SAMUS: Adapting Segment Anything Model for Clinically-Friendly and Generalizable Ultrasound Image Segmentation

摘要

SAM 作为计算机视觉领域的一个通用模型，具有卓越的分割能力以及强大的零样本泛化能力。由于目前医学图像分割领域大多数为专有模型，同时，医学数据集的标注耗时耗力且需要标注者具有极高的专业知识，因此，针对医学图像分割设计一个通用模型是有意义的。本文首先介绍了 SAM 的模型架构，同时回顾了当前基于 SAM 的医学图像分割的研究，然后详细介绍了本文复现的 SAMUS 的模型架构。在复现过程中，首先使用数据增强操作来提高 BUSI 数据集的多样性，使得模型具有更强的泛化能力；其次，本文使用 Dice 系数以及 Iou 交并比两个分割领域的常见的评价指标对模型性能进行评估。SAMUS 在 BUSI 数据集上的 Dice 系数为 88.27，Iou 交并比为 81.21，原论文中的 Dice 系数为 85.77。此外，本文借鉴了 HQ-SAM 的想法，设计了 HQ-SAMUS 模型架构，并在 BUSI 数据集上验证了该模型的有效性，在同等实验环境下，其相较于 SAMUS 的 Dice 系数提高了 0.36，Iou 交并比提高了 0.46.

关键词： SAM 模型；医学图像分割；乳腺超声图像；SAMUS 模型；HQ-SAMUS 模型

1 引言

在网络规模的数据集上预训练的大型语言模型具有强大的零样本和少样本泛化能力，正在彻底改变 NLP 领域 [1]，常见的基础模型有 BERT、GPT-3、CLIP 等。这些基础模型通常使用提示工程在新的数据分布上解决一系列的下游分割任务，使得模型能够在训练期间没有见到的任务和数据分布中进行推广和泛化。目前图像分割领域的模型基本上都是专有模型，而基础模型将带来解决分割问题的新范式，特别是可以帮助科研人员提升在解决专有任务时的效率，所以针对图像分割领域的基础模型进行探索是十分必要的。Segment Anything Model [2] 作为图像分割领域的一个基础模型，在人工构造的迄今为止最大的 SA1B 数据集上进行训练，具有卓越的分割能力以及强大的零样本泛化能力。SAM 模型可以根据用户提示，包括点、边界框和粗掩码，自动分割出相应的对象，同时 SAM 具有歧义意识，当提示具有歧义性时，能够给出合理的分割结果。通过简单的提示，SAM 可以解决新数据分布上的一系列下游分割问题，能够轻松地适应各种分割应用程序，具有极大的现实意义。分割是医学成像分析中的一项基本任务，它涉及识别和描绘各种医学图像（如器官、病变和组织）中的感兴趣区域（ROI）。准确的分割结果对疾病诊断、治疗计划和疾病进展监测等临床应用至关重要 [3, 4]。人工分割

长期以来一直是描绘解剖结构和病理区域的“金标准”，但这一过程耗时耗力，且通常需要标注者具有极高的专业知识。基于深度学习的模型能够学习复杂的图像特征，并在各种任务中提供准确的分割结果，在医学图像分割中显示出巨大的前景。然而，目前常见的医学图像分割模型是针对特定的分割任务进行设计和训练的，当应用于新任务或不同类型的成像数据时，需要重新训练，给临床使用带来了极大的不便。虽然 SAM 模型具有强大的表征能力，但由于自然图像和医学图像之间存在的领域差异，使得其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 [5]。为了弥补领域差异并使 SAM 能够有效地适应于医学图像领域，目前有许多研究通过调整 SAM 来设计与训练通用的医学图像分割模型。MedSAM [6] 通过冻结图像编码器和提示编码器，以可接受的成本在医学图像上训练 SAM。MSA [7] 在 ViT 图像编码器的每个 Transformer 层上添加两个上下转换器来引入特定于任务的信息。因此，通过微调 SAM 的模型结构，或者针对医学数据集重新训练 SAM 模型等方式，使得 SAM 适应于医学图像分割领域是一个十分具有前景的研究方向。

2 相关工作

2.1 SAM 模型架构

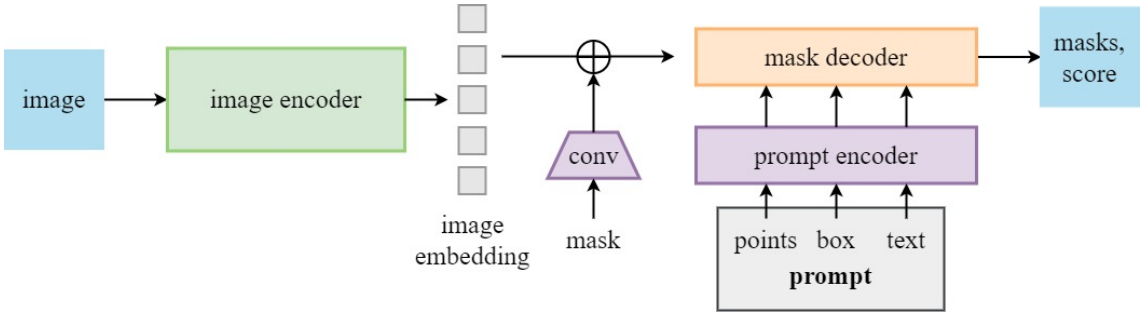

图 1. SAM 模型总体架构

SAM 模型的主要目的是为图像分割领域建立一个基础模型，该模型基于一个包含 1100 万张多样化的、高分辨率的图像和 11 亿个高质量的分割掩码的 SA-1B 数据集上进行预训练，具有较强的分割能力以及强大的零样本泛化能力，该模型由图像编码器、特征编码器以及掩膜解码器三个部分组成（图 1）。首先使用图像编码器从输入图像中提取图像特征，然后使用提示编码器从提示（点、边界框、文本以及粗掩码）中提取提示特征，最后使用一个轻量级的掩膜解码器将两者信息进行融合得到相应的掩膜图以及置信分数。各模块阐述如下：

（1）image encoder：该模块的输入为图像，输出为提取的图像特征，骨干网络可以选择 vit-h、vit-l 或 vit-b，三个骨干网络的差别仅在于模型大小不同，同时该模块采用 MAE 方式进行预训练，以保证模型能够适应高分辨率的图像；

（2）prompt encoder：该模块的输入为提示，提示类型可以为点、边界框、文本以及粗掩码，输出为提取的提示特征。其中，使用位置编码器对点和边界框特征进行提取，使用 CLIP 的文本编码器对文本信息进行特征提取，使用 CNN 模块对密集提示（粗掩码）进行特征提取；

（3）mask decoder：该模块的输入为前两个编码器中提取的图像特征、提示特征，以及额

外加上的 output tokens 和 output tokens，类似于 vit 中的 cls token，用于输出最后的结果；输出为预测的掩膜结果以及相应的置信分数。图像特征和 tokens 首先经过由自注意力以及交叉注意力操作组成的模块进行特征更新。其次，以 tokens 作为 q，与更新后的图像特征进行交叉注意力操作，得到最终的 tokens。在获得输出时，共有两个分支，一个分支是对 tokens 直接进行 MLP 操作，得到对应的置信分数，另一个分支是对 tokens 进行 MLP 操作后，与经过两层转置卷积的图像特征进行矩阵乘法操作得到最终的 mask 掩膜图。SAM 的掩膜解码器结构如图 2所示。

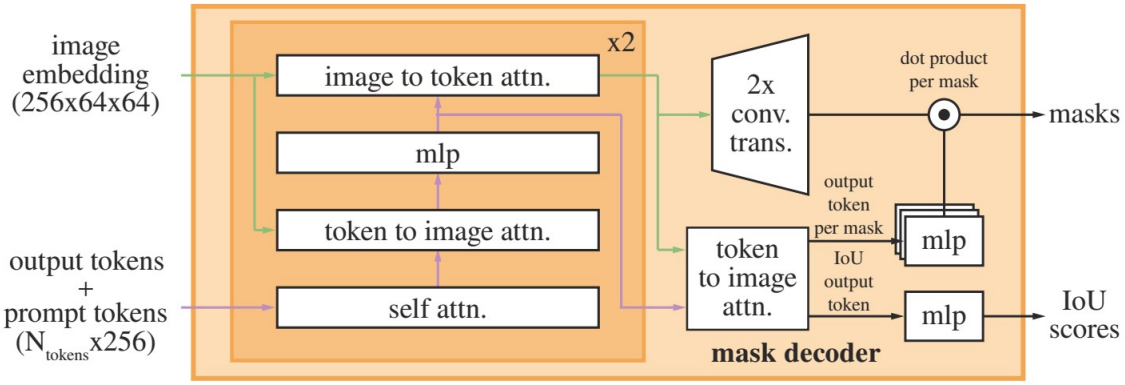

图 2. SAM 模型总体架构

2.2 视觉调优

为了充分利用计算机视觉领域基础模型在大规模数据集上学习到的强大的分割能力，当前研究针对基础模型适应于下游任务提出了一系列的视觉调优方法。视觉调优方法可以分为五大类，分别是微调、参数调优、重新映射调优、提示调优和自适应调优 [8]。具体来说，微调方法包括对预训练模型的整个参数集进行调整，以及对预训练模型的特定部分进行选择性微调 [6]。参数调优方法则直接修改模型参数的权重或偏差 [9]。重新映射方法通过知识蒸馏、基于权重的重新映射或基于体系结构的重新映射的方式，将学习到的信息从预训练模型迁移到下游模型 [10]。提示调优和适配器调优是视觉调优中最常使用的两种调优方式。提示调优通过将一组可学习参数与输入结合或设计一个子网络来生成视觉提示，将下游任务的知识引入到预训练模型中 [11]。适配器调优是通过将额外的可学习参数与固定的预训练模型结合起来，以促进下游任务的学习 [7]。

2.3 SAM 在医学图像分割中的应用

由于自然图像和医学图像之间存在显著的领域差距，在最新的针对 SAM 在医学图像分割上的评估工作表明，无论是否使用提示或者使用何种提示，SAM 的 zero-shot 泛化能力都不足以直接部署在医学图像上。He 等人 [12] 评估了 SAM 在 12 个医学数据集上的分割精度，发现 SAM 的零样本泛化能力明显落后于在特定医学数据集上训练的专有模型，在某些任务中，SAM 的性能差距高达 70%。Huang 等人 [5] 在考虑了不同类型的提示对医学图像分割性能的影响的情况下，也得到了类似的观察结果。由于 SAM 模型在大规模的自然图像数据集上进行训练，所以其具有鲁棒泛化的潜力，也因此使其不局限于任何特定的医学成像模式，即

如果 SAM 微调被证明对一种类型的医学成像有效，那么同样的方法很有可能也适用于其他模式。在 SAM 适应自动医学图像分割的背景下，最近的一些研究采用了参数高效迁移学习（PETL）技术，如 LoRA 或 Adapters，均在自动分割中表现出较好的性能。SAMed [13] 在图像编码器上使用了低秩（low-rank-based，LoRA）策略，以更低的计算成本对 SAM 进行调优，使其更适用于医学图像分割。SAMUS [14] 通过可重叠的 patch embedding、位置适配器、特征适配器等操作，设计了一个在超声图像分割领域中的通用模型，并在综合超声数据集上对其有效性进行了全面评估。然而，以往的研究方法大部分侧重于二维图像信息的自适应，忽略了医学图像中有价值的三维信息，主要包括医学体积数据中至关重要的三维空间信息和医学视频数据中的时间信息。MA-SAM [15] 通过在 SAM 架构内将一系列 3D 适配器集成到 2D Transforme-Block 块中，使得模型能够高效地捕获医学数据中的体积或时间信息。

3 SAMUS 模型

3.1 本文方法概述

目前大多数基于 SAM 大模型的医学图像分割模型采用与 SAM 一致的分块操作，即在对图像特征建模之前，对输入图像执行无重叠的 16 倍 tokenization 操作，这破坏了对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形状、模糊边界、小尺寸或低对比度的临床图像。此外，大部分模型规定输入图像的分辨率为 1024*1024 或 512*512，通过 tokenization 操作后生成的长序列导致大量的计算，对临床使用不够友好。同时，在 SAM 及相关的基于 SAM 大模型的医学图像分割模型中，图像编码器的 ViT 只关注图像的全局信息，而不关注图像的局部信息，导致了图像细节信息的损失。针对以上问题，在论文 SAMUS [14] 中，作者基于 SAM 模型，针对医学数据中的超声图像设计了一个通用的分割模型，该模型将 SAM 卓越的分割性能和强大的泛化能力转移到医学图像分割领域，同时降低了计算的复杂度。SAMUS 继承了 SAM 的 ViT 图像编码器、提示编码器和掩码解码器，并对图像编码器进行了一定的设计，SAMUS 的总体结构如图 3所示。

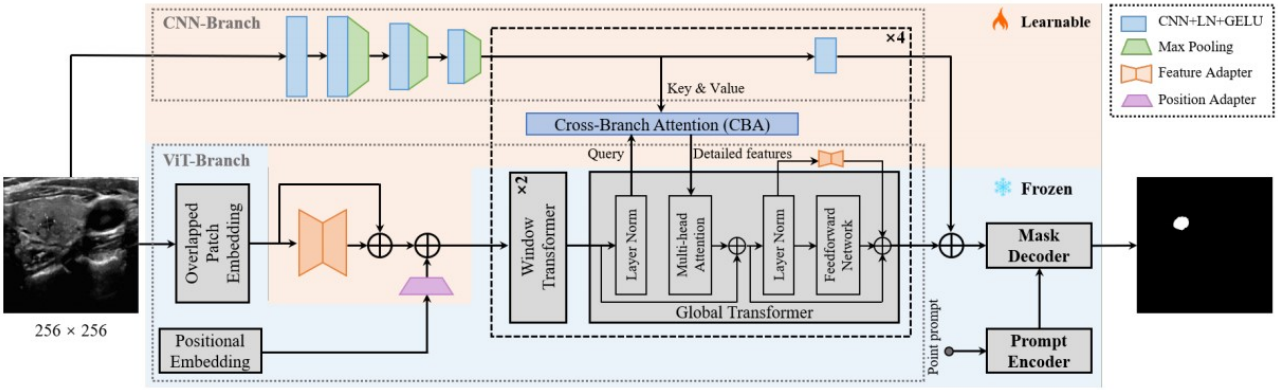

图 3. SAMUS 模型总体架构

相比于 SAM 模型，SAMUS 在整体上与 SAM 保持一致，在 SAMUS 的训练过程中，SAM 的原始参数全部冻结，该操作不但使得模型的训练参数较少，同时还保留了 SAM 强大的特征提取能力。论文的创新点主要在于引入了五个新模块。各模块详细介绍如下：

（1）Overlapped Patch Embedding：该模块的模型参数与 SAM 的 Patch Embedding 的参数相同，步长变为原来的一半，使用可重叠的 tokenization，保留了对识别小目标和边界至关重要的局部信息；

（2）Position Adapter：该模块的模型参数为可学习的参数，由下采样以及卷积操作实现，主要作用是为了匹配嵌入序列的分辨率及调整位置嵌入，以帮助 ViT-Branch 更好地处理较小的输入图像；

（3）Feature Adapter：feature adapter 出现在 image encoder 的两个地方，第一处为 tokenization 与 ViT 图像编码器之间，主要作用是使小尺寸图像的输入特征与 SAM 预训练的 ViT 图像编码器的所需特征分布对齐；剩余的 feature adapter 位于 global transformer 中的前馈网络的残差连接上，主要作用是使 ViT 编码器学习医学图像特征，使得模型更适应于医学图像。此外，feature adapter 主要由下采样，卷积以及上采样操作实现；

（4）CNN-Branch：使用简单的 CNN 架构学习图像的局部信息，学习到的局部信息通过 CBA 模块注入到 Global Transformer 中，使用简单的 CNN 架构学习图像的局部特征是为了防止过拟合；

（5）Cross-Branch Attention（CBA）：该模块使用交叉注意力机制将从 CNN-Branch 中学习到的局部信息注入到 Global Transformer 中，使 ViT image encoder 能够获得图像的局部信息。

3.2 数据增强

通常情况下，在深度学习中，训练样本数量越多，模型的训练效果越好，泛化能力越强。数据增强操作可以增加训练样本的数量，提高模型的泛化能力，同时生成更具多样性的样本，使得模型具有更强的鲁棒性。常见的数据增强操作有：随机裁剪、随机翻转、添加高斯噪声等。由于 BUSI 数据集较小，所以在训练前，本文使用一系列的数据增强方式以提高数据集的样本数量及样本的多样性，使得模型具有更强的泛化能力。以下为 SAMUS 模型使用的数据增强操作：

（1）Gamma 增强 & 随机改变对比度

Gamma 增强和随机改变对比度均用于调整图像的对比度。对于 Gamma 增强，通常情况下，Gamma 值越大，图像的对比度越高，反之则越小。在代码实现部分，本文以一定的概率，对图像进行 Gamma 增强操作，产生具有不同对比度的图像；对于随机改变对比度，该数据增强方式对原图 image 进行操作，首先随机生成一个（0，1）的值，如果小于预先定义的 self.p_contr，则执行随机改变对比度操作。其次，随机生成一个（0.8，2.0）的值，对图像进行对比度增强。

（2）随机缩放和裁剪（random scale and crop）

对原图 image 以及标签 ground truth 进行随机缩放并随机裁剪以调整到原始大小的数据增强操作。首先随机生成一个（0，1）的值，如果小于预先定义的 self.p_scale，则执行缩放和裁剪。其次，随机生成一个（1，1.3）的值，作为图像和标签的缩放比例。最后，对图像和标签进行随机裁剪，使其与原始大小保持一致。

（3）随机旋转（random rotation）

对原图 image 以及标签 ground truth 进行随机旋转。首先随机生成一个（0，1）的值，如果小于预先定义的 self.p_rota，则执行随机旋转。其次，随机生成一个（-30，30）的值，作

为图像和标签的旋转角度。

3.3 评价指标

模型的评价指标是用来评估模型性能优劣的一个定量指标，通常情况下，针对不同的深度学习任务，需要设计不同的评价指标。对于分类任务，常见的评价指标有准确率、召回率、F1 分数等；对于回归任务，常见的评价指标有均方误差 MSE、平均绝对误差 MAE 等；对于分割任务，常见的评价指标有像素准确率 PA、交并比 IoU 等。由于本任务是针对医学图像进行分割，属于像素级分类任务，因此本文使用图像分割领域常见的 Dice 相似系数和 IoU 交并比作为模型的评价指标。Dice 系数是一种集合相似度度量指标，通常用于计算两个样本的相似度，取值范围为，分割结果最好时的取值为 1，最差时的取值为 0，在本文中用于衡量分割结果与真实值之间的重合度。IoU 交并比的定义为两个集合之间交集与并集之间的比值，在本文中表示为真实值与预测值之间的交集同真实值与预测值之间的并集的比值。Dice 系数与 IoU 交并比的数学公式表示如下。

$$\text{Dice} = \frac{2|X \bigcap Y|}{|X| + |Y|} \tag{1}$$

$$\text{IoU} = \frac{|X \bigcap Y|}{|X \bigcup Y|} \tag{2}$$

其中，X 表示预测结果为正例的部分，Y 表示标签为正例的部分，$|X \bigcup Y|$ 表示 X 与 Y 的交集的元素个数，$|X \bigcap Y|$ 表示 X 与 Y 的并集的元素个数，$|X|$ 和 $|Y|$ 分别表示 X 和 Y 的元素个数。

4 复现细节

4.1 BUSI 数据集介绍

BUSI 数据集 [16] 全称为 “Breast Ultrasound Images Dataset”，该数据集共收集了 780 张年龄为 25-75 岁的妇女的乳腺超声图像。该数据集包含了 normal（正常）、benign（良性）、malignant（恶性）三个类别，每种类别的图像数量分别为 133、437、210 张，平均分辨率为 500*500，同时，每张图像都有相应的专家标注好的标签。BUSI 数据集展示如图 4所示。

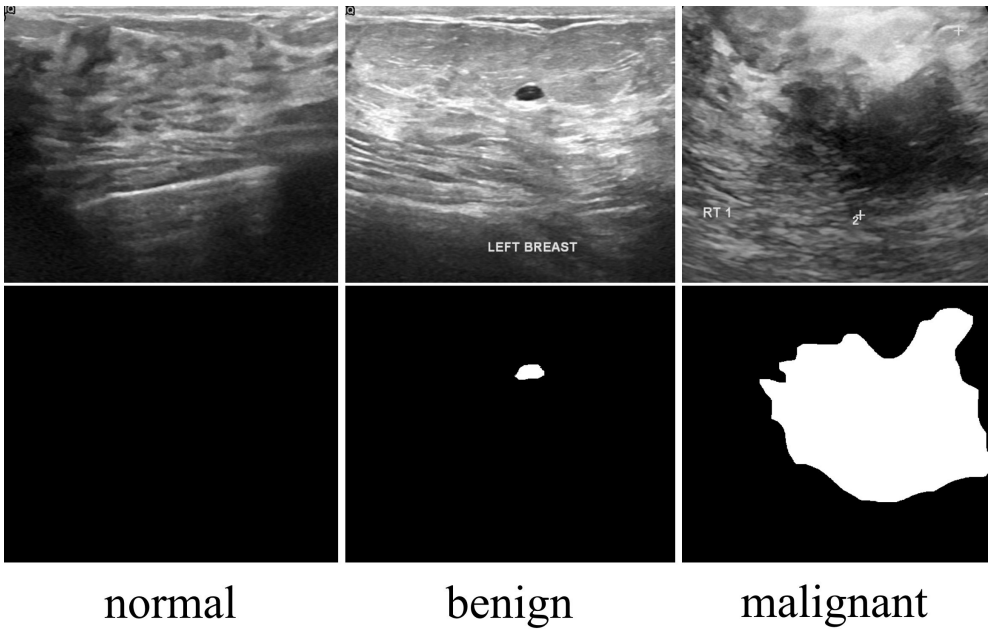

图 4. BUSI 数据集展示

由于 BUSI 数据集是超声图像，所以具有多噪声、低对比度、边缘模糊、形状复杂、多尺度（包含小尺寸物体）以及位置多变（包含多个病灶区域）等超声图像的固有缺点。

图 5. 多尺度问题展示

4.2 与已有开源代码对比

本文参考了 SAMUS 开源的代码，但是重写了 dataset 的定义、模型训练、模型评估以及模型预测部分的代码。同时，本文也借鉴了 HQ-SAM 的解码器部分代码，并修改原 SAMUS 的解码器部分，本文将此模型命名为 HQ-SAMUS。

4.3 创新点

由于 SAMUS 模型存在无法分割多个病灶区域，即难以分割小物体区域，以及分割区域边缘较模糊的问题。根据经验可知，在深度网络中，底层特征通常包含边缘、纹理等局部细节信息，而高层特征包含丰富的全局语义信息，将低级信息注入到高级语义信息中，有利于模型对多病灶区域以及模糊边缘进行分割。HQ-SMA [17] 的设计初衷是因为 image encoder 在对图像不断进行特征提取的过程中，会造成边缘等底层信息的丢失，所以该模型提出通过融合底层特征和高级特征，以生成精细化的分割结果。但是该模型是通过新增加一个 tokens 来生成精细化的分割结果，而不是在 mask decoder 之前对图像特征进行底层信息的注入，该操

作既保留了 SAM 在大规模数据集上学习到的知识，同时也将损失的低级特征与高级特征进行融合。因此，本文借鉴 HQ-SAM 的思想，通过新增一个 tokens 以学习更精细化的图像分割结果，在保留 SAM 在大规模数据集上学习到的知识的情况下，向高级特征中注入低级特征，进而生成精细化的分割结果。HQ-SAMUS 的图像编码器部分与 SAMUS 一致，整体模型结构如图 6所示。

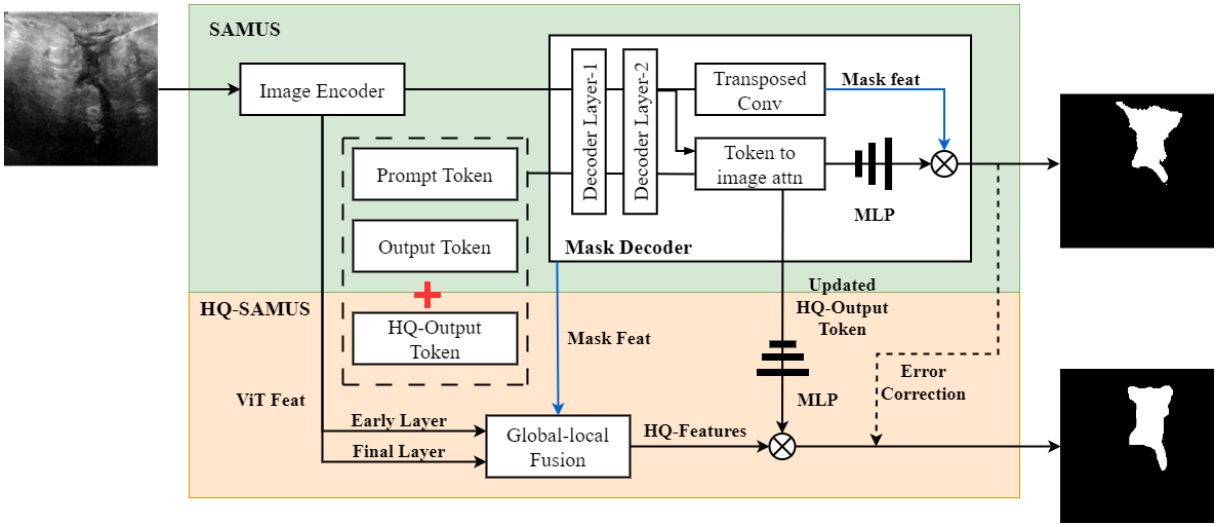

图 6. HQ-SAMUS 模型结构

5 实验结果分析

5.1 SAMUS 复现结果展示

首先对 BUSI 数据集进行划分，训练集、验证集以及测试集的划分比例为 7：2：1，在服务器上共训练了 400 个 epoch，其中第 300 个 epoch 的评价指标最高，Dice 系数为 88.27，Iou 交并比为 81.21，SAMUS 论文中的 Dice 系数为 85.77. 可视化结果如下图所示（第一行为预测结果，第二行为 gt，后同）。

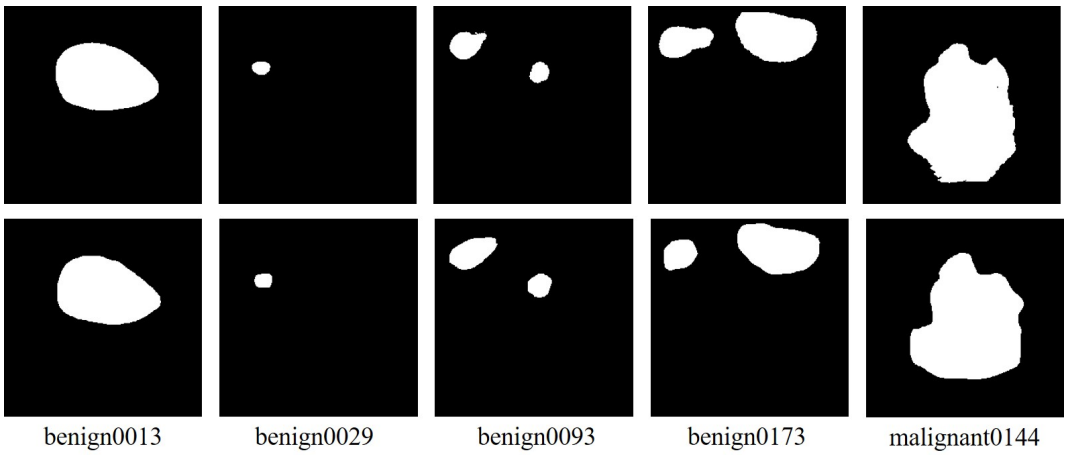

图 7. SAMUS 复现分割成功案例展示

虽然 SAMUS 能在 BUSI 数据集上取得较高的 Dice 结果，但是在一些图像上仍无法获得

较好的分割效果，主要表现在：无法分割多个病灶区域、分割区域连通性较差、分割边缘较模糊。

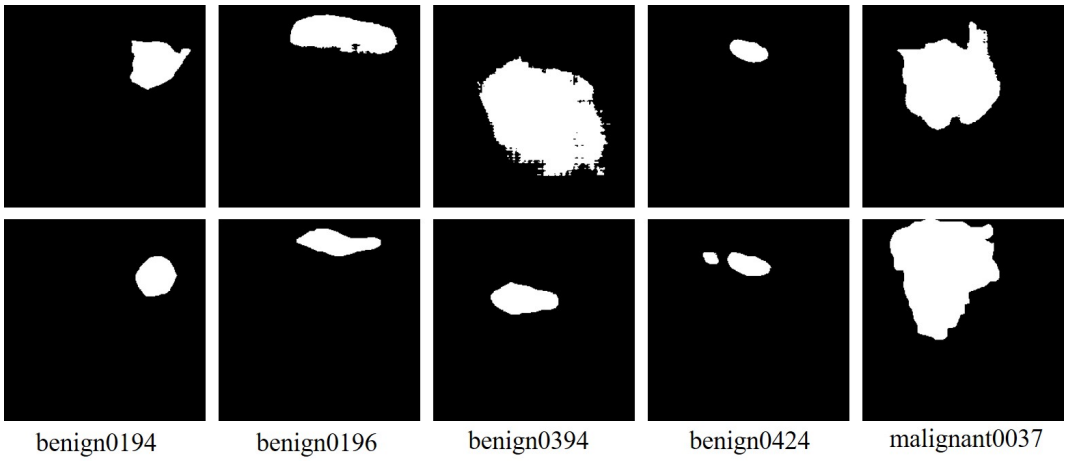

图 8. SAMUS 复现分割失败案例展示

针对以上问题，可以从以下两个方面进行改进：第一个为向掩码解码器中注入图像的边缘信息以及加入监督边缘区域分割效果的损失函数，第二个为增强分割区域的连通性。

5.2 HQ-SAMUS 复现结果展示

实验环境与 SAMUS 相同的情况下，对 HQ-SAMUS 进行训练，定量结果如表1所示。

表 1. 实验结果

Model_Experiment	Dice	Iou
SAMUS	88.27	81.21
HQ-SAMUS	**88.63**	**81.67**

HQ-SAMUS 模型的分割效果如下图所示。

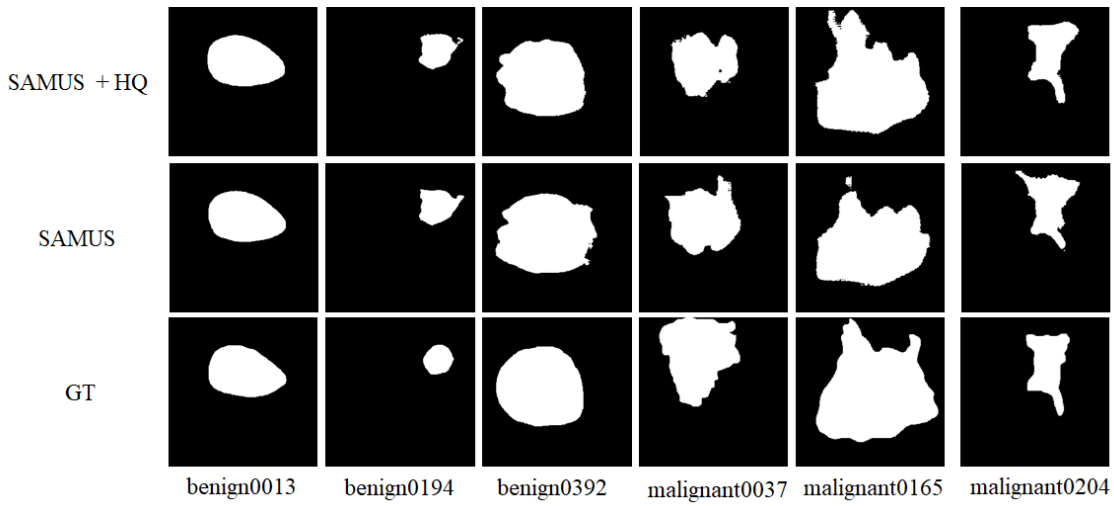

图 9. HQ-SAMUS 分割效果图

6 总结与展望

本文首先详细介绍了复现的 SAMUS 的模型架构，SAMUS 在整体上继承了 SAM 的 ViT 图像编码器、提示编码器和掩码解码器，同时，其针对 SAM 直接应用于医学图像分割中存在的问题，设计了五个有效的新模块，分别为可重叠的 Patch Embedding、位置适配器、特征适配器、CNN 分支、跨分支注意力机制。在复现过程中，首先使用数据增强操作来提高 BUSI 数据集的样本数量以及样本多样性，使得模型具有更强的鲁棒性。在模型评估方面，本文使用 Dice 系数以及 Iou 交并比两个分割领域的常见的评价指标对模型性能进行评估。复现结果显示，SAMUS 在 BUSI 数据集上的 Dice 系数为 88.27，Iou 交并比为 81.21，原论文中的 Dice 系数为 85.77。通过可视化 SAMUS 在 BUSI 数据集上的分割效果发现，在分割过程中存在无法分割多个病灶区域、分割区域连通性较差、分割边缘较模糊的问题，可能原因之一为在模型学习过程中，损失了有助于边缘分割的低级信息。因此，本文借鉴了 HQ-SAM 的想法，设计了 HQ-SAMUS 模型架构，通过新增加一个 tokens 以生成精细化的分割结果，该操作既保留了 SAM 在大规模数据集上学习到的知识，也将损失的低级特征与高级特征进行融合。同时，本文在 BUSI 数据集上验证了该模型的有效性，在同等实验环境下，其相较于 SAMUS 的 Dice 系数提高了 0.36，Iou 交并比提高了 0.46。

参考文献

[1] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[2] Kirillov A, Mintun E, Ravi N, et al. Segment anything[J]. ar**v preprint ar**v:2304.02643, 2023.

[3] De Fauw J, Ledsam J R, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease[J]. Nature medicine, 2018, 24(9): 1342-1350.

[4] Ouyang D, He B, Ghorbani A, et al. Video-based AI for beat-to-beat assessment of cardiac function[J]. Nature, 2020, 580(7802): 252-256.

[5] Huang Y, Yang X, Liu L, et al. Segment anything model for medical images?[J]. Medical Image Analysis, 2023: 103061.

[6] Ma J, Wang B. Segment anything in medical images[J]. ar**v preprint ar**v:2304.12306, 2023.

[7] Wu J, Fu R, Fang H, et al. Medical sam adapter: Adapting segment anything model for medical image segmentation[J]. ar**v preprint ar**v:2304.12620, 2023.

[8] Yu B X B, Chang J, Wang H, et al. Visual Tuning[J]. ar**v preprint ar**v:2305.06061, 2023.

[9] Lian D, Zhou D, Feng J, et al. Scaling shifting your features: A new baseline for efficient model tuning[J]. Advances in Neural Information Processing Systems, 2022, 35: 109-123.

[10] Zhao B, Cui Q, Song R, et al. Decoupled knowledge distillation[C]//Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2022: 11953-11962.

[11] Wang H, Chang J, Luo X, et al. Lion: Implicit vision prompt tuning[J]. ar**v preprint ar**v:2303.09992, 2023.

[12] He S, Bao R, Li J, et al. Accuracy of segment-anything model (sam) in medical image segmentation tasks[J]. ar**v preprint ar**v:2304.09324, 2023.

[13] Zhang K, Liu D. Customized segment anything model for medical image segmentation[J]. ar**v preprint ar**v:2304.13785, 2023.

[14] Lin X, **ang Y, Zhang L, et al. SAMUS: Adapting Segment Anything Model for Clinically-Friendly and Generalizable Ultrasound Image Segmentation[J]. ar**v preprint ar**v:2309.06824, 2023.

[15] Chen C, Miao J, Wu D, et al. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation[J]. ar**v preprint ar**v:2309.08842, 2023.

[16] Al-Dhabyani W, Gomaa M, Khaled H, et al. Dataset of breast ultrasound images[J]. Data in brief, 2020, 28: 104863.

[17] Ke L, Ye M, Danelljan M, et al. Segment Anything in High Quality[J]. ar**v preprint ar**v:2306.01567, 2023.