Beyond Adapting SAM: Towards End-to-End Ultrasound Image Segmentation via Auto Prompting

摘要

端到端的医学图像分割对于以任务特定模型为主的计算机辅助诊断具有重要价值，但通常泛化性较差，对于不同的数据集需要定制模型，这在临床应用上带来了很大的不便。随着最近用于通用图像分割的分段任何模型（SAM）所带来的突破，人们已经做出了广泛的努力来使 SAM 适应医学成像，但效果仍然不尽如人意。SAMUS 是一种专为超声图像分割而定制的通用模型，并进一步使其能够以端到端的方式工作，称为 AutoSAMUS。在 SAMUS 中，引入了并行 CNN 分支，通过跨分支注意力来补充局部信息，并联合使用特征适配器和位置适配器来使 SAM 从自然域适应到超声域，同时降低训练复杂度。AutoSAMUS 是通过引入自动提示生成器（APG）来代替 SAMUS 的手动提示编码器来自动生成提示嵌入来实现的。此外，我加入时空建模的方法，增加时序信息，从而更好的指导分割，提高分割结果。实验证明了 SAMUS 和 AutoSAMUS 在医学图像分割方面比原 SAM 模型有更好的性能。并通过在超声心动图数据集上的实验评估证明，SAMUS 相比其他先进的特定任务和基于 SAM 的基础模型的优越性。相信基于 SAM 的自动提示模型有潜力成为端到端医学图像分割的新范例，值得更多探索。

关键词：SAM 模型；自动提示；医学图像分割；超声心动图；SAMUS 模型

1 引言

医学图像分割是识别和突出医学图像中特定器官、组织和病变的关键技术，是计算机辅助诊断系统的组成部分 [9]。人们已经提出了许多用于医学图像分割的深度学习模型，展示了巨大的潜力 [5,14]。然而，这些模型是针对特定对象量身定制的，应用于其他对象时需要训练新的模型参数，给任务多样化的临床应用带来很大不便。

分割任何模型（SAM）作为视觉分割的通用基础模型，由于其跨不同对象的卓越分割能力和强大的零样本泛化能力而赢得了相当多的赞誉 [7]。根据用户提示，包括点、边界框和粗掩模，SAM 能够分割相应的对象。因此，通过简单的提示，SAM 可以毫不费力地适应各种分段应用。这种范例能够将多个单独的医学图像分割任务集成到一个统一的框架（即通用模型）中，极大地促进了临床部署 [6]。

尽管构建了迄今为止最大的数据集 SA-1B，但由于缺乏可靠的临床注释，SAM 在医学领域的性能迅速下降 [6]。目前已经提出了一些基础模型，通过在医学数据集上调整 SAM 来使

SAM 适应医学图像分割 [10,15]。然而，与 SAM 相同，它们在特征建模之前对输入图像执行无重叠的 16× 标记化，这破坏了对于识别小目标和边界至关重要的局部信息，使得它们难以分割具有复杂/线状形状的临床对象，边界弱、尺寸小或对比度低。此外，这些基于 SAM 的模型需要手动提供与任务相关的提示来生成相应的掩模，从而形成半自动分割流程。在处理某些临床任务时，这种范式不够灵活。

具体来说，SAMUS 继承了 SAM 的 ViT 图像编码器、提示编码器和掩模解码器，并对图像编码器进行了量身定制的设计。首先，通过减少所需的输入大小来缩短 ViT 分支的序列长度，以降低计算复杂度。然后，开发了一个特征适配器和一个位置适配器来对 ViT 图像编码器从自然领域到医学领域进行微调。为了补充 ViT 图像编码器中的本地信息，还引入了一个并行 CNN 分支图像编码器，与 ViT 分支一起运行，并提出了一个跨分支注意模块来启用 ViT 分支中的每个补丁从 CNN 分支吸收本地信息。并且，在 SANUS 的基础上，引入具有可学习任务标记的自动提示生成器（APG），以取代 SAMUS 的手动提示编码器来生成与任务相关的提示嵌入，称为 AutoSAMUS。

2 相关工作

2.1 将 SAM 应用于医学图像分割

SAM 在自然图像中表现出了卓越的性能，但在一些医学图像分割任务中表现不佳，尤其是在形状复杂、边界模糊、尺寸小或对比度低的物体上 [6]。为了弥补这一差距并使 SAM 能够有效地适应医学图像领域，人们提出了几种将视觉调整技术应用于 SAM 的方法。具体来说，MedSAM 通过冻结提示编码器来在医学图像上训练 SAM，重点是调整图像编码器和掩模解码器 [10]。SAMed 在图像编码器上应用基于低秩的策略，以较低的计算成本调整 SAM，使其更适合医学图像分割 [18]。MSA 在 ViT 图像编码器和掩模解码器上采用 down-ReLU-up 适配器来引入医疗信息 [15]。与当前基于 SAM 的通用模型相比，所提出的 SAMUS 更注重补充局部特征并实现端到端自动分割。

2.2 SAM 中的提示

Vanilla SAM 在精确的空间提示（例如点、边界框和掩模）的驱动下生成与任务相关的掩模。为了自动获取这些空间提示，一些方法引入了单独的输入相关网络。具体来说，Adapter-Shadow、SAC 和 UV-SAM 分别使用 EfficientNet、U-Net 和 SegFormer 来生成用于制作空间提示的粗略掩模 [2,11,19]。这种单独网络引入的参数与特定于任务的方法处于同一水平，使得通用模型变得麻烦。Polyp-SAM++ 使用 Grounding DINO 从文本提示生成边界框提示 [13]。自适应 SAM 和 SP-SAM 通过 CLIP 将文本提示编码为提示嵌入 [16,17]。尽管这些方法可以有效地利用文本信息，但医学场景中缺乏文本图像数据来调整在自然场景上训练的文本编码器。SurgicalSAM 提出了一种基于原型的类提示编码器来生成密集和稀疏提示嵌入 [17]。自动提示 SAM 通过构建 Up-Down 全卷积层开发了自动提示编码器 [8]。这些提示编码器与掩码解码器深度耦合，导致难以通过基于 SAM 的模型的多目标学习构建鲁棒的特征表示。相比之下，所提出的 APG 是一个轻量级且独立的模块，并且高度可扩展至其他基于 SAM 的基础模型。

2.3 时空建模的方法

主流视频时间建模方法包括多帧聚合和时空记忆网络。多帧聚合通过聚合相邻帧的语义信息来学习时间特征。相比之下，时空记忆通过沿时间维度传播语义信息来对视频时间信息进行建模。虽然多帧聚合被广泛使用，但其 GPU 内存需求随着视频长度的增加而迅速增加，限制了其在长视频处理中的应用。相比之下，时空记忆网络可以在保证时间建模的同时显着减少内存消耗，使其更适合扩展到医学视频分析等领域。时空记忆网络（STM）首先由 Oh 等人提出 [12] 用于视频对象分割任务。随后的方法，包括 STCN [4]、XMem [3] 和 XMem++ [1]，已经证明了通用视频分割的巨大潜力。然而，这些方法需要一个带注释的参考关键帧来对视频进行分割，这对超声心动图的任务来说很困难。

3 本文方法

3.1 SAMUS 模型概述

如图 1所示：SAMUS 的整体架构继承自 SAM，保留了提示编码器和掩码解码器的结构和参数。相比之下，图像编码器经过精心修改，以解决局部特征不足和计算内存消耗过多的挑战，使其更加适合临床。主要修改包括减少所需的输入大小、重叠补丁嵌入、向 ViT 分支引入适配器、添加 CNN 分支以及引入跨分支注意力。具体来说，输入空间分辨率从 1024 × 1024 像素缩小到 256 × 256 像素，由于 Transformer 中的输入序列较短，因此 GPU 内存成本大幅降低。重叠的补丁嵌入使用与 SAM 中的补丁嵌入相同的参数，而其补丁步长是原始步长的一半，很好地保留了补丁边界的信息。ViT 分支中的适配器包括一个位置适配器和五个特征适配器。由于输入大小较小，位置适配器用于适应较短序列中的全局位置嵌入。第一个特征适配器遵循重叠补丁嵌入，将输入特征与预训练 ViT 图像编码器所需的特征分布对齐。其余的特征适配器连接到全局变换器中前馈网络的剩余连接，以微调预训练的图像编码器。就 CNN 分支而言，它与 ViT 分支并行，通过 CBA 模块为后者提供补充的局部信息，CBA 模块以 ViT 分支特征作为查询，并与 CNN 分支的特征建立全局依赖关系。需要注意的是，CBA 仅集成到每个全局变压器中。最后，将两个分支的输出组合起来作为 SAMUS 的最终编码图像嵌入 Fi。

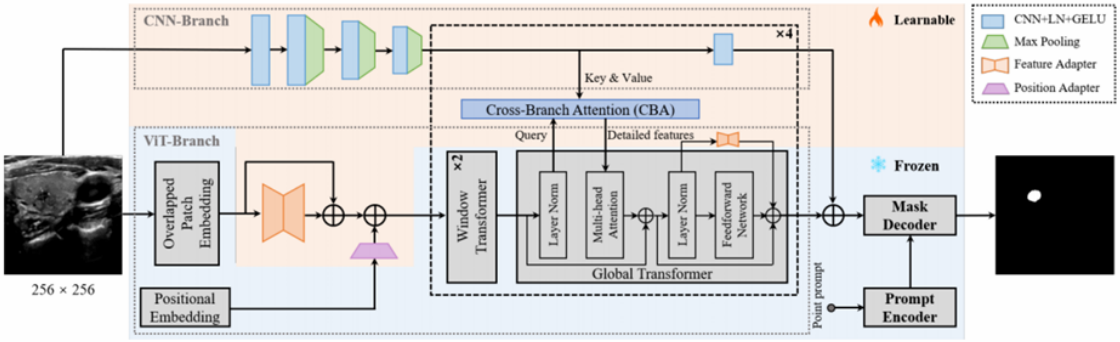

图 1. SAMUS 模型框架示意图

3.2 ViT 分支中的适配器

为了促进 SAM 的训练图像编码器（即 ViT 分支）泛化到较小的输入尺寸和医学图像领域，SAMUS 引入了一个位置适配器和五个特征适配器。这些适配器可以有效地调整 ViT 分支，同时只需要更少的参数。具体来说，位置适配器负责调整位置嵌入以匹配嵌入序列的分辨率。它首先通过最大池化对位置嵌入进行下采样，步长和内核大小为 2，实现与嵌入序列相同的分辨率。然后，应用内核大小为 3 × 3 的卷积运算来调整位置嵌入，进一步帮助 ViT 分支更好地处理较小的输入。所有特征适配器都具有相同的结构，包括三个组件：向下线性投影、激活函数和向上线性投影。每个特征适配器的流程可以表述为：

$$\mathcal{A}(x) = \mathcal{G}\left(xE_d\right)E_u \tag{1}$$

其中 $\mathcal{G}$ 表示 GELU 激活函数，$E_d \in R^{d\times\frac{d}{4}}$ 和 $E_u \in R^{\frac{d}{4}\times d}$ 是投影矩阵，d 是特征嵌入的维度。通过这些简单的操作，特征适配器使 ViT 分支能够更好地适应医学图像域的特征分布。

3.3 AutoSAMUS

AutoSAMUS 是在 SAMUS 基础上扩展的端到端自动提示框架,通过用 APG 替代 SAMUS 的手动提示编码器来实现。如图 2 所示，APG 的输入由输出标记 $T_o \in R^{5\times d}$ 和图像嵌入 F_i 组成，其中 T_o 是从掩模解码器中提取的冻结参数，F_i 是图像编码器的输出。为了指示分割任务，APG 引入了可学习的任务标记 $T_t \in R^{k\times d}$，用于自动生成与任务相关的提示嵌入，其中 k 是任务标记的数量。首先，使用交叉注意力和多层感知器（MLP）的组合来耦合任务标记和输出标记，公式为：

$$\mathcal{C}\left(T_t, T_o\right) = \textbf{MLP}\left(\sigma\left(\frac{T_tW_q\left(T_oW_k\right)^T}{\sqrt{d}}\right)\left(T_oW_v\right)\right), \tag{2}$$

其中 MLP 由两个线性层组成。W_q、W_k 和 $W_v \in R^{d\times d}$ 是可学习的权重矩阵。然后，更新后的任务标记表示为：

$$T_{t_1} = \mathcal{C}\left(\mathcal{C}\left(T_t, T_o\right), \mathcal{C}\left(T_o, T_t\right)\right)W + T_t \tag{3}$$

其中 W $\in R^{d\times d}$ 是投影矩阵。类似地，在公式 3 中通过交换等式 T_t 和 T_o 的位置，可以计算出更新后的输出标记 T_{o_1}。接下来，为了使图像嵌入适应任务域并使任务标记了解图像信息，我们在图像嵌入和组合标记之间执行公式 3 中定义的组合操作 $\mathcal{C}$, T = [T_{t_1} , T_{o_1}]。之后，更新的图像嵌入和组合标记表示为 $F_i' = \mathcal{C}(\mathcal{C}(F_i, T), \mathcal{C}(T, F_i))$ 和 $T' = \mathcal{C}(\mathcal{C}(T, F_i), \mathcal{C}(F_i, T))$ 。最后，基于 T' 和 F_i'，APG 生成稀疏提示嵌入 $P_s = T'[:k, :]$ 和密集提示嵌入 $P_d = M(F_i')$ 以提示冻结掩码解码器，其中 M 表示由通道为 $\frac{d}{4}$, $\frac{d}{4}$, $\frac{d}{4}$ 和 d 的四个单卷积块组成的操作序列。此外，更新后的图像嵌入 F_i' 将替代原始图像嵌入参与掩模解码。

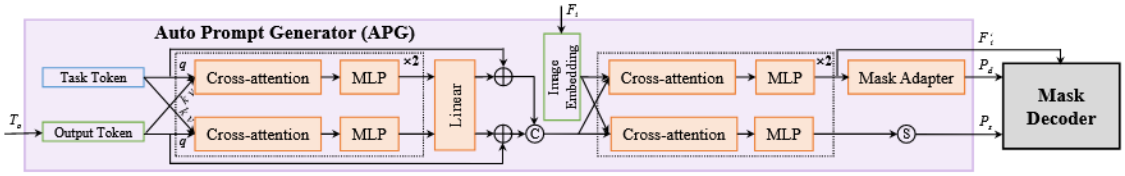

图 2. 自动提示生成器示意图

4 复现细节

4.1 与已有开源代码对比

本文的复现过程使用了 SAMUS 原论文中的模型的框架代码作为基础框架，在此基础上进行模块创新。SAMUS 原论文的代码可以参考https://github.com/xianlin7/SAMUS 。

4.2 实验环境搭建

本文参照原论文的实验环境搭建步骤进行配置，详情参照代码中的 readme.md 文件，在此不再赘述。

4.3 多尺度特征融合

本文创新点是通过对图片进行多尺度的特征提取，再进行特征融合，从而提取出更多信息。同时，通过时空建模建立时序信息，并将其与 SAMUS 中的 CNN 分支相结合，它用于提供更多的特征信息，从而指导分割。多尺度特征融合的具体过程见如下代码：

```
class Multi_Scale_Feature_Fusion_Module(nn.Module):
    def __init__(self, in_channel, out_channel):
        super(Multi_Scale_Feature_Fusion_Module, self).__init__()
        self.relu = nn.ReLU(True)
        self.branch0 = nn.Sequential(BasicConv2d(in_channel, out_channel, 1))
        self.branch1 = nn.Sequential(BasicConv2d(out_channel, out_channel,
        kernel_size=3, stride=1, padding=6, dilation=6))
        self.branch2 = nn.Sequential(BasicConv2d(out_channel, out_channel,
        kernel_size=3, stride=1, padding=12, dilation=12))
        self.branch3 = nn.Sequential(BasicConv2d(out_channel, out_channel,
        kernel_size=3, stride=1, padding=18, dilation=18))
        self.in_conv = nn.Conv2d(in_channel, out_channel, 1)
    def forward(self, x):
        x0 = self.branch0(x)
        sq_x = self.in_conv(x)
        x1 = self.branch1(sq_x)
        x2 = self.branch2(sq_x + x1)
```

```
x3 = self.branch3(sq_x + x2)
x = self.relu(x0 + x3)
return x
```

4.4 时空建模

涉及论文 idea 实现，暂不开源，请谅解。

5 实验结果分析

在本节中，我们将分析提出的 Multi-scale feature fusion 模块在医学超声图像分割中的有效性。

我们在 CAMUS 数据集上进行消融实验。实验使用 Adam 优化器对网络进行优化。初始学习率设置为 0.0005，批量大小设置为 1，epoch 数设置为 200。使用 dice loss 和 mse loss 作为损失函数进行训练，在 200 轮训练过程中保存 loss 最小的模型用于测试，使用两个个常见的分割评估指标，Dice 分数和 Hausdorff 距离进行测试评估。与 SAMUS 模型本身进行分割性能对比，结果如表 1 所示。

Components	CAMUS	
MSFF	Dice	HD
×	91.13	11.76
✓	92.07	3.47

表 1. 消融实验结果

从表中可以看出，我们的方法比 SAMUS 本身在 Dice 分数上高出约 0.95%，在霍斯多夫距离上有所下降。我们将成功归因于以下原因: 多尺度特征信息的融合使模型增加了对超声图像形状边界等局部信息的关注，针对超声图像对比度低和边界模糊的特点弥补了 Transformer 只能关注全局信息的不足。

6 总结与展望

深度学习在医学图像分割中的应用潜力巨大，但这些模型通常是针对特定类型的医学图像设计的，因此在面对不同数据集时，通常需要对模型结构进行调整以适应数据集的特性，这给临床实践带来了不小的挑战。而视觉大模型（如 SAM）在自然图像分割领域展现了出色的泛化能力，但由于缺乏医学领域的专业知识，在医学图像分割任务中的表现不尽如人意。

SAMUS 在此背景下应运而生，它以大规模图像分割模型——Segmentation Anything Model (SAM) 为基础，探索了一种针对医学图像分割的定制化大规模模型的新研究范式。SAMUS 通过在 SAM 的框架中加入 CNN 分支图像编码器和跨分支注意力模块，弥补了 Transformer 在局部信息提取上的不足。此外，SAMUS 还利用特征适配器和位置适配器对图像编码器进行微调，使其更好地适应下游任务。实验结果表明，SAMUS 相较于原始 SAM 模型，在医学图像分割任务中表现出了更优越的性能，验证了其在医学图像分割中的有效性。进一步地，我

在 SAMUS 的基础上尝试融入超声图像中的多尺度信息，增强模型对局部信息的关注，从而提升了分割效果。

未来，随着技术的不断进步和数据集的日益丰富，各种基于 SAM 的微调模型有望在医学图像分割领域大量涌现，成为医生和医疗研究人员的强大工具，为患者提供更高质量的医疗服务。例如，在手术规划、肿瘤检测、器官定位等临床应用中，基于大模型的分割方法将能够提供更精确和个性化的辅助，帮助医生作出更加准确的诊断与治疗决策。此外，这类模型还可以与临床决策支持系统结合，实现更加智能的医学影像分析，推动个性化和精准医疗的发展。

参考文献

[1] Maksym Bekuzarov, Ariana Bermudez, Joon-Young Lee, and Hao Li. Xmem++: Production-level video segmentation from few annotated frames. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 635–644, 2023.

[2] R. Biswas. Polyp-sam++: Can a text guided sam perform better for polyp segmentation? *arXiv preprint arXiv:2308.06623*, 2023.

[3] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022.

[4] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021.

[5] X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu. Missformer: An effective transformer for 2d medical image segmentation. *IEEE Trans. Med. Imag.*, 42(5):1484–1494, 2022.

[6] Y. Huang and et al. Segment anything model for medical images? *Med. Image Anal.*, 92:103061, 2024.

[7] A. Kirillov and et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[8] C. Li, P. Khanduri, Y. Qiang, R. I. Sultan, I. Chetty, and D. Zhu. Autoprompting sam for mobile friendly 3d medical image segmentation. *arXiv preprint arXiv:2308.14936*, 2023.

[9] X. Liu, L. Song, S. Liu, and Y. Zhang. A review of deep-learning-based medical image segmentation methods. *Sustainability*, 13(3):1224, 2021.

[10] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang. Segment anything in medical images. *Nat. Commun.*, 15(1):654, 2024.

[11] S. Na, Y. Guo, F. Jiang, H. Ma, and J. Huang. Segment any cell: A sam-based auto-prompting fine-tuning framework for nuclei segmentation. *arXiv preprint arXiv:2401.13220*, 2024.

[12] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9226–9235, 2019.

[13] J. N. Paranjape, N. G. Nair, S. Sikder, S. S. Vedula, and V. M. Patel. Adaptivesam: Towards efficient tuning of sam for surgical scene segmentation. *arXiv preprint arXiv:2308.03726*, 2023.

[14] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *MICCAI 2015*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015.

[15] J. Wu and et al. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[16] W. Yue and et al. Part to whole: Collaborative prompting for surgical instrument segmentation. *arXiv preprint arXiv:2312.14481*, 2023.

[17] W. Yue, J. Zhang, K. Hu, Y. Xia, J. Luo, and Z. Wang. Surgicalsam: Efficient class promptable surgical instrument segmentation. *arXiv preprint arXiv:2308.08746*, 2023.

[18] K. Zhang and D. Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

[19] X. Zhang, Y. Liu, Y. Lin, Q. Liao, and Y. Li. Uv-sam: Adapting segment anything model for urban village identification. *arXiv preprint arXiv:2401.08083*, 2024.