息肉视频分割：一种深度学习的角度

Ge-Peng Ji[1], Guohao Xiao[2], Yu-Cheng Chou[3], Deng-Ping Fan[✉4], Kai Zhao[5], Geng Chen[6] and Luc Van Gool[4]

摘要

作者的团队提出了深度学习时代的第一个综合的视频息肉分割 (VPS) 的研究。多年来，由于缺乏具有细粒度分割注释的大规模数据集，VPS 的发展并不能轻松地向前推进。为了解决这个问题，作者的团队首先引入了一个高质量的逐帧标注的名为 SUN-SEG 的 VPS 数据集，这个数据集包含了来自著名的 SUN-database 的 158690 个结肠镜检查视频帧。SUN-SEG 提供了涵盖不同类型的附加注释，包括属性、对象掩码、边界、点画和多边形掩码等。其次，作者的团队提出来一个名为 PNS+ 的息肉视频分割的基线网络，PNS+ 由一个全局编码器、一个局部编码器和标准化的自注意力 (NS) 模块组成。全局和局部编码器接收一个帧和多个连续帧来提取长期和短期的时空特征表示，然后通过两个 NS 模块逐步细化时空特征。大量实验表明，PNS+ 实现了最佳的性能和实时推理速度 (170fps)，使其成为了解决 VPS 任务的强有力的基线网络。具体来说，作者将 PNS+ 和 13 个具有代表性的息肉/对象分割模型进行对比，在多项指标上达到了最佳的效果。

关键词：息肉视频分割；数据集；标准化自注意力机制

1 引言

结直肠癌 (CRC) 作为第二大致命癌症和第三大常见恶性肿瘤，据估计每年会导致数百万例的发病和死亡。CRC 患者的生存率在疾病的第一阶段超过 95%，但在第四和第五阶段急剧下降至低于 35%[1]。因此，通过结肠镜检查和乙状结肠镜检查等筛查技术对阳性 CRC 病例进行早期诊断对于提高存活率至关重要。为了预防 CRC，医生可以切除有风险转变为癌症的结肠息肉。然而，这一过程很大程度上依赖于医生的经验，并存在较高的息肉缺失率[2]。

近年来，人工智能 (AI) 技术被应用于医学结肠镜检查中候选病变息肉的自动检测。然而，由于有限的注释数据这个问题，开发具有令人满意的准确率的人工智能模型仍然具有挑战性。深度学习模型通常渴望得到一个具有密集注释标签的大规模视频数据集。此外，在评估这些方法的实际性能时，还缺少一个由 VPS 社区商定的基准测试。为此，作者的团队提出了一项系统研究，以促进视频息肉分割 (VPS) 深度学习模型的开发。

作者的团队精心制作了一个名为 SUN-SEG 的大规模 VPS 数据集，其中包含从 SUN-database[3]中选择的 158,690 帧，同时提供了多种标签，包括属性、对象掩码、边界、涂鸦和多边形，这些标签可以进一步支持结肠镜检查诊断、定位和衍生任务的发展。

作者的团队同时设计了一个简单但有效的名为 PNS+ 的 VPS 基线网络，该网络由一个局部编码器、一个全局编码器和两个标准化的自注意力 (NS) 组成。全局和局部编码器分别从第一个锚点帧和多个连续帧中提取短期和长期的时空表示。当需要在提取的特征中耦合上下文语义信息时，NS 模块会动态地更新感受野来完成这个任务。大量实验表明，PNS+ 在具有挑战性的 SUN-SEG 数据集上表现出了最佳的性能。

最后，为了全面了解 VPS 任务的发展情况，作者的团队通过评估 13 种尖端息肉/对象分割方法进行了第一个大规模基准测试。基于测试结果 (即 5 个基于图像的方法和 8 个基于视频的方法)，作者的团队认为 VPS 任务并没有很好地完成，并为进一步的探索留下了很大的空间。

2 相关工作

本节主要从图像的息肉分割和视频的息肉分割两个方面来回顾最近在计算机辅助息肉诊断方面的努力。图像的息肉分割指的是单独对一帧图像进行特征处理并得到最后的分割结果，而视频息肉分割则是对一个视频序列 (多帧图像) 进行处理，分割过程中会利用帧与帧之间的时空信息来进行最后的分割预测。

2.1 图像息肉分割 (IPS)

最近几年，基于 CNN 的深度学习方法在图像处理领域取得了巨大的成功。Brandao 等人[4]采用了完全卷积网络 (FCN) 和预先训练的模型来分割息肉。后来，Akbari 等人[5]引入了一种改进的 FCN 来提高分割精度。受 UNet[6]在生物医学图像分割方面的巨大成功的启发，研究者设计出了 UNet++[7]和 ResUNet[8]来进行息肉分割，以提高性能。此外，PolypSeg[9]、ACS[10]、ColonSegNet[11]和 SCR-Net[12]探讨了统一架构在自适应学习语义上下文中的有效性。SANet[13]和 MSNet[14]作为新提出的方法，分别设计了浅层注意模块和减法单元，以实现精确和高效的分割。此外，一些工作还通过引入额外的约束来提升精度，三种约束分别是施加显式边界监督[15]、引入隐式边界监督[16]和探索模糊区域的不确定性[17]。

最近，Transformers[18]因其强大的建模能力而越来越受欢迎，越来越多的研究者将 Transformer 应用于计算机视觉领域的任务，并取得了不错的效果。TransFuse[19]将 Transformer 和 CNN 结合，采用一种分支并行的方案来捕获全局依赖和底层空间细节。此外，TransFuse 还设计了一个 BiFusion 模块来融合来自两个分支的多层次特征。Segtran[20]提出了一个压缩注意力模块来对原本的自注意力模块进行正则化，接着提出扩展模块来学习多样化的表示，同时提出一种位置编码方案来施加归纳连续性偏差。基于 PVT[21]，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。

2.2 视频息肉分割 (VPS)

尽管 IPS 方法的进展很迅速，并且取得了不错的结果，但 IPS 方法存在着固有的局限性，即忽视了视频序列中有价值的时间线索。因此，研究者们一直致力于结合视频连续帧之间的时空特征。一种混合 2/3D CNN 框架被提出用于聚合时空相关性，并获得了更好的分割效果。然而，核的大小限制了帧之间的空间相关性，限制了息肉快速运动的精确分割。为了缓解上述问题，PNSNet[22]引入了一个归一化的自注意 (NS) 块来有效地学习具有邻域相关性的时空表示。在本次复现工作中，作者的团队深入研究了一种更有效的基于 NS 块的全局-局部学习策略，它可以充分利用长期和短期的时空依赖性。

3 基于 NS 和全局-局部学习的 PNS+

3.1 PNS+ 框架

本小节将对 PNS+ 进行描述，描述内容包括标准化的自我注意力机制、全局到局部的学习策略，以及其它的实现细节。PNS+ 框架如图 1所示。

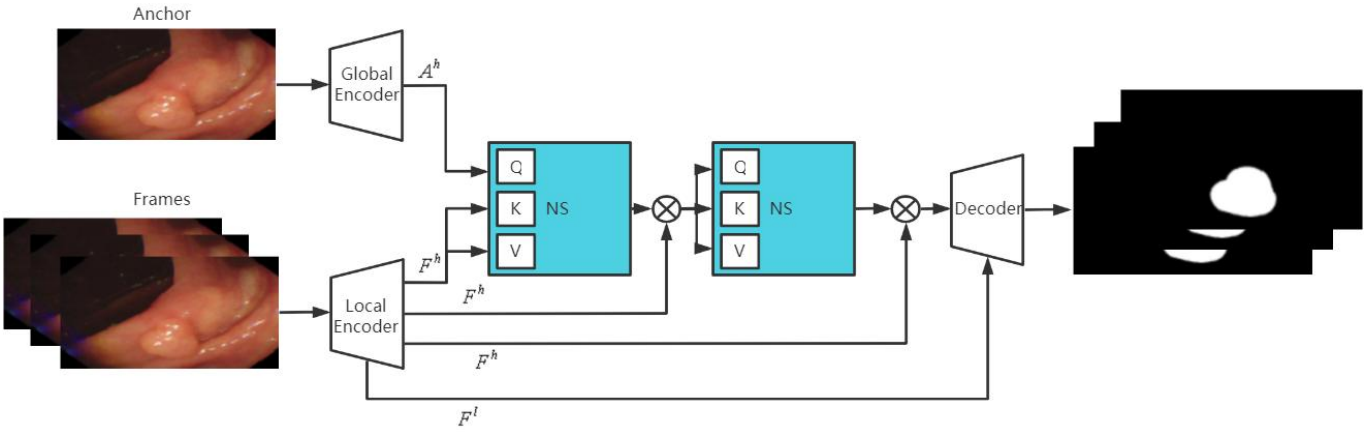

图 1: PNS+ 框架图

3.2 标准化自注意力模块

近年来，自我注意力机制[23]在许多流行的计算机视觉任务中得到了广泛的应用。根据作者团队最初的研究结果，由于在不同的拍摄角度和速度下可以捕捉到息肉的多尺度特性，因此将原始的自我注意机制引入到 VPS 任务中并不能获得令人满意的结果 (即较高的精度和速度)。因此，作者团队提出了一个归一化的自注意 (NS) 块，其动机是动态更新感受野，对于基于自我注意力机制的网络很重要，NS 模块如图 2 所示。NS 模块的计算过程可表示为：

$$\begin{aligned} Y \in \mathbb{R}^{T\times H\times W\times C} &= NS(Q,K,V) \\ &= (M^T W_T) \odot M^S \end{aligned} \tag{1}$$

其中，W_T 代表一组可学习的参数，$\odot$ 代表通道级别的哈达玛积运算。

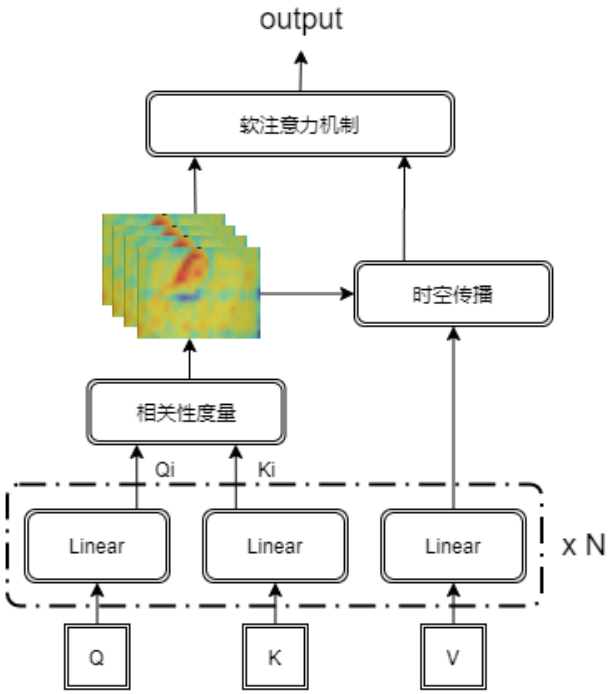

图 2: NS 模块

3.2.1 维度拆分机制

具体来说，给定维度为 $\mathbb{R}^{T\times H\times W\times C}$ 的三个候选特征 (即查询特征 Q、键特征 K 和值特征 V)，NS 模块利用三个线性嵌入函数 $\theta(\cdot)$、$\varphi(\cdot)$ 和 $g(\cdot)$ 来生成相应的注意特征。这些函数可以通过一个内核大

小为 1×1×1 的卷积层来实现，具体规则可表示为：

$$Q_i = F^G\langle\theta\left(Q\right)\rangle, K_i = F^G\langle\phi\left(K\right)\rangle, V_i = F^G\langle g\left(V\right)\rangle \tag{2}$$

其中，函数 F^G 表示将注意特征分成 N 个组的操作，最终得到三个不同的特征：查询特征 Q_i、键特征 K_i 和值特征 V_i，其中，$i = \{1, 2, \cdots, N\}$。

3.2.2 相关性度量

为了对连续帧之间的时空关系进行建模，测量查询特征 Q_i 和键特征 K_i 之间的相似性变得尤为重要。具体来说，NS 模块引入了 N 个相关性度量块来计算目标像素的约束领域的时空亲和矩阵，相关性度量块能够捕获更多的帧与帧之间的时空相关性。

定义 X^q 为 Q_i 在 (x,y,z) 处的像素点，相关性度量可以表示为 X^q 在 K_i 的约束领域内进行点采样的过程，其公式可表示为：

$$F^s\langle X^q, K_i\rangle = \sum_{m=x-kd_i}^{x+kd_i}\sum_{n=y-kd_i}^{y+kd_i}\sum_{t=1}^{T} K_i(m,n,t) \tag{3}$$

其中，$1 \leqslant x \leqslant H$,$1 \leqslant y \leqslant W$,$1 \leqslant t \leqslant T$, 同时 $F^s\langle X^q, K_i\rangle \in \mathbb{R}^{T(2k+1)^2\times\frac{C}{N}}$。因此，相关性度量的约束领域的大小取决于核大小 k、扩张率 d_i 和帧数 T。

基于上述的点采样函数，NS 模块以自适应点采样的方式获得亲和矩阵 M_i^A，用于度量目标像素及其周围时空内容的相关性。其公式可表示为：

$$M_i^A = Softmax(\frac{Q_i F^S\langle X^q, K_i\rangle^T}{\sqrt{C/N}}) \tag{4}$$

其中，$M_i^A \in \mathbb{R}^{THW\times T(2k+1)^2}$，$\sqrt{C/N}$ 是平衡多头注意力的比例因子。

3.2.3 时空传播

基于亲和矩阵和与相关性度量类似的方法，NS 模块还计算了时间聚合过程中受约束领域影响的时空聚合特征 $M_i^T \in \mathbb{R}^{THW\times\frac{C}{N}}$。其计算公式可表示为：

$$M_i^T = M_i^A F^S\langle X^a, V_i\rangle \tag{5}$$

其中，$X^a \in M_i^A$。

3.2.4 软注意力机制

我们首先沿着通道维数将一组亲和矩阵 M_i^A 进行连接，以生成 M^A。软注意力特征的计算公式可表示为：

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

其中，$M^A \in \mathbb{R}^{THW\times T(2k+1)^2 N}$，且 $Max(\cdot)$ 函数计算通道级别的最大值。然后，使用同样的聚合方式，NS 模块对所有的时空聚合特征 M_i^T 进行拼接来得到 M^T。

3.3 全局-局部学习策略

作者的团队提出了一种全新的全局到局部的学习策略，它在任意时间距离实现长期和短期的时空交互，产生了一种简单但能有效聚合时空相关性的学习策略。具体来说，在全局时间水平上附加一个时空学习分支，将长期的时空相关性引入网络的学习过程。

3.3.1 全局解码器

将一个视频序列的第一帧 $I_1 \in \mathbb{R}^{H\times W\times 3}$ 作为描点帧，全局解码器用于提取锚点帧的特征。使用与PraNet[16]类似的特征提取方法，全局解码器的选择是 Res2Net-50[24]，用其 conv4_6 层的输出作为提取的特征。同时为了减少后续的计算量，作者团队引入了 RFB 模块来进行特征通道数的缩减，最终得到锚点帧特征 $A^h \in \mathbb{R}^{H^h\times W^h\times C^h}$。

3.3.2 局部解码器

局部解码器以一段连续的视频帧 $F_\delta \in \mathbb{R}^{T\times H\times W\times C}$ 作为输入。与全局解码器的特征提取方式相似，我们使用 Res2Net-50 来提取两组特征，一组以 Res2Net-50 的 conv3_4 的输出作为低层特征 $F^l \in \mathbb{R}^{T\times H^l\times W^l\times C^l}$，另一组以 Res2Net-50 的 conv4_6 的输出作为高层特征 $F^h \in \mathbb{R}^{T\times H^h\times W^h\times C^h}$。其中，$H^l = \frac{H}{4}, W^l = \frac{W}{4}, C^l = 24, H^h = \frac{H}{8}, W^l = \frac{W}{8}, C^h = 32$。

3.3.3 全局-局部时空相关性传播

直观地说，全局-局部时空相关性传播的目的是建立锚点帧和高级短期特征以及低级长期特征之间的像素相似性，这可以视作全局时空依赖性的建模。首先是建立全局的时空依赖性，计算公式可表示为：

$$Z^g \in \mathbb{R}^{T\times H^h\times W^h\times C^h} = NS(A^h, F^h, F^h) \oplus X^h \tag{7}$$

其中，$\oplus$ 代表残差式的元素级相加，目的是让 NS 模块的内部梯度具有更好的传播稳定性。紧接着，将 Z^g 作为第二个 NS 模块的输入，目的是将长期依赖的 Z^g 传播到一个局部领域 (即与 F_δ 产生联系)。计算公式可表示为：

$$Z^l \in \mathbb{R}^{T\times H^h\times W^h\times C^h} = NS(Z^g, Z^g, Z^g) \oplus X^h \oplus Z^g \tag{8}$$

3.4 损失函数

给定一个预测 P^s 和相应的真实标签 (GT)G_s，PNS+ 使用二元交叉熵作为损失函数对网络的训练过程进行优化，其公式可表示为：

$$\mathcal{L}_{bce} = -\sum[G_s \log P_s + (1 - G_s) \log (1 - P_s)] \tag{9}$$

4 复现细节

4.1 与已有开源代码对比

本次复现工作是在作者提供的开源代码的基础上进行的。主要创新点是引入了一个特征记忆模块来对精度进行优化，引入特征记忆模块的动机是 PNS+ 框架在进行正向推理时仅仅用到了一个视频序列的片段，而没有使用所有已经出现过的帧来进行推理。

4.2 实验环境搭建

本次复现工作基于 Pytorch1.1 实现，使用单个 8GB 内存的 GeForce RTX 2080 GPU 在 SUN-SEG 数据集上训练 15 个轮次，batchsize 的选择为 8。图像的输入大小定义为 256×448，输入模型之前，所有图像都会进行归一化操作，目的是加速模型的训练过程。优化器的选择是 Adam，其中，初始学习率设定为 0.0001，参数衰减率设定为 0.0001。

4.3 创新点

特征记忆模块首先将 Res2Net-50 的 conv4_6 层的输出特征进行一次特征编码，所有历史帧的特征都会进行一次特征编码，接着使用一系列的矩阵操作来实现所有历史帧特征的融合，最后和当前帧的特征进行连接，得到最后的 memory 特征。memory 特征会输入 PNS+ 的 Decoder 模块进行预测的辅助。特征记忆模块如图 3 所示。

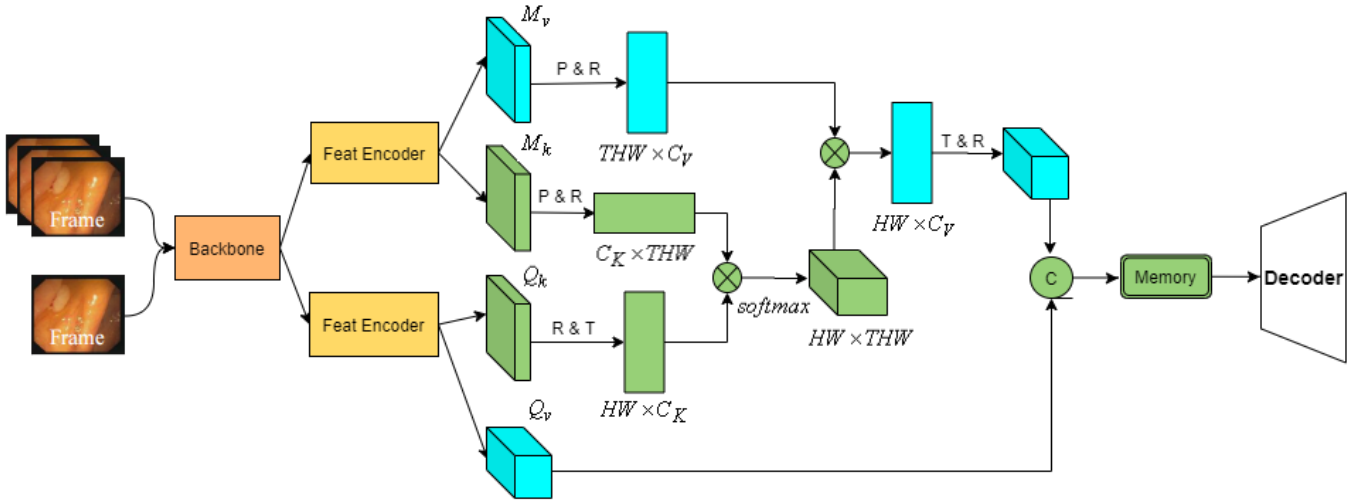

图 3: 特征记忆模块

5 实验结果分析

根据作者提供的训练参数的设置，复现模型在 15 个训练轮次后达到收敛程度，损失函数降到了 0.006 左右。但是复现模型和改进后的复现模型未能达到作者论文中的精度，但是添加了特征记忆模块后的改进模型在精度上超过了原本的复现模型。我在复现工作中选择了 $S_\alpha, E_\phi^{mn}, F_\beta^w, Dice$ 作为比较指标。我将改进后的模型 (FM-PNS+) 和原本的复现模型 (PNS+) 在 SUN-SEG-Easy/-Hard(Seen) 上进行了对比，具体结果如表 1。

表 1: FM-PNS+ 与 PNS+ 的精度比较

method	SUN-SEG-Easy (Seen)				SUN-SEG-Hard (Seen)			
	S_α	E_ϕ^{mn}	F_β^w	$Dice$	S_α	E_ϕ^{mn}	F_β^w	$Dice$
PNS+	0.886	0.898	0.792	0.746	0.855	0.883	0.751	0.712
FM-PNS+	**0.894**	**0.912**	**0.823**	**0.762**	**0.863**	**0.891**	**0.788**	**0.733**

同时，我提供了一些 FM-PNS+ 和 PNS+ 的可视化的比较结果，如图 4 所示。

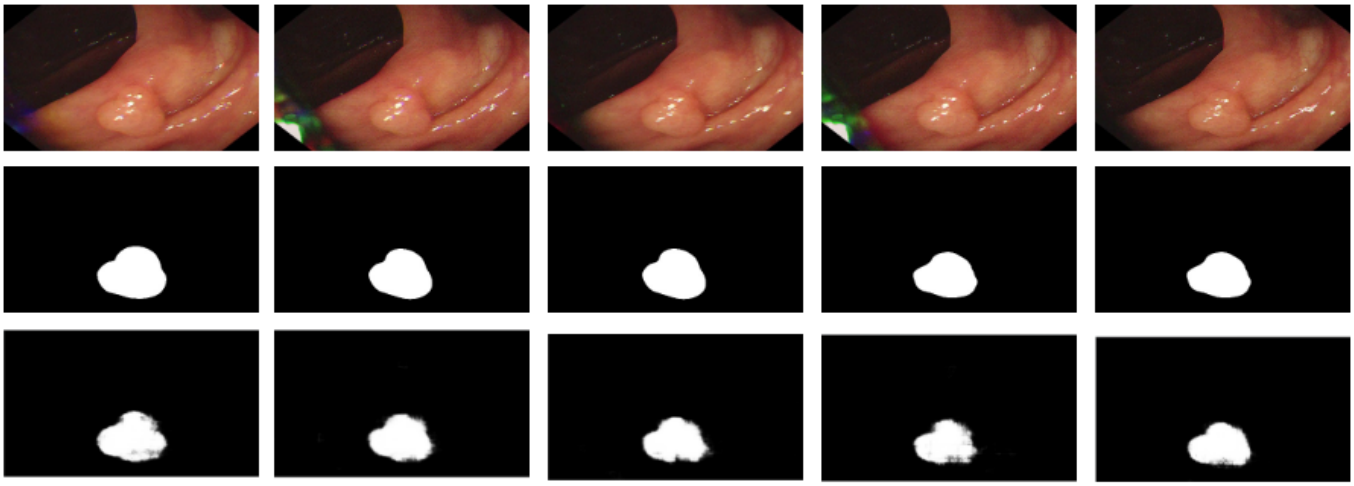

图 4: FM-PNS+ 与 PNS+ 的可视化比较

6 总结与展望

本次前沿技术的复现工作到此圆满完成。在本次复现工作中，我对视频息肉分割任务的发展历程有了更进一步的了解，明白了以往的工作障碍是没有一个标准的视频级的数据集，本次复现工作的一个重大贡献就是制作了第一个大型的视频级的息肉分割数据集。同时，我深入地了解了 PNS+ 的设计框架，并且提出了一个能够聚合所有历史帧特征的特征记忆模块来提升模型的分割精度，实验表明，特征记忆模块的引入对于分割精度有一定的提升。本次复现工作的唯一缺陷是原复现模型 (PNS+) 和改进模型 (FM-PNS+) 都未能达到作者论文中的精度，我猜测是模型超参数的设置和训练策略的选择有一定缺陷，后续的工作是找到一组合适的超参数和训练策略来提升模型的分割精度。

参考文献

[1] BERNAL J, SÁNCHEZ J, VILARINO F. Towards automatic polyp detection with a polyp appearance model[J]. Pattern Recognition, 2012, 45(9): 3166-3182.

[2] PUYAL J G B, BHATIA K K, BRANDAO P, et al. Endoscopic polyp segmentation using a hybrid 2D/3D CNN[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2020: 295-305.

[3] MISAWA M, KUDO S E, MORI Y, et al. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video)[J]. Gastrointestinal endoscopy, 2021, 93(4): 960-967.

[4] BRANDAO P, MAZOMENOS E, CIUTI G, et al. Fully convolutional neural networks for polyp segmentation in colonoscopy[C]//Medical Imaging 2017: Computer-Aided Diagnosis: vol. 10134. 2017: 101-107.

[5] AKBARI M, MOHREKESH M, NASR-ESFAHANI E, et al. Polyp segmentation in colonoscopy images using fully convolutional network[C]//2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2018: 69-72.

[6] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. 2015: 234-241.

[7] ZHOU Z, SIDDIQUEE M M R, TAJBAKHSH N, et al. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation[J]. IEEE transactions on medical imaging, 2019, 39(6): 1856-1867.

[8] JHA D, SMEDSRUD P H, RIEGLER M A, et al. Resunet++: An advanced architecture for medical image segmentation[C]//2019 IEEE International Symposium on Multimedia (ISM). 2019: 225-2255.

[9] ZHONG J, WANG W, WU H, et al. Polypseg: An efficient context-aware network for polyp segmentation from colonoscopy videos[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2020: 285-294.

[10] ZHANG R, LI G, LI Z, et al. Adaptive context selection for polyp segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2020: 253-262.

[11] JHA D, ALI S, TOMAR N K, et al. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning[J]. Ieee Access, 2021, 9: 40496-40510.

[12] WU H, ZHONG J, WANG W, et al. Precise yet efficient semantic calibration and refinement in convnets for real-time polyp segmentation from colonoscopy videos[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 35: 4. 2021: 2916-2924.

[13] WEI J, HU Y, ZHANG R, et al. Shallow attention network for polyp segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2021: 699-708.

[14] ZHAO X, ZHANG L, LU H. Automatic polyp segmentation via multi-scale subtraction network[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2021: 120-130.

[15] MURUGESAN B, SARVESWARAN K, SHANKARANARAYANA S M, et al. Psi-Net: Shape and boundary aware joint multi-task deep network for medical image segmentation[C]//2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2019: 7223-7226.

[16] FAN D P, JI G P, ZHOU T, et al. Pranet: Parallel reverse attention network for polyp segmentation [C]//International conference on medical image computing and computer-assisted intervention. 2020: 263-273.

[17] KIM T, LEE H, KIM D. Uacanet: Uncertainty augmented context attention for polyp segmentation[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 2167-2175.

[18] SHAMSHAD F, KHAN S, ZAMIR S W, et al. Transformers in medical imaging: A survey[J]. arXiv preprint arXiv:2201.09873, 2022.

[19] ZHANG Y, LIU H, HU Q. Transfuse: Fusing transformers and cnns for medical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2021: 14-24.

[20] LI S, SUI X, LUO X, et al. Medical image segmentation using squeeze-and-expansion transformers[J]. arXiv preprint arXiv:2105.09511, 2021.

[21] WANG W, XIE E, LI X, et al. Pvt v2: Improved baselines with pyramid vision transformer[J]. Computational Visual Media, 2022, 8(3): 415-424.

[22] JI G P, CHOU Y C, FAN D P, et al. Progressively normalized self-attention network for video polyp segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2021: 142-152.

[23] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7794-7803.

[24] BORJI A. Saliency prediction in the deep learning era: Successes and limitations[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 43(2): 679-700.