# WeakPolyp: You Only Look Bounding Box for Polyp Segmentation

**摘要**

受限于昂贵的像素级标签，息肉分割模型受到数据短缺和泛化能力受损的困扰。相比之下，息肉边界框注释更便宜且更易于访问。因此，为了降低标记成本，该论文建议学习完全基于边界框注释的弱监督息肉分割模型（即 WeakPolyp）。本文是对原论文的一次复现以及一点小的改进。然而，粗边界框包含太多噪声。为了避免干扰，该论文引入了掩模到框（M2B）转换。通过监督预测的外框掩码而不是预测本身，M2B 极大地减轻了粗略标签和精确预测之间的不匹配。但是，M2B 只提供稀疏监督，导致预测不唯一。因此，该论文进一步提出了用于密集监督的尺度一致性（SC）损失。通过在不同尺度上显式地对齐同一图像的预测，SC 损失很大程度上减少了预测的变化。请注意，该论文的 WeakPolyp 是即插即用模型，可以轻松移植到其他有吸引力的骨干网。此外，所提出的模块仅在训练期间使用，不会给推理带来计算成本。大量的实验证明了该方法提出的 WeakPolyp 的有效性，它令人惊讶地实现了与完全监督模型相当的性能，根本不需要掩模注释。

**关键词：**息肉分割；弱监督；结直肠癌

## 1 引言

结直肠癌（CRC）已成为全球健康的主要威胁 [5] [6] [7] [11] [9]。由于大多数结直肠癌起源于结直肠息肉，因此有必要对息肉进行早期筛查 [11] [10] [9] [8]。

之前的方法都受到完全监督，并且需要像素级注释。然而，逐像素标记既耗时又昂贵，这阻碍了实际的临床使用。此外，许多息肉没有明确的边界。像素级标记不可避免地会引入主观噪音。为了解决上述局限性，迫切需要一种通用的息肉分割模型。在该论文中通过仅使用粗边界框注释的弱监督息肉分割模型（名为 WeakPolyp）来实现这一目标。图 1(a) 显示了该论文的 WeakPolyp 模型和完全监督模型之间的差异。与完全监督的相比，WeakPolyp 只需要每个息肉的边界框，从而大大降低了标记成本。更有意义的是，WeakPolyp 可以利用现有的大规模息肉检测数据集来辅助息肉分割任务。最后，WeakPolyp 不需要对息肉边界进行标记，避免了源头的主观噪声。所有这些优点使得 WeakPolyp 在临床上更加实用。然而，边界框注释比像素级注释粗糙得多，无法描述息肉的形状。简单地采用这些框注释作为监督会引入过多的背景噪声，从而导致模型不理想。作为一种解决方案，BoxPolyp 仅以高确定性监督像素。

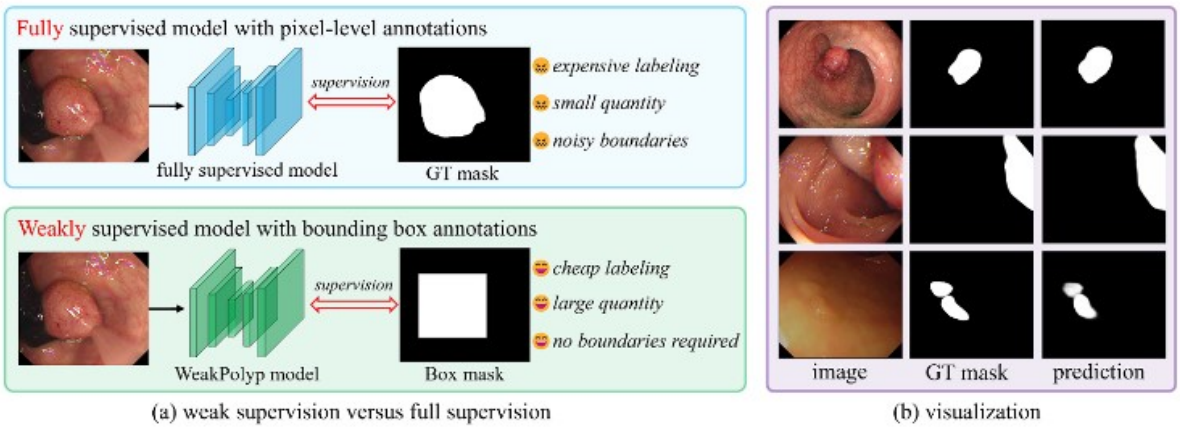


图 1. (a) 完全监督模型与 WeakPolyp 之间的比较。(b) WeakPolyp 预测的可视化。

然而，它需要一个完全监督的模型来预测不确定性图。与 BoxPolyp [12] 不同，该论文的 WeakPolyp 完全遵循弱监督形式，不需要额外的模型或注释。令人惊讶的是，仅通过重新设计监督损失而不对模型结构进行任何更改，WeakPolyp 就实现了与完全监督的对应模型相当的性能。图 1(b) 可视化了 WeakPolyp 的一些预测结果。WeakPolyp 主要由两个新颖的组件实现：mask-to-box (M2B) 转换和尺度一致性 (SC) 损失。在实践中，M2B 用于通过投影和反投影将预测掩模转换为盒状掩模。然后，这个转换后的掩模由边界框注释监督。这种间接监督避免了注释的盒形偏差的误导。然而，预测掩模中的许多区域在投影中丢失，因此得不到监督。为了充分探索这些区域，该论文提出 SC 损失来提供像素级的自我监督，同时根本不需要注释。具体来说，SC 损失显着减少了同一图像在不同尺度下的预测之间的距离。通过强制特征对齐，抑制预测的过度多样性，从而提高模型泛化能力。总之，这篇论文的贡献有三方面：

（1）完全基于边界框注释构建 WeakPolyp 模型，这大大降低了标记成本并实现了与完全监督相当的性能。

（2）提出 M2B 变换来减轻预测和监督之间的不匹配，并设计 SC 损失来提高模型针对预测变化的鲁棒性。

（3）提出的 WeakPolyp 是一个即插即用的选项，它可以提高不同骨干网下息肉分割模型的性能。

## 2 相关工作

已经提出了几种方法来从结肠镜检查图像中定位像素级息肉区域。它们可以分为两大类。

### 2.1 基于 CNN 的方法

Brandao 等人采用带有预训练模型的全卷积网络（FCN）来分割息肉。后来，Akbari 等人引入了改进的 FCN 来提高分割精度。受到 UNet 在生物医学图像分割方面取得巨大成功的启发，UNet++ [1] 和 ResUNet [2] 被用于息肉分割以提高性能。此外，PolypSeg、ACS、ColonSegNet 和 SCR-Net 探索了 UNetenhanced 架构在自适应学习语义上下文方面的有效性。作为新提出的方法，SANNe 和 MSNet 分别设计了浅层注意力模块和减法单元，以实现

精确高效的分割。此外，一些工作选择通过三种主流方式引入额外的约束：施加显式边界监督，引入隐式边界感知表示，以及探索模糊区域的不确定性。

### 2.2 基于 Transformer 的方法

最近，Transformers [3] 因其强大的建模能力而越来越受欢迎。TransFuse 结合了 Transformer 和 CNN，称为并行分支方案，用于捕获全局依赖性和低级空间细节。此外，BiFusion 模块旨在融合两个分支的多级特征。Segtran 提出了一个压缩注意力块来规范自注意力，并且扩展块学习多样化的表示。提出了一种位置编码方案来施加归纳连续性偏差。基于 PVT [4]，Dong 等人引入了一个具有三个紧密组件的模型，即级联融合、伪装识别和相似性聚合模块。

## 3 本文方法

图 2 描述了 WeakPolyp 的组件，包括分割阶段和监督阶段。对于分割阶段，该论文采用 Res2Net 作为主干。对于输入图像 I RH×W,Res2Net 提取四个尺度的特征 fi|i = 1, ..., 4 ，分辨率为 [H/2(i+1),W/2(i+1)]。考虑到计算成本，仅使用 f2、f3 和 f4。为了融合它们，该论文首先应用 1×1 卷积层来统一 f2、f3、f4 的通道，然后使用双线性上采样来统一它们的分辨率。变换为相同大小后，f2、f3、f4 相加并输入到一个 1×1 卷积层中进行最终预测。该论文的贡献主要在于监督阶段，而不是分割阶段，包括掩模到框（M2B）转换和尺度一致性（SC）损失。值得注意的是，M2B 和 SC 都独立于具体的模型结构。

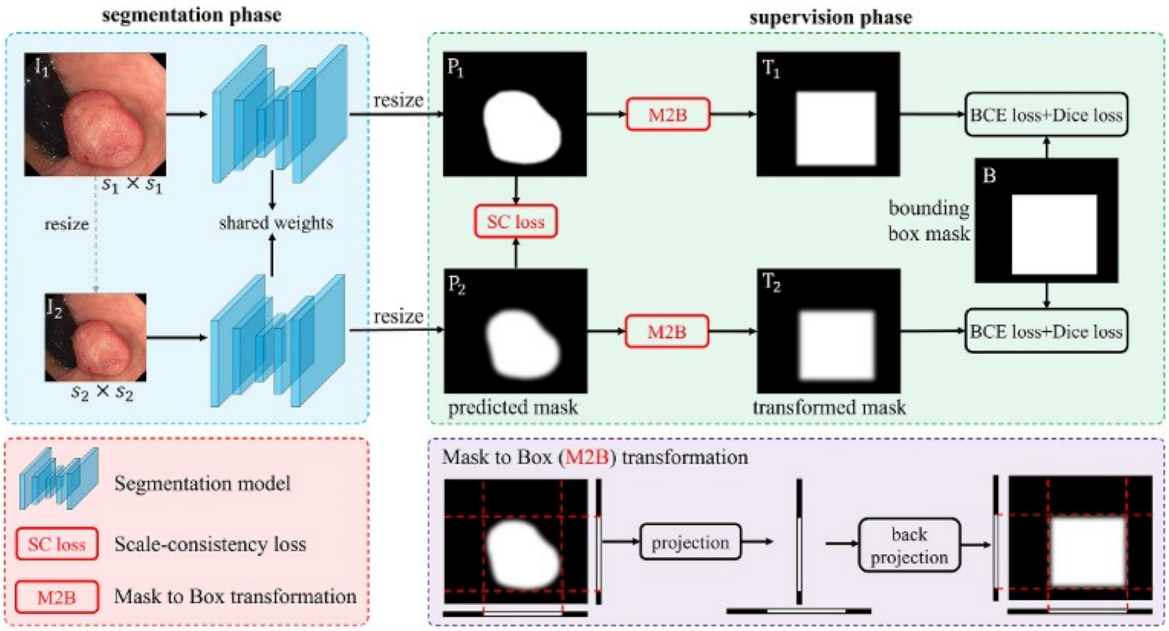


图 2. WeakPolyp 模型的框架

对于每个输入图像 I，该论文首先将其调整为两个不同的尺度：I1 Rs1×s1 和 I2 Rs2×s2 。然后，I1 和 I2 被发送到分割模型并得到两个预测掩模 P1 和 P2，两者都已调整为相同大小。接下来，提出了 SC 损失来减少 P1 和 P2 之间的距离，这有助于抑制预测的变化。最后，为了适合边界框注释 (B)，P1 和 P2 被发送到 M2B 并转换为盒状掩模 T1 和 T2。通过 T1/T2 和 B，该论文计算二元交叉熵（BCE）损失和 Dice 损失，而不用担心噪声干扰。

### 3.1 Mask-to-Box (M2B) Transformation

实现弱监督息肉分割的一种简单方法是使用边界框注释 B 来监督预测掩模 P1/P2。不幸的是，以这种方式训练的模型泛化能力很差。因为 B 中存在很强的盒形偏差。使用这种偏差进行训练，模型被迫预测盒形掩模，无法保持息肉的轮廓。为了解决这个问题，该论文创新性地使用 B 来监督 P1/P2 的边界框掩模（即 T1/T2），而不是 P1/P2 本身。这种间接监督将 P1/P2 与 B 分开，使得 P1/P2 在获得息肉的位置和范围的同时不受 B 的形状偏差的影响。但如何实现从 P1/P2 到 T1/T2 的转变呢？该论文设计了 M2B 模块，它分为两个步骤：投影和反投影，如图 2 所示。投影。如方程式所示 1，给定一个预测掩码 P [0,1]H×W ，该论文将其水平和垂直投影为两个向量 Pw [0, 1]1×W 和 Ph [0, 1]H×1。在此投影中，该论文不使用均值池，而是使用最大池来选取 P 中每行/列的最大值。因为 max pooling 可以完全去除息肉的形状信息。投影后，Pw 和 Ph 中仅存储息肉的位置和范围。

$$P_\omega = \max(P, axis = 0) \in [0,1]^{1*W}, P_h = \max(P, axis = 1) \in [0,1]^{H*1} \quad (1)$$

反投影。基于 Pw 和 Ph，该论文通过反投影构建息肉的边界框掩模。如方程式所示 2，Pw 和 Ph 首先被重复为 P'w 和 P'h，其大小与 P 相同。然后，该论文逐元素取 P'w 和 P'h 的最小值以获得边界框掩模 T 。如图 2 所示，T 不再包含息肉的轮廓。

$$\begin{aligned} P'_\omega &= repeat(P_\omega, H, axis = 0) \in [0,1]^{H*W} \\ P'_h &= repeat(P_h, W, axis = 0) \in [0,1]^{H*W} \\ T &= \min(P'_\omega, P'_h) \in [0,1]^{H*W} \end{aligned} \quad (2)$$

监督。通过 M2B，P1 和 P2 分别变换为 T1 和 T2。由于 T1/T2 和 B 都是盒状掩模，因此该论文直接计算它们之间的监督损失，而不用担心盒状偏差的误导。具体来说，该论文按照 [19,5] 采用 BCE 损失 LBCE 和 Dice 损失 LDice 进行模型监督，如式 3 所示。

$$\mathcal{L}_{Sum} = \frac{\mathcal{L}_{BCE}(T_1, B) + \mathcal{L}_{BCE}(T_2, B)}{2} + \frac{\mathcal{L}_{Dice}(T_1, B) + \mathcal{L}_{Dice}(T_2, B)}{2} \quad (3)$$

优先事项。通过简单的变换，M2B 将噪声监督转变为无噪声监督，从而使预测的掩模能够保留息肉的轮廓。值得注意的是，M2B 是可微分的，可以使用 PyTorch 轻松实现并插入模型中参与梯度反向传播。

### 3.2 Scale Consistency (SC) Loss

在 M2B 中，P 中的大多数像素在投影中被忽略，因此只有少数具有高响应值的像素参与监督损失。这种稀疏的监督可能会导致非唯一的预测。如图 3 所示，经过 M2B 投影，可以将

具有不同响应值的五个预测掩模转换为相同的边界框掩模。因此，该论文考虑引入 SC 损失来实现无注释的密集监督，从而降低了预测的自由度。

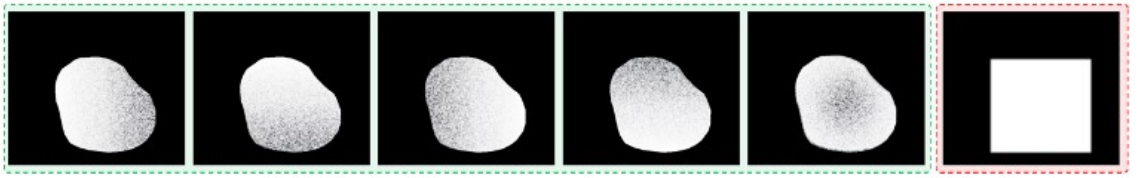


图 3. 不同的预测可能对应于相同的边界框掩模。

方法。如图 2 所示，由于预测的非唯一性以及 I1 和 I2 之间的尺度差异，P1 和 P2 的响应值不同。但 P1 和 P2 来自同一个图像 I1。它们应该是完全相同的。鉴于此，如等式所示 4，该论文明确地构建了密集监督 LSC 减小 P1 和 P2 之间的距离，其中 (i, j) 是像素坐标。请注意，LSC 仅涉及边界框内的像素，以更多地强调息肉区域。尽管 LSC 很简单，但它带来了像素级约束来补偿 LSum 的稀疏性，从而减少了预测的多样性。

$$\mathcal{L}_{SC} = \frac{\sum_{(i,j)\in box} |P_1^{i,j} - P_2^{i,j}|}{\sum_{(i,j)\in box} 1} \tag{4}$$

### 3.3 Total Loss

如方程式所示 5。将 LSum 和 LSC 结合起来，得到 WeakPolyp 模型。请注意，WeakPolyp 只是替换了监督损失，而不对模型结构进行任何更改。因此具有通用性，可以移植到其他模型上。此外，LSum 和 LSC 仅在训练期间使用。在推断中，它们将被删除，因此对模型的速度没有影响。

## 4 复现细节

### 4.1 与已有开源代码对比

本论文提供了源代码，但是存在部分细节及参数没有公开的问题，需要对代码进行调通同时配置参数。因此，我先对论文进行复现工作，下载模型提供的权重及数据集，并对数据集进行论文所述的扩充处理，补充代码细节并配置参数，多次实验选择最优值。此外，在成功复现该论文的基础上，我也对模型进行改进，进行了一个小小的改进。

### 4.2 数据集

SUN-SEG 中的结肠镜检查视频来自昭和大学和名古屋大学数据库(也称为 SUN-database)，这是用于检测任务的最大视频息肉数据集。训练集包含 112 个视频剪辑，总共 19,544 帧，测试集分为四个子测试集，即 SUN-SEG-Seen-Easy (33 个剪辑/4,719 帧)、SUN-SEG-Seen-Hard

(17 个剪辑/3,882 帧)、SUN-SEG-Unseen-Easy （86 个剪辑/12,351 帧）和 SUNSEG-Unseen-Hard（37 个剪辑/8,640 帧)。Easy/Hard 表示样本分割的难度级别，Seen 表示片段来自与训练集相同的视频但不重叠，Unseen 表示片段来自与训练集不重叠的视频。

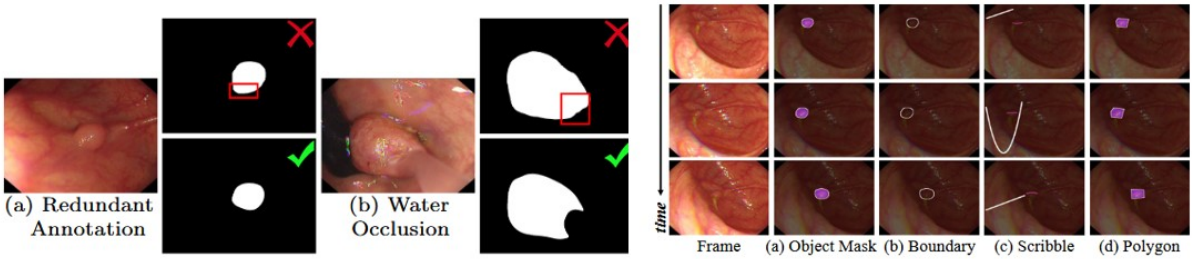


图 4. SUN-SEG 数据集展示

息肉视频分割面临多个挑战。首先，结肠镜检查中摄像机移动而息肉和背景固定，导致光流模式在息肉和正常组织之间不明显，使得基于光流的分割方法无效。其次，结肠镜视频中相邻帧之间的显著变化，源于近距离观察和快速相机轨迹，导致基于全局注意力块的特征聚合方法受到不稳定短期特征的影响。最后，复杂的光照环境和低质量帧要求长时间跨度的建模以捕捉可靠帧，但现有方法通常只关注较短的时间范围，无法有效处理低质量帧的情况。

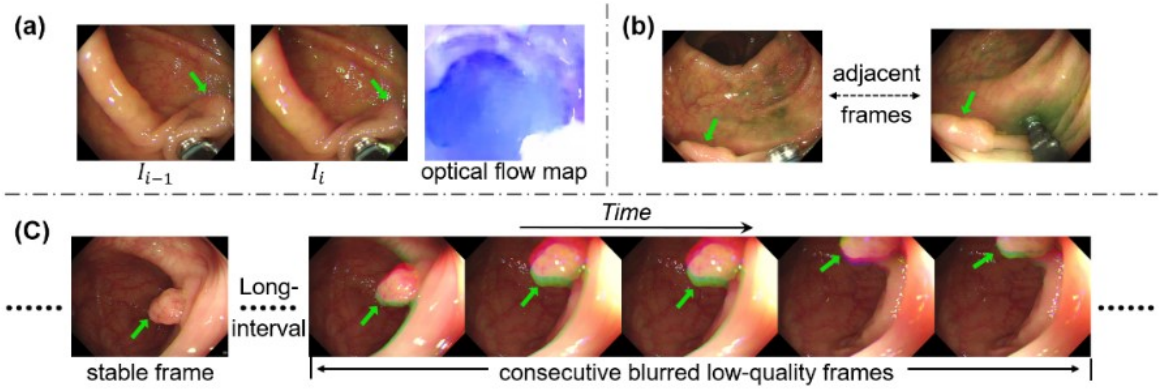


图 5. 息肉视频分割的三大挑战

### 4.3 性能指标

本次实验采用 mIoU 和 Dice 来评估模型的性能指标，旨在量模型在每个类别上的分割准确性以及整体分割性能，提供了对整体性能的综合评估。

IoU 是预测区域与真实标注区域的交集与并集之比, 对于多个类别，mloU 计算如下:

$$mIoU = \frac{1}{N}\sum_{i=1}^{N} IoU_i \qquad (5)$$

其中，N 是类别的总数，IoUi 是第 i 个类别的 IoU。mloU 的取值范围是（0,1)，值越接近 1 表示模型的分割性能越好。

Dice 也称为 F1 分数，是另一种评估分割质量的指标。它基于预测区域和真实区域的交集与它们各自大小的总和来计算。DICE 系数的计算公式为：

$$Dice = \frac{2\times|X\cap Y|}{|X|+|Y|} \tag{6}$$

其中，X 表示预测的分割区域，Y 表示真实的分割区域（ground truth）。|X Y| 是预测区域和真实区域交集的像素数量，|X| 和 |Y| 分别是预测区域和真实区域的像素数量。

### 4.4 创新点

由于本论文采用的方法中，主要是对 Backone 中提取出三个特征进行融合，刚好 Mask2Former 架构的 Decoder 也是三个特征进行融合输入，进行了一个简单的替换。

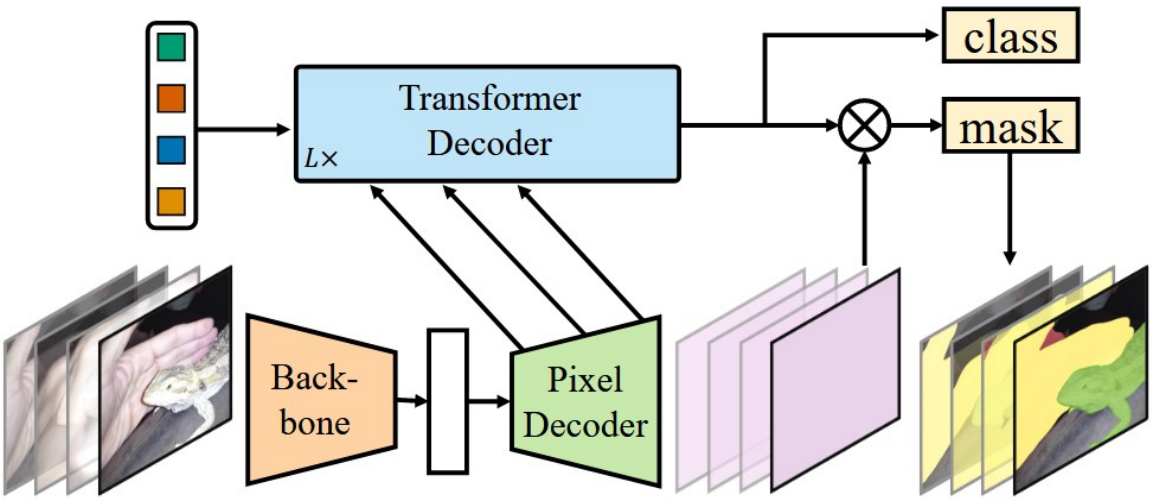


图 6. Mask2Former 架构

## 5 实验结果分析

在 WeakPolyp 的复现过程中，由于实验设备所用设备不同，且设置的 batchsize 与原论文不同，在 SUN-SEG 实验结果显示其精度与原论文存在一定的的差异。结果表如表 1 所示和可视化效果图如图 7 所示。

*表 1 实验结果以及与原论文对比以及改进效果*

| | Model | Easy Testing | | Hard Testing | |
|---|---|---|---|---|---|
| | | Dice | IOU | Dice | IOU |
| *论文* | PVTv2-B2 | 85.3 | 78.1 | 85.4 | 77.7 |
| *复现* | PVTv2-B2 | 86.77 | 80.02 | 85.94 | 78.40 |
| *改进* | PVTv2-B2 | 87.40 | 80.74 | 86.44 | 78.99 |

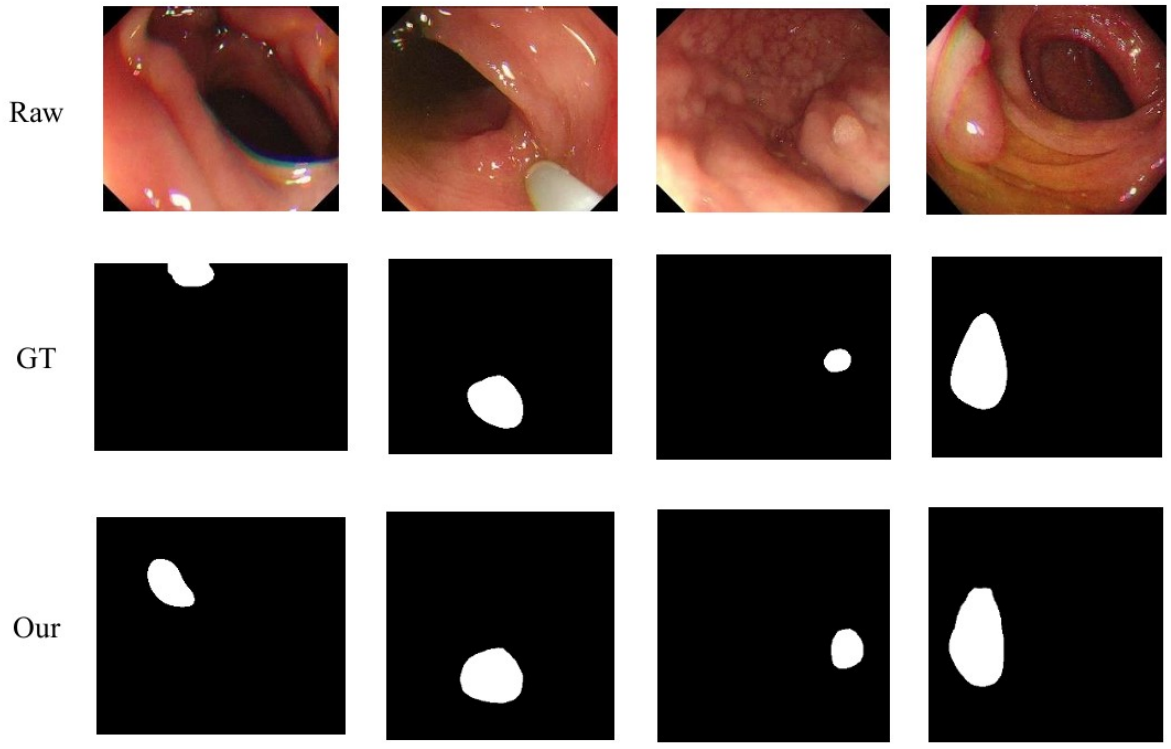


图 7. 复现案例可视化

更换新的解码器后，Dice 和 mIoU 的提升可以归因于其增强的上下文关系捕捉能力，这使得模型能够更准确地识别和定位息肉等小而复杂的结构。新解码器可用了多尺度特征融合技术，同时考虑不同分辨率的图像信息，提高了分割的精确度。此外，级联注意解码器（CASCADE）通过利用分层 Vision Transformer 的多尺度特性，显著提高了评分。新解码器还可能通过改进特征映射的细化，提供了更强的特征表示，有助于模型更好地定位器官或病变的边界。

## 6 总结与展望

受限于昂贵的标记成本，像素级注释不易获得，这阻碍了息肉分割领域的发展。该论文提出了完全基于边界框注释的 WeakPolyp 模型。WeakPolyp 不需要像素级标注，从而避免了主观噪声标签的干扰。更重要的是，WeakPolyp 甚至达到了与完全监督模型相当的性能，显示了弱监督学习在息肉分割领域的巨大潜力。

在本文中，我主要对该论文进行了基本的复现，和进行了一定的改进，未来，将在弱监督息肉分割中引入时间信息，以进一步减少模型对标签的依赖。

## 参考文献


[1] Mengjun Cheng, Zishang Kong, Guoli Song, Yonghong Tian, Yongsheng Liang, and Jie Chen. Learnable oriented-derivative network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 720–730. Springer, 2021.

[2] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. arxiv 2021. *arXiv preprint arXiv:2108.06932*.

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

[4] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019.

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

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

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

[8] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.

[9] Yutian Shen, Xiao Jia, and Max Q-H Meng. Hrenet: A hard region enhancement network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 559–568. Springer, 2021.

[10] Nikhil Kumar Tomar, Debesh Jha, Ulas Bagci, and Sharib Ali. Tganet: Text-guided attention for improved polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 151–160. Springer, 2022.

[11] Jinfeng Wang, Qiming Huang, Feilong Tang, Jia Meng, Jionglong Su, and Sifan Song. Stepwise feature fusion: Local guides global. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 110–120. Springer, 2022.

[12] Jun Wei, Yiwen Hu, Guanbin Li, Shuguang Cui, S Kevin Zhou, and Zhen Li. Boxpolyp: boost generalized polyp segmentation using extra coarse bounding box annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 67–77. Springer, 2022.