中国科学：信息科学

SCIENTIA SINICA Informationis

论文

Polyp-PVT：基于金字塔视觉特征变换器的息肉分割技术

董波[1]，王文海[2]，李金鹏[3]，范登平[†*]

†. 南开大学，天津 300350，中国

1. 浙江大学，杭州 310027，中国

2. 南京大学，南京 210093，中国

3. 起源人工智能研究院，阿布扎比 000000，阿联酋

* 通信作者. E-mail: dengpfan@gmail.com

摘要 大多数息肉分割技术使用 U 形结构的 CNN 作为其骨干网络，因此在编码器和解码器之间交换信息时会需要面对两个关键问题：1）考虑不同层级特征之间的贡献差异；2）设计一种有效的机制来融合这些特征. 与现有的基于 CNN 的方法不同，本文采用了 Transformer 作为编码器，它可以学习更强大和鲁棒的特征表示. 此外，考虑到息肉图像采集的影响以及息肉高度伪装的特性，本文引入了三个高效的模块，级联融合模块（CFM）、伪装识别模块（CIM）和相似性聚合模块（SAM）. 其中，CFM 用于从高层特征中收集息肉的语义和位置信息，而 CIM 用于捕获隐藏在低层特征中的息肉信息. 在 SAM 的帮助下，本文将具有高级语义位置信息的息肉区域的像素特征扩展到整个息肉区域，从而有效地融合跨层特征. 本文设计的模型称为 Polyp-PVT，它有效地抑制了特征中的噪声并显着提高了模型的表达能力. 通过在五个广泛使用的数据集上进行的大量实验，结果表明，与现有方法相比本文所提出的模型对各种具有挑战性的场景（如：外观变化、小物体）更加鲁棒，并取得了当前最先进的性能. 模型开源代码：https://github.com/DengPingFan/Polyp-PVT.

关键词 息肉分割，金字塔视觉特征变换器，结肠镜检查

1 引言

结肠镜检查可以及时发现并切除结直肠息肉，从而防止疾病进一步扩散，是检测结直肠病变的黄金标准. 息肉分割作为医学图像分析中的一项基础任务，旨在结肠镜检查中准确定位息肉，这对直肠癌的临床预防具有重要意义. 传统的息肉分割方法主要依赖低层特征，如：纹理 [20]、几何特征 [48]、线性迭代聚类超像素 [47] 等. 然而，这些方法往往会产生低质量的分割结果，并且泛化能力较差. 随着深度学习在医学图像分析的发展，息肉分割取得了可喜的进展. 特别是 U-Net [57] 因

引用格式： 董波，王文海，李金鹏，范登平. 基于金字塔视觉特征变换器的息肉分割技术. 中国科学：信息科学, 在审文章
Dong B., Wang W., Li J., Fan D.-P.. Polyp Segmentation with Pyramid Vision Transformers (in Chinese). Sci Sin Inform, for review

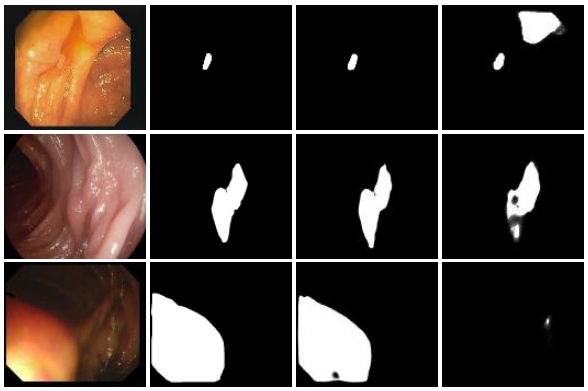


图 1　本文模型和 SANet [76] 方法在具有不同种挑战的样本上的分割结果，例如伪装物（第一行和第二行）和数据采集的影响（第三行）. 上述示例从上到下分别来自 ClinicDB [5]，ETIS [59] 和 ColonDB [63] 数据集. 这些示例结果体现了本文模型的准确性和稳定性.

Figure 1 The segmentation examples of our model and SANet [76] with different challenge cases, *e.g.*, camouflage (1^{st} and 2^{nd} rows) and image acquisition influence (3^{rd} row). The images from top to bottom are from ClinicDB [5], ETIS [59], and ColonDB [63], which show that our model has better generalization ability.

其能够采用多级特征重构高分辨率预测结果而引起了广泛关注. 最近，PraNet [18] 采用两阶段分割方法，先利用并行解码器预测粗糙区域，然后采用注意力机制恢复息肉的边缘和内部结构以进行细粒度分割. ThresholdNet [23] 是一种基于混合流形的置信度引导数据增强方法，用于解决标注数据有限和数据分布不平衡带来的问题.

虽然这些方法与传统方法相比，准确性和泛化能力都有很大提高，但它们仍然难以定位息肉的边界，如图 1所示，原因如下：(1) 图像噪声. 在数据采集过程中，镜头在肠道内旋转以获取不同角度的息肉图像，这会导致运动模糊和反射问题. 因此大大增加了息肉检测的难度；(2) 伪装属性. 息肉的颜色和质地与周围组织非常相似，对比度低，为它们提供了强大的伪装属性 [15,17]，难以识别；(3) 多中心数据. 当前的模型难以泛化到具有不同域/分布的多中心数据.

为了解决上述问题，本文的主要学术贡献归纳为：

- 本文提出了一个新的息肉分割框架，称为 **Polyp-PVT**. 与现有的基于 CNN 的方法不同，本文采用金字塔视觉特征变换器作为编码器，以提取更强大和鲁棒的特征.

- 为了支持本文的框架，本文引入了三个简单的模块. 具体来说，级联融合模块（CFM）通过渐进式融合方式从高级特征中收集息肉的语义和位置信息. 同时，伪装识别模块（CIM）用于捕获伪装在低级特征中的息肉线索，使用注意力机制更多地关注潜在的息肉，从而减少低级特征中的错误信息. 本文进一步使用了配备非局部和图卷积层的相似性聚合模块（SAM），以便从息肉区域挖掘局部像素和全局语义线索.

- 最后，本文对五个（即：Kvasir-SEG [34]、ClinicDB [5]、ColonDB [63]、Endoscene [71] 和 ETIS [59]）具有挑战性的基准数据集进行了广泛的实验，以便评估本文 Polyp-PVT 的性能. 在 ColonDB 上，本文的方法实现了 0.808 的平均 Dice（mDic），比现有的最先进方法 SANet [76] 高 5.5%. 在 ETIS 上，本文模型取得了了 0.787 的 mDic，比 SANet [76] 高 3.7%.

2 相关研究现状

2.1 息肉分割

传统方法. 计算机辅助检测是人工检测的有效替代方案，文献 [56] 针对无线胶囊内镜成像中溃疡、息肉、肿瘤的检测进行了详细调研. 早期的息肉分割解决方案主要基于低级特征，例如纹理 [48]、几何特征 [48] 或者线性迭代聚类超像素 [47]. 然而，由于息肉与周围组织之间的高度相似性，这些方法具有很高的漏检或误检风险.

深度学习方法. 深度学习技术 [26, 41, 43, 60, 73] 进一步加速了息肉分割任务的发展. Akbari 等人 [2] 提出了一种使用全卷积神经网络的息肉分割模型，其分割结果明显优于传统解决方案. Brandao 等人 [9] 使用阴影策略中的形状来恢复深度，将结果合并到 RGB 模型中以提供更丰富的特征表示. 最近，基于编码器-解码器的模型，如 U-Net [57]、UNet++ [92] 以及 ResUNet++ [35]，以优异的性能逐渐占据了该领域的主导地位. Sun 等人 [62] 引入了扩张卷积来提取和聚合具有分辨率保留的高级语义特征，以改进编码器网络. Psi-Net [51] 引入了多任务分割模型，结合轮廓预测和距离图估计来辅助分割掩码预测. Hemin 等人 [54] 基于 Mask R-CNN [25] 首先尝试使用更深的特征提取器进行息肉分割.

与基于 U-Net [3, 57, 92] 的方法不同，PraNet [18] 使用反向注意模块来挖掘具有全局特征图的边界信息，该全局特征图由来自高层特征的并行部分解码器生成. Polyp-Net [4] 提出了一种带有局部梯度加权嵌入水平集的双树小波池 CNN，有效地避免了高信号区域的错误信息，从而显着降低了误报率. Rahim 等人 [55] 提出对相同的隐藏层使用不同的卷积核，通过 MISH 和校正线性单元激活函数进行更深的特征提取，以实现深度特征传播和平滑的非单调性. 此外，他们采用联合广义交点，克服了尺度不变性、旋转和形状差异. Jha 等人 [31] 设计了一种名为 ColonSNet 的实时息肉分割方法. Ahmed 等人 [1] 首次将生成对抗网络应用于息肉分割领域. Thambawita 等人 [66] 本文的另一个有趣的想法是将基于金字塔的特征增强引入息肉分割任务. 此外，Tomar 等人 [69] 设计了一个基于 ResUNet++ 的双解码器注意力息肉分割网络. 最近，MSEG [29] 改进了 PraNet 并提出了一种简单的编码器-解码器结构. 具体来说，他们使用 Hardnet [10] 替代了原来的 Res2Net50 骨干网络，去掉了 attention 机制，实现了更快更准确的息肉分割. 作为早期的尝试，Transfuse [90] 是第一个采用将 CNN 和 Transformer 以并行方式结合的双分支架构. DCRNet [87] 使用外部和内部上下文关系模块来分别估计相同和不同图像中每个位置与所有其他位置之间的相似度. MSNet [81] 引入了多尺度减法网络，以消除多尺度特征之间的冗余和互补信息. 对息肉分割进行全面审查超出了本文的范围. 在表 1中，本文对与相关的代表性作品进行了简要总结.

2.2 视觉 Transformer

Transformers 使用多头自注意力（MHSA）层来模拟长期依赖关系. 与卷积层不同，MHSA 层具有动态权重和全局感受野，使其更加灵活和有效. Transformer [70] 最早由 Vaswani 提出，用于机器翻译任务，此后在自然语言处理领域产生了广泛的影响. 为了将 Transformer 应用于计算机视觉任务，Dosovitskiy 等人 [13] 提出了视觉特征的 Transformer（ViT），这是第一个用于图像分类的视觉 Transformer 模型. ViT 将一张图像分成多个 patch，转换为 token 后依次送到 Transformer 编码

表 1　息肉分割方法调研. CL = ClinicDB [5], EL = ETIS-Larib [59], C6 = CVC-612 [5], AM = ASU-Mayo [64, 91], ES = EndoScene [71], DB = ColonDB [63], ED = Endotect 2020, KS = Kvasir-SEG [34], KCS = Kvasir Capsule-SEG [61], PraNetD = 和 PraNet [18] 使用相同的数据集，IS = 图像分割，VS = 视频分割，CF = 分类，OD = 对象检测.

Table 1 A survey on polyp segmentation. CL = ClinicDB [5], EL = ETIS-Larib [59], C6 = CVC-612 [5], AM = ASU-Mayo [64,91], ES = EndoScene [71], DB = ColonDB [63], ED = Endotect 2020, KS = Kvasir-SEG [34], KCS = Kvasir Capsule-SEG [61], PraNetD = same datasets used in PraNet [18], IS = image segmentation, VS = video segmentation, CF = classfication, OD = object detection.

No.	Model	Publication	Year	Code	Type	Dataset	Core Components
1	CSCPD [20]	IJPRAI	2014	N/A	IS	Private data	Adaptive-scale candidate
2	APD [48]	TMI	2014	N/A	IS	Private data	Geometrical analysis, binary classifier
3	SBCP [47]	SPMB	2017	N/A	IS	Private data	Superpixel
4	FCN [2]	EMBC	2018	N/A	IS	DB	FCN and patch selection
5	D-FCN [9]	JMRR	2018	N/A	IS	CL, EL, AM, and DB	FCN and Shape-from-Shading
6	UNet++ [92]	DLMIA	2018	PyTorch	IS	AM	Skip pathways and deep supervision
7	Psi-Net [51]	EMBC	2019	PyTorch	IS	Endovis	Shape and boundary aware
8	MR-CNN [54]	ISMICT	2019	N/A	IS	C6, EL, and DB	Deep feature extractors
9	UDC [62]	ICMLA	2019	N/A	IS	C6 and EL	Dilation convolution
10	ThresholdNet [23]	TMI	2020	PyTorch	IS	ES and WCE	Learn to threshold
11	MI2GAN [84]	MICCAI	2020	N/A	IS	C6 and EL	Confidence-guided manifold mixup GAN based model
12	ACSNet [89]	MICCAI	2020	PyTorch	IS	ES and KS	Adaptive context selection
13	PraNet [18]	MICCAI	2020	PyTorch	IS	PraNetD	Parallel partial decoder attention
14	Auto-Polyp [1]	MediaEval	2020	N/A	IS	KS	Image-to-image translation
15	APS [68]	MediaEval	2020	N/A	IS	KS	Variants of U-shaped structure
16	PFA [66]	MediaEval	2020	PyTorch	IS	KS	Pyramid focus augmentation
17	MMT [32]	MediaEval	2020	N/A	IS	KS	Competition introduction
18	U-Net-ResNet50 [3]	MediaEval	2020	N/A	IS	KS	Variants of U-shaped structure
19	Survey [56]	CMIG	2021	N/A	CF	Private data	Classification
20	Polyp-Net [4]	TIM	2020	N/A	IS	DB and CL	Multimodel fusion network
21	Polyp-DCNN [55]	BSPC	2021	N/A	OD	EL	Convolutional neural network
22	EU-Net [53]	CRV	2021	PyTorch	IS	PraNetD	Semantic information enhancement
23	DSAS [46]	MIDL	2021	Matlab	IS	KS	Stochastic activation selection
24	U-Net-MobileNetV2 [8]	arXiv	2021	N/A	IS	KS	Variants of U-shaped structure
25	MSEG [29]	arXiv	2021	PyTorch	IS	PraNetD	Hardnet and partial decoder
26	FSSNet [38]	arXiv	2021	N/A	IS	C6 and KS	Meta-learning
27	AG-CUResNeSt [58]	RIVF	2021	N/A	IS	PraNetD	ResNeSt, attention gates
28	MPAPS [86]	JBHI	2021	PyTorch	IS	DB, KS, and EL	Mutual-prototype adaptation network
29	ResUNet++ [33]	JBHI	2021	PyTorch	IS, VS	PraNetD and AM	ResUNet++, CRF and TTA
30	NanoNet [36]	CBMS	2021	PyTorch	IS, VS	ED, KS, and KCS	Real-Time polyp segmentation
31	ColonSegNet [31]	Access	2021	PyTorch	IS	KS	Residual block and SENet
32	Segtran [40]	IJCAI	2021	PyTorch	IS	C6 and KS	Transformer
33	DDANet [69]	ICPR	2021	PyTorch	IS	KS	Dual decoder attention network
34	UACANet [39]	ACM MM	2021	PyTorch	IS	PraNetD	Uncertainty augmented Context attention network
35	DivergentNet [67]	ISBI	2021	PyTorch	IS	EndoCV 2021	Combine multiple models
36	DWHieraSeg [82]	MIA	2021	PyTorch	IS	ES	Dynamic-weighting
37	Transfuse [90]	MICCAI	2021	N/A	IS	PraNetD	Transformer and CNN
38	SANet [76]	MICCAI	2021	PyTorch	IS	PraNetD	Shallow attention network
39	PNS-Net [37]	MICCAI	2021	PyTorch	VS	C6, KS, ES, and AM	Progressively normalized Self-attention network
40	DCRNet [87]	ISBI	2022	PyTorch	IS	ES, KS, and PICCOLO	Within-image and cross-image contextual relations

器中，然后用 MLP 进行图像分类. HVT [52] 是基于分层渐进池化的方法来压缩一个 token 的序列长度，减少 ViT 中的冗余和计算次数. 基于池的视觉 Transformer [27] 借鉴了 CNN 的原理，即随着深度的增加，特征图通道的数量增加，空间维度减少. Yuan 等人 [88] 指出，ViT 中简单的 token 结构无法捕获重要的局部特征，例如边缘和线条，这会降低训练效率并导致冗余注意力机制. 因此提出了 T2T ViT 使用逐层 token 到 token 的转换来逐渐合并相邻 token 并建模局部特征，同时减少 token 的长度. TNT [24] 采用了适用于细粒度图像任务的 Transformer，进一步划分原始图像 patch，以更小的单元进行 self-attention 机制计算. 同时，外部和内部转换器用于提取全局和局部特征.

为了适应语义分割等密集预测任务，多种方案 [12,22,42,72,74,79,85] 将 CNN 的金字塔结构引入到 Transformer 主干的设计中. 例如，基于 PVT 的模型 [72,74] 使用具有四个阶段的分层变换器，表明纯变换器主干可以与其 CNN 对应物一样通用，并且在检测和分割任务中表现更好. 在这项工作中，本文设计了一个新的、基于 Transformer 的息肉分割框架，即使在极端情况下也能准确定位息肉的边界.

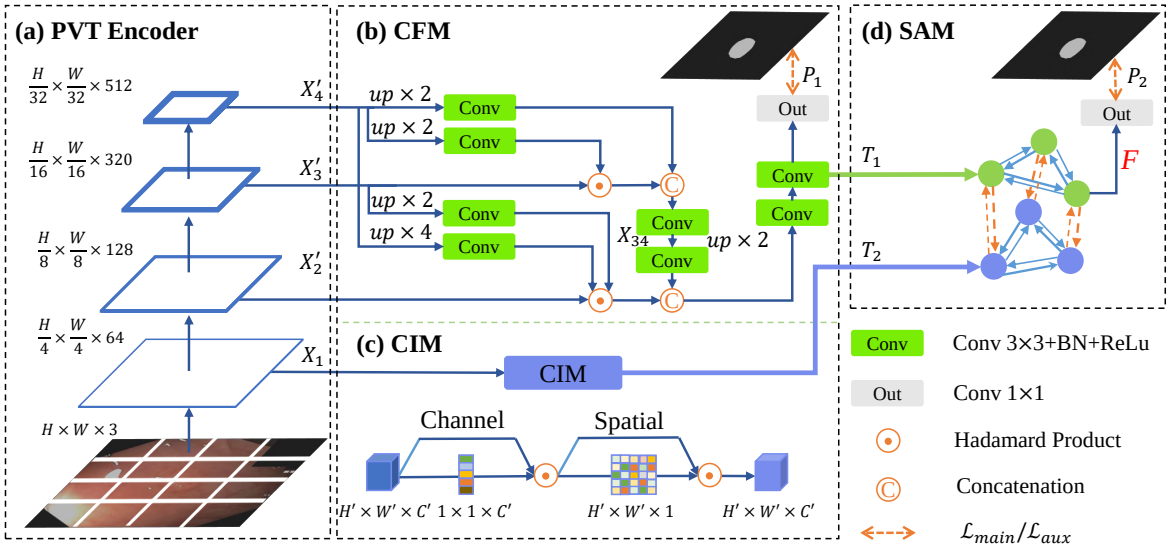


图 2　本文模型框架包括（a）作为编码器网络的金字塔视觉特征变换器（PVT），（b）用于融合高级特征的级联融合模块（CFM），（c）用于过滤低级信息的伪装识别模块（CIM）以及（d）用于整合最终输出的高级和低级特征的相似性聚合模块（SAM）.

Figure 2 Framework of the proposed Polyp-PVT, which consists of a pyramid vision Transformer (PVT) (a) as the encoder network, (b) cascaded fusion module (CFM) for fusing the high-level feature, (c) camouflage identification module (CIM) to filter out the low-level information, and (d) similarity aggregation module (SAM) for integrating the high- and low-level features for the final output.

3　本文的 Polyp-PVT 方法

3.1　整体架构

如图 2所示, 本文的 Polyp-PVT 由四个关键模块组成：即基于视觉特征变换器的编码器、级联融合模块（CFM）、伪装识别模块（CIM）和相似性聚合模块（SAM）. 具体来说，视觉特征变换器用于从输入图像中提取多尺度长距离依赖特征. CFM 用于收集语义线索并以渐进方式聚合高级特征来定位息肉. CIM 旨在去除噪声并增强息肉的低级表示，如：纹理、颜色和边缘. SAM 用于融合 CIM 和 CFM 提供的低级和高级特征，有效地将息肉信息从像素级传递到整个区域.

给定输入图像 $I \in \mathbb{R}^{H\times W\times 3}$，本文使用基于 Transformer 的骨干网络 PVT [74] 提取包含四个不同尺度特征的特征金字塔 $X_i \in \mathbb{R}^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i}$，其中 $C_i \in \{64, 128, 320, 512\}$、$i \in \{1, 2, 3, 4\}$. 然后，本文通过三个卷积单元将三个高级特征 X_2、X_3 和 X_4 的通道调整为 32 并将它们（$X_2^{'}$、$X_3^{'}$ 和 $X_4^{'}$）输入到 CFM 进行融合，得到特征图 $T_1 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 32}$. 同时，低级特征 X_1 通过 CIM 被转换为 $T_2 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 64}$. 之后，$T_1$ 和 T_2 通过 SAM 对齐和融合，产生最终的特征图 $F \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 32}$. 最后，$F$ 被送入 1×1 卷积层以预测息肉分割结果 P_2. 本文使用 P_1 和 P_2 的总和作为最终预测. 在训练期间，本文使用主要损失函数 $\mathcal{L}_{\texttt{main}}$ 和辅助损失函数 $\mathcal{L}_{\texttt{aux}}$ 来优化模型. 其中，主要损失函数 $\mathcal{L}_{\texttt{main}}$ 用于计算最终分割结果 P_2 和真值图（GT）之间的损失，用于优化最终的息肉分割结果. 而辅助损失函数 $\mathcal{L}_{\texttt{aux}}$ 用于监督 CFM 生成的中间结果 P_1.

3.2 基于视觉特征变换器的编码器

由于图像采集不受控的因素，息肉图像往往包含明显的噪声，例如运动模糊和反射. 最近的一些工作 [7,83] 发现视觉转换器 [13,72,74] 表现出比 CNN 更强的性能和更好的抗输入干扰性 [26,60]. 受此启发，本文使用视觉特征变换器 PVT [74] 作为本文的骨干网络来提取更强大的息肉分割特征. 与 [13,42] 使用固定的 “柱状” 结构或移位窗口方式不同，PVT 是一个金字塔结构，它采用空间缩减注意力操作来计算，从而减少计算资源的消耗.

本文所提出的模型是独立于骨干网络的，其他著名的 Transformer 骨干网络在本文的框架中也是可行的. 具体来说，本文采用了 PVTv2 [72]，它是 PVT 的改进版本，具有更强大的特征提取能力. 为了使 PVTv2 适应息肉分割任务，本文移除了最后一个分类层，并将由四个不同阶段产生多尺度特征图（即，X_1、X_2、X_3 和 X_4）用于息肉分割. 在这些特征图中，X_1 提供了息肉的详细外观信息，而 X_2、X_3 和 X_4 提供了高级特征.

3.3 级联融合模块

为了平衡精度和计算资源，本文遵循最近流行的做法 [18,80] 来实现级联融合模块（CFM）. 在介绍这个模块前，本文先将 $\mathcal{F}(\cdot)$ 定义为由窗口大小为 3×3 以及 padding 为 1 的卷积层，批量归一化 [30] 和激活函数 [21] 组成的卷积单元. 如图 2（b）所示，CFM 主要由两个级联部分组成. 第一部分，本文将最高级特征图 $X_4^{'}$ 上采样到与 $X_3^{'}$ 相同的大小，再将结果通过两个卷积单元$\mathcal{F}_1(\cdot)$ 和 $\mathcal{F}_2(\cdot)$ 生成特征图 X_4^1 和 X_4^2. 然后，将 X_4^1 和 $X_3^{'}$ 相乘，并将相乘的结果与 X_4^2 在通道维度拼接起来. 最后，使用卷积单元 $\mathcal{F}_3(\cdot)$ 来平滑拼接后的特征，得到融合的特征图 $X_{34} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times 32}$. 该过程可以概括为公式 1.

$$X_{34} = \mathcal{F}_3(\mathrm{Concat}(\mathcal{F}_1(X_4^{'}) \odot X_3^{'}, \mathcal{F}_2(X_4^{'}))), \tag{1}$$

其中 “$\odot$” 表示 Hadamard 积，$\mathrm{Concat}(\cdot)$ 是沿通道维度的拼接操作.

如公式 2所示，第二部分遵循与第一部分类似的过程. 首先，将特征图 $X_4^{'}$，$X_3^{'}$，X_{34} 上采样到与 $X_2^{'}$ 相同的大小，并使用卷积单元 $\mathcal{F}_4(\cdot)$、$\mathcal{F}_5(\cdot)$ 和 $\mathcal{F}_6(\cdot)$ 对上采样的特征图进行平滑. 然后，将平滑后的特征图 $X_4^{'}$ 和 $X_3^{'}$ 与 $X_2^{'}$ 相乘，并将相乘后的结果与上采样并平滑后的 X_{34} 拼接. 最后，将拼接后的特征图输入两个卷积单元 $\mathcal{F}_7(\cdot)$ 和 $\mathcal{F}_8(\cdot)$ 以减少维度，得到 $T_1 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 32}$，即 CFM 的输出.

$$T_1 = \mathcal{F}_8(\mathcal{F}_7(\mathrm{Concat}(\mathcal{F}_4(X_4^{'}) \odot \mathcal{F}_5(X_3^{'}) \odot X_2^{'}, \mathcal{F}_6(X_{34})))), \tag{2}$$

3.4 伪装识别模块

低级特征通常包含丰富的细节信息，例如纹理、颜色和边缘细节. 然而，息肉在外观上往往与背景非常相似. 因此，本文需要一个强大的提取器来识别息肉的细节.

如图 2 (c) 所示, 本文引入了一个伪装识别模块 (CIM) 来从低级特征图 X_1 的不同维度捕获息肉的细节. 具体来说，CIM 包含一个通道注意力操作 [78] $\mathrm{Att}_c(\cdot)$ 和一个空间注意力操作 [28] $\mathrm{Att}_s(\cdot)$, 计算公式为:

$$T_2 = \mathrm{Att}_s\left(\mathrm{Att}_c\left(X_1\right)\right), \tag{3}$$

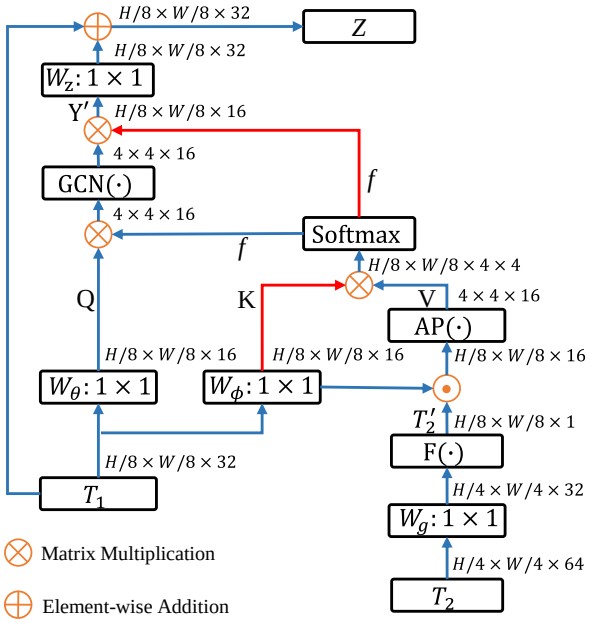


图 3　SAM 模块的详细信息 [65]. 红色箭头表示转置. 该模块由 GCN 和 non-local 组成. 它将具有高级语义的位置线索的息肉区域的像素特征扩展到整个区域.

Figure 3 Details of the SAM. The red arrow indicates a transpose. It is composed of GCN and non-local, which extend the pixel features of polyp regions with high-level semantic location cues to the entire region.

其中，通道注意力操作 $\mathrm{Att}_c(\cdot)$ 可以写成如下：

$$\mathrm{Att}_c(x) = \sigma\left(\mathcal{H}_1\left(P_{\max}\left(x\right)\right) + \mathcal{H}_2\left(P_{\mathrm{avg}}\left(x\right)\right)\right) \odot x, \tag{4}$$

其中，x 是输入张量，$\sigma(\cdot)$ 是 Softmax 函数. $P_{\max}(\cdot)$ 和 $P_{\mathrm{avg}}(\cdot)$ 分别表示自适应最大池化和自适应平均池化函数. $\mathcal{H}_1(\cdot)$ 和 $\mathcal{H}_2(\cdot)$ 共享参数，先通过一个 1×1 的卷积层将通道维度减少 6 倍，然后通过一个 ReLU 层和另一个 1×1 的卷积层来恢复原始通道维度. 另外，空间注意力操作 $\mathrm{Att}_s(\cdot)$ 为：

$$\mathrm{Att}_s(x) = \sigma(\mathcal{G}(\mathrm{Concat}(R_{\max}(x), R_{\mathrm{avg}}(x)))) \odot x, \tag{5}$$

其中，$R_{\max}(\cdot)$ 和 $R_{\mathrm{avg}}(\cdot)$ 分别表示沿通道维度获得的最大值和平均值. $\mathcal{G}(\cdot)$ 是一个窗口大小为 7×7 以及 padding 为 3 卷积层.

3.5　相似度聚合模块

借鉴人脸解析模型中基于边缘的图投影策略 [65], 本文也将 non-local [75] 操作引入到图卷积 [45] 中来探索来自 CIM 的低级局部特征与来自 CFM 的高级语义特征之间的关系，从而有效地实现相似性聚合模块（SAM）功能. 该模块将具有高级语义位置信息的息肉区域从像素特征扩展到整个息肉区域，从而有效地融合跨层特征.

给定包含高级语义信息的特征图 T_1 和具有丰富外观细节的 T_2，利用自注意力机制进行融合. 首先，在 T_1 上应用两个线性映射函数 $W_\theta(\cdot)$ 和 $W_\phi(\cdot)$ 进行降维，得到特征图 $Q \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 16}$ 和

$K \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 16}$. 本文采用窗口大小为 1×1 的卷积层做线性映射. 这个过程可以表示为:

$$Q = W_\theta(T_1), \qquad K = W_\phi(T_1). \tag{6}$$

对于 T_2, 使用卷积单元 $W_g(\cdot)$ 将通道维度减少到 32 并将其插值到与 T_1 相同的大小. 然后, 在通道维度上进行 Softmax 并选择第二个通道[1)]作为注意力图, 生成 $T_2^{'} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 1}$. 这些操作在图 3中表示为 $\mathrm{F}(\cdot)$. 接下来, 计算 K 和 $T_2^{'}$ 之间的 Hadamard 积. 该操作可以为不同的像素分配不同的权重, 以增加边缘像素的权重. 之后, 使用自适应池化操作来减少特征的位移, 并对其应用中心裁剪以获得特征图 $V \in \mathbb{R}^{4\times 4\times 16}$. 该流程可以表述如下:

$$V = \mathrm{AP}(K \odot \mathrm{F}(W_g(T_2))), \tag{7}$$

其中 $\mathrm{AP}(\cdot)$ 表示池化和裁剪操作.

然后, 通过内积建立 V 和 K 中每个像素之间的相关性, 公式如下:

$$f = \sigma(V \otimes K^{\mathsf{T}}), \tag{8}$$

其中 "$\otimes$" 表示内积运算. K^{T} 是 K 的转置, f 是相关注意力图.

得到相关注意力图 f 后, 将其与特征图 Q 相乘并将结果特征送到图卷积层 $\mathrm{GCN}(\cdot)$ 得到 $G \in \mathbb{R}^{4\times 4\times 16}$. 同 [65] 一样, 本文计算 f 和 G 之间的内积为公式 9, 将图卷积特征重构为原始结构特征:

$$Y' = f^{\mathsf{T}} \otimes \mathrm{GCN}(f \otimes Q). \tag{9}$$

重构后的特征图 Y' 通过卷积层 $W_z(\cdot)$ 和 1×1 内核大小调整为与 Y 相同的通道大小, 然后合并 T_1 获得 SAM 的最终输出 $Z \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 32}$. 公式 10总结了这个过程的细节:

$$Z = T_1 + W_z(Y'). \tag{10}$$

3.6 损失函数

本文的损失函数如下:

$$\mathcal{L} = \mathcal{L}_{\mathrm{main}} + \mathcal{L}_{\mathrm{aux}}, \tag{11}$$

其中 $\mathcal{L}_{\mathrm{main}}$ 和 $\mathcal{L}_{\mathrm{aux}}$ 分别是主要损失和辅助损失. 主要损失 $\mathcal{L}_{\mathrm{main}}$ 在最终分割结果 P_2 和真值图 G 之间计算, 可以写为:

$$\mathcal{L}_{\mathrm{main}} = \mathcal{L}^w_{\mathrm{IoU}}(P_2, G) + \mathcal{L}^w_{\mathrm{BCE}}(P_2, G). \tag{12}$$

辅助损失 $\mathcal{L}_{\mathrm{aux}}$ 在 CFM 的中间结果 P_1 和真值图 G 之间计算得出, 可以表示为:

$$\mathcal{L}_{\mathrm{aux}} = \mathcal{L}^w_{\mathrm{IoU}}(P_1, G) + \mathcal{L}^w_{\mathrm{BCE}}(P_1, G). \tag{13}$$

$\mathcal{L}^w_{\mathrm{IoU}}(\cdot)$ 和 $\mathcal{L}^w_{\mathrm{BCE}}(\cdot)$ 是加权交叉联合 (IoU) 损失 [77] 和加权二元交叉熵 (BCE) 损失 [77], 它们在全局结构 (对象级) 和局部细节 (像素级) 两个方面约束预测结果. 与标准的 BCE 损失函数不同, 它不是平等地对待所有像素, $\mathcal{L}^w_{\mathrm{BCE}}(\cdot)$ 考虑每个像素的重要性并为困难像素分配更高的权重. 此外, 与标准 IoU 损失相比, $\mathcal{L}^w_{\mathrm{IoU}}(\cdot)$ 也更关注困难像素.

1) 选择策略与 [65] 相同.

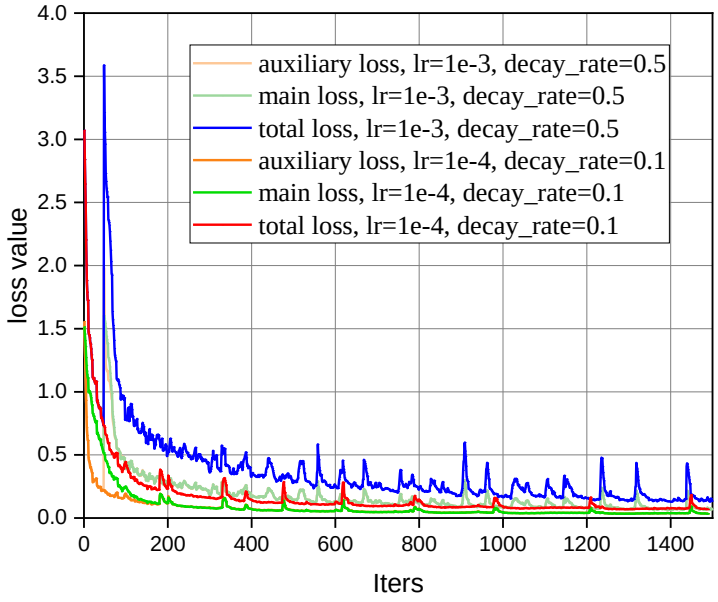


图 4　不同训练参数设置下的损失函数的数值曲线.

Figure 4　Loss curves under different training parameter settings.

表 2　训练阶段的参数设置.

Table 2　Parameter setting during the training stage.

优化器	学习率（lr）	多尺度	Clip
AdamW	1e-4	[0.75,1,1.25]	0.5
衰减率	权重衰减	epoch	输入大小
0.1	1e-4	100	352×352

3.7　技术细节

本文使用 PyTorch 框架实现 Polyp-PVT 模型，并使用 Tesla V100 来加速计算. 考虑到每个息肉图像大小的差异，本文在训练阶段采用多尺度策略 [18,29]. 为了更新网络参数，本文使用了广泛用于 Transformer 网络的 [42,72,74] AdamW 优化器 [44]. 学习率设置为 1e-4，权重衰减也调整为 1e-4. 此外，本文将输入图像的大小调整为 352×352，其中 mini-batch 大小为 16，epoch 设置为 100. 关于训练损失曲线、参数设置和网络参数的更多详细信息分别显示在图 4、表 2和表 3中. 总训练时间接近 3 小时（30 个 epochs）达到最佳性能. 测试阶段，本文只将图像大小调整为 352×352，没有任何后处理优化策略.

4　实验

4.1　评估指标

本文采用了六种广泛使用的评估指标，包括 Dice [50]、IoU、平均绝对误差（MAE）、加权 F-measure（F_β^w）[49]，S-measure（S_α）[11]，和 E-measure（E_ξ）[14,16] 来评估模型性能. 在这些指标中，Dice 和 IoU 是区域层面的相似性度量，主要关注分割对象的内部一致性. 本文计算了 Dice 和

表 3　每个模块的网络参数. 请注意，编码器参数与 PVT 相同，没有任何变化. BasicConv2d 和 Conv2d 的参数为 [in_channel，out_channel，kernel_size，padding]，GCN 的参数为 [num_state，num_node].

Table 3 Network parameters of each module. Note that the encoder parameters are the same as PVT without any changes. BasicConv2d and Conv2d with the parameters [in_channel, out_channel, kernel_size, padding] and GCN [num_state, num_node].

编码器		**SAM**	
patch_size	[4]	AvgPool2d	[6]
embed_dims	[64, 128, 320, 512]	Conv2d	[32,16,1,1]
num_heads	[1, 2, 5, 8]	Conv2d	[32,16,1,1]
mlp_ratios	[8, 8, 4, 4]	Conv2d	[16,32,1,1]
depths	[3, 4, 18, 3]	GCN	[16,16]
sr_ratios	[8, 4, 2, 1]	BasicConv2d	[64,32,1,0]
drop_rate	[0]		
drop_path_rate	[0.1]		
CFM		**CIM**	
BasicConv2d	[32,32,3,1]	AvgPool2d	[1]
BasicConv2d	[32,32,3,1]	AvgPool2d	[1]
BasicConv2d	[32,32,3,1]	Conv2d	[64,4,1,0]
BasicConv2d	[32,32,3,1]	ReLU	
BasicConv2d	[64,64,3,1]	Conv2d	[4,64,1,0]
BasicConv2d	[64,64,3,1]	Sigmoid	
BasicConv2d	[96,96,3,1]	Conv2d	[2,1,7,3]
BasicConv2d	[96,32,3,1]	Sigmoid	

表 4　测试数据集的定量结果，即 Kvasir-SEG 和 ClinicDB 数据集. 最好的结果用粗体表示.

Table 4 Quantitative results of the test datasets, *i.e.*, Kvasir-SEG and ClinicDB. The best results are in **boldface**

	Kvasir-SEG [34]							ClinicDB [5]						
Model	mDic	mIoU	F_β^w	S_α	mE_ξ	$maxE_\xi$	MAE	mDic	mIoU	F_β^w	S_α	mE_ξ	$maxE_\xi$	MAE
MICCAI'15 U-Net	0.818	0.746	0.794	0.858	0.881	0.893	0.055	0.823	0.755	0.811	0.889	0.913	0.954	0.019
DLMIA'18 UNet++	0.821	0.743	0.808	0.862	0.886	0.909	0.048	0.794	0.729	0.785	0.873	0.891	0.931	0.022
MICCAI'19 SFA	0.723	0.611	0.670	0.782	0.834	0.849	0.075	0.700	0.607	0.647	0.793	0.840	0.885	0.042
arXiv'21 MSEG	0.897	0.839	0.885	0.912	0.942	0.948	0.028	0.909	0.864	0.907	0.938	0.961	0.969	0.007
arXiv'21 DCRNet	0.886	0.825	0.868	0.911	0.933	0.941	0.035	0.896	0.844	0.890	0.933	0.964	0.978	0.010
MICCAI'20 ACSNet	0.898	0.838	0.882	0.920	0.941	0.952	0.032	0.882	0.826	0.873	0.927	0.947	0.959	0.011
MICCAI'20 PraNet	0.898	0.840	0.885	0.915	0.944	0.948	0.030	0.899	0.849	0.896	0.936	0.963	0.979	0.009
CRV'21 EU-Net	0.908	0.854	0.893	0.917	0.951	0.954	0.028	0.902	0.846	0.891	0.936	0.959	0.965	0.011
MICCAI'21 SANet	0.904	0.847	0.892	0.915	0.949	0.953	0.028	0.916	0.859	0.909	0.939	0.971	0.976	0.012
Polyp-PVT (Ours)	**0.917**	**0.864**	**0.911**	**0.925**	**0.956**	**0.962**	**0.023**	**0.937**	**0.889**	**0.936**	**0.949**	**0.985**	**0.989**	**0.006**

IoU 的平均值，分别表示为 mDic 和 mIoU. MAE 是逐像素比较指标，表示预测值与真实值之间的绝对误差的平均值. 加权 F-measure（F_β^w）综合考虑了查全率和查准率，消除了常规指标中对每个像素同等对待的影响. S-measure（S_α）侧重于区域和对象级别的目标前景的结构相似性. E-measure（E_ξ）用于评估像素和图像级别的分割结果. 本文报告 E-measure 的平均值和最大值，分别表示为 mE_ξ 和 $maxE_\xi$. 评估代码来源于 https://github.com/DengPingFan/PraNet.

4.2　数据集和对比模型

数据集. 和 PraNet [18] 一样, 本文也采用五个具有挑战性的公共数据集, 包括 Kvasir-SEG [34]、ClinicDB [5]、ColonDB [63]、Endoscene [71] 和 ETIS [59] 来验证本文框架的有效性.

模型. 本文收集了息肉分割领域的九开源模型进行比较，分别为 U-Net [57]、PraNet [18]、SFA [19]、UNet++ [92]、MSEG [29]、ACSNet [89]、DCRNet [87]、EU-Net [53] 和 SANet [76]. 为了公平比较，本文使用他们的开源代码在相同的训练和测试集上进行评估. 唯独 SFA 结果是使用公开的测试模型生成的.

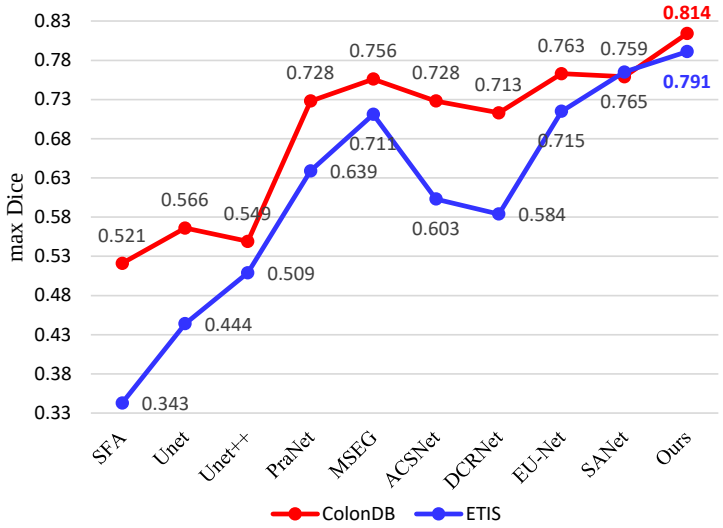


图 5　模型的泛化能力的评价. 在 ColonDB 和 ETIS 数据集上，本文方法取得了最高的 Dice 分数.

Figure 5　Evaluation of model generalization ability. We provide the max Dice results on ColonDB and ETIS.

表 5　测试数据集 ColonDB 和 ETIS 的定量结果. SFA 结果用公布的测试代码生成.

Table 5　Quantitative results of the test datasets ColonDB and ETIS. The SFA result is generated using the published code.

	ColonDB [63]							ETIS [59]						
Model	mDic	mIoU	F_β^w	S_α	mE_ξ	$maxE_\xi$	MAE	mDic	mIoU	F_β^w	S_α	mE_ξ	$maxE_\xi$	MAE
MICCAI'15 U-Net	0.512	0.444	0.498	0.712	0.696	0.776	0.061	0.398	0.335	0.366	0.684	0.643	0.740	0.036
DLMIA'18 UNet++	0.483	0.410	0.467	0.691	0.680	0.760	0.064	0.401	0.344	0.390	0.683	0.629	0.776	0.035
MICCAI'19 SFA	0.469	0.347	0.379	0.634	0.675	0.764	0.094	0.297	0.217	0.231	0.557	0.531	0.632	0.109
MICCAI'20 ACSNet	0.716	0.649	0.697	0.829	0.839	0.851	0.039	0.578	0.509	0.530	0.754	0.737	0.764	0.059
arXiv'21 MSEG	0.735	0.666	0.724	0.834	0.859	0.875	0.038	0.700	0.630	0.671	0.828	0.854	0.890	0.015
arXiv'21 DCRNet	0.704	0.631	0.684	0.821	0.840	0.848	0.052	0.556	0.496	0.506	0.736	0.742	0.773	0.096
MICCAI'20 PraNet	0.712	0.640	0.699	0.820	0.847	0.872	0.043	0.628	0.567	0.600	0.794	0.808	0.841	0.031
CRV'21 EU-Net	0.756	0.681	0.730	0.831	0.863	0.872	0.045	0.687	0.609	0.636	0.793	0.807	0.841	0.067
MICCAI'21 SANet	0.753	0.670	0.726	0.837	0.869	0.878	0.043	0.750	0.654	0.685	0.849	0.881	0.897	0.015
Polyp-PVT (Ours)	**0.808**	**0.727**	**0.795**	**0.865**	**0.913**	**0.919**	**0.031**	**0.787**	**0.706**	**0.750**	**0.871**	**0.906**	**0.910**	**0.013**

4.3　学习能力分析

实验设置. 本文使用 ClinicDB 和 Kvasir-SEG 数据集来评估 Polyp-PVT 的学习能力. ClinicDB 包含 612 张图像，这些图像是从 31 个结肠镜检查视频中提取的. Kvasir-SEG 则从 Kvasir 数据集中的息肉类中收集的，包括 1,000 张息肉图像. 依照 PraNet 中的设置，本文采用来自 ClinicDB 和 Kvasir-SEG 数据集的 900 和 548 幅图像作为训练集，其余 64 和 100 幅图像分别作为测试集.

结果. 从表 4中可以看出，本文模型优于当前的方法，表明它具有更好的学习能力. 在 Kvasir-SEG 数据集上，本文模型的 mDic 分数比次优的模型 SANet 高 1.3%，比 PraNet 高 1.9%. 在 ClinicDB 数据集上，本文模型的 mDic 分数比 SANet 高 2.1%，比 PraNet 高 3.8%.

4.4　泛化能力分析

实验设置. 为了验证模型的泛化性能，本文在三个训练集中不可见的（即多中心）数据集上进行了测试，即：ETIS、ColonDB 和 EndoScene. ETIS 有 196 张图像，ColonDB 有 380 张图像，

表 6 测试数据集 Endoscene 的定量结果. SFA 结果用公开的测试代码生成.

Table 6 Quantitative results of the test dataset Endoscene. The SFA result is generated using the published code.

	Endoscene [71]						
Model	mDic	mIoU	F_β^w	S_α	mE_ξ	$maxE_\xi$	MAE
MICCAI'15 U-Net	0.710	0.627	0.684	0.843	0.847	0.875	0.022
DLMIA'18 UNet++	0.707	0.624	0.687	0.839	0.834	0.898	0.018
MICCAI'19 SFA	0.467	0.329	0.341	0.640	0.644	0.817	0.065
MICCAI'20 ACSNet	0.863	0.787	0.825	0.923	0.939	0.968	0.013
arXiv'21 MSEG	0.874	0.804	0.852	0.924	0.948	0.957	0.009
arXiv'21 DCRNet	0.856	0.788	0.830	0.921	0.943	0.960	0.010
MICCAI'20 PraNet	0.871	0.797	0.843	0.925	0.950	0.972	0.010
CRV'21 EU-Net	0.837	0.765	0.805	0.904	0.919	0.933	0.015
MICCAI'21 SANet	0.888	0.815	0.859	0.928	0.962	0.972	0.008
Polyp-PVT (Ours)	**0.900**	**0.833**	**0.884**	**0.935**	**0.973**	**0.981**	**0.007**

表 7 本文模型和对比模型平均 Dice（mDic）的标准差（SD）.

Table 7 The standard deviation (SD) of the mean dice (mDic) of our model and the comparison models.

数据集	Kvasir-SEG	ClinicDB	ColonDB	ETIS	Endoscene
指标	mDic ± SD	mDic ± SD	mDic ± SD	mDic ± SD	mDic ± SD
MICCAI'15 U-Net	.818 ± .039	.823 ± .047	.483 ± .034	.398 ± .033	.710 ± .049
DLMIA'18 UNet++	.821 ± .040	.794 ± .044	.456 ± .037	.401 ± .057	.707 ± .053
MICCAI'19 SFA	.723 ± .052	.701 ± .054	.444 ± .037	.297 ± .025	.468 ± .050
arXiv'21 MSEG	.897 ± .041	.910 ± .048	.735 ± .039	.700 ± .039	.874 ± .051
MICCAI'20 ACSNet	.898 ± .045	.882 ± .048	.716 ± .040	.578 ± .035	.863 ± .055
arXiv'21 DCRNet	.886 ± .043	.896 ± .049	.704 ± .039	.556 ± .039	.857 ± .052
MICCAI'20 PraNet	.898 ± .041	.899 ± .048	.712 ± .038	.628 ± .036	.871 ± .051
EU-Net	.908 ± .042	.902 ± .048	.756 ± .040	.687 ± .039	.837 ± .049
MICCAI'21 SANet	.904 ± .042	.916 ± .049	.752 ± .040	.750 ± .047	.888 ± .054
Polyp-PVT (Ours)	**.917** ± .042	**.937** ± .050	**.808** ± .043	**.787** ± .044	**.900** ± .052

EndoScene 有 60 张图像. 值得注意的是，这些数据集中的图像属于不同的医疗中心. 换言之，模型没有在训练数据中见到过这些数据，这与 ClinicDB 和 Kvasir-SEG 的验证方法不同.

结果. 结果展示在表 5和表 6中. 可以看出，与现有模型相比，本文的 Polyp-PVT 取得了良好的泛化性能. 在 ColonDB 上，它分别领先次优的 SANet 和经典的 PraNet 模型 5.5% 和 9.6%. 在 ETIS 上，本文分别超过 SANet 和 PraNet 模型 3.7% 和 15.9%. 此外，在 EndoScene 上，本文模型分别比 SANet 和 PraNet 高出 1.2% 和 2.9%. 为了进一步证明 Polyp-PVT 的泛化能力，本文在图 5中展示了最大 Dice 结果，可以发现本文模型的指标在数据集 ColonDB 和 ETIS 上均有稳定提升. 此外，本文在表 7中显示了本文模型与其他模型之间的平均 Dice（mDic）的标准偏差（SD）. 可以看出，本文模型和比较模型在 SD 上没有太大差异，它们都是稳定和平衡的.

CIM 的有效性. 为了展示 CIM 的性能，本文将 CIM 从 Polyp-PVT 中删除，表示为 "Polyp-PVT（w/o CIM)" . 如表 8所示，该变体的性能比完整的 Polyp-PVT 差. 具体来说，去除 CIM 会导致 mDic 减少 1.8%. 显然，缺失 CIM 会引入显着的噪声（请参阅图 8）.

4.5 定性分析

图 6 和图 7展示了本文模型和对比模型的可视化结果. Polyp-PVT 的结果有两个优点. 1）本文模型能适应不同条件下的数据. 即在不同的采集环境下保持稳定的识别和分割能力，如不同的光照、对比度、反射、运动模糊等；2）模型分割结果具有内部一致性，预测边缘更接近真值图.

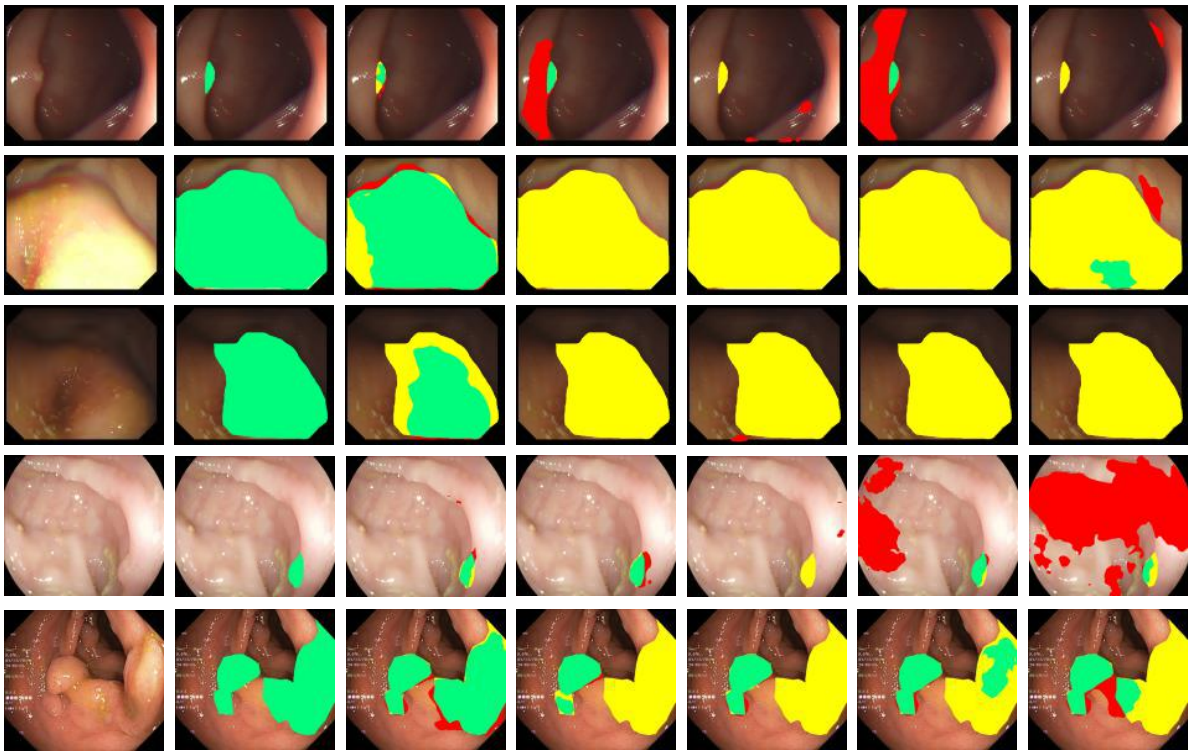


图 6　当前模型的可视化结果. 绿色表示正确的息肉. 黄色是漏检的息肉. 红色是错误的预测. 可以看出本文所提出的模型可以准确地定位和分割不同大小的息肉.

Figure 6 Visualization results with the current models. Green indicates a correct polyp. Yellow is the missed polyp. Red is the wrong prediction. As we can see, the proposed model can accurately locate and segment polyps, regardless of the number of size.

表 8　消融实验的定量结果.

Table 8 Quantitative results for ablation studies.

数据集	指标	Bas.	w/o CFM	w/o CIM	w/o SAM	Final
Endoscene	mDic	0.869	0.892	0.882	0.874	0.900
	mIoU	0.792	0.826	0.808	0.801	0.833
ClinicDB	mDic	0.903	0.915	0.930	0.930	0.937
	mIoU	0.847	0.865	0.881	0.877	0.889
ColonDB	mDic	0.796	0.802	0.805	0.779	0.808
	mIoU	0.707	0.721	0.724	0.696	0.727
ETIS	mDic	0.759	0.771	0.785	0.778	0.787
	mIoU	0.668	0.690	0.711	0.693	0.706
Kvasir-SEG	mDic	0.910	0.922	0.910	0.910	0.917
	mIoU	0.856	0.872	0.858	0.853	0.864

4.6　消融实验

本文详细描述了每个组件对整个模型的有效性. 训练、测试和超参数设置与3.7 节中提到的相同. 结果显示在表 8中.

组成. 本文使用 PVTv2 [72] 作为基线（Bas.），并通过从完整的 Polyp-PVT 中移除或替换组件并将变体与标准版本（“Polyp-PVT = PVT+CFM+CIM+SAM”）进行比较来评估模块的有效性.

CFM 的有效性. 为了分析 CFM 的有效性，本文训练了一个“Polyp-PVT（w/o CFM）”版

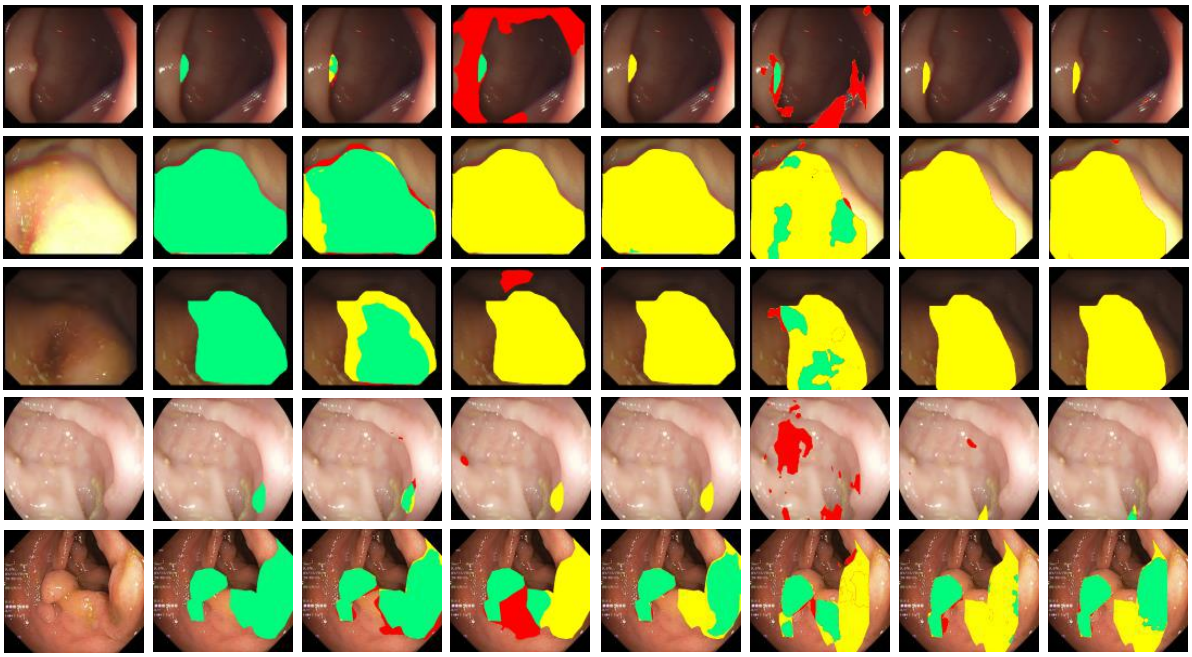


图 7　当前模型的可视化结果.

Figure 7　Visualization results with the current models.

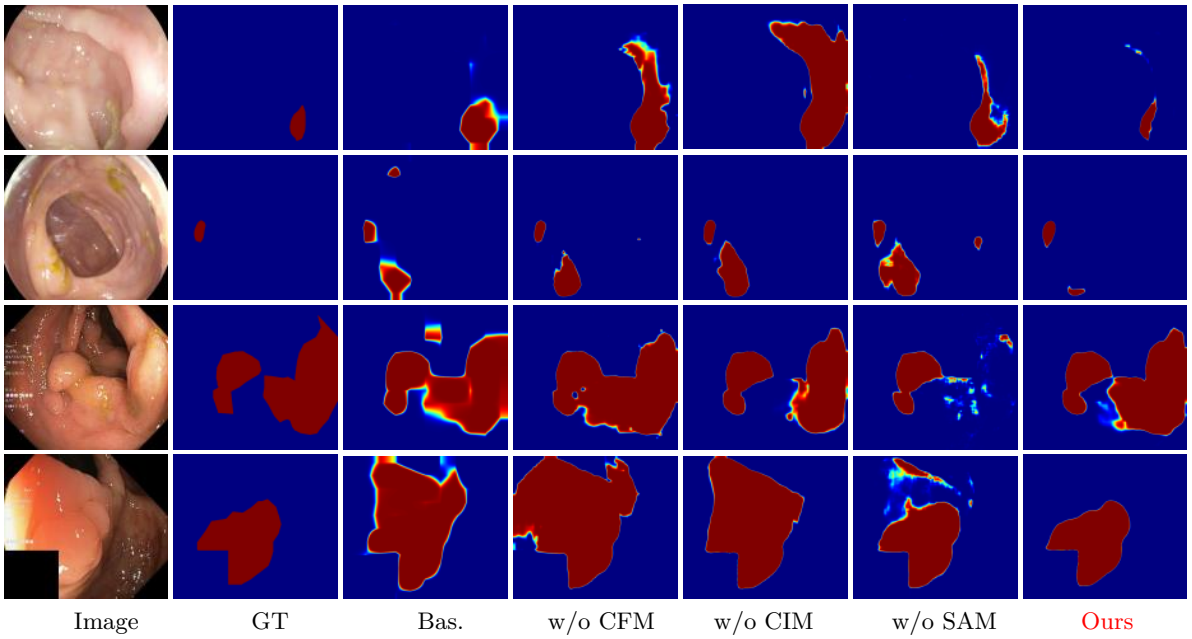


图 8　消融实验结果的可视化. 输出结果被转换为热图. 可见移除任何模块都会导致漏检的或者错误的结果.

Figure 8　Visualization of the ablation study results, which are converted from the output into heat maps. As can be seen, removing any module leads to missed or incorrectly detected results.

本. 表 8表明，与标准 Polyp-PVT 相比，没有 CFM 的模型在所有五个数据集上都急剧下降. 特别是，mDic 在 ClinicDB 上从 0.937 降低到 0.915.

表 9 在视频息肉数据集 CVC-612-T 和 CVC-612-V 上的分割结果，其中最好的结果被加粗表示.

Table 9 The result of video polyp segmentation on the *i.e.*, CVC-612-T and CVC-612-V, where the best results are in **boldface**.

	CVC-612-T [5]							CVC-612-V [5]						
模型	mDic	mIoU	F_β^w	S_α	mE_ξ	$maxE_\xi$	MAE	mDic	mIoU	F_β^w	S_α	mE_ξ	$maxE_\xi$	MAE
MICCAI'15 U-Net	0.711	0.618	0.694	0.810	0.836	0.853	0.058	0.709	0.597	0.680	0.826	0.855	0.872	0.023
TMI'19 UNet++	0.697	0.603	0.688	0.800	0.817	0.865	0.059	0.668	0.557	0.642	0.805	0.830	0.846	0.025
ISM'19 ResUNet++	0.616	0.512	0.604	0.727	0.758	0.760	0.084	0.750	0.646	0.717	0.829	0.877	0.879	0.023
MICCAI'20 ACSNet	0.780	0.697	0.772	0.838	0.864	0.866	0.053	0.801	0.710	0.765	0.847	0.887	0.890	0.054
MICCAI'20 PraNet	0.833	0.767	0.834	0.886	0.904	**0.926**	0.038	0.857	0.793	0.855	0.915	0.936	0.965	0.013
MICCAI'21 PNS-Net	0.837	0.765	0.838	0.903	0.903	0.923	0.038	0.851	0.769	0.836	0.923	0.944	0.962	0.012
Polyp-PVT (Ours)	**0.846**	**0.776**	**0.850**	**0.895**	**0.908**	**0.926**	**0.037**	**0.882**	**0.810**	**0.874**	**0.924**	**0.963**	**0.967**	**0.012**

表 10 在视频息肉数据集 CVC-300-TV 上的分割结果.

Table 10 Video polyp segmentation results on the CVC-300-TV.

	CVC-300-TV [6]						
模型	mDic	mIoU	F_β^w	S_α	mE_ξ	$maxE_\xi$	MAE
MICCAI'15 U-Net	0.631	0.516	0.567	0.793	0.826	0.849	0.027
TMI'19 UNet++	0.638	0.527	0.581	0.796	0.831	0.847	0.024
ISM'19 ResUNet++	0.533	0.410	0.469	0.703	0.718	0.720	0.052
MICCAI'20 ACSNet	0.732	0.627	0.703	0.837	0.871	0.875	0.016
MICCAI'20 PraNet	0.716	0.624	0.700	0.833	0.852	0.904	0.016
MICCAI'21 PNS-Net	0.813	0.710	0.778	0.909	0.921	0.942	0.013
Polyp-PVT (Ours)	**0.880**	**0.802**	**0.869**	**0.915**	**0.961**	**0.965**	**0.011**

SAM 的有效性. 类似地，通过将 SAM 模块从整个 Polyp-PVT 中移除并用元素相加操作替换它来测试 SAM 模块的有效性，这表示为 "Polyp-PVT （w/o SAM）" . 在 ColonDB 上，完整 Polyp-PVT 的性能在 mDic 和 mIoU 方面分别提高了 2.9% 和 3.1%. 图 8更直观地展示了 SAM 的优势, 缺少 SAM 会导致细节错误甚至漏检.

4.7 视频息肉分割

为了验证所提出模型的优越性，本文对视频息肉分割数据集进行了实验. 为了公平比较，本文使用与 PNS-Net 相同的训练数据集重新训练本文模型，并使用相同的测试集 [37] 评估. 本文在三个标准基准（CVC-300-TV [6]，CVC-612-T [5]，和 CVC-612-V [5]）上比较本文模型。并且对比了 6 种经典的方法，包括 U-Net [57]、UNet++ [92]、ResUNet++ [35]、ACSNet [89]、PraNet [18] 和 PNS-Net [37]，结果展示在表 9和表 10. 这里比较方法的所有预测图均由 PNS-Net 提供. 如结果所示，本文的方法非常有竞争力，并且在 CVC-612-V 和 CVC-300-TV 上分别领先现有最好的 PNS-Net 模型 3.1% 和 6.7%mDice 得分.

4.8 局限性

尽管本文的 Polyp-PVT 超越了现有算法，但在某些情况下仍然表现不佳. 本文在图 9中展示了一些失败的案例. 可以看出，一个主要限制是无法检测具有重叠光影的准确息肉边界（第一行）. 本文模型可以识别息肉的位置信息（第一行中的绿色掩膜），但它将边缘的光影部分视为息肉（第一行中的红色掩膜）. 更致命的是，本文观察到反射点在图像中非常突出，导致模型错误地将反射点预测为息肉（第二行和第三行中的红色掩模）. 因此，本文推测模型的预测可能仅基于这些点. 本文认为一种简单的方法是将输入图像转换为灰度图像，可以消除光影的反射和重叠，辅助模型进行判断.

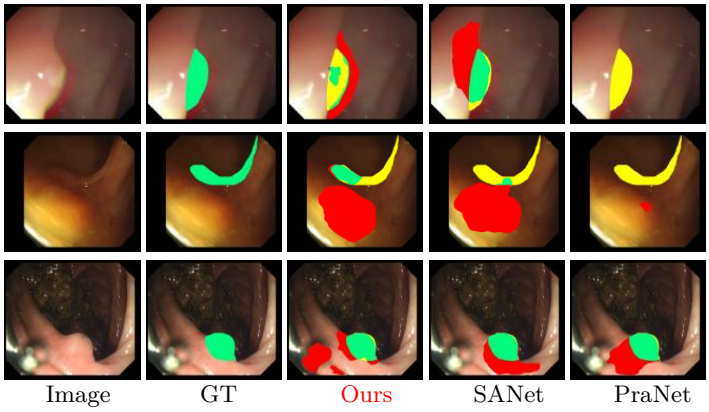


图 9　一些失败示例的可视化. 绿色表示正确的息肉. 黄色是漏检息肉. 红色是错误的预测.

Figure 9 Visualization of failure cases. Green indicates a correct polyp. Yellow is the missed polyp. Red is the wrong prediction.

5　总结

本文提出了一种新的息肉分割框架，名为 **Polyp-PVT**，它利用金字塔视觉特征变换器作为编码器，以显式提取更强大和鲁棒的特征. 大量实验表明 Polyp-PVT 在五个具有挑战性的数据集上始终优于当前的前沿模型，无需任何预处理/后处理. 特别是对于未见过的 ColonDB 数据集，所提出的模型的平均 Dice 得分首次达到 0.8 以上. 有趣的是，本文在视频息肉分割任务方面也超越了当前最先进的 PNS-Net，展示了出色的学习能力. 具体来说，本文通过引入三个简单的组件来获得上述成果，即：级联融合模块（CFM）、伪装识别模块（CIM）和相似性聚合模块（SAM），它们有效地提取了高层次和低层次息肉特征，并有效地融合它们以获得最终输出. 本文希望这项研究能够激发更多新颖的想法来解决息肉分割任务.

参考文献

1 A. M. A. Ahmed. Generative adversarial networks for automatic polyp segmentation. In *MediaEvalW*, 2020.

2 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 *IEEE EMBC*, 2018.

3 S. Alam, N. K. Tomar, A. Thakur, D. Jha, and A. Rauniyar. Automatic polyp segmentation using u-net-resnet50. In *MediaEvalW*, 2020.

4 D. Banik, K. Roy, D. Bhattacharjee, M. Nasipuri, and O. Krejcar. Polyp-net: A multimodel fusion network for polyp segmentation. *IEEE TIM*, 70:1–12, 2020.

5 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. *CMIG*, 43:99–111, 2015.

6 J. Bernal, J. Sánchez, and F. Vilarino. Towards automatic polyp detection with a polyp appearance model. *PR*, 45(9):3166–3182, 2012.

7 S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit. Understanding robustness of transformers for image classification. In *ICCV*, 2021.

8 M. V. Branch and A. S. Carvalho. Polyp segmentation in colonoscopy images using u-net-mobilenetv2. *arXiv preprint arXiv:2103.15715*, 2021.

9 P. Brandao, O. Zisimopoulos, E. Mazomenos, G. Ciuti, J. Bernal, M. Visentini-Scarzanella, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo, et al. Towards a computed-aided diagnosis system in colonoscopy: automatic polyp segmentation using convolution neural networks. *JMRR*, 3(02):1840002, 2018.

10 P. Chao, C.-Y. Kao, Y.-S. Ruan, C.-H. Huang, and Y.-L. Lin. Hardnet: A low memory traffic network. In *CVPR*, 2019.

11 M.-M. Chen and D.-P. Fan. Structure-measure: A new way to evaluate foreground maps. *IJCV*, 129:2622–2638, 2021.

12 X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen. Twins: Revisiting the design of spatial attention in vision transformers. *arXiv preprint arXiv:2104.13840*, 2021.

13 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.

14 D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, 2018.

15 D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao. Concealed object detection. *IEEE TPAMI*, 2021.

16 D.-P. Fan, G.-P. Ji, X. Qin, and M.-M. Cheng. Cognitive vision inspired object segmentation metric and loss function. *SSI*, 2021.

17 D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao. Camouflaged object detection. In *CVPR*, 2020.

18 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.

19 Y. Fang, C. Chen, Y. Yuan, and K.-y. Tong. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In *MICCAI*, 2019.

20 M. Fiori, P. Musé, and G. Sapiro. A complete system for candidate polyps detection in virtual colonoscopy. *IJPRAI*, 28(07):1460014, 2014.

21 X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *AISTATS*, 2011.

22 B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *ICCV*, 2021.

23 X. Guo, C. Yang, Y. Liu, and Y. Yuan. Learn to threshold: Thresholdnet with confidence-guided manifold mixup for polyp segmentation. *IEEE TMI*, 40(4):1134–1146, 2020.

24 K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang. Transformer in transformer. *arXiv preprint arXiv:2103.00112*, 2021.

25 K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017.

26 K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

27 B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh. Rethinking spatial dimensions of vision transformers. In *ICCV*, 2021.

28 J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

29 C.-H. Huang, H.-Y. Wu, and Y.-L. Lin. Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. *arXiv preprint arXiv:2101.07172*, 2021.

30 S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

31 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*, 9:40496–40510, 2021.

32 D. Jha, S. Hicks, K. Emanuelsen, H. D. Johansen, D. Johansen, T. de Lange, M. A. Riegler, and P. Halvorsen. Medico multimedia task at mediaeval 2020: Automatic polyp segmentation. In *MediaEvalW*, 2020.

33 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 resunet++, conditional random field and test-time augmentation. *IEEE JBHI*, 25(6):2029–2040, 2021.

34 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 *MMM*, 2020.

35 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.

36 D. Jha, N. K. Tomar, S. Ali, M. A. Riegler, H. D. Johansen, D. Johansen, T. de Lange, and P. Halvorsen. Nanonet: Real-time polyp segmentation in video capsule endoscopy and colonoscopy. In *IEEE CBMS*, 2021.

37 G.-P. Ji, Y.-C. Chou, D.-P. Fan, G. Chen, D. Jha, H. Fu, and L. Shao. Pns-net: Progressively normalized self-attention network for video polyp segmentation. In *MICCAI*, 2021.

38 R. Khadga, D. Jha, S. Ali, S. Hicks, V. Thambawita, M. A. Riegler, and P. Halvorsen. Few-shot segmentation of medical images based on meta-learning with implicit gradients. *arXiv preprint arXiv:2106.03223*, 2021.

39 T. Kim, H. Lee, and D. Kim. Uacanet: Uncertainty augmented context attention for polyp semgnetaion. In *ACM MM*, 2021.

40 S. Li, X. Sui, X. Luo, X. Xu, L. Yong, and R. S. M. Goh. Medical image segmentation using squeeze-and-expansion transformers. In *IJCAI*, 2021.

41 X. Li, W. Wang, X. Hu, and J. Yang. Selective kernel networks. In *CVPR*, 2019.

42 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 *ICCV*, 2021.

43 J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

44 I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

45 Y. Lu, Y. Chen, D. Zhao, and J. Chen. Graph-fcn for image semantic segmentation. In *ISNN*, 2019.

46 A. Lumini, L. Nanni, and G. Maguolo. Deep ensembles based on stochastic activation selection for polyp segmentation. In *MIDL*, 2021.

47 O. H. Maghsoudi. Superpixel based segmentation and classification of polyps in wireless capsule endoscopy. In *IEEE SPMB*, 2017.

48 A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y.-H. R. Tsai. Automated polyp detection in colon capsule endoscopy. *IEEE TMI*, 33(7):1488–1502, 2014.

49 R. Margolin, L. Zelnik-Manor, and A. Tal. How to evaluate foreground maps? In *CVPR*, 2014.

50 F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016.

51 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 *IEEE EMBC*, 2019.

52 Z. Pan, B. Zhuang, J. Liu, H. He, and J. Cai. Scalable visual transformers with hierarchical pooling. In *ICCV*, 2021.

53 K. Patel, A. M. Bur, and G. Wang. Enhanced u-net: A feature enhancement network for polyp segmentation. In *CRV*, 2021.

54 H. A. Qadir, Y. Shin, J. Solhusvik, J. Bergsland, L. Aabakken, and I. Balasingham. Polyp detection and segmentation using mask r-cnn: Does a deeper feature extractor cnn always perform better? In *ISMICT*, 2019.

55 T. Rahim, S. A. Hassan, and S. Y. Shin. A deep convolutional neural network for the detection of polyps in colonoscopy images. *BSPC*, 68:102654, 2021.

56 T. Rahim, M. A. Usman, and S. Y. Shin. A survey on contemporary computer-aided tumor, polyp, and ulcer detection methods in wireless capsule endoscopy imaging. *CMIG*, page 101767, 2020.

57 O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

58 D. V. Sang, T. Q. Chung, P. N. Lan, D. V. Hang, D. Van Long, and N. T. Thuy. Ag-curesnest: A novel method for colon polyp segmentation. In *IEEE RIVF*, 2021.

59 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. *IJCARS*, 9(2):283–293, 2014.

60 K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

61 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*, 8(1):1–10, 2021.

62 X. Sun, P. Zhang, D. Wang, Y. Cao, and B. Liu. Colorectal polyp segmentation by u-net with dilation convolution. In *IEEE ICMLA*, 2019.

63 N. Tajbakhsh, S. R. Gurudu, and J. Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE TMI*, 35(2):630–644, 2015.

64 N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE TMI*, 35(5):1299–1312, 2016.

65 G. Te, Y. Liu, W. Hu, H. Shi, and T. Mei. Edge-aware graph representation learning and reasoning for face parsing. In *ECCV*, 2020.

66 V. Thambawita, S. Hicks, P. Halvorsen, and M. A. Riegler. Pyramid-focus-augmentation: Medical image segmentation with step-wise focus. In *MediaEvalW*, 2020.

67 V. Thambawita, S. A. Hicks, P. Halvorsen, and M. A. Riegler. Divergentnets: Medical image segmentation by network ensemble. In *ISBI & EndoCV*, 2021.

68 N. K. Tomar. Automatic polyp segmentation using fully convolutional neural network. In *MediaEvalW*, 2020.

69 N. K. Tomar, D. Jha, S. Ali, H. D. Johansen, D. Johansen, M. A. Riegler, and P. Halvorsen. Ddanet: Dual decoder attention network for automatic polyp segmentation. In *ICPRW*, 2021.

70 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

71 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. *JHE*, 2017, 2017.

72 W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pvtv2: Improved baselines with pyramid vision transformer. *arXiv preprint arXiv:2106.13797*, 2021.

73 W. Wang, X. Li, T. Lu, and J. Yang. Mixed link networks. In *IJCAI*, 2018.

74 W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021.

75 X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, 2018.

76 J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, and S. Cui. Shallow attention network for polyp segmentation. In *MICCAI*, 2021.

77 J. Wei, S. Wang, and Q. Huang. F^3net: Fusion, feedback and focus for salient object detection. In *AAAI*, 2020.

78 S. Woo, J. Park, J.-Y. Lee, and I. So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018.

79 H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang. Cvt: Introducing convolutions to vision transformers. In *ICCV*, 2021.

80 Z. Wu, L. Su, and Q. Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, 2019.

81 Z. Xiaoqi, Z. Lihe, and L. Huchuan. Automatic polyp segmentation via multi-scale subtraction network. In *MICCAI*, 2021.

82 G. Xiaoqing, Y. Chen, and Y. Yixuan. Dynamic-weighting hierarchical segmentation network for medical images. *MIA*, page 102196, 2021.

83 E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021.

84 X. Xie, J. Chen, Y. Li, L. Shen, K. Ma, and Y. Zheng. Mi^2gan: Generative adversarial network for medical image domain adaptation using mutual information constraint. In *MICCAI*, 2020.

85 W. Xu, Y. Xu, T. Chang, and Z. Tu. Co-scale conv-attentional image transformers. In *ICCV*, 2021.

86 C. Yang, X. Guo, M. Zhu, B. Ibragimov, and Y. Yuan. Mutual-prototype adaptation for cross-domain polyp segmentation. *IEEE JBHI*, 2021.

87 Z. Yin, K. Liang, Z. Ma, and J. Guo. Duplex contextual relation network for polyp segmentation. 2022.

88 L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *ICCV*, 2021.

89 R. Zhang, G. Li, Z. Li, S. Cui, D. Qian, and Y. Yu. Adaptive context selection for polyp segmentation. In *MICCAI*, 2020.

90 Y. Zhang, H. Liu, and Q. Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *MICCAI*, 2021.

91 Z. Zhou, J. Shin, L. Zhang, S. Gurudu, M. Gotway, and J. Liang. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *CVPR*, 2017.

92 Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation. In *DLMIA*. 2018.

Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers

Bo DONG[1], Wenhai WANG[2], Jinpeng LI[3] & Deng-Ping FAN[†*]

†. *Nankai University, Tianjin* 300350*, China*;
1. *Zhejiang University, Hangzhou* 310027*, China*;
2. *Nanjing University, Nanjing* 210093*, China*;
3. *IIAI, Abu Dhabi* 000000*, UAE*
* Corresponding author. E-mail: dengpfan@gmail.com

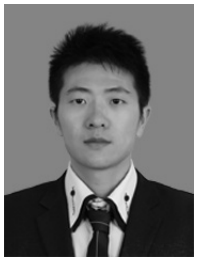
Abstract Most polyp segmentation methods use CNNs with U shape structure as their backbone, facing two key issues when exchanging information between the encoder and decoder: 1) taking into account the differences in contribution between different-level features; and 2) designing an effective mechanism for fusing these features. Different from existing CNN-based methods, we adopt a transformer encoder, which learns more powerful and robust feature representations. In addition, considering the image acquisition influence and elusive properties of polyps, we introduce three efficient modules, including a cascaded fusion module (CFM), a camouflage identification module (CIM), and a similarity aggregation module (SAM). Among these, the CFM is used to collect polyps' semantic and location information from high-level features. At the same time the CIM is applied to capture polyp information disguised in low-level features. With the help of the SAM, we extend the pixel features of the polyp area with high-level semantic position information to the entire polyp area, thereby effectively fusing cross-level features. The proposed model, named Polyp-PVT, effectively suppresses noises in the features and significantly improves their expressive capabilities. Extensive experiments on five widely adopted datasets show that the proposed model is more robust to various challenging situations (*e.g.*, appearance changes, small objects) than existing methods, and achieves the new state-of-the-art performance. Source code: https://github.com/DengPingFan/Polyp-PVT.

Keywords Polyp segmentation, pyramid vision transformer, colonoscopy

Bo DONG was born in 1998. He received the B.Sc. degree from the School of Optical-Electronic Information and Computer Engineering, University of Shanghai for Science and Technology. He is currently a master student from the College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China. His research interest includes deep learning and computer vision.

Jinpeng LI was born in 1992. He received the B.Sc. degree from the Department of Computer Science, Northeastern University and the M.Phil. degree from the Department of Imaging and Interventional Radiology, The Chinese University of Hong Kong. His research interests include pedestrian detection, person search and medical image analysis. He has published about 5 papers in vision conferences such as CVPR, AAAI, ACM MM, *etc.*

Wenhai Wang was born in 1994. received his B.S. degrees from Nanjing University of Science and Technology, China in 2016. He received his PhD degree from the Department of Computer Science, Nanjing University in 2021. His main research interests include scene text detection, deep neural networks exploration, object detection, instance and semantic segmentation. He has published about 20 papers in vision journals and conferences such as T-PAMI, CVPR, ICCV, ECCV, *etc.*



Deng-Ping FAN was born in 1988. received his PhD degree from the Nankai University in 2019. He joined Inception Institute of Artificial Intelligence (IIAI) in 2019. He has published about 25 top journal and conference papers such as TPAMI, TIP, CVPR, ICCV, ECCV, etc. His research interests include computer vision and visual attention, especially on RGB salient object detection (SOD), RGB-D SOD, Video SOD, Co-SOD. He won the Best Paper Finalist Award at IEEE CVPR 2019, the Best Paper Award Nominee at IEEE CVPR 2020.