Computer Science and Application 计算机科学与应用, 2025, 15(3), 29-38
Published Online March 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.153055

Hans汉斯

基于SAM的零样本多模态舌体分割方法

钟甫广[1]，邓森耀[1]，曾军英[1]，冯　跃[1]，钟甫东[1]，贾旭东[2*]

[1]五邑大学电子与信息工程学院，广东 江门
[2]加州州立大学北岭分校计算机科学与工程学院，美国 洛杉矶

收稿日期：2025年2月3日；录用日期：2025年3月4日；发布日期：2025年3月12日

摘　要

舌诊通过观察舌体特征评估健康状态，而舌体分割作为智能舌诊的关键步骤，需要准确分离舌体与背景，为后续特征提取和健康分析奠定基础。然而，舌体分割目前面临着两大挑战：一是数据的稀缺性，二是现有的分割大模型(如SAM模型)对人工提示的依赖性。为了解决以上问题，本文提出了一种零样本多模态的分割方法。该方法结合SAM模型和多模态提示技术，通过两阶段框架实现：1) 初步分割和相似度聚类，利用SAM模型生成初步分割结果，并通过相似度聚类解码器筛选潜在有效分割；2) 精细化分割，利用多模态大语言模型分析舌体特征，生成精确点提示，再次输入到SAM模型中以实现高精度分割。该方法在无需特定任务训练或标注数据的情况下，实现了SAM模型在舌诊领域的智能分割应用。实验结果显示，相比于原始的SAM模型，该方法在三个舌诊数据集上的mIoU指标分别提升了27.3%，18.2%，29.7%。

关键词

舌体分割，零样本学习，多模态大语言模型，相似度聚类，医学图像处理

Zero-Shot Multimodal Tongue Image Segmentation Based on SAM

Fuguang Zhong[1], Senyao Deng[1], Junying Zeng[1], Yue Feng[1], Fudong Zhong[1], Xudong Jia[2*]

[1]School of Electronics and Information Engineering, Wuyi University, Jiangmen Guangdong
[2]School of Computer Science and Engineering, California State University, Northridge, Los Angeles USA

Received: Feb. 3rd, 2025; accepted: Mar. 4th, 2025; published: Mar. 12th, 2025

Abstract

Tongue diagnosis assesses health status by observing tongue characteristics, and tongue segmentation, as a key step in intelligent tongue diagnosis, requires accurately separating the tongue body

*通讯作者。

文章引用: 钟甫广, 邓森耀, 曾军英, 冯跃, 钟甫东, 贾旭东. 基于SAM的零样本多模态舌体分割方法[J]. 计算机科学与应用, 2025, 15(3): 29-38. DOI: 10.12677/csa.2025.153055

from the background to lay a foundation for subsequent feature extraction and health analysis. However, tongue segmentation currently faces two main challenges: data scarcity and the dependency of existing large segmentation models (such as the segment anything model) on manual prompts. To address these issues, this paper proposes a zero-shot multimodal segmentation method. This method combines the SAM model with multimodal prompt techniques and implemented in a two-stage framework: 1) initial segmentation and similarity clustering, where the SAM model generates initial segmentation results, followed by a similarity clustering decoder to filter out potentially effective segmentations; 2) refined segmentation, where a multimodal large language model analyzes tongue characteristics to generate precise point prompts, which are re-entered into the SAM model to achieve high-precision segmentation. This method enables intelligent segmentation with the SAM model in tongue diagnosis without the need for task-specific training or annotated data. Experimental results show that, compared to the original SAM model, this method improves the mIoU metric on three tongue diagnosis datasets by 27.3%, 18.2%, and 29.7%, respectively.

Keywords

Tongue Image Segmentation, Zero-Shot Learning, Multimodal Large Language Model, Similarity Clustering, Medical Image Processing

1. 引言

医学图像语义分割在计算机辅助诊断和智慧医疗中发挥着至关重要的作用，极大地提高了诊断的效率和准确率[1]。在此背景下，传统中医的舌诊正面临新的发展机遇。舌诊通过观察舌色、舌形、舌苔等特征评估全身健康状态和疾病信息，具有独特的诊断优势。研究表明，舌象与多种疾病有显著关联，拓展了舌诊在疾病预测和健康评估中的应用前景。舌体分割是舌诊客观化的重要基础步骤，其准确性直接影响后续诊断的可靠性[2]。因此，实现高精度的舌体自动化分割成为舌诊客观化的重要挑战。

舌体分割的主要挑战在于数据的稀缺性[3]。与其他医学图像领域相比，舌体数据的标准化数据集较少，制约了深度学习模型的应用和性能表现。舌体采集和标注需要专业中医知识，耗时费力。此外，舌体图像的采集缺乏统一标准，易受光照和拍摄角度等因素影响，导致图像质量不稳定，增加了数据标准化的难度。尽管传统机器学习和深度学习方法在医学图像分割中取得了显著成果[4]，但这些方法依赖大量精确标注的数据，在舌诊领域尤为困难且成本高昂。舌体与周围组织(如嘴唇)的边界模糊，且不同个体在舌体大小、形状和颜色上存在显著差异，进一步增加了分割任务的复杂性。

在此背景下，Segment Anything Model (SAM) [5]作为一个基于 Transformer [6]架构的可提示基础模型，凭借其在大规模 SA-1B 数据集上的训练，展现了强大的零样本泛化能力，为图像分割任务开辟了新的可能性。Transformer 架构基于自注意力机制，能够有效捕捉图像中的长距离依赖关系，实现更精准的边界识别和特征提取。然而，直接将 SAM 应用于舌体分割仍面临挑战。首先，SAM 作为交互式模型，通常需要用户提供提示(如点、框或掩码)来引导分割，处理大量舌体图像时效率低下。此外，舌体边界模糊，SAM 在没有精确提示的情况下难以准确定位边界，导致分割结果不理想。而且，SAM 的分割结果依赖用户输入的提示，不同提示可能导致结果不一致，影响后续分析和诊断。

为提升 SAM 在医学图像分割中的性能，研究人员进行了多方面的改进。例如，Chai 等[7]提出的

SAMM 方法结合了额外的卷积神经网络(CNN)作为补充编码器，在多个医学图像分割数据集上显著优于原始 SAM 模型。Shi 等[8]则针对多模态 MRI 图像的胶质瘤分割任务，提出了一种基于多模态融合的跨模态注意力适配器，有效整合来自不同 MRI 序列的信息，显著提升了分割效果。然而，这些改进方法仍存在局限性，大多数方法依赖额外的训练数据和计算资源，且未解决智能化分割的需求，尤其是在零样本或少样本分割场景中的应用不足。

本文提出了一种基于 SAM 的零样本多模态舌体分割方法(Zero-shot Multimodal Tongue Image Segmentation, ZMT-SAM)，通过结合 SAM 模型与多模态大语言模型提示技术，在无需额外训练数据的情况下实现对舌体的高精度自动化分割。与现有方法相比，ZMT-SAM 克服了数据稀缺的问题，显著提升了分割精度与智能化程度。

以下是该方法的主要创新点：

1) ZMT-SAM 模型采用无监督方法，不依赖特定任务的训练或标注数据。通过直接利用预训练的 SAM 模型进行初步分割，解决了舌诊领域数据稀缺的问题，展示了强大的零样本泛化能力。本文引入了相似度聚类解码器，将 SAM 生成的分割掩码按照相似度进行分类，筛选出潜在的有效分割结果，为后续精细分割奠定基础。

2) ZMT-SAM 模型利用多模态大语言模型生成舌体的详细描述，并将这些描述转化为精确的点提示。这些点提示被重新输入到 SAM 中，再次分割复杂的舌体，实现更高的精度。

3) 在多个舌体数据集上的实验结果表明，ZMT-SAM 不仅在分割精度上显著优于现有方法，还展现出良好的鲁棒性和泛化能力。

2. 方法论

2.1. 零样本多模态舌体分割方法

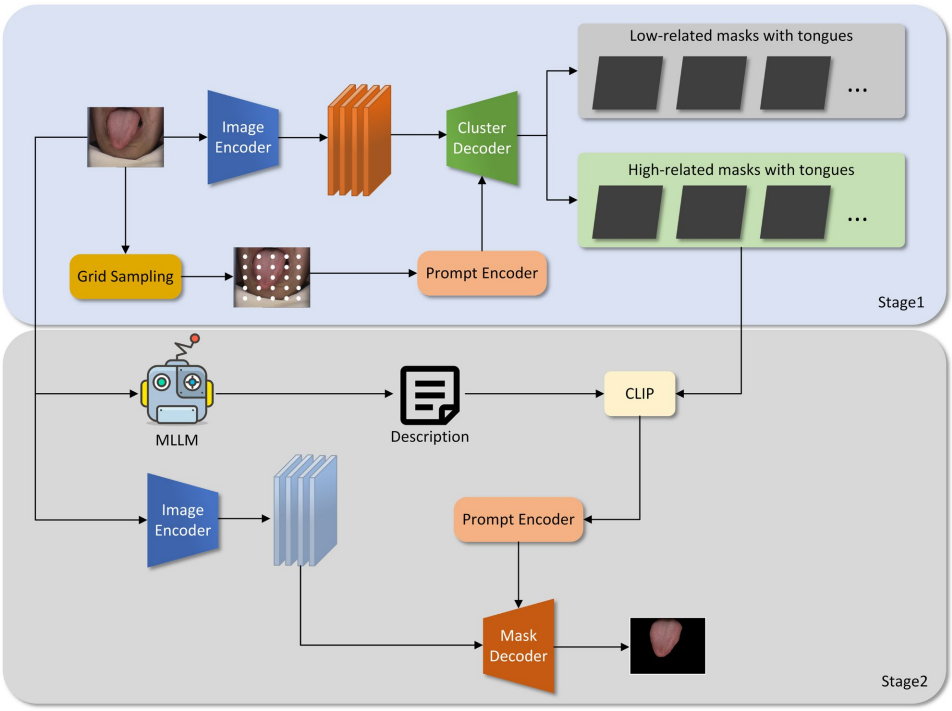

Figure 1. Structure of the ZMT-SAM
图 1. ZMT-SAM 模型架构

本文设计的基于 SAM 的零样本多模态舌体分割方法如图 1 所示，分为两个阶段。

在第一阶段，着重于初步分割和相似度聚类。首先，舌体图像和提示图像分别通过图像编码器和提示编码器进行处理。图像编码器用于提取图像中的视觉特征，例如边缘、纹理和形状，而提示编码器则将提示信息转换为特征向量，以帮助模型识别图像中的目标区域。通过这种方式，视觉信息与提示信息得以有效融合，从而在后续的分割任务中实现更高的准确性和精度。这些编码结果输入到 SAM 模型中，生成网格化的初步分割结果。接着，通过相似度量聚类解码器对这些初步分割结果进行处理，将掩码划分为与舌体高度相关和低相关度的两类，从而有效筛选出最可能包含舌体区域的分割结果。

在第二阶段，专注于精细化分割。首先，第一阶段筛选出的高相似度掩码被输入到 CLIP [9]模型中。CLIP 模型通过将图像和文本描述映射到同一表示空间，融合视觉和语义信息。同时，原始舌体图像被送入多模态大语言模型(MLLM [10])进行分析，以生成关于舌体特征的详细文本描述，并进一步输入到 CLIP 模型中，通过结合这些文本描述和图像信息来生成更精确的提示。

最后，这些精确的点提示与原始舌体图像一起被重新输入到 SAM 模型中。在 SAM 内部，点提示通过提示编码器处理，而原始图像则再次通过图像编码器处理。这些编码结果输入到 SAM 的分割解码器中，最终输出高精度的舌体分割结果。

2.2. 相似度聚类解码器

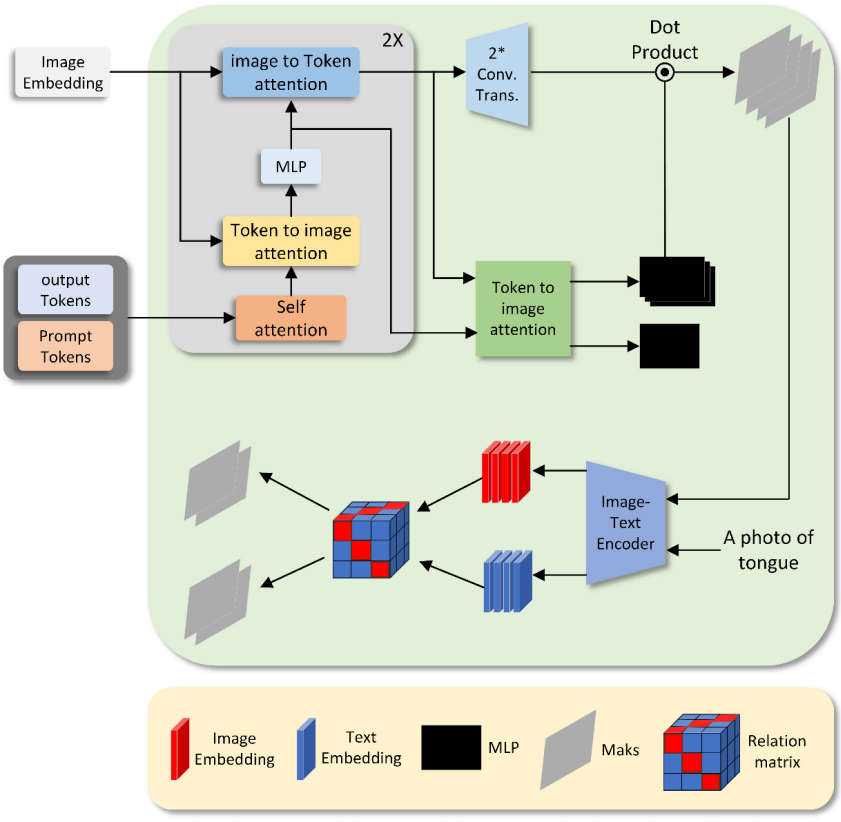

Figure 2. The structure of the similarity clustering decoder.
图 2. 相似度聚类解码器结构

为了优化 SAM 在舌体分割中的初步结果，我们提出了一种创新的分类与筛选策略，通过相似度聚类模块(图 2)对 SAM 生成的多个潜在掩码进行分类。该模块旨在将掩码分为与舌体高度相关的掩码和非舌

体区域的掩码，提升分割精度和效率。这一无监督聚类方法不依赖大量标注数据，显著提高了分割效果。为了更好地理解相似度聚类模块的工作原理，接下来我们详细介绍其内部运行机制。

掩码解码器通过融合图像编码和提示编码，输出高质量的分割掩码。首先，将输出标记与提示标记拼接，接着，这些标记通过一个两层 Transformer 结构进行深度融合。标记首先通过自注意力层进行处理，自注意力机制计算标记之间的关系以提取重要的上下文信息。公式如下：

$$\mathrm{SelfAttention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{soft\,max}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right) \tag{1}$$

其中，Q、K、V 分别是查询矩阵、键矩阵和值矩阵，d_k 是键向量的维度。自注意力机制通过计算标记之间的关系，提取重要的上下文信息。随后，标记作为交叉注意力中的查询，与图像嵌入进行交叉注意力操作，从而更新标记。交叉注意力的公式为：

$$\mathrm{CrossAttention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{soft\,max}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right) \tag{2}$$

在这个过程里，Q 是标记嵌入，K、V 是图像嵌入。通过交叉注意力机制，标记嵌入和图像嵌入之间的信息得以有效融合。接下来，经过两层的 MLP 层，标记得到进一步更新和优化。同时，模型将图像嵌入作为查询，再次与标记进行交叉注意力操作，这一过程进一步更新了图像嵌入。经过两次这种双向交叉注意力操作后，模型将标记作为查询，与经过更新的图像嵌入进行最终的交叉注意力操作，输出最终优化后的标记。MLP 的非线性变换公式如下：

$$\mathrm{MLP}(x) = \sigma(xW_1 + b_1)W_2 + b_2 \tag{3}$$

其中，W 和是权重矩阵，b 是偏置向量，σ 是激活函数。更新后的图像嵌入通过两层转置卷积进行上采样，卷积核大小为 2 × 2，步长为 2。这种上采样方式能够进一步提升图像嵌入的分辨率与细节信息。卷积操作的公式如下：

$$y = \mathrm{ConvTranspose}\ 2d(x, W_1, b_1, s, k) * W_2 + b_2 \tag{4}$$

$$Y = \sigma(\mathrm{LN}(y)) \tag{5}$$

ConvTranspose $2d$ 为转置卷积操作，LN 为层归一化，σ 为激活函数，W 为转置卷积的权重，b 为转置卷积的偏置，s 为步长，k 为卷积核大小。通过上采样后的图像嵌入与最终的优化标记进行整合。在此过程中，掩码标记从输出标记中分离，并通过一个三层的 MLP 层调整通道数，使其与最终输出的图像嵌入保持一致。

本文进一步引入了创新性的相似度聚类模块。这个模块对掩码解码器输出的多个候选掩码进行深入的相似度分析。通过采用先进的聚类算法，我们成功地将这些掩码分为两个关键组：高相似度组和低相似度组。高相似度组通常包含与舌体形状、纹理和位置高度相关的掩码，这些掩码对于准确识别和分割舌体至关重要。相比之下，低相似度组可能包含背景、口腔其他部位或其他非关键区域的掩码，这些信息在舌体分割任务中相对次要。

为了实现这一划分，我们在图像处理部分采用了 Vision Transformer [11] (ViT)的架构。舌体图像首先被划分为固定大小的图像块。这种分割方法可以像处理自然语言中的词元序列一样处理图像数据。接下来，每个图像块都被展平成一维向量，并通过一个可学习的线性投影层转换为固定维度的嵌入向量。这个维度是模型的隐藏维度。为了保留图像块在原始图像中的空间信息，需要为每个块向量添加位置编码。经过位置编码的图像块向量随后被送入一系列 32 层的 Transformer 编码器层。在这里，多头注意力机制发挥了关键作用，允许模型同时关注图像的不同方面，大大增强了特征提取的多样性和丰富性。多头注

意力公式如下：

$$\mathrm{MHA}(Q,K,V)=\mathrm{Concat}(\mathrm{head}_i,\cdots,\mathrm{head}_h)W \tag{6}$$

$$\mathrm{head}_i=\mathrm{Attention}(QW_i^Q,KW_i^K,VW_i^V) \tag{7}$$

这种全局上下文的捕获使得模型能够理解图像的整体结构和局部细节之间的关系。最后每个图像块的输出向量会合并成一个单一的图像表示生成最终的高维图像嵌入。这个嵌入捕获了输入图像的全面视觉特征。我们称之为向量 **A**。与此同时，在文本处理流程中(“一张关于舌头的照片”文本输入)，输入的文本被送入一个 12 层的 Transformer 编码器层。尽管层数较少，但这个文本 Transformer 同样采用了多头注意力机制来理解单词之间的关系和上下文信息。通过这 12 层的处理，文本被转化为另一个高维向量，我们称之为向量 **B**。将向量 **A** 与向量 **B** 计算相似度，然后根据相似度的高低，将这些掩码分为高相似度组和低相似度组。公式如图 3 所示：

$$Similarity(\mathbf{A},\mathbf{B})=\frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{8}$$

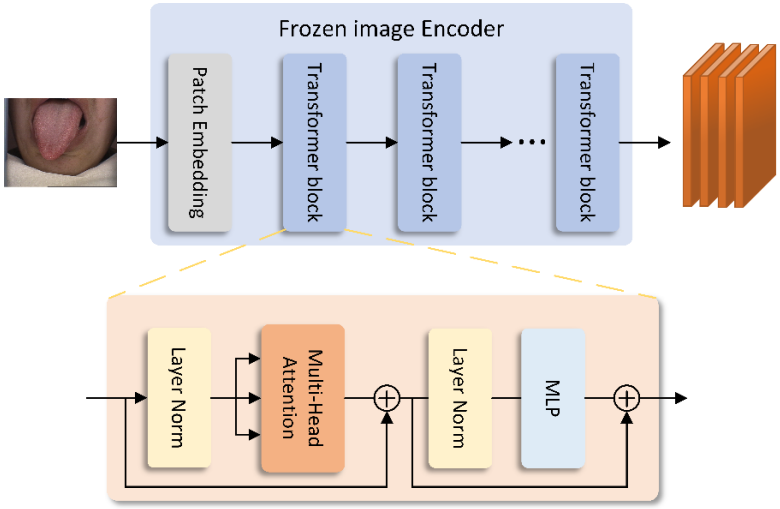

Figure 3. The structure of the Vision Transformer (ViT) encoder model
图 3. VIT 编码器模型架构图

2.3. 基于提示学习的精准舌体分割方法

提示学习已成为自然语言处理和计算机视觉领域的重要突破，为解决复杂任务提供了一种高效的新方法。通过精心设计的提示，可以引导预训练模型完成各种下游任务，无需进行大规模的模型微调。

在本文的舌诊分割框架中，提示学习展现了其强大的潜力和灵活性。核心方法是利用多模态大语言模型分析舌体图像。首先，将高质量的舌诊图片输入到经过大规模多模态预训练模型中，通过设计精确的提示，如“请描述舌头形状，特别注意与口腔其他部位的区别”，引导模型聚焦重要特征。模型生成的文本描述为分割任务提供丰富的语义信息，随后输入到 CLIP 模型中。CLIP 将文本和图像映射到同一表示空间，在高相似度组掩码中得到最相似的掩码，然后生成精确的点提示。这些点提示与原始图像一起输入到 SAM 模型，最终实现高精度的舌体区域分割。基于提示学习的方法充分发挥了大语言模型在语义理解和跨模态匹配中的优势，能够捕捉到传统方法可能忽视的细微特征，同时增强了分割过程的可解释性。

在本节中，我们将详细介绍我们的零样本舌体分割的统一框架。首先我们将解释 SAM 是如何为图像

的每个部分生成掩码并完成过滤，然后解释我们是如何将获得的掩码通过 CLIP 检索到我们需要的舌头分割图片。

3. 实验设置

3.1. 数据集

为了全面验证模型有效性，本文采用三个特点各异的数据集。数据集一[12]：哈尔滨工业大学提供的公开舌体分割数据集，包含 300 张标准化 RGB 舌体图像(768 × 576 像素)及人工标注掩码。数据集二[13]：从网络收集的 1000 张不同尺寸舌体图像，使用 labelme 工具进行人工标注。数据集三：268 张来自专业舌诊仪器的高质量舌体图像。

3.2. 实验指标

Table 1. Zero-shot tongue segmentation experiment results
表 1. 零样本舌体分割实验结果

方法	数据集一			数据集二			数据集三		
	mIoU %	mPA %	Acc %	mIoU %	mPA %	Acc %	mIoU %	mPA %	Acc %
Unet	62.9	86.6	79.7	59.0	79.8	74.4	44.4	86.2	76.2
Unet++	85.5	93.4	94.4	73.4	87.4	85.3	54.9	90.7	87.6
PSPNet	41.4	52.0	79.0	54.3	67.0	77.4	48.1	50.0	86.1
DeepLabV3	82.1	93.0	92.6	69.3	82.4	83.2	46.8	88.5	79.0
DeepLabV3+	89.1	94.4	96.0	79.1	87.1	89.7	62.5	90.8	92.8
MAnet	61.0	80.3	79.7	58.5	77.7	74.3	59.4	90.8	91.0
Linknet	43.0	53.4	79.6	64.3	75.8	81.6	60.2	65.7	86.1
FPN	83.3	94.1	93.1	72.5	83.1	85.7	65.5	70.5	86.9
PAN	55.9	64.4	84.1	58.7	70.5	86.8	56.9	58.6	84.1
ZMT-SAM	93.5	96.6	97.0	88.5	93.5	93.2	91.3	95.1	97.4

为了评估模型的性能，本文采用了三个评估指标：(Mean IoU Score)、(Mean Pixel Accuracy)和(Accuracy)。公式如下：

$$\mathrm{IoU} = \frac{TP}{TP + FN + FP} \tag{9}$$

$$\mathrm{mIoU} = \frac{\sum_i^N \mathrm{IoU}_i}{N} \tag{10}$$

$$\mathrm{CPA} = \frac{TP}{TP + FP} \tag{11}$$

$$\mathrm{mPA} = \frac{\sum_i^N \mathrm{CPA}_i}{N} \tag{12}$$

$$\mathrm{Acc} = \frac{TP + TN}{TP + TN + FN + FP} \tag{13}$$

TP 和 *TN* 分别表示预测为正且实际为正、预测为负且实际为负的像素数量。*FN* 和 *FP* 分别表示预测为负但实际为正、预测为正但实际为负的像素数量。*N* 是类别的总数。IoU_i 是第 *i* 类别的 IoU 值。 CPA_i 是第 *i* 类别的分类精度。

本文以 SAM 作为基准模型，与多个经典的图像分割模型进行对比。实验结果如表 1 所示。对比模型包括 Unet [14]、Unet++ [15]、PSPNet [16]、DeepLabV3 [17]、DeepLabV3+ [18]、Manet [19]、Linknet [20]、FPN [21]和 PAN [22]。本次实验将在零样本情况下进行测试。

3.3. 实验环境

实验使用了 NVIDIA GeForce RTX 3060 12G 显卡。在软件环境方面，采用了 PyTorch 2.12 作为主要的深度学习框架，并结合 TorchVision 0.16.2 处理图像相关任务。为加速计算，实验使用了 CUDA 11.8 进行 GPU 加速。此外，实验系统运行在 Windows 11 操作系统上。

4. 实验结果与分析

4.1. 舌体分割实验

这些结果表明，我们提出的方法在所有指标上均取得了最佳成绩，证明了其在该领域中的有效性和优越性。基于实验结果，本文提出的方法在三个不同特征的数据集上均表现出卓越性能。在标准化公开数据集一上，本方法的 mIoU、mPA 和 Acc 分别达到 93.5%、96.6%和 97.0%，显著优于其他对比模型，验证了其在理想条件下的高精度分割能力。在模拟实际应用场景的网络收集数据集上，尽管图像更加复杂多样，本方法仍保持了 88.5%、93.5%和 93.2%的高水平表现，展现了其鲁棒性和对非标准化、多样化图像的适应能力。相比之下，其他模型在该数据集上的性能显著下降，进一步突显了本方法的优势。在专业舌诊仪器的高质量数据集上，本方法的 mIoU、mPA 和 Acc 分别为 91.3%、95.1%和 97.4%，表明其在接近临床应用的标准化环境中同样具有极高的准确性和实用性。

本文提出的方法在三个不同特征的数据集上均表现出卓越性能，尤其在零样本学习场景下表现突出。展示了出色的跨数据集泛化能力。

4.2. 消融实验

为了验证本文方法的有效性，进行了消融实验，结果如表 2 所示。实验对比了三种方法：基础的 SAM 模型、结合 CLIP 的 SAM 模型(SAM + CLIP)，以及本文提出的完整方法。

Table 2. Ablation experiment results
表 2. 消融实验结果

方法	数据集一			数据集二			数据集三		
	mIoU %	mPA %	Acc %	mIoU %	mPA%	Acc %	mIoU %	mPA %	Acc %
SAM	66.2	72.2	81.4	70.3	81.5	84.4	61.6	63.1	75.2
SAM + CLIP	76.7	88.1	83.8	80.9	87.2	88.7	90.5	95.1	96.7
ZMT-SAM	93.5	96.6	97.0	88.5	93.5	93.2	91.3	95.1	97.4

实验结果表明，本文方法在所有数据集上均取得了最佳性能。在数据集一上，本方法的 mIoU、mPA 和 Acc 分别为 93.5%、96.6%和 97.0%，显著优于其他两种方法，验证了其在标准化数据集上的卓越表现。在更具挑战性的数据集二上，本方法的 mIoU、mPA 和 Acc 分别达到 88.5%、93.5%和 93.2%，依然保持

领先，展现了强大的鲁棒性和适应能力。在数据集三上，本方法的 mIoU、mPA 和 Acc 分别为 91.3%、95.1%和 97.4%，不仅优于其他两种方法，还展现了其在高质量专业数据集上的优异性能。

综上所述，消融实验结果清晰地表明，本文方法的每个组成部分均对最终性能有实质性贡献。基础的 SAM 模型提供了坚实的分割基础，CLIP 的引入增强了模型的语义理解，而本文的相似度聚类解码器和其他优化策略进一步提升了模型在舌体分割任务中的整体性能，确保了其在不同数据集上的出色表现及跨数据集的泛化能力。

5. 结论

本文提出了一种基于 SAM 的零样本多模态舌体分割方法，结合了相似度聚类解码器和多模态大语言模型提示技术，实现了对舌体的高精度自动分割。该方法在无需任务特定训练数据的情况下，有效解决了舌诊领域数据稀缺和对人工提示依赖的问题。实验结果表明，ZMT-SAM 在多个舌体数据集上均表现出卓越性能，尤其在零样本学习场景中展现了出色的鲁棒性和泛化能力，显著优于传统的分割模型。消融实验进一步证明了相似度聚类解码器和多模态提示策略对分割精度的提升作用。未来研究将集中于进一步优化模型的处理速度与扩展性，探索其在更多中医望诊图像数据中的应用潜力，以推动中医诊断的智能化和现代化发展。

参考文献

[1] 清华, 孙水发, 吴义熔. 基于短距离跳跃连接的 U2-Net+医学图像语义分割[J/OL]. 现代电子技术: 1-9. http://kns.cnki.net/kcms/detail/61.1224.TN.20240705.1143.002.html, 2024-10-25.

[2] 梁淑芬, 陈琛, 冯跃, 等. 基于一种局部图像增强和改进分水岭的舌体分割算法[J]. 现代电子技术, 2021, 44(16): 138-144.

[3] Li, L., Luo, Z., Zhang, M., Cai, Y., Li, C. and Li, S. (2020) An Iterative Transfer Learning Framework for Cross-Domain Tongue Segmentation. *Concurrency and Computation: Practice and Experience*, **32**, e5714. https://doi.org/10.1002/cpe.5714

[4] Zhang, X., Bian, H., Cai, Y., Zhang, K. and Li, H. (2022) An Improved Tongue Image Segmentation Algorithm Based on Deeplabv3+ Framework. *IET Image Processing*, **16**, 1473-1485. https://doi.org/10.1049/ipr2.12425

[5] 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

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., *et al*. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[7] Chai, S., Jain, R.K., Teng, S., Liu, J., Li, Y., Tateyama, T., *et al*. (2023) Ladder Fine-Tuning Approach for SAM Integrating Complementary Network. arXiv: 2306.12737. https://arxiv.org/abs/2306.12737

[8] Shi, X., Chai, S., Li, Y., Cheng, J., Bai, J., Zhao, G., *et al*. (2023) Cross-Modality Attention Adapter: A Glioma Segmentation Fine-Tuning Method for SAM Using Multimodal Brain MR Images. arXiv: 2307.01124. https://arxiv.org/abs/2307.01124

[9] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., *et al*. (2021) Learning Transferable Visual Models from Natural Language Supervision. arXiv: 2103.00020. https://doi.org/10.48550/arXiv.2103.00020

[10] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., *et al*. (2024) A Survey on Multimodal Large Language Models. arXiv: 2306.13549. https://arxiv.org/abs/2306.13549v2

[11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., *et al*. (2021) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv: 2010.11929. https://arxiv.org/abs/2010.11929v2

[12] TongeImageDataset. https://github.com/BioHit/TongeImageDataset

[13] TongueSAM: An Universal Tongue Segmentation Model Based on SAM with Zero-Shot. https://github.com/cshan-github/tonguesam

[14] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation.

Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, 5-9 October 2015, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[15] Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N. and Liang, J. (2018) UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Granada, 20 September 2018, 3-11. https://doi.org/10.1007/978-3-030-00889-5_1

[16] Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J. (2017) Pyramid Scene Parsing Network. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6230-6239. https://doi.org/10.1109/cvpr.2017.660

[17] Chen, L.-C. Papandreou, G., Schroff, F. and Adam, H. (2017) Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv: 1706.05587. https://arxiv.org/abs/1706.05587v3

[18] Chen, L., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. (2018) Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *Computer Vision—ECCV* 2018, Munich, 8-14 September 2018, 833-851. https://doi.org/10.1007/978-3-030-01234-2_49

[19] Li, R., Zheng, S., Zhang, C., Duan, C., Su, J., Wang, L., *et al*. (2022) Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, **60**, 1-13. https://doi.org/10.1109/tgrs.2021.3093977

[20] Chaurasia, A. and Culurciello, E. (2017) LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation. 2017 *IEEE Visual Communications and Image Processing* (*VCIP*), St. Petersburg, 10-13 December 2017, 1-4. https://doi.org/10.1109/vcip.2017.8305148

[21] Kirillov, A., Girshick, R., He, K. and Dollar, P. (2019) Panoptic Feature Pyramid Networks. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 6392-6401. https://doi.org/10.1109/cvpr.2019.00656

[22] Li, H., Xiong, P., An, J. and Wang, L. (2018) Pyramid Attention Network for Semantic Segmentation. arXiv: 1805.10180. https://arxiv.org/abs/1805.10180v3