Segment Anything 系列论文阅读报告

周炜杰 20231610

阅读的论文：Segment Anything

https://arxiv.org/abs/2304.02643

1. 背景及意义

1.1 背景

在自然语言处理（NLP）领域，在大规模网络数据集上预训练的大语言模型（LLM）展现出强大的零样本泛化和少样本泛化能力。这些“基础模型”能处理训练时没有见过的任务。提示工程强化了大语言模型的零样本泛化和少样本泛化能力。

而在 CV 领域，CLIP 和 ALIGN 等研究探索了基础模型，但这些模型主要关注图像与文本的对齐，而 CV 领域包括大量其他任务。对于图像分割这样的任务，不仅缺少统一的基础模型，也缺少大量的训练数据。

受基础模型和提示工程在 NLP 领域的巨大成功启发，针对图像分割领域基础模型缺失与数据匮乏这一背景，SAM 尝试建立图像分割的视觉基础模型。

1.2 任务

SAM 的核心任务是“可提示分割”（promptable segmentation）要求给定任何形式（点、框或语义信息等）的分割提示，模型返回一个有效的分割掩码。

如同语言提示词一样，SAM 的提示词也存在歧义。例如，一个落在衬衫上的点既有可能指代“衬衫”，也有可能指代“穿衬衫的人”。因此，我们要求模型至少能输出一个针对该提示的分割掩码。

1.3 现状分析

LLM 在 NLP 领域取得的了巨大的成功，CV 领域也需要探索基础模型，但在 SAM 之前，CV 领域的基础模型主要集中在视觉-语言对齐层面。尽管在语义理解上取得了进展，但这些模型缺乏对像素级精细结构的处理能力，对于图像分割任务，当时缺乏一个具备海量知识的能处理通用任务的基础模型。

在 SAM 之前，图像分割任务，例如语义分割，实例分割和全景分割等任务是割裂的，每种任务需要专门的架构设计，缺乏统一的建模方式。全监督训练的模

型通常只能分割训练集中预定义的类别，无法处理未知物体。

1.4 挑战和问题

数据匮乏：与 NLP 领域天然的拥有海量的互联网数据不同，世界上不存在大规模的高质量图像-掩码对。而现有的分割数据集规模小，而且往往只标注了特定类别，缺乏训练基础模型所需要的规模和多样性。

提示的歧义：与 NLP 领域的提示词一样，图像分割中的提示也具有歧义。如果只允许唯一输出，模型在面对歧义时会 “平均化”多种可能的数据，带来错误的掩码。因此模型必须具备输出多个结果的能力。

1.5 价值和意义

SAM 提出了首个分割领域的基础模型，通过大规模预训练达到了强大的零样本泛化能力。它证明了在 CV 领域中，通过提示引导模型输出也同样有效。同时，SAM 将以往割裂的语义分割、实例分割等分割任务统一到一个框架下。

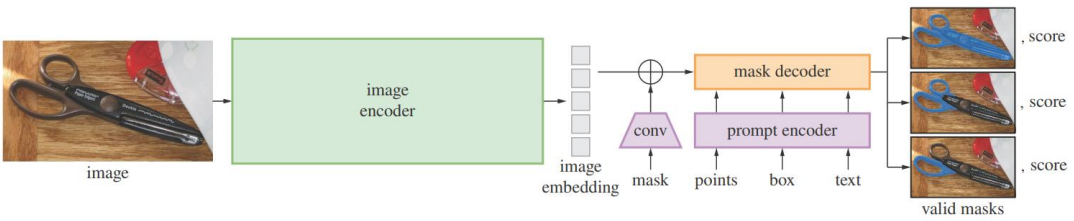
为了完成这一任务，构建了 SA-1B 数据集，推动了视觉模型的发展。

2. 方法具体描述

2.1 流程，整体框架

SAM 的整体框架由图像编码器，提示编码器和掩码解码器三个模块组成。图像编码器是一个参数量巨大的重型网络（基于 ViT），将图像映射为高位特征嵌入。提示编码器是个轻型模块，将用户输入的提示转化为嵌入向量。掩码解码器是个轻量级的 Transformer 解码器，它接收图像嵌入和提示嵌入，输出最终的分割掩码。

工作流程如下图所示。其中，图像编码器负责进行图像特征提取，在处理每张图片是仅运行一次。生成的图像嵌入可以被缓存，供后续的提示交互重复使用，减少计算成本。用户的输入的提示通过提示编码器实时转换为提示嵌入。图像嵌入和提示嵌入共同输入到一个掩码解码器。该解码器极其轻量，能在 CPU 环境下 50ms 内输出分割掩码。



2.2 创新

SAM 在方法上的创新主要包括以下三方面。

第一方面是对歧义的处理，针对一个提示可能对应多个有效物体掩码的问题，SAM 设计了针对单个提示并行预测 3 个掩码（整体、部分和子部分）的方案，并对每个掩码按置信度进行排序。

另一方面是效率的提升，通过将图像编码（大量计算）与提示编码（少量计算）解耦，使得模型在图像特征提取后，能满足交互式分割的实时性要求。

为解决分割数据稀缺的问题，SAM 采用了一个包含三个阶段的数据生产流水线。首先是辅助手动阶段，利用早期版本的 SAM 模型通过点击生成掩码；随后进入半自动阶段，模型已经能够自动分割图像中显著的物体，标注员的任务转变为标注模型未识别出的低置信度或不显著物体；最终是全自动阶段，利用网格点提示激发模型对图像进行全要素分割，并配合后处理机制生成高质量掩码。

2.3 理论，算法

图像端使用了一个基于 MAE 预训练的重型 ViT 作为图像编码器。为了适应分割任务对细节的高要求，该编码器经过了针对高分辨率输入的最小化适配，采用了窗口注意力与全局注意力相结合的机制。

提示端则采用了一个轻量级的提示编码器，它将稀疏的几何提示（点、框）通过位置编码与可学习的类型嵌入相加映射为向量，将文本提示通过 CLIP 的文本编码器映射为向量，并将稠密的掩码提示通过卷积网络处理后与图像嵌入进行逐元素相加，从而将多模态的交互信息统一映射到高维嵌入空间。

在解码层面，设计了一个极其轻量的掩码解码器（Mask Decoder），其结构源自修改后的 Transformer 解码器块。该模块引入了一种双向交叉注意力机制：通过提示对图像的注意力聚焦特定区域，图像对提示的注意力检索提示的语义表示。经过两层解码块的处理后，算法并没有直接生成像素级的掩码图，而是采用了一种动态预测机制：解码器输出一个上采样后的图像特征图和一个与输出 Token 对应的向量，该向量被映射为一个动态线性分类器的权重。最终，通过计算该分类器与图像特征图的点积，生成每个像素的前景概率图。

在训练算法上，采用了 Focal Loss 和 Dice Loss 的线性组合来监督生成质量，使用“最小损失”，即在训练时仅对预测结果中与真值 IoU 最高的掩码计算

梯度。这种算法设计保证了模型不会在面对歧义时通过取平均带来错误，而是学会输出多样化且合理的分割结果。

3. 实验分析与讨论

3.1 实验设置，数据

SAM 训练使用的数据集是 SA-1B 数据集。该数据集包含 1100 万张图像和 11 亿个掩码。其中最为关键的特征是，99.1%的掩码是由 SAM 模型全自动生成的合成数据（对应上文提到的“全自动阶段”）。为了验证 SAM 作为基础模型的零样本迁移能力，使用 23 个数据集作为测试集，这些测试集没有用在 SAM 的训练过程中，包括水下、航拍、医学影像等领域。在边缘检测等其他任务，使用 BSDS500 等专门的数据集。

使用 AdamW 优化器，在 256 个 GPU 上训练。图像输入分辨率为 1024*1024。因为数据量足够大，不进行强烈的数据增强。为了处理歧义性，模型针对每个提示预测 3 个掩码。

分割的核心指标是 mIoU，包括标准 mIoU 和 Oracle mIoU。Oracle mIoU 针对 SAM 输出的 3 个掩码，选取与真值 IoU 最高的那一个进行计算。它剥离了歧义性问题，专注于评估模型能否分割出正确的物体。

由于分割任务，特别是单点提示分割，存在歧义，低 mIoU 不代表分割错误。因此引入了人类视觉评分，由人类对掩码质量进行评分，范围从 1 到 10。用于补充 mIoU。

对于特定任务，使用特定任务指标。例如，边缘检测使用 BSDS500 数据集的标准指标：ODS (Optimal Dataset Scale), OIS (Optimal Image Scale), AP (Average Precision), R50 (Recall at 50% precision)。

3.2 实验对比结果与分析（对比实验，消融实验，超参实验等）

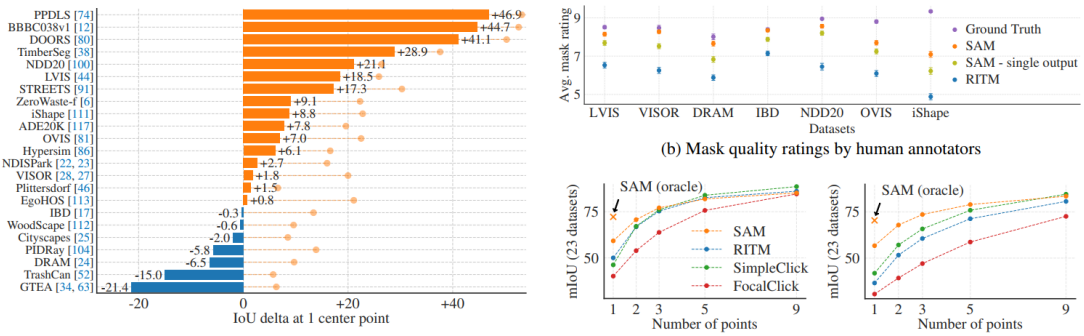
零样本单点分割性能分析

在核心的提示分割任务中，实验采用了包含 23 个不同领域数据集的测试基准，评估模型在单点提示下的表现。对比最先进的交互式分割模型 RITM，SAM 在 23 个数据集中的 16 个上均优于 RITM。考虑到单点提示的歧义性，实验引入了 Oracle mIoU 指标，从 SAM 输出的三个候选掩码中选取与真值最匹配的一个进行评估，SAM 在这一指标下，在所有 23 个数据集上均大幅超越了 RITM。

在人类评估指标上，结果显示，即使在部分 SAM 自动指标较低的数据集上，人类标注员对 SAM 生成掩码的质量评分（1-10 分制）也始终显著高于 RITM。

实验进一步引入了 SimpleClick 和 FocalClick 作为额外的基线模型进行对比，全面评估模型在单点与多点提示下的交互式分割能力。在单点提示下，这些基线模型的表现均低于 RITM 和 SAM，带有 Oracle 设置的 SAM 展现出远超其他模型的性能。然而，随着提示点数量从 1 逐渐增加至 9，观察发现各方法之间的性能差距呈现收敛趋势。这符合预期，因为更多的提示点消除了歧义性，降低了任务难度，而 SAM 的核心优势正是在于处理极少样本提示时的鲁棒性和对歧义的解析能力。

在上述实验的基础上，将中心点采样修改为随机点采样。在单点分割时，SAM 与基线模型之间的性能差距进一步扩大。无论采用何种采样方式，SAM 均能维持相当的性能水平，展现了极高的稳定性。



(a) SAM *vs.* RITM [92] on 23 datasets

(b) Mask quality ratings by human annotators

(c) Center points (default)

(d) Random points

Figure 9: Point to mask evaluation on 23 datasets. (a) Mean IoU of SAM and the strongest single point segmenter, RITM [92]. Due to ambiguity, a single mask may not match ground truth; circles show "oracle" results of the most relevant of SAM's 3 predictions. (b) Per-dataset comparison of mask quality ratings by annotators from 1 (worst) to 10 (best). All methods use the ground truth mask center as the prompt. (c, d) mIoU with varying number of points. SAM significantly outperforms prior interactive segmenters with 1 point and is on par with more points. Low absolute mIoU at 1 point is the result of ambiguity.

下游任务的零样本迁移

为了验证作为基础模型的泛化性，在未经过专门训练的边缘检测、对象建议和实例分割等任务上评估 SAM。在 BSDS500 边缘检测任务中，生成的边缘图在定性上非常合理，定量指标上实现了极高的召回率（R50 为 0.928），优于传统的零样本方法。尽管其精度略低于全监督方法，但这主要是因为 SAM 输出了所有合理的边缘，而未学习特定数据集特定的标注偏见。

method	year	ODS	OIS	AP	R50
HED [108]	2015	.788	.808	.840	.923
EDETR [79]	2022	.840	.858	.896	.930
zero-shot transfer methods:					
Sobel filter	1968	.539	-	-	-
Canny [13]	1986	.600	.640	.580	-
Felz-Hutt [35]	2004	.610	.640	.560	-
SAM	2023	.768	.786	.794	.928

Table 3: Zero-shot transfer to edge detection on BSDS500.

在 LVIS 数据集的对象建议任务中，SAM 在中大型物体及稀有类别（Rare categories）上的表现优于经过全监督训练的 ViTDet-H，在小型物体和常见类别上，SAM 略逊于 ViTDet-H。

	mask AR@1000						
method	all	small	med.	large	freq.	com.	rare
ViTDet-H [62]	63.0	51.7	80.8	87.0	63.1	63.3	58.3
zero-shot transfer methods:							
SAM – single out.	54.9	42.8	76.7	74.4	54.7	59.8	62.0
SAM	59.3	45.5	81.6	86.9	59.1	63.9	65.8

Table 4: Object proposal generation on LVIS v1. SAM is applied zero-shot, *i.e.* it was not trained for object proposal generation nor did it access LVIS images or annotations.

在实例分割任务中，当使用 ViTDet 的检测框作为提示时，SAM 的 Mask AP 虽然略低于全监督基线，但进一步的人类视觉评估和定性分析揭示，SAM 的掩码边界通常比 ViTDet 更清晰、更精确，并且能够正确处理遮挡，其 AP 较低反而部分归因于 COCO/LVIS 数据集标注本身存在的偏差。

	COCO [66]				LVIS v1 [44]			
method	AP	AP^S	AP^M	AP^L	AP	AP^S	AP^M	AP^L
ViTDet-H [62]	51.0	32.0	54.3	68.9	46.6	35.0	58.0	66.3
zero-shot transfer methods (segmentation module only):								
SAM	46.5	30.8	51.0	61.7	44.7	32.5	57.6	65.5

Table 5: Instance segmentation results. SAM is prompted with ViTDet boxes to do zero-shot segmentation. The fully-supervised ViTDet outperforms SAM, but the gap shrinks on the higher-quality LVIS masks. Interestingly, SAM outperforms ViTDet according to human ratings (see Fig. 11).

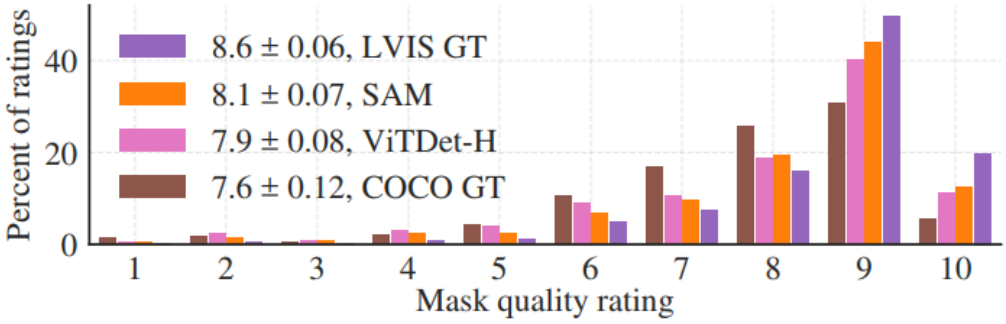


Figure 11: Mask quality rating distribution from our human study for ViTDet and SAM, both applied to LVIS ground truth boxes. We also report LVIS and COCO ground truth quality. The legend shows rating means and 95% confidence intervals. Despite its lower AP (Table 5), SAM has higher ratings than ViTDet, suggesting that ViTDet exploits biases in the COCO and LVIS training data.

通过定性分析，验证 SAM 处理文本提示的能力。实验表明，SAM 能够根据简单的文本提示或短语成功分割对象。当仅靠文本提示无法准确分割时，额外添加一个点提示通常能有效修正结果。

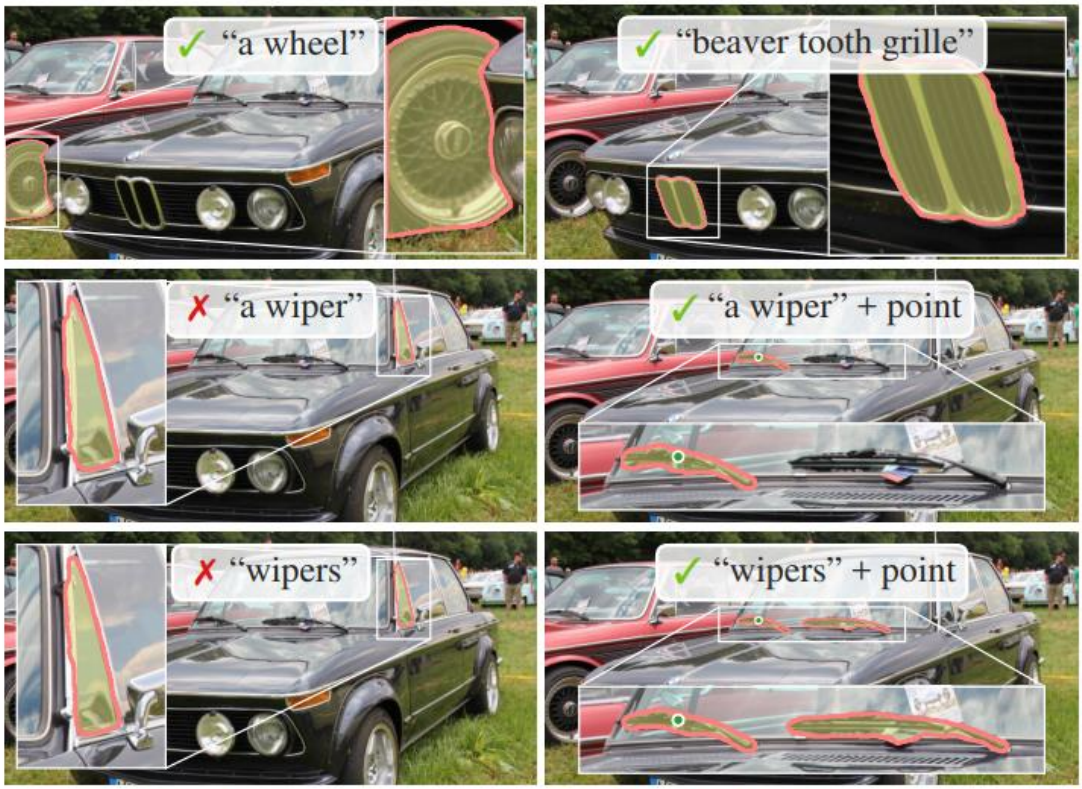


Figure 12: Zero-shot text-to-mask. SAM can work with simple and nuanced text prompts. When SAM fails to make a correct prediction, an additional point prompt can help.

Results. We show qualitative results in Fig. 12. SAM can segment objects based on simple text prompts like "a wheel" as well as phrases like "beaver tooth grille". When SAM fails to pick the right object from a text prompt only, an additional point often fixes the prediction, similar to [31].

消融实验

针对数据引擎不同阶段及数据规模的影响，实验进行了详尽的消融分析。首先，关于数据来源的消融显示，随着数据引擎从“手动”到“全自动”阶段的推进，模型性能稳步提升。值得注意的是，仅使用“全自动”阶段生成的掩码训练的模型，其性能与使用所有阶段数据混合训练的模型几乎相当，这证明了大规模

自动生成数据的有效性。

其次，在数据量的扩展性上，实验发现当训练数据从完整的 1100 万张（11M）下采样至 100 万张（10%）时，模型在 23 个数据集上的性能下降非常小。然而，当数据量进一步减少至 10 万张（1%）时，性能则出现显著滑坡。这表明，虽然 SA-1B 数据集的规模带来了收益，但在约 100 万张图像的量级上，模型对于一般特征的学习已趋于饱和，展现了极高的数据效率。

在模型架构方面，实验对比了基于 ViT-B、ViT-L 和 ViT-H 的图像编码器。结果表明，从 ViT-B 扩展到 ViT-H 能带来显著的性能提升，但在 ViT-L 到 ViT-H 的过程中，收益不高。ViT-L 可能是一个在性能与计算成本之间较好的平衡点。

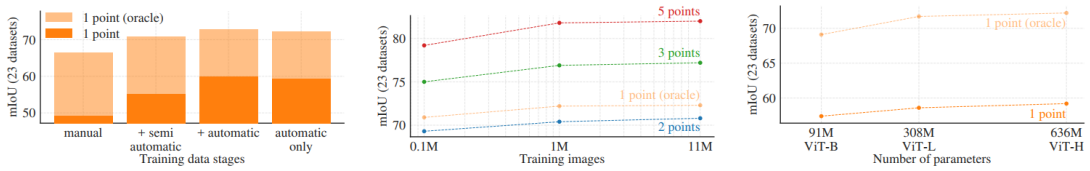


Figure 13: Ablation studies of our data engine stages, image encoder scaling, and training data scaling. (Left) Each data engine stage leads to improvements on our 23 dataset suite, and training with only the automatic data (our default) yields similar results to using data from all three stages. (Middle) SAM trained with ~10% of SA-1B and full SA-1B is comparable. We train with all 11M images by default, but using 1M images is a reasonable practical setting. (Right) Scaling SAM's image encoder shows meaningful, yet saturating gains. Nevertheless, smaller image encoders may be preferred in certain settings.

3.3 结果讨论

实验结果表明，SAM 具备基础模型的特性，符合基础模型的定义：在大规模数据集上训练，通过提示工程适应下游任务，在零样本泛化中表现出优异性能。SAM 在显微镜、水下等完全未见过的领域表现出的鲁棒性，证明了从大规模数据中学习通用物体概念的有效性。

SAM 被证明可以作为一个更大系统的模块，例如与 CLIP 结合完成文本到掩码的分割，以及使用检测框作为提示，完成实力分割任务。

实验数据支持了 SAM 的多掩码输出设计。在交互式分割中，提供“合理的选项”比提供“唯一的平均值”更符合实际需求。

SA-1B 的成功表明，即使是由模型生成的“伪标签”，只要规模足够大且质量足够高，也能训练出高准度的基础模型。这为解决 CV 领域数据匮乏问题指明了方向。

4. 阅读心得

4.1 和课堂的哪一部分知识相关

本文使用了 Transformer 架构和注意力机制，使用了 ViT 模型，这些都是课上学过的。本文是计算机视觉领域图像分割领域的作品，这也是在课上讲过的。SAM 使用的数据引擎体现了半监督学习的思想，展示了通过生成伪标签来构造大量数据训练模型，突破数据匮乏的限制，提示模型性能。

4.2 方法的优势和劣势

SAM 最大的贡献是建立了 CV 领域图像分割基础模型的范式。其泛化能力，尤其是零样本泛化能力极强。同时，SAM 统一了图像分割领域的若干下游任务。其次，SAM 的交互性能设计极其出色，通过图像嵌入与提示嵌入的解耦，实现了端侧的低延迟响应，极大地提升了用户体验。此外，歧义处理实现了模型在模糊提示，例如单点提示下的高准确率。

站在 2026 年来看，SAM 也有不少局限性，我们通过 SAM 系列后续的演进也能看出来。首先是时空能力的确实，SAM 1 本质上是一个静态图像处理模型，无法利用视频的时序上下文信息。在用于视频时，会导致物体类别频繁切换和闪烁。其次是语义认知不足。SAM 1 的分割是类别无关的，无法回答“分割的是什么”，也无法响应高级语义概念的指令。最后，虽然解码器轻量，但是图像编码器计算开销仍然巨大，在端侧部署仍有困难。

4.3 如何进行改进

我们可以通过 SAM 2 与 SAM 3 的进展来探讨 SAM 1 的改进。针对视频任务，可以引入对历史帧的记忆，让模型能够记住过去的信息，实现时序的目标跟踪，将图像分割拓展为时空分割。

SAM 1 在语义上的尝试优先，对语义信息仅仅进行了浅度耦合，未来可以尝试各种方式加强视觉信息与语义信息的耦合，提升模型语义能力。

在效率优化方面，可以尝试更有效的 Transformer 来代替标准的 ViT。