据集，早期的评测基准CVC-EndoSceneStill[11]将CVC-ColonDB[2]和CVC-ClinicDB[7]进行组合。ETIS-Larib[6]包含来自32个结肠镜视频的196个样本，其中每个视频大约5帧。EDD2020[15]包含来自5个不同机构和多个胃肠器官的386张结肠镜图像。其为疾病检测、定位和分割提供标签。PICCOLO[18]也从40个视频中收集了带有稀疏标注的3,433图像帧。可以看到，上述的视频数据集都采用了抽样标注的策略；而且碍于劳动密集型的标注流程，每个视频片段仍然缺少逐帧的掩码标注。作为首个具有密集掩码标注的视频数据集，ASU-Mayo[9]包含了来自38个视频中的36,458个连续帧，然而其仅提供了其中10个视频的3,856个掩码标注。近期，PolypGen[23]收集了一个包含300多名患者的

多中心数据集，并带有3,788个掩码和包围盒的单帧和连续帧标签。与现有的工作不同，本文引入了SUN-SEG。这是第一个用于视频息肉分割任务的高质量密集标注的数据集。其包含了丰富的手工标签：目标掩码、边缘标签、线标签、多边形标签和属性标签。希望这项工作可以推动结肠镜检查诊断、定位和衍生任务的发展。

2.2 结肠镜相关方法

早期的解决方案[2, 24–26]致力于通过挖掘手工特征（如颜色、形状、纹理和超像素）来识别结肠息肉。然而，由于手工特征对于非均匀息肉的表征能力有限，且息肉与难样本之间较为相似[27]。相比之下，数据驱动下的人工智能技术可以借力于较强的学习能力来处理这些困难情况。本节主要介绍最新的图像/视频息肉分割技术[28]，将息肉分类[29, 30]和检测[31, 32]的系统性回顾工作留到未来完成。

2.2.1 图像息肉分割（IPS）

若干方法从结肠镜检查图像中定位像素级的息肉区域，它们可被分为如下两大类。**（a）基于卷积神经网络的方法：** Brandao等人[33]采用了一个全卷积网络（FCN）与一个预训练模型来分割息肉。随后，Akbari等人[34]引入了一种改进型FCN网络来提高分割精度。受到UNet[35]在生物医学图像分割中巨大成功的启发，UNet++[36]和ResUNet[37]被应用于息肉分割以提高性能。此外，基于UNet的增强型结构的方法(如PolypSeg[38]、ACS[39]、ColonSegNet[40]和SCR-Net[41])可以自适应地学习上下文语义。作为新提出的方法，SANet[42]和MSNet[43]分别设计了浅层注意力模块和减法单元，实现了精确、高效的分割。此外，一些工作通过三种主流方式来引入辅助约束：运用显式边缘监督[44–48]、引入隐式边缘表征[49–51]和探索二义区域的不确定性[52]。**（b）基于Transformer的方法：** 近年来，Transformer[53]因其强大的建模能力而越来越受欢迎。TransFuse[54]结合了Transformer和卷积神经网络，称其为并行分支方案，用于捕获全局依赖关系和底层空间细节，并设计了BiFusion模块来融合两个分支的多层次特征。Segtran[55]提出了一种压缩注意力块来正则化自注意力，然后使用拓展模块来学习多样化表征，并提出一个位置编码方法来施加归纳连续性偏置。在PVT[56]的基础上，Dong等人[57]提出了一种基于级联融合、伪装识别和相似度聚合三个紧密组件的模型。

2.2.2 视频息肉分割（VPS）

尽管IPS领域取得了长足的进展，但其忽视了结肠镜检查视频中丰富的线索。因而存在本质上的局限性。为此，学者们致力于将连续视频帧之间的时空特征结合起来。Puyal等人[3]提出了一种混合的2/3D卷积神经网络框架来聚合时空相关性，并获得较好的分割结果。然而，卷积核的大小限制了帧间的空间相关性，从而限制了快速运动息肉的精确分割。为了缓解上述问题，本文的会议版本PNSNet[5]提出了一种归一化自注意力（NS）模块来有效地学习时空表征的邻域相关性。本文基于归一化自注意力模块，深入研究了一种更高效的全局至局部学习策略。该策略能够充分利用长期和短期的时空依赖关系。

3 视频息肉分割数据集

本节从数据收集、数据组织、专业标注和数据集统计等四个方面详细介绍了SUN-SEG数据集。

3.1 数据组织

SUN-SEG中的结肠镜视频来源于昭和大学与名古屋大学数据库（又称为SUN-database[4]）。

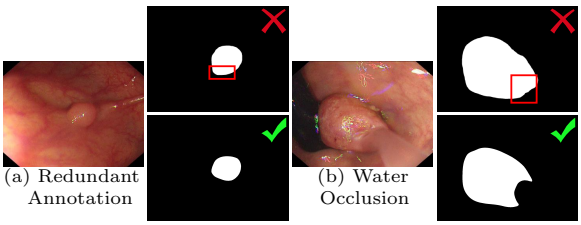


图1 高品控的数据标注流程。例如：拒绝了样例（a），因其边缘与息肉的不一致，并拒绝了样例（b），因其水柱重叠区域被错误地标注。

该数据集是用于检测任务的最大规模结肠镜视频息肉数据集。采用SUN-database作为本文的数据源有如下两个优点：**（a）挑战性场景：** 视频由高清内窥镜（CF-HQ290ZI和CF-H290ECI，Olympus公司）和摄像机（IMH-10，Olympus公司）拍摄，在动态场景下提供不同息肉尺寸大小的视频片段，如在不同对焦距离和速度下的成像。**（c）可靠的病理定位：** 最初的分类信息和包围盒标签由三位研究助理提供，并且由两位具有医学专业领域知识的内窥镜专家进行再次核验。

原始SUN-database具有113个结肠镜检查视频，包括100个阳性视频与49,136个息肉视频帧，13个阴性视频与109,554张非息肉帧[1]。本文将该数据集手工裁剪成378个阳性短片段和728个阴性短片段，同时保持视频帧之间内在的时序连续关系。这样的数据预处理可确保每个视频在实时帧率（即：30 fps）下有大约3~11秒的持续时长，从而提高了各种算法和设备的容错率。因此，重新组织后的SUN-SEG数据集总共包含了1,106个视频片段和158,690视频帧，为构建一个具有代表性的评测基准提供了坚实的基础。

[1]上述统计数据来自于Website，这与原始论文[4]中所报告数据有一定的差异。此外，SUN-database只能用于研究或教育的非商业用途，在获得原作者许可后可以获取自由访问。

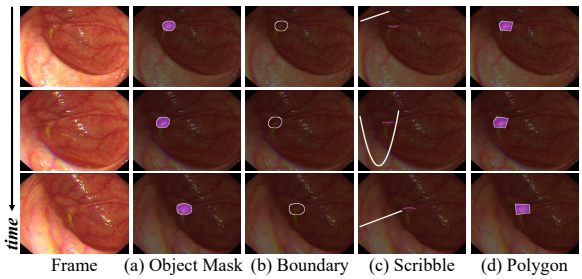


图2 SUN-SEG中的每个视频帧的多样化标签，包括目标掩码（a）、边缘标签（b）和两个弱标签，即：线标签（c）和多边形标签（d）。更多的细节请参见第3.2节。

3.2 专业标注

本文遵循文献[58]采用了相似的标注流程。根据SUN-database所提供的原始包围盒标签，十位富有经验的标注者使用Adobe Photoshop软件来提供各式的标签。接着，三位结肠镜相关的研究人员重新核验了这些初始标签的质量和正确性。图1展示了高质量控制标准背后的两个（即：拒绝和通过）典型样例。除了SUN[4]所提供的原始病理标签，如病理模式（例如：低度恶性腺瘤、增生性息肉等）、形状（例如：带蒂、近带蒂等）和位置（例如：盲肠、升结肠等），本文在SUN-SEG数据集中进一步以多样化的标签对其进行扩展。新扩展的标签由以下五个层次组成：视觉属性标签→ 目标掩码→ 边缘标签→ 线标签→ 多边形标签。挑选的样本和相应的标签请参见图2。其详细描述[2]如下：

- 视觉属性： 根据视频的视觉特性，本文给出基于视频粒度的9个视觉属性。其分类标准的详细描述请参见表2。
- 目标掩码： 正确分析病变区域对临床医师十分有帮助。因此，如图2 (a)所示，本文为每个视频帧提供了像素级别的目标掩码。本文在目标掩码的基础上进一步细化了原始包围盒的坐

[2]更为完整的标签描述请参见：https://github.com/GewelsJI/VPS/blob/main/docs/DATA_DESCRIPTION.md。

表2 视觉属性及其描述列表。

视觉属性	描述
SI	医疗器械： 内窥镜手术过程中的仪器，例如：夹子、镊子、手术刀和电极。
IB	模糊的边缘： 目标周围的前景和背景区域具有相似的颜色。
HO	明显的目标： 目标区域有显著区分的颜色。
GH	伪影： 目标快速移动或刷新率不足导致边缘的RGB颜色异常。
FM	高速运动： 在某视频片段中目标的逐帧平均运动大于20像素，可计算为相邻帧之间息肉质心的欧氏距离。
SO	小目标： 在一个视频片段中，目标大小与图像区域之间平均占比小于0.05。
LO	大目标： 在一个视频片段中，目标大小与图像区域之间平均占比大于0.15。
OC	遮挡： 息肉目标被部分或完全遮挡。
OV	视线越界： 息肉目标被图像边界剪裁。
SV	尺度变化： 在一个视频片段中，目标掩码和其包围盒的平均面积比率小于0.5。

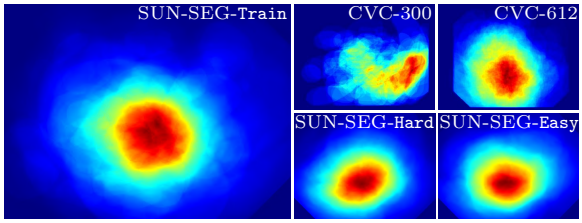


图3 在CVC-300数据集、CVC-612数据集和SUN-SEG-`Train/-Easy/-Hard`子数据集上所计算的中心偏置[60]。

标，使其与目标更贴合，从而提供更为可靠的定位标签。

- 边缘标签： 图2 (b)展示了通过计算目标掩码的图像梯度所生成的息肉边缘。
- 线标签： 此外，本文还提供了两个弱标签，以促进在数据不足的情况下的研究工作。对于图2 (c)中的线标签，本文使用两条高阶曲线分别表示前景（紫色曲线）和背景（白色曲线）。它们由线性函数或者二次函数进行随机生成，从而保证了不同标注者不同的主观性。
- 多边形标签： 类似地，在图2 (d)中，本文使用Douglas-Peucker算法[59]来随机地找到拟合目标边缘的外切多边形或内接多边形。

3.3 数据集统计

为了更好地展示，本节讨论了SUN-SEG的三个子数据集中几个重要的统计数据。关于SUN-SEG数据集切分的更多细节请参见第4.4.1节。

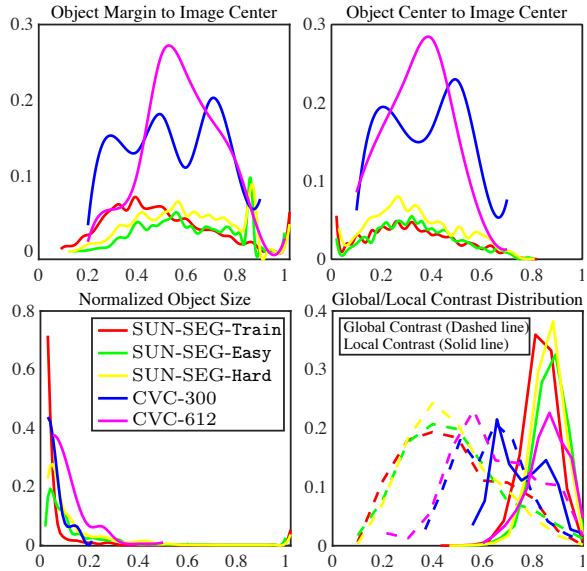


图4 现有的视频息肉分割数据集（CVC-300和CVC-612）与SUN-SEG-`Train/-Easy/-Hard`三个子数据集的统计曲线。注意到，水平轴和垂直轴分别表示频率及其统计指标。这些曲线体现了本文数据集的多样性。

- 中心偏置： 与通用目标检测任务不同，因为目标通常不在图像的中心，医学图像通常具有较高的中心偏置。为了描述中心偏置的程度，本文计算了每个数据集真值图的平均数据分布。图3和图4（上部分）显示了SUN-SEG的三个子数据集的中心偏置比CVC-300/-612低。
- 息肉尺度： 结肠镜检查是一种基于相机自运动的视频采集。这与通用领域中通过固定摄像机拍摄运动目标（例如：不可数事物和可数目标）不同。因此，息肉的大小变化和摄像机不

规则运动导致了其尺度不一。息肉会在视野中部分甚至完全消失。图4（左下）展示了在五个视频息肉分割数据集上的息肉尺度统计比较。

- 全局/局部对比： 为展示结肠息肉识别的困难程度，如图4（右下）所示，本文使用全局和局部对比策略[61]来进行定量描述。

4 视频息肉分割基线方法

本节首先在第4.1节明确了任务的定义。接着描述了PNS+的细节，包括归一化自注意力模块（请参见第4.2节）、全局至局部学习策略（请参见第4.3节）以及实现细节（请参见第4.4节）。

4.1 任务定义

本文主要研究视频息肉分割任务。它可以被定义为一个二元视频目标分割任务，即识别息肉与非息肉区域。具体来说，本文的目标是建立一个算法模型，为每帧分配逐像素的概率预测（即：一个从0到1的非二值掩码）。此外，本文将其他类型任务（例如：视频息肉检测）留给未来进行探索。

4.2 归一化自注意力模块

近年来，自注意力机制[62]在诸多流行的计算机视觉任务中得到了广泛的应用。本文初步研究发现，在不同拍摄角度和速度下所获取的息肉具有多尺度特性；那么，简单地将原始自注意力机制引入到视频息肉分割任务中将无法获得令人满意的结果（高精度和高速度）。若直接使用朴素的自注意力方案（例如：非局部网络[62]），会产生较高的计算成本并限制了推理速度。如图5（右）所示，由于动态更新感受野对基于自注意力的网络非常重要，因此本文引入了归一化自注意力（NS）模块。NS模块包含五个关键步骤，具体内容如下：

4.2.1 增强法则

受到视频显著性目标检测模型[63]启发，本文运用三个策略：包含**通道分离**、**查询相关**和**归一化**法则来降低计算代价并提高准确率。

（a）通道分离法则： 具体来说，给定三个大小为$\mathbb{R}^{T\times H\times W\times C}$的输入特征（即：查询特征$Q$、键特征$K$和值特征$V$），本文运用三个线性嵌入函数$\theta(\cdot)$、$\phi(\cdot)$和$g(\cdot)$来生成相应的注意力特征。该函数可以使用$1\times 1\times 1$核大小的卷积层实现[62]。注意到，$T$、$H$、$W$和$C$分别代表给定特征的帧数、高度、宽度和通道数。该法则可被表示为：

$$Q_i = \mathcal{F}^G\langle\theta(Q)\rangle, K_i = \mathcal{F}^G\langle\phi(K)\rangle, V_i = \mathcal{F}^G\langle g(V)\rangle, \tag{1}$$

其中$\mathcal{F}^G$函数表示沿着通道维度将每个注意力特征划分为N组的操作，用于生成分离的查询特征、分离的键特征和分离的值特征，即：$\{Q_i, K_i, V_i\} \in \mathbb{R}^{T\times H\times W\times \frac{C}{N}}$，其中$i = \{1, 2, \cdots, N\}$。

（b）查询相关法则： 为了建模连续帧之间的时空关系，则需要度量分离的查询特征$\{Q_i\}_{i=1}^N$和分离的键特征$\{K_i\}_{i=1}^N$之间的相似性。受到文献[63]启发，本文引入N个关联性度量模块（即：查询相关法则），用于计算目标像素点在受约束邻域内的时空相似度矩阵。如文献[62]所述，关联性度量模块可用于捕获T帧内关于目标对象更多的相关性，而不是计算查询位置和键特征所有位置之间的响应。具体而言，得到了特征Q_i在(x, y, z)位置的查询像素$\mathbf{X}^q$所对应的K_i约束邻域，该邻域可以通过一个点采样函数$\mathcal{F}^S$来获得。计算公式如下：

$$\mathcal{F}^S\langle\mathbf{X}^q, K_i\rangle = \Sigma_{m=x-kd_i}^{x+kd_i}\Sigma_{n=y-kd_i}^{y+kd_i}\Sigma_{t=1}^{T}K_i(m, n, t), \tag{2}$$

其中$1 \le x \le H$、$1 \le y \le W$、$1 \le z \le T$和$\mathcal{F}^S\langle\mathbf{X}^q, K_i\rangle \in \mathbb{R}^{T(2k+1)^2\times\frac{C}{N}}$。因此，约束邻域的范围取决于不同时空感受野的设定，即它

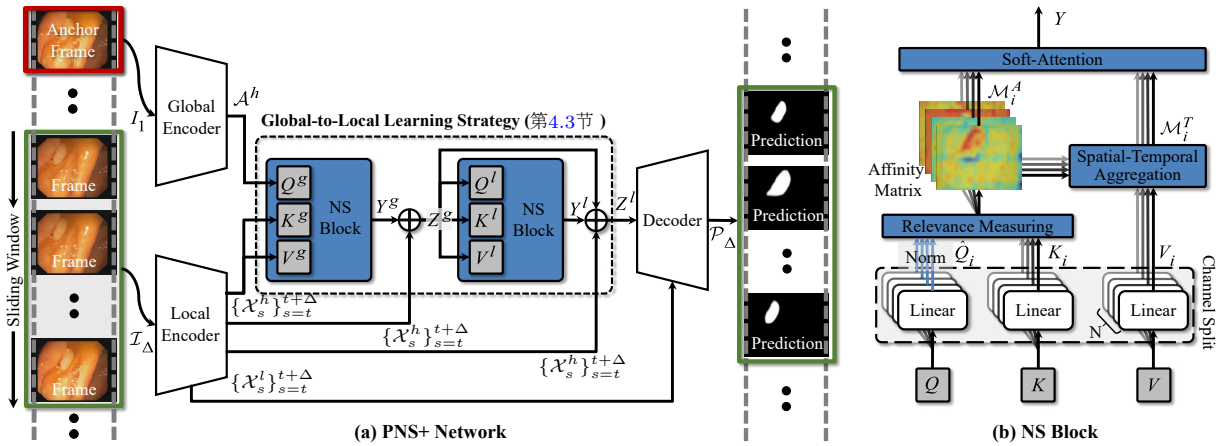


图5 基于归一化自注意力（NS）模块（b）的PNS+网络的框架流程图（a）。

们分别具有不同的核大小k、在第i组时的膨胀率d_i和帧数T。

（c）归一化法则： 然而，输入特征Q_i在前馈过程存在内部协变量偏移问题[64]，这将导致层参数不能动态地适应下一个小批次数据。因此，本文通过以下方式来保持特征Q_i的固定分布：

$$\hat{Q}_i = \texttt{Norm}(Q_i), \tag{3}$$

其中$\texttt{Norm}(\cdot)$是通过沿时间维度的层归一化[65]操作来实现的。

4.2.2 关联性度量

相似度矩阵$\mathcal{M}_i^A$用于度量以自适应的点采样方式而来的目标像素及其周围时空内容的相似性（请参见公式(2)），该过程被定义为：

$$\mathcal{M}_i^A = \texttt{Softmax}(\frac{\hat{Q}_i \mathcal{F}^S\langle \hat{\mathbf{X}}^q, K_i\rangle^{\mathbf{T}}}{\sqrt{C/N}}), \text{ when } \hat{\mathbf{X}}^q \in \hat{Q}_i, \tag{4}$$

其中$\mathcal{M}_i^A \in \mathbb{R}^{THW\times T(2k+1)^2}$。$\sqrt{C/N}$代表缩放因子，用于平衡多头注意力。

4.2.3 时空聚合

与关联性度量类似，本文也在时间聚合过程中计算了约束邻域内的时空聚合特征$\mathcal{M}_i^T \in \mathbb{R}^{THW\times \frac{C}{N}}$。该过程可以被表述为：

$$\mathcal{M}_i^T = \mathcal{M}_i^A \mathcal{F}^S\langle \mathbf{X}^a, V_i\rangle, \text{ when } \mathbf{X}^a \in \mathcal{M}_i^A. \tag{5}$$

4.2.4 软注意力

本文使用软注意力模块去融合来自相似度矩阵的组特征$\mathcal{M}_i^A$和聚合特征$\mathcal{M}_i^T$。在融合过程中，相关的时空模式应当被增强，而弱相关的时空模式应当被抑制。首先，沿着通道维度拼接一组相似度矩阵$\mathcal{M}_i^A$，从而得到$\mathcal{M}^A$。软注意力图$\mathcal{M}^S$则由下式计算而来：

$$\mathcal{M}^S \in \mathbb{R}^{THW\times 1} = \texttt{Max}(\mathcal{M}^A), \tag{6}$$

其中$\mathcal{M}^A \in \mathbb{R}^{THW\times T(2k+1)^2 N}$和$\texttt{Max}(\cdot)$函数计算了通道维度的最大值。接着，本文在通道维度拼接一组时空聚合的组特征$\mathcal{M}_i^T$以生成$\mathcal{M}^T$。

4.2.5 归一化自注意力

最后，本文的归一化自注意力模块（即$\texttt{NS}(\cdot,\cdot,\cdot)$函数）可被定义为：

$$Y \in \mathbb{R}^{T\times H\times W\times C} = \texttt{NS}(Q,K,V) = (\mathcal{M}^T\mathbf{W}_T)\circledast\mathcal{M}^S, \tag{7}$$

其中$\mathbf{W}_T$代表可学习的权值，$\circledast$代表在通道维度使用哈达玛乘积操作。

4.3 全局至局部学习策略

通过建立候选特征间的密集关联，非局部算子[62]展现了其对于短期时空相关性建模的潜力。然而，由于计算资源的限制，这种机制在建模长期的时空依赖关系时仍然受限，即网络只能处理其中一小段视频。为了解决这个问题，本文提出了一种新颖的学习策略，用于在任意时间距离长度下实现长期和短期时空信息的传播，从而得到一个更为有效的框架：PNS+。其包括如下五个步骤。

4.3.1 全局编码器

本方案采用第一帧$I_1 \in \mathbb{R}^{H' \times W' \times 3}$作为锚帧（即：全局参考），并用来计算锚帧和滑动窗口内连续帧之间的依赖关系。遵循文献[49]，本文使用相同的骨架网络（即：Res2Net-50[66]），并从conv4_6层来提取特征。为了减轻计算负担，本文采用一个类似RFB[67]的模块，用来减少特征通道维度并生成锚特征$\mathcal{A}^h \in \mathbb{R}^{H^h \times W^h \times C^h}$。

4.3.2 局部编码器

局部编码器从滑动窗口中选取一段连续的帧$\mathcal{I}_\Delta = \{I_s\}_{s=t}^{t+\Delta} \in \mathbb{R}^{H' \times W' \times 3}$ $(t>1)$作为输入。与全局编码器类似，本方案从Res2Net-50骨架网络的conv3_4和conv4_6层提取两组短期特征，并使用通道缩减模块来生成底层短期特征$\{\mathcal{X}_s^l\}_{s=t}^{t+\Delta} \in \mathbb{R}^{H^l \times W^l \times C^l}$和高层短期特征$\{\mathcal{X}_s^h\}_{s=t}^{t+\Delta} \in \mathbb{R}^{H^h \times W^h \times C^h}$。本文默认实现设定为：$H^l = \frac{H'}{4}$、$W^l = \frac{W'}{4}$、$C^l = 24$、$H^h = \frac{H'}{8}$、$W^h = \frac{W'}{8}$和$C^h = 32$。

4.3.3 全局时空建模

如图5所示，本方案利用第一个NS模块来建模任意时间距离下的长期关联性，该过程接受一个四维的时序特征作为输入，因此有：

$$\begin{aligned} \tilde{\mathcal{X}}^h \in \mathbb{R}^{\Delta \times H^h \times W^h \times C^h} &\Leftarrow \{\mathcal{X}_s^h\}_{s=t}^{t+\Delta} \in \mathbb{R}^{H^h \times W^h \times C^h}, \\ \tilde{\mathcal{A}}^h \in \mathbb{R}^{1 \times H^h \times W^h \times C^h} &\Leftarrow \mathcal{A}^h \in \mathbb{R}^{H^h \times W^h \times C^h}, \end{aligned} \tag{8}$$

其中，$\Leftarrow$表示将候选特征重构为时间形式，从而生成一个四维张量。随后，本方案将锚特征和高层短期特征分别用作查询条目（即：$Q^g = \tilde{\mathcal{A}}^h$）和键条目及值条目（即：$K^g = \tilde{\mathcal{X}}^h$ 和$V^g = \tilde{\mathcal{X}}^h$）。直观来看，其目标是在锚特征和高层局部特征之间建立像素级别的相似度，可被视为对全局时空关联性的建模。这个过程可被定义为：

$$Z^g \in \mathbb{R}^{\Delta \times H^h \times W^h \times C^h} = \texttt{NS}(\tilde{\mathcal{A}}^h, \tilde{\mathcal{X}}^h, \tilde{\mathcal{X}}^h) \oplus \tilde{\mathcal{X}}^h, \tag{9}$$

其中$\oplus$表示逐元素加法的残差操作[68]。该操作为第一个NS模块内的梯度传播提供了良好的收敛稳定性，使得它可以轻易地插入预训练网络中。

4.3.4 全局至局部传播

此外，本方案希望将长期依赖Z^g传播到局部邻域（即：滑动窗口中的视频帧）之中。因此，可将Z^g作为第二个NS模块的输入，即：查询特征$Q^l = Z^g$、键特征$K^l = Z^g$、值特征$V^l = Z^g$。因此有：

$$Z^l = \texttt{NS}(Z^g, Z^g, Z^g) \oplus Z^g \oplus \tilde{\mathcal{X}}^h. \tag{10}$$

通过引入两个残差连接的方式，可以保持第二个NS模块的内部梯度稳定性（即：$\oplus Z^g$）和外部梯度稳定性（即：$\oplus \tilde{\mathcal{X}}^h$）。

4.3.5 解码器和优化目标

最后，本文使用一个两阶段UNet型解码器$\mathcal{F}^D$，将来自局部编码器的底层短期特征$\mathcal{X}_s^l$和来自第二个NS模块的时空特征Z^l聚合起来。在聚合之前，本文将特征Z^l恢复为空间形式，即$\{Z_s^l\}_{s=t}^{t+\Delta}$。解

码器的预测用以下方法计算:

$$\mathcal{P}_{\Delta} = \{P_s\}_{s=t}^{t+\Delta} = \mathcal{F}^{D}\langle\{\mathcal{X}_s^l\}_{s=t}^{t+\Delta}, \{Z_s^l\}_{s=t}^{t+\Delta}\rangle. \quad (11)$$

为此，给定时间戳为s的预测图P_s和与其相应的真值图（GT）G_s，本文利用二值交叉熵损失函数进行优化，其表达式为:

$$\mathcal{L}_{bce} = -\sum[G_s \log(P_s) + (1 - G_s)\log(1 - P_s))]. \quad (12)$$

4.4 实现细节

4.4.1 数据集

本文随机切分出40%的SUN-SEG数据用于训练，即包含了112个视频（19,544帧）的SUN-SEG-`Train`。剩余的数据用于测试，其包含119个视频片段（17,070帧）的SUN-SEG-`Easy`和包含54个视频片段（12,522帧）的SUN-SEG-`Hard`，其是根据在每个病理类别中难度级别进行切分而来。具体来说，两种结肠镜检查场景（即：Seen和Unseen）[3]包含在上述的两个测试数据集中：SUN-SEG-`Easy`（Seen：33个视频片段和Unseen：86个视频片段）和SUN-SEG-`Hard`（Seen：17个视频片段和Unseen：37个视频片段），以得到更具细粒度的实验分析。

4.4.2 训练细节

本文在配备Intel Xeon (R) CPU E5-2690v4 × 24和四张NVIDIA Tesla V100 16 GB GPU的服务器平台上，使用SUN-SEG-`Train`数据集来训练本文的模型。在训练前加载Res2Net-50[66]的ImageNet预训练权值，新添加的层都经过Kaiming权值初始化。将批大小设置为24，这需要大约5小时和15个迭代周期来达到模型收敛。对于每一小批次数据，选择某个视频中的第一帧作为锚帧，并从同一视频中随机选取连续的五帧(即：Δ=5)。Adam优化器的初始学习速率和权值衰减分别设置为3e-4和1e-4。默认情况下，本文将注意力组的数量设置为N=4。对于第一个NS模块，本文设置卷积核大小k=3和膨胀率d_i=$\{3, 4, 3, 4\}$以捕获更大的感受野及更长期的表征。对于第二个NS模块，本文设置相同的卷积核大小k=3并降低膨胀率d_i=$\{1, 2, 1, 2\}$以主要关注短期的关联。

[3]Seen表示测试数据集中的帧样本来自训练集中相同的息肉，而Unseen表示训练集中不存在该息肉。

4.4.3 推理阶段

本文在SUN-SEG-`Easy`和SUN-SEG-`Hard`数据子集上的Seen和Unseen场景评测了所提出的PNS+模型。与训练阶段类似的是，在推理过程中本文也选择第一帧作为锚帧，并从视频中选取5帧（Δ=5），同时将其图片大小缩放为256×448。本文使用网络的输出$\mathcal{P}_{\Delta}$作为最终的输出，其后接一个*Sigmoid*函数。本文的PNS+在单张V100显卡上取得了约170fps的推理速度，且无需任何启发式的后处理技术（例如：DenseCRF[69]）。

5 视频息肉分割评测基准

5.1 评测协议

5.1.1 对比模型

为了实现系统的评估，本文挑选了13个近期的息肉/目标分割方法作为视频息肉分割评测基准的对比模型，其中包括5个基于图像的方法（即：UNet [35]、UNet++ [36]、ACSNet [39]、PraNet [49]、SANet [42]）和8个基于视频的方法（即：COSNet [70]、MAT [71]、PCSA [63]、2/3D [3]、AMD [72]、DCF [73]、FSNet [74]和PNSNet [5]）。为了公平比较，所有的对比模型在其相应的默认训练设置下，使用

表3 在两个具有结肠镜检查Seen场景的测试子数据集上的定量比较结果。

Model	SUN-SEG-Easy (Seen)				SUN-SEG-Hard (Seen)			
	$\mathcal{S}_\alpha$	E_ϕ^{mn}	F_β^w	Dice	$\mathcal{S}_\alpha$	E_ϕ^{mn}	F_β^w	Dice
COSNet	0.845	0.836	0.727	0.804	0.785	0.772	0.626	0.725
MAT	0.879	0.861	0.731	0.833	0.840	0.821	0.652	0.776
PCSA	0.852	0.835	0.681	0.779	0.772	0.759	0.566	0.679
2/3D	0.895	0.909	0.819	0.856	0.849	0.868	0.753	0.809
AMD	0.471	0.526	0.114	0.245	0.480	0.536	0.115	0.231
DCF	0.572	0.591	0.357	0.398	0.603	0.602	0.385	0.443
FSNet	0.890	0.895	0.818	0.873	0.848	0.859	0.755	0.828
PNSNet	0.906	0.910	0.836	0.861	0.870	0.892	0.787	0.823
PNS+	**0.917**	**0.924**	**0.848**	**0.888**	**0.887**	**0.929**	**0.806**	**0.855**

与PNS+相同的数据集训练到收敛为止。值得注意的是，本文专注于探索SUN-SEG数据集中的阳性样本，而把阴性样本（无息肉）留到未来的工作中。

5.1.2 评测指标

为了更深入地了解模型性能，本文使用以下六个不同的指标来评估时间戳s下的预测图P_s和真值图G_s，包括：（a）Dice指标（Dice $= \frac{2\times|P_s\cap G_s|}{|P_s\cup G_s|}$）用于评估预测图和真实图之间的相似性，并惩罚假阳/阴性。$\cap$、$\cup$和$|\cdot|$分别表示区域内的交集、并集和像素数目。（b）像素级别的灵敏度（Sen $= \frac{|P_s\cap G_s|}{|G_s|}$）用于评估整体病变区域的真阳性预测。由于结肠镜检查目是筛查息肉需要有较低的息肉缺失率，有息肉的患者应该极有可能被识别出来。因此，可以通过灵敏度来惩罚假阴性预测，即正确检测息肉的能力。（c）F指标[75]（$F_\beta = \frac{(1+\beta^2)\times\mathrm{Prc}\times\mathrm{Rcl}}{\beta^2\times(\mathrm{Prc}+\mathrm{Rcl})}$）被广泛地用于评估二值掩码，以$\beta$为权重对精确度和召回率进行调和均值计算，将准确率（Prc $= \frac{|P_s\cap G_s|}{|P_s|}$）和召回率（Rcl $= \frac{|P_s\cap G_s|}{|G_s|}$）综合考量以进行更全面的评估。（d）根据文献[76, 77]所述，加权F指标[78]（$F_\beta^w = \frac{(1+\beta^2)\times\mathrm{Prc}^w\times\mathrm{Rcl}^w}{\beta^2\times(\mathrm{Prc}^w+\mathrm{Rcl}^w)}$）修正了Dice指标和$F_\beta$中的“同等重要的缺陷”，以提供了更可靠的评估结果。本文遵循文献[79]建议，分别将F_β和F_β^w的参数β^2设置为0.3和1。（e）不同于以上像素级别评测指标，结构指标[80]（$\mathcal{S}_\alpha = \alpha\times\mathcal{S}_o(P_S,G_S)+(1-\alpha)\times\mathcal{S}_r(P_S,G_S)$）分别用于衡量目标$\mathcal{S}_o$和区域$\mathcal{S}_r$的结构相似性。本文默认使用参数$\alpha=0.5$。（f）Fan等人提出了一个基于人类视觉感知的指标：增强匹配指标[81]（$E_\phi = \frac{1}{W\times H}\sum_x^W\sum_y^H\phi(P_s(x,y),G_s(x,y))$），其中$\phi$是增强匹配矩阵。$W$和$H$是真值图$G_s$的宽度和高度。该指标本质上即适合在结肠镜检查中评估息肉异质的位置和形状。

如在第4.1节中提到的，模型生成连续的浮点数预测，因此需要将浮点值透过从0到255的阈值转换为二进制值。具体地，本文提供了Dice指标的最大值和E_ϕ、F_β和灵敏度在不同阈值下二值指标的平均值。一键测评工具箱可在https://github.com/GewelsJI/VPS/tree/main/eval中获取。

5.2 定量结果对比

基于上述评测协议，本文对两个测试子数据集（即：SUN-SEG-Easy和 SUN-SEG-Hard）进行了全面的视频息肉分割基准评测，其包括以下三个方面：

5.2.1 学习能力

值得注意的是，基于图像的模型是逐帧进行训练和推理的。为了更好地揭示结肠镜检查视频的时空学习能力，本文在两个Seen子集上验证了基于视频的对比模型的性能。对于这些显示在表3中的子数据集，本文的PNS+ 超越了基于视频的最优方法，例如：SUN-SEG-Easy（Seen）上的Dice指标：PNSNet（0.861）*vs.* PNS+（0.888）和SUN-SEG-Hard（Seen）上的F_ϕ^{mn}指标：PNSNet（0.892）*vs.* PNS+（0.929）。以上结果表明，本文的模型具有较强的学习能力，能够准确地分割息肉。

表4 两个测试子数据集在结肠镜检查的Unseen场景上的定量结果比较。'R/T'表示使用作者提供的代码重新训练的非公开模型。最好的分数会以**粗体**显示。

	模型	出版单位	代码	SUN-SEG-Easy (Unseen)						SUN-SEG-Hard (Unseen)					
				$\mathcal{S}_\alpha$	E_ϕ^{mn}	F_β^w	F_β^{mn}	Dice	Sen	$\mathcal{S}_\alpha$	E_ϕ^{mn}	F_β^w	F_β^{mn}	Dice	Sen
IMAGE	UNet[35]	MICCAI_{15}	Link	0.669	0.677	0.459	0.528	0.530	0.420	0.670	0.679	0.457	0.527	0.542	0.429
	UNet++[36]	TMI_{18}	Link	0.684	0.687	0.491	0.553	0.559	0.457	0.685	0.697	0.480	0.544	0.554	0.467
	ACSNet[39]	MICCAI_{20}	Link	0.782	0.779	0.642	0.688	0.713	0.601	0.783	0.787	0.636	0.684	0.708	0.618
	PraNet[49]	MICCAI_{20}	Link	0.733	0.753	0.572	0.632	0.621	0.524	0.717	0.735	0.544	0.607	0.598	0.512
	SANet[42]	MICCAI_{21}	Link	0.720	0.745	0.566	0.634	0.649	0.521	0.706	0.743	0.526	0.580	0.598	0.505
VIDEO	COSNet[70]	TPAMI_{19}	Link	0.654	0.600	0.431	0.496	0.596	0.359	0.670	0.627	0.443	0.506	0.606	0.380
	MAT[71]	TIP_{20}	Link	0.770	0.737	0.575	0.641	0.710	0.542	0.785	0.755	0.578	0.645	0.712	0.579
	PCSA[63]	AAAI_{20}	Link	0.680	0.660	0.451	0.519	0.592	0.398	0.682	0.660	0.442	0.510	0.584	0.415
	2/3D[3]	MICCAI_{20}	R/T	0.786	0.777	0.652	0.708	0.722	0.603	0.786	0.775	0.634	0.688	0.706	0.607
	AMD[72]	NeurIPS_{21}	Link	0.474	0.533	0.133	0.146	0.266	0.222	0.472	0.527	0.128	0.141	0.252	0.213
	DCF[73]	ICCV_{21}	Link	0.523	0.514	0.270	0.312	0.325	0.340	0.514	0.522	0.263	0.303	0.317	0.364
	FSNet[74]	ICCV_{21}	Link	0.725	0.695	0.551	0.630	0.702	0.493	0.724	0.694	0.541	0.611	0.699	0.491
	PNSNet[5]	MICCAI_{21}	Link	0.767	0.744	0.616	0.664	0.676	0.574	0.767	0.755	0.609	0.656	0.675	0.579
	PNS+	**OURS_{22}**	Link	**0.806**	**0.798**	**0.676**	**0.730**	**0.756**	**0.630**	**0.797**	**0.793**	**0.653**	**0.709**	**0.737**	**0.623**

表5 在SUN-SEG-Easy和SUN-SEG-Hard（Unseen）上基于视觉属性的结构指标（$\mathcal{S}_\alpha$）比较。

	SUN-SEG-Easy (Unseen)										SUN-SEG-Hard (Unseen)									
	SI	**IB**	**HO**	**GH**	**FM**	**SO**	**LO**	**OC**	**OV**	**SV**	**SI**	**IB**	**HO**	**GH**	**FM**	**SO**	**LO**	**OC**	**OV**	**SV**
UNet	0.675	0.548	0.768	0.715	0.633	0.593	0.648	0.670	0.643	0.620	0.618	0.619	0.663	0.676	0.713	0.689	0.633	0.658	0.659	0.658
UNet++	0.701	0.542	0.782	0.739	0.647	0.591	0.678	0.683	0.665	0.617	0.654	0.604	0.665	0.696	0.714	0.681	0.660	0.676	0.677	0.678
ACSNet	0.789	0.612	0.896	0.820	0.704	0.663	0.787	0.770	0.759	0.705	0.770	0.681	0.828	0.795	0.817	0.738	0.810	**0.828**	0.806	0.759
PraNet	0.745	0.585	0.821	0.772	0.673	0.611	0.722	0.722	0.703	0.653	0.673	0.635	0.725	0.720	0.755	0.691	0.666	0.714	0.708	0.703
SANet	0.724	0.582	0.854	0.760	0.676	0.615	0.703	0.701	0.711	0.680	0.658	0.565	0.738	0.709	0.760	0.692	0.733	0.729	0.727	0.693
COSNet	0.663	0.531	0.786	0.684	0.610	0.549	0.637	0.648	0.613	0.617	0.641	0.593	0.727	0.668	0.690	0.637	0.694	0.707	0.666	0.625
MAT	0.772	0.664	0.873	0.789	0.706	**0.691**	0.755	0.738	0.746	0.715	**0.772**	0.701	0.801	0.776	0.782	0.780	0.791	0.795	0.789	0.750
PCSA	0.676	0.563	0.759	0.708	0.628	0.610	0.634	0.662	0.656	0.616	0.656	0.591	0.692	0.683	0.706	0.671	0.612	0.677	0.665	0.663
2/3D	0.809	0.625	**0.899**	0.835	0.728	0.667	**0.820**	**0.783**	0.778	0.719	0.768	0.662	**0.865**	0.784	0.797	0.737	**0.853**	0.827	**0.808**	0.765
AMD	0.476	0.461	0.471	0.481	0.484	0.466	0.447	0.467	0.442	0.498	0.471	0.468	0.447	0.473	0.468	0.469	0.453	0.487	0.462	0.481
DCF	0.465	0.485	0.479	0.505	0.541	0.495	0.362	0.484	0.492	0.495	0.441	0.508	0.422	0.498	0.587	0.556	0.351	0.470	0.494	0.540
FSNet	0.719	0.603	0.810	0.752	0.694	0.632	0.686	0.711	0.691	0.665	0.662	0.648	0.743	0.713	0.774	0.723	0.701	0.728	0.728	0.694
PNSNet	0.789	0.592	0.871	0.820	0.723	0.619	0.768	0.749	0.751	0.705	0.746	0.631	0.803	0.780	0.778	0.743	0.805	0.790	0.794	0.758
PNS+	**0.819**	**0.667**	0.883	**0.844**	**0.738**	0.690	0.796	0.782	**0.798**	**0.734**	0.770	**0.703**	0.817	**0.801**	**0.823**	**0.793**	0.792	0.808	0.807	**0.795**

5.2.2 泛化能力

为了验证模型的泛化性，本文在两个测试子数据集上进行了实验，这些测试子数据集带有未知的结肠镜检查场景。如表4所示，本文在六个指标中展示了与其他最新的基于图像和视频的对比模型的性能比较。可以看到本文的PNS+ 在与基于图像和视频的方法的比较中取得了很大的提升，例如：SUN-SEG-Easy（Unseen）上的Dice指标：ACSNet（0.713）*vs.* 2/3D (0.722) *vs.* PNS+（0.756）和SUN-SEG-Hard（Unseen）上的F_β^w：ACSNet（0.636）*vs.* 2/3D（0.634）*vs.* PNS+（0.653）。有趣的是，本文观察到PNSNet模型的性能在两个Unseen数据集上显著下降，这侧面地反应了本文所提出的全局到局部学习策略具有更好的泛化能力，特别对于一个具有较大时间跨度的视频片段。

5.2.3 基于属性的性能

最后，本文分析了在表2中所呈现的基于视觉属性的比较。基于$\mathcal{S}_\alpha$指标，表5揭示了PNS+模型在四个属性（即：IB、GH、FM和SV）上始终优于其他的对比模型。更具体地说，如表5所示，大多数方法在IB属性上无法成功进行视频息肉分割任务，因为结肠息肉具有模糊的边界。相反

表6 PNS+中核心组件的消融实验。详细分析请参见第5.4节。

No.	VARIANTS Base	N	Soft	Norm	Strategy	SUN-SEG-Easy (Unseen) $\mathcal{S}_\alpha$	E_ϕ^{mn}	F_β^w	Dice	SUN-SEG-Hard (Unseen) $\mathcal{S}_\alpha$	E_ϕ^{mn}	F_β^w	Dice
#01	✓	-	-	-	-	0.729	0.718	0.571	0.616	0.726	0.720	0.559	0.603
#02	✓	1	✓	✓	L	0.782	0.766	0.631	0.722	0.783	0.775	0.629	0.715
#03	✓	2	✓	✓	L	0.773	0.760	0.625	0.720	0.785	0.784	0.631	0.719
#04	✓	4	✓	✓	L	0.786	0.777	0.651	0.741	0.792	0.789	0.649	0.735
#05	✓	8	✓	✓	L	0.774	0.762	0.627	0.724	0.775	0.774	0.619	0.708
#06	✓	4	-	✓	L	0.782	0.775	0.639	0.722	0.785	0.786	0.637	0.715
#07	✓	4	✓	-	L	0.755	0.752	0.587	0.705	0.754	0.751	0.579	0.694
#08	✓	4	✓	✓	L→L	0.748	0.717	0.577	0.705	0.760	0.741	0.587	0.693
#09	✓	4	✓	✓	L→G	0.788	0.780	0.645	0.741	0.776	0.768	0.618	0.715
#10	✓	4	✓	✓	G→G	0.778	0.763	0.627	0.726	0.767	0.753	0.599	0.694
#OUR	✓	4	✓	✓	G→L	**0.806**	**0.798**	**0.676**	**0.756**	**0.797**	**0.793**	**0.653**	**0.737**

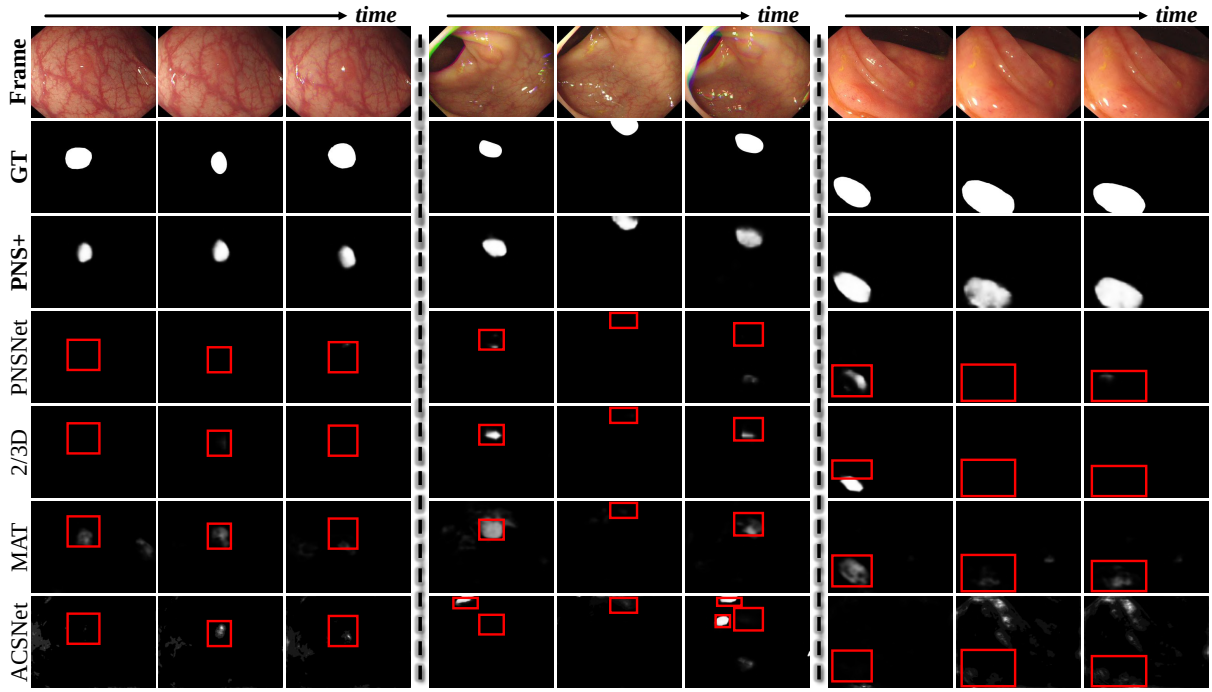
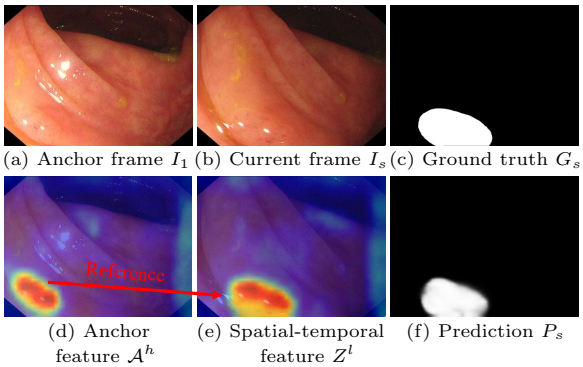


图6 在三个序列（从左到右依次为：case14_3、case30和case3_2）上对所提出的PNS+模型和四个具有代表性的对比模型进行定性结果的可视化。红框代表错误或遗漏的预测。推荐读者浏览项目主页上完整的动态对比图。

的，PNS+模型在SUN-SEG-Easy（Unseen）的这个具有挑战性的IB属性上取得了最好的性能（$\mathcal{S}_\alpha = 0.667$）。这个发现也与图6中显示的结果一致。同样地，SO属性也呈现出较低的分数（例如：SUN-SEG-Easy（Unseen）：$\mathcal{S}_\alpha$=0.667），这表明这两个属性是结肠镜检查中最具挑战性的难题。相反地，HO和LO属性相较其他属性始终保持较高的性能，使得息肉更容易被发现。上述现象符合本文的预期，因为在这些相对简单的场景中分布偏差较小。请参见第5.5节中对于具有挑战性场景更多的可视化分析。

5.3 定性结果对比

如图6所示，本文展示了四个经典模型（PNSNet、2/3D、MAT、ACSNet）和PNS+模型的可视化结果对比。在最后四行，这些对比模型未能对与背景具有相同纹理的息肉提供完整的



(a) Anchor frame I_1 (b) Current frame I_s (c) Ground truth G_s

(d) Anchor feature $\mathcal{A}^h$ (e) Spatial-temporal feature Z^l (f) Prediction P_s

图7 关键数据流的特征可视化。红色箭头表示使用锚特征$\mathcal{A}^h$来指导时空帧Z^l的表征。更多细节请参考第5.4.5节。

分割结果。相比之下，本文在第三行的模型可以在这种具有挑战性的情况（即：具有不同大小且与背景具有相同纹理的息肉）下准确地定位和分割息肉。

5.4 消融实验

为了验证本文核心设计的贡献，本文进行了广泛的消融实验，并将结果汇总在表6中。

5.4.1 基础网络的贡献

本文利用Res2Net-50[66]骨架网络初始化了一个UNet型变体#01，其可以被看作是一种基于图像的方法来生成逐帧预测。本文观察到**#OUR**在SUN-SEG-Easy（Unseen）上显著提高了性能（$\mathcal{S}_\alpha$指标: +7.7%）。

5.4.2 通道分离的贡献:

为探究公式(1)中通道分离法则的最佳版本，本节衍生出四个具有不同通道分离数目的变体模型：#02（N=1）、#03（N=2）、#04（N=4）和#05（N=8）。这些结果表明，过小的（#02 & #03）和过大的（#05）通道分离数目会破坏通道级别的信息，使知识在不同通道中发生崩塌。相比之下，本文采用合适的尺度（即：N=4），在SUN-SEG-Hard（Unseen）上与变体#05相比性能更好（如：Dice指标：2.7%↑）。这种权衡尺度将使本文模型更多聚焦于息肉相关的信息，同时抑制不相关内容。

5.4.3 软注意力的贡献

本文进一步移除软注意力，并观察到在SUN-SEG-Easy（Unseen）数据子集上，使用软注意力的变体模型#04普遍比没有软注意力的变体模型#06要好（如：Dice指标: 1.9%↑）。这一改进表明，为了提高性能，需要引入软注意操作来合成聚合特征和相似度矩阵。

5.4.4 归一化操作的贡献

通过比较变体模型#04和变体模型#07，本文还研究了归一化操作所带来的效果提升。本文观察到在SUN-SEG-Hard（Unseen）子数据集上变体模型#04通常优于变体模型#07（Dice指标：4.1%↑）。结果表明，通过沿时间维度的层归一化来稳定查询条目的分布可以解决注意力机制中的内部协变量偏移问题。

5.4.5 不同的学习策略

最后，本文通过派生出三个变体模型#08（局部至局部）、#09（局部至全局）、#10（全局至全局）和**#OUR**（全局至局部）来检验所提出学习策略（请参见第4.3节）的有效性。举例来说，变体模型#09会先结合局部时空线索，然后引入全局线索，这被称为局部至全局学习策略（L→G）。由于缺乏全局上下文，当关注局部信息时，变体模型#08在SUN-SEG-Easy（Unseen）（$\mathcal{S}_\alpha$指标: 5.8%↓）上显著退步。另一方面，如果只关注全局信息，变体模型#10的性能在SUN-SEG-Hard（Unseen）上明显下降（如：F_β^w指标: 5.4%↓）。相反地，由于将长期线索传播到短期邻域，使用全局至局部策

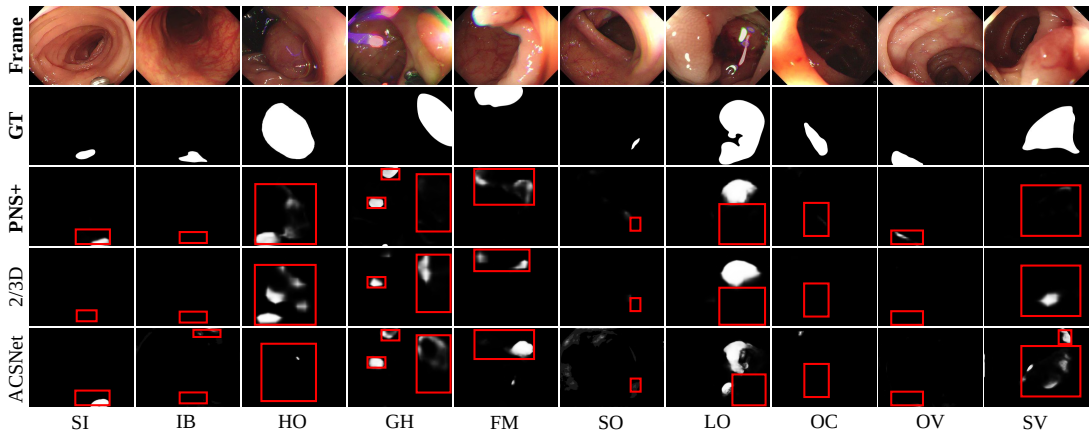


图8 取自10个视觉属性的、具有挑战性的样本。更多的分析可以参考第5.5节。

略的**#OUR**在SUN-SEG-Hard（Unseen）上优于变体#09（如：F_β^w指标: 3.5%↑）。

通过可视化关键数据流，本文进一步验证了全局到局部学习策略的有效性。如图7所示，第一列和第二列分别表示从全局编码器提取的锚特征$\mathcal{A}^h$和从第二个NS模块提取的时空特征Z^l。注意到，当前帧I_s是从连续帧I_Δ中随机选取的。它表明尽管当前帧I_s由于不确定的边界（IB属性）很难识别，PNS+模型仍可以在锚帧I_1的协助下传播长期依赖。值得注意的是，在图6的最右列，PNSNet模型因为缺乏全局到局部学习策略而无法准确定位息肉，与之相比，本文的模型成功地利用锚帧的全局参考来检测息肉。

5.5 问题和挑战

本节讨论具有挑战性属性中的一些常见问题，其可视化结果展示于图8。值得注意的是，视频息肉分割在医学成像领域是一个新兴且颇具挑战性的方向，其整体性能精度还不够高。本文观察到现有的前沿模型（例如：ACSNet和2/3D）和本文的基线模型（PNS+）在特定情况下（包括LO属性、HO属性、SI属性、GH属性和SV属性），仍然缺乏足够的鲁棒性。对于HO属性（第三列）和LO属性（第八列），由于外观变化较大，三个模型都无法捕获整个息肉。此外，手术器械（第一列）和光学伪影（第四列）上的假阳/阴性预测（红框标记）表明在这种困难的情况下，这些模型在缺乏准确的息肉相关表征时无法有效地学习语义信息。此外，SV属性（最后一列）的错误预测，是因为训练集中各种息肉形状的数据不足导致的。上述缺陷激励着我们探索更鲁棒的学习范式，以提高视频息肉分割的准确性。

本文还观察到，当病变区域与肠壁颜色相似或其尺寸太小以至于超过了输入图像的视场时，三个模型均无法准确地定位息肉区域。因此，通过伪装模式发掘技术[58, 82]来提高IB属性和SO属性的检测能力，有很大的探索空间。最后，由于缺乏对时间维度的理解将导致FM属性、OV属性和OC属性的错误预测。这里以OV属性和OC属性为例，由于遮挡在整个视频片段中是不连续的，更彻底地运用时间线索应该能够缓解由于肠壁或图像边界遮挡而导致的性能下降。综上所述，这些颇具挑战性的场景是其他方法会面临的共同难题，并导致性能大幅下降，这值得进一步探索。

6 潜在方向

本节重点介绍了在深度学习时代背景下结肠镜检查研究的几个潜在发展趋势。

- **高精度诊断**：如表4所示，本文观察到前沿的方法在SUN-SEG-Hard数据集上表现仍不满意（例如：灵敏度指标<0.63）。本文认为高精度的视频息肉分割算法将推动临床医学辅助诊断技术的发展。
- **数据受限学习**：在特定临床应用中，受限条件下探索有效的学习策略是非常有前景的，例如：弱/无/自监督学习和知识蒸馏等技术。
- **隐私保护AI**：视频息肉分割的智能系统必须在从训练到生产和管理的整个生命周期中保护数据，这会推动联邦学习等基础技术的发展。
- **可信赖AI**：人工智能引导的决策是如何制定的、又是什么样的决定性因素在扮演重要的角色，这对理解深度网络的内涵起着至关重要的作用。换句话说，视频息肉分割模型应该是因果性的、透明的、可解释的和互动性的，这能激发更可信的发展，例如：文献[83]。

上述视频息肉分割的潜在方向仍尚未解决。幸运的是，有一些经典工作可供参考，这提供了一个潜在的基础，以便迁移到本研究社区内。

7 结语

本文首次从深度学习的角度对视频息肉分割（VPS）进行了全方位的研究。首先提出了一个大规模的视频息肉分割数据集SUN-SEG，通过拓展经典的SUN-database得到各种标签，例如：属性标签、目标掩码、边缘标签、线标签和多边形标签。本文接着设计了一个简单且高效的基线模型，名为PNS+，用于从结肠镜检查视频中分割结肠息肉。PNS+模型基于归一化自注意力模块，通过一种全新的全局至局部学习策略，充分利用了长期和短期的时空线索。本文进一步贡献了第一个全面的评测基准，包含了13个前沿的息肉/目标分割方法。实验结果表明，PNS+模型在所有对比模型中取得了最佳性能。最后，本文概述了深度学习时代背景下结肠镜相关研究未来的几个潜在方向。我们希望本研究能够推动其他相关医学视频分析技术的发展。

References

[1] G.-P. Ji, G. Xiao, Y.-C. Chou, D.-P. Fan, K. Zhao, G. Chen, H. Fu, and L. Van Gool, "Video polyp segmentation: A deep learning perspective," [Online], 2022, Available: https://arxiv.org/abs/2203.14291.

[2] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognition*, vol. 45, no. 9, pp. 3166–3182, 2012, DOI: 10.1016/j.patcog.2012.03.002.

[3] 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 *International Conference on Medical Image Computing and Computer Assisted Intervention*. Lima, Peru: Springer, 2020, pp. 295–305, DOI: 10.1007/978-3-030-59725-2_29.

[4] 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)," *Gastrointestinal endoscopy*, vol. 93, no. 4, pp. 960–967, 2021, DOI: 10.1016/j.gie.2020.07.060.

[5] 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 *International Conference on Medical Image Computing and Computer Assisted Intervention*. Strasbourg, France: Springer, 2021, pp. 142–152, DOI: 10.1007/978-3-030-87193-2_14.

[6] 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," *International journal of computer assisted radiology and surgery*, vol. 9, no. 2, pp. 283–293, 2014, DOI: 10.1007/s11548-013-0926-3.

[7] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015, DOI: 10.1016/j.compmedimag.2015.02.007.

[8] P. Mesejo, D. Pizarro, A. Abergel, O. Rouquette, S. Beorchia, L. Poincloux, and A. Bartoli, "Computer-aided classification of gastrointestinal lesions in regular colonoscopy," *Transactions on Medical Imaging*, vol. 35, no. 9, pp. 2051–2063, 2016, DOI: 10.1109/TMI.2016.2547947.

[9] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, 2016, DOI: 10.1109/TMI.2015.2487997.

[10] "Gastrointestinal Image ANAlysis (GIANA) Challenge," https://endovissub2017-giana.grand-challenge.org/home/.

[11] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *Journal of Healthcare Engineering*, vol. 2017, p. 4037190, 2017, DOI: 10.1155/2017/4037190.

[12] A. Koulaouzidis, D. K. Iakovidis, D. E. Yung, E. Rondonotti, U. Kopylov, J. N. Plevris, E. Toth, A. Eliakim, G. W. Johansson, W. Marlicz *et al.*, "Kid project: an internet-based digital video atlas of capsule endoscopy for research purposes," *Endoscopy International Open*, vol. 5, no. 6, p. E477, 2017, DOI: 10.1055/s-0043-105488.

[13] D. K. Iakovidis, S. V. Georgakopoulos, M. Vasilakakis, A. Koulaouzidis, and V. P. Plagianakos, "Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification," *Transactions on Medical Imaging*, vol. 37, no. 10, pp. 2196–2210, 2018, DOI: 10.1109/TMI.2018.2837002.

[14] 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 *Multimedia Systems Conference*. Taipei, Taiwan: ACM, 2017, pp. 164–169, DOI: 10.1145/3083187.3083212.

[15] 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," [Online], 2020, available: https://arxiv.org/abs/2003.03376.

[16] 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," *Scientific Data*, vol. 7, no. 1, pp. 1–14, 2020, DOI: 10.1038/s41597-020-00622-y.

[17] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *International Conference on Multimedia Modeling*. Daejeon, Korea: Springer, 2020, pp. 451–462, DOI: 10.1007/978-3-030-37734-2_37.

[18] L. F. Sánchez-Peralta, J. B. Pagador, A. Picón, Á. J. Calderón, F. Polo, N. Andraka, 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," *Applied Sciences*, vol. 10, no. 23, p. 8501, 2020, DOI: 10.3390/app10238501.

[19] 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," *Scientific Data*, vol. 8, no. 1, pp. 1–10, 2021, DOI: 10.1038/s41597-021-00920-z.

[20] 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," *BMC Medical Imaging*, vol. 20, no. 1, pp. 1–14, 2020, DOI: 10.1186/s12880-020-00482-3.

[21] Y. Ma, X. Chen, K. Cheng, Y. Li, and B. Sun, "Ldpolypvideo benchmark: A large-scale colonoscopy video dataset of diverse polyps," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Strasbourg, France: Springer, 2021, pp. 387–396, DOI: 10.1007/978-3-030-87240-3_37.

[22] 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," *Plos one*, vol. 16, no. 8, p. e0255809, 2021, DOI: 10.1371/journal.pone.0255809.

[23] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O. E. Salem, D. Lamarque, C. Daul, K. V. Anonsen, M. A. Riegler *et al.*, "Polypgen: A multi-center polyp detection

and segmentation dataset for generalisability assessment," [Online], 2021, available: https://arxiv.org/abs/2106.04463.

[24] 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 *International Conference on Pattern Recognition*. Hong Kong, China: IEEE, 2006, pp. 695–698, DOI: 10.1109/ICPR.2006.268.

[25] A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y.-H. R. Tsai, "Automated polyp detection in colon capsule endoscopy," *Transactions on Medical Imaging*, vol. 33, no. 7, pp. 1488–1502, 2014, DOI: 10.1109/TMI.2014.2314959.

[26] O. H. Maghsoudi, "Superpixel based segmentation and classification of polyps in wireless capsule endoscopy," in *Signal Processing in Medicine and Biology Symposium*. Philadelphia, PA, USA: IEEE, 2017, pp. 1–4, DOI: 10.1109/SPMB.2017.8257027.

[27] 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," *Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 65–75, 2016, DOI: 10.1109/JBHI.2016.2637004.

[28] W. Tavanapong, J. Oh, M. Riegler, M. I. Khaleel, B. Mitta, and P. C. De Groen, "Artificial intelligence for colonoscopy: Past, present, and future," *Journal of Biomedical and Health Informatics*, pp. 1–1, 2022, DOI: 10.1109/JBHI.2022.3160098.

[29] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two-stream deep feature modelling for automated video endoscopy data analysis," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Lima, Peru: Springer, 2020, pp. 742–751, DOI: 10.1007/978-3-030-59716-0_71.

[30] 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," *Medical image analysis*, vol. 62, p. 101653, 2020, DOI: 10.1016/j.media.2020.101653.

[31] R. Zhang, Y. Zheng, C. C. Poon, D. Shen, and J. Y. Lau, "Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker," *Pattern Recognition*, vol. 83, pp. 209–219, 2018, DOI: 10.1016/j.patcog.2018.05.026.

[32] 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 *International Conference on Medical Image Computing and Computer Assisted Intervention*. Strasbourg, France: Springer, 2021, pp. 302–312, DOI: 10.1007/978-3-030-87240-3_29.

[33] P. Brandao, E. Mazomenos, G. Ciuti, R. Caliò, F. Bianchi, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo, and D. Stoyanov,

"Fully convolutional neural networks for polyp segmentation in colonoscopy," in *Medical Imaging 2017: Computer-Aided Diagnosis*. Orlando, FL, USA: SPIE, 2017, pp. 101–107, DOI: 10.1117/12.2254361.

[34] M. Akbari, M. Mohrekesh, E. Nasr-Esfahani, S. R. Soroushmehr, N. Karimi, S. Samavi, and K. Najarian, "Polyp segmentation in colonoscopy images using fully convolutional network," in *Engineering in Medicine and Biology Society*. Honolulu, HI, USA: IEEE, 2018, pp. 69–72, DOI: 10.1109/EMBC.2018.8512197.

[35] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Munich, Germany: Springer, 2015, pp. 234–241, DOI: 10.1007/978-3-319-24574-4_28.

[36] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019, DOI: 10.1109/TMI.2019.2959609.

[37] 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 *International Symposium on Multimedia*. San Diego, CA, USA: IEEE, 2019, pp. 225–2255, DOI: 10.1109/ISM46123.2019.00049.

[38] J. Zhong, W. Wang, H. Wu, Z. Wen, and J. Qin, "Polypseg: An efficient context-aware network for polyp segmentation from colonoscopy videos," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Lima, Peru: Springer, 2020, pp. 285–294, DOI: 10.1007/978-3-030-59725-2_28.

[39] R. Zhang, G. Li, Z. Li, S. Cui, D. Qian, and Y. Yu, "Adaptive context selection for polyp segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Lima, Peru: Springer, 2020, pp. 253–262, DOI: 10.1007/978-3-030-59725-2_25.

[40] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, and P. Halvorsen, "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *IEEE Access*, vol. 9, pp. 40 496–40 510, 2021, DOI: 10.1109/ACCESS.2021.3063716.

[41] 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 Conference on Artificial Intelligence*. [Online]: AAAI Press, 2021, pp. 2916–2924.

[42] J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, and S. Cui, "Shallow attention network for polyp segmentation," in *International Conference on Medical Image Computing and*

Computer Assisted Intervention. Strasbourg, France: Springer, 2021, pp. 699–708, DOI: 10.1007/978-3-030-87193-2_66.

[43] X. Zhao, L. Zhang, and H. Lu, "Automatic polyp segmentation via multi-scale subtraction network," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Strasbourg, France: Springer, 2021, pp. 120–130, DOI: 10.1007/978-3-030-87193-2_12.

[44] B. Murugesan, K. Sarveswaran, S. M. Shankaranarayana, K. Ram, J. Joseph, and M. Sivaprakasam, "Psi-Net: Shape and boundary aware joint multi-task deep network for medical image segmentation," in *Engineering in Medicine and Biology Society*. Germany, Germany: IEEE, 2019, pp. 7223–7226, DOI: 10.1109/EMBC.2019.8857339.

[45] R. Wang, S. Chen, C. Ji, J. Fan, and Y. Li, "Boundary-aware context neural network for medical image segmentation," *Medical Image Analysis*, vol. 78, p. 102395, 2022, DOI: 10.1016/j.media.2022.102395.

[46] Y. Fang, C. Chen, Y. Yuan, and K.-y. Tong, "Selective feature aggregation network with area-boundary constraints for polyp segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Shenzhen, China: Springer, 2019, pp. 302–310, DOI: 10.1007/978-3-030-32239-7_34.

[47] Y. Shen, X. Jia, and M. Q.-H. Meng, "Hrenet: A hard region enhancement network for polyp segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Strasbourg, France: Springer, 2021, pp. 559–568, DOI: 10.1007/978-3-030-87193-2_53.

[48] G.-P. Ji, L. Zhu, M. Zhuge, and K. Fu, "Fast camouflaged object detection via edge-based reversible re-calibration network," *Pattern Recognition*, vol. 123, p. 108414, 2022, DOI: https://doi.org/10.1016/j.patcog.2021.108414.

[49] 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 *International Conference on Medical Image Computing and Computer Assisted Intervention*. Lima, Peru: Springer, 2020, pp. 263–273, DOI: https://doi.org/10.1007/978-3-030-59725-2_26.

[50] T.-C. Nguyen, T.-P. Nguyen, G.-H. Diep, A.-H. Tran-Dinh, T. V. Nguyen, and M.-T. Tran, "Ccbanet: Cascading context and balancing attention for polyp segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Strasbourg, France: Springer, 2021, pp. 633–643, DOI: 10.1007/978-3-030-87193-2_60.

[51] M. Cheng, Z. Kong, G. Song, Y. Tian, Y. Liang, and J. Chen, "Learnable oriented-derivative network for polyp segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Strasbourg, France: Springer, 2021, pp. 720–730, DOI:

10.1007/978-3-030-87193-2_68.

[52] T. Kim, H. Lee, and D. Kim, "Uacanet: Uncertainty augmented context attention for polyp segmentation," in *Multimedia*. Chengdu, China: ACM, 2021, pp. 2167–2175, DOI: 10.1145/3474085.3475375.

[53] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," [Online], 2022, available: https://arxiv.org/abs/2201.09873.

[54] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Strasbourg, France: Springer, 2021, pp. 14–24, DOI: 10.1007/978-3-030-87193-2_2.

[55] S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, and R. S. M. Goh, "Medical image segmentation using squeeze-and-expansion transformers," in *International Joint Conference on Artificial Intelligence*. Montreal, Canada: IJCAI, 2021, DOI: 10.24963/ijcai.2021/112.

[56] 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," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022, DOI: 10.1007/s41095-022-0274-8.

[57] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-pvt: Polyp segmentation with pyramid vision transformers," [Online], 2021, available: https://arxiv.org/abs/2108.06932.

[58] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *Transactions on pattern analysis and machine intelligence*, pp. 1–1, 2021, DOI: 10.1109/TPAMI.2021.3085766.

[59] U. Ramer, "An iterative procedure for the polygonal approximation of plane curves," *Computer graphics and image processing*, vol. 1, no. 3, pp. 244–256, 1972, DOI: 10.1016/S0146-664X(72)80017-0.

[60] D.-P. Fan, J. Zhang, G. Xu, M.-M. Cheng, and L. Shao, "Salient objects in clutter," *Transactions on pattern analysis and machine intelligence*, pp. 1–1, 2022, DOI: 10.1109/TPAMI.2022.3166451.

[61] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks," *Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2075–2089, 2020, DOI: 10.1109/TNNLS.2020.2996406.

[62] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Conference on computer vision and pattern recognition*. Salt Lake City, UT, USA: IEEE, 2018, pp. 7794–7803, DOI: 10.1109/CVPR.2018.00813.

[63] Y. Gu, L. Wang, Z. Wang, Y. Liu, M.-M. Cheng, and S.-P. Lu, "Pyramid constrained self-attention network for fast video

salient object detection," in *AAAI Conference on Artificial Intelligence*, vol. 34. New York, New York, USA: AAAI Press, 2020, pp. 10 869–10 876, DOI: 10.1609/aaai.v34i07.6718.

[64] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, and H. Lu, "Normalized and geometry-aware self-attention network for image captioning," in *Conference on computer vision and pattern recognition*. Seattle, WA, USA: IEEE, 2020, pp. 10 327–10 336, DOI: 10.1109/CVPR42600.2020.01034.

[65] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," in *NIPS 2016 Deep Learning Symposium*. Barcelona, Spain: Curran Associates, Inc., 2016.

[66] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *Transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019, DOI: 10.1109/TPAMI.2019.2938758.

[67] S. Liu, D. Huang *et al.*, "Receptive field block net for accurate and fast object detection," in *European conference on computer vision*. Munich, Germany: Springer, 2018, pp. 385–400, DOI: 10.1007/978-3-030-01252-6_24.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conference on computer vision and pattern recognition*. Las Vegas, NV, USA: IEEE, 2016, pp. 770–778, DOI: 10.1109/CVPR.2016.90.

[69] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, vol. 24. Granada, Spain: Curran Associates, Inc., 2011, pp. 109–117.

[70] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *Conference on computer vision and pattern recognition*. Long Beach, CA, USA: IEEE, 2019, pp. 3623–3632, DOI: 10.1109/CVPR.2019.00374.

[71] T. Zhou, J. Li, S. Wang, R. Tao, and J. Shen, "Matnet: Motion-attentive transition network for zero-shot video object segmentation," *Transactions on image processing*, vol. 29, pp. 8326–8338, 2020, DOI: 10.1109/TIP.2020.3013162.

[72] R. Liu, Z. Wu, S. Yu, and S. Lin, "The emergence of objectness: Learning zero-shot segmentation from videos," in *Advances in neural information processing systems*. [Online]: Curran Associates, Inc., 2021.

[73] M. Zhang, J. Liu, Y. Wang, Y. Piao, S. Yao, W. Ji, J. Li, H. Lu, and Z. Luo, "Dynamic context-sensitive filtering network for video salient object detection," in *International conference on computer vision*. [Online]: IEEE, 2021, pp. 1553–1563, DOI: 10.1109/ICCV48922.2021.00158.

[74] G.-P. Ji, K. Fu, Z. Wu, D.-P. Fan, J. Shen, and L. Shao, "Full-duplex strategy for video object segmentation," in *International conference on computer vision*. [Online]: IEEE, 2021, pp. 4922–4933, DOI: 10.1109/ICCV48922.2021.00488.

[75] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Conference on computer vision and pattern recognition*. Miami, FL, USA: IEEE, 2009, pp. 1597–1604, DOI: 10.1109/CVPR.2009.5206596.

[76] D.-P. Fan, G.-P. Ji, X. Qin, and M.-M. Cheng, "Cognitive vision inspired object segmentation metric and loss function," *SCIENTIA SINICA Informationis*, vol. 6, p. 6, 2021, DOI: 10.1155/2017/4037190.

[77] M.-M. Cheng and D.-P. Fan, "Structure-measure: A new way to evaluate foreground maps," *International journal of computer vision*, vol. 129, no. 9, pp. 2622–2638, 2021, DOI: 10.1007/s11263-021-01490-8.

[78] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Conference on computer vision and pattern recognition*. Columbus, OH, USA: IEEE, 2014, pp. 248–255, DOI: 10.1109/CVPR.2014.39.

[79] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *Transactions on image processing*, vol. 24, no. 12, pp. 5706–5722, 2015, DOI: 10.1109/TIP.2015.2487833.

[80] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *International conference on computer vision*. Venice, Italy: IEEE, 2017, pp. 4548–4557, DOI: 10.1109/ICCV.2017.487.

[81] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *International Joint Conference on Artificial Intelligence*. Stockholm, Sweden: IJCAI, 2018, pp. 698–704, DOI: 10.24963/ijcai.2018/97.

[82] 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," *Machine Intelligence Research*, 2022.

[83] K. Zou, X. Yuan, X. Shen, M. Wang, and H. Fu, "Tbrats: Trusted brain tumor segmentation," [Online], 2022, available: https://arxiv.org/abs/2206.09309.