一种用于视频息肉分割的渐进式归一化自注意力网络

季葛鹏[1,2], 周昱程[2], 范登平[1,✉], 陈耿[1], 付华柱[1], Debesh Jha[3], and 邵岭[1]

[1] 起源人工智能研究院 [2] 武汉大学 [3] SimulaMet

摘要 现有的视频息肉分割 (VPS) 模型常常使用卷积神经网络 (CNNs) 提取特征。然而，CNNs 因其有限的感受野，不能在连续的视频帧之中充分地挖掘全局时空信息，产生假阳性分割结果。本文提出一种新颖的 ***PNS-Net***(渐进式归一化自注意力网络)，它能够在单张 RTX 2080 显卡上以 **~140fps** 实时的速度高效地从息肉视频中学习表征，并且无需任何后处理技术。本文的 *PNS-Net* 仅仅基于一个基础的归一化自注意力模块，它完全由循环结构和卷积神经网络构成。在具有挑战性的 VPS 数据集上的实验验证了本文的 *PNS-Net* 取得了卓越的性能。我们还进行了广泛的实验来研究通道分离、软注意和渐进学习策略的有效性。本文的 *PNS-Net* 在不同设定下均取得理想效果，使其成为一个可行的 VPS 任务解决方案。

Keywords: 归一化自注意力 · 息肉分割 · 结肠镜检查

1 引言

结直肠癌 (Colorectal Cancer, CRC) 的早期诊断对提高结直肠癌患者的生存率至关重要。事实上，所处第一阶段的结直肠癌的生存率超过 95%，而处于第四和第五阶段 [4] 则下降到 35% 以下。目前，结肠镜检查已广泛应用于临床之中，并已成为筛查结直肠癌的标准方法。在结肠镜检查期间，医生用内窥镜检查肠道以识别息肉，如果不加以治疗，息肉可能发展为结直肠癌。在临床实践之中，结肠镜检查在很大程度上依赖于医生的经验，息肉的漏诊率很高 [18]。这类局限性可通过自动息肉分割技术解决，该技术可以在不需要医生干预的情况下，从结肠镜图像/视频中分割息肉。然而，由于息肉与所处环境在边界处的对比度较低，且息肉的形状变化较大，因此准确并实时的息肉分割是一项具有挑战性的任务。

现有工作为克服这些挑战作出了诸多尝试。在早期的研究中，基于学习的方法主要依赖手工提取的特征 [16,20]，例如：颜色、形状、纹理、外观或

季葛鹏和周昱程对本文具有同等贡献度；通讯作者：范登平 (dengpfan@gmail.com)；代码链接：http://dpfan.net/pranet/

其组合。这类方法通常训练一个分类器来从结肠镜图像中分离出息肉。然而，由于手工提取特征在描述异质息肉以及息肉与难样本之间的极度相似性时表征能力有限，所以通常存在检测精度低的问题。最近研究显示，基于深度学习方法已被用于息肉分割 [24,26]。虽然这类方法取得了一定的进展，但它们只是使用矩形标记框来检测息肉，因此不能准确地定位边界。为了解决这个问题，Brandao 等人 [5] 采用了一个带有预训练的全卷积神经网络 (FCN) 来识别和分割息肉。随后，Akbari 等人 [1] 引入了一种改进的 FCN 来提高息肉分割的精度。受到 UNet [19] 网络在生物医学图像分割任务中成功的启发，UNet++ [28] 和 ResUNet [13] 被用于对息肉区域进行分割，并取得了良好的效果。还有一些方法侧重于区域-边界的约束。例如，Psi-Net [17] 同时使用息肉的边界和区域信息。Fang 等人 [9] 引入了一个三阶段的选择性特征聚合网络。ACSNet [25] 网络使用了一个基于上下文选择的自适应编码器-解码器框架。Zhong [27] 提出了一种基于自适应尺度和全局语义上下文的内容感知网络。近期，PraNet [8] 模型作为图像息肉分割的黄金标准，在反向注意模块中利用区域和边界线索，取得了卓越性能。然而，这些方法只针对静态图像进行训练和评估，并专注于静态信息，从而忽略了内镜镜视频中丰富的时序信息，这有助于得到更好预测结果。为此，Puyal 等人 [18] 提出了一种 2D 与 3D 混合的 CNN 架构。模型融合了时空关联性并获得了更好的分割结果。但是帧间的空间相关性受到卷积核大小的限制，不利于视频的快速而准确分割。

最近，自注意力网络 [22] 在视频目标分割 [10]、图像超分辨率 [23] 学习等计算机视觉任务中表现出了优异的性能。受此启发，本文提出了一种新的自注意框架，即渐进式归一化自注意力网络 (***PNS-Net***) 来解决视频息肉分割 (VPS) 任务。本文的贡献如下：

- 与现有的基于卷积神经网络模型不同的是，本文提出的模型框架引入了一个全新的视角即自注意力模型来解决 VPS 任务。
- 为充分利用时间和空间线索，本文提出了一个简单的归一化自注意力块 (NS)。NS 块灵活 (与骨架无关) 且高效，使其能很容易地被嵌入到当前基于卷积神经网络的编码器-解码器架构中，从而获得更好的性能。
- 本文在具有挑战性的 VPS 数据集上评估了所提出的模型，并将其与两种经典方法 (即：UNet [19] 和 UNet++ [28]) 和三种前沿模型 (即：ResUNet [13]、ACSNet [25] 和 PraNet [8]) 进行了比较。实验结果表明

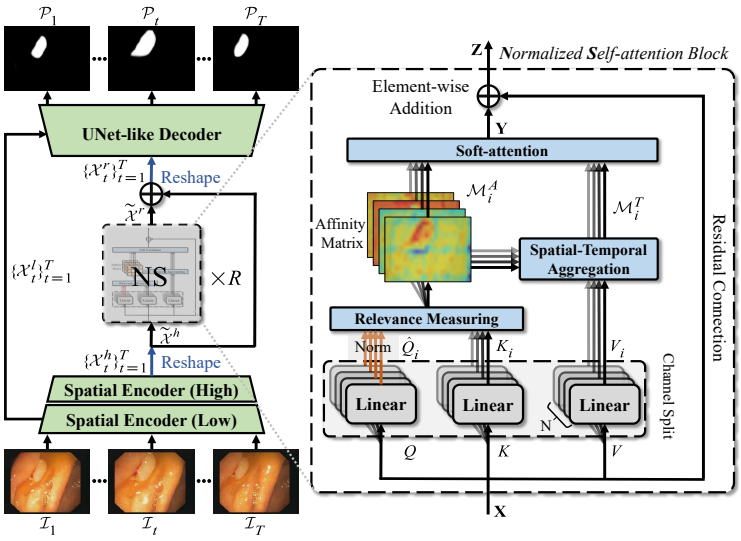

图 1: 本文的 *PNS-Net* 模型的流程图，由堆叠 (×R) 学习策略 (请参见 § 2.2) 的归一化自注意力块 (请参见 § 2.1) 构成。

该模型具有较高的实时性。所有的训练数据、模型、结果和评估工具都将开源以推动该领域的发展。

2 方法

2.1 归一化自注意力 (NS)

动机： 近年来，自注意力机制 [22] 在许多流行的计算机视觉任务中得到了广泛的应用。然而，在本文初步的研究中发现，在 VPS 任务中引入原始的自注意力机制并没有取得令人满意的结果 (即：高精度和高推理速度)。

分析： 对于 VPS 任务，各种尺寸的息肉以不同的速度移动。因此，动态地更新网络的感受野是很重要的。此外，自注意力（如非局部网络 [22]）需要较高的计算和内存资源，这限制了快速和密集的预测任务的推理速度。受到最近视频显著性目标检测模型 [10] 的启发，本文分别利用**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。

通道分离法则：具体来说，给定输入特征 (即： $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times C}$)，它提取自 T 个视频帧，尺寸为 $H \times W$，通道数为 C。首先使用三个嵌入函数 $\theta(\cdot)$、$\phi(\cdot)$ 和 $g(\cdot)$ 分别生成对应的注意力特征，该过程可以使用 $1 \times 1 \times 1$ 大

小的卷积层实现 [22]。该过程可被表示为:

$$Q = \theta(\mathbf{X}), K = \phi(\mathbf{X}), V = g(\mathbf{X}). \tag{1}$$

然后，将每一个注意力特征 $\{Q, K, V\} \in \mathbb{R}^{T\times H\times W\times C}$ 沿着特征维度分离成 N 组，并生成查询特征、键特征和值特征，即： $\{Q_i, K_i, V_i\} \in \mathbb{R}^{T\times H\times W\times \frac{C}{N}}$，其中 $i = \{1, 2, \cdots, N\}$。

查询相关法则：为了提取连续视频帧之间的时空关联，需要度量查询特征 Q_i 与键特征 K_i 之间的相似性。受到 [10] 启发，本文引入 N 个关联性度量块 (即： 查询相关法则)，用于计算查询特征目标像素点与其在键特征受约束邻域内的时空相似度矩阵。如 [22] 所述，关联性度量块可以在 T 帧内捕获关于目标对象的更多相关性，而不是计算查询位置和特征中所有位置之间的响应。更为具体地，给定一个具有固定核大小 k 和空洞率 $d_i = 2i - 1$ 的卷积，得到 Q_i 中每一个位置为 (x, y, z) 的查询像素 $\mathbf{X}^q$ 与在 K_i 特征中所对应的受约束邻域，该邻域可由一个采样函数 $\mathcal{F}^S$ 所得。计算公式如下:

$$\mathcal{F}^S\langle \mathbf{X}^q, K_i\rangle \in \mathbb{R}^{T(2k+1)^2\times \frac{C}{N}} = \Sigma_{m=x-kd_i}^{x+kd_i}\Sigma_{n=y-kd_i}^{y+kd_i}\Sigma_{t=1}^{T} K_i(m, n, t), \tag{2}$$

其中 $1 \le x \le H$、$1 \le y \le W$ 和 $1 \le z \le T$。因此，受约束邻域的大小取决于不同的时空感受野，分别具有不同的卷积核大小 k、膨胀率 d_i 和帧数 T。

归一化法则：然而，输入 Q_i 的前馈存在内部协变量偏移的问题 [11]，导致层参数不能动态地适应下一个小批次数据。因此，本文通过以下方式保持 Q_i 特征的固定分布:

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

其中的 `Norm` 是通过层归一化 [2] 操作 沿时间维度实现的。

相关性度量： 最终，相似度矩阵由下式计算而来:

$$\mathcal{M}_i^A \in \mathbb{R}^{THW\times T(2k+1)^2} = \texttt{Softmax}\left(\frac{\hat{Q}_i\mathcal{F}^S\langle \hat{\mathbf{X}}^q, K_i\rangle^{\mathbf{T}}}{\sqrt{C/N}}\right), \text{ when } \hat{\mathbf{X}}^q \in \hat{Q}_i, \tag{4}$$

其中 $\sqrt{C/N}$ 代表放缩因子，用于平衡多头注意力 [21]。

时空聚合： 与相关性度量方式相似，该方案也计算了时空聚合特征 $\mathcal{M}_i^T$ 在约束邻域内的时域聚合。该过程可以被表述为:

$$\mathcal{M}_i^T \in \mathbb{R}^{THW\times \frac{C}{N}} = \mathcal{M}_i^A\mathcal{F}^S\langle \mathbf{X}^a, V_i\rangle, \text{ when } \mathbf{X}^a \in \mathcal{M}_i^A, \tag{5}$$

软注意力： 本方法使用了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，应加强相关的时空模式，抑制弱相关

的时空模式。首先沿着通道维度拼接一组相似度矩阵 $\mathcal{M}_i^A$，用于生成 $\mathcal{M}^A$。因此，软注意力图 $\mathcal{M}^S$ 可由下式计算而来：

$$\mathcal{M}^S \in \mathbb{R}^{THW\times 1} \leftarrow \texttt{max}\mathcal{M}^A, \text{ when } \mathcal{M}^A \in \mathbb{R}^{THW\times T(2k+1)^2N}, \tag{6}$$

其中 max 函数计算了通道维度上的最大值。然后沿着通道维度拼接一组时空聚合特征 $\mathcal{M}_i^T$，用于生成 $\mathcal{M}^T$。

归一化自注意力： 最终，本方案的 NS 块可以定义为：

$$\mathbf{Z} \in \mathbb{R}^{T\times H\times W\times C} = \mathbf{X} + \mathbf{Y} = \mathbf{X} + (\mathcal{M}^T\mathbf{W}_T) \circledast \mathcal{M}^S, \tag{7}$$

其中 $\mathbf{W}_T$ 代表可学习的权值，$\circledast$ 代表在通道维度的哈达玛积。

2.2 渐进式学习策略

编码器： 为了公平比较，本方案使用了与 PraNet [8] 相同的骨架网络 (即：Res2Net-50)。给定一个含有 T 帧的视频片段作为输入 (即：$\{\mathcal{I}\}_{t=1}^T \in \mathbb{R}^{H'\times W'\times 3}$)，首先将其输入空间编码器中，并从 conv3_4 和 conv4_6 层中分别提取出两个空间特征。为了减轻计算负担，本文采用一个近似 RFB 结构 [15] 的模块来减少特征通道。从而得到两个空间特征，即底层特征 (即：$\{\mathcal{X}_t^l\}_{t=1}^T \in \mathbb{R}^{H^l\times W^l\times C^l}$) 和高层特征 (即：$\{\mathcal{X}_t^h\}_{t=1}^T \in \mathbb{R}^{H^h\times W^h\times C^h}$)[1]。

渐进式归一化自注意力： 大多数注意力策略的旨在提纯候选特征，如一阶 [8] 和二阶 [22,21] 函数。因此，高层特征中的强语义信息可能在网络的前向传递过程中逐渐分散。为了缓解这种情况，本文在 NS 块中引入了一个渐进式残差学习策略。具体而言，首先将与连续输入帧相对应的高层特征 $\{\mathcal{X}_t^h\}_{t=1}^T$ 重构为时序特征，该特征可以视为一个四维张量 (即：$\widetilde{\mathcal{X}}^h \in \mathbb{R}^{T\times H^h\times W^h\times C^h}$)。然后以渐进式的方式来堆叠归一化自注意力，用于提纯 $\widetilde{\mathcal{X}}^h$：

$$\widetilde{\mathcal{X}}^r \in \mathbb{R}^{T\times H^h\times W^h\times C^h} = \texttt{NS}^{\times R}(\widetilde{\mathcal{X}}^h) = \texttt{NS}^{\times R}(\mathcal{F}^R(\{\mathcal{X}_t^h\}_{t=1}^T)), \tag{8}$$

其中 $\texttt{NS}^{\times R}$ 表示在提纯过程中堆叠了 R 个归一化自注意力块。$\mathcal{F}^R$ 代表时间维度的特征重构函数。为了使得该模块轻松地嵌入预训练网络中，常用的解决方案是添加一个残差学习过程。最终，提纯后的时空特征可由下式产生：

$$\{\mathcal{X}_t^r\}_{t=1}^T \in \mathbb{R}^{H^h\times W^h\times C^h} = \mathcal{F}^R(\widetilde{\mathcal{X}}^h + \widetilde{\mathcal{X}}^r). \tag{9}$$

[1] 本文设定 $H^l = \frac{H'}{4}$、$W^l = \frac{W'}{4}$、$C^l = 24$、$H^h = \frac{H'}{8}$、$W^h = \frac{W'}{8}$ 和 $C^h = 32$。

表 1: 在不同数据集上的定量结果

		2018∼2019			2020		2021
	Metrics	UNet MICCAI [19]	UNet++ TMI [28]	ResUNet ISM [13]	ACSNet MICCAI [25]	PraNet MICCAI [8]	***PNS-Net* (OUR)**
	Speed	108fps	45fps	20fps	35fps	97fps	**140fps**
CVC-300-TV	maxDice↑	0.639	0.649	0.535	0.738	0.739	**0.840**
	maxSpe↑	0.963	0.944	0.852	0.987	0.993	**0.996**
	maxIoU↑	0.525	0.539	0.412	0.632	0.645	**0.745**
	$S_\alpha \uparrow$	0.793	0.796	0.703	0.837	0.833	**0.909**
	$E_\phi \uparrow$	0.826	0.831	0.718	0.871	0.852	**0.921**
	$M \downarrow$	0.027	0.024	0.052	0.016	0.016	**0.013**
CVC-612-V	maxDice↑	0.725	0.684	0.752	0.804	0.869	**0.873**
	maxSpe↑	0.971	0.952	0.939	0.929	0.983	**0.991**
	maxIoU↑	0.610	0.570	0.648	0.712	0.799	**0.800**
	$S_\alpha \uparrow$	0.826	0.805	0.829	0.847	0.915	**0.923**
	$E_\phi \uparrow$	0.855	0.830	0.877	0.887	0.936	**0.944**
	$M \downarrow$	0.023	0.025	0.023	0.054	0.013	**0.012**
CVC-612-T	maxDice↑	0.729	0.740	0.617	0.782	0.852	**0.860**
	maxSpe↑	0.971	0.975	0.950	0.975	0.986	**0.992**
	maxIoU↑	0.635	0.635	0.514	0.700	0.786	**0.795**
	$S_\alpha \uparrow$	0.810	0.800	0.727	0.838	0.886	**0.903**
	$E_\phi \uparrow$	0.836	0.817	0.758	0.864	**0.904**	0.903
	$M \downarrow$	0.058	0.059	0.084	0.053	0.038	**0.038**

解码器和学习策略： 本方案通过一个两阶段的类 UNet 解码器 $\mathcal{F}^D$ 融合来自空间编码器的底层特征 $\{\mathcal{X}_t^l\}_{t=1}^T$ 和来自 PNS 块的时空特征 $\{\mathcal{X}_t^r\}_{t=1}^T$。因此，该方法的输出可由 $\{\mathcal{P}_t\}_{t=1}^T = \mathcal{F}^D(\{\mathcal{X}_t^l\}_{t=1}^T, \{\mathcal{X}_t^r\}_{t=1}^T)$ 计算而来。在学习过程中采用了标准的二值交叉熵损失函数。

3 实验

3.1 实现细节

数据集： 本文在实验中使用了四个常用的息肉数据集，包括基于图像的 (即： Kvasir [12]) 和基于视频的 (即： CVC-300 [4]、CVC-612 [3] 和 ASU-Mayo [20]) Kvasir 是大型且具有挑战性的数据集，其包含了 1,000 张带有逐像素标注的息肉图像样本。整个 Kvasir 数据集被用于训练。ASU-Mayo 包含 10 个来自正常人的视频负样本和 10 个来自病人的视频正样本。本文仅使用正样本用于训练。与 [4,3] 采用同一套协议，本文将来自 CVC-300 数据集 (12 个视频片段) 和 CVC-612 数据集 (29 视频片段) 按照 60% 训练集、20% 验证集和 20% 测试集的比例进行划分。

训练： 由于受到视频训练数据的限制，本文充分使用了大规模的息肉图像数据去捕获息肉和场景中的表观信息。因此，本文使用下列两步来训练模型：*i) 预训练阶段*：从 *PNS-Net* 中移除归一化自注意力 (NS) 块并使用基于图像的息肉数据集 (即： Kvasir [12]) 和基于视频的息肉数据集的训练集部分 (即： CVC-300 [4]、CVC-612 [3] 和 ASU-Mayo [20])。Adam 算法的初始学习速率和权重衰减均设定为 1e-4。*PNS-Net* 模型的静态部分在 100 个训练周期后收敛。*ii) 微调阶段*：将 NS 块嵌入 *PNS-Net* 模型中，并使用视频息肉数据集 (ASU-Mayo 和 CVC-300 与 CVC-612 的训练集) 对整个网络进行微调。将注意力组数设定为 $N = 4$，堆叠的归一化自注意力块数量设定为 $R = 2$，卷积核尺寸设定为 $k = 3$。初始学习率设定为 1e-4，并对整个模型进行一个周期的微调。虽然密集标注的 VPS 数据很少，但所提出的模型仍具有良好的泛化性能。

测试和运行时间： 为了测试 *PNS-Net* 模型的性能，本文在具有挑战性的数据集上进行了验证，包括：CVC-612 的测试集 (即： CVC-612-T), CVC-612 的验证集 (即： CVC-612-V) 和 CVC-300 测试/验证集 (即： CVC-300-TV)。在推理过程中，本文从息肉片段中采样了 T=5 帧并将其缩放到 256×448 大小作为输入。本文使用网络的输出 $\mathcal{P}_t$ 作为最终的输出，后接一个 *sigmoid* 函数。本文的 *PNS-Net* 在单张 RTX 2080 显卡上取得了约 140fps 的推理速度，无需任何后处理技术 (CRF [14])。表. 1列出了对比模型的速度。

3.2 视频息肉分割的评估

基线模型： 为了公平的对比，本文使用了与 *PNS-Net* 模型相同的数据，在默认设定下重新训练了五个先进的息肉分割基线 (即： UNet [19]、UNet++ [28]、ResUNet [13]、ACSNet [25] 和 PraNet [8])。

指标： 指标包括：(1) 最大 Dice (maxDice) 用于度量两组数据之间的相似度；(2) 最大特异性 (maxSpe) 是指被判定为阴性的样本的百分比；(3) 最大交并比 (maxIoU) 测量两个掩模图像之间的重叠率; (4) 结构指标 [6] (S_α) 评估基于区域的和基于目标的结构相似度。(5) 增强匹配指标 [7] (E_ϕ) 度量像素级别的匹配量和图像级别的统计量。(6) 平均绝对误差 (M) 度量预测图像和真值图像之间像素级别的误差。

定性对比： 图. 2展示了本文方法在 CVC-612-T 数据集上的息肉分割结果。本文模型能够在诸如不同的尺寸、同质区域和不同纹理等多种困难场景下准确地定位和分割息肉。

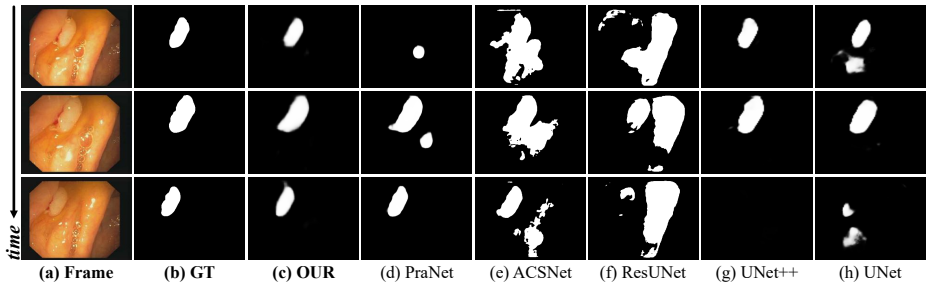

图 2: CVC-612-T [3] 数据集上的定量结果展示。更多可视化结果请参见补充材料 (即： PDF 文档和视频)。

表 2: 消融实验。更多细节请参见 § 3.3

No.	Variants					CVC-300-TV				CVC-612-T			
	Base	N	Soft	Norm	R	maxDice↑	maxIoU↑	S_α ↑	E_ϕ ↑	maxDice↑	maxIoU↑	S_α ↑	E_ϕ ↑
#1	✓					0.778	0.665	0.850	0.858	0.850	0.778	0.896	0.885
#2	✓	1			1	0.755	0.650	0.865	0.844	0.850	0.779	0.896	0.891
#3	✓	2			1	0.790	0.679	0.876	0.872	0.825	0.746	0.870	0.856
#4	✓	4			1	0.809	0.709	0.893	0.884	0.834	0.760	0.881	0.867
#5	✓	8			1	0.763	0.663	0.867	0.842	0.787	0.702	0.841	0.829
#6	✓	4	✓		1	0.829	0.729	0.896	0.903	0.852	0.784	0.895	0.897
#7	✓	4	✓	✓	1	0.827	0.732	0.897	0.898	0.856	0.792	0.898	0.896
#8	✓	4	✓	✓	2	**0.840**	**0.745**	**0.909**	**0.921**	**0.860**	**0.795**	**0.903**	**0.903**
#9	✓	4	✓	✓	3	0.737	0.609	0.793	0.751	0.732	0.613	0.776	0.728

定量对比： 表. 1列出了定量对比结果。本文在测试集上进行了三组实验来验证模型的性能。CVC-300-TV 由验证集和测试集组成，共包含 6 个视频。CVC-612-V 和 CVC-612-T 各自包含五个视频。在 CVC-300 上，所有的基线方法都表现很差，本文的 *PNS-Net* 模型在所有指标上都取得了显著性能，并很大程度上超过 (max Dice:∼10%) 了所有的前沿方法。在 CVC-612-V 和 CVC-612-T 数据集上，本文的 *PNS-Net* 模型 始终优于其他前沿算法。

3.3 消融实验

通道分离的有效性： 本文研究了不同尺度下通道分离法则的贡献度。结果在表. 2的 #2 行到 #5 行中列出。观察到 #4 实验 (N=4) 在所有指标上均优于其他的设定 (即： #2、#3 和 #5) 。这些提升表明，不合适的感受野 (RF) 尺寸会损害时序信息的挖掘能力，因为较大的感受野会更多地关注全局内容而不是局部的运动信息。另一方面，当通道分离数太小时，模型无法捕捉到以不同速度运动的多尺度息肉。

软注意力的有效性： 我们进一步研究软注意力机制的作用。如表. 2所示，在 CVC-612-T 数据集上使用软注意力块的 #6 通常比 #4 表现要好。这表明引入软注意力块来融合聚合特征和相似度矩阵是提高性能的必要条件。

NS 块数量的有效性： 为了获得在不同数量归一化自注意力块下的设定，本文派生出三个变体模型，分别为 #7、#8 和 #9。我们观察到，在 CVC-300-TV 和 CVC-612-T 数据集的所有指标下，设定 $R = 2$ 的 #8 (*PNS-Net* 设定) 明显优于 #7 和 #9。这一提升说明了过多的 NS 块迭代可能会导致对小数据集产生过拟合 (#9)。相比之下，该模型无法通过单一的残差块来缓解高阶特征的扩散问题。根据经验，在更大的数据集上训练时，建议增加 NS 块的数量。

4 结论

本文提出了一个基于自注意力的框架 (*PNS-Net*)，其以超高推理速度 (~140fps) 从结肠镜检查视频中准确地分割出息肉。这种基本的归一化自注意力块能够轻易地嵌入到现有的基于卷积神经网络的架构之中。实验表明，本文的 *PNS-Net* 模型在现有的公共数据集上在六个指标下取得了最佳性能。此外，大量的消融实验表明，*PNS-Net* 模型中的核心组件均有效。我们希望本文的 *PNS-Net* 模型能够在 VPS 任务和其它密切相关的基于视频的医学分割任务中作为一个催化剂，进而推动相关研究。未来，我们打算在更大规模的 VPS 数据集上探索 *PNS-Net* 模型的性能。

参考文献

1. Akbari, M., Mohrekesh, M., Nasr-Esfahani, E., Soroushmehr, S.R., Karimi, N., Samavi, S., Najarian, K.: Polyp segmentation in colonoscopy images using fully convolutional network. In: IEEE EMBC. pp. 69–72 (2018)
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. CMIG **43**, 99–111 (2015)
4. Bernal, J., Sánchez, J., Vilarino, F.: Towards automatic polyp detection with a polyp appearance model. PR **45**(9), 3166–3182 (2012)

5. Brandao, P., Mazomenos, E., Ciuti, G., Caliò, R., Bianchi, F., Menciassi, A., Dario, P., Koulaouzidis, A., Arezzo, A., Stoyanov, D.: Fully convolutional neural networks for polyp segmentation in colonoscopy. In: MICAD. vol. 10134, p. 101340F (2017)
6. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: IEEE ICCV. pp. 4548–4557 (2017)
7. Fan, D.P., Ji, G.P., Qin, X., Cheng, M.M.: Cognitive vision inspired object segmentation metric and loss function. SSI (2020)
8. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: MICCAI. pp. 263—273 (2020)
9. Fang, Y., Chen, C., Yuan, Y., Tong, K.y.: Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: MICCAI. pp. 302–310. Springer (2019)
10. Gu, Y., Wang, L., Wang, Z., Liu, Y., Cheng, M.M., Lu, S.P.: Pyramid constrained self-attention network for fast video salient object detection. In: AAAI. vol. 34, pp. 10869–10876 (2020)
11. Guo, L., Liu, J., Zhu, X., Yao, P., Lu, S., Lu, H.: Normalized and geometry-aware self-attention network for image captioning. In: IEEE CVPR. pp. 10327–10336 (2020)
12. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MMM. pp. 451–462. Springer (2020)
13. Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D.: Resunet++: An advanced architecture for medical image segmentation. In: IEEE ISM. pp. 225–2255 (2019)
14. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. NIPS **24**, 109–117 (2011)
15. Liu, S., Huang, D., et al.: Receptive field block net for accurate and fast object detection. In: ECCV. pp. 385–400 (2018)
16. Mamonov, A.V., Figueiredo, I.N., Figueiredo, P.N., Tsai, Y.H.R.: Automated polyp detection in colon capsule endoscopy. IEEE TMI **33**(7), 1488–1502 (2014)
17. Murugesan, B., Sarveswaran, K., Shankaranarayana, S.M., Ram, K., Joseph, J., Sivaprakasam, M.: Psi-Net: Shape and boundary aware joint multi-task deep network for medical image segmentation. In: IEEE EMBC. pp. 7223–7226 (2019)
18. Puyal, J.G.B., Bhatia, K.K., Brandao, P., Ahmad, O.F., Toth, D., Kader, R., Lovat, L., Mountney, P., Stoyanov, D.: Endoscopic polyp segmentation using a hybrid 2d/3d cnn. In: MICCAI. pp. 295–305. Springer (2020)

19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
20. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE TMI **35**(2), 630–644 (2015)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
22. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: IEEE CVPR. pp. 7794–7803 (2018)
23. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: IEEE CVPR. pp. 5791–5800 (2020)
24. Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A.: Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. IEEE JBHI **21**(1), 65–75 (2016)
25. Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., Yu, Y.: Adaptive context selection for polyp segmentation. In: MICCAI. pp. 253–262. Springer (2020)
26. Zhang, R., Zheng, Y., Poon, C.C., Shen, D., Lau, J.Y.: Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. PR **83**, 209–219 (2018)
27. Zhong, J., Wang, W., Wu, H., Wen, Z., Qin, J.: Polypseg: An efficient context-aware network for polyp segmentation from colonoscopy videos. In: MICCAI. pp. 285–294. Springer (2020)
28. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. IEEE TMI pp. 3–11 (2019)