Modeling and Simulation 建模与仿真, 2025, 14(3), 337-347
Published Online March 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.143227

Hans汉斯

SAM2U-Net：基于SAM2与U-Net的医学图像分割模型

候俊伟[1]，刘　磊[1,2]

[1]上海理工大学，光电信息与计算机工程学院，上海
[2]上海理工大学，管理学院，上海

收稿日期：2025年2月22日；录用日期：2025年3月15日；发布日期：2025年3月24日

摘　要

医学图像分割在医学诊断中起着关键作用。尽管新兴视觉模型在各种医学分割任务中表现优异，但大多仅针对特定任务设计，缺乏普适性。本研究提出了一种新型SAM2学习模型，旨在实现通用医学图像分割。该模型基于U型架构，创新性地将SAM2的Hiera骨干网络与CNN模块并行结合，通过多尺度特征提取机制增强分割精度。在甲状腺结节诊断、结肠镜息肉分割等六个数据集上的实验表明，本模型在Dice系数和IoU上平均提高了1.46%，优于现有方法。结果证实，该模型能有效提取医学图像的病理特征，实现准确的区域分割，为广泛的临床诊断任务提供支持。

关键词

甲状腺结节，息肉分割，深度学习，超声图像，分割算法，医学图像，U型架构，Transformer模块，并行网络

SAM2U-Net: A U-Shaped Medical Image Segmentation Model Based on SAM2 and U-Net

Junwei Hou[1], Lei Liu[1,2]

[1]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai
[2]School of Management, University of Shanghai for Science and Technology, Shanghai

Received: Feb. 22nd, 2025; accepted: Mar. 15th, 2025; published: Mar. 24th, 2025

文章引用: 候俊伟, 刘磊. SAM2U-Net: 基于 SAM2 与 U-Net 的医学图像分割模型[J]. 建模与仿真, 2025, 14(3): 337-347.
DOI: 10.12677/mos.2025.143227

Abstract

Medical image segmentation plays a crucial role in medical diagnosis. Although emerging visual models perform excellently in various medical segmentation tasks, most are designed for specific tasks and lack universality. This study proposes a novel SAM2 learning model aimed at achieving general medical image segmentation. The model is based on a U-shaped architecture and innovatively combines the Hiera backbone network of SAM2 with CNN modules in parallel, enhancing segmentation accuracy through a multi-scale feature extraction mechanism. Experiments on six datasets, including thyroid nodule diagnosis and colonoscopic polyp segmentation, demonstrate that this model improves the average Dice coefficient and IoU by 1.45% compared to existing methods. The results confirm that the model can effectively extract pathological features from medical images, achieving accurate regional segmentation and supporting a wide range of clinical diagnostic tasks.

Keywords

Thyroid Nodule, Polyp Segmentation, Deep Learning, Ultrasound Imaging, Segmentation Algorithm, Medical Imaging, U-Shaped Architecture, Transformer Module, Parallel Network

1. 引言

在传统的医学图像分割领域，主要依赖数字图像处理技术进行分割，但并未针对特定医学问题训练专业模型或方法。例如，常用的阈值法[1]、区域生长法[2]和边缘检测法[3]等，核心在于分析图像中像素值的统计特征，并根据一定规则进行目标区域的分割。然而，这些规则往往是通用的，缺乏针对特定医学问题的专业知识。因此，在图像复杂且对结果精度要求较高的医学领域，这些传统方法的分割效果相对较差，难以满足临床应用的需求。

深度学习的兴起为图像分割领域带来了革命性的变革。得益于计算机性能的大幅提升和多个里程碑式模型结构的提出，深度学习模型在公开的图像分割任务中的表现大幅超越了传统分割方法。卷积神经网络(CNN) [4]作为深度学习的代表算法之一，在各种医学分割任务中展现出卓越的性能，包括甲状腺结节分割、视网膜图像分割[5] [6]等。U-Net [7]作为一种开创性的医学图像分割架构，在多种分割任务中取得了令人瞩目的成果。然而，卷积神经网络受限于卷积核的特性，主要擅长对局部信息进行建模，在捕捉全局上下文信息方面存在不足。

视觉基础模型(Vision Foundation Models, VFMs)的出现为图像分割领域开辟了新的研究方向。其中，特别值得关注的是 Segment Anything Model (SAM) [8]及其改进版本 SAM2 [9]。SAM2 在 SAM 的基础上，利用更大规模的数据集进行训练，并在模型架构上进行了优化。然而，尽管有这些改进，当没有提供手动提示时，SAM2 仍可能产生与分割类别无关的结果，这一缺陷在一定程度上限制了其在医学分割等下游任务中的直接应用。为了使 SAM 更好地适应下游任务，研究者们提出了几种方法，包括使用适配器[10]进行参数高效微调，以及集成额外的条件输入(如文本提示[11]或上下文样本[12])。一些研究人员还探索了将 SAM 转换为 U 型架构的可能性[7]。然而，受视觉编码器(ViT) [13]结构的限制，这些方法往往缺乏处理复杂分割任务

所需的层次结构。随着 SAM2 的引入，其优秀的分层骨干网络为设计新型 U 型网络开辟了新的途径。

尽管现有研究在医学图像分割领域取得了显著进展，但仍面临着如何同时有效捕获局部细节和全局上下文信息的挑战。特别是在处理复杂的医学图像时，现有模型往往难以在保持高分辨率的同时准确识别病变区域。为了解决这一问题，本研究提出了一种新颖的并行编码器结构，旨在结合 CNN 的局部特征提取能力和 SAM2 的全局上下文捕获能力。

本文提出的并行编码器结构通过融合 CNN 和 SAM2 的优势，有效解决了传统单分支编码器在复杂医学图像分割任务中的局限性。具体而言，本研究的主要贡献包括：1) 设计了一种新型并行编码器架构，同时处理局部细节和全局上下文信息；2) 创新性地结合了 CNN 的局部特征提取能力和 SAM2 预训练 Hiera 骨干网络的多尺度特征捕获能力；3) 提出了一种更加灵活和高效的医学图像分割解决方案，适用于处理复杂的病变区域。通过这种设计，本研究显著提升模型在多样化医学图像分割任务中的表现，为临床诊断和治疗规划提供更加可靠的技术支持。

2. 基于 SAM2 与 U-Net 的网络分割模型

本文提出的 SAM2U-Net 分割网络网络结构如图 1 所示。

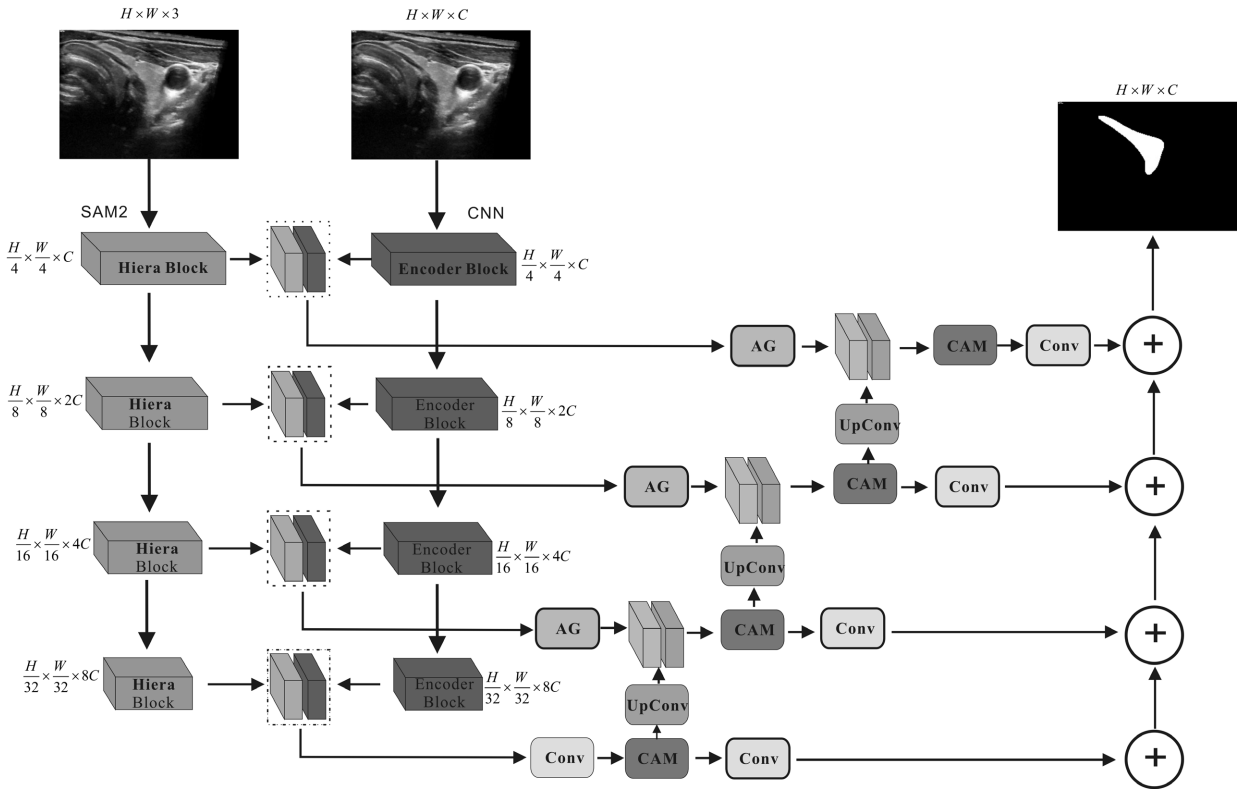

Figure 1. SAM2U-Net structure
图 1. SAM2U-Net 结构图

2.1. SAM2 与 U-Net 特征融合编码器

特征融合编码器通过引入两个并行的特征提取分支，使得网络能够同时处理局部和全局信息。对于 SAM2 分支，编码器使用由 SAM2 预训练的 Hiera [14]主干，结构如图 2 所示。

与 SAM1 中使用的纯 ViT 编码器相比，Hiera 采用分层结构，可以捕获多尺度特征，更适合设计 U

型网络。具体来说，给定一个输入图像，Hiera 将输出四个层次特征 F_i，输出如式(1)所示：

$$F_{SAM\,2}^{i} = \mathbb{R}^{C_i \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}} \tag{1}$$

在 CNN 分支采用 ResNet-50 网络捕获局部细节特征。处理流程如下：对于输入图像，首先经过一个 7×7 初始卷积层，生成第一层特征图。接着执行一次 3×3 最大池化(Max Pooling)，将特征图尺寸进一步缩小。随后，对特征图依次通过 3 次、4 次、6 次、3 次 Bottleneck 残差块处理，残差块堆叠逐层提取高层次特征，输出用于后续全局信息融合。每一层的计算方法如公式(2)表示：

$$F^{i}{}_{CNN} = RL_i\left(X\right) \in \mathbb{R}^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times 2^{i-1}C} \tag{2}$$

其中，$\mathbb{R}^{H\times W\times 3}$ 为输入图像，$RL(\cdot)$ 表示在第 i 层的 CNN 分支特征。

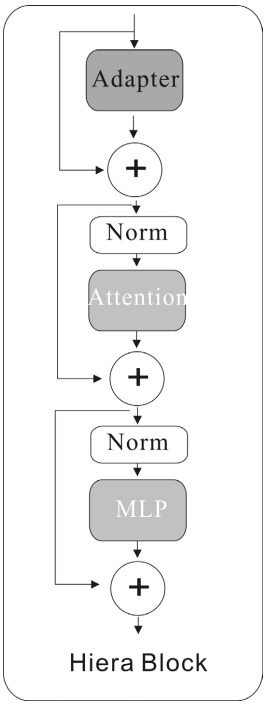

Figure 2. Hiera structure
图 2. Hiera 结构图

2.2. 级联特征解码器

解码器部分利用注意力门(Attention Gate, AG)机制，将上一层解码器的上采样特征与当前层通道注意力模块的输出相结合，实现多层特征的融合。然后，使用卷积注意(CAM)模块对融合后的特征进行处理，增强特征表达的能力。接着，将每个 CAM 模块的输出通过卷积层发送到预测端，进行初步预测。最终，将来自四个不同预测头的结果进行汇总，产生最终的分割图。这样，通过多层次特征的融合与逐层精细处理，模型能够生成更加精准的分割结果。

注意力门(AG)是一种用于选择性关注重要区域并抑制不相关背景信息的机制，结构如图 3 所示。基于输入图像的空间信息逐步抑制不相关背景区域的特征响应，从而增强模型对目标区域的感知能力。其核心思想是通过注意力机制对不同区域进行加权，使模型能够专注于对分割任务有用的特征，而忽略对结果无关或干扰的部分。这在处理复杂背景或者多目标场景时尤为有效。

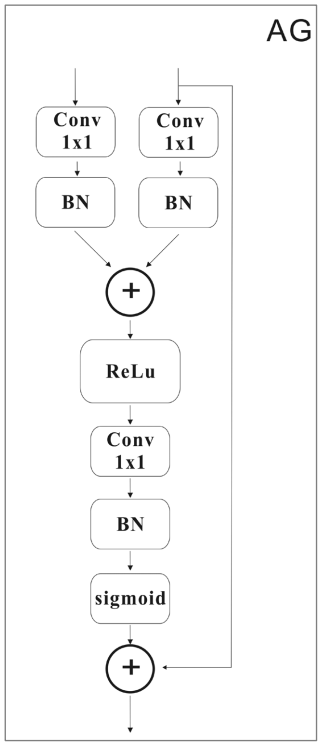

Figure 3. AG structure
图 3. AG 结构图

注意门通常由两个输入分支组成：一个是当前层的通道注意力输出，另一个是来自上一层的特征图，由公式(3) (4)给出：

$$q_{att}(g,x)=\sigma_1\left(BN\left(C_g(g)+BN\left(C_x(x)\right)\right)\right) \tag{3}$$

$$AG(g,x)=x*\sigma_2\left(BN\left(C\left(q_{att}(g,x)\right)\right)\right) \tag{4}$$

这里 $\sigma_1(\cdot)$ 表示 ReLu 激活函数，$\sigma_2(\cdot)$ 表示 Sigmoid 激活函数，$C_g(\cdot)$、$C_x(\cdot)$、$C(\cdot)$ 表示卷积操作，$BN(\cdot)$ 为 Batch Normalization 操作，g 和 x 分别表示上一层上采样后的特征和通道注意力输出的特征。

卷积注意模块(CAM)通过结合卷积与注意力机制来细化特征映射，增强模型对关键区域的关注。CAM 模块主要由通道注意力(CA)、空间注意力(SA) [15]和卷积模块组成，共同提升特征的表达能力，结构图如图 4 所示。

CA 模块通过在高度和宽度两个空间维度上分别进行注意力计算，能够更精确地捕获图像中的空间分布特征，从而实现对特征间依赖关系的全面把握。在特征提取上，为缓解二维空间在全局池化上容易造成的位置信息丢失，CA 将通道注意分解为两个并行的单维特征，有效地将空间坐标信息整合到生成的注意力图中。如式(5)所示：

$$CA(x)=\sigma_2\left(C_2\left(\sigma_1\left(C_1(P_m(x))\right)\right)+C_2\left(\sigma_1\left(C_1\left(P_a(x)\right)\right)\right)\otimes x \tag{5}$$

SA 的计算包括全局平均池化和最大池化，分别提取全局的平均信息和最大值信息，并将这些信息进行融合，生成注意力权重。通过这种方式，空间注意力模块能够动态调整每个位置的响应，提升特征图在空间上的辨别力。其计算公式如式(6)所示：其中 $\sigma(\cdot)$ 为 Sigmoid 激活函数，$C_m(\cdot)$ 和 $C_a(\cdot)$ 分别表示所取得最大值和平均值，$C(\cdot)$ 是一个 7×7 且 padding 为 3 的卷积层。

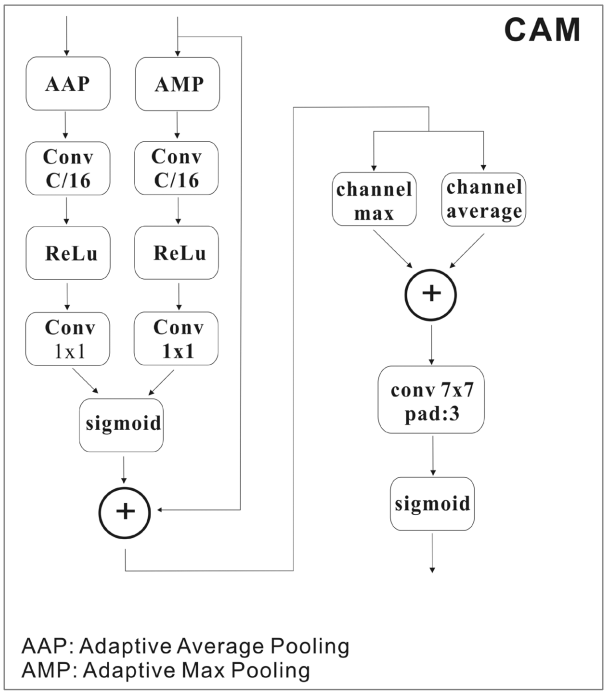

Figure 4. CAM structure
图 4. CAM 结构图

$$SA(x)=\sigma\left(C\left(C_m(x)+C_a(x)\right)\right)\otimes x \tag{6}$$

其中 $\sigma(\cdot)$ 为 Sigmoid 激活函数，$C_m(\cdot)$ 和 $C_a(\cdot)$ 分别表示所取得最大值和平均值，$C(\cdot)$ 是一个 7×7 且 padding 为 3 的卷积层。

卷积块是 CAM 模块中另一个关键部分，通常由一系列卷积层、归一化和非线性激活函数组成。ConvBlock 的主要任务是进一步提取和细化特征图中的局部信息。卷积层通过滑动窗口的方式，捕获局部的空间特征和模式，而归一化则能够稳定训练过程，加速网络收敛。其表达式如式(7)所示：

$$\text{ConvBlock}(x)=\sigma\left(\text{BN}\left(C\left(\sigma\left(\text{BN}\left(C(x)\right)\right)\right)\right)\right) \tag{7}$$

其中 $\sigma(\cdot)$ 为 ReLu 激活函数，$\text{BN}(\cdot)$ 为 Batch normalization 层，$C(\cdot)$ 是一个 3×3 卷积层。

在网络的最后阶段，模型通过对不同层次的特征图分别进行 4 倍、8 倍、16 倍和 32 倍上采样，统一尺寸为原始图片，再将四个预测结果融合来生成最终的预测图。这一过程通过加性运算(element-wise addition)来综合来自不同分辨率和层次的预测结果。通过对四个预测结果进行加性运算，模型能够将不同层次的特征信息有效融合，生成包含丰富细节和语义信息的最终预测图。如式(8)所示：

$$\text{output}=\sigma\left(p1+p2+p3+p4\right) \tag{8}$$

其中，$p1$、$p2$、$p3$、$p4$ 分别为四个预测头生成的特征映射。通过逐元素加法将这四个预测结果进行融合，相加后的结果能够综合来自不同层次的特征信息。接着，对融合后的结果应用 Sigmoid 激活函数，从而生成最终的预测输出图。

3. SAM2U-Net 模型参数设置与实验结果分析

为了验证本模型的有效性，将本文所提模型分别在甲状腺数据集和息肉分割数据集与其他先进模型

进行对比实验。以下部分为数据集进行描述。1) DDTI [16]：包含来自 299 名患者的 347 组甲状腺超声图像，并由放射科专家在 XML 文件中对可疑甲状腺病变进行了详细注释和诊断描述。2) TG3K [17]：包括 3493 张带有像素标签的超声图像，其中包含来自各种设备和视图的高质量结节掩膜标记。3) Kvasir-SEG [18]：针对结肠息肉像素级分割的内窥镜数据集，包括来自结肠镜检查视频序列的 1000 个息肉图像及其标签。4) CVC-ClinicDB [19]：包含来自 29 个不同序列的结肠镜检查视频的 612 张图像。5) ColonDB [20]：ColonDB 数据集强调了多样性和真实世界的挑战，包括不同类型的息肉和各种内镜检查条件。6) ETIS [21]：数据集提供了热成像图像和相应的息肉标注，总数为 196 张。为了进行模型训练和评估，数据集被划分为训练集、验证集和测试集，分别占总数据的 70%、20%和 10%，确保了模型性能的充分验证与测试。

3.1. SAM2U-Net 模型参数设置与性能评估指标

本文采用 Adam 优化器[22]对网络模型进行优化，Adam 中的权重衰减(weight decay)机制有助于防止模型过拟合。在医学图像分割中，由于医学图像的多样性和复杂性，模型很容易过度学习训练数据中的噪声或特定样本的特征，导致在测试数据上性能下降。初始学习率设置为 0.0001，权重衰减率为 0.1，批量大小为 8。选择 Dice Loss [23]作为损失函数，因其在医学图像分割任务中表现出色，能够有效处理前景和背景不平衡问题。所用模型及所有对比模型均基于 PyTorch1.13.1 实现，并在 Windows 11 环境下训练，使用 NVIDIA RTX3090 GPU (24GB 显存)加速计算。

多项深度学习网络的测试结果表明，当训练轮次(epoch)设置为 50 时，最终获得的权重在测试集上能够实现最佳预测精度。更多的训练轮次可能导致网络过度拟合，而少于 50 轮次则可能导致训练不足。图 5 展示了本文模型在甲状腺数据集 DDTI 上的训练损失(Loss)曲线，其中横轴表示训练的轮次，纵轴表示对应的损失值。因此，本文所有实验均将训练轮次设置为 50。

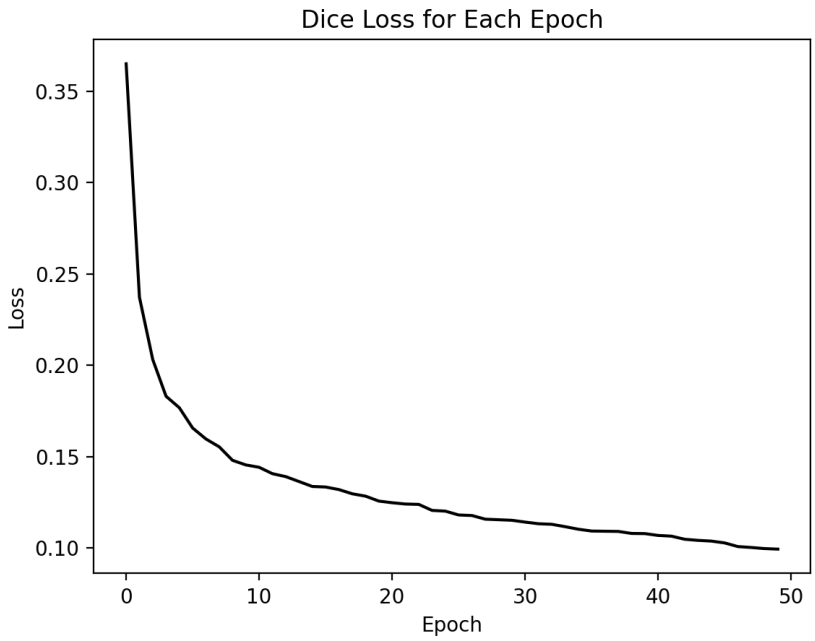

Figure 5. Dice loss for 50 epoch
图 5. 50 epoch 时 dice loss 曲线

本文使用 Dice 系数[24]、交并比(intersection over Union, IoU) [25]评估指标来评估本模型的性能。Dice 系数是一种集合相似度度量，用于计算分割结果与标签之间的相似性，能够有效测量两个边界

的重叠度。其计算公式如式(9)所示。此外，生成的预测图像与标签之间的重叠比率称为交并比(IoU)，即它们的交集与并集的比率。理想情况下，当预测图像与标签完全重叠时，Dice 系数和 IoU 的值均为 1，表示完美的分割效果。计算公式如式(10)所示。

$$\mathrm{Dice}=\frac{2\left|X\cap Y\right|}{\left|X\right|+\left|Y\right|} \tag{9}$$

$$\mathrm{IoU}=\frac{\left|X\cap Y\right|}{\left|X\cup Y\right|} \tag{10}$$

其中，$X\cap Y$ 表示 X 和 Y 样本之间的交集，$\left|\cdot\right|$ 表示元素的数量。

3.2. 分割模型性能对比分析

在本节中，我们对几种先进的图像分割方法进行了全面评估，以验证我们模型的性能。表 1 展示了在甲状腺 DDTI 数据集与 TG3K 数据集上的实验结果。

Table 1. Results of different models in the thyroid nodule segmentation task
表 1. 不同模型在甲状腺结节分割任务中的结果

Methods	DDTI		TG3K	
	mDice	mIoU	mDice	mIoU
U-Net	0.870	0.772	0.702	0.611
DeepLabv3 [26]	0.878	0.785	0.736	0.636
DeepLabv3+ [27]	0.887	0.708	0.747	0.652
CCSDG [28]	0.873	0.775	0.740	0.649
META-Unet [29]	0.896	0.812	0.739	0.656
SAM2U-Net	**0.911**	**0.832**	**0.779**	**0.683**

Table 2. Results of different models in the polyp segmentation task
表 2. 不同模型在息肉分割任务中的分割结果

Methods	Kvasir		ClinicDB		mDice		mIoU	
	mDice	mDice	mDice	mIoU	mDice	mIoU	mDice	mIoU
U-Net	0.818	0.746	0.823	0.755	0.504	0.436	0.398	0.335
UNet++	0.821	0.744	0.794	0.729	0.482	0.408	0.401	0.344
MSEG+ [30]	0.897	0.839	0.909	0.864	0.735	0.666	0.700	0.630
SANet [31]	0.904	0.847	0.906	0.859	0.752	0.669	0.750	0.654
MSNet [32]	0.905	0.849	0.908	0.869	0.747	0.668	0.720	0.650
SSFormer [33]	0.917	0.864	0.906	0.855	0.802	0.721	0.805	0.720
CFA-Net [34]	0.915	0.861	**0.933**	**0.883**	0.743	0.665	0.832	0.655
SAM-Path [35]	0.828	0.730	0.750	0.644	0.632	0.516	0.844	0.756
Yolo-SAM	0.852	0.742	0.895	0.810	0.893	0.808	0.933	0.875
SAM2U-Net	**0.928**	**0.879**	0.925	0.876	**0.904**	**0.842**	**0.937**	**0.882**

我们的 SAM2U-net 模型与 5 个公开可用的先进甲状腺结节分割方法进行了定量比较。结果显示，SAM2U-Net 在 Dice 和 IoU 指标上均优于其他模型。特别是在 TG3K 数据集上，与最先进的 META-Unet

相比，我们的模型在 Dice 指标上提升了 4.0% (从 0.739 提升到 0.779)，在 mIoU 上提升了 2.7% (从 0.656 提升到 0.683)。这一显著提升归因于模型独特的并行编码器结构，有效结合了局部细节和全局信息。

表 2 展示了我们提出的 SAM2U-Net 模型与其他先进方法在息肉分割任务上的性能比较。结果表明，SAM2U-Net 在四个广泛使用的数据集中的三个都展现出了卓越的性能。在 Kvasir 数据集上，SAM2U-Net 达到了 0.928 的 mDice 和 0.879 的 mIoU，这两个指标都优于所有比较方法。在 ColonDB 和 ETIS 数据集上，分别为 0.904 和 0.937 的 mDice，均为最佳的性能表现。在 ClinicDB 数据集上，达到了 0.925 的 mDice 和 0.876 的 mIoU。虽然在这个数据集上 CFA-Net 的 mDice 略高(0.933)，但 SAM2-Net 仍然优于大多数其他方法，显示了其在不同数据集上的稳定性和泛化能力。

SAM2-Net 的优异表现来源于其创新的网络架构，该架构能有效地结合局部细节和全局上下文信息。特别是在处理复杂或不规则形状的息肉时，模型展现出了明显的优势。

3.3. 消融实验

为了评估 Hiera 骨干网络尺寸对 SAM2U-Net 性能的影响，我们进行了一系列消融实验。表 3 展示了不同 Hiera 版本在 DDTI 数据集上的分割结果以及模型的网络计算量。

Table 3. Segmentation results of different Hiera models
表 3. 不同 Hiera 分割结果

Backbones	DDTI		complexity	
	mDice	mIoU	Params	FLOPs
Hiera-Tiny	0.879	0.786	76.2M	7.1G
Hiera-Small	0.893	0.809	76.4M	8.6G
Hiera-Base+	0.911	0.832	77.1M	14.6G
Hiera-Large	0.897	0.813	79.5M	44.5G

实验结果清楚地表明，随着骨干网络尺寸的增加，模型性能普遍提高。具体来说：在 DDTI 数据集上：mDice 从 Hiera-Tiny 的 0.879 提升到 Hiera-base+的 0.911，提高了 3.2 个百分点。mIoU 从 0.786 提升到 0.831，提高了 4.5 个百分点。当骨干网络尺寸进一步增大时，并没有取得更佳的分割效果，但计算成本(FLOPs)却显著增加。

4. 结论

本研究构建了新型医学图像分割架构 SAM2U-Net，创新性地整合了卷积神经网络(CNN)与 Segment Anything Model 2 (SAM2)的多分辨率分层架构。该模型通过 CNN 模块实现病灶区域局部纹理特征的精确提取，同时利用 SAM2 的多尺度感知机制完成全局语义特征的鲁棒表征，二者的协同作用显著提升了复杂解剖结构及异质性病变区域的分割精度。特别地，引入的自适应通道注意力模块实现了局部–全局特征的动态权重分配，通过特征通道的显著性重标定优化了多源信息的融合效能。

基于多中心医学影像数据集的验证结果表明，SAM2U-Net 在分割性能指标方面展现出显著优势。在 DDTI 基准数据集上的实验验证表明，该模型在 Dice 系数(0.911)和交并比(0.832)等核心指标上均显著超越现有先进方法，其中 Dice 系数较最优模型提升 1.5 个百分点。跨数据集验证实验进一步证实，该架构对多种医学影像均保持稳定的分割性能，其创新性的特征融合机制有效解决了传统方法在复杂背景下存在的边界模糊和区域漏分问题。该技术突破为临床医学影像的自动化分析提供了新的范式，其高精度分割能力可有效降低人工判读的主观偏差，为临床诊疗决策优化提供可靠依据。

参考文献

[1] Otsu, N. (1979) A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, **9**, 62-66. https://doi.org/10.1109/tsmc.1979.4310076

[2] Zhang, Y. (2006) An Overview of Image and Video Segmentation in the Last 40 Years. In: Zhang, Y.J., Ed., *Advances in Image and Video Segmentation*, IGI Global, 1-16. https://doi.org/10.4018/978-1-59140-753-9.ch001

[3] Khan, J.F., Bhuiyan, S.M.A. and Adhami, R.R. (2011) Image Segmentation and Shape Analysis for Road-Sign Detection. *IEEE Transactions on Intelligent Transportation Systems*, **12**, 83-96. https://doi.org/10.1109/tits.2010.2073466

[4] Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**, 2278-2324. https://doi.org/10.1109/5.726791

[5] Fu, Z., Li, J. and Hua, Z. (2022) Deau-net: Attention Networks Based on Dual Encoder for Medical Image Segmentation. *Computers in Biology and Medicine*, **150**, Article ID: 106197. https://doi.org/10.1016/j.compbiomed.2022.106197

[6] 司明明, 陈玮, 胡春燕, 等. 融合 Resnet50 和 U-Net 的眼底彩色血管图像分割[J]. 电子科技, 2021, 34(8): 19-24.

[7] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W. and Frangi, A., Eds., *Medical Image Computing and Computer-Assisted Intervention—MICCAI* 2015, Springer, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[8] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., *et al*. (2023) Segment Anything. 2023 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 3992-4003. https://doi.org/10.1109/iccv51070.2023.00371

[9] Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., *et al*. (2024) Sam 2: Segment Anything in Images and Videos. arXiv: 2408.00714.

[10] Chen, T., Zhu, L., Ding, C., Cao, R., Wang, Y., Zhang, S., *et al* (2023) SAM-Adapter: Adapting Segment Anything in Underperformed Scenes. 2023 *IEEE*/*CVF International Conference on Computer Vision Workshops* (*ICCVW*), Paris, 2-6 October 2023, 3359-3367. https://doi.org/10.1109/iccvw60793.2023.00361

[11] Huang, D., Xiong, X., Ma, J., Li, J., Jie, Z., Ma, L., *et al*. (2024) AlignSAM: Aligning Segment Anything Model to Open Context via Reinforcement Learning. 2024 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 3205-3215. https://doi.org/10.1109/cvpr52733.2024.00309

[12] Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X. and Shen, C. (2024) Matcher: Segment Anything with One Shot Using All-Purpose Feature Matching. ICLR.

[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Deh-ghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2021) An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. ICLR.

[14] Ravi, N., Gabeur, V., Hu, Y.T., *et al*. (2024) Sam 2: Segment Anything in Images and Videos. arXiv preprint arXiv:2408.00714.

[15] Woo, S., Park, J., Lee, J. and Kweon, I.S. (2018) CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018, Springer, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[16] Vasques, B.I., Pereira, J.M., Santos, A. and Neves, L.A. (2017) A Public Dataset for Thyroid Ultrasound Image Analysis. *Proceedings of the IEEE International Conference on Image Processing* (*ICIP*), Beijing, 17-20 September 2017.

[17] Gong, H., Chen, G., Wang, R., Xie, X., Mao, M., Yu, Y., *et al*. (2021) Multi-Task Learning for Thyroid Nodule Segmentation with Thyroid Region Prior. 2021 *IEEE* 18*th International Symposium on Biomedical Imaging* (*ISBI*), Nice, 13-16 April 2021, 257-261. https://doi.org/10.1109/isbi48211.2021.9434087

[18] Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., *et al*. (2019) Kvasir-SEG: A Segmented Polyp Dataset. In: Ro, Y., *et al.*, Eds., *MultiMedia Modeling*, *MMM* 2020, Springer, 451-462. https://doi.org/10.1007/978-3-030-37734-2_37

[19] Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C. and Vilariño, F. (2015) WM-DOVA Maps for Accurate Polyp Highlighting in Colonoscopy: Validation Vs. Saliency Maps from Physicians. *Computerized Medical Imaging and Graphics*, **43**, 99-111. https://doi.org/10.1016/j.compmedimag.2015.02.007

[20] Tajbakhsh, N., Gurudu, S.R. and Liang, J. (2016) Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. *IEEE Transactions on Medical Imaging*, **35**, 630-644. https://doi.org/10.1109/tmi.2015.2487997

[21] Silva, J., Histace, A., Romain, O., Dray, X. and Granado, B. (2013) Toward Embedded Detection of Polyps in WCE Images for Early Diagnosis of Colorectal Cancer. *International Journal of Computer Assisted Radiology and Surgery*, **9**, 283-293. https://doi.org/10.1007/s11548-013-0926-3

[22] Loshchilov, I. and Hutter, F. (2019) Decoupled Weight Decay Regularization. *International Conference on Learning*

Representations (*ICLR*), New Orleans, 6-9 May 2019.

[23] Milletari, F., Navab, N. and Ahmadi, S. (2016) V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 2016 *Fourth International Conference on* 3*D Vision* (3*DV*), Stanford, 25-28 October 2016, 565-571. https://doi.org/10.1109/3dv.2016.79

[24] Dice, L.R. (1945) Measures of the Amount of Ecologic Association between Species. *Ecology*, **26**, 297-302. https://doi.org/10.2307/1932409

[25] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J. and Zisserman, A. (2009) The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, **88**, 303-338. https://doi.org/10.1007/s11263-009-0275-4

[26] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L. (2017) Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv: 1706.05587.

[27] Chen, L., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. (2018) Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018., Springer, 833-851. https://doi.org/10.1007/978-3-030-01234-2_49

[28] Zhang, Y., Zhang, Q., Wang, X. and Zhang, Y. (2022) Devil Is in Channels: Contrastive Single Domain Generalization for Medical Image Segmentation. arXiv: 2209.07211.

[29] Trinh, Q.H. (2023) Meta-Polyp: A Baseline for Efficient Polyp Segmentation. arXiv: 2305.07848.

[30] Huang, C.H., Wu, H.Y. and Lin, Y.L. (2021) HarDNet-MSEG: A Simple Encoder-Decoder Polyp Segmentation Neural Network that Achieves over 0.9 Mean Dice and 86 FPS. arXiv: 2101.07172.

[31] Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K. and Cui, S. (2021) Shallow Attention Network for Polyp Segmentation. In: de Bruijne, M., *et al.*, Eds., *Medical Image Computing and Computer Assisted Intervention—MICCAI* 2021, Springer, 699-708. https://doi.org/10.1007/978-3-030-87193-2_66

[32] Zhao, X., Zhang, L. and Lu, H. (2021) Automatic Polyp Segmentation via Multi-Scale Subtraction Network. In: de Bruijne, M., *et al.*, Eds., *Medical Image Computing and Computer Assisted Intervention—MICCAI* 2021, Springer, 120-130. https://doi.org/10.1007/978-3-030-87193-2_12

[33] Wang, J., Huang, Q., Tang, F., Meng, J., Su, J. and Song, S. (2022) Stepwise Feature Fusion: Local Guides Global. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S. and Li, S., Eds., *Medical Image Computing and Computer Assisted Intervention—MICCAI* 2022, Springer, 110-120. https://doi.org/10.1007/978-3-031-16437-8_11

[34] Zhou, T., Zhou, Y., He, K., Gong, C., Yang, J., Fu, H., *et al*. (2023) Cross-Level Feature Aggregation Network for Polyp Segmentation. *Pattern Recognition*, **140**, Article ID: 109555. https://doi.org/10.1016/j.patcog.2023.109555

[35] Zhang, J., Ma, K., Kapse, S., Saltz, J., Vakalopoulou, M., Prasanna, P., *et al*. (2023) SAM-Path: A Segment Anything Model for Semantic Segmentation in Digital Pathology. In: Celebi, M.E., *et al*., Eds., *Medical Image Computing and Computer Assisted Intervention—MICCAI* 2023 Workshops, Springer, 161-170. https://doi.org/10.1007/978-3-031-47401-9_16