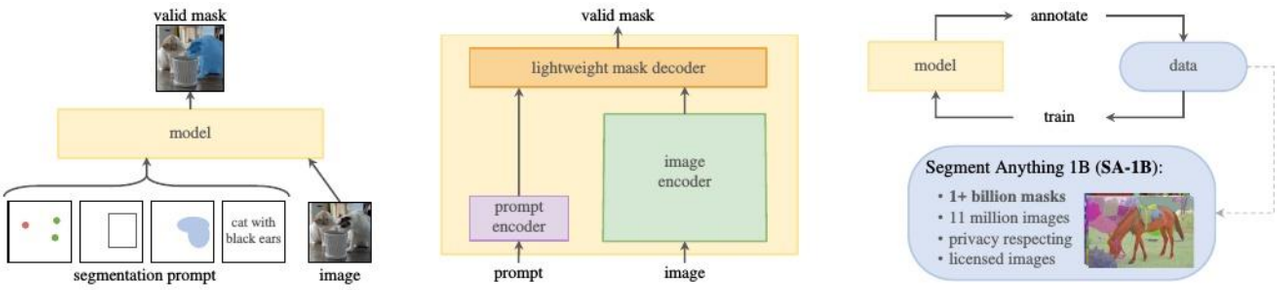
任意片段

亚历山大基里洛夫 [1,2,4] [null]Eric Mintun[2] [null]Nikhila 拉维 [1,2] [null]汉字毛 [2] [null] 克洛艾罗兰 [3] [null] 劳拉 Gustafson[3] [null] 肖太 [3]
[null]Spencer Whitehead 亚历山大 C. Berg Lo Wan-Yen [4] [null] 罗斯 Girshick[4]

[1null] 项目领导 [2null] 联合第一作者 [3null] 平等的贡献 [4 null] 定向铅

Meta AI Research， FAIR



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

图 1：我们的目标是通过引入三个相互关联的组件来构建分割的基础模型：一个可提示的分割任务，一个分割模型（SAM），它为数据注释提供动力，并通过提示工程实现对一系列任务的零采样传输，以及一个用于收集 SA-1B 的数据引擎，我们的数据集超过 10 亿个掩码。

摘要

*我们介绍任意片段（SA）项目：一个新的用于图像分割的任务、模型和数据集。在数据收集循环中使用我们的高效模型，我们建立了迄今为止（到目前为止）最大的分割数据集，在 11M 许可和尊重隐私的图像上拥有超过 10 亿个掩模。该模型经过设计和训练，具有提示性，因此它可以将零拍摄转移到新的图像分布和任务。我们评估了它在许多任务上的能力，发现它的零射击性能令人印象深刻——通常与之前的完全监督结果相竞争，甚至优于前者。我们在 seg-anything.com 上发布了任意片段模型（SAM）和对应的 1B 掩模和 11M 图像的数据集（SA-1B），以促进对计算机视觉基础模型的研究。***我们建议在以下网址阅读全文：arxiv.org/abs/2304.02643。**

1.介绍

在 web 规模的数据集上预训练的大型语言模型正在以强大的零射击和少射击泛化[10]彻底改变 NLP。这些“基础模型”[8]可以泛化到训练期间所看到的任务和数据分布之外。这种能力通常通过提示工程来实现，其中使用手工制作的文本来提示语言模型为手头的任务生成有效的文本响应。当使用来自网络的大量文本语料库进行缩放和训练时，这些模型的零和少镜头性能与（even）相比惊人地好 null 匹配（在某些情况下）微调模型[10,20]。经验趋势表明，这种行为随着模型规模、数据集大小和总训练计算而改善[54,10,20,49]。

基础模型也在计算机视觉中进行了探索，尽管程度较低。也许最突出的例子是对齐来自网络的配对文本和图像。例如，CLIP[80]和 ALIGN b[53]使用对比学习来训练对齐两种模式的文本和图像编码器。经过训练后，工程文本提示可以实现对新颖视觉概念和数据分布的零射击泛化。这样的编码器还可以与其他模块有效地组合以实现下游任务，例如图像生成（例如 DALL • E[81]）。虽然在视觉和语言编码器方面已经取得了很大的进展，但计算机视觉包括超出此范围的广泛问题，并且对于其中的许多问题，并不存在丰富的训练数据。

在这项工作中，我们的目标是建立一个用于图像分割的基础模型。也就是说，我们寻求开发一个可提示的模型，并使用能够实现强大泛化的任务在广泛的数据集上对其进行预训练。有了这个模型，我们的目标是使用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，我们解决了以下关于图像分割的问题：

1.什么任务可以实现零射击泛化？

2.对应的模型架构是什么？

3.哪些数据可以为这项任务和模型提供动力？

这些问题纠缠在一起，需要综合解决。我们首先定义一个提示的分割任务，它足够通用，可以提供强大的预训练目标，并实现广泛的下游应用。这个任务需要一个支持灵活提示的模型，并且可以在提示时实时输出分段掩码，以允许交互使用。为了训练我们的模型，我们需要一个多样化的、大规模的数据源。不幸的是，目前还没有用于分割的 web 级数据源；为了解决这个问题，我们构建了一个“数据引擎”，即我们在使用我们的高效模型来协助数据收集和使用新收集的数据来改进模型之间进行迭代。接下来，我们介绍每个相互关联的组件，然后是我们创建的数据集和证明我们方法有效性的实验。

任务(§2)。在 NLP 和最近的计算机视觉中，基础模型是一个很有前途的发展，它可以通过使用“提示”技术对新的数据集和任务执行零射击和少射击学习。受这条工作线的启发，我们提出了提示分割任务，其目标是在给定任何分割提示的情况下返回一个有效的分割掩码（见图 1a）。提示符只是指定图像中要分割的内容，例如，提示符可以包括识别对象的空间或文本信息。对有效输出掩码的要求意味着，即使提示是模糊的，并且可以引用多个对象（例如，衬衫上的一个点可能表示衬衫或穿着它的人），输出也应该是这些对象中至少一个对象的合理掩码。我们使用提示分割任务作为预训练目标，并通过提示工程来解决一般的下游分割任务。

模型(§3)。提示分割任务和现实世界使用的目标对模型架构施加了约束。特别是，模型必须支持灵活的提示，需要在平摊实时中计算掩码以允许交互使用，并且必须具有歧义意识。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，然后将两个信息源组合在一个轻量级的掩码解码器中，该解码器预测分割掩码。我们将此模型称为分段任意模型（Segment Anything model，简称 SAM）（见图 1b）。通过将 SAM 分为图像编码器和快速提示编码器/掩码解码器，可以使用不同的提示重复使用相同的图像嵌入（并平摊其成本）。给定一个图像嵌入，提示编码器和掩码解码器在 web 浏览器中从一个提示中预测出一个掩码，时间为 50ms。我们专注于点、框和掩码提示，并通过自由格式的文本提示呈现初始结果。为了使 SAM 能够识别歧义，我们将其设计为预测单个提示的多个掩码，从而允许 SAM 自然地处理歧义，例如衬衫与人的例子。

null **数据引擎（§4）。**为了实现对新数据分布的强泛化，我们发现有必要在一个大而多样的掩码集上训练 SAM，而不是已经存在的任何分割数据集。虽然基础模型的典型方法是在线获取数据[80]，但掩码并不是天然丰富的，因此我们需要一种替代策略。我们的解决方案是构建一个“数据引擎”，即我们与模型在环数据集注释共同开发我们的模型（见图 1c）。我们的数据引擎有三个阶段：辅助手动、半自动和全自动。在第一阶段，SAM 帮助注释者注释掩码，类似于经典的交互式分段设置。在第二阶段，SAM 可以通过提示可能的对象位置来自动为对象子集生成掩码，而注释器则专注于注释剩余的对象，从而帮助增加掩码的多样性。在最后阶段，我们用前景点的规则网格提示 SAM，平均每张图像产生 100 个高质量的蒙版。

数据集(§5)。我们的最终数据集 SA-1B 包括来自 11M 许可和隐私保护图像的超过 1B 个掩码（见图 2）。SA-1B 是使用我们的数据引擎的最后阶段完全自动收集的，比任何现有的分割数据集都多 400 个掩码[64,43,115,58]，并且经过我们的广泛验证，掩码具有高质量和多样性。除了用于训练 SAM 的鲁棒性和通用性之外，我们希望 SA-1B 成为旨在建立新基础模型的研究的宝贵资源。

实验(§6)。我们广泛评估 SAM。首先，使用 23 个不同的新分割数据集，我们发现 SAM 从单个前景点产生高质量的掩模，通常仅略低于手动注释的地面真值。其次，我们在使用提示工程的零射击传输协议下的各种下游任务上发现了一致的强定量和定性结果，包括边缘检测、对象提议生成、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 可以使用开箱即用的快速工程来解决 SAM 训练数据之外涉及物体和图像分布的各种任务。然而，正如我们在§7 中讨论的那样，改进的空间仍然存在。

负责任的人工智能。我们在补充中提供了模型/数据集卡，并报告了在使用 SA-1B 和 SAM 时可能存在的公平性问题和偏差。SA-1B 中的图像跨越了地理和经济上不同的区域，我们发现 SAM 在不同人群中的表现相似。总之，我们希望这将使我们的工作对现实世界的用例更加公平。

释放。我们正在发布用于研究目的的 SA-1B 数据集，并在一个宽松的开放许可（Apache 2.0）下在 https://segment-anything.com 上提供 SAM。我们还通过在线演示展示了 SAM 的功能。

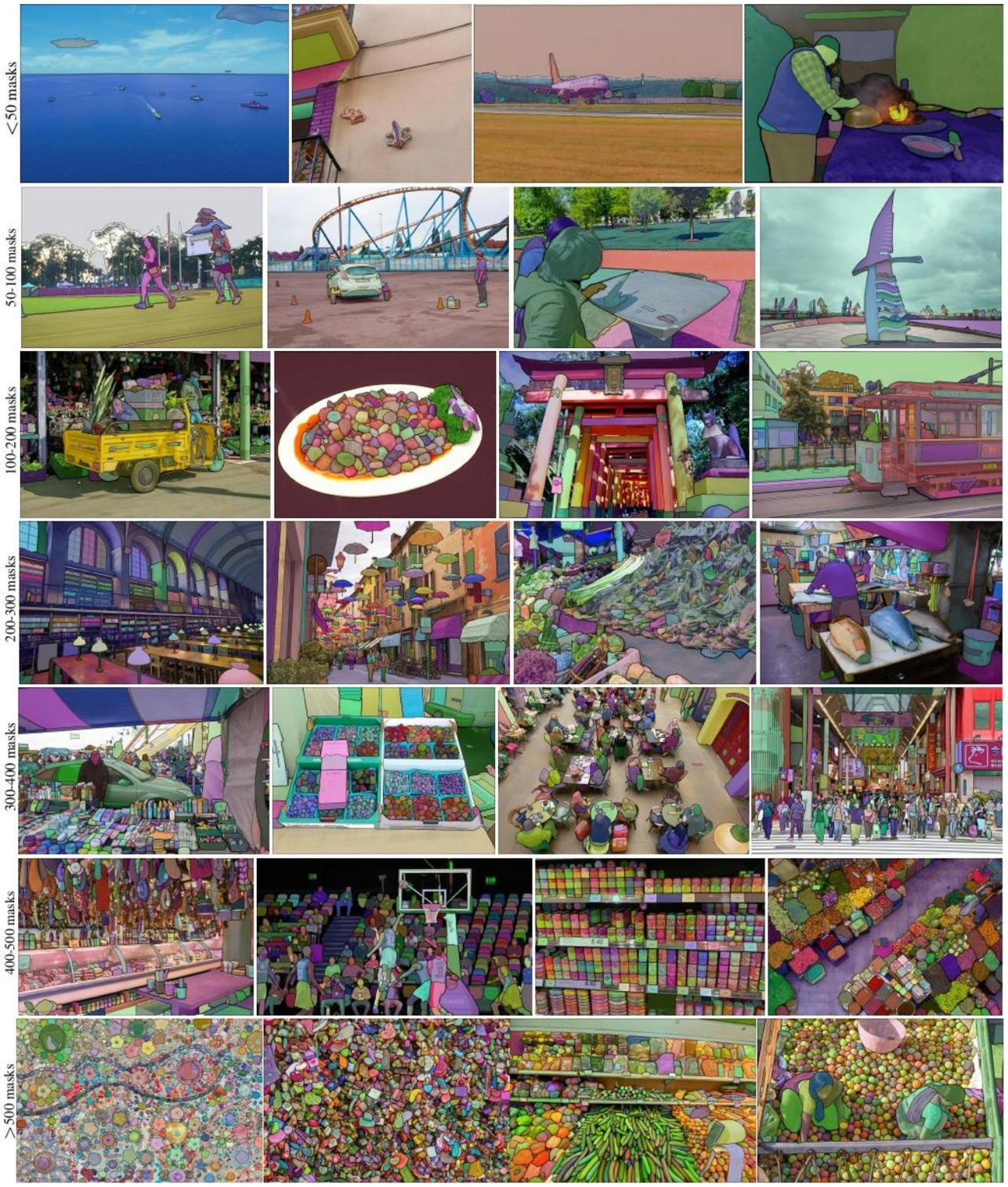


图 2：来自我们新引入的数据集 SA-1B 的带有叠加掩码的示例图像。SA-1B 包含 11M 不同的、高分辨率的、许可的和隐私保护的图像和 1.1B 高质量的分割蒙版。这些面具是由 SAM 完全自动标注的，正如我们通过人工评分和大量实验验证的那样，它们具有高质量和多样性。我们根据每张图像的掩码数量对图像进行分组，以实现可视化（每张图像平均有 100 个掩码）。

2.分割任何任务

我们从 NLP 中获得灵感，其中下一个令牌预测任务用于基础模型预训练，并通过提示工程[10]解决各种下游任务。为了构建分割的基础模型，我们的目标是定义一个具有类似能力的任务。

的任务。我们首先将提示的概念从 NLP 转化为分割，其中提示可以是一组前景/背景点，一个粗略的框或掩码，自由格式的文本，或者一般情况下，任何指示图像中分割内容的信息。那么，提示式分割任务就是在给定任何提示的情况下返回一个有效的分割掩码。“有效”掩码的要求仅仅意味着，即使提示是模糊的，并且可以引用多个对象（例如，回想一下衬衫 vs 人的例子，见图 3），输出应该是这些对象中至少一个的合理掩码。这一要求类似于期望语言模型对歧义提示输出连贯的响应。我们选择这个任务，是因为它引出了一种自然的预训练算法，以及一种通过提示将零射击转移到下游分割任务的通用方法。

训练。提示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（例如，点、框、掩码），并将模型的掩码预测与地面事实进行比较。我们从交互式分割中采用了这种方法[107,68]，尽管与交互式分割不同，交互式分割的目的是在足够的用户输入后最终预测一个有效的掩码，我们的目标是始终预测任何提示的有效掩码，即使提示是模糊的。这确保了预训练的模型在涉及歧义的用例中是有效的，包括我们的数据引擎 §4 所要求的自动注释。我们注意到，在这项任务中表现良好是具有挑战性的，需要专门的建模和训练损失选择，我们在 §3 中讨论过。

Zero-shot 转移。直观地说，我们的预训练任务赋予了模型在推理时对任何提示做出适当响应的能力，因此下游任务可以通过工程适当提示来解决。例如，如果有一个针对猫的边界盒检测器，则可以通过将检测器的盒输出作为提示提供给我们的模型来解决猫实例分割问题。一般来说，各种各样的实际分割任务都可以作为提示。除了自动数据集标记之外，我们在 §6 的实验中探索了五个不同的示例任务。

相关的任务。分割是一个很广阔的领域：有交互式分割[55,107]、边缘检测[3]、超像素化[83]、目标提议生成[2]、前景分割[92]、语义分割[88]、实例分割[64]、全视分割[57]等。我们的提示式分割任务的目标是产生

null

图 3：每列显示 SAM 从一个不明确的点提示（绿色圆圈）生成的 3 个有效掩码。

一个功能广泛的模型，可以通过提示工程适应许多（尽管不是全部）现有的和新的分割任务。这种能力是任务泛化[25]的一种形式。请注意，这与之前在多任务分割系统上的工作不同。在多任务系统中，单个模型执行一组固定的任务，例如联合语义、实例和全视分割[112,18,52]，但训练和测试任务是相同的。我们工作中的一个重要区别是，为提示分割训练的模型可以在推理时通过充当更大系统中的组件来执行新的、不同的任务，例如，为了执行实例分割，将提示分割模型与现有的对象检测器相结合。

讨论。提示和组合是强大的工具，使单个模型能够以可扩展的方式使用，潜在地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，例如，CLIP[80]是 DALL • E[81]图像生成系统的文本-图像对齐组件。我们预计，由提示工程等技术驱动的可组合系统设计，将比专门为固定任务集训练的系统实现更广泛的应用。通过组合的镜头来比较提示式和交互式分割也很有趣：虽然交互式分割模型是在设计时考虑到人类用户的，但正如我们将演示的那样，为提示式分割训练的模型也可以组成一个更大的算法系统。

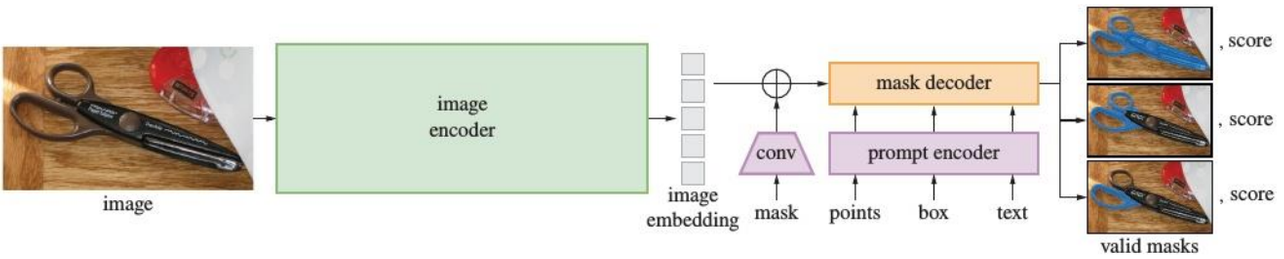


图 4：分段任意模型（SAM）概述。重量级图像编码器输出图像嵌入，然后可以通过各种输入提示高效查询，以平摊实时速度生成对象掩码。对于对应于多个对象的模糊提示，SAM 可以输出多个有效掩码和相关的置信度分数。

3.分段任意模型

接下来，我们描述了用于即时分割的分段任意模型（SAM）。SAM 有三个组件，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掩码解码器。我们建立在 Transformer 视觉模型[13,32,19,60]的基础上，对（平摊）实时性能进行了特定的权衡。我们在这里高层次地描述了这些组件，详细信息见 § B。

图像编码器。在可扩展性和强大的预训练方法的激励下，我们使用了一个 MAE[46]预训练视觉变压器（ViT）[32]，最低限度地适应处理高分辨率输入[60]。图像编码器每幅图像运行一次，可以在提示模型之前应用。

提示编码器。我们考虑两组提示：稀疏（点、框、文本）和密集（掩码）。我们通过位置编码来表示点和框[93]，并对每个提示类型和自由格式文本的学习嵌入求和，使用 CLIP[80]的现成文本编码器。密集提示（即掩码）使用卷积进行嵌入，并在图像嵌入中对元素进行求和。

面具解码器。掩码解码器有效地将图像嵌入、提示嵌入和输出令牌映射到掩码。这种设计受到[13,19]的启发，采用了对 Transformer 解码器块[101]的修改，然后是动态掩码预测头。我们修改后的解码器块在两个方向上使用提示自注意和交叉注意（提示到图像嵌入，反之亦然）来更新所有嵌入。在运行两个块之后，我们对图像嵌入进行上采样，MLP 将输出标记映射到动态线性分类器，然后该分类器计算每个图像位置的掩码前景概率。

解决歧义。有了一个输出，如果给出一个模糊的提示，该模型将对多个有效掩码进行平均。为了解决这个问题，我们修改了模型来预测单个提示符的多个输出掩码（见图 3）。我们发现 3 个掩码输出足以解决大多数常见情况（嵌套掩码通常最多有三个深度：整体、部分和子部分）。在训练期间，我们只对最小值进行反向支撑 nullLoss [14,44,62] over mask。为了对口罩进行排名，该模型预测每个口罩的置信度分数（即估计的 IoU）。

效率。整体模型设计很大程度上是由效率驱动的。给定预先计算的图像嵌入，提示编码器和掩码解码器在 web 浏览器中运行，在 CPU 上，大约 50ms。这种运行时性能使我们的模型能够无缝、实时地进行交互式提示。

损失和训练。我们用[13]中使用的焦点损失[63]和骰子损失[71]的线性组合来监督掩模预测。我们使用几何提示的混合来训练可提示的分割任务（文本提示见 § 6.2）。接下来[90,36]，我们通过每个掩码 11 轮随机抽样提示来模拟一个交互式设置，允许 SAM 无缝集成到我们的数据引擎中。

4.细分任何数据引擎

由于分割掩码在互联网上并不丰富，我们构建了一个数据引擎来收集我们的 1.1B 掩码数据集 SA-1B。数据引擎有三个阶段：(1)模型辅助的手动注释阶段，(2)混合了自动预测掩码和模型辅助注释的半自动阶段，以及(3)我们的模型在没有注释器输入的情况下生成掩码的全自动阶段。接下来我们将详细介绍每一个阶段。

Assisted-manual 阶段。在第一阶段，类似于经典的交互式分割，一组专业的注释者使用基于浏览器的交互式分割工具，通过点击前景/背景对象点来标记蒙版。可以使用像素精确的“画笔”和“橡皮擦”工具对蒙版进行细化。我们的模型辅助注释直接在浏览器中实时运行（使用预先计算的图像嵌入），从而实现真正的交互式体验。我们没有对标记对象施加语义约束，注释者可以自由地标记“东西”和“东西”[1]。我们建议注释者标记他们可以命名或描述的对象，但没有收集这些名称或描述。注释者被要求按显著性顺序标记对象，并被鼓励在一个蒙版花费超过 30 秒的时间进行注释后继续进行下一个图像。

在此阶段开始时，使用公共分割数据集训练 SAM。在充分标注数据后，仅使用新标注的掩码重新训练 SAM。随着收集到的蒙版越来越多，图像编码器从 vitb 缩放到 vith，其他架构细节也随之演变；我们总共重新训练了我们的模型 6 次。随着模型的改进，每个掩模的平均标注时间从 34 秒减少到 14 秒。我们注意到，14 秒比 COCO 的蒙版标注快 6.5[64]，只比带极值点的边界框标注慢 2[714,69]。随着 SAM 的改进，每张图像的平均掩模数从 20 个增加到 44 个。总体而言，我们在这一阶段从 120k 张图像中收集了 430 万个掩模。

半自动的阶段。在这个阶段，我们的目标是增加遮罩的多样性，以提高我们的模型分割任何东西的能力。为了将注释器集中在不太突出的对象上，我们首先自动检测自信蒙版。然后，我们向注释者展示了预先填充了这些蒙版的图像，并要求他们注释任何额外的未注释对象。为了检测自信蒙版，我们使用通用的“对象”类别在所有第一阶段蒙版上训练了一个边界框检测器[82]。在这个阶段，我们在 180k 张图像中额外收集了 5.9 万个遮罩（总共 1020 万个遮罩）。与第一阶段一样，我们定期在新收集的数据上重新训练我们的模型（5 次）。每个掩码的平均标注时间恢复到 34 秒（不包括自动掩码），因为这些对象的标记更具挑战性。每张图像的平均蒙版数量从 44 个增加到 72 个（包括自动蒙版）。

全自动舞台。在最后阶段，标注是全自动的。这是可行的，因为我们的模型有两个主要的增强。首先，在这个阶段开始的时候，我们已经收集了足够的遮罩来极大地改进模型，包括上一阶段的各种遮罩。其次，到这个阶段，我们已经开发出了模糊感知模型，即使在模糊的情况下，我们也可以预测有效的蒙版。具体来说，我们用 32 × 32 规则网格的点提示模型，并为每个点预测一组可能对应于有效对象的蒙版。使用模糊感知模型，如果一个点位于部分或子部分上，我们的模型将返回子部分、部分和整个对象。我们的模型的 IoU 预测模块用于选择置信掩模；此外，我们只识别和选择稳定的掩码（如果阈值在 0.5 和 0.5 +的概率图上得到相似的掩码，我们认为掩码是稳定的）。最后，在选择了自信和稳定的掩码后，我们应用非最大抑制（NMS）来过滤重复。为了进一步提高较小蒙版的质量，我们还处理了多个重叠的放大图像裁剪。关于这一阶段的进一步细节，请参见§C。我们对数据集中的所有 11M 图像应用了全自动掩码生成，总共产生了 11 亿个高质量的掩码。接下来，我们描述和分析结果数据集 SA-1B。

null

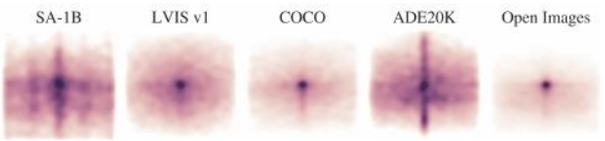


图 5：图像大小的归一化掩码中心分布。

5.分段任意数据集

我们的数据集 SA-1B 由 11M 不同的、高分辨率的、许可的、保护隐私的图像和 11 亿个高质量的分割蒙版组成，这些图像是由我们的数据引擎收集的。我们将 SA-1B 与现有数据集进行比较，并分析掩码质量和属性。我们正在发布 SA-1B，以帮助计算机视觉基础模型的未来发展。我们注意到，SA-1B 将在有利的许可协议下发布，用于某些研究用途，并保护研究人员。

图像。我们从一个直接与摄影师合作的供应商那里获得了一组新的 11M 图像。这些图像是高分辨率的（平均 3300 × 4950 像素），由此产生的数据大小可能会带来可访问性和存储方面的挑战。因此，我们发布了下采样图像，其最短边设置为 1500 像素。即使在降采样之后，我们的图像的分辨率也比许多现有的视觉数据集高得多（例如，COCO[64]图像是£480 × 640 像素）。请注意，今天大多数模型都是在低得多的分辨率输入上运行的。在发布的图像中，人脸和车牌已经被模糊化。

面具。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。我们直接将它们与专业注释进行比较，并查看各种掩码属性如何与突出的分割数据集进行比较。我们的主要结论，正如下面的分析和§6 中的实验所证实的那样，我们的自动蒙版对于训练模型来说是高质量和有效的。基于这些发现，SA-1B 只包括自动生成的掩模。

面具的质量。为了估计蒙版质量，我们随机抽取了 500 张图片（£50k 蒙版），并要求我们的专业注释者提高这些图片中所有蒙版的质量。注释者使用了我们的模型和像素精确的“画笔”和“橡皮擦”编辑工具来做到这一点。这个过程产生了一对自动预测和专业校正的蒙版。我们计算了每对对之间的 IoU，发现 94%的对 IoU 大于 90%（97%的对 IoU 大于 75%）。相比之下，先前的研究估计注释者之间的一致性为 85-91% IoU[43,58]。我们在§6 中的实验通过人类评分证实，相对于各种数据集，掩码质量很高，并且在自动掩码上训练我们的模型几乎与使用数据引擎生成的所有掩码一样好。

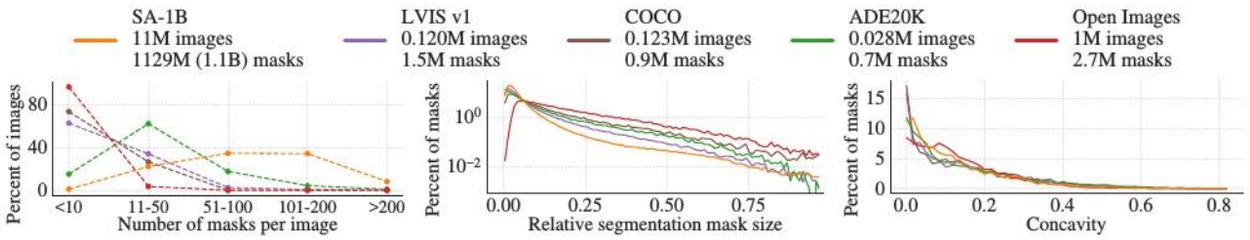


图 6：数据集掩码属性。图例引用了每个数据集中的图像和掩码的数量。注意，SA-1B 比现有最大的分割数据集 Open 图像[58]有 11 个、400 个、更多的图像和遮罩。

面具的属性。在图 5 中，我们绘制了与现有最大的分割数据集相比，SA-1B 中目标中心的空间分布。所有数据集中都存在常见的拍照者偏差。我们观察到，与 LVIS v1[43]和 ADE20K[115]这两个分布最相似的数据集相比，SA-1B 具有更大的图像角点覆盖率，而 COCO[64]和 Open Images V5[58]具有更突出的中心偏差。在图 6（图例）中，我们按大小对这些数据集进行了比较。SA-1B 比第二大的“Open Images”多 11 个图像和 400 个蒙版。平均来说，它比 Open Images 每张图像多 36 个遮罩。在这方面最接近的数据集，ADE20K，每张图像的掩模仍然少 3.5 个。图 6（左）绘制了每张图像的掩码分布。接下来，我们在图 6（中）中查看图像相对掩码大小（掩码面积除以图像面积的平方根）。正如预期的那样，由于我们的数据集在每张图像上有更多的掩码，因此它也倾向于包含更大比例的中小型相对大小的掩码。最后，为了分析形状复杂性，我们在图 6（右）中查看掩模的凹凸度（1 减去掩模面积除以掩模的凸壳面积）。由于形状复杂性与掩模尺寸相关，我们首先通过从分类掩模尺寸中进行分层抽样来控制数据集的掩模尺寸分布。我们观察到，我们的掩模的凹凸度分布与其他数据集大致相似。

6.零射转移实验

在本节中，我们提出了零射击转移实验与 SAM，片段任何模型。我们考虑了五个任务，其中四个与用于训练 SAM 的提示分割任务有很大不同。这些实验在训练过程中没有看到的数据集和任务上评估 SAM（我们使用的“零射击转移”遵循了 CLIP[80]中的用法）。数据集可能包括新的图像分布，如水下或以自我为中心的图像，据我们所知，没有出现在 SA-1B 中。

我们的实验从测试提示分割的核心目标开始：从任何提示产生有效的掩码。我们强调单个前景点提示的挑战性场景，因为它比其他更具体的提示更有可能是模糊的。接下来，我们提出了一系列遍历低、中、高-的实验 null 层次的图像理解，大致平行于该领域的历史发展。具体来说，我们提示 SAM(1)执行边缘检测，(2)分割所有内容，即对象建议生成，(3)分割检测到的对象，即实例分割，以及(4)作为概念验证，从自由格式文本中分割对象。这四个任务与 SAM 训练的提示分割任务有很大的不同，并且是通过提示工程实现的。我们报告了零射击单点有效掩码评估和零射击文本，以掩盖主要文本中的概念验证。我们建议读者参考我们在零镜头边缘检测、对象提议和实例分割方面的实验补充。此外，我们在补编中报告了一组消融。我们根据训练数据的大小和组成以及图像编码器架构来分析 SAM 的性能。

实现。除非另有说明：(1)SAM 使用 MAE[46]预训练的 viti - h[32]图像编码器；(2)SAM 在 SA-1B 上训练，注意该数据集仅包括从我们的数据引擎的最后阶段自动生成的掩码。对于所有其他模型和训练细节，如超参数，请参阅 § B。

6.1.零射单点有效掩模评估

的任务。我们评估从单个前景点分割一个对象。这个任务是病态的，因为一个点可以引用多个对象。大多数数据集中的 Ground truth mask 不会枚举所有可能的 mask，这可能会使自动度量不可靠。因此，我们用一项人类研究来补充标准的 mIoU 度量（即预测和地面真实掩码之间所有 IoUs 的平均值），其中注释者将掩码质量从 1（无意义）到 10（像素完美）进行评分。看到 § E。1、§ F 和 § H 了解更多细节。

默认情况下，我们从地面真实掩模的“中心”采样点（在掩模内部距离变换的最大值处），遵循交互式分割中的标准评估协议[90]。由于 SAM 能够预测多个掩码，我们默认只评估模型最自信的掩码。基线都是单掩码方法。我们主要与 RITM[90]进行比较，RITM 是一种强大的交互式分割器，与其他强大的基线相比，它在我们的基准上表现最好[65,17]。

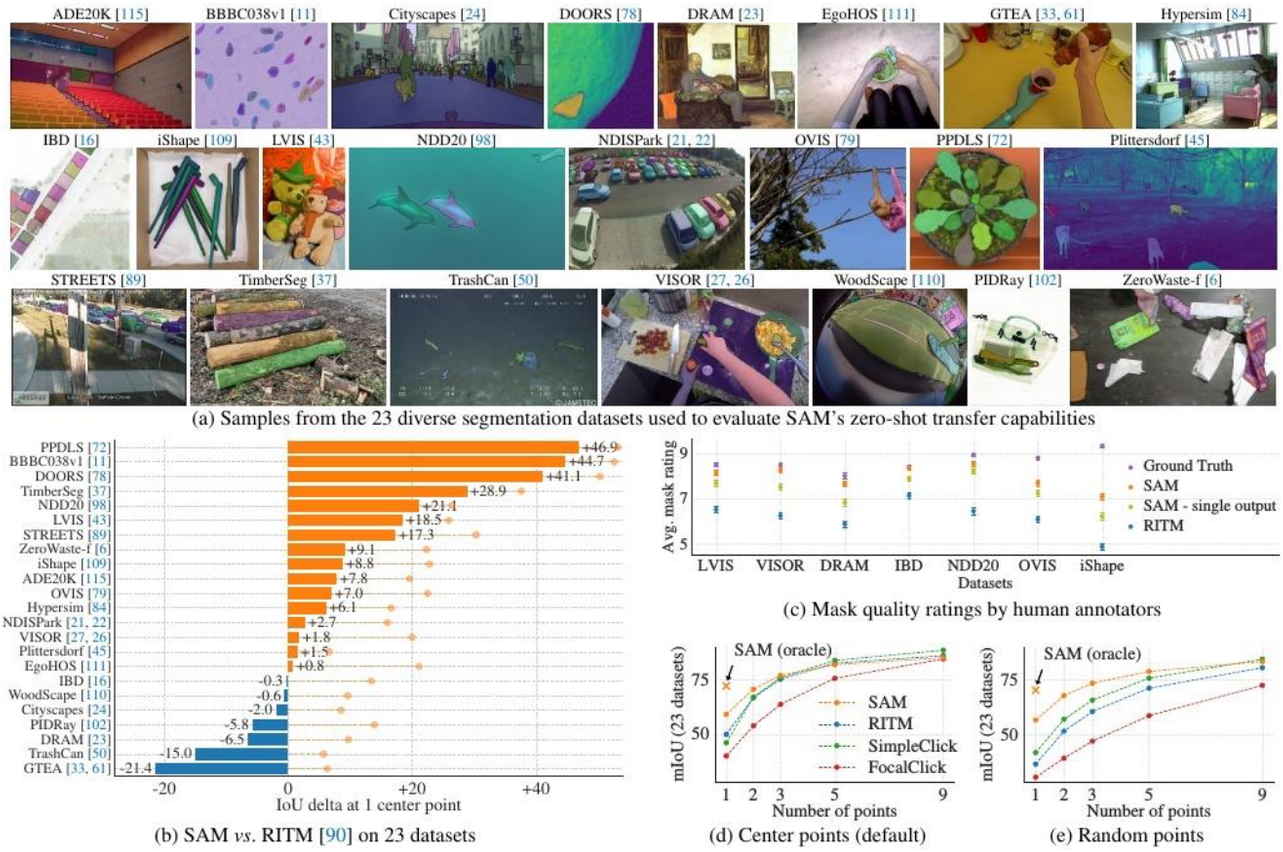


图 7：指向 23 个数据集的掩码评估。(a)数据集样本。(b) SAM 和最强单点分段 RITM 的平均 IoU[90]。由于模糊性，单个掩模可能与地面真实值不匹配；圆圈表示 SAM 3 个预测中最相关的“oracle”结果。(c)注释者对每个数据集的掩码质量评分的比较，从 1（最差）到 10（最好）。Mask center 作为提示符。(d, e）不同点数的 mIoU。SAM 显著优于先前的 1 点交互式分割器，并且与更多点相当。1 点的绝对 mIoU 低是模糊性的结果。

数据集。我们使用了一套新编译的 23 个数据集，这些数据集具有不同的图像分布，详见附录表 4。我们使用所有 23 个数据集进行 mIoU 评估。对于人类研究，我们使用图 7c 中列出的子集（由于此类研究的资源需求）。这个子集包括两个数据集，其中 SAM 根据自动度量优于 RITM 和低于这些数据集。

结果。首先，我们将使用 mIoU 对全套 23 个数据集进行自动评估。我们将图 7b 中每个数据集的结果与 RITM 进行比较。SAM 在 23 个数据集中的 16 个上产生了更高的结果，高出了 IoU 47 个。我们还提出了一个“oracle”结果，其中通过将 SAM 的 3 个掩码与基础真实值进行比较来选择最相关的掩码，而不是选择最自信的掩码。这揭示了模糊性对自动评估的影响。特别是，使用 oracle 执行歧义解析，SAM 在所有数据集上的性能都优于 RITM。

人体研究的结果如图 7c 所示。误差条为 95%置信区间(所有差异均显著；详情见 § F)。我们观察到注释-null 研究人员一致认为 SAM 口罩的质量远远高于最严格的基准 RITM。具有单一输出掩码的精简版“不存在歧义”的 SAM 一直具有较低的评级。SAM 的平均评分在 7 到 9 之间，这与定性评分准则相对应：“高分（7-9）：对象是可识别的，错误小而罕见（例如，遗漏了一个小的，严重模糊的断开的组件，……）。这些结果表明，SAM 已经学会了从单个点分割有效掩模。请注意，对于像 DRAM 和 IBD 这样的数据集，SAM 在自动指标上表现较差，但它在人类研究中始终获得较高的评级。

图 7d 显示了额外的基线，SimpleClick[65] 和 FocalClick[17]。随着点数从 1 增加到 9，我们观察到方法之间的差距减小。随着任务变得更简单，这是可以预料到的；此外，SAM 没有针对非常高的 IoU 制度进行优化。最后，在图 7e 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 与基线之间的差距越来越大，SAM 在任何一种抽样方法下都能够获得可比的结果。

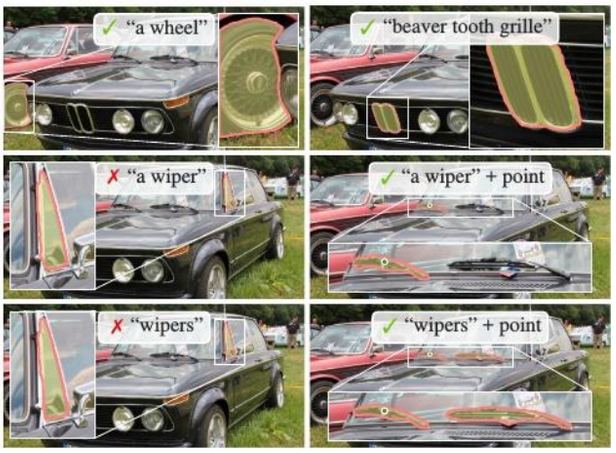


图 8：零镜头文本转掩码。SAM 可以处理简单而微妙的文本提示。当 SAM 无法做出正确的预测时，额外的点数提示可以提供帮助。

6.2.Zero-Shot Text-to-Mask

的方法。这个实验是 SAM 从自由格式文本提示中分割对象的能力的概念验证。虽然我们在之前的所有实验中都使用了完全相同的 SAM，但对于这个 SAM 的训练过程进行了修改，使其能够感知文本，但不需要新的文本注释。具体来说，对于每个人工收集的面积大于 100-的掩码 [2 null] 提取 CLIP 图像嵌入。然后，在训练期间，我们用提取的 CLIP 图像嵌入提示 SAM 作为其第一次交互。这里的关键观察是，由于 CLIP 的图像嵌入被训练成与其文本嵌入对齐，因此我们可以使用图像嵌入进行训练，但使用文本嵌入进行推理。也就是说，在推理时，我们通过 CLIP 的文本编码器运行文本，然后将结果文本嵌入作为 SAM 的提示（参见 § E）。详情见 § e . 5)。

结果。我们在图 8 中展示了定性结果。SAM 可以根据简单的文本提示（如“车轮”）和短语（如“海狸牙格栅”）来分割对象。当 SAM 无法仅从文本提示中选择正确的对象时，一个额外的点通常会修复预测，类似于[30]。

7.讨论

基础模型。从机器学习的早期开始，预训练模型就已经适应了下游任务[97]。近年来，随着对规模的日益重视，这种范式变得越来越重要，这种模型最近被（重新）标榜为“基础模型”：即“在大规模的广泛数据上训练并适应广泛的下游任务”的模型[8]。我们的工作与这个定义很好地相关，尽管我们注意到图像分割的基础模型本质上是一个有限的范围，因为它代表了计算机视觉的一个重要的，但分数的子集。我们

null 我们还将我们的方法的一个方面与[8]进行了对比，[8]强调自监督学习在基础模型中的作用。虽然我们的模型是用自监督技术（MAE[46]）初始化的，但它的绝大部分能力来自于大规模的监督训练。在数据引擎可以扩展可用注释的情况下，比如我们的，监督训练提供了一个有效的解决方案。

组合性。预先训练的模型可以提供甚至超出训练时想象的新功能。一个突出的例子是 CLIP[80]如何在更大的系统中作为组件使用，如 DALL • E[81]。我们的目标是使用 SAM 使这种组合更加简单。我们的目标是通过要求 SAM 预测广泛分割提示的有效掩码来实现这一目标。其效果是在 SAM 和其他组件之间创建一个可靠的接口。例如，MCC[104]可以很容易地使用 SAM 来分割感兴趣的物体，并从单个 RGB-D 图像中实现对未见物体的强泛化以进行 3D 重建。在另一个例子中，SAM 可以通过可穿戴设备检测到的注视点来提示，从而实现新的应用。由于 SAM 能够推广到新的领域，比如以自我为中心的图像，这样的系统不需要额外的训练就可以工作。

的局限性。虽然 SAM 总体上表现良好，但它并不完美。它可能会错过精细的结构，有时会产生小的不连接的组件，并且不会像“放大”的计算密集型方法（例如[17]）那样清晰地产生边界。一般来说，当提供许多点时，我们期望专用的交互式分割方法优于 SAM，例如[65]。与这些方法不同，SAM 是为通用性和使用广度而设计的，而不是为高 IoU 交互式分割而设计的。此外，SAM 可以实时处理提示，但是当使用重型图像编码器时，SAM 的整体性能不是实时的。我们对文本到掩码任务的尝试是探索性的，并不是完全健壮的，尽管我们相信它可以通过更多的努力得到改进。虽然 SAM 可以执行许多任务，但尚不清楚如何设计实现语义和全景分割的简单提示。最后，还有一些特定于领域的工具，例如[7]，我们希望它们在各自的领域中表现优于 SAM。

结论。Segment Anything 项目是将图像分割提升到基础模型时代的一次尝试。我们的主要贡献是一个新的任务（提示分割），模型（SAM）和数据集（SA-1B），使这一飞跃成为可能。SAM 是否达到基础模型的地位还有待观察，因为它在社区中是如何使用的，但无论我们期望这项工作的前景如何，超过 1B 个掩模的发布，以及我们及时的分割模型将有助于铺平前进的道路。

参考文献

[1]爱德华·H·阿德尔森。论看东西：人类和机器对材料的感知。*《人类视觉与电子成像》VI，2001*。5

[2] Bogdan Alexe， Thomas Deselaers 和 Vittorio Ferrari。什么是对象？*CVPR, 2010 年*。4、19

[3] Pablo Arbel´aez, Michael Maire， Charless Fowlkes 和 Jitendra Malik。轮廓检测和分层图像分割。*TPAMI, 2010 年*。4、19、28

[4]吉米·雷巴、杰米·瑞安·基罗斯和杰弗里·E·辛顿。层正常化。*农业学报：1607.06450,2016*。13

[5]包杭波，李东，魏福如。BEiT：图像转换器的 BERT 预训练。*农业学报：2106.08254,2021*。15

[6] Dina Bashkirova, Mohamed Abdelfattah, Ziliang Zhu, James Akl, Fadi Alladkani, Ping Hu, Vitaly Ablavsky, Berk Calli， Sarah Adel Bargal 和 Kate Saenko。ZeroWaste 数据集：朝向杂乱场景中的可变形物体分割。*CVPR, 2022 年*。8、18

[qh]斯图尔特·伯格、多米尼克·库特拉、托尔本·克鲁格、克里斯托弗·n·斯特拉赫、伯恩哈德·x·考斯勒、卡斯滕·豪博尔德、马丁·席格、珍妮兹·阿莱斯、托尔斯滕·贝尔、马库斯·鲁迪、凯末尔·埃伦、杰米·i·塞万提斯、徐、芬恩·博滕穆勒、阿德里安·沃尔尼、张冲、乌尔里希·歌德、弗雷德·a·哈姆普雷希特和安娜·克雷舒克。Ilastik：用于（生物）图像分析的交互式机器学习。*Nature Methods, 2019*。9

[10] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, 迈克尔 S Bernstein, Jeannette Bohg, Antoine Bosselut， Emma Brunskill 等。论基础模型的机遇与风险。*农业学报：2108.07258,2021*。1、9

[9]古斯塔夫·布雷德尔，克里斯汀·坦纳和安德·科努科格鲁。分段编辑网络的迭代交互训练。*MICCAI, 2018 年*。15

[10]汤姆·布朗、本杰明·曼恩、尼克·莱德、梅勒妮·苏比亚、贾里德·D·卡普兰、普拉弗拉·达里瓦尔、阿文德·尼拉坎坦、普拉纳夫·希亚姆、吉里什·萨斯特里、阿曼达·阿斯凯尔、桑迪尼·阿加瓦尔、阿里尔·赫伯特-沃斯、格雷琴·克鲁格、汤姆·亨尼根、雷温·查尔德、阿迪蒂亚·拉梅什、丹尼尔·齐格勒、杰弗里·吴、克莱门斯·温特、克里斯·黑塞、马克·陈、埃里克·西格勒、马特乌斯·利特温、斯科特·格雷、本杰明·切斯、杰克·克拉克、克里斯托弗·伯纳、萨姆·麦坎迪什、亚历克·拉德福德、伊利亚·苏茨克维尔和达ario Amodei。语言模型是少数几次的学习者。*NeurIPS, 2020 年*。1、4

[qh] Juan C. Caicedo, Allen Goodman, Kyle W. Karhohs, Beth A. Ci- mini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, Mohammad Rohban， Shan- tanu Singh 和 Anne E. Carpenter。跨成像实验的核分割：2018 年数据科学碗。*Nature Methods, 2019*。8,17,18

[12]约翰·坎尼。边缘检测的计算方法。*TPAMI, 1986 年*。19

[13] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier， Alexander Kirillov 和 Sergey Zagoruyko。端到端变形金刚对象检测。*大会,2020 年*。5、13、14

[14]纪尧姆·夏皮亚、马蒂亚斯·霍夫曼和伯恩哈德·谢尔科夫。通过多模态预测实现的自动图像着色。*大会,2008 年*。5、14

[15] Neelima Chavali, Harsh Agrawal， Aroma Mahendru 和 Dhruv Batra。对象-提案评估协议是“可游戏的”。*CVPR, 2016 年*。19、20

[10]陈，徐，卢淑芳，梁荣华，南连亮。MVS 建筑的三维实例分割。*IEEE 地球科学与遥感学报，2022*。8、17、18、22、23、24

[10]陈，赵志燕，张一磊，段曼妮，齐冬莲，赵恒双。FocalClick：走向实用的交互式图像分割。*CVPR, 2022 年*。7、8、9、17、19

[18]程博文、伊山·米斯拉、亚历山大·G·施维因、亚历山大·基里洛夫、罗希特·吉达尔。用于通用图像分割的 mask -attention mask transformer。*CVPR, 2022 年*。4

null[19]郑博文、亚历克斯·施维因和亚历山大·基里洛夫。逐像素分类并不是语义分割所需要的全部。*NeurIPS, 2021 年*。5,13,14

[20] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton， Sebastian Gehrmann 等。PaLM：用路径（pathways）缩放语言建模。*农业学报：2204.02311,2022*。1

[21]卢卡·钱皮、卡洛斯·圣地亚哥、若昂·科斯泰拉、克劳迪奥·热纳罗和朱塞佩·阿马托。面向交通密度估计的领域自适应。*计算机视觉、成像与计算机图形学理论与应用国际联合会议，2021*。8、18

[22]卢卡·钱皮、卡洛斯·圣地亚哥、若昂·科斯泰拉、克劳迪奥·热纳罗和朱塞佩·阿马托。昼夜实例分割公园（NDIS- park）数据集：用于停车场车辆检测、分割和计数的白天和夜间拍摄的图像集合。*Zen- odo, 2022*。8、18

[23] Nadav Cohen， Yael Newman 和 Ariel Shamir。艺术绘画中的语义分割。*计算机图形学论坛，2022*。8、17、18、22、23、24

100 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*。8,17,18

[25]布鲁诺·达席尔瓦、乔治·科尼达里斯和安德鲁·巴托。学习参数化技能。*ICML, 2012 年*。4

[26]迪马·达门、黑兹尔·道蒂、乔瓦尼·玛丽亚·法里内拉、安东尼奥·福尔纳里、马建、伊万杰洛斯·卡扎科斯、戴维·莫尔蒂森蒂、乔纳森·门罗、托比·佩雷特、威尔·普莱斯和迈克尔·雷。重新缩放以自我为中心的愿景：EPIC- KITCHENS-100 的收藏、管道和挑战。*IJCV, 2022 年*。8、18、22、23、24

[27] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler， David Fouhey 和 Dima Damen。EPIC-KITCHENS VISOR 基准：视频分割和对象关系。*NeurIPS, 2022 年*。8,17,18,22,23,24

[28] Terrance 德弗里斯，Ishan Misra， Changhan 王和 laurence Van der Maaten。对象识别对每个人都适用吗？*2019 年 CVPR 研讨会*。16

[29] Mark D´ıaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei， Vinodkumar Prabhakaran 和 Emily Denton。Crowd-WorkSheets：计算众包数据集注释底层的个人和集体身份。*ACM 公平、问责和透明度会议，2022 年*。24

[30]丁恒辉，斯科特·科恩，布莱恩·普莱斯，姜旭东。PhraseClick：朝着通过短语和点击实现灵活的交互细分。*大会,2020 年*。9

[31]Piotr Doll 和 C·劳伦斯·齐特尼克。使用结构化森林的快速边缘检测。*TPAMI, 2014 年*。19

[qh]阿列克谢·多索维茨基，卢卡斯·拜尔，亚历山大·科列斯尼科夫，德克·魏森博恩，翟晓华，托马斯·乌特希纳，穆斯塔法·德哈加尼，马蒂亚斯·明德勒，格奥尔格·海戈尔德，西尔万·盖利，雅各布·乌斯科瑞特，尼尔·霍尔斯比。一张图片抵得上 16x16 个单词：用于大规模图像识别的《变形金刚》。*ICLR, 2021 年*。5,7,13

[qh] Alireza Fathi，任晓峰，詹姆斯 M.雷格。学习在以自我为中心的活动中识别物体。*CVPR, 2011*。8,17,18

[34] Pedro F Felzenszwalb 和 Daniel P hutenlocher。高效的基于图的图像分割。*IJCV, 2004 年*。19

托马斯 B.菲茨帕特里克。太阳反应性皮肤类型的有效性和实用性 i 至 vi. 皮肤病学档案，1988。17

[36] Marco Forte, 布莱恩·普莱斯, 斯科特·科恩， Ning 徐和法郎，ois Piti ' e。在交互式分割中达到 99%的准确率。*农业学报：2003.07932,2020*。5,14,15

[37]让-米歇尔·福丹、奥利维尔·加马什、文森特·格龙丹、法郎·约·波默洛和菲利普·吉古埃。林业作业中自主原木抓取的实例分割。*——2022 人*。8、18

[38] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daum ' e Iii 和 Kate

克劳福德。数据集的数据表。*ACM 通讯，2021*。24

[39] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, chung - yi Lin, Ekin D Cubuk, Quoc V Le，和 Barret Zoph。简单复制粘贴是一种用于实例分割的强数据增强方法。*CVPR， 2021 年*。13、15、21

[40]Ross Girshick、杰夫·多纳休、特雷弗·达雷尔和吉腾德拉·马利克。丰富的特征层次结构，用于精确的对象检测和语义分割。*CVPR, 2014*。19

[10] Priya Goyal, Piotr Doll, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, 安德鲁 Tulloch, Yangqing 贾文奇, 何开明。精确的、大的小批量 SGD: 1 小时内训练 ImageNet。*农业学报:1706.02677, 2017*。15

b[42]克里斯汀·格劳曼、安德鲁·韦斯特伯里、尤金·伯恩、扎卡里·查维斯、安东尼奥·弗尔纳里、罗希特·格达尔、杰克逊·伯格、姜旭东、刘星宇、刘星宇、米格尔·马丁、图沙尔·纳加拉扬、伊利亚·拉多萨沃维奇、桑托什·库马尔·拉马克里希南、菲奥娜·瑞安、贾扬特·夏尔马、迈克尔·雷、徐、钟聪徐、赵陈鑫磊、西丹特·班萨尔、德鲁夫·巴特拉、文森特·卡蒂利尔、肖恩·克兰、杜天都、莫里·杜拉蒂、阿克沙伊·埃拉帕利、克里斯托弗·费希滕霍费尔、阿德里亚诺·弗拉戈梅尼、傅其琛、克里斯蒂安·富根、亚伯拉罕·格布雷泽西、克里斯蒂娜·冈萨雷斯、詹姆斯·希利斯、黄旭华、黄旭华、贾文奇、韦斯利·邱、贾希姆·柯拉尔、萨特威克·科图尔、阿努拉格·库马尔、费德里科·兰迪尼、Chao 李、李阳浩、李、Karttikeya Mangalam、Raghava Mod- hugu、Jonathan Munro、Tullie Murrell、西康 Takumi、Will Price、Paola Ruiz Puentes、Merey Ramazanova、Leda Sari、Kiran Somasundaram、Audrey Southerland、Sugano Yusuke、陶瑞杰、吴新迪、吴宇晨、八木卓uma、朱云毅、Pablo Arbelaez、David Crandall、Dima Damen、Giovanni Maria Farinella、Bernard Ghanem、Vamsi Krishna Ithapu、C. V. Jawahar、hanbyl Joo、Kris Kitani、李、Richard Newcombe、Aude Oliva、Hyun Soo Park、詹姆斯 M. Rehg、Yoichi Sato、shijianbo 、 Mike Zheng Shou 、 Antonio Torralba 、 Lorenzo Torresani、Mingfei Yan 和 Jitendra Malik。Ego4D：环游世界 3000 小时的 Egocentric 视频。*CVPR, 2022*。18

b[43] Agrim Gupta， Piotr Dollar 和 Ross Girshick。LVIS：用于大词汇实例分割的数据集。*CVPR, 2019*。2、6、7、8、17、18、19、21、23

Abner Guzman-Rivera, Dhruv Batra，和 Pushmeet Kohli。选择学习（Multiple choice learning）：学习产生多个结构化输出。*NeurIPS, 2012 年*。5、14

[45] tim Haucke, Hjalmar S. K ¨ uhl 和 Volker Steinhage。SOCRATES：利用立体视觉引入视觉野生动物监测的深度。*传感器,2022*。8、18

[10]何开明，陈鑫磊，谢思宁，李阳浩，Piotr Doll, Ross Girshick。蒙面自动编码器是可扩展的视觉学习器。*CVPR, 2022*。5,7,9,13,15

bbb 何开明，张翔宇，任少卿，孙健。用于图像识别的深度残差学习。*CVPR, 2016 年*。14

[48]丹·亨德里克斯和凯文·金普尔。高斯误差线性单位（gelus）。*《农业学报》，2016*。13

[49] 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。*农业学报，2022*。1

[qh]洪仲锡，迈克尔·富尔顿，朱纳德·萨塔尔。垃圾桶：面向海洋垃圾视觉检测的语义分割数据集。*农业学报：2007.08097,2020*。8,17,18

[51]黄高，孙瑜，刘庄，Daniel Sedra, Kilian Q Weinberger。具有随机深度的深度网络。*大会,2016 年*。15

[52] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov，和 Humphrey Shi。Oneformer：一个统治通用图像分割的变压器。*农业学报:2211.06220, 2022*。4

[53]贾超、杨茵飞、夏烨、陈怡婷、Zarana Parekh、范晓辉、乐国、宋允萱、李震、Tom Duerig。

null 基于噪声文本监督的视觉和视觉语言表征学习的扩展。*ICML, 2021 年*。1

[54]贾里德·卡普兰，山姆·麦坎迪什，汤姆·亨尼根，汤姆·B·布朗，本杰明·切斯，雷温·查尔德，斯科特·格雷，亚历克·雷德福，杰弗里·吴和达里奥·阿莫代。神经语言模型的缩放定律。*农业学报：2001.08361,2020*。1

[55]迈克尔·卡斯，安德鲁·威特金和德米特里·特佐普洛斯。蛇：主动轮廓模型。*IJCV, 1988 年*。4

[10]金大勋，林宗义，Anelia Angelova，权仁素，郭卫成。在不学习分类的情况下学习开放世界对象建议。*IEEE 机器人与自动化通讯，2022*。19

[57] Alexander Kirillov, 何开明, Ross Girshick， Carsten Rother 和 Piotr Doll。展示全景的分割。*CVPR, 2019 年*。4

[58]阿丽娜·库兹涅佐娃、哈桑·罗姆、尼尔·奥尔德林、贾斯帕·乌伊林斯、伊万·克拉辛、乔迪·蓬图塞、沙哈布·卡马里、斯特凡·波波夫、马泰奥·马洛奇、Alexander·科列斯尼科夫、汤姆·杜瑞格和维托里奥·法拉利。开放图像数据集 v4：统一的图像分类、对象检测、大规模的视觉关系检测。*IJCV, 2020 年*。2,6,7,16

[59] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt， 和 Thomas Dandres。量化机器学习的碳排放。*《农业学报》，2019*。28

[10]李阳浩，毛涵子，Ross Girshick，何开明。探索用于目标检测的 plain vision transformer 骨干。*大会,2022 年*。5、13、19、21、22、23

[61]李寅，叶哲帆，James M. Rehg。探究自我中心行为。*CVPR, 2015 年*。8、18

[62]李竹文，陈奇峰，Vladlen Koltun。具有潜在多样性的交互式图像分割。*CVPR, 2018 年*。5、14、17

[63]林宗义，郭雅娟，Ross Girshick，何开明，Piotr Doll。密集目标检测中的焦损失。*ICCV, 2017 年*。5、14

[64]林宗义，迈克尔 Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll, C . Lawrence Zitnick。Microsoft COCO: context 中的 Common objects。*大会,2014 年*。2、4、6、7、16、17、18、21

[65]刘琴、徐振林、贝尔塔修斯、马克·尼塔默尔。Sim- pleClick：基于简单视觉变形器的交互式图像分割。*农业学报（英文版）：2210.11006,2022*。7,8,9,17

[66]伊利亚·洛什奇洛夫和弗兰克·哈特。解耦权衰减正则化。*ICLR, 2019 年*。15

[67]凯茜·H·卢卡斯、丹尼尔·OB·琼斯、凯瑟琳·J·霍利黑德、罗伯特·H·康登、卡洛斯·M·杜阿尔特、威廉·M·格雷厄姆、凯莉·L·罗宾逊、凯莉·A·皮特、马克·席尔德豪尔和吉姆·雷格兹。全球海洋中胶状浮游动物生物量：地理变异和环境驱动因素。*全球生态与生物地理，2014*。18

[68] Sabarinath Mahadevan, Paul Voigtlaender，和 Bastian Leibe。迭代训练的交互式分割。*BMVC, 2018 年*。4、15

[69] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, Luc Van Gool。深度极值切割：从极值点到对象分割。*CVPR, 2018 年*。6

[70] David Martin, Charless Fowlkes， Doron Tal 和 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*。19 日,28 日

[71]福斯托·米勒塔里、纳西尔·纳瓦布和赛义德·艾哈迈德·艾哈迈迪。V-Net：用于体积医学图像分割的全卷积神经网络。*3 dv, 2016*。5、14

[72] Massimo Minervini, Andreas Fischbach, Hanno Scharr, Sotirios A. Tsaftaris。基于图像的植物表型分析的细粒度注释数据集。*Pattern Recognition Letters, 2016*。8、18

[73] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman，Ben 哈钦森，Elena Spitzer， inoluwa Deborah Raji 和 Timnit Gebru。模特报道用的模特卡。*2019 年公平、问责和透明度会议论文集*。24 日,28 日

[74] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller，和 Vittorio Ferrari。极致点击，实现高效对象标注。*ICCV, 2017 年*。6

[75]大卫·帕特森，约瑟夫·冈萨雷斯，郭乐，陈亮，路易斯-米克尔·芒吉亚，丹尼尔·罗斯柴尔德，大卫·苏，莫德·特西耶，杰夫·迪恩。碳排放与大型神经网络训练。*arXiv: 2104.10350, 2021*。28

[76]李晓明，李晓明，李晓明，等。基于双向语言模型的半监督序列标注。*计算语言学协会第 55 届年会论文集，2017*。16

[77]蒲梦阳、黄亚平、刘玉明、关清基、凌海斌。EDTER：带变压器的边缘检测。*CVPR, 2022 年*。19

[78]马蒂亚·普利亚蒂和弗朗西斯科·托普托。DOORS: Dataset fOr bOuldeRs Segmentation.。*Zenodo, 2022 年*。8、18

[79]齐继阳、高燕、胡尧、王兴刚、刘小雨、白翔、塞尔日·贝隆吉、艾伦·尤维尔、菲利普·托尔、宋白。闭塞视频实例分割：一个基准。*IJCV, 2022 年*。8、18、22、23、24

[80] 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转移的视觉模型。*ICML, 2021 年*。1、2、4、5、7、9、13、21

[81] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford， Mark Chen 和 Ilya Sutskever。零射击文本到图像的生成。*ICML, 2021 年*。1,4,9

[82]任少卿，何开明，孙健。更快的 R-CNN：用区域建议网络走向实时目标检测。*NeurIPS, 2015 年*。6 日 19

[83]Ren。学习用于分词的分类模型。*ICCV, 2003 年*。4

[84]迈克·罗伯茨，杰森·拉马布拉姆，阿努拉格·兰詹，阿图利特·库马尔，米格尔·安吉尔·包蒂斯塔，内森·帕赞，拉斯·韦伯和约书亚·m·萨斯金德。Hypersim：用于整体室内场景理解的逼真合成数据集。*ICCV, 2021 年*。8,17,18

[85] Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari，和 Caroline Pantofaru。为公平向更具包容性的人物注释迈进了一步。*2021 年 AAAI/ACM AI、伦理与社会会议论文集，2021 年*。16

[86] Sefik Ilkin Serengil 和 Alper Ozpinar。LightFace：一个混合深度人脸识别框架。*ASYU, 2020 年*。26

[87] 塞菲克·伊尔金·塞伦吉尔和阿尔珀·奥斯皮纳尔。HyperExtended LightFace：一个面部属性分析框架。*ICEET, 2021 年*。26

[88]杰米·肖顿、约翰·温、卡斯滕·罗瑟和安东尼奥·克里明-伊西。TextonBoost：用于多类物体识别和分割的关节外观、形状和上下文建模。*大会,2006 年*。4

[89]科里·斯奈德和杜明。STREETS：一种新颖的交通流摄像机网络数据集。*NeurIPS, 2019 年*。8、18

[90]康斯坦丁·索菲尤克、伊利亚·A·彼得罗夫、安东·科努申。基于掩码指导的交互式分割迭代训练的复兴。*ICIP, 2022 年*。5、7、8、14、15、17、22、23、28

[91] 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。*《机器学习研究杂志》，2014 年*。14

[92]李志刚，李志刚。实时跟踪的自适应背景混合模型。*CVPR, 1999 年*。4

[93] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara fridovicch - keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ra- mamoorthi, Jonathan Barron, Ren Ng。傅里叶特征让网络在低维域学习高频函数。*NeurIPS, 2020 年*。5、13

[94]唐岩松、田毅、陆继文、冯建江、周杰。RGB-D 自我中心视频中的动作识别。*ICIP, 2017 年*。18

null[95]唐岩松、王子安、陆继文、冯建江、周杰。用于 RGB-D 自我中心动作识别的多流深度神经网络。*IEEE 视频技术电路与系统学报，2019*。18

[96]世界银行。世界按收入和地区划分，2022 年。/ the-world-by-income-and-region.html https://datatopics.worldbank.org/world-development-指标。16

[97]塞巴斯蒂安·特伦。学习第 n 个东西比学习第一个东西容易吗？*NeurIPS, 1995 年*。9

[98]卡梅隆·特罗特，乔治亚·阿特金森，马特·夏普，克尔斯滕·理查森，a·斯蒂芬·麦高夫，尼克·赖特，本·伯维尔，佩尔·伯格伦。NDD20：用于粗粒度和细粒度分类的大规模少拍海豚数据集。*农业学报：2005.13359,2020*。8,17,18,22,23,24

[99] 美国环境保护署。温室气体当量计算器。https://www.epa.gov/energy/greenhouse-气体当量计算器，2022。28

[100] Koen EA van de Sande, Jasper RR Uijlings, Theo Gevers, Arnold WM Smeulders。分割作为对象识别的选择性搜索。*ICCV, 2011 年*。19

[101] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser，和 Illia Polosukhin。注意力就是你所需的一切。*NeurIPS, 2017 年*。5、13

[102]王伯英，张立波，温龙银，刘祥龙，吴超远。走向现实世界违禁物品检测：大规模 x 射线基准。*CVPR, 2021 年*。8,17,18

[103]王维耀，王恒，王伟耀，王伯英，Jitendra Malik。开放世界实例分割：从习得的两两亲和中挖掘伪地真。*CVPR, 2022 年*。19

[104]吴超远，刘志强，马立祺。面向 3D 重建的多视图压缩编码。*CVPR, 2023 年*。9

[105]肖剑雄，詹姆斯·海斯，克里斯塔·爱辛格，奥德·奥利瓦，安东尼奥·托拉尔巴。SUN 数据库：从修道院到动物园的大规模场景识别。*CVPR, 2010 年*。18

[106]谢思宁，涂卓文。整体嵌套边缘检测。*ICCV, 2015 年*。19

[107]徐宁，杨继梅，杨继梅。深度交互对象选择。*CVPR, 2016 年*。4、17

[108]杨磊，克林特·奇纳米，李飞飞，邓佳，奥尔格·罗斯-萨科夫斯基。迈向更公平的数据集：过滤和平衡 imagenet 层次结构中 people 子树的分布。*2020 年公平、问责和透明度会议论文集，2020 年*。17

[109]杨磊，魏燕，何一生，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。*农业学报：2109.15068,2021*。8,18,22,23,24

[110]陈建军，陈建军，陈建军，陈建军，陈建军，陈建军，陈建军，陈建军。WoodScape：用于自动驾驶的多任务、多摄像头鱼眼数据集。*ICCV, 2019 年*。8、18

[111]张凌志，周生豪，Simon Stent，史建波。细粒度以自我为中心的手-物分割：数据集、模型和应用。*大会,2022 年*。8,17,18

[112]张立波，庞江淼，陈凯。K-Net：走向统一的图像分割。*NeurIPS, 2021 年*。4

[113]赵介宇，王伯英，张凯伟。男性也喜欢购物：利用语料库层面约束减少性别偏见放大。*农业学报:1707.09457, 2017*。16

[114]周博雷、阿加塔·拉佩德里扎、阿迪提亚·科斯拉、奥德·奥利瓦和安东尼奥·托拉尔巴。Places：用于场景识别的 1000 万张图像数据库。*TPAMI, 2017 年*。18

[115]周伯雷、赵航、泽维尔·普伊格、肖特特、桑雅·菲德勒、阿德拉·巴里鲁索、安东尼奥·托拉尔巴。通过 ADE20K 数据集对场景的语义理解。*IJCV, 2019 年*。2,7,8,18