



本科生毕业论文（设计）

|  |
| --- |
| 医学图像中的 SAM |

作者

Jun Ma

Yuting He

Feifei Li

Lin Han

Cheny

You

Bo Wa

摘 要

医学影像分割是临床实践中的一个重要组成部分，有助于准确诊断、治疗规划和疾病监测。然而，目前的方法主要依赖于定制的模型，这些模型在不同任务中表现出有限的通用性。在本研究中，我们介绍了 MedSAM，这是首个专为通用医学图像分割设计的基础模型。MedSAM 利用由 100 多万张图像组成的精心策划的数据集的力量，不仅超越了现有的最先进的分割基础模型，而且表现出与专业模型相当甚至更优的性能。此外，MedSAM 还能精确提取用于肿瘤负荷量化的重要生物标记物。通过在广泛的任务范围内提供准确高效的分割，MedSAM 在加快诊断工具的发展和个性化治疗方案方面具有巨大的潜力。

目录

[医学图像中的 SAM - 0 -](#_Toc165858321)

[第1章 简介 - 1 -](#_Toc165858322)

[第2章 结果 - 2 -](#_Toc165858323)

[第3章 讨论 - 5 -](#_Toc165858324)

[第4章 方法 - 6 -](#_Toc165858325)

[4.1 研究设计 - 6 -](#_Toc165858326)

[4.2 数据集策划和预处理 - 7 -](#_Toc165858327)

[4.3 网络架构 - 8 -](#_Toc165858328)

[4.4 训练协议和实验设置 - 8 -](#_Toc165858329)

[第5章 评估指标 - 10 -](#_Toc165858330)

[5.1 统计分析 - 10 -](#_Toc165858331)

[5.2 数据可用性 - 10 -](#_Toc165858332)

[5.3 代码可用性 - 10 -](#_Toc165858333)

[5.4 致谢 - 11 -](#_Toc165858334)

# 简介

分割是医学影像分析中的一项基本任务，涉及在各种医学图像中识别和描绘感兴趣的区域（ROI），如器官、病变和组织。准确的分割对于许多临床应用至关重要，包括疾病诊断、治疗计划和疾病进展监测[1]、[2]。手动分割一直是描绘解剖结构和病理区域的黄金标准，但这个过程耗时、劳动密集，通常需要高度的专业知识。半自动或全自动分割方法可以显著减少所需的时间和劳动，提高一致性，并使大规模数据集的分析成为可能。基于深度学习的模型在医学图像分割中显示出巨大的潜力，因为它们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，从分割特定的解剖结构到识别病理区域[3]。然而，许多当前医学图像分割模型的一个显著限制是它们的任务特定性。这些模型通常为特定分割任务设计和训练，当应用于新任务或不同类型的成像数据时，它们的性能可能会显著下降。这种缺乏普遍性对这些模型在临床实践中的更广泛应用构成了重大障碍。相比之下，自然图像分割领域的最新进展见证了分割基础模型的出现[4]、[5]，在各种分割任务中展示了显著的多功能性和性能。然而，它们在医学图像分割中的应用一直具有挑战性，因为存在显著的领域差距[6]（补充相关工作）。因此，医学图像分割中对通用模型的需求日益增长：可以一次训练，然后应用于广泛的分割任务的模型。这样的模型不仅在模型容量方面展现出更高的多功能性，而且可能在不同任务中带来更一致的结果，从共享的底层架构和训练过程中受益。受到“分割任何事物模型”（SAM）[4]的显著普遍性激励，我们介绍了MedSAM，第一个通用医学图像分割的基础模型。MedSAM是在前所未有的规模上从SAM模型适应而来的，拥有超过一百万对医学图像-掩码对。我们通过在70多个内部验证任务和40多个外部验证任务上的全面实验，对MedSAM进行了彻底评估，涵盖了各种解剖结构、病理状况和医学成像方式。实验结果表明，MedSAM一致性地超越了最先进的（SOTA）分割基础模型，同时实现了与专家模型相当或甚至超越的性能。这些结果突出了MedSAM作为医学图像分割的强大工具的潜力。

# 结果

MedSAM旨在成为通用医学图像分割的基础模型。构建这样一个模型的一个关键方面是能够适应成像条件、解剖结构和病理状况的广泛变化。为了应对这一挑战，我们策划了一个多样化且大规模的医学图像分割数据集，包含1,090,486对医学图像-掩码对，涵盖了15种成像方式、30多种癌症类型以及多种成像协议（图1a，补充表1-4）。这个大规模数据集允许MedSAM学习医学图像的丰富表示，捕捉不同成像方式下广泛的解剖结构和病变。图1b提供了数据集中不同医学成像方式的图像分布概览，按其总数排名。显然，计算机断层扫描（CT）、磁共振成像（MRI）和内窥镜检查是主要的成像方式，反映了它们在临床实践中的普遍性。CT和MRI图像提供了3D体结构的详细横截面视图，使它们成为非侵入性诊断成像不可或缺的工具。内窥镜检查虽然更具侵入性，但能够直接观察器官内部，对于诊断胃肠和泌尿系统疾病非常有价值。尽管这些成像方式很普遍，但像超声、病理、眼底、皮肤镜、乳腺摄影和光学相干断层扫描（OCT）等其他方式在临床实践中也扮演着重要角色。这些方式的多样性及其相应的分割目标强调了需要通用且有效的分割模型，这些模型能够处理与每种方式相关的独特特征。

图1b提供了数据集中不同医学成像方式的图像分布概览，按其总数排名。显然，计算机断层扫描（CT）、磁共振成像（MRI）和内窥镜检查是主要的成像方式，反映了它们在临床实践中的普遍性。CT和MRI图像提供了3D体结构的详细横截面视图，使它们成为非侵入性诊断成像不可或缺的工具。内窥镜检查虽然更具侵入性，但能够直接观察器官内部，对于诊断胃肠和泌尿系统疾病非常有价值。尽管这些成像方式很普遍，但像超声、病理、眼底、皮肤镜、乳腺摄影和光学相干断层扫描（OCT）等其他方式在临床实践中也扮演着重要角色。这些方式的多样性及其相应的分割目标强调了需要通用且有效的分割模型，这些模型能够处理与每种方式相关的独特特征。

另一个关键考虑是选择合适的分割提示和网络架构。虽然全自动分割基础模型的概念很吸引人，但它充满了挑战，使其变得不切实际。其中一个主要挑战是分割任务固有的可变性。例如，给定一个肝癌CT图像，分割任务可以根据具体的临床情况而变化。例如，一个临床医生可能对分割肝脏肿瘤感兴趣，而另一个可能需要分割整个肝脏和周围器官。此外，成像方式的多样性也带来了另一个挑战。像CT和MR这样的成像方式产生3D图像，而像X射线和超声这样的成像方式则产生2D图像。任务定义和成像方式的这些可变性使得设计一个能够准确预见并满足不同用户多样化需求的全自动模型变得复杂。

考虑到这些挑战，我们认为开发一个可提示的2D分割模型是一个更实际的方法。该模型可以根据用户提供的提示轻松适应特定任务，提供增强的灵活性和适应性。它还能通过将3D图像作为一系列2D切片来处理，同时处理2D和3D图像。典型的用户提示包括点和边界框，我们在补充图1中展示了一些使用不同提示的分割示例。可以发现，基于点的提示存在歧义，需要多次用户干预，而边界框提示可以清晰地指定感兴趣区域（ROI），仅需少量用户干预，减少歧义并消除试错。我们遵循SAM [4]中的网络架构，包括一个图像编码器、一个提示编码器和一个掩码解码器（图1c）。图像编码器[7]将输入图像映射到高维图像嵌入空间。提示编码器将用户绘制的边界框通过位置编码[8]转换为特征表示。最后，掩码解码器使用交叉注意力[9]（方法）融合图像嵌入和提示特征。

我们通过内部验证和外部验证评估了MedSAM，并将其与最先进的分割基础模型SAM [4]和专家U-Net模型[3]进行了比较。内部验证包含超过70个分割任务（补充表5-8，图2-4），图2a显示了12个代表性分割任务的Dice相似系数（DSC）得分。总体而言，尽管SAM在一些RGB图像分割任务上表现出色，如皮肤癌分割（88.8%）在皮肤镜图像和息肉（94.1%）分割在内窥镜图像中，但它在大多数CT、MR和灰度图像分割任务上表现不佳。这可能归因于SAM在多种RGB图像上的训练，以及由于其独特的外观，皮肤镜和内窥镜图像中的许多分割目标相对容易分割。MedSAM和U-Net在大多数分割任务上都大幅度超越了SAM（p < 0.05），这是预期的，因为它们在医学图像数据集上的调整或训练。与U-Net专家模型相比，MedSAM在大多数任务上仍然表现更好。例如，MedSAM在涉及脑内出血CT、胶质瘤MR T1、气胸X线摄影和息肉内窥镜图像的分割任务中分别实现了94.0%（四分位数范围（IQR）：91.5-94.9%）、94.4%（IQR：91.6-95.8%）、81.5%（IQR：75.1-86.8%）和98.4%（IQR：97.9-98.9%）的中位数DSC得分，分别比U-Net专家模型的性能高出5%、6.6%、5.1%和3.6%。在几个RGB图像分割任务上，例如皮肤癌分割，U-Net和MedSAM之间的性能相当（95.1%对95.2%）。这些分割目标通常具有清晰的边界和良好的对比度，因此相对容易分割。值得注意的是，U-Net是针对每个类别单独训练的（方法），但MedSAM是一个通用模型，只训练了一次。图2b展示了SAM、U-Net和MedSAM在CT、MR、超声和内窥镜图像上的一些分割示例。SAM倾向于分割对比度高或边界清晰的区域，这容易犯下分割不足或过度分割的错误。虽然U-Net专家模型提供了更好的分割质量，但它们仍然难以处理具有弱边界的目标。相比之下，MedSAM能够准确分割各种成像条件下的广泛目标，即使是具有弱或缺失边界的对象（补充图5-7）。

外部验证包括超过30个分割任务，所有这些任务都来自新数据集或未见过的分割目标（补充表9-11，图2，8-9）。图2c显示了12个典型分割任务的DSC得分。SAM在大多数CT和MR分割任务上继续表现出较低的性能，而U-Net专家模型并没有一致地超越SAM（例如，在CT图像中的肺癌分割（55.8%对64.2%）），这表明它们在未见数据集上的泛化能力有限。相比之下，MedSAM持续提供了优越的性能。例如，MedSAM在鼻咽癌分割任务上获得了90.3%（IQR：87.8-93.2%）的中位数DSC得分，比SAM和专家U-Net分别提高了53.3%和24.5%。值得注意的是，MedSAM在一些未见过的成像方式上也取得了更好的性能（例如，腹部T1 Inphase和Outphase），比SAM和专家U-Net模型提高了3-7%。在灰度和RGB图像分割任务上，MedSAM和U-Net专家模型在中位数DSC得分方面的性能相当，但MedSAM的异常值更少。图2d展示了四个定性评估的分割示例，揭示了尽管所有方法都有处理简单分割目标的能力，但MedSAM在分割具有挑战性的目标上表现更好，例如CT图像中的肝癌和MR图像中的宫颈癌（补充图10）。此外，我们对SAM和MedSAM之间的图像嵌入的显著性图进行了可视化和比较分析（补充图11）。值得注意的是，MedSAM的特征展示了更丰富的语义信息，特别是与高度相关的解剖结构有关。总的来说，这些结果表明MedSAM在新数据集上具有强大的泛化能力。

除了广泛的适用性，我们进一步展示了MedSAM促进了肿瘤负担的精确量化，这是肿瘤学实践中的一个关键生物标志物[10]（图2e）。具体来说，我们使用MedSAM分割结果计算了肾脏、结肠、肝脏和胰腺癌的肿瘤体积，并将其与专家分割得出的体积进行了比较。从MedSAM和专家评估得到的肿瘤体积显示出很高的皮尔逊相关性（r = 0.99），强调了MedSAM的分割结果可以有效用于精确的肿瘤负担量化。最后，我们将MedSAM的性能与六位人类专家在前列腺分割方面的表现进行了比较（方法）。发现MedSAM的表现与四位人类专家相当，甚至超过了两位专家，突显了其作为临床实践中医学图像分割的强大工具的潜力。

# 讨论

我们介绍了MedSAM，这是一个由深度学习驱动的基础模型，旨在对多种医学成像方式下的广泛解剖结构和病变进行分割。MedSAM在一个精心组装的大规模数据集上进行训练，该数据集包含超过一百万对医学图像掩码对。其可提示的配置在自动化和定制化之间取得了最佳平衡，使MedSAM成为一个通用的医学图像分割工具。

通过包含内部和外部验证的全面评估，MedSAM已经证明了其在分割多样化目标和处理新数据及任务方面的强大的泛化能力。其性能不仅显著超过了现有的最先进分割基础模型，而且还能与专家模型相媲美甚至超越。通过提供解剖结构和病理区域的精确划分，MedSAM有助于计算各种定量测量值，这些测量值作为生物标志物具有重要作用。例如，在肿瘤学领域，MedSAM可以在生成精确的肿瘤分割结果方面发挥关键作用，从而实现后续的肿瘤体积计算，这是评估疾病进展和治疗反应的关键生物标志物。

尽管MedSAM拥有强大的能力，但它确实存在某些局限性。其中一个限制是训练集中成像方式的不平衡，其中CT、MRI和内窥镜图像在数据集中占主导地位。这可能会影响模型对代表性较低的成像方式，如乳腺摄影的性能。另一个局限性是它在分割类似血管分支结构方面的困难，因为在这种情况下，边界框提示可能是模糊的。例如，在眼底图像中，动脉和静脉共享相同的边界框。然而，这些局限性并不减少MedSAM的实用性。由于MedSAM已经从大规模训练集中学习了丰富且具有代表性的医学图像特征，它可以被微调，以有效地分割来自代表性较低的成像方式或类似血管这样的复杂结构的新任务。

总之，这项研究突出了构建一个能够处理众多分割任务的单一基础模型的可行性，从而消除了对特定任务模型的需求。作为医学图像分割中的第一个基础模型，MedSAM具有巨大的潜力，可以加速新诊断和治疗工具的发展，并最终有助于改善患者护理[11]。

# 方法

## 研究设计

分割是许多基于医学图像的临床分析任务中的一个重要步骤。例如，在脑肿瘤成像中，MR图像的分割可以帮助确定肿瘤的位置、大小和类型，这对于手术规划和预后至关重要[12]。在心脏成像中，超声心动图或MRI中左心室等结构的分割对于评估心脏功能和诊断心力衰竭等疾病是必不可少的[13]。在肺部成像中，胸部X光或CT图像中肺野的分割对于诊断和监测慢性阻塞性肺病（COPD）和COVID-19等疾病至关重要[14]。在过去的几十年中，医学图像分割领域见证了众多方法的发展[15]。然而，许多现有方法的一个显著限制是它们的任务特定性和数据集特定性，使它们无法推广到新数据集和目标上。这一限制阻碍了它们在临床实践的广泛应用。

最近的深度学习领域的发展，特别是引入了像Segment Anything Model（SAM）[4]这样的基础模型，已经显示出在解决医学图像分割中的泛化挑战方面的巨大潜力。基础模型利用大量的训练数据和强大的架构来捕捉图像中的复杂模式和关系。为了研究SAM在医学领域的适用性，我们使用SAM对一个代表性的腹部CT图像进行了分割实验。SAM提供了三种主要的分割模式：全自动分割、边界框模式和点模式。尽管文本提示被纳入了SAM的训练流程中，但需要注意的是，这是一个概念验证，而不是SAM官方代码库中公开可用的功能。补充图1展示了三种分割模式所得的结果。这些结果是使用在线演示生成的，可在 https://segment-anything.com/demo 访问。在“自动分割所有”模式下，SAM根据图像强度将整个图像划分为六个不同的区域（补充图1b）。然而，由于两个主要原因，这种分割结果的实用性受到限制。首先，分割区域缺乏语义标签，使得解释特定解剖结构变得具有挑战性。其次，在临床场景中，医疗保健专业人员主要关注有意义的感兴趣区域（ROIs），如肝脏、肾脏、脾脏和病变。

另一方面，基于边界框的分割模式展示了有希望的结果，特别是对于右肾的分割，这是通过提供左上角和右下角的点（补充图2c）实现的。对于基于点的分割模式（补充图1d），我们最初提供了一个代表右肾中心的单一前景点。然而，SAM过度分割了整个腹部。为了纠正这个问题，我们在过度分割的区域内引入了一个背景点。这个调整导致分割掩码缩小，仅包括肝脏和右肾。最后，通过在肝脏上添加另一个背景点，我们得到了所需的肾脏分割。

总结来说，在使用SAM进行医学图像分割时，分割所有模式经常产生缺乏实际用途的分区，而基于点的模式可能存在歧义，并且需要多次迭代进行预测和校正。相反，基于边界框的模式通过精确定义感兴趣区域（ROI）并始终产生合理的分割结果，提供了明显的优势，消除了反复试错的需要。然而，尽管具有前景潜力，但最近的研究显示，SAM在各种医学图像分割任务中提供令人满意的分割结果方面遇到了挑战。鉴于这些限制，本研究的目标是开发一个强大的分割基础模型，能够有效地解决广泛的分割目标和多样化的成像方式。随后的小节提供了关于训练和（内部和外部）验证集、网络架构、训练协议以及与最新基准比较的关键方面的全面概述。

## 数据集策划和预处理

我们通过整合来自互联网上各种来源公开可用的医学图像分割数据集的图像，策划了一个全面的数据集。这些来源包括位于 https://www.cancerimagingarchive.net/ 的癌症影像档案库（TCIA）、位于 https://www.kaggle.com/ 的Kaggle、位于 https://grand-challenge.org/challenges/ 的Grand-Challenge、位于 https://www.nature.com/sdata/ 的科学数据、位于 https://codalab.lisn.upsaclay.fr/ 的CodaLab，以及医学图像计算和计算机辅助干预学会（MICCAI）在 http://www.miccai.org/ 内的分割挑战。所使用的数据集的完整列表在补充表1-4中呈现。

原始的3D数据集包括DICOM、nrrd或mhd格式的计算机断层扫描（CT）和磁共振（MR）图像。为了确保与开发医学图像深度学习模型的一致性和兼容性，我们将图像转换为广泛使用的NifTI格式。此外，灰度图像（如X射线和超声图像）以及RGB图像（包括内窥镜、皮肤镜、眼底和病理图像）被转换为png格式。我们应用了几个专门的标准来提高数据集的质量和一致性，包括不完整的图像和具有分支结构的分割目标、不准确的注释以及微小体积。值得注意的是，不同成像方式的图像强度变化显著。例如，CT图像的强度值范围从-2000到2000，而MR图像的范围为0到3000。在内窥镜和超声图像中，强度值通常从0到255。为了便于稳定训练，我们对所有图像执行了强度归一化，确保它们共享相同的强度范围。

对于CT图像，我们最初使用典型的窗口宽度和水平值来归一化豪斯菲尔德单位，如 https://radiopaedia.org/articles/windowing-ct 中所述。随后，强度值被重新缩放到[0, 255]的范围。对于MR、X射线、超声、乳腺摄影和光学相干断层扫描（OCT）图像，我们在将它们重新缩放到[0, 255]的范围之前，将强度值限制在0.95和99.5百分位数之间的范围内。对于RGB图像（例如，内窥镜、皮肤镜、眼底和病理图像），如果它们已经在预期的强度范围[0, 255]内，则它们的强度保持不变。然而，如果它们超出了这个范围，我们使用最大最小值归一化将强度值重新缩放到[0, 255]。最后，为了满足模型的输入要求，所有图像都被调整到统一的大小1024×1024×3。对于全幻灯片病理图像，我们使用滑动窗口方法提取了图像块。对于3D CT和MR图像，每个2D切片被调整到1024×1024的大小，并且通道被重复三次以保持一致性。剩余的2D图像直接调整到1024×1024×3的大小。调整图像大小时使用了双三次插值，而最近邻插值被用于调整掩码的大小，以保留它们的精确边界并避免引入不希望的伪影。这些标准化程序确保了所有图像的一致性和兼容性，并促进了它们在模型训练和评估流水线的后续阶段的无缝集成。

## 网络架构

本研究使用的网络建立在变换器架构[9]上，该架构在自然语言处理[16]和图像识别任务[7]等多个领域展示了显著的有效性。具体来说，该网络包括一个基于视觉变换器（ViT）的图像编码器，负责提取图像特征，一个用于整合用户交互（边界框）的提示编码器，以及一个使用图像嵌入、提示嵌入和输出令牌生成分割结果和置信度分数的掩码解码器。

为了在分割性能和计算效率之间取得平衡，我们采用了基础ViT模型作为图像编码器，因为广泛的评估表明，更大的ViT模型，如ViT Large和ViT Huge，仅提供了微小的准确性提升[4]，同时显著增加了计算需求。具体来说，基础ViT模型由12个变换层[9]组成，每个块包括一个多头自注意力块和一个包含层归一化的多层感知器（MLP）块[17]。使用掩码自编码器建模[18]进行预训练，随后在SAM数据集[4]上进行全监督训练。输入图像（1024×1024×3）被重塑成大小为16×16×3的展平2D块序列，通过图像编码器后得到64×64的特征尺寸图像嵌入，这是16倍的下采样。提示编码器将边界框提示的角点映射到256维的向量嵌入[8]。特别是，每个边界框由左上角点和右下角点的嵌入对表示。为了在计算出图像嵌入后促进实时用户交互，采用了轻量级掩码解码器架构。它包括两个变换层[9]用于融合图像嵌入和提示编码，以及两个转置卷积层以将嵌入分辨率增强到256×256。随后，嵌入经过sigmoid激活，然后通过双线性插值以匹配输入尺寸。

## 训练协议和实验设置

在数据预处理期间，我们获得了1,090,486对医学图像-掩码对用于模型开发（不包括外部验证集表1-4）。对于内部验证，我们将数据集随机分成80%、10%和10%，分别用于训练、调整和验证。这种设置允许我们在训练期间监控模型在调整集上的性能，并调整其参数以防止过拟合。对于外部验证，我们使用了在训练期间模型未见过的保留数据集。这些数据集为模型的泛化能力提供了严格的测试，因为它们代表了模型以前未遇到过的新患者、成像条件以及潜在的新分割任务。通过评估MedSAM在这些未见数据集上的性能，我们可以真实地了解MedSAM在现实世界临床设置中可能的表现，其中它将需要处理数据中的广泛变异性和不可预测性。训练和验证是独立的。

模型使用预训练的SAM模型和ViT-Base模型进行初始化。我们固定了提示编码器，因为它已经可以编码边界框提示。在训练期间，图像编码器和掩码解码器中的所有可训练参数都进行了更新。具体来说，图像编码器和掩码解码器的可训练参数数量分别为89,670,912和4,058,340。边界框提示是通过在真值掩码上加上0-20像素的随机扰动来模拟的。损失函数是Dice损失和交叉熵损失的未加权和，这在各种分割任务中已被证明是稳健的[3]。具体来说，设S、G分别表示分割结果和真值。si、gi分别表示体素i的预测分割和真值。N是图像I中体素的数量。交叉熵损失定义为：

而Dice损失定义为：

最终损失L定义为：

该网络通过AdamW [19]优化器（β1 = 0.9, β2 = 0.999）进行优化，初始学习率为1e-4，权重衰减为0.01。批量大小为160，没有使用数据增强。该模型在20个A100（80G）GPU上训练了100个周期，并选择了最后的检查点作为最终模型。

此外，为了全面评估MedSAM的性能，我们与最先进的分割基础模型SAM [4]和专家模型进行了比较分析。具体来说，我们将训练图像分为四类：CT图像、MR图像、灰度图像（包括胸部X射线（CXR）、超声、乳腺摄影和OCT图像），以及RGB图像（包括病理学、内窥镜和皮肤镜图像）。对于每个类别，我们基于nnU-Net训练了一个专家模型，它在许多分割任务上都取得了SOTA性能[3]。为了生成U-Net的训练数据，我们裁剪了边界框内的图像和相应的掩码。为了进行公平比较，MedSAM和U-Net专家模型都在同一数据分割上进行了训练。主要的区别在于训练方法：MedSAM在完整的训练集上进行了一次训练，而U-Net专家模型则分别在每个对应于一个类别的子集上进行了训练。

除了与SAM和U-Net专家模型的比较分析外，我们还通过将MedSAM与六位专家在前列腺MR图像分割数据集（52个案例）上的表现进行比较，进一步评估了MedSAM的性能。对于每个案例，六位专家提供了他们各自的分割结果，真值基于多数投票确定。我们计算了每个案例和专家的DSC得分，然后将它们与MedSAM的结果进行比较。

# 评估指标

我们遵循Metric Reload [20]中的建议，使用Dice相似系数（DSC）和归一化表面距离（NSD）来定量评估分割结果。DSC是一个基于区域的分割度量，旨在评估真实值和分割结果之间的区域重叠，定义为：

NSD是一个基于边界的度量，旨在评估在给定容差下真实值和分割结果之间的边界共识，定义为：

其中,分别表示在容差 τ 下真实值和分割表面的边界区域。在本文中，我们将容差 τ 设置为2。

## 统计分析

为了统计分析并比较多上述三种方法（MedSAM、SAM和专家模型）的性能，我们采用了Wilcoxon符号秩检验。这种非参数检验非常适合比较配对样本，特别是当数据不符合正态分布假设时非常有用。这种分析使我们能够确定是否有任何方法在分割性能上表现出统计上的优越性，与其他方法相比，为我们提供了关于三种评估方法——SAM、U-Net专家模型和MedSAM的比较有效性的宝贵见解。Wilcoxon符号秩检验的结果标记在DSC和NSD得分表（补充表5-11）上。

## 数据可用性

本研究中的所有数据集均来自公共数据集。下载链接已在补充表12中提供。

## 代码可用性

所有代码均使用Python (3.10)实现，使用Pytorch (2.0)作为基础的深度学习框架。我们还使用了多个Python包进行数据分析和结果可视化，包括SimpleITK (2.2.1)、nibabel (5.1.0)、torchvision (0.15.2)、numpy (1.24.3)、scikit-image (0.20.0)、opencv-python (4.7.0)、scipy (1.10.1)、pandas (2.0.2)、matplotlib (3.7.1) 和 plotly (5.15.0)。Biorender被用来创建图1a。训练脚本、推理脚本和训练好的模型已在 https://github.com/bowang-lab/MedSAM 上公开可用。

## 致谢

本文的作者非常感谢所有为社区提供公共医学图像的数据所有者。我们还要感谢Meta AI将“Segment Anything”的源代码公开提供给社区。

图表、致谢及参考文献已略去(见原文)。

原文出处：MA J, HE Y, LI F, 等. Segment Anything in Medical Images[J].