

分割万物

亚历山大·基里洛夫^{1,2,4} 埃里克·明顿² 尼基拉·拉维^{1,2} 韩子毛²

谢特·肖³ 斯宾塞·怀特黑德 亚历山大·C·伯格 万延·罗

¹项目负责人 ²共同第一作者 ³同等贡献

克洛伊·罗兰³ 劳拉·古斯塔夫森³

彼得·多拉尔·罗斯·吉斯克

4方向引线

Meta AI 研究院, FAIR

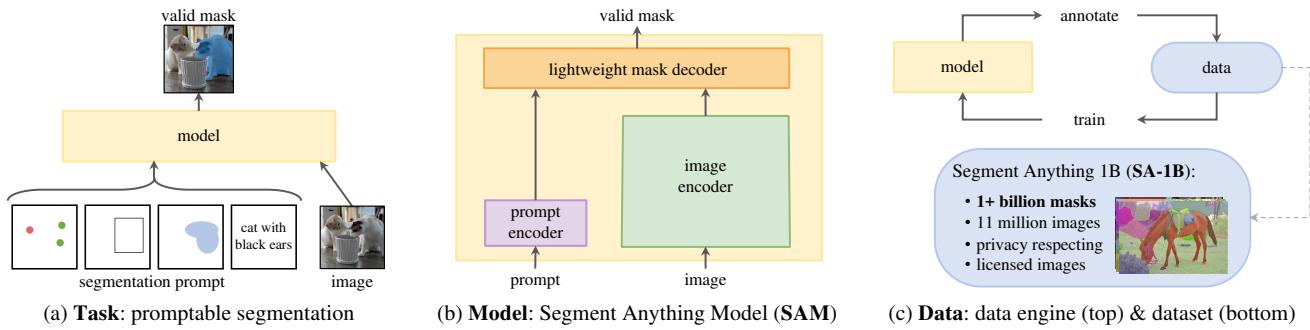


图1：我们旨在通过引入三个相互关联的组件构建一个用于分割的基础模型：一个可提示分割模型task，一个支持数据标注并通过提示工程实现一系列任务零样本迁移的分割模型model (SAM)，以及一个用于收集SA-1B（我们包含超过10亿掩码的数据集）的数据引擎。

摘要

We introduce the Segment Anything (SA) project: a new task, model, and dataset for image segmentation. Using our efficient model in a data collection loop, we built the largest segmentation dataset to date (by far), with over 1 billion masks on 11M licensed and privacy respecting images. The model is designed and trained to be promptable, so it can transfer zero-shot to new image distributions and tasks. We evaluate its capabilities on numerous tasks and find that its zero-shot performance is impressive – often competitive with or even superior to prior fully supervised results. We are releasing the Segment Anything Model (SAM) and corresponding dataset (SA-1B) of 1B masks and 11M images at <https://segment-anything.com> to foster research into foundation models for computer vision.

1. 引言

基于网络规模数据集预训练的大型语言模型，凭借强大的零样本和少样本泛化能力正在革新自然语言处理领域[10]。这些“基础模型”[8]能够泛化至训练时未见过的任务和数据分布。该能力通常通过{v*}实现——即使用人工编写的文本来提示语言模型，使其针对当前任务生成有效的文本响应。当利用海量网络文本语料进行规模化训练时，这些模型的零样本和少样本性能甚至可以媲美（乃至超越）

在某些情况下匹配）经过微调的模型[10, 21]。实证趋势表明，这种行为随着模型规模、数据集大小和总训练计算量的增加而改善[56, 10, 21, 51]。

基础模型在计算机视觉领域也有所探索，尽管程度相对较低。或许最显著的例证是利用网络中的配对文本和图像进行对齐。例如，CLIP [82] 和 ALIGN [55] 采用对比学习训练文本与图像编码器，使两种模态相互通对齐。训练完成后，通过精心设计的文本提示可实现对新视觉概念和数据分布的零样本泛化。此类编码器还能与其他模块高效结合，以支持下游任务，例如图像生成（e.g., 如 DALL-E [83]）。尽管视觉与语言编码器已取得长足进展，但计算机视觉涵盖的问题远不止于此，且其中许多领域缺乏充足的训练数据。

在这项工作中，我们的目标是构建 a foundation model for image segmentation。也就是说，我们致力于开发一个可提示的模型，并通过一项能够实现强大泛化能力的任务，在广泛的数据集上进行预训练。借助该模型，我们旨在通过提示工程解决新数据分布上的一系列下游分割问题。

该计划的成功取决于三个组成部分：任务、模型和数据。为了开发它们，我们针对图像分割提出以下问题：

1. 什么任务能够实现零样本泛化？
2. 对应的模型架构是什么？
3. 哪些数据能够支撑该任务与模型？

这些问题相互交织，需要一个综合性的解决方案。我们首先定义一个 *promptable segmentation* 任务，该任务具有足够的通用性，既能提供强大的预训练目标，又能支持广泛的下游应用。这项任务要求模型支持灵活的提示输入，并能在收到提示时实时输出分割掩码，以实现交互式使用。为了训练我们的模型，我们需要一个多样化、大规模的数据来源。遗憾的是，目前没有网络规模的分割数据源；为了解决这个问题，我们构建了一个“数据引擎”，*i.e.*，即在利用高效模型辅助数据收集与使用新收集的数据改进模型之间进行迭代。接下来我们将逐一介绍这些相互关联的组成部分，随后展示我们创建的数据集以及证明方法有效性的实验。

任务（§2）。在自然语言处理以及最近的计算机视觉领域，基础模型是一个有前景的发展，通常通过使用“提示”技术，能够对新数据集和任务进行零样本和少样本学习。受这一系列工作的启发，我们提出了 *promptable segmentation task*，其目标是在给定任何分割 *prompt*（见图1a）的情况下返回一个 *valid* 分割掩码。提示简单地指定了图像中要分割的内容，*e.g.* 例如，提示可以包含识别物体的空间或文本信息。有效输出掩码的要求意味着，即使提示模糊且可能指向多个对象（例如，衬衫上的一个点可能表示衬衫或穿着它的人），输出也应为至少其中一个对象提供合理的掩码。我们使用可提示分割任务作为预训练目标，并通过提示工程来解决通用的下游分割任务。

模型（§3）。可提示的分割任务和实际应用目标对模型架构施加了约束。具体而言，模型必须支持 *flexible prompts*，需要以摊销的 *real-time* 计算掩码以实现交互式使用，并且必须是 *ambiguity-aware*。令人惊讶的是，我们发现一个简单的设计满足了所有三个约束：一个强大的图像编码器计算图像嵌入，一个提示编码器嵌入提示，然后这两个信息源在一个轻量级的掩码解码器中结合，预测分割掩码。我们将此模型称为 Segment Anything Model，或 SAM（见图 1b）。通过将 SAM 分离为图像编码器和快速的提示编码器/掩码解码器，相同的图像嵌入可以（并摊销其成本）与不同的提示一起重复使用。给定图像嵌入，提示编码器和掩码解码器在 Web 浏览器中 ~50ms 内从提示预测掩码。我们专注于点、框和掩码提示，并展示了自由形式文本提示的初步结果。为了使 SAM 能够感知歧义，我们设计它能够为单个提示预测多个掩码，使 SAM 能够自然地处理歧义，例如衬衫 *vs.* 人物示例。

数据引擎（§4）。为实现对新数据分布的强大泛化能力，我们发现有必要在大量多样化的掩码数据集上训练 SAM，这超出了现有任何分割数据集的规模。虽然基础模型的典型方法是在线获取数据[82]，但掩码并非自然丰富，因此我们需要替代策略。我们的解决方案是构建一个“数据引擎”，即通过模型在环的数据标注与模型协同开发（见图1c）。该数据引擎包含三个阶段：辅助标注阶段、半自动标注阶段和全自动标注阶段。在第一阶段，SAM辅助标注员标注掩码，类似于经典的交互式分割设置。在第二阶段，SAM能通过可能的目标位置提示自动生成部分对象的掩码，标注员则专注于标注剩余对象，从而提升掩码多样性。在最终阶段，我们使用前景点规则网格提示 SAM，每张图像平均生成~100个高质量掩码。

数据集（§5）。我们的最终数据集 SA-1B 包含了来自 11M 张经授权且保护隐私的图像中的 1B 多个掩码（见图2）。SA-1B 完全通过我们数据引擎的最终阶段自动收集，其掩码数量比任何现有分割数据集[66, 44, 117, 60]多出 400× 倍，并且我们经过广泛验证，这些掩码具有高质量和多样性。除了用于训练 SAM 以实现鲁棒性和泛化性之外，我们希望 SA-1B 能成为旨在构建新基础模型的研究的宝贵资源。

负责任的人工智能（§6）。我们研究并报告了使用 SA-1B 和 SAM 时可能存在的公平性问题与偏见。SA-1B 中的图像覆盖了地理和经济上多样化的国家集合，我们发现 SAM 在不同人群中的表现相似。我们希望这些工作能使我们的研究在现实应用场景中更加公平。我们在附录中提供了模型和数据集卡片。

实验（§7）。我们对 SAM 进行了广泛评估。首先，在一个包含 23 个分割数据集的全新多样化测试套件中，我们发现 SAM 仅通过单个前景点就能生成高质量掩码，其效果通常仅略低于人工标注的真实值。其次，通过提示工程在零样本迁移协议下，我们在多种下游任务中持续观察到强劲的定量与定性结果，包括边缘检测、目标候选框生成、实例分割以及文本到掩码预测的初步探索。这些结果表明，SAM 可通过提示工程直接应用于解决涉及目标与图像分布的多种任务，甚至超越其训练数据范围。然而，如 §8 所讨论，其性能仍有提升空间。

发布。我们为研究目的发布了 SA-1B 数据集，并在 <https://segment-anything.com> 上以宽松的开源许可（Apache 2.0）提供 SAM 模型。我们还通过在线演示展示了 SAM 的功能。

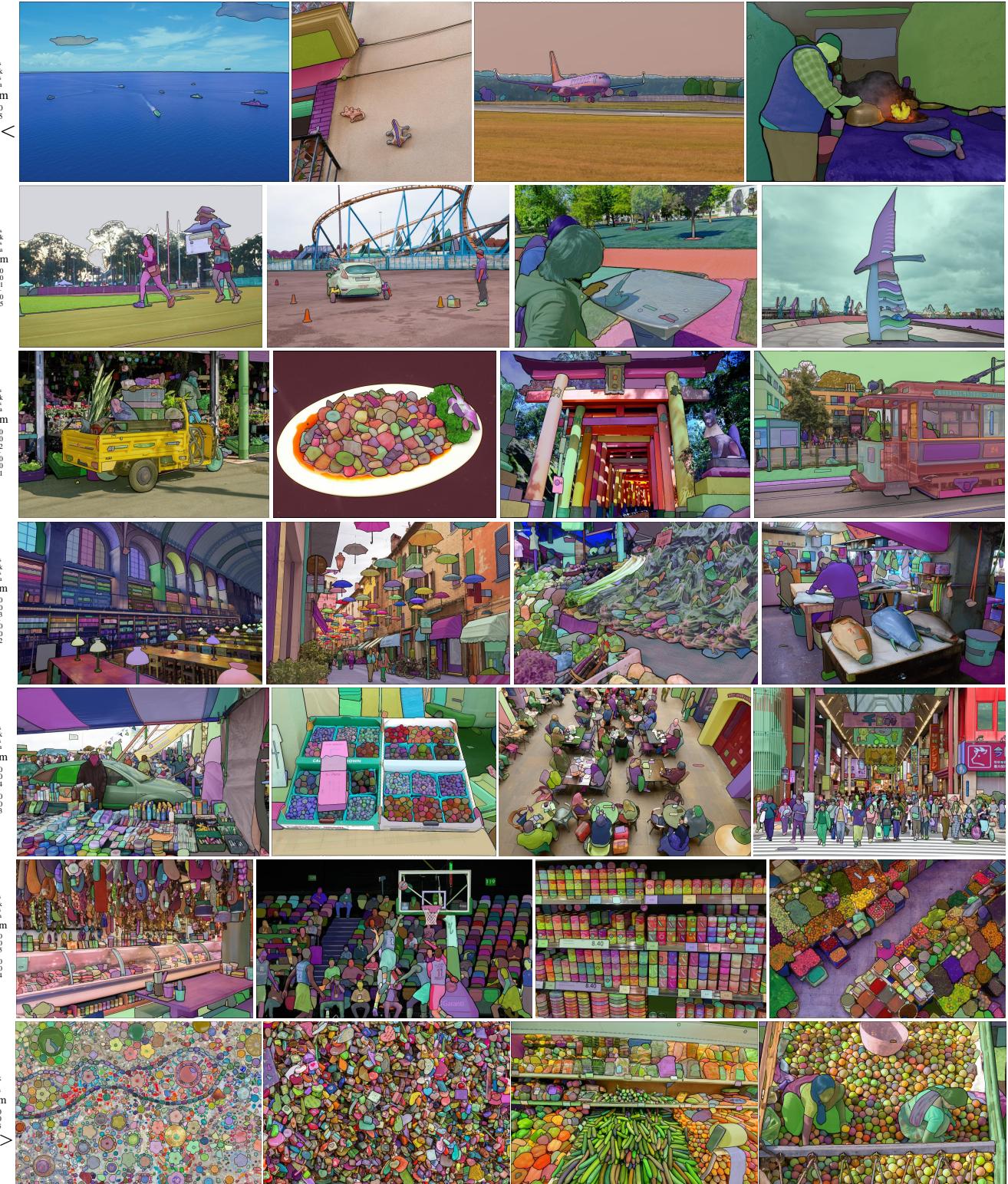


图2：来自我们新引入的数据集SA-1B的带有叠加掩码的示例图像。SA-1B包含1100万张多样化、高分辨率、已获授权且保护隐私的图像，以及11亿个高质量分割掩码。这些掩码由SAM通过*fully automatically*标注，并且正如我们通过人工评估和大量实验所验证的，它们具有高质量和多样性。我们按每张图像的掩码数量对图像进行分组以便可视化（平均每张图像有~100个掩码）。

2. 分割任意物体任务

我们从自然语言处理领域获得启发，该领域通过下一词元预测任务进行基础模型预训练*and*，并借助提示工程解决多样化的下游任务[10]。为了构建适用于分割任务的基础模型，我们致力于定义一种具备类似能力的任务。

任务。我们首先将提示的概念从自然语言处理领域迁移到分割领域，其中提示可以是一组前景/背景点、一个粗略的边界框或掩码、自由形式的文本，或者广义上任何指示图像中需要分割内容的信息。那么，*promptable segmentation task* 就是在给定任意 *prompt* 的情况下返回一个 *valid* 分割掩码。对“有效”掩码的要求仅仅意味着，即使提示是 *ambiguous* 且可能指向多个对象 (*e.g.*, 回想一下衬衫 *vs* 人物的例子, 参见图3)，输出也应当至少为这些对象中的 *one* 提供一个合理的掩码。这一要求类似于期望语言模型对模糊提示输出连贯的响应。我们选择此任务是因为它引出了一个自然的预训练算法 *and*，这是一种通过提示实现下游分割任务零样本迁移的通用方法。

预训练。可提示的分割任务提出了一种自然的预训练算法，该算法为每个训练样本模拟一系列提示 (*e.g.*, 点、框、掩码)，并将模型的掩码预测与真实标注进行比较。我们借鉴了交互式分割的方法 [109, 70]，但与之不同的是，交互式分割的目标是在获得足够用户输入后最终预测出有效的掩码，而我们的目标是为 *any prompt* 始终预测一个 *valid mask*，即使提示是 *ambiguous*。这确保了预训练模型在涉及模糊性的使用场景中（包括我们的数据引擎 §4 所要求的自动标注）是有效的。我们注意到，在此任务上表现出色具有挑战性，需要专门的建模和训练损失函数选择，我们将在 §3 中讨论这些内容。

零样本迁移。直观上，我们的预训练任务赋予模型在推理时对任何提示做出恰当响应的能力，因此下游任务可以通过设计合适的提示来解决。例如，若已有针对猫的边界框检测器，只需将检测器的框输出作为提示输入我们的模型，即可解决猫实例分割问题。总体而言，大量实际分割任务均可转化为提示工程。除自动数据集标注外，我们还在第7节的实验中探索了五种不同的示例任务。

相关任务。分割是一个广泛的领域：包括交互式分割[57, 109]、边缘检测[3]、超像素化[85]、目标候选区域生成[2]、前景分割[94]、语义分割[90]、实例分割[66]、全景分割[59]、etc。我们的可提示分割任务旨在生成



图3：每列展示了SAM从单个模糊点提示（绿色圆圈）生成的3个有效掩码。

一个能力广泛的模型，能够通过提示工程适应 *many* (虽非全部)现有及 *new* 分割任务。这种能力是任务泛化的一种形式[26]。需注意，这与以往多任务分割系统的研究不同。在多任务系统中，单一模型执行 *fixed* 一组任务，*e.g.* 例如联合语义、实例和全景分割[114, 19, 54]，但训练任务与测试任务相同。我们工作中的一项重要区别在于：为可提示分割训练的模型能在推理时作为 *component* 嵌入更大系统中执行全新不同的任务，*e.g.* 例如进行实例分割时，可提示分割模型会 *combined* 与现有目标检测器结合使用。

讨论。提示与组合是强大的工具，使得单个模型能够以可扩展的方式被使用，甚至可能完成模型设计时未知的任务。这种方法类似于其他基础模型的使用方式，*e.g.*，例如CLIP[82]作为DALL-E[83]图像生成系统中的文本-图像对齐组件。我们预计，通过提示工程等技术驱动的可组合系统设计，将比专门针对固定任务集训练的系统支持更广泛的应用。从组合的角度比较可提示分割与交互式分割也很有趣：交互式分割模型在设计时考虑了人类用户，而训练用于可提示分割的模型也可以被组合到更大的算法系统中，正如我们将展示的那样。

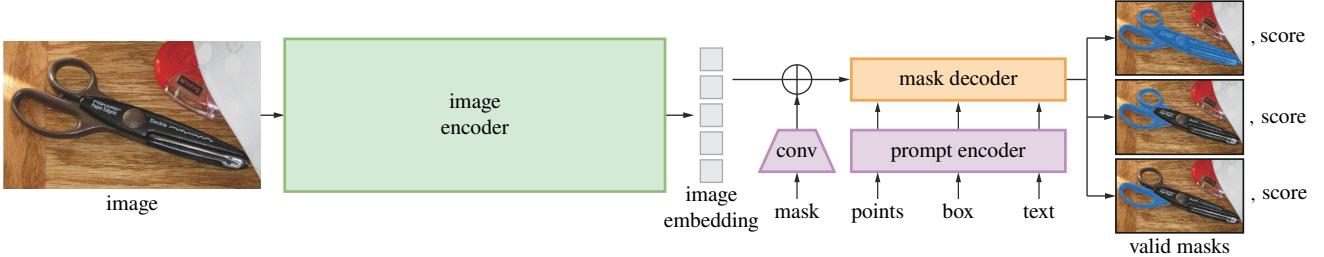


图4：Segment Anything Model (SAM) 概览。一个重型图像编码器输出图像嵌入，随后可通过多种输入提示进行高效查询，以在均摊实时速度下生成物体掩码。对于对应多个物体的模糊提示，SAM能够输出多个有效掩码及相应的置信度分数。

3. 分割一切模型

接下来，我们将介绍用于提示式分割的Segment Anything模型 (SAM)。SAM包含三个组件，如图4所示：一个图像编码器、一个灵活的提示编码器和一个快速的掩码解码器。我们基于Transformer视觉模型[14, 33, 20, 62]构建，并在（摊销）实时性能方面进行了特定权衡。此处我们概要描述这些组件，详细内容见§A节。

图像编码器。出于可扩展性和强大的预训练方法的考虑，我们采用了一个经过MAE [47]预训练的视觉Transformer (ViT) [33]，并对其进行了最小程度的调整以处理高分辨率输入[62]。该图像编码器每张图像仅运行一次，且可在提示模型前应用。

提示编码器。我们考虑两组提示：*sparse* (点、框、文本) 和 *dense* (掩码)。我们通过位置编码[95]与每种提示类型的学习嵌入相加来表示点和框，并使用CLIP[82]中的现成文本编码器处理自由格式文本。密集提示 (*i.e.* 例如掩码) 通过卷积进行嵌入，并与图像嵌入逐元素相加。

掩码解码器。掩码解码器高效地将图像嵌入、提示嵌入和输出令牌映射到掩码。该设计受[14, 20]启发，采用改进的Transformer解码器块[103]和动态掩码预测头。我们改进的解码器块使用提示自注意力和双向交叉注意力（提示到图像嵌入及反向）来更新*all*嵌入。运行两个块后，我们对图像嵌入进行上采样，并通过一个MLP将输出令牌映射到动态线性分类器，随后计算每个图像位置的前景掩码概率。

解决歧义问题。当面对一个模糊提示时，如果只有一个输出，模型会对多个有效掩码进行平均处理。为了解决这个问题，我们修改了模型，使其能够针对单个提示预测多个输出掩码（见图3）。我们发现，输出3个掩码足以应对大多数常见情况（嵌套掩码通常最多为三层：整体、部分和子部分）。在训练过程中，我们仅对最小损失进行反向传播。

损失 [15, 45, 64] 针对掩码。为了对掩码进行排序，模型为每个掩码预测一个置信度分数 ($\{v^*\}$ ，即估计的IoU)。整体模型设计主要受效率驱动。在给定预算的图像嵌入后，提示编码器和掩码解码器可在网络浏览器的CPU上以~50毫秒的速度运行。这种运行时性能使我们的模型能够实现无缝、实时的交互式提示。

损失与训练。我们使用[14]中采用的焦点损失[65]和骰子损失[73]的线性组合来监督掩码预测。我们通过混合几何提示（文本提示见§7.5）来训练可提示分割任务。遵循[92, 37]的方法，我们通过每轮掩码随机采样11轮提示来模拟交互式设置，使SAM能够无缝集成到我们的数据引擎中。

4. 分割一切数据引擎

由于互联网上分割掩码并不丰富，我们构建了一个数据引擎来收集我们的11亿掩码数据集SA-1B。该数据引擎分为三个阶段：(1) 模型辅助人工标注阶段，(2) 自动预测掩码与模型辅助标注相结合的半自动阶段，以及(3) 完全自动阶段——在此阶段中，我们的模型无需标注者输入即可生成掩码。接下来我们将逐一详述每个阶段。

辅助手动阶段。在第一阶段，类似于经典的交互式分割，一组专业标注员使用基于浏览器的交互式分割工具（由SAM驱动）通过点击前景/背景物体点来标注掩码。掩码可以使用像素级精确的“画笔”和“橡皮擦”工具进行细化。我们的模型辅助标注直接在浏览器中实时运行（使用预算的图像嵌入），实现了真正的交互式体验。我们对标注物体没有施加语义约束，标注员可以自由标注“背景物质”和“前景物体”[1]。我们建议标注员标注他们能够命名或描述的物体，但并未收集这些名称或描述。标注员被要求按显著程度顺序标注物体，并鼓励他们在单个掩码标注超过30秒后继续处理下一张图像。

在此阶段开始时，SAM使用常见的公共分割数据集进行训练。经过充分的数据标注后，仅使用新标注的掩码对SAM进行重新训练。随着收集到的掩码数量增加，图像编码器从ViT-B扩展到ViT-H，其他架构细节也逐步优化；我们总共对模型进行了6次重新训练。随着模型性能提升，每个掩码的平均标注时间从34秒减少到14秒。我们注意到，14秒的标注速度比COCO数据集的掩码标注快 $6.5\times$ ，仅比使用极值点的边界框标注慢 $2\times$ 。随着SAM的改进，每张图像的平均掩码数量从20个增加到44个。总体而言，在此阶段我们从12万张图像中收集了430万个掩码。

半自动化阶段。在此阶段，我们旨在提升掩码的

diversity

，以增强模型分割任意对象的能力。为使标注者聚焦于较不显著的对象，我们首先自动检测出高置信度的掩码，随后向标注者展示已预填充这些掩码的图像，并要求他们标注任何其他未标注的对象。为检测高置信度掩码，我们使用通用“物体”类别，基于所有第一阶段掩码训练了一个边界框检测器[84]。在此阶段，我们从18万张图像中额外收集了590万个掩码（掩码总数达到1020万个）。与第一阶段类似，我们定期使用新收集的数据重新训练模型（共5次）。由于这些对象标注难度更高，每个掩码的平均标注时间回升至34秒（不含自动生成的掩码）。每张图像的平均掩码数量从44个增加到72个（包含自动生成的掩码）。

全自动阶段。在最后阶段，标注工作实现了

fully automatic

。这得益于我们模型的两项重大改进：首先，在此阶段开始时，我们已收集到足够的掩码来大幅提升模型性能，其中包括前一阶段获取的多样化掩码；其次，至此我们已开发出能够感知歧义的模型，即使在模糊情况下也能预测出有效的掩码。具体而言，我们以 32×32 的规则点阵提示模型，并为每个点预测一组可能对应有效物体的掩码。通过这种歧义感知模型，若某点位于部件或子部件上，模型将返回子部件、部件及完整物体的掩码。我们模型的IoU预测模块用于筛选

confident

个掩码；此外，我们仅识别并选取

stable

个稳定掩码（当概率图在 $0.5 - \delta$ 和 $0.5 + \delta$ 阈值下生成相似掩码时，我们视该掩码为稳定）。最终，在选取高置信度且稳定的掩码后，我们应用非极大值抑制（NMS）过滤重复结果。为提升较小掩码的质量，我们还对多个重叠的放大图像区域进行了处理。此阶段的更多细节详见§B。我们将全自动掩码生成技术应用于数据集中所有1100万张图像，共产生11亿个高质量掩码。接下来我们将描述并分析由此构建的SA-1B数据集。

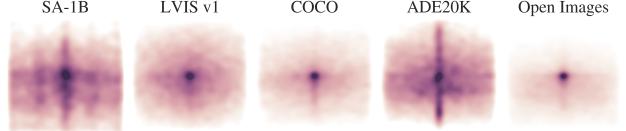


图5：图像尺寸归一化的掩码中心分布。

5. 分割任意数据集

我们的数据集SA-1B包含1100万张多样化、高分辨率、经过授权且保护隐私的图像，以及通过我们的数据引擎收集的11亿个高质量分割掩码。我们将SA-1B与现有数据集进行比较，并分析掩码的质量与特性。我们公开发布SA-1B，旨在助力未来计算机视觉基础模型的开发。需要说明的是，SA-1B将在有利于特定研究用途的许可协议下发布，并为研究人员提供相应保护。

图像。我们从一个直接与摄影师合作的供应商处获得了一套新的1100万张图片的授权。这些图片分辨率很高（平均为 3300×4950 像素），由此产生的数据量可能会带来访问和存储方面的挑战。因此，我们发布的是经过下采样的图像，其最短边设置为1500像素。即使在下采样之后，我们的图像分辨率仍显著高于许多现有的视觉数据集（e.g.例如，COCO [66] 的图像为 $\sim 480\times 640$ 像素）。请注意，目前大多数模型处理的输入分辨率要低得多。已发布图像中的人脸和车辆牌照已做模糊处理。

掩码。我们的数据引擎生成了11亿个掩码，其中99.1%是完全自动生成的。因此，自动掩码的质量至关重要。我们将其直接与专业标注进行比较，并研究各种掩码属性与主流分割数据集的对比情况。我们的主要结论是，如下文分析和第7节实验所证实，我们的自动掩码质量高，能有效用于模型训练。基于这些发现，SA-1B *only includes automatically generated masks.*

掩码质量。为评估掩码质量，我们随机抽取了500张图像（对应~5万个掩码），并请专业标注员使用我们的模型及像素级精度的“画笔”与“橡皮擦”编辑工具对这些图像中的所有掩码进行质量优化。该流程生成了自动预测掩码与专业修正掩码的配对数据。通过计算每对掩码的交并比（IoU），我们发现94%的配对掩码IoU超过90%（97%的配对掩码IoU超过75%）。作为对比，先前研究估计人工标注者间一致性仅为85-91% IoU [44, 60]。我们在§7节中的实验通过人工评分证实：相较于多种数据集，我们的掩码质量更高，且使用自动掩码训练模型的效果几乎与采用数据引擎生成的全部掩码相当。

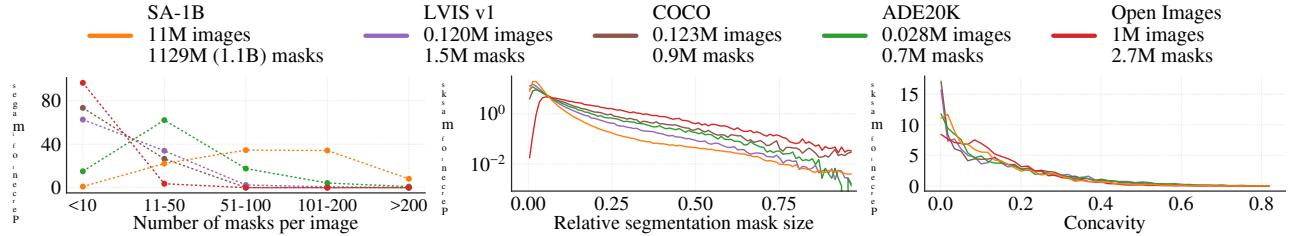


图6：数据集掩码属性。图例标注了每个数据集的图像和掩码数量。请注意，SA-1B 比现有最大的分割数据集 Open Images [60] 多出 $11 \times$ 张图像和 $400 \times$ 个掩码。

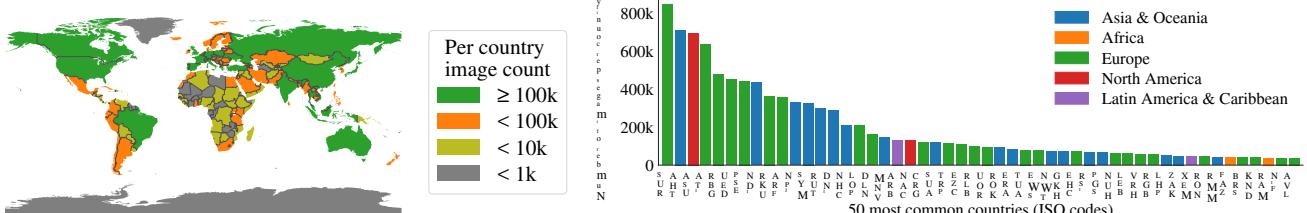


图7：SA-1B图像估计地理分布。世界上大多数国家在SA-1B中拥有超过1000张图像，而图像数量最多的三个国家来自世界不同地区。

掩码属性。在图5中，我们绘制了SA-1B中物体中心的空间分布，并与现有最大的分割数据集进行了比较。所有数据集中都存在常见的摄影师偏差。我们观察到，与分布最相似的两个数据集LVIS v1 [44]和ADE20K [117]相比，SA-1B对图像角落的覆盖更广，而COCO [66]和Open Images V5 [60]则表现出更明显的中心偏差。在图6（图例）中，我们按规模比较了这些数据集。SA-1B的图像数量比第二大数据集Open Images多 $11\times$ ，掩码数量多 $400\times$ 。平均而言，其每张图像的掩码数量比Open Images多 $36\times$ 。在这方面最接近的数据集ADE20K，每张图像的掩码数量仍比SA-1B少 $3.5\times$ 。图6（左）绘制了每张图像掩码数量的分布。接下来，我们在图6（中）分析了图像相对掩码大小（掩码面积的平方根除以图像面积）。正如预期，由于我们的数据集每张图像包含更多掩码，其中小和中等相对尺寸掩码的占比也倾向于更高。最后，为分析形状复杂性，我们在图6（右）中考察了掩码的凹度（1减去掩码面积除以其凸包面积）。由于形状复杂性与掩码尺寸相关，我们首先通过对分组的掩码尺寸进行分层抽样，以控制数据集的掩码尺寸分布。我们观察到，SA-1B掩码的凹度分布与其他数据集总体相似。

6. 分割一切模型RAI分析

接下来，我们通过研究使用SA-1B和SAM时潜在的公平性问题与偏见，对我们的工作进行了负责任人工智能（RAI）分析。我们重点关注SA-1B的地理与收入分布，以及SAM在受保护人物属性上的公平性。我们还在\$F中提供了数据集、数据标注和模型卡片。

# countries	SA-1B		% images			
	#imgs	#masks	SA-1B	COCO	O.I.	
Africa	54	300k	28M	2.8%	3.0%	1.7%
Asia & Oceania	70	3.9M	423M	36.2%	11.4%	14.3%
Europe	47	5.4M	540M	49.8%	34.2%	36.2%
Latin America & Carib.	42	380k	36M	3.5%	3.1%	5.0%
North America	4	830k	80M	7.7%	48.3%	42.8%
high income countries	81	5.8M	598M	54.0%	89.1%	87.5%
middle income countries	108	4.9M	499M	45.0%	10.5%	12.0%
low income countries	28	100k	9.4M	0.9%	0.4%	0.5%

表1：地理与收入代表性对比。SA-1B数据集在欧洲、亚洲及大洋洲地区以及中等收入国家中具有更高的代表性。来自非洲、拉丁美洲和加勒比地区以及低收入国家的图像在所有数据集中均呈现不足。

地理与收入代表性。我们采用标准方法推断图像拍摄所在国家（见§C）。图7展示了SA-1B中各国家图像数量分布（左图）及图像数量最多的50个国家（右图）。值得注意的是，排名前三的国家来自世界不同地区。随后在表1中，我们对比了SA-1B、COCO[66]和Open Images[60]在地理分布与收入层次上的代表性。SA-1B在欧洲、亚洲及大洋洲地区，以及中等收入国家的图像占比显著更高。所有数据集对非洲及低收入国家的呈现均不足。需要指出的是，在SA-1B中，包括非洲在内的所有地区都拥有至少2800万个掩码，较以往任何数据集的掩码数量高出 $10\times$ 倍。最后我们观察到，各区域和收入层次中每张图像的平均掩码数量（未展示）保持相对稳定（每图94–108个掩码）。

	mIoU at		mIoU at	
	1 point	3 points	1 point	3 points
<i>perceived gender presentation</i>				
feminine	54.4 ± 1.7	90.4 ± 0.6	1	52.9 ± 2.2
masculine	55.7 ± 1.7	90.1 ± 0.6	2	51.5 ± 1.4
<i>perceived age group</i>				
older	62.9 ± 6.7	92.6 ± 1.3	3	52.2 ± 1.9
middle	54.5 ± 1.3	90.2 ± 0.5	4	51.5 ± 2.7
young	54.2 ± 2.2	91.2 ± 0.7	5	52.4 ± 4.2
			6	56.7 ± 6.3
				91.2 ± 2.4

表2：SAM在不同感知性别表现、年龄组和肤色人群中的分割性能。图中显示了95%置信区间。在每个分组内，除较年长vs组外，所有置信区间均相互重叠。

在人群分割中的公平性。我们通过测量SAM在不同群体间的性能差异，调查了在感知性别表现、感知年龄组和感知肤色方面潜在的公平性问题。我们使用More Inclusive Annotations for People (MIAP) [87]数据集进行性别表现和年龄的分析，并使用专有数据集进行肤色分析（见§C）。我们的评估采用模拟交互式分割，随机采样1个和3个点（见§D）。表2（左上）展示了感知性别表现的结果。我们注意到，女性在检测和分割数据集中代表性不足[115]，但观察到SAM在不同群体间的表现相似。我们在表2（左下）中对感知年龄重复了分析，指出那些感知为更年轻和更年长的人在大规模数据集中代表性不足[110]。SAM在感知为更年长的人上表现最佳（尽管置信区间较大）。最后，我们在表2（右侧）对感知肤色重复了分析，指出在大规模数据集中，较浅肤色的人被过度代表，而较深肤色的人代表性不足[110]。由于MIAP不包含感知肤色的标注，我们使用了一个专有数据集，其中包含感知的Fitzpatrick皮肤类型[36]标注，范围从1（最浅肤色）到6（最深肤色）。虽然均值略有变化，但我们未发现群体间存在显著差异。我们认为这些发现源于任务本身的性质，并承认当SAM作为更大系统的组件使用时，可能会出现偏差。最后，在§C中，我们将分析扩展到服装分割，发现存在跨感知性别表现的偏差迹象。

7. 零样本迁移实验

在本节中，我们展示了使用SAM (Segment Anything Model) 进行的zero-shot transfer实验。我们考虑了五项任务，其中四项与训练SAM时使用的可提示分割任务有显著差异。这些实验在训练过程中未见过的数据集和任务上评估了SAM。

在训练过程中（我们对“零样本迁移”的使用遵循了CLIP[82]中的定义）。这些数据集可能包含新颖的图像分布，例如水下或第一人称视角图像 (e.g. 图8)，据我们所知，这些图像并未出现在SA-1B中。

我们的实验首先测试了可提示分割的核心目标：从任何提示中生成有效的掩码。我们重点考察了具有挑战性的前景点提示single场景，因为与其他更具体的提示相比，这种提示更容易产生歧义。接下来，我们进行了一系列实验，涵盖低层、中层和高层图像理解任务，大致与该领域的历史发展脉络平行。具体而言，我们通过提示SAM实现以下功能：(1) 边缘检测，(2) 分割所有目标 (*i.e.* 对象提议生成)，(3) 分割检测到的对象 (*i.e.* 实例分割)，以及(4) 作为概念验证，通过自由文本提示分割目标。这四项任务与SAM训练所用的可提示分割任务存在显著差异，均通过提示工程实现。最后我们通过消融研究完成实验验证。

实现。除非另有说明：(1) SAM使用MAE [47]预训练的ViT-H [33]图像编码器；(2) SAM在SA-1B数据集上训练，需注意该数据集仅包含我们数据引擎最终阶段自动生成的掩码。所有其他模型与训练细节（如超参数）请参阅§A。

7.1. 零样本单点有效掩码评估

任务。我们评估从single前景点分割物体的能力。由于单个点可能对应多个物体，这一任务本身是不适当的。大多数数据集中的真实掩码并未枚举all所有可能的掩码，这可能导致自动评估指标不可靠。因此，我们在标准mIoU指标 (*i.e.*，即预测掩码与真实掩码间所有IoU的平均值) 之外，补充进行了人工评估：标注者根据掩码质量从1（无意义）到10（像素级完美）进行评分。更多细节参见§D.1、§E及§G章节。

默认情况下，我们按照交互式分割的标准评估协议[92]，从真实掩码的“中心”（掩码内部距离变换的最大值处）采样点。由于SAM能够预测多个掩码，默认情况下我们仅评估模型置信度最高的掩码。基线方法均为单掩码预测模型。我们主要与RITM[92]进行对比——这是一个强大的交互式分割器，在我们的基准测试中相较于其他强基线方法[67, 18]表现最佳。

数据集。我们采用了一个新整理的包含23个数据集、涵盖多样化图像分布的测试集。图8列出了这些数据集并展示了每个数据集的示例样本（更多细节见附录表7）。我们使用全部23个数据集进行mIoU评估。对于人工评估研究，我们使用图9b所列的子集（由于此类研究对资源要求较高）。该子集既包含SAM在自动指标评估中优于RITM的数据集，也包含其表现不及RITM的数据集。

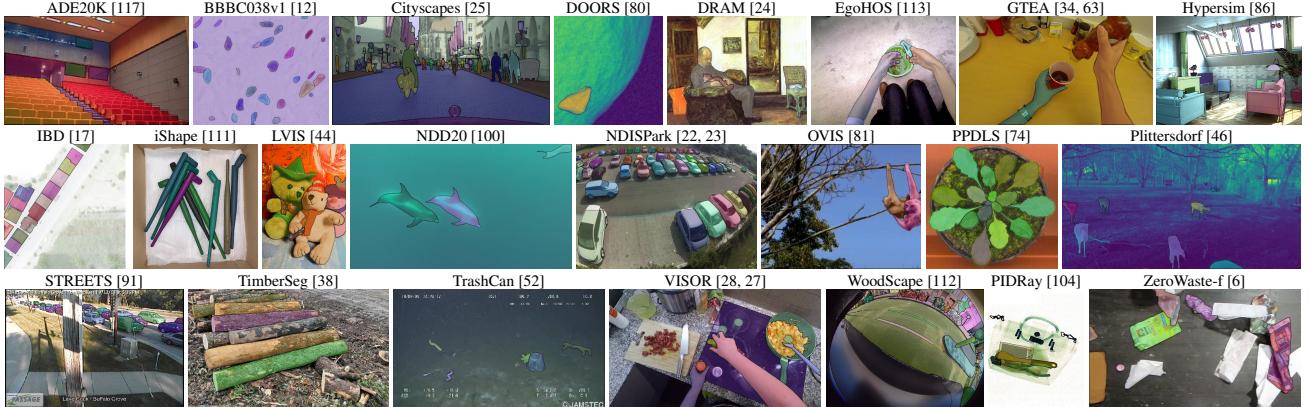
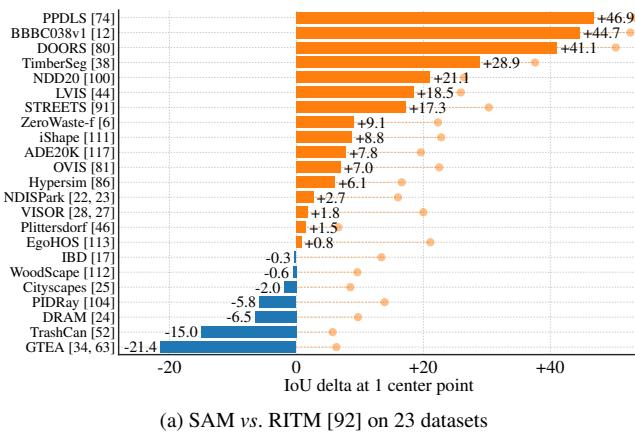
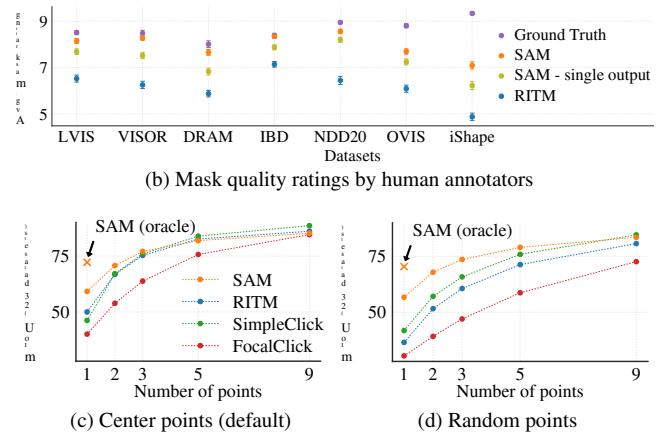


图8：用于评估SAM零样本迁移能力的23个多样化分割数据集样本。



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



(b) Mask quality ratings by human annotators

(c) Center points (default)

(d) Random points

图9：在23个数据集上的点选掩码评估。(a) SAM与最强单点分割器RITM[92]的平均交并比。由于歧义性，单个掩码可能无法匹配真实标注；圆圈显示SAM 3个预测中最相关结果的“理想”性能。(b) 各数据集中标注者对掩码质量评分（1分为最差，10分为最佳）的对比。所有方法均使用真实掩码中心点作为提示。(c, d) 不同点数下的平均交并比。SAM在单点提示时显著优于先前的交互式分割器，多点提示时性能相当。单点提示时绝对平均交并比较低是歧义性导致的结果。

结果。首先，我们使用mIoU在完整的23个数据集套件上进行自动评估。我们在图9a中与RITM进行了逐数据集的结果对比。SAM在23个数据集中的16个上取得了更高的结果，最高领先~47 IoU。我们还展示了一个“预言”结果，其中通过将SAM生成的3个掩码与真实标注进行比较，选择最相关的一个，而非选择置信度最高的掩码。这揭示了模糊性对自动评估的影响。具体而言，通过预言机制解决模糊性后，SAM在all个数据集上超越了RITM。

人类研究的结果如图9b所示。误差条表示平均掩码评分的95%置信区间（所有差异均显著；详见§E）。我们观察到标注者一致认为SAM生成的掩码质量显著高于最强基线RITM。而采用单输出掩码的消融版“无歧义感知”SAM评分持续较低，但仍高于RITM。SAM的平均评分

介于7到9之间，这对应着定性评级指南：“

*A high score (7-9): The object is identifiable and errors are small and rare (例如
.., missing a small,*

heavily obscured disconnected component, ...).”这些结果表明，SAM已学会从单点分割出有效掩码。值得注意的是，对于像DRAM和IBD这样的数据集，SAM在自动指标上表现较差，*it receives consistently higher ratings in the human study.*

图9c展示了额外的基线方法，SimpleClick [67]和FocalClick [18]，它们在单点性能上低于RITM和SAM。随着点数从1增加到9，我们观察到各方法之间的差距逐渐缩小。这是预料之中的，因为任务变得更简单；同时，SAM并未针对极高IoU区域进行优化。最后，在图9d中，我们将默认的中心点采样替换为随机点采样。我们观察到SAM与基线方法之间的差距增大，并且SAM能够在任一种采样方法下取得可比的结果。

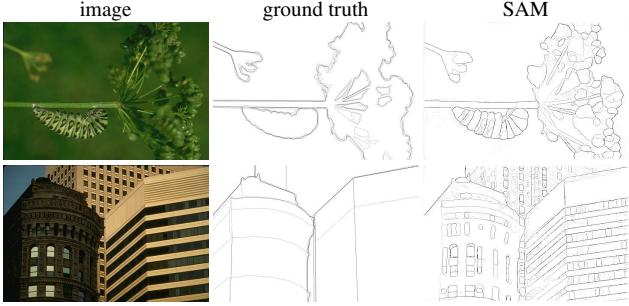


图10：BSDS500上的零样本边缘预测。SAM并未被训练用于预测边缘图，且在训练过程中未接触过BSDS图像或标注。

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

表3：在BSDS500上进行边缘检测的零样本迁移。

7.2. 零样本边缘检测

方法。我们在经典的边缘检测任务上使用BSDS500 [72, 3] 评估SAM。我们采用简化版的自动掩码生成流程。具体而言，我们使用 16×16 的规则前景点网格提示SAM，从而生成768个预测掩码（每个点3个）。通过非极大值抑制（NMS）去除冗余掩码。随后，通过对未阈值化的掩码概率图进行Sobel滤波及标准的轻量级后处理（包括边缘NMS）来计算边缘图（详见§D.2）。

结果。我们在图10中展示了代表性的边缘检测图（更多结果见图15）。从定性角度看，我们观察到尽管SAM并非针对边缘检测任务进行训练，但其生成的边缘图仍具有合理性。与真实标注相比，SAM预测出了更多边缘，其中包含BSDS500数据集中未标注的合理边缘。这种偏差在表3中得到了量化体现：在50%精度（R50）下的召回率较高，但这是以牺牲精度为代价的。相较于那些专门学习BSDS500数据偏置（ $\{v^*\}$ ，即哪些边缘需要抑制）的先进方法，SAM自然存在差距。然而，与HED[108]等开创性深度学习方法（同样基于BSDS500训练）相比，SAM表现良好，并且显著优于早期（尽管已过时）的零样本迁移方法。

7.3. 零样本目标候选框生成

方法。接下来，我们在物体提议生成这一中层任务上评估SAM[2, 102]。该任务在物体检测研究中曾发挥重要作用，常作为

method	all	mask AR@1000					
		small	med.	large	freq.	com.	rare
ViTDet-H [62]	63.0	51.7	80.8	87.0	63.1	63.3	58.3
<i>zero-shot transfer methods:</i>							
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

表4：LVIS v1上的物体提议生成。SAM以零样本方式应用，*i.e.* 它并未针对物体提议生成进行训练，也未接触过LVIS图像或标注。

在开创性系统中的中间步骤（*e.g.*, [102, 41, 84]）。为生成物体候选框，我们运行了一个略微修改的自动掩码生成流程版本，并将生成的掩码作为候选框输出（详见§D.3）。

我们在LVIS v1 [44]上计算标准平均召回率（AR）指标。我们重点关注LVIS，因为其庞大的类别数量构成了一个具有挑战性的测试。我们与一个*strong*基线进行比较，该基线实现为ViTDet [62]检测器（采用级联Mask R-CNN [48, 11] ViT-H）。需要指出的是，这个“基线”对应的是“伪装成提案生成器的检测器”（DMP）方法[16]，该方法已被证明能够人为提升AR指标，因此这是一个极具挑战性的比较基准。

结果。在表4中，我们毫不意外地发现，使用ViTDet-H的检测结果作为物体提议（*i.e.*，即操纵AR指标的DMP方法[16]）整体表现最佳。然而，SAM在多项指标上的表现相当出色。值得注意的是，它在中型和大型物体以及稀有和常见物体上的表现均优于ViTDet-H。实际上，SAM仅在小型物体和频繁出现物体上表现不及ViTDet-H——由于ViTDet-H在LVIS数据集上训练，能轻易学习到LVIS特有的标注偏差，而SAM则不具备这一条件。我们还对比了SAM的消融版本（“单一输出”），该版本忽略了歧义性处理，在所有AR指标上均显著逊色于完整版SAM。

7.4. 零样本实例分割

方法。转向更高层次的视觉任务，我们使用SAM作为实例分割器的分割模块。实现方式很简单：我们先运行一个目标检测器（使用之前提到的ViTDet），然后用其输出的边界框提示SAM。这展示了 $\{v^*\}$ SAM在更大系统中的应用。

结果。我们在表5中比较了SAM和ViTDet在COCO和LVIS数据集上预测的掩码。从掩码AP指标来看，我们在两个数据集上都观察到了差距，其中SAM的表现相当接近ViTDet，但确实仍落后于后者。通过可视化输出，我们观察到SAM的掩码在质量上通常优于ViTDet，具有更清晰的边界（参见§D.4和图16）。为了深入探究这一现象，我们额外进行了一项人工评估研究，要求标注者按照之前使用的1到10质量评分标准对ViTDet掩码和SAM掩码进行评级。在图11中，我们发现SAM在人工评估中 *consistently* 优于ViTDet。

method	COCO [66]				LVIS v1 [44]			
	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
<i>zero-shot transfer methods (segmentation module only):</i>								
SAM	46.5	30.8	51.0	61.7	44.7	32.5	57.6	65.5

表5：实例分割结果。SAM通过ViTDet提供的边界框提示进行零样本分割。完全监督的ViTDet表现优于SAM，但在更高质量的LVIS掩码上差距缩小。有趣的是，根据人工评分（见图11），SAM的表现超越了ViTDet。

◦

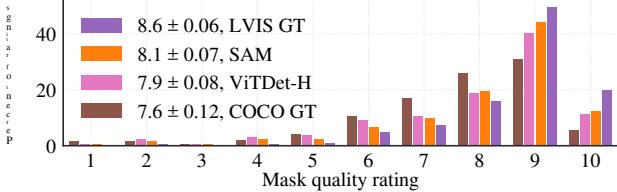


图11：针对ViTDet和SAM（均应用于LVIS真实标注框）的人工研究中掩码质量评分分布。我们还报告了LVIS和COCO真实标注的质量。图例显示了评分均值及95%置信区间。尽管SAM的AP较低（表5），但其评分高于ViTDet，这表明ViTDet利用了COCO和LVIS训练数据中的偏差。

我们假设在COCO数据集上，由于掩码AP差距较大且真实标注质量相对较低（正如人工研究所证实的那样），ViTDet模型学习了COCO掩码的特定偏差。而SAM作为一种零样本方法，无法利用这些（通常不受欢迎）偏差。LVIS数据集拥有更高质量的真实标注，但仍存在特定的特性（e.g., 掩码不包含孔洞，按构造方式仅为简单多边形）以及针对模态vs非模态掩码的偏差。同样，SAM未经训练以学习这些偏差，而ViTDet则能够利用它们。

7.5. 零样本文本到掩码

方法。最后，我们考虑一个更高层次的任务：从自由文本中分割对象。这个实验是对SAM处理文本提示能力的概念验证。虽然我们在之前的所有实验中都使用了完全相同的SAM，但在这个实验中，我们修改了SAM的训练过程，使其具备文本感知能力，但这种方式不需要新的文本标注。具体来说，对于每个面积大于 100^2 的手动收集掩码，我们提取其CLIP *image*嵌入。然后，在训练过程中，我们将提取的CLIP图像嵌入作为SAM的首次交互提示。这里的关键观察是，由于CLIP的*image*嵌入被训练为与其*text*嵌入对齐，我们可以使用图像嵌入进行训练，但在推理时使用文本嵌入。也就是说，在推理时，我们将文本输入CLIP的文本编码器，然后将生成的文本嵌入作为提示输入SAM（详见§D.5）。

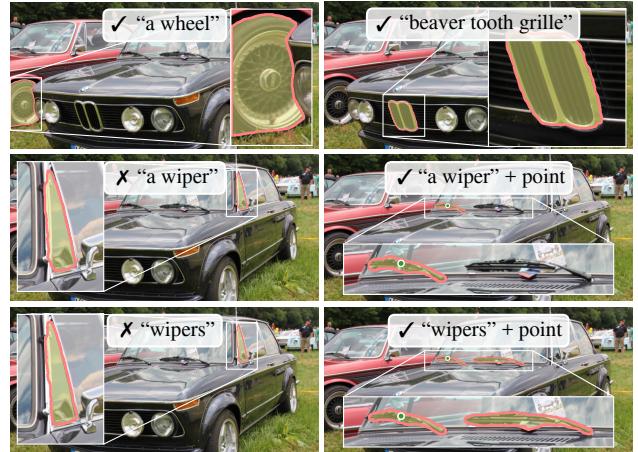


图12：零样本文本到掩码。SAM能够处理简单和细微的文本提示。当SAM未能做出正确预测时，额外的点提示可以提供帮助。

结果。我们在图12中展示了定性结果。SAM能够根据简单的文本提示（如“一个轮子”）以及短语（如“海狸牙齿格栅”）分割物体。当SAM仅凭文本提示未能选择正确对象时，通常添加一个额外点即可修正预测，这与文献[31]的方法类似。

7.6. 消融实验

我们在包含23个数据集的数据集套件上，使用单中心点提示协议进行了多项消融实验。需要指出的是，单个点可能具有歧义，而这种歧义性可能未在仅包含每个点单个掩码的真值中体现。由于SAM在零样本迁移设置下运行，其排名最高的掩码vs与数据标注准则产生的掩码之间可能存在系统性偏差。因此，我们额外报告了相对于真值的最佳掩码（即“oracle”结果）◦

图13（左）展示了SAM在数据引擎各阶段累积数据上训练时的性能表现。我们观察到每个阶段都能提升mIoU值。当使用全部三个阶段的数据进行训练时，自动生成掩码的数量远超人工和半自动掩码。为解决这一问题，我们发现训练时将人工与半自动掩码过采样 $10\times$ 倍能获得最佳效果。但这种设置会使训练过程复杂化。因此我们测试了第四种方案——仅使用自动生成的掩码。在此数据配置下，SAM的性能仅略低于使用全部数据的情况（相差~0.5 mIoU）。基于此，为简化训练配置，我们默认仅采用自动生成的掩码。

在图13（中）我们考察了数据量的影响。完整的SA-1B包含1100万张图像，我们在此消融实验中均匀地子采样至100万和10万张。当图像数量为10万时，我们观察到所有设置下的平均交并比（mIoU）均大幅下降。然而，使用100万张图像（约占完整数据集的10%）时，我们观察到与使用完整数据集相当的结果。这一数据规模仍包含约1亿个掩码，对许多实际应用场景而言可能是一个可行的设置。

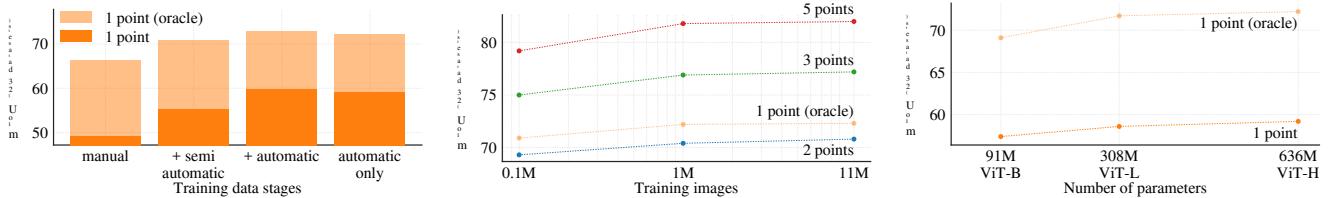


图13：我们的数据引擎阶段、图像编码器缩放和训练数据缩放的消融研究。（左）每个数据引擎阶段都在我们的2个数据集套件上带来了改进，仅使用自动数据（我们的默认设置）进行训练的结果与使用所有三个阶段的数据相似。（中）使用SA-1B的{v*}10%和完整SA-1B训练的SAM表现相当。我们默认使用全部1100万张图像进行训练，但使用100万张图像是一个合理的实际设置。（右）缩放SAM的图像编码器显示出有意义但趋于饱和的收益。尽管如此，在某些设置中，较小的图像编码器可能更受青睐。

最后，图13（右）展示了使用ViT-B、ViT-L和ViT-H图像编码器的结果。ViT-H相比ViT-B有显著提升，但相对于ViT-L仅有边际改善。目前进一步扩大图像编码器规模似乎收效甚微。

8. 讨论

基础模型。自机器学习早期以来，预训练模型已被应用于下游任务[99]。近年来，随着对规模化的日益重视，这一范式变得越来越重要，此类模型最近被（重新）定义为“基础模型”：*i.e.* 即“在大规模广泛数据上训练，并能适应广泛下游任务”的模型[8]。我们的工作与这一定义高度契合，尽管我们注意到，图像分割的基础模型本质上范围有限，因为它代表了计算机视觉中一个重要但局部的子集。我们还将自己方法的一个方面与[8]进行对比，后者强调*self-supervised*学习在基础模型中的作用。虽然我们的模型通过自监督技术（MAE [47]）初始化，但其绝大部分能力来自大规模*supervised*训练。在数据引擎能够扩展可用标注的情况下（如我们的方法），监督训练提供了一种有效的解决方案。

组合性。预训练模型能够催生新的能力，甚至超越训练时预设的范畴。一个典型例子是CLIP[82]如何作为*component*被应用于DALL-E[83]等更大系统中。我们的目标是通过SAM使这类组合变得直观便捷。为此，我们要求SAM能够针对各种分割提示预测出有效掩码，从而在SAM与其他组件间建立可靠的交互接口。例如，MCC[106]可轻松调用SAM分割目标物体，仅凭单张RGB-D图像就能实现针对未见物体的强泛化三维重建。另一案例中，SAM可通过可穿戴设备检测的注视点进行提示，从而开启全新应用场景。得益于SAM对第一视角图像等新领域的泛化能力，此类系统无需额外训练即可运行。

局限性。尽管SAM在一般情况下表现良好，但它并非完美无缺。它可能遗漏细微结构，有时会产生细小的不连贯组件幻觉，并且无法像那些需要“放大”处理的、计算密集型方法那样生成清晰的边界，*e.g.* [18]。通常，当提供多个点位时，我们预计专用的交互式分割方法会优于SAM，*e.g.* [67]。与这些方法不同，SAM的设计初衷是追求通用性和广泛适用性，而非高IoU的交互式分割。此外，SAM虽然能实时处理提示，但在使用重型图像编码器时，其整体性能仍无法达到实时标准。我们在文本到掩码任务上的尝试是探索性的，尚未完全稳健，但我们相信通过更多努力可以改进这一点。尽管SAM能执行多种任务，但目前尚不清楚如何设计简单的提示来实现语义分割和全景分割。最后，在某些特定领域存在专业工具（例如[7]），我们预计这些工具在各自领域会优于SAM。

结论。Segment Anything项目旨在将图像分割提升至基础模型时代。我们的主要贡献在于提出了新的任务（可提示分割）、模型（SAM）和数据集（SA-1B），使这一跨越成为可能。SAM能否成为基础模型仍有待社区使用实践的检验，但无论如何，我们相信这项工作的视角、超过10亿掩码的公开以及可提示分割模型{v*}的发布，都将为未来发展铺平道路。

致谢。我们要感谢Aaron Adcock和Jitendra Malik的有益讨论。感谢Vaibhav Aggarwal和Yanghao Li在模型扩展方面的帮助。感谢Cheng-Yang Fu、Jiabo Hu和Robert Kuo在数据标注平台搭建中的支持。感谢Allen Goodman和Bram Wasti在优化模型网页版过程中的协助。最后，感谢Morteza Behrooz、Ashley Gabriel、Ahuva Goldstand、Sumanth Gurram、Somya Jain、Devansh Kukreja、Joshua Lane、Lilian Luong、Mallika Malhotra、William Ngan、Omkar Parkhi、Nikhil Raina、Dirk Rowe、Neil Sejoor、Vanessa Stark、Bala Varadarajan和Zachary Winstrom在演示程序、数据集查看器以及其他资源与工具开发中提供的帮助。

参考文献

- [1] Edward H Adelson。论观察物质：人类与机器对材料的感知。*Human vision and electronic imaging VI*, 2001年。5 [2] Bogdan Alexe、Thomas Deselaers 和 Vittorio Ferrari。何为物体？*CVPR*, 2010年。4, 10 [3] Pablo Arbeláez、Michael Maire、Charles Fowlkes 和 Jitendra Malik。轮廓检测与层次化图像分割。*TPAMI*, 2010年。4, 10, 21, 28 [4] Jimmy Lei Ba、Jamie Ryan Kiros 和 Geoffrey E Hinton。层归一化。*arXiv:1607.06450*, 2016年。16 [5] 鲍航博、董力和韦福如。BEiT：图像Transformer的BERT预训练。*arXiv:2106.08254*, 2021年。17 [6] Dina Bashkirova、Mohamed Abdelfattah、Ziliang Zhu、James Akl、Fadi Alladkani、Ping Hu、Vitaly Ablavsky、Berk Cali、Sarah Adel Bargal 和 Kate Saenko。ZeroWaste数据集：面向杂乱场景中的可变形物体分割。*CVPR*, 2022年。9, 20 [7] Stuart Berg、Dominik Kutra、Thorben Kroeger、Christoph N. Straehle、Bernhard X. Kausler、Carsten Haubold、Martin Schiegg、Janez Ales、Thorsten Beier、Markus Rudy、Kemal Eren、Jaime I. Cervantes、Buote Xu、Fynn Beuttenmueller、Adrian Wolny、Chong Zhang、Ullrich Koethe、Fred A. Hamprecht 和 Anna Kreshuk。ilastik：用于（生物）图像分析的交互式机器学习。*Nature Methods*, 2019年。12 [8] Rishi Bommasani、Drew A Hudson、Ehsan Adeli、Russ Altman、Simran Arora、Sydney von Arx、Michael S Bernstein、Jeannette Bohg、Antoin e Bosselut、Emma Brunskill 等。论基础模型的机遇与风险。*arXiv:2108.07258*, 2021年。1, 12 [9] Gustav Bredell、Christine Tanner 和 Ender Konukoglu。用于分割编辑网络的迭代交互训练。*MICCAI*, 2018年。17 [10] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel Ziegler、Jeffrey Wu、Clemens Winter、Chris Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam Mc Candlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。语言模型是小样本学习者。*NeurIPS*, 2020年。1, 4 [11] 蔡兆伟 和 Nuno Vasconcelos。Cascade R-CNN：深入高质量物体检测。*CVPR*, 2018年。10 [12] Juan C. Caicedo、Allen Goodman、Kyle W. Karhohs、Beth A. Cimini、Jeanelle Ackerman、Marzieh Haghighi、CherKeng Heng、Tim Becker、Minh Doan、Claire McQuin、Mohammad Rohban、Shantanu Singh 和 Anne E. Carpenter。跨成像实验的细胞核分割：2018年数据科学碗。*Nature Methods*, 2019年。9, 19, 20 [13] John Cannon。边缘检测的计算方法。*TPAMI*, 1986年。10, 21 [14] Nicolas Carion、Francisco Massa、Gabriel Synnaeve、Nicolas Usunier、Alexander Kirillov 和 Sergey Zagoruyko。基于Transformer的端到端物体检测。*ECCV*, 2020年。5, 16, 17 [15] Guillaume Charpiat、Matthias Hofmann 和 Bernhard Schölkopf。通过多模态预测实现的自动图像着色。*ECCV*, 2008年。5, 17 [16] Neelima Chavali、Harsh Agrawal、Aroma Mahendru 和 Dhruv Batra。物体提议评估协议是“可博弈的”。*CVPR*, 2016年。10, 21 [17] 陈佳洲、徐阳辉、陆淑芳、梁荣华和南亮亮。MVS建筑物的3D实例分割。*IEEE Transactions on Geoscience and Remote Sensing*, 2022年。9, 19, 20, 23, 24 [18] 陈曦、赵志燕、张一雷、段曼妮、齐东连 和 赵恒爽。FocalClick：迈向实用的交互式图像分割。*CVPR*, 2022年。8, 9, 12, 19 [19] Bowen Cheng、Ishan Misra、Alexander G Schwing、Alexander Kirillov 和 Rohit Girdhar。用于通用图像分割的掩码注意力掩码变换器。*CVPR*, 2022年。4 [20] Bowen Cheng、Alex Schwing 和 Alexander Kirillov。逐像素分类并非语义分割的全部所需。*NeurIPS*, 2021年。5, 16, 17 [21] Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann 等。PaLM：通过Pathways扩展语言建模。*arXiv:2204.02311*, 2022年。1 [22] Luca Ciampi、Carlos Santiago、Joao Costeira、Claudio Gennaro 和 Giuseppe Amato。交通密度估计的领域自适应。
- International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2021年。9, 20 [23] Luca Ciampi、Carlos Santiago、Joao Costeira、Claudio Gennaro 和 Giuseppe Amato。昼夜实例分割停车场（NDIS-Park）数据集：用于停车场区域车辆检测、分割和计数的白天与夜间图像集。*Zenodo*, 2022年。9, 20 [24] Nadav Cohen、Yael Newman 和 Ariel Shamir。艺术绘画中的语义分割。*Computer Graphics Forum*, 2022年。9, 19, 20, 23, 24 [25] Marius Cordts、Mohamed Omran、Sebastian Ramos、Timo Rehfeld、Markus Enzweiler、Rodrigo Benenson、Uwe Franke、Stefan Roth 和 Bernt Schiele。用于语义城市场景理解的Cityscapes数据集。*CVPR*, 2016年。9, 19, 20 [26] Bruno da Silva、George Konidaris 和 Andrew Barto。学习参数化技能。*ICML*, 2012年。4 [27] Dima Damen、Hazel Doughty、Giovanni Maria Farinella、Antonino Furnari、Jian Ma、Evangelos Kazakos、Davide Moltisanti、Jonathan Munro、Toby Perrett、Will Price 和 Michael Wray。重新缩放以自我为中心的视觉：EPIC-KITCHENS-100的数据收集、流程与挑战。*IJCV*, 2022年。9, 20, 23, 24 [28] Ahmad Darkhalil、Dandan Shan、Bin Zhu、Jian Ma、Amlan Kar、Richard Higgins、Sanja Fidler、David Fouhey 和 Dima Damen。EPIC-KITCHENS VISOR基准：视频分割与物体关系。*NeurIPS*, 2022年。9, 19, 20, 23, 24 [29] Terrance De Vries、Ishan Misra、Changhan Wang 和 Laurens Van der Maaten。物体识别对每个人都有效吗？*CVPR workshops*, 2019年。18 [30] Mark D'az、Ian Kvlichan、Rachel Rosen、Dylan Baker、Razvan Amironesei、Vinodkumar Prabhakaran 和 Emily Denton。CrowdWorkSheets：考量众包数据集标注背后的个体与集体身份。*ACM Conference on Fairness, Accountability, and Transparency*, 2022年。25 [31] Henghui Ding、Scott Cohen、Brian Price 和 Xudong Jiang。PhraseClick：通过短语和点击实现灵活的交互式分割。*ECCV*, 2020年。11 [32] Piotr Dollár 和 C Lawrence Zitnick。使用结构化森林的快速边缘检测。*TPAMI*, 2014年。21 [33] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszkoreit 和 Neil Houlsby。一张图像值16x16个词：大规模图像识别的变换器。*ICLR*, 2021年。5, 8, 16 [34] Alireza Fathi、Xiaofeng Ren 和 James M. Rehg。学习识别以自我为中心活动中的物体。*CVPR*, 2011年。9, 19, 20 [35] Pedro F Felzenszwalb 和 Daniel P Huttenlocher。高效的基于图的图像分割。*IJCV*, 2004年。10 [36] Thomas B. Fitzpatrick。日光反应性皮肤类型 I 至 VI 的有效性与实用性。*Archives of Dermatology*, 1988年。8 [37] Marco Forte、Brian Price、Scott Cohen、Ning Xu 和 François Pitié。在交互式分割中达到99%的准确率。*arXiv:2003.07932*, 2020年。5, 17 [38] Jean-Michel Fortin、Olivier Gamache、Vincent Grondin、François Pomerleau 和 Philippe Giguère。林业作业中用于自主抓取原木的实例分割。*IROS*, 2022年。9, 20

- [39] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, Kate Crawford. 数据集的数据表。 *Communications of the ACM*, 2021. 25[40] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, Barret Zoph. 简单的复制粘贴是实例分割的一种强大数据增强方法。 *CVPR*, 2021. 16, 18, 22[41] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik. 用于精确目标检测和语义分割的丰富特征层次结构。 *CVPR*, 2014. 10[42] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, Kaiming He. 精确的大批量SGD：在1小时内训练ImageNet。 *arXiv:1706.02677*, 2017. 17[43] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragnani, Qichen Fu, Christian Fuegen, Abraham Gebrselassie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Leslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhuguri, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, Jitendra Malik. Ego4D：3000小时第一人称视角视频环游世界。 *CVPR*, 2022. 20[44] Agrim Gupta, Piotr Dollár, Ross Girshick. LVIS：一个用于大词汇量实例分割的数据集。 *CVPR*, 2019. 2, 6, 7, 9, 10, 11, 19, 20, 21, 24[45] Abner Guzman-Rivera, Dhruv Batra, Pushmeet Kohli. 多选学习：学习生成多个结构化输出。 *NeurIPS*, 2012. 5, 17[46] Timm Haucke, Hjalmar S. Kühl, Volker Steinhage. SOCRATES：利用立体视觉在视觉野生动物监测中引入深度。 *Sensors*, 2022. 9, 20[47] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, Ross Girshick. 掩码自编码器是可扩展的视觉学习器。 *CVPR*, 2022. 5, 8, 12, 16, 17[48] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick. Mask R-CNN。 *ICCV*, 2017. 10[49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. 用于图像识别的深度残差学习。 *CVPR*, 2016. 16[50] Dan Hendrycks, Kevin Gimpel. 高斯误差线性单元（GELUs）。 *arXiv:1606.08415*, 2016. 16[51] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, 等. 训练计算最优的大型语言模型。 *arXiv:2203.15556*, 2022. 1[52] Jungseok Hong, Michael Fulton, Junae Sattar. TrashCan：一个面向海洋垃圾视觉检测的语义分割数据集。 *arXiv:2007.08097*, 2020. 9, 19, 20[53] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, Kilian Q Weinberger. 具有随机深度的深度网络。 *ECCV*, 2016. 17[54] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, Humphrey Shi. Oneformer：一个统治通用图像分割的Transformer。 *arXiv:2211.06220*, 2022. 4
- [55] 贾超、杨寅飞、夏晔、陈怡婷、Zarana Parekh、Hieu Pham、Quoc Le、Sung Yun-Hsuan、李臻、Tom Duerig。利用噪声文本监督扩展视觉和视觉-语言表征学习。 *ICML*, 2021年。1[56] Jared Kaplun, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Dario Amodei。神经语言模型的缩放定律。 *arXiv:2001.08361*, 2020年。1[57] Michael Kass, Andrew Witkin, Demetri Terzopoulos。Snakes：主动轮廓模型。 *IJCV*, 1988年。4[58] Dahun Kim, 林腾毅、Anelia Angelova, In So Kweon, Weicheng Kuo。学习开放世界物体提议而无需学习分类。 *IEEE Robotics and Automation Letters*, 2022年。21[59] Alexander Kirillov, 何恺明、Ross Girshick, Carsten Rother, Piotr Dollár。全景分割。 *CVPR*, 2019年。4[60] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, Vittorio Ferrari。Open Images数据集 v4：大规模统一的图像分类、物体检测和视觉关系检测。 *IJCV*, 2020年。2, 6, 7, 18, 19[61] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, Thomas Dandres。量化机器学习的碳排放。 *arXiv:1910.09700*, 2019年。28[62] 李阳浩、Hanzi Mao, Ross Girshick, 何恺明。探索用于物体检测的朴素视觉Transformer骨干网络。 *ECCV*, 2022年。5, 10, 11, 16, 21, 23, 24[63] 李寅、叶哲凡、James M. Rehg。探究自我中心行为。 *CVPR*, 2015年。9, 20[64] 李竹文、陈启峰、Vladlen Koltun。具有潜在多样性的交互式图像分割。 *CVPR*, 2018年。5, 17, 19[65] 林腾毅、Priya Goyal, Ross Girshick, 何恺明、Piotr Dollár。用于密集物体检测的焦点损失。 *ICCV*, 2017年。5, 17[66] 林腾毅、Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。 *ECCV*, 2014年。2, 4, 6, 7, 11, 18, 19, 20[67] 刘钦、Zhenlin Xu、Gedas Bertasius, Marc Niethammer。SimpleClick：使用简单视觉Transformer进行交互式图像分割。 *arXiv:2210.11006*, 2022年。8, 9, 12, 19[68] Ilya Loshchilov, Frank Hutter。解耦权重衰减正则化。 *ICLR*, 2019年。17[69] Cathy H Lucas, Daniel OB Jones, Catherine J Hollyhead, Robert H Condon, Carlos M Duarte, William M Graham, Kelly L Robinson, Kylie A Pitt, Mark Schildhauer, Jim Regetz。全球海洋中的胶质浮游动物生物量：地理变异与环境驱动因素。 *Global Ecology and Biogeography*, 2014年。20[70] Sabarinath Mahadevan, Paul Voigtlaender, Bastian Leibe。迭代训练的交互式分割。 *BMVC*, 2018年。4, 17[71] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, Luc Van Gool。深度极限切割：从极值点到物体分割。 *CVPR*, 2018年。6[72] David Martin, Charless Fowlkes, Doron Tal, Jitendra Malik。人类分割自然图像数据库及其在评估分割算法和测量生态统计中的应用。 *ICCV*, 2001年。10, 21, 28[73] Fausto Milletari, Nassir Navab, Seyed-Ahmad Ahmadi。V-Net：用于体积医学图像分割的全卷积神经网络。 *3DV*, 2016年。5, 17[74] Massimo Minervini, Andreas Fischbach, Hannu Scharr, Sotirios A. Tsaftaris。用于基于图像的植物表型分析的细粒度标注数据集。 *Pattern Recognition Letters*, 2016年。9, 20[75] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru。模型报告用的模型卡片。 *Proceedings of the conference on fairness, accountability, and transparency*, 2019年。25, 28

[76] Dim P Papadopoulos、Jasper RR Uijlings、Frank Keller 和 Vittorio Ferrari。用于高效对象标注的极限点击。*ICCV*, 2017年。6[77] David Patterson、Joseph Gonzalez、Quoc Le、Chen Liang、Lluis-Miquel Munguia、Daniel Rothchild、David So、Maud Texier 和 Jeff Dean。碳排放与大型神经网络训练。*arXiv:2104.10350*, 2021年。28[78] Matthew E Peters、Waleed Ammar、Chandra Bhagavatula 和 Russell Power。使用双向语言模型的半监督序列标注。

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017年。18[79] Mengyang Pu、Yaping Huang、Yuming Liu、Qingji Guan 和 Haibin Ling。EDTER：使用Transformer进行边缘检测。*CVPR*, 2022年。10[80] Matti a Pugliatti 和 Francesco Toppo。DOORS：用于巨石分割的数据集。*Zenodo*, 2022年。9, 20[81] Jiyang Qi、Yan Gao、Yao Hu、Xinggang Wang、Xiaoyu Liu、Xiang Bai、Serge Belongie、Alan Yuille、Philip Torr 和 Song Bai。遮挡视频实例分割：一项基准测试。*ICCV*, 2022年。9, 20, 23, 24[82] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。从自然语言监督中学习可迁移的视觉模型。*ICML*, 2021年。1, 2, 4, 5, 8, 12, 16, 22[83] Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen 和 Ilya Sutskever。零样本文本到图像生成。*ICML*, 2021年。1, 4, 12[84] Shaoqing Ren、Kaiming He、Ross Girshick 和 Jian Sun。Faster R-CNN：基于区域提议网络实现实时目标检测。*NeurIPS*, 2015年。6, 10[85] Xiaofeng Ren 和 Jitendra Malik。学习用于分割的分类模型。*ICCV*, 2003年。4[86] Mike Roberts、Jason Ramapuram、Anurag Ranjan、Atulit Kumar、Miguel Angel Bautista、Nathan Paczan、Russ Webb 和 Joshua M. Susskind。Hypersim：用于整体室内场景理解的光照真实合成数据集。*ICCV*, 2021年。9, 19, 20[87] Candice Schumann、Susanna Ricco、Utsav Prabhu、Vittorio Ferrari 和 Caroline Pantofaru。迈向更包容的、促进公平性的人物标注。

Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021年。8, 19[88] Sefik Ilkin Serengil 和 Alper Ozpinar。LightFace：一个混合深度人脸识别框架。*ASYU*, 2020年。26[89] Sefik Ilkin Serengil 和 Alper Ozpinar。HyperExtended LightFace：一个人脸属性分析框架。*ICEET*, 2021年。26[90] Jamie Shotton、John Winn、Carsten Rother 和 Antonio Criminisi。TextonBoost：用于多类对象识别与分割的外观、形状和上下文联合建模。*ECCV*, 2006年。4[91] Corey Snyder 和 Minh Do。STREETS：一个新颖的交通流摄像头网络数据集。*NeurIPS*, 2019年。9, 20[92] Konstantin Sofiuk、Ilya A Petrov 和 Anton Konushin。通过掩码引导复兴交互式分割的迭代训练。*ICIP*, 2022年。5, 8, 9, 17, 19, 23, 24, 28[93] Nitish Srivastava、Geoffrey Hinton、Alex Krizhevsky、Ilya Sutskever 和 Ruslan Salakhutdinov。Dropout：一种防止神经网络过拟合的简单方法。*The Journal of Machine Learning Research*, 2014年。16[94] Chris Stauffer 和 W Eric L Grimson。用于实时跟踪的自适应背景混合模型。*CVPR*, 1999年。4[95] Matthew Tancik、Pratul Srinivasan、Ben Mildenhall、Sara Fridovich-Keil、Nithin Raghavan、Utkarsh Singhal、Ravi Ramamoorthi、Jonathan Barron 和 Ren Ng。傅里叶特征让网络在低维域中学习高频函数。*NeurIPS*, 2020年。5, 16[96] Yansong Tang、Yi Tian、Jiwen Lu、Jianjiang Feng 和 Jie Zhou。RG-B-D 第一人称视频中的动作识别。*ICIP*, 2017年。20

[97] 唐岩松, 王梓安, 卢继文, 冯建江, 周杰。用于RGB-D第一人称动作识别的多流深度神经网络。

IEEE Transactions on Circuits and Systems for Video Technology, 2019。20[98] 世界银行。按收入和地区划分的世界, 2022年。<https://datatopics.worldbank.org/world-development-indicators/the-world-by-income-and-region.html>。18[99] Sebastian Thrun。学习第n件事比学习第一件事更容易吗? *NeurIPS*, 1995。12[100] Cameron Trotter, Georgia Atkinson, Matt Sharpe, Kirsten Richardson, A. Stephen McGough, Nick Wright, Ben Burville, Per Berggren, NDD20：一个用于粗粒度和细粒度分类的大规模少样本海豚数据集。*arXiv:2005.13359*, 2020。9, 19, 20, 23, 24[101] 美国环境保护署。温室气体等价物计算器。<https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator>, 2022。28[102] Koen EA van de Sande, Jasper RR Uijlings, Theo Gevers, Arnold W M Smeulders。分割作为目标识别的选择性搜索。*ICCV*, 2011。10[103] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin。注意力就是你所需要的一切。*NeurIPS*, 2017。5, 16[104] 王博颖, 张立波, 文龙吟, 刘祥龙, 吴彦军。面向真实世界违禁品检测：一个大规模X射线基准。*CVPR*, 2021。9, 19, 20[105] Weiyao Wang, Matt Feiszli, Heng Wang, Jitendra Malik, Du Tran。开放世界实例分割：利用学习到的成对亲和性伪真值。*CVPR*, 2022。21[106] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, Georgia Gkioxari。用于3D重建的多视角压缩编码。*CVPR*, 2023。12[107] Jianxiong Xiao, James Hays, Krista Ehinger, Aude Oliva, Antonio Torralba。SUN数据库：从修道院到动物园的大规模场景识别。*CVPR*, 2010。20[108] 谢赛宁, 涂卓文。整体嵌套边缘检测。*ICCV*, 2015。10[109] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, Thomas S Huang。深度交互式对象选择。*CVPR*, 2016。4, 19[110] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, Olga Russakovsky。迈向更公平的数据集：过滤和平衡ImageNet层次结构中人物子树的分布。*Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020。8[111] 杨磊, 魏延子, 何宜升, 孙伟, 黄振航, 黄海斌, 范浩强。iShape：迈向不规则形状实例分割的第一步。*arXiv:2109.15068*, 2021。9, 20, 23, 24[112] Senthil Yogamani, Ciarán Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O’Dea, Michal Uricár, Stefan Milz, Martin Simon, Karl Amende, 等。WoodScape：一个用于自动驾驶的多任务、多相机鱼眼数据集。*ICCV*, 2019。9, 20[113] 张凌志, 周圣皓, Simon Stent, 石建波。细粒度第一人称手-物分割：数据集、模型与应用。*ECCV*, 2022。9, 19, 20[114] 张文武, 庞江森, 陈恺, Loy Chen Change。K-Net：迈向统一的图像分割。*NeurIPS*, 2021。4[115] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang。男人也喜欢购物：使用语料库级约束减少性别偏见放大。*arXiv:1707.09457*, 2017。8[116] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, Antonio Torralba。Places：一个用于场景识别的1000万图像数据库。*TPAMI*, 2017。20[117] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, Antonio Torralba。通过ADE20K数据集对场景进行语义理解。*IJCV*, 2019。2, 7, 9, 20

附录

目录:

- §A: Segment Anything 模型与任务详解
- §B: 自动掩码生成详情
- §C: RAI 补充详情
- §D: 实验实施细节
- §E: 人类研究实验设计
- §F: 数据集、标注与模型卡片
- §G: 标注指南

A. 分割任意模型与任务详情

图像编码器。通常，图像编码器可以是任何输出 $C \times H \times W$ 图像嵌入的网络。出于可扩展性和利用强大预训练模型的考虑，我们采用 MAE [47] 预训练的 Vision Transformer (ViT) [33] 并仅进行最小调整以处理高分辨率输入，具体为 ViT-H/16 模型，其采用 14×14 窗口注意力机制和四个等间距的全局注意力块（遵循 [62] 的方法）。图像编码器的输出是输入图像的 $16 \times$ 降采样嵌入。由于我们的运行时目标是实时处理每个提示，因此可以承担较高的图像编码器 FLOPs 开销，因为每个图像仅需计算一次编码，*not* 每个提示。

遵循标准做法 (*e.g.*, [40])，我们通过重新缩放图像并填充较短边，使用 1024×1024 的输入分辨率。因此，图像嵌入为 64×64 。为降低通道维度，依据[62]，我们使用 1×1 卷积将通道数降至 256，随后接一个 3×3 卷积（同样保持 256 通道）。每个卷积后均进行层归一化处理[4]。

提示编码器。稀疏提示被映射为 256 维向量嵌入，具体方式如下：点的表示由其位置的位置编码[95]与两个可学习嵌入之一相加而成，这两个嵌入分别指示该点属于前景或背景。框的表示由一对嵌入构成：(1) 其左上角位置编码与表示“左上角”的可学习嵌入相加；(2) 采用相同结构，但使用表示“右下角”的可学习嵌入。最后，为表示自由文本，我们使用 CLIP[82] 的文本编码器（原则上任何文本编码器均可使用）。本节后续内容将主要关注几何提示，文本提示的详细讨论见§D.5。

密集提示 (*i.e.*, 如掩码) 与图像具有空间对应关系。我们以比输入图像低 $4 \times$ 倍的分辨率输入掩码，随后通过两个 2×2 、步长为 2 的卷积层（输出通道数分别为 4 和 16）进一步下采样 $4 \times$ 倍。最后一层 1×1 卷积将通道维度映射至 256。每层之间均使用 GELU 激活函数[50]和层归一化进行分隔。掩码

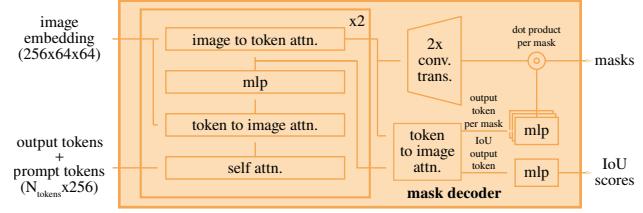


图14：轻量级掩码解码器的细节。一个双层解码器通过交叉注意力更新图像嵌入和提示令牌。随后，图像嵌入被上采样，更新后的输出令牌用于动态预测掩码。（为图示清晰未展示：在每个注意力层中，位置编码被添加到图像嵌入中，且完整的原始提示令牌（包括位置编码）被重新添加到令牌查询和键中。）

图像嵌入随后按元素相加。如果没有掩码提示，则向每个图像嵌入位置添加一个代表“无掩码”的学习嵌入。

轻量级掩码解码器。该模块高效地将图像嵌入和一组提示嵌入映射到输出掩码。为了融合这些输入，我们借鉴了Transformer分割模型[14, 20]的思路，并改进了标准Transformer解码器[103]。在应用解码器之前，我们首先在提示嵌入集合中插入一个可学习的输出令牌嵌入，该嵌入将用于解码器输出，类似于[33]中的[class]令牌。为简化表述，我们将这些嵌入（包括图像嵌入的{v*}）统称为“令牌”。

我们的解码器设计如图14所示。每个解码器层执行4个步骤：(1) 对标记进行自注意力计算，(2) 从标记（作为查询）到图像嵌入的交叉注意力计算，(3) 逐点MLP更新每个标记，以及(4) 从图像嵌入（作为查询）到标记的交叉注意力计算。最后一步利用提示信息更新图像嵌入。在交叉注意力计算期间，图像嵌入被视为一组 64^2 个 256 维向量。每个自注意力/交叉注意力模块和 MLP 均包含残差连接[49]、层归一化以及训练时 0.1 的 dropout[93]。下一层解码器接收来自前一层的更新后标记和更新后图像嵌入。我们使用两层解码器结构。

为确保解码器能够获取关键的几何信息，每当图像嵌入参与注意力层时，都会向其添加位置编码。此外，每当原始提示标记（包括其位置编码）参与注意力层时，它们都会被重新添加到更新后的标记中。这使得模型能够强烈依赖于提示标记的几何位置和类型。

运行解码器后，我们通过两个转置卷积将更新后的图像嵌入上采样 $4 \times$ 倍。

层（现在它相对于输入图像已下采样 $4\times$ ）。接着，这些标记再次关注图像嵌入，并将更新后的输出标记嵌入传递给一个小型3层MLP，该MLP输出一个向量，与上采样图像嵌入的通道维度相匹配。最后，我们通过上采样图像嵌入与MLP输出之间的空间逐点乘积来预测掩码。

Transformer的嵌入维度为256。其MLP模块的内部维度较大，达到2048，但MLP仅应用于提示词元，这些词元数量相对较少（通常不超过20个）。然而，在交叉注意力层中，由于我们使用 64×64 的图像嵌入，为了计算效率，我们将查询、键和值的通道维度降低了 $2\times$ 至128。所有注意力层均使用8个头。

用于放大输出图像嵌入的转置卷积为 2×2 ，步长为2，输出通道维度分别为64和32，并采用GELU激活函数。它们之间通过层归一化进行分隔。

使模型具备歧义感知能力。如前所述，单个输入提示可能存在歧义，即对应多个有效掩码，而模型将学会对这些掩码进行平均处理。我们通过一个简单的修改消除了这个问题：不再预测单个掩码，而是使用少量输出标记同时预测多个掩码。默认情况下我们预测三个掩码，因为我们观察到三个层级（整体、部分和子部分）通常足以描述嵌套掩码。在训练期间，我们计算真实标注与每个预测掩码之间的损失（稍后详述），但仅反向传播损失最低的预测结果。这是多输出模型常用的技术[15, 45, 64]。为便于实际应用中对预测掩码进行排序，我们增加了一个小型头部（基于额外输出标记运行），用于估计每个预测掩码与其覆盖对象之间的交并比 $\{v^*\}$ 。

使用多个提示时，模糊性会大大减少，三个输出掩码通常会变得相似。为了在训练时最小化退化损失的计算，并确保单一明确掩码获得常规梯度信号，当给出多个提示时，我们只预测一个掩码。这是通过添加第四个输出标记来实现的，用于额外的掩码预测。这个第四掩码在单个提示时从不返回，而在多个提示时是唯一返回的掩码。

损失函数。我们采用[20, 14]的方法，使用焦点损失[65]和骰子损失[73]的线性组合来监督掩码预测，焦点损失与骰子损失的比例为20:1。与[20, 14]不同，我们发现在每个解码层后进行辅助深度监督并无帮助。IoU预测头通过计算IoU预测值与预测掩码相对于真实掩码的IoU之间的均方误差损失进行训练。该损失以1.0的固定比例系数与掩码损失相加。

训练算法。遵循近期方法[92, 37]，我们在训练过程中模拟交互式分割设置。首先，以相等概率为目标掩码随机选择前景点或边界框。点从真实掩码中均匀采样。边界框采用真实掩码的边界框，并在每个坐标上添加随机噪声，其标准差等于框边长的10%，最大不超过20像素。这种噪声配置在实例分割（围绕目标物体生成紧密边界框）和交互式分割（用户可能绘制宽松边界框）等应用场景间实现了合理的折衷。

在根据第一个提示做出预测后，后续点会从前一个掩码预测与真实掩码之间的误差区域中均匀选取。若误差区域为假阴性，则新点为前景点；若为假阳性，则为背景点。我们还将上一轮迭代的掩码预测作为额外提示输入模型。为向新一轮迭代提供最完整的信息，我们提供未经过阈值处理的掩码逻辑值而非二值化掩码。当返回多个掩码时，传递给下一轮迭代并用于采样下一点的掩码是预测IoU最高的那个。

我们发现，在迭代采样8个点后（我们已测试至16个点），收益开始递减。此外，为鼓励模型从提供的掩码中获益，我们还增加了两次不额外采样点的迭代。其中一次迭代随机插入在8个迭代采样点之间，另一次始终置于末尾。这形成了总计11次迭代：一次初始输入提示采样、8次迭代采样点，以及两次未向模型提供新外部信息的迭代，使模型能够学习优化自身的掩码预测。我们注意到，采用相对较多的迭代次数是可行的，因为我们的轻量级掩码解码器仅需图像编码器不到1%的计算量，因此每次迭代仅增加少量开销。这与先前每次优化器更新仅执行一步或少量交互步骤的交互方法[70, 9, 37, 92]形成鲜明对比。

训练方案。我们使用AdamW [68]优化器（ $\beta = 0.9$, $\beta = 0.999$ ）和线性学习率预热[42]（250次迭代），并采用分步学习率衰减计划。预热后的初始学习率为 $8e-4$ 。我们训练90k次迭代（约2个SA-1B周期），并在60k次迭代和86666次迭代时将学习率降低10倍。批处理大小为256张图像。为规范SAM，我们设置权重衰减为0.1，并采用丢弃路径[53]（比率0.4）。我们使用分层学习率衰减[5]（系数0.8）。未应用数据增强。SAM的初始化基于MAE [47]预训练的ViT-H模型。由于图像编码器较大且输入尺寸为 1024×1024 ，我们在256个GPU上分布式训练。为限制GPU内存——

在内存使用方面，我们训练时每个GPU最多使用64个随机采样的掩码。此外，我们发现对SA-1B掩码进行轻度过滤——丢弃任何覆盖图像超过90%的掩码——能在质量上提升结果。

对于训练中的消融实验及其他变体 (*e.g.*, 文本到掩码 §D.5)，我们按如下方式调整上述默认方案：当仅使用第一和第二阶段数据引擎的数据进行训练时，我们采用尺度范围为[0.1, 2.0]的大尺度抖动[40]对输入进行增强。直观而言，在训练数据较为有限时，数据增强可能更有益。为训练ViT-B和ViT-L模型，我们使用18万次迭代，批大小为128，并分配到128个GPU上执行。针对ViT-B/L模型，我们分别设置 $lr = 8e^{-4}/4e^{-4}$ 、 $ld = 0.6/0.8$ 、 $wd = 0.1$ 、 $dp = 0.6/0.4$ 。

B. 自动掩码生成细节

这里我们讨论数据引擎全自动阶段的细节，该阶段用于生成已发布的SA-1B。

裁剪。掩码生成于完整图像上 32×32 点的规则网格，以及20个额外放大的图像裁剪区域，这些区域分别来自 2×2 和 4×4 部分重叠窗口，并分别使用 16×16 和 8×8 规则点网格。原始高分辨率图像被用于裁剪（这是我们唯一使用它们的情况）。我们移除了触及裁剪区域内部边界的掩码。我们分两个阶段应用了标准的基于贪婪框的非极大值抑制（为提升效率而使用框）：首先在每个裁剪区域内进行，其次在不同裁剪区域之间进行。在裁剪区域内应用NMS时，我们使用模型预测的交并比来对掩码排序。在不同裁剪区域间应用NMS时，我们根据掩码的来源裁剪区域，从最放大（*i.e.*，来自 4×4 裁剪区域）到最不放大（*i.e.*，原始图像）对掩码进行排序。在两种情况下，我们使用的NMS阈值均为0.7。

过滤。我们使用了三种过滤器来提升掩码质量。首先，为保留仅*confident*掩码，我们依据模型预测的IoU分数以88.0为阈值进行筛选。其次，为保留仅*stable*掩码，我们通过在不同阈值下对同一基础软掩码进行二值化，比较生成的两个二值掩码。仅当预测结果（*i.e.*，即逻辑值以0为阈值生成的二值掩码）与其对应的-1和+1阈值掩码之间的IoU大于或等于95.0时，我们才保留该预测。第三，我们注意到自动生成的掩码偶尔会覆盖整张图像。这类掩码通常缺乏实际意义，因此我们通过移除覆盖图像面积95%及以上的掩码进行过滤。所有过滤阈值的选择均旨在同时获得大量掩码与高质量掩码，该标准由专业标注人员依据§5所述方法进行评估。

后处理。我们观察到两种错误类型，通过后处理可以轻松缓解。首先，大约4%的掩码包含微小的伪影成分。为解决这一问题，我们移除了面积小于

超过100像素（包括如果最大组件低于此阈值，则移除整个掩码）。其次，另外估计有4%的掩码包含小的伪影孔洞。为解决这些问题，我们填充了面积小于100像素的孔洞。孔洞通过反转掩码的组件来识别。

自动掩码生成模型。我们训练了一个特殊版本的SAM，用于全自动掩码生成，该版本牺牲了一定的推理速度以提升掩码生成性能。我们在此指出默认SAM与用于数据生成的版本之间的差异：后者仅基于手动和半自动数据进行训练，训练时长更长（177656次迭代而非9万次），并采用大规模抖动数据增强[40]；模拟交互式训练仅使用点和掩码提示（无框提示），且每个掩码在训练期间仅采样4个点（从默认的9点减少至4点，这加快了训练迭代速度且对单点性能无影响，但若使用更多点评估会损害mIoU）；最后，其掩码解码器使用3层结构而非2层。

SA-1B示例。我们在图2中展示了SA-1B样本。更多示例请参阅我们的数据集浏览器。

C. RAI附加详情

推断SA-1B的地理信息。虽然SA-1B中的图像未附带地理标签，但每张图像都配有描述其内容和拍摄地点的说明文字。我们使用基于Elmo的命名实体识别模型[78]从这些说明文字中推断出近似的地理位置。每个提取出的地点实体都会被映射到所有匹配的国家、省份和城市。通过首先考虑匹配的国家，然后是省份，最后是城市，将说明文字映射到单一国家。我们注意到这种方法存在模糊性和潜在的偏见（{v*}，例如“Georgia”可能指国家或美国州份）。因此，我们使用提取的位置来分析整个数据集，但不会发布推断出的位置。根据图像提供商的要求，说明文字也不会公开。

推断COCO和Open Images的地理信息。COCO [66]和Open Images [60]数据集未提供地理位置信息。依据[29]的方法，我们使用Flickr API检索地理元数据。我们获取了COCO训练集中24%的图像位置（19,562张图像），对于Open Images则获取了训练集中18%的图像位置（仅考虑带标注掩码的图像，共493,517张）。需要说明的是，所获地理信息为近似值，且带有此类信息的图像样本可能无法完全代表完整数据集的分布。

推断收入信息。我们利用每张图片推断出的国家，参照世界银行[98]定义的收入等级来查找对应的收入水平。我们将中高收入和中低收入等级合并为一个单一的中等收入等级。

	mIoU at		mIoU at	
	1 point	3 points	1 point	3 points
<i>perceived gender presentation</i>				
feminine	76.3 ± 1.1	90.7 ± 0.5	older	81.9 ± 3.8
masculine	81.0 ± 1.2	92.3 ± 0.4	middle	78.2 ± 0.8
			young	77.3 ± 2.7
				91.5 ± 0.9

表6: SAM在感知性别呈现和年龄组中分割服装的性能。感知性别的区间是互斥的，其中男性类别的mIoU更高。年龄组的置信区间存在重叠。

公平性在人物分割中的考量。为了研究SAM在人物分割中的公平性，我们采用了更具包容性的人物标注数据集（MIAP）[87]中的测试集标注，该数据集基于Open Images [60]，使我们能够比较SAM在不同感知性别表现和感知年龄组上的性能。MIAP提供边界框标注，而本次分析需要真实掩码。为获取真实掩码，我们从Open Images中选取每个人物类别掩码，若其对应边界框与MIAP标注边界框的边长相对误差在1%以内，则予以保留，最终得到3.9k个掩码。

服装分割的公平性。我们将第6节的分析延伸至服装分割领域，考察SAM在服装分割上的表现与穿着者属性之间的关系。我们使用了Open Images数据集中所有位于MIAP人体检测框内、且属于服装大类的6.5万个真实标注掩码。表6展示了模型在不同感知性别表达与年龄组间的性能对比。研究发现，SAM在呈现显著男性化特征的穿着者服装分割上表现更优（95%置信区间无重叠）。当评估点从1个增至3个时，这种差异逐渐缩小。不同感知年龄组间的差异则未达显著水平。结果表明，在使用单点提示进行服装分割时，模型存在针对感知性别表达的偏差，我们建议SAM用户注意这一局限性。

D. 实验实施细节

D.1. 零样本单点有效掩码评估

数据集。我们构建了一个新的分割基准，利用先前工作中的23个多样化分割数据集套件，评估我们模型的零样本迁移能力。每个数据集的描述见表7。示例可参见正文图8。该套件涵盖多个领域，包括第一视角[34, 28, 113]、显微成像[12]、X射线[104]、水下[52, 100]、航拍[17]、仿真[86]、驾驶[25]和绘画[24]图像。为提升评估效率，我们对包含超过1.5万个掩码的数据集进行了子采样。具体而言，我们随机选取图像，使采样图像中的掩码总数达到~10k。我们对所有数据集中的人脸进行了模糊处理。

点采样。我们的默认点采样遵循交互式分割中的标准实践[109, 64, 92]。第一个点被确定性选择为距离物体边界最远的点。随后的每个点则是选择在真实标注与先前预测之间的误差区域边界上距离最远的点。部分实验（在指定处）采用更具挑战性的采样策略：第一个点选用random点，而非确定性选择的“中心”点。后续各点仍按上述方式选取。这一设定能更好地反映首点未必可靠地位于掩码中心的使用场景，例如通过视线注视进行提示的情况。

评估。我们测量经过N个点提示后的预测与真实掩码之间的IoU，其中 $N = \{1, 2, 3, 5, 9\}$ ，且点通过上述任一策略迭代采样。每个数据集的mIoU是该数据集中所有对象掩码IoU的平均值。最后，我们通过汇总所有23个数据集的每数据集mIoU来计算核心指标。我们的评估与标准交互式分割评估协议不同，后者衡量达到X% IoU所需的平均点数（最多20个点）。我们专注于仅使用一个或少数几个点后的预测，因为我们的许多应用场景仅涉及单个或极少量提示。基于我们注重实时提示处理的应用需求，我们预期最佳交互式分割模型在使用大量点数时应优于SAM。

基线方法。我们采用三种近期表现优异的交互式分割基线：RITM [92]、FocalClick [18] 和 SimpleClick [67]。针对每种方法，均使用作者公开的、基于最广泛数据集训练的最大模型。对于RITM，采用作者提出的基于COCO [66]和LVIS [44]联合数据集训练的HRNet32 I-T-M模型。对于FocalClick，采用在包含8个不同分割数据集的“联合数据集”[18]上训练的SegFormerB3-S2模型。对于SimpleClick，采用基于COCO和LVIS联合数据集训练的ViT-H448模型。我们遵循建议的默认数据预处理策略（*i.e.*，包括数据增强或图像尺寸调整），在评估过程中未修改或调整任何参数。实验发现，在单点标注的23个数据集测试中，RITM优于其他基线方法，因此将其作为默认基线。在进行多点标注评估时，我们将汇报所有基线的结果。

单点歧义与神谕评估。除了在N点提示后的IoU，我们还报告了SAM在单点上的“神谕”性能，通过从SAM的三个预测掩码中评估与真实情况最匹配的预测掩码（而非默认使用SAM自身排名第一的掩码）。这一方法通过放宽从多个有效对象中猜测唯一正确掩码的要求，解决了可能的单点提示歧义问题。

dataset	abbreviation & link	image type	description	mask type	source split	# images sampled	# masks sampled
Plant Phenotyping Datasets Leaf Segmentation [74]	PPDLS	Plants	Leaf segmentation for images of tobacco and aral plants.	Instance	N/A	182	2347
BBBC038v1 from Broad Bioimage Benchmark Collection [12]	BBBC038v1	Microscopy	Biological images of cells in a variety of settings testing robustness in nuclei segmentation.	Instance	Train	227	10506
Dataset fOr bOuldeRs Segmentation [80]	DOORS	Boulders	Segmentation masks of single boulders positioned on the surface of a spherical mesh.	Instance	DS1	10000	10000
TimberSeg 1.0 [38]	TimberSeg	Logs	Segmentation masks of individual logs in piles of timber in various environments and conditions. Images are taken from an operator's point-of-view.	Instance	N/A	220	2487
Northumberland Dolphin Dataset 2020 [100]	ND20	Underwater	Segmentation masks of two different dolphin species in images taken above and under water.	Instance	N/A	4402	6100
Large Vocabulary Instance Segmentation [44]	LVIS	Scenes	Additional annotations for the COCO [66] dataset to enable the study of long-tailed object detection and segmentation.	Instance	Validation (v0.5)	945	9642
STREETS [91]	STREETS	Traffic camera	Segmentation masks of cars in traffic camera footage.	Instance	N/A	819	9854
ZeroWaste-f [6]	ZeroWaste-f	Recycling	Segmentation masks in cluttered scenes of deformed recycling waste.	Instance	Train	2947	6155
iShape [111]	iShape	Irregular shapes	Segmentation masks of irregular shapes like antennas, logs, fences, and hangers.	Instance	Validation	754	9742
ADE20K [117]	ADE20K	Scenes	Object and part segmentation masks for images from SUN [107] and Places [116] datasets.	Instance	Validation	302	10128
Occluded Video Instance Segmentation [81]	OVIS	Occlusions	Instance segmentation masks in videos, focusing on objects that are occluded.	Instance	Train	2044	10011
Hypersim [86]	Hypersim	Simulation	Photorealistic synthetic dataset of indoor scenes with instance masks.	Instance	Evermotion archinteriors volumes 1-55 excluding 20,25,40,49	338	9445
Night and Day Instance Segmented Park [22, 23]	NDISPARK	Parking lots	Images of parking lots from video footage taken at day and night during different weather conditions and camera angles for vehicle segmentation.	Instance	Train	111	2577
EPIC-KITCHENS VISOR [28, 27]	VISOR	Egocentric	Segmentation masks for hands and active objects in ego-centric video from the cooking dataset EPIC-KITCHENS [27].	Instance	Validation	1864	10141
Plittersdorf dataset [46]	Plittersdorf	Stereo images	Segmentation masks of wildlife in images taken with the SOCRATES stereo camera trap.	Instance	Train, validation, test	187	546
Egocentric Hand-Object Segmentation [113]	EgoHOS	Egocentric	Fine-grained egocentric hand-object segmentation dataset. Dataset contains mask annotations for existing datasets.	Instance	Train (including only Ego4D [43] and THU-READ [97, 96])	2940	9961
InstanceBuilding 2D [17]	IBD	Drones	High-resolution drone UAV images annotated with roof instance segmentation masks.	Instance	Train (2D annotations)	467	11953
WoodScape [112]	WoodScape	Fisheye driving	Fisheye driving dataset with segmentation masks. Images are taken from four surround-view cameras.	Instance	Set 1	107	10266
Cityscapes [25]	Cityscapes	Driving	Stereo video of street scenes with segmentation masks.	Panoptic	Validation	293	9973
PIDRay [104]	PIDRay	X-ray	Segmentation masks of prohibited items in X-ray images of baggage.	Instance	Test (hard)	3733	8892
Diverse Realism in Art Movements [24]	DRAM	Paintings	Domain adaptation dataset for semantic segmentation of art paintings.	Semantic	Test	718	1179
TrashCan [52]	TrashCan	Underwater	Segmentation masks of trash in images taken by underwater ROVs. Images are sourced from the J-EDI [69] dataset.	Instance	Train (instance task)	5936	9540
Georgia Tech Egocentric Activity Datasets [34, 63]	GTEA	Egocentric	Videos are composed of four different subjects performing seven types of daily activities with segmentation masks of hands.	Instance	Train (segmenting hands task)	652	1208

表7：用于评估点提示零样本分割的数据集。这23个数据集涵盖广泛领域，详见“图像类型”列。为提高评估效率，我们对包含超过1.5万个掩码的数据集进行子采样。具体而言，我们随机抽取图像，使图像中掩码总数保持在~10k。

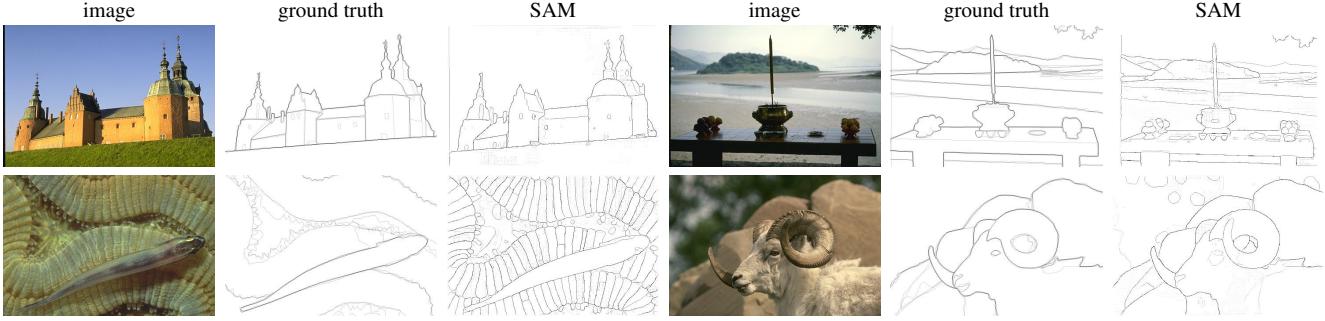


图15：BSDS500上零样本边缘预测的额外可视化。请记住，SAM并未针对预测边缘图进行训练，且在训练过程中未接触过BSDS图像及其标注。

D.2. 零样本边缘检测

数据集与评估指标。我们在BSDS500 [72, 3] 上进行零样本边缘检测实验。每张图像的真实标注来自五位不同标注者的手动标注。我们使用边缘检测的四个标准指标 [3, 32] 在包含200张图像的测试子集上报告结果：最优数据集尺度（ODS）、最优图像尺度（OIS）、平均精度（AP）以及50%精度下的召回率（R50）。

方法。对于零样本迁移，我们使用了一个简化版的自动掩码生成流程。我们通过一个 16×16 的规则前景点网格提示SAM，这产生了768个预测掩码（每个点三个）。我们不通过预测IoU或稳定性进行筛选。冗余掩码通过非极大值抑制（NMS）移除。然后，我们对剩余掩码的未阈值化概率图应用Sobel滤波器，并将不与掩码外边界像素相交的值设为零。最后，我们对所有预测进行逐像素取最大值，将结果线性归一化到[0,1]区间，并应用边缘NMS [13]来细化边缘。

可视化。在图15中，我们展示了SAM零样本边缘预测的更多示例。这些定性示例进一步说明，尽管SAM未经边缘检测训练，其输出仍倾向于生成合理的边缘图。我们可以看到，这些边缘能够与人工标注良好对齐。然而如前所述，由于SAM并非针对边缘检测任务进行训练，它并未学习BSDS500数据集的偏差，因此其输出的边缘数量往往多于真实标注中包含的边缘。

D.3. 零样本目标提议

数据集与评估指标。我们在LVIS v1验证集[44]上报告了1000个候选框时掩码的标准平均召回率（AR）指标。由于LVIS包含1203个物体类别的高质量掩码，这为物体候选框生成提供了具有挑战性的测试。我们重点关注AR@1000，因为我们的模型具有开放世界特性，可能会产生大量超出LVIS中1203个类别的有效掩码。为衡量模型在常见、普通和稀有类别上的性能——

在类别方面，我们使用AR@1000，但仅针对包含相应LVIS类别的基础真值集进行测量。

基线。我们采用级联ViTDet-H作为基线，这是[62]中在LVIS数据集上平均精度（AP）最强的模型。如正文所述，在域内训练的目标检测器可以“利用”平均召回率（AR）[16]，预计其作为基线会比其他专注于开放世界提议或分割的模型[58, 105]更强。为生成1000个提议，我们禁用了三个级联阶段的分数阈值设置，并将每个阶段的最大预测数量提升至1000。

方法。我们采用改进版的SAM自动掩码生成流程进行零样本迁移。首先，为保持推理时间与ViTDet相当，我们不对图像裁剪进行处理。其次，我们移除了基于预测IoU和稳定性的筛选步骤。这保留了可调节的两个参数以获取每张图像~1000个掩码：输入点网格和用于重复掩码抑制的NMS阈值。我们选择 64×64 点网格和0.9的NMS阈值，平均每张图像可生成~900个掩码。在评估阶段，若单张图像生成的掩码数量超过1000个，则根据置信度与稳定性分数的平均值进行排序，并截取前1000个候选掩码。

我们假设，SAM能够输出多个掩码的能力对于这项任务尤其宝贵，因为召回率应当受益于从单个输入点生成的多个尺度的提议。为了验证这一点，我们与一个消融版本的SAM进行比较，该版本仅输出单个掩码而非三个（SAM - 单输出）。由于该模型生成的掩码较少，我们进一步将采样点数量和NMS阈值分别提高至 128×128 和0.95，从而平均每张图像获得~950个掩码。此外，单输出SAM不会生成在自动掩码生成流程中用于NMS排序掩码的IoU分数，因此掩码改为随机排序。测试表明，这与更复杂的掩码排序方法（例如使用掩码的最大logit值作为模型置信度的代理）具有相似的性能。

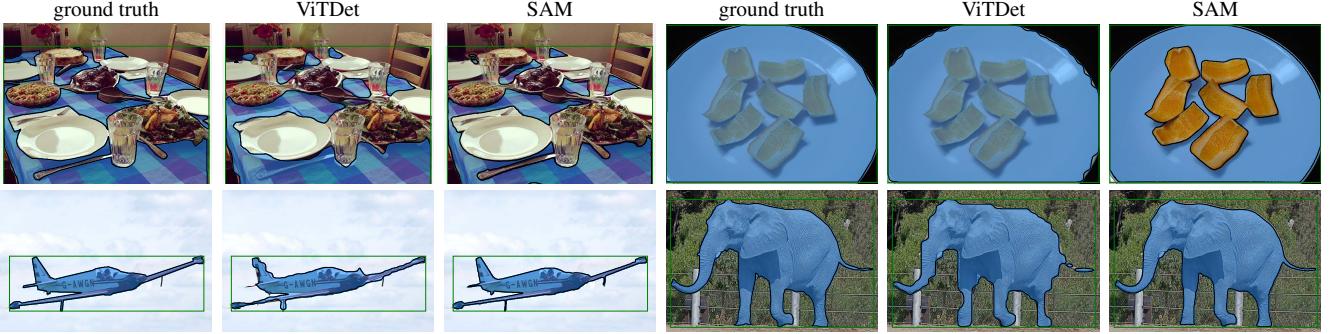


图16：LVIS v1上的零样本实例分割。SAM生成的掩码质量高于ViTDet。作为零样本模型，SAM没有机会学习特定训练数据的偏差；例如右上角所示，SAM做出了模态预测，而LVIS中的真实标注是非模态的，因为LVIS中的掩码标注没有孔洞。

D.4. 零样本实例分割

方法。对于零样本实例分割，我们使用全监督ViTDet在COCO和LVIS v1验证集上输出的边界框提示SAM。我们通过将置信度最高的预测掩码与边界框提示一同反馈给掩码解码器进行额外一轮掩码优化迭代，以生成最终预测结果。图16展示了在LVIS数据集上预测的零样本实例分割效果。与ViTDet相比，SAM倾向于生成边界更清晰的高质量掩码。我们通过§7.4中的人工研究证实了这一观察。值得注意的是，作为零样本模型，SAM无法学习数据集中存在的标注偏差。例如，我们观察到SAM对餐盘做出了合理的模态预测，而LVIS掩码因设计限制不允许包含孔洞，因此餐盘在标注中被处理为非模态形式。

D.5. 零样本文本到掩码

模型与训练。我们使用公开可用的最大CLIP模型[82]（ViT-L/14@336px）计算文本和图像嵌入，并在使用前对其进行 ℓ^2 归一化处理。为训练SAM，我们采用数据引擎前两个阶段生成的掩码，同时舍弃所有面积小于 100^2 像素的掩码。该模型通过大规模抖动增强[40]进行训练，迭代12万次，批处理大小为128。其余训练参数均遵循默认设置。

生成训练提示。为提取输入提示，我们首先将每个掩码周围的边界框按 $1\times$ 到 $2\times$ 的随机比例扩展，将扩展后的框进行方形裁剪以保持其宽高比，并调整至 336×36 像素。在将裁剪图像输入CLIP图像编码器前，我们以50%的概率将掩码外部的像素置零。为确保嵌入专注于目标对象，我们在最后一层使用掩码注意力机制，将输出令牌的注意力限制在掩码内的图像位置。最终，我们的提示即为输出令牌嵌入。训练时，我们首先提供基于CLIP的提示，随后添加额外的迭代点提示以优化预测结果。



图17：对SAM潜在空间中掩码嵌入相似度进行阈值处理的可视化。查询由洋红色框标示；顶行显示低阈值下的匹配结果，底行显示高阈值下的匹配结果。即使SAM未经过显式语义监督训练，同一图像中最相似的掩码嵌入通常与查询掩码嵌入在语义上具有相似性。

推理。在推理过程中，我们使用未经修改的CLIP文本编码器为SAM生成提示。我们基于CLIP已对齐文本与图像嵌入的特性，这使得我们能够在训练时不依赖任何显式文本监督，而在推理时使用基于文本的提示。

D.6. 探索SAM的潜在空间

最后，我们进行了一项初步研究，以定性探索SAM学习到的潜在空间。具体而言，我们关注的是，尽管SAM未在显式的语义监督下训练，它是否仍能在其表征中捕捉到语义信息。为此，我们通过以下步骤计算 $\{v^*\}$ ：从SAM中提取围绕掩码及其水平翻转版本的图像裁剪的图像嵌入，将图像嵌入与二值掩码相乘，并在空间位置上取平均值。在图17中，我们展示了同一图像中查询掩码及（潜在空间中）相似掩码的3个示例。我们观察到

每个查询的最近邻显示出一些形状和语义上的相似性，尽管并不完美。虽然这些结果是初步的，但它们表明SAM生成的表示可能对多种用途有益，例如进一步的数据标注、理解数据集内容，或作为下游任务的特征。

E. 人类研究实验设计

这里我们详细介绍了用于评估§7.1和§7.4中掩码质量的人类研究细节。该人类研究的目的在于解决使用与真实标注的交并比（IoU）作为预测掩码质量衡量标准的两点局限性。首先，对于如单点这样的模糊输入，模型可能因返回一个与真实标注不同但有效的物体掩码而受到严重惩罚。其次，真实标注掩码可能包含各种偏差，例如边缘质量的系统性误差，或对遮挡物体进行模态或非模态分割的决策。在特定领域训练的模型可能学会这些偏差，从而获得更高的IoU，却未必能生成更好的掩码。通过人工评审，我们可以获得一种独立于底层真实标注掩码的掩码质量衡量方法，以缓解这些问题。

模型。对于单点评估，我们使用RITM[92]、单输出SAM以及SAM来验证两个假设。首先，我们假设当给定单个点时，即使与真实标注的IoU等指标未能体现差异，SAM生成的掩码在视觉质量上仍优于基线交互式分割模型。其次，我们假设SAM区分歧义掩码的能力能够提升单点输入的掩码质量，因为单输出SAM可能返回对歧义掩码进行平均处理的结果。

在实例分割实验中，我们评估了级联ViTDet-H [62] 和SAM，以验证以下假设：即使SAM由于无法学习验证数据集的特定标注偏差而导致AP值较低，其生成的掩码在视觉上仍具有更高质量。

数据集。对于单点实验，我们从23个数据集中选取了7个，因为完整套件规模过大，不便于人工评估。我们选择了LVIS v0.5 [17]、VISOR [28, 27]、DRAM [24]、IBD [17]、NDD20 [100]、OVIS [81]和iShape [111]，这些数据集提供了多样化的图像集合，涵盖场景级、以自我为中心、手绘、俯拍、水下及合成图像。此外，该集合既包含SAM在IoU指标上优于RITM的数据集，也包含反之的数据集。对于实例分割实验，我们使用LVIS v1验证集，以便与在LVIS上训练的ViTDet进行直接比较。

方法。我们将模型生成的掩码呈现给专业标注员，并要求他们根据提供的指南对每个掩码进行评分（完整指南见§G）。标注员来自同一公司——

一家为数据引擎收集手动标注掩码的公司。标注员可以访问一张图像、单个模型的预测掩码以及模型的输入（单点或单框），并被要求根据三个标准评估掩码：该掩码是否对应有效物体？掩码边界是否清晰？以及掩码是否与输入对应？随后他们提交了1-10分的评分，以表示掩码的整体质量。

得分为1表示掩码完全不对应任何物体；低分（2-4分）表示掩码存在严重错误，例如包含大范围的其他物体区域或出现大面积无意义的边界；中等分数（5-6分）表示掩码基本合理但仍存在显著的语义或边界错误；高分（7-9分）表示掩码仅存在细微边界错误；而10分则代表掩码没有可见错误。标注人员获得了五种不同的视图，每种视图都旨在帮助识别不同类型错误。

在单点实验中，每个数据集随机选取1000个掩码，这些掩码来自用于基准测试零样本交互式分割的相同子集（有关这些子集的详细信息，请参见§D.1）。模型输入为最中心点，该点计算为掩码边缘距离变换的最大值。在实例分割实验中，从LVIS v1验证集中选取1000个掩码，模型输入为LVIS真实标注框。所有实验中，尺寸小于 24^2 像素的掩码均被排除在采样之外，以避免向评估者展示过小而难以准确判断的掩码。出于内存和显示原因，在预测掩码前，大型图像被重新缩放至最大边长为2000像素。所有实验中，相同的输入被馈送到每个模型以生成预测掩码。

为了比较，每个数据集的真实掩码也被提交进行评分。在单点实验中，每个数据集总共产生了4000个评分任务（RITM、SAM单输出、SAM和真实掩码各100个）；在实例分割实验中，则产生了3000个总任务（ViTDet、SAM和真实掩码）。

对于每个数据集，这些任务以随机顺序插入队列，由30名标注员从中领取任务。在审查研究的初步测试中，我们将每个任务分配给五名不同的标注员，发现评分具有合理的一致性：五名标注员评分的平均标准差为0.83。此外，标注公司部署了质量保证测试员，抽查部分结果以检测是否严重偏离指南。因此，在我们的实验中，每个任务（*i.e.*，即对单张图像中的一个口罩进行评分）仅由一名标注员完成。每名标注员处理每个任务的平均时间为90秒，虽超出我们最初设定的30秒目标，但仍足以在7个选定数据集上收集大量评分。

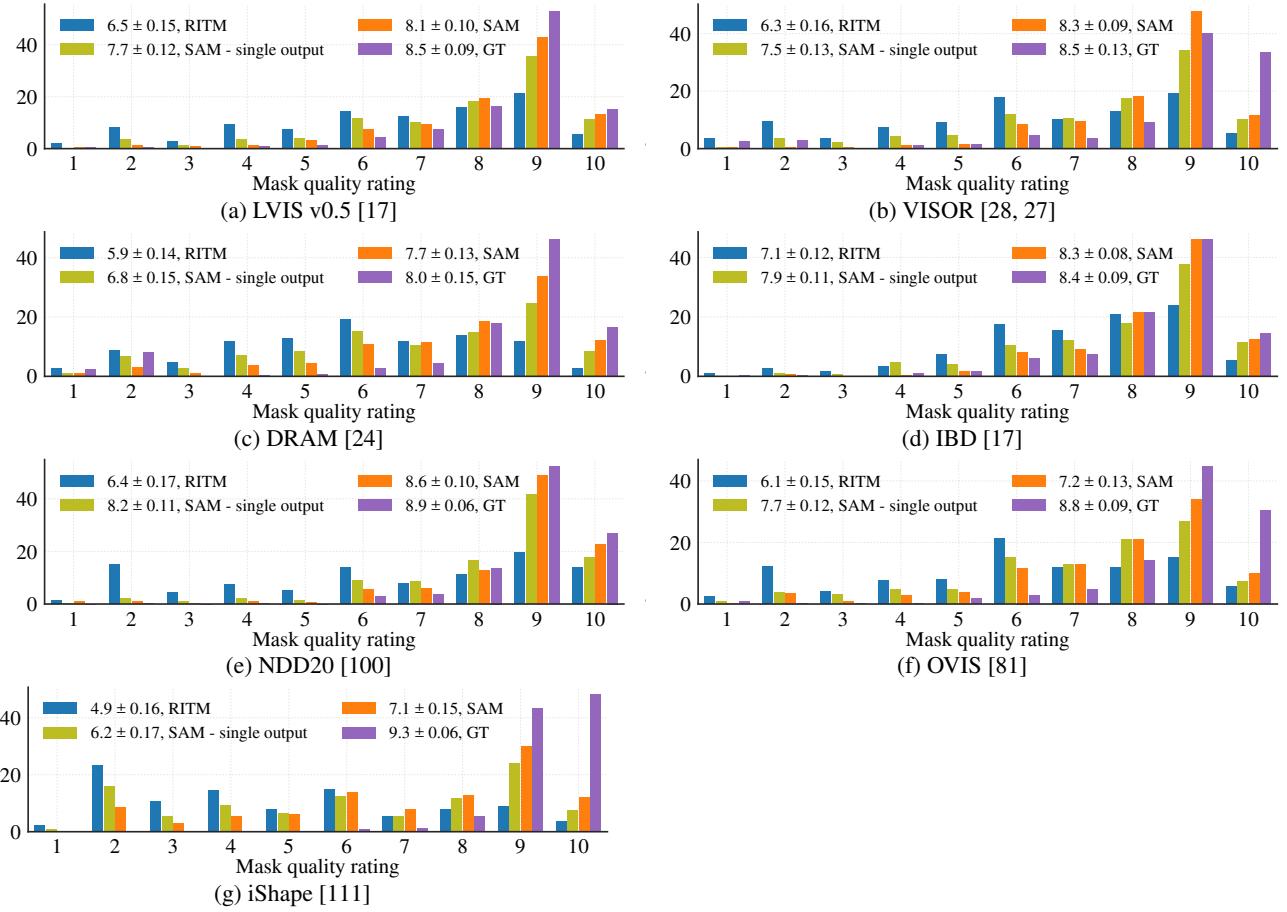


图18：我们人工评估研究中各数据集的掩码质量评分分布。

dataset	SAM > baseline		SAM > SAM single out.	
	p-value	CI ₉₉ ($\Delta\mu$)	p-value	CI ₉₉ ($\Delta\mu$)
<i>point input (RITM [92] baseline):</i>				
LVIS v0.5 [44]	4e-69	(1.40, 1.84)	2e-11	(0.29, 0.64)
VISOR [28, 27]	7e-98	(1.81, 2.24)	7e-26	(0.58, 0.94)
DRAM [24]	1e-76	(1.54, 2.00)	2e-24	(0.62, 1.03)
IBD [17]	2e-57	(1.03, 1.39)	1e-15	(0.32, 0.62)
NDD20 [100]	2e-86	(1.88, 2.37)	5e-08	(0.19, 0.55)
OVIS [81]	2e-64	(1.38, 1.84)	3e-10	(0.27, 0.63)
iShape [111]	2e-88	(1.97, 2.47)	7e-23	(0.65, 1.10)
<i>box input (ViTDet-H [62] baseline):</i>				
LVIS v1 [44]	2e-05	(0.11, 0.42)	N/A	N/A

表8：统计检验显示SAM的掩码质量评分显著高于基线和单输出SAM。P值通过配对t检验计算，而平均分数差异的置信区间则通过10k样本的配对自助法计算。所有p值均显著，且所有置信区间均不包含零值。

结果。图18展示了单点实验中每个数据集的评分直方图。我们进行了统计

对两个假设进行检验：(1) SAM 比基线模型 (RITM 或 ViTDet) 获得更高分数；(2) SAM 比单输出 SAM 获得更高分数。通过模型分数均值的配对 t 检验计算 p 值，并辅以 1 万次样本的配对自助法检验来求取均值差异的 99% 置信区间。表 8 展示了这些检验的 p 值与置信区间。所有统计检验均具有高度显著性，且所有置信区间均不包含零值。

例如，在实例分割任务中，主文本的图11展示了评分直方图。为了与COCO真实标注进行比较，我们额外纳入了794份在人类评审流程测试期间收集的COCO真实标注掩膜评分。这些掩膜以与LVIS结果完全相同的设置呈现给评分者。为公平比较，图11中LVIS的结果已对每个模型和真实标注统一子采样至相同的794个输入。在表8中，我们使用全部1000份评分进行统计检验，结果表明SAM在掩膜质量上相比ViTDet的提升具有统计显著性。

F. 数据集、标注与模型卡片

在§F.1中，我们按照[39]的格式，以问答列表的形式提供了SA-1B的数据集卡片。接着，在§F.2中我们依据 CrowdWorksheets[30]的规范，同样以问答列表的形式，为第4章所述数据引擎的前两个阶段提供了数据标注卡片。我们在表9中遵循[75]的框架提供了模型卡片。

F.1. SA-1B 数据集卡片

动机

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.* 我们的数据集对视觉社区的贡献体现在四个方面：(1) 我们发布了包含1100万张图像和11亿个掩码的数据集，这是迄今为止规模最大的分割数据集。(2) 我们发布的数据集注重隐私保护：所有图像中的人脸和车牌均已进行模糊处理。(3) 本数据集采用广泛的使用条款授权，具体内容可在<https://ai.facebook.com/datasets/segment-anything> 查看。(4) 该数据的地理分布比以往数据集更具多样性，我们希望这将推动社区在构建更公平、更均衡的模型方面向前迈进。

2. *Who created the dataset (例如, .. which team, research group) and on behalf of which entity (例如, .. company, institution, organization)?* 该数据集由Meta AI的FAIR团队创建。基础图像收集自第三方图片公司并已获授权。

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.* Meta AI 资助了该数据集的创建。

4. *Any other comments?* 否。

组成

1. *What do the instances that comprise the dataset represent (例如, .. documents, photos, people, countries)? Are there multiple types of instances (例如, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.* 数据集中的所有实例均为照片。照片的主题多样；常见的照片主题包括：地点、物体、场景。所有照片均不相同，但存在一些针对同一主题拍摄的照片组。

2. *How many instances are there in total (of each type, if appropriate)?* 有110万张图片。

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (例如, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (例如, to cover a more diverse range of instances, because instances were withheld or unavailable).* 该数据集由从图片供应商处获得授权的图像构成，包含了所有已授权的实例。这些图像均为照片，*i.e.*，而非艺术作品，尽管存在少数例外。数据集中包含了为每张图像生成的所有掩码。我们保留了~2k张随机选取的图像用于测试目的。

4. *What data does each instance consist of? "Raw" data (例如, .. unprocessed text or images) or features? In either case, please provide a description.* 数据集中的每个实例都是一张图像。这些图像经过处理，模糊了人脸和车牌，以保护图像中人物的身份。

5. *Is there a label or target associated with each instance? If so, please provide a description.* 每张图像都带有掩码标注。这些掩码没有关联的类别或文本信息。平均每张图像包含~100个掩码，掩码总数为~11亿个。

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (例如, because it was unavailable). This does not include intentionally removed information, but might include, 例如, redacted text.* 是的。每张图片都附有一段简短的说明文字，以自由文本形式描述照片的内容和地点。根据我们与照片提供方的协议，我们不得公开这些说明文字。不过，我们在论文中使用了这些文字来分析数据集的地理分布。

7. *Are relationships between individual instances made explicit (例如, .. users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.* 不，数据集中实例之间没有已知的关系。

8. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.* **Errors:** 这些掩码由分割模型生成，因此掩码可能存在错误或不一致之处。**Redundancies:** 尽管没有两张图像完全相同，但存在同一主体在相近时间拍摄的图像实例。

9. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (例如, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (即, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (例如, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.* 该数据集是自包含的。

10. *Does the dataset contain data that might be considered confidential (例如, .. data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.* 否。

11. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.* 我们采取了两项安全措施来防止不良内容出现：(1) 照片均从图片供应商处获得授权，且必须符合该供应商的服务条款。我们要求供应商过滤掉所有授权图片中的不良内容。(2) 如果用户在数据集中发现不良图片，我们欢迎通过segment-anything@meta.com 邮箱举报该图片以便移除。尽管已采取这些措施，我们仍观察到少量图片包含抗议或其他集会场景，这些场景涉及可能引发冒犯的各类宗教信仰或政治观点。我们未能制定出可完全过滤此类图片的策略，因此依赖用户举报这类内容。

12. *Does the dataset identify any subpopulations (例如, .. by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.* 该数据集未识别照片中人物的任何亚群体。

13. *Is it possible to identify individuals (即, one or more natural persons), either directly or indirectly (即, in combination with other data) from the dataset? If so, please describe how.* 所有图像均经过人脸模糊处理模型，以移除任何个人身份信息。若用户发现任何匿名化问题，欢迎通过segment-anything@meta.com 报告问题及对应图像编号。

14. *Does the dataset contain data that might be considered sensitive in any way (例如, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.* 该数据集包含抗议活动或其他可能暗示宗教信仰、政治观点或工会成员身份的集会场景。然而，数据集中所有人的面部均已通过模糊处理进行匿名化，因此无法识别数据集中的任何个人。

15. *Any other comments?* 否。

收集过程

1. *How was the data associated with each instance acquired? Was the data directly observable (例如, raw text, movie ratings), reported by subjects (例如, .. survey responses), or indirectly inferred/derived from other data (例如, .. part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.* 每张图像对应的已发布掩码均由我们的分割模型SAM自动推断得出。通过模型辅助人工标注收集的掩码将不予发布。质量验证方法如§5所述。

2. *What mechanisms or procedures were used to collect the data (例如, .. hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?* 数据集中的图像已获得图片提供商的授权。它们均为摄影师使用不同相机拍摄的照片。

3. If the dataset is a sample from a larger set, what was the sampling strategy (例如, deterministic, probabilistic with specific sampling probabilities)? 我们为测试目的保留了~2k张随机选择的图像。其余已获授权的图像均包含在数据集中。4. Who was involved in the data collection process (例如, students, crowdworkers, contractors) and how were they compensated (例如, how much were crowdworkers paid)? 发布的掩码由SAM自动推断生成。关于模型辅助人工标注流程的详细信息, 请参阅§F.2章节的数据标注说明卡。请注意这些掩码将不会发布。5.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (例如, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. 已获授权的照片拍摄日期跨度很大, 最早可追溯至2022年。6.

Were any ethical review processes conducted (例如, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation. If the dataset does not relate to people, you may skip the remaining questions in this section. 我们进行了内部隐私审查, 以评估并确定如何降低照片中人物隐私的潜在风险。对面部和车牌进行模糊处理可保护照片中人物的隐私。7.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (例如, websites)? 我们从第三方图片供应商处获得了数据授权。8.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself. 图像授权自第三方供应商, 该供应商已就收集个人所需的通知和同意事项作出适当声明。此外, 所有可识别信息 (e.g.面部、车牌) 均已进行模糊处理。根据数据集许可条款, 禁止尝试识别图像中特定个体或建立关联。9.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented. 图像授权自第三方供应商, 该供应商已就收集个人所需的通知和同意事项作出适当声明。此外, 所有图像中的可识别信息 (e.g.面部、车牌) 均已进行模糊处理。为免疑义, 根据数据集许可条款, 禁止尝试识别图像中特定个体或建立关联。10.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate). 我们邀请用户通过segment-anything@meta.com邮箱联系, 申请移除相关图像。11.

Has an analysis of the potential impact of the dataset and its use on data subjects (例如, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation. 为消除对照片被收录者的潜在影响, 所有可识别信息 (面部、车牌) 均已进行模糊处理。

12. Any other comments? 不。

预处理 / 清洗 / 标注

1. Was any preprocessing / cleaning / labeling of the data done (例如, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section. 我们将高分辨率授权图像的短边调整为1500像素, 并仅对图像进行处理以移除照片中的任何可识别和个人信息 (如人脸、车牌)。2.

Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (例如, to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data. 不, 出于安全考虑和尊重隐私, 我们已删除相关数据, 因此不会发布未经修改的照片。

3. Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point. 我们使用了

RetinaFace [88, 89] 模型 (<https://github.com/serengil/retinaface>) 用于检测人脸。用于模糊车牌号的模型尚未公开。

用途

1. Has the dataset been used for any tasks already? If so, please provide a description. 该数据集被用于训练我们的分割模型SAM。

2. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point. 不。但所有数据集的使用者都必须引用它, 因此其使用情况可通过引文追踪工具进行跟踪。

3. What (other) tasks could the dataset be used for? 我们旨在将该数据集构建为大规模分割数据集。同时, 我们诚邀研究界为该数据集补充更多标注信息。

4. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (例如, stereotyping, quality of service issues) or other risks or harms (例如, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms? 我们在§6中对数据集的大致地理和收入水平覆盖范围进行了分析。虽然我们相信目前我们的数据集比大多数公开存在的数据集更具代表性, 但我们承认我们并未在所有群体中实现均衡覆盖, 并鼓励用户注意其模型使用此数据集可能学到的潜在偏见。5.

Are there tasks for which the dataset should not be used? If so, please provide a description. 包括禁止使用案例在内的数据集完整使用条款可在 <http://ai.facebook.com/datasets/segment-anything> 找到。

6. Any other comments? 否。

分布

1. Will the dataset be distributed to third parties outside of the entity (例如, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description. 该数据集将可供研究社区使用。

2. How will the dataset will be distributed (例如, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)? 数据集可在 <https://ai.facebook.com/datasets/segment-anything> 获取。
3. When will the dataset be distributed? 该数据集将于2023年发布。

4. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions. 是的。该数据集的许可协议和使用条款可在 <https://ai.facebook.com/datasets/segment-anything> 找到。用户必须在下载或使用数据集前同意使用条款。

5. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions. SA-1B 数据集的完整使用条款和使用限制可在 <https://ai.facebook.com/datasets/segment-anything> 查看。

6. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation. SA-1B 数据集的使用许可和限制可在 <https://ai.facebook.com/datasets/segment-anything> 查看。

7. Any other comments? 否。

维护

1. Who will be supporting/hosting/maintaining the dataset? 该数据集将托管于 <https://ai.facebook.com/datasets/segment-anything> 并由 Meta AI 维护。2. How can the owner/curator/manager of the dataset be contacted (例如, email address)? 请发送邮件至 segment-anything@meta.com。3.

Is there an erratum? If so, please provide a link or other access point. 否。4. Will the dataset be updated (例如, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (例如, mailing list,

GitHub)? 为了帮助使用SA-1B的研究实现可复现性, 唯一的更新将是移除已报告的图像。

5. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (例如, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced. 数据保留没有限制。我们已采取措施从所有人物图像中移除个人信息。用户可以在此处举报内容以便可能被移除: segment-anything@meta.com。

6. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers. 不, 由于唯一的更新将是移除潜在有害内容, 我们不会保留包含该内容的旧版本。

7. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description. 我们鼓励用户为SA-1B收集更多标注。任何生成标注的用户将负责托管和分发其标注数据。

8. Any other comments? 否。

F.2. 数据标注卡片

任务表述

1. At a high level, what are the subjective aspects of your task? 对图像中存在的物体进行分割本质上是一项主观任务。例如, 一位标注者可能将两只靴子分割为一个掩码, 而另一位标注者可能将每只靴子单独分割。根据标注者的技能水平, 掩码的质量和每张图像的掩码数量在不同标注者之间存在差异。尽管任务存在这些主观因素, 我们仍认为高效标注是可行的, 因为数据是以逐掩码的方式进行标注的, 主要关注数据的多样性而非完整性。

2. What assumptions do you make about annotators? 我们的标注员全职投入标注工作, 人员流失率极低。这使得我们能够定期培训标注员, 提供反馈并解答他们的问题。具体而言: (1) 通过明确阐述工作目标并提供清晰的指导(包括任务可视化说明和视频录像), 标注员获得了充分的情境理解, 从而能够合理认知并执行任务。(2) 共享目标与关键成果, 并与标注员保持每周例会, 这有效提升了标注员在长期工作中提高标注质量与数量的可能性。

3. How did you choose the specific wording of your task instructions? What steps, if any, were taken to verify the clarity of task instructions and wording for annotators? 由于我们的任务是标注图像, 标注指南中包含了视觉示例。我们的研究团队完成了30项标注任务, 以识别使用标注工具时可能遇到的明显挑战, 共同决定如何处理复杂案例, 并完善指南。研究团队每周与标注员举行反馈会议。团队执行任务的视频会实时与标注员共享, 随后进行问答环节。标注员可以在反馈会议期间及异步沟通中, 就不明确的方面提出反馈。

4. What, if any, risks did your task pose for annotators and were they informed of the risks prior to engagement with the task? 未识别出风险。图像在标注阶段前已过滤不良内容。

5. What are the precise instructions that were provided to annotators? 我们仅提供高级层指导: 给定一张图像, 我们的目标是分割出所有可能的物体。标注者需为每个可识别的潜在物体生成掩码。物体可通过我们的交互式分割工具进行分割, 具体方式包括: 使用修正性前景/背景点击来添加/移除掩码区域, 或在物体周围绘制边界框。掩码可通过像素级精确工具进行细化处理。

选择标注

1. Are there certain perspectives that should be privileged? If so, how did you seek these perspectives out? 我们选择与曾参与其他视觉标注任务的标注员合作。

2. Are there certain perspectives that would be harmful to include? If so, how did you screen these perspectives out? 否。

3. Were sociodemographic characteristics used to select annotators for your task? If so, please detail the process. 否。

4. If you have any aggregated socio-demographic statistics about your annotator pool, please describe. Do you have reason to believe that socio-demographic characteristics of annotators may have impacted how they annotated the data? Why or why not? 我们与130名标注员进行了合作。所有标注员均位于肯尼亚。我们认为标注员的社会人口特征并未对标注数据产生实质性影响。

5. Consider the intended context of use of the dataset and the individuals and communities that may be impacted by a model trained on this dataset. Are these communities represented in your annotator pool? Segment Anything 1B (SA-1B) 数据集仅限研究用途。如第6节所述, SA-1B数据集是地理多样性最丰富的分割数据集之一。此外, 我们在第6节中分析了基于该数据集训练的模型在负责任人工智能方面的考量。

平台与基础设施选择

1. What annotation platform did you utilize? At a high level, what considerations informed your decision to choose this platform? Did the chosen platform sufficiently meet the requirements you outlined for annotator pools? Are any aspects not covered? 我们使用了专有的标注平台。2.

What, if any, communication channels did your chosen platform offer to facilitate communication with annotators? How did this channel of communication influence the annotation process and/or resulting annotations? 我们每周手动审核标注结果, 并向标注员反馈意见。我们会沟通常见的错误或不一致之处及相应的修正方法。此外, 标注质量保障团队每日也会向标注员提供改进建议。除每周的反馈会议外, 标注员还可通过共享表格和聊天群组与研究团队保持沟通, 这一流程显著提升了标注的平均速度与质量。3.

How much were annotators compensated? Did you consider any particular pay standards, when determining their compensation? If so, please describe. 标注员的报酬按服务商设定的时薪结算。该服务商为共益企业(Certified B Corporation) 认证机构。

数据集分析与评估

1. How do you define the quality of annotations in your context, and how did you assess the quality in the dataset you constructed? 标注员首先接受培训。他们参加了由供应商主导的为期一天的培训课程, 随后被要求标注来自训练队列的大量样本。在供应商质量保证团队与研究团队协作下, 通过人工抽查标注员的掩码以确保质量后, 标注员方可从培训阶段晋级至生产阶段。平均而言, 标注员在晋级前需经历为期一周的培训。生产质量评估遵循类似流程: 供应商质量保证团队与研究团队每周人工审核标注结果, 并同步反馈意见。2.

Have you conducted any analysis on disagreement patterns? If so, what analyses did you use and what were the major findings? Did you analyze potential sources of disagreement? 我们在每周与标注员的会议中指出了常见错误。3.

How do the individual annotator responses relate to the final labels released in the dataset? 这些标注数据仅用于训练早期版本的SAM模型, 目前我们暂无公开这些数据的计划。

数据集发布与维护

1. Do you have reason to believe the annotations in this dataset may change over time? Do you plan to update your dataset? 不, 除非是为了移除令人反感的图像。2.

Are there any conditions or definitions that, if changed, could impact the utility of your dataset? 我们不这么认为。

3. Will you attempt to track, impose limitations on, or otherwise influence how your dataset is used? If so, how? SA-1B数据集将根据一份允许用于特定研究目的并为研究者提供保护的许可协议发布。研究人员必须同意该许可协议的条款方可访问数据集。

4. Were annotators informed about how the data is externalized? If changes to the dataset are made, will they be informed? 不, 我们目前不计划发布手动标注数据。

5. Is there a process by which annotators can later choose to withdraw their data from the dataset? If so, please detail. 否。

Model Overview

Name	SAM or Segment Anything Model
Version	1.0
Date	2023
Organization	The FAIR team of Meta AI
Mode type	Promptable segmentation model
Architecture	See §3
Repository	https://github.com/facebookresearch/segment-anything
Citation	https://research.facebook.com/publications/segment-anything
License	Apache 2.0

Intended Use

Primary intended uses	SAM is intended to be used for any prompt-based segmentation task. We explored its use in <i>segmenting objects from a point</i> (§7.1), <i>edge detection</i> (§7.2), <i>segmenting all objects</i> (§7.3), and <i>segmenting detected objects</i> (§7.4). We explored how SAM can integrate with other vision models to <i>segment objects from text</i> (§7.5).
Primary intended users	SAM was primarily developed for research. The license for SAM can be found at https://github.com/facebookresearch/segment-anything .
Out-of-scope use cases	See terms of use for SAM found at https://github.com/facebookresearch/segment-anything . See <i>Use Cases</i> under <i>Ethical Considerations</i> .
Caveats and recommendations	SAM has impressive zero-shot performance across a wide range of tasks. We note, however, that in the zero-shot setting there may be multiple valid ground truth masks for a given input. We recommend users take this into consideration when using SAM for zero-shot segmentation. SAM can miss fine structures and can hallucinate small disconnected components. See §8 for a discussion of limitations.

Relevant Factors

Groups	SAM was designed to segment any object. This includes <i>stuff</i> and <i>things</i> .
Instrumentation and environment	We benchmarked SAM on a diverse set of datasets and found that SAM can handle a variety of visual data including <i>simulations, paintings, underwater images, microscopy images, driving data, stereo images, fish-eye images</i> . See §D.1 and Table 7 for information on the benchmarks used.

Metrics

Model performance measures	<p>We evaluated SAM on a variety of metrics based on the downstream task in our experiments.</p> <ul style="list-style-type: none"> • <i>mIoU</i>: We used the mean intersection-over-union after a given number of prompts to evaluate the segmentation quality of a mask when prompted with points. • <i>Human evaluation</i>: We performed a human study (detailed in §E) to evaluate the real world performance of SAM. We compared the masks generated by SAM to a baseline state-of-the-art interactive segmentation model, RITM [92], using a perceptual quality scale from 1 to 10. • <i>AP</i>: We used average precision to evaluate instance segmentation for a given box and edge detection. • <i>AR@1000</i>: We used average recall to evaluate object proposal generation. • <i>ODS, OIS, AP, R50</i>: We used the standard edge detection evaluation metrics from BSDS500 [72, 3].
----------------------------	---

Evaluation Data

Data sources | See §D.1.

Training Data

Data source | See Data Card in §F.1.

Ethical Considerations

Data	We trained SAM on licensed images. The images were filtered for objectionable content by the provider, but we acknowledge the possibility of false negatives. We performed a geographic analysis of the SA-1B dataset in §6. While SA-1B is more geographically diverse than many of its predecessors, we acknowledge that some geographic regions and economic groups are underrepresented.
Cost and impact of compute	SAM was trained on 256 A100 GPUs for 68 hours. We acknowledge the environmental impact and cost of training large scale models. The environmental impact of training the released SAM model is approximately 6963 kWh resulting in an estimated 2.8 metric tons of carbon dioxide given the specific data center used, using the calculation described in [77] and the ML CO ₂ Impact calculator [61]. This is equivalent to ~7k miles driven by the average gasoline-powered passenger vehicle in the US [101]. We released the SAM models to both reduce the need for retraining and lower the barrier to entry for large scale vision research.
Risks and harms	We evaluated SAM for fairness in §6. Downstream use cases of SAM will create their own potential for biases and fairness concerns. As such we recommend users run their own fairness evaluation when using SAM for their specific use case.
Use cases	We implore users to use their best judgement for downstream use of the model.

表9：SAM模型卡片，遵循[75]中详述的流程。

We have several models that, when provided with a click or a box as input, output a mask. We would like to compare the quality of these models by rating the quality of their masks on many examples.

The user will be asked to rate the quality of the mask annotation.

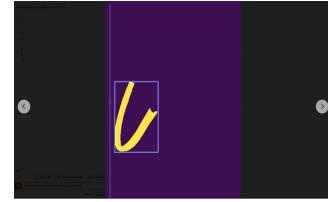
- Each job reviews one mask in one image.
- On the right, there will be five image thumbnails in two rows. Each thumbnail can be mouse-overed to show the image at a larger size. Clicking on the thumbnail will make it full screen, and clicking again will return to the original screen.
- The images show the mask in three different views. On the top row: (left) the image with both the object and the mask overlaid on the image, and (right) the mask alone. On the bottom row: (left) a zoomed in view of the object without a mask, and (right) a zoomed in view of the mask overlaid on the image. These views are provided to make it easier to see differences between them.
- The mask will be red when overlaid on the image.
- When shown by itself, the mask is yellow, and the background is purple.
- If the mask is yellow, then the user can click on the mask to draw a box. This is the input to the model, as if you had clicked at this location or drawn this box.
- On the left, there are buttons labeled 1-10. This is used to rate the quality of the shown mask.

Objective and Setup



Example interface page. There will be five images on the right and a question box on the left.

Mouse over an image to show the full image.



Click on an image to make it full screen. The arrows will cycle between images. Click again to return to previous view.

The first image on the top row shows the image without a mask. A blue point will be on the object of interest, or a blue and white box will surround it.

The second image on the top row shows the mask for the object in red.

The third image on the top row shows the mask only. The mask is in yellow and the background is purple.

The second image on the bottom row shows a zoomed in view of the object with a mask. The mask is in red.

On the left are buttons to rate the mask quality, with selections 1-10.

Task

Does the mask have a good boundary?

- Errors in the boundary can include:
 - Incorrect holes in the mask
 - Incorrect pixels included separated from the main part of the mask
 - Poor edge quality, where the mask does not exactly match the edge of the object
 - Including a nearby obscuring foreground object (a mask that covers obscuring objects is fine, and one that does not cover obscuring objects is fine, but one that does some of both has an error)
 - Providing a small mask is not an error, as long as the mask still matches the edges of the object.

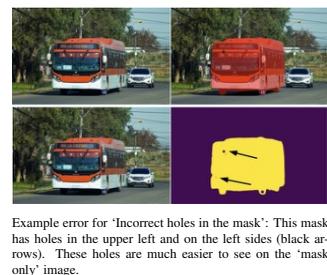
Judging Mask Quality (2 of 3)

Does the mask correspond to the provided point or box?

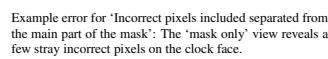
- For points:
 - The point needs to be on the mask.
 - The size or position of the object with respect to the point does not matter (a point on someone's gloved hand can correspond to the glove or to the entire person, both are valid masks).
- For boxes:
 - The object needs to be the best object that is the size of the box (a box around someone's entire head but the mask is of their hair, this is an error; their hair is in the box but is not the correct object).
 - If the box clearly corresponds to a given object but is slightly smaller than it, it is okay if the mask only partially overlaps a box (if a box around a person misses their extended hand, the mask can still include their hand even if the mask goes outside the box).

Judging Mask Quality (3 of 3)

Example error of 'Include an arbitrary part of a collection': In top image, the point is on one orange rind, but the mask covers two orange rinds. This is a mask error: the mask covers an arbitrary number of objects in the collection, and should either cover one orange rind or all of them. In the bottom image, the box is around both vegetables. Since this is the best match to the box, this is not a mask error.



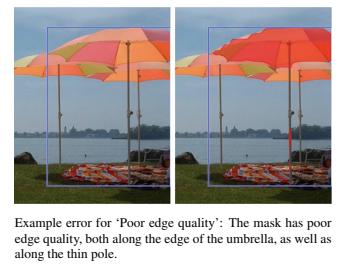
Example error for 'Include holes in the mask': This mask has holes in the upper left and on the left sides (black arrows). These holes are much easier to see on the 'mask only' image.



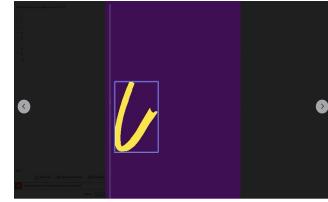
Example error for 'Incorrect pixels included separated from the main part of the mask': The 'mask only' view reveals a few stray incorrect pixels on the clock face.



Example error of 'Missing a part of an object': The mask is missing a disconnected part of the object: the back half of the zebra, and the right portion of the plate.



Example error for 'Poor edge quality': The mask has poor edge quality, both along the edge of the umbrella, as well as along the thin pole.



Click on an image to make it full screen. The arrows will cycle between images. Click again to return to previous view.

The first image on the bottom row shows a zoomed in view of the object without a mask.

Does the mask correspond to an actual object?

- Valid objects can include:
 - Entire single objects (such as a person, shirt, or tree)
 - Logical parts of objects (a chair leg, a car door, a tabletop)
 - Collection of objects (a pile of books, a crowd of people)
 - Stuff (the ground, the sky)
- Example errors a mask may have. The severity of these errors may be minor or major:
 - Include a piece of another object (the mask of a person including the arm of a nearby person)
 - Miss a part of an object (the mask covers only one part of a building obscured by a tree in the foreground)
 - Combine two unrelated things (a single mask covers both a dog and a pen on a table)
 - Include an arbitrary part of a collection (a pen input to a model is a collection of one object, but the collection applies to a pile of many pens). If a box surrounds an arbitrary collection, it is not an error to provide a mask for these objects.
- If you are unsure, a good rule-of-thumb is: can you name the object in question? However, some things that are hard to name may still be good objects (an unusual component of a machine, something at the edge of the image for which it is hard to determine what it is).

Judging Mask Quality (1 of 3)

图19：此处我们提供了用于人工审核掩码质量的完整标注指南。部分图像经过轻微编辑，面部已做模糊处理以便发布。建议缩放查看（共2部分，第1部分）。

G. 标注指南

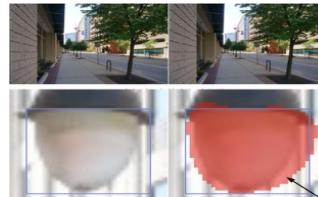
我们在图19和图20中提供了用于人工审查掩码质量的完整标注指南。



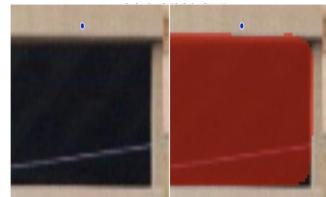
“结合两个不相关事物”的示例：点表示蜥蜴，但面具同时覆盖了蜥蜴和鸟。这是一个面具错误。



“未能一致处理遮挡前景物体”的示例错误：右侧的柱子（蓝色箭头）被排除在掩码之外，而左侧的柱子（黑色箭头）却被包含在物体中。掩码应当同时包含或同时排除这两者。



“小掩码像素化”示例：该掩码的边界不完美，因为它在黑色箭头处超出了物体。然而，掩码的“块状”图案并非错误，因为在如此放大时，图像本身也以相同方式呈现块状效果。



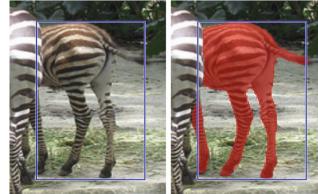
与所提供点一致性的示例错误：掩码与蓝点不符，因此这是掩码错误。



与所提供点一致性的示例：对于此输入点，但徽标（左侧）和容器（右侧）均为有效对象，因为蓝色点仅同时位于两者之上。两个掩码均无掩码错误。



与方框一致性的示例：方框包围了橙子碗，但掩码仅针对单个橙子。这是一个掩码错误。

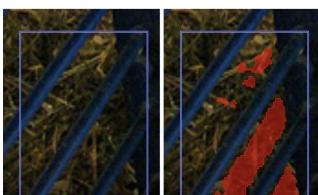


与方框一致性的示例：方框的形状与斑马匹配。即使掩码略微延伸到方框外部以包含斑马的左腿，这也不是错误。

Overall mask quality is subjective, each of the above errors may hurt mask quality only a little or a lot, depending on how serious the error is. Please use your best judgment when choosing mask scores, and try to stay consistent from mask-to-mask. Here are some general guidelines for what different scores should correspond to:

- A score of 1: It is not possible to tell what object this mask corresponds to. This includes the case that there is no mask at all.
- A low score (2-4): The object is mostly identifiable, but the mask quality is extremely poor (e.g. large regions of the mask cover other objects; large regions of the object missing; extremely spiky mask boundaries that cut through the middle of the object).
- A mid score (5-6): The object is identifiable and the boundary is mostly correct, but there are one or two significant unaligned part of the object; object is not a significant part of another object; very poor boundary quality in one area of the object but not the entire object.
- A high score (7-9): The object is identifiable and errors are small and rare (missing a small, less highly obscured disconnected component; having small regions where the mask boundary does not quite match the object boundary).
- A score of 10: The mask is pixel-perfect; it has no identifiable errors at all.

掩码评分



得分为1的掩码示例：不清楚这个掩码对应什么物体。



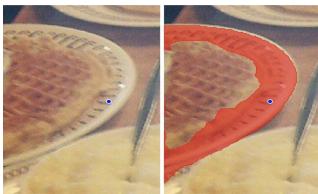
低分（2-4）掩码示例：主体可识别，但掩码包含了另一物体的大面积错误部分。



低分（2-4分）掩码示例：主体可识别，但物体有大量随机部分缺失。



低至中等分数（4-5）的掩码示例：物体可识别且边缘全部正确，但掩码错误地包含了左侧人物的手。



中等分数（5-6）的掩码示例：掩码与板片明显对应，但与华夫格的边界相当模糊。



中等分数（5-6）的掩码示例：物体易于识别，且大部分边缘合理。然而，存在一个显著的不连通部分（框架内的手臂）几乎完全缺失，同时该区域存在斑驳的像素。



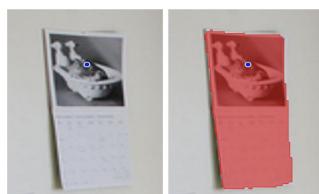
中等至高评分（6-8）的掩码示例：该掩码存在两处较小的边界缺陷区域，分别位于掩码顶部和右下角。



中等至高分数（6-8）的掩码示例：花环是一个有效对象，其尺寸与盒子相符（整个花环+钟表也将是一个有效对象）。然而，钟表上存在错误的杂散掩码像素。



高分（7-9）掩码示例：马的轮廓几乎完全正确，除了其后腿右侧。掩码始终包含马佩戴的所有装备，且边界划分合理。



一个得分非常高的掩码示例（~9）：仅在掩码边缘存微小错误。块状的“像素化”并非错误，因为在此尺度下图像本身也是块状的。



得分非常高（9-10）的掩码示例：掩码仅在右下边缘有非常细微的错误。



得分非常高9(10)的掩码示例：掩码边缘仅有微小瑕疵。

图20：此处我们提供了用于人工审核掩码质量的完整标注指南。部分图片已稍作编辑，面部已进行模糊处理以便发布。建议缩放查看（共2部分，第2部分）。