arXiv:2301.03992v1 [cs.CV] 10 Jan 2023

视觉变换器是优秀的掩码自动标注器

Shiyi Lan[1] Xitong Yang[2] Zhiding Yu[1] Zuxuan Wu[3] Jose M.Alvarez[1] AnimaAnandkumar[1,41]
NVIDIA [2]Meta AI，FAIR [3]复旦大学 [4]加州理工学院https://github.com/NVlabs/mask-auto-labeler



图 1. **掩码示例** **掩码自动标注器在 COCO 上生成的伪标签**。仅将人工标注的边界框 re
用作训练期间的监督 以获得这些结果。

摘要

我们提出掩码自动标注器（*MAL*），这是一种高质量的基于 *Transformer* 的掩码自动标注框架，仅使用框标注进行实例分割。*MAL* 将边界框裁剪图像作为输入，并有条件地生成其掩码伪标签。我们证明，视觉 *Transformer* 是出色的掩码自动标注器。我们的方法显著缩小了自动标注与人工标注在掩码质量方面的差距。使用 *MAL* 生成的掩码训练的实例分割模型，其性能几乎可以媲美全监督对应模型，最高可达到全监督模型性能的 *97.4%*。最佳模型在 *COCO* 实例分割（*2017* 测试开发集）上取得了 *44.1%* 的平均精度均值，显著优于当前最先进的框监督方法。定性结果表明，在某些情况下，*MAL* 生成的掩码甚至优于人工标注。

1. 引言

过去十年间，计算机视觉取得了显著进展。实例分割等任务使我们能够以像素级精度定位并分割物体。然而，这些任务严重依赖大量昂贵的人工掩码标注。例如，在创建 COCO 数据集时，掩码标注耗费了约 5.5 万个工时，占总标注时间的约 79%。此外，人类也会犯错。人工标注往往与物体的实际边界不吻合。对于复杂物体，如果缺乏质量控制，人工标注的质量往往会显著下降。由于质量控制成本高昂且难以实施，一些其他大规模检测数据集，例如 OpenImages [2]和 Objects365 [3]，，仅包含部分实例分割标签，甚至完全不包含此类标签。

鉴于这些局限，人们对框监督实例分割的研究兴趣日益浓厚，其目标是直接根据边界框监督预测目标掩码。近期的框监督实例分割方法 [4–8]展现出良好的性能。这些方法的出现，挑战了长期以来认为训练实例分割模型需要掩码标注的观点。然而，当前最先进的方法与全监督基准之间仍存在不可忽视的差距。

我们的贡献：为解决框监督实例分割问题，我们提出了一种两阶段框架，包括掩码自动标注阶段和实例分割训练阶段（见图2）。我们提出了一种基于 Transformer 的掩码自动标注框架——掩码自动标注器（MAL），该框架以感兴趣区域（RoI）图像作为输入

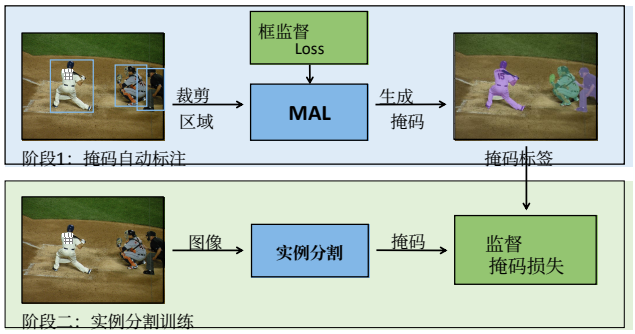


图2. 边界框监督实例分割两阶段框架概览。第一阶段，我们使用边界框监督训练掩码自动标注器，并有条件地为训练图像中的裁剪区域生成掩码（上）。随后，我们使用生成的掩码训练实例分割模型（下）。

并在边界框内有条件地生成高质量掩码（如图1所示）。我们的贡献可概括如下：

- 我们的两阶段框架设计灵活，可兼容任何实例分割架构。与现有方法不同，我们的框架简单明了，且不依赖特定的实例分割模块设计。

- 我们发现，将视觉变换器（ViT）用作图像编码器，能取得出人意料的出色自动标注效果。

  我们还证明，MAL 中的一些特定设计对于取得出色的自动标注效果至关重要，例如我们的基于注意力的解码器、结合框扩展的多实例学习，以及类别无关训练。得益于这些组件，MAL 有时甚至能在标注质量上超越人类。

- 使用 MAL 生成的掩码进行训练时，实例分割模型在 COCO 和 LVIS 上的表现最高可达全监督性能的 97.4%。我们的结果显著缩小了边界框监督方法与全监督方法之间的差距。我们还通过标注训练期间未见过的新类别，展示了 MAL 出色的开放词汇泛化能力。

我们的方法大幅优于现有所有最先进的框监督实例分割方法。这可能归功于视觉变换器的优良表征及其展现出的新兴特性，例如有意义的分组 [9], ；我们观察到，对物体的注意力可能会显著助益于我们的任务（如图所示。6）。我们还推测，我们的类别无关训练设计使 MAL 能够专注于学习通用分组而非专注于类别信息。我们的出色结果为在真实场景中消除实例分割对昂贵人工标注的需求铺平了道路。

2. 相关工作

2.1. 视觉变换器

Transformer最初是在自然语言处理中提出的 [10]。视觉变换器 [11]（ViT）随后兴起，成为极具竞争力的视觉识别模型，它们以多头自注意力（MHSA）取代卷积，作为基本构建模块。近年来，这些模型在许多视觉识别任务中展现出极具竞争力的性能 [12]。我们将现有的视觉变换器大致分为两类：普通视觉变换器和层次化视觉变换器。

标准视觉变换器。标准视觉Transformer [11] 是最早的视觉Transformer。标准视觉Transformer的结构最为简单，由一个分词嵌入层和一系列MHSA层组成。然而，全局MHSA层计算开销较大，通常还面临显著的优化问题。为提升其性能，人们提出了许多设计和训练方案，以数据高效的方式训练ViT [9, 13–19]。

层次化视觉变换器。层次化视觉变换器 [12, 20–22] 采用金字塔形架构，通过多尺度设计，使其能够胜任图像分类之外的其他任务。与普通视觉变换器相比，这类视觉变换器 [20, 21] 将多头自注意力层划分为多个层次阶段。阶段之间设有空间降采样层，例如最大池化层。这些架构通常会结合卷积层 [23]，并经常采用高效的自注意力设计来处理较长的序列。

2.2. 实例分割

实例分割是一项预测物体边界框和掩码的视觉识别任务。

全监督实例分割。在这种设定下，边界框和实例级掩码都作为监督信号提供。早期工作 [24–27]采用两阶段架构，在第一阶段生成候选框或分割候选区域，然后在第二阶段生成最终分割结果和分类信息。后来，实例分割模型大致分为两类：一些方法延续两阶段设计的思路，并将其扩展为多阶段架构 [28, 29]；另一些方法则简化架构，提出单阶段实例分割，例如 YOLACT [30]、SOLO [31, 32]、CondInst [33], 和 PolarMask [34, 35]。近期，DETR 和可变形 DETR [36, 37] 展现了基于查询的方法在目标检测中的巨大潜力。随后，MaxDeepLab [38]、MaskFormer [39],

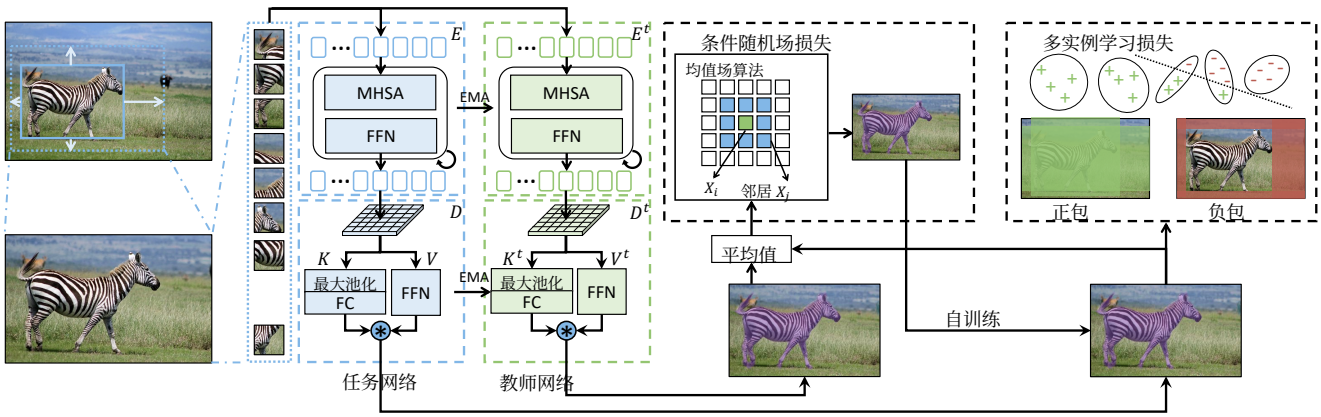


图 3. MAL架构概览。我们展示了掩码自动标注器的架构。掩码自动标注器以裁剪图像作为输入。掩码自动标注器由两个对称的网络组成，任务网络和教师网络。每个网络都包含图像编码器 E（或 E^t）和掩码解码器 D（或 D^t）。我们使用指数移动平均（EMA）更新教师网络的权重。我们采用多实例学习（MIL）损失和条件随机场（CRF）损失。CRF 损失对教师网络和任务网络的平均掩码预测进行处理，以使训练更加稳定，并生成用于自训练的精细掩码。

PanopticSegFormer [40]，Mask2Former [41]和 Mask DINO [42]沿着这一方向提出，进一步拓展了实例分割的边界。另一方面，实例分割也受益于更强大的骨干网络设计，例如 Swin Transformer [12, 22]，ViTDet [43],和 ConvNeXt [44]。

弱监督实例分割。弱监督实例分割主要有两种方式：使用图像级标签进行学习，以及使用框级标签进行学习。前者利用图像级类别信息执行实例分割 [45–49]，，后者则使用框监督。Hsu 等人 [4]利用紧框先验。之后，Box-Inst [5]提出利用颜色平滑性来提高准确率。此外，DiscoBox [7]提出在该任务中同时利用颜色平滑性和图像间对应关系。其他后续工作 [6, 8] 也利用了紧框先验和颜色平滑先验。

2.3. 深度学习解释

人们对深入理解深度网络的兴趣，促使许多研究致力于探索深度神经网络的可解释性。例如，类别激活图（CAM） [50] 和 Grad-CAM [51]能够可视化卷积神经网络（CNN）在图像分类训练过程中逐渐形成的定位能力。这种能力也推动了大量弱监督定位研究，并揭示了其与一般弱监督学习之间的深层联系，这在一定程度上启发了本文的解码器设计。DINO [9] 进一步表明，在使用视觉变换器进行自监督学习时，会涌现出具有语义意义的视觉分组。此外，FAN [52] 表明，视觉变换器中这类涌现特性与其鲁棒性有关。

3. 方法

我们的方法不同于以往的框监督实例分割框架 [4–8] ，后者会同时学习检测和实例分割。如图2所示，我们采用两阶段框架：阶段一中，一个网络专注于生成掩码伪标签，另一个网络则专注于在阶段二学习实例分割 [24, 28, 41, 43] 。我们提出的自动标注框架用于阶段一，以生成高质量的掩码伪标签。

我们提出这一两阶段框架，是因为它具有以下优点：

- 我们可以放宽阶段一的学习约束，仅专注于掩码伪标签。因此，在这一阶段，我们可以使用感兴趣区域（RoI）图像，而非未裁剪图像作为输入。这一改变使我们能够提高小目标的分辨率，并采用第 3.1 节提到的强训练技术，从而提升掩码质量。

- 我们可以在阶段一和阶段二采用不同的图像编码器和掩码解码器，以获得更高的性能。我们通过实验发现，阶段一和阶段二的图像编码器与掩码解码器适用的架构有所不同。消融研究见表 3 和4。

- 我们可以在阶段二使用 MAL 生成的掩码直接训练最先进的全监督实例分割模型。这使我们的方法比以往依赖特定架构的框监督实例分割方法更加灵活 [4–8]。

由于阶段二沿用之前的标准流程，这里无需重新介绍，因此我们将在接下来的小节中重点介绍阶段一（MAL）。

3.1. RoI 输入生成

大多数框监督实例分割方法 [4–7]都使用完整图像进行训练。然而，我们发现，使用 RoI 图像可能对框监督实例分割更有益。此外，我们比较了两种直观的 RoI 图像采样策略，以获取前景和背景像素，并详细说明了效果更好的策略——边界框扩展。

使用 RoI 输入的优势。 使用 RoI 图像作为输入有两个优点。首先，将 RoI 图像用作输入天然适合处理小目标，因为无论物体多小，RoI 图像都会被放大，从而避免低分辨率带来的问题。其次，使用 RoI 输入能让MAL专注于学习分割，避免分心于学习其他复杂任务，例如目标检测。**RoI 采**

样策略。 采样策略应确保纳入正像素和负像素。我们提出两种简单直接的采样策略：

- 第一种策略是使用边界框裁剪图像，作为正输入。对于负输入，我们使用随机生成且仅包含背景像素的边界框来裁剪图像。采用裁剪策略时，MAL无法生成良好的掩码伪标签。我们观察到，网络往往会学习到平凡解（将所有像素都预测为前景或背景）。

- 第二种策略是随机扩展边界框并纳入背景像素，从扩展后的行和列中选取负包。我们在图3 中直观展示了正/负包的定义方式，并在第3.3节中详细说明。 **这一精细设计对 MAL 的运行至关重要**，因为它能防止 MAL 学到平凡解。若没有这一设计，生成的掩码往往会填满整个边界框。

边界框扩展细节。给定一张未裁剪的图像 $\boldsymbol{I}^u \in \mathbb{R}^{C\times H^u\times W^u}$ 以及边界框 $\boldsymbol{b}=(x_0,y_0,x_1,y_1)$，其中表示左上角和右下角的 x、y 坐标。为获取背景像素，我们随机将边界框 $\boldsymbol{b}$ 扩展至 $\boldsymbol{b'}=(x_c+\beta_x(x_0-x_c)\ y_c+\beta'_x(y_0-y_c)\ x_c+\beta_y(x_1-x_c)\ y_c+\beta'_y(y_1-y_c))$，其中 $x_c=(x_0+x_1)/2$, $y_c=(y_0+y_1)/2$。为生成 $\beta_x,\beta'_x,\beta_y,\beta'_y$ 的随机值，我们针对 x 和 y 方向随机生成$\theta_x,\theta_y\in[0,\theta]$，其中 θ 是边界框扩展率的上限。接下来，我们随机生成$\beta_x\in[0,\theta_x]$和 $\beta_y\in[0,\theta_y]$。最后，我们将 β'_x 设为$\theta_x-\beta_x$ 并且 β'_y 作为 $\theta_y-\beta_y$。最后，我们使用 $\boldsymbol{b'}$ 裁剪图像并获得裁剪后的图像 $\boldsymbol{I}^t$。我们在表 θ 中对5进行了消融研究。最后，我们将裁剪后的图像调整大小$\boldsymbol{I}^t$ 至 $C\times H^c\times W^c$，使其与输入图像 $\boldsymbol{I}^c$的尺寸相同。

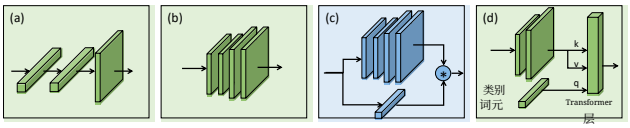


图 4。（a）全连接解码器（b）全卷积解码器（c）基于注意力的解码器（用于 MAL）（d）基于查询的解码器。

3.2. MAL架构

MAL 可分为两个对称的网络：任务网络和教师网络。任务网络由一个图像编码器组成，记作 E，以及一个掩码解码器，记作 D，如图3所示。教师网络的架构与任务网络相同。我们分别将任务网络和教师网络的分割输出记作 $\boldsymbol{m}$，$\boldsymbol{m}^t\in\{0,1\}^N$。

图像编码器。我们使用标准视觉Transformer [11] 作为图像编码器，并移除标准视觉Transformer的分类头。我们在第4.4节比较了不同的图像编码器。我们还尝试在标准视觉Transformer之上使用特征金字塔网络，例如 FPN [53], ，但这会导致性能下降。在 ViTDet [43]中也发现了类似的结论。

掩码解码器。对于掩码解码器 D，我们使用一个受 YOLACT [30]，启发的简单注意力网络，其中包括一个实例感知头 K 和一个逐像素头V，其中 $D(E(\boldsymbol{I}))=K(E(\boldsymbol{I}))\cdot V(E(\boldsymbol{I}))$，而“ $\cdot$”表示内积算子。

对于实例感知头 K，我们使用一个最大池化层，后接一个全连接层。 K 的输入通道维度等于 E的输出通道维度。 K 的输出通道维度为 256。

对于逐像素头 V，我们使用四个连续的卷积层。每层后面都接一个 ReLU层。在第二个和第三个卷积层之间，我们插入一个双线性插值层，将特征分辨率提高 2 倍。输入通道维度等于 E的输出通道维度。隐藏通道和输出通道均设为 256 维。我们还在第4.5节中比较了掩码解码器的不同设计方案。

指数移动平均（EMA）教师。我们不直接训练教师网络，而是借鉴 MOCO 的做法，利用指数移动平均（EMA）根据任务网络的参数更新教师网络的参数 [54]。使用 EMA 教师的目的是消除训练中的损失爆炸问题，因为优化标准视觉变换器通常并非易事 [13, 14, 16]。移除教师网络后，我们没有观察到基于DeiT-small的MAL性能显著下降或提升。不过，在 MAL 中使用更大规模的图像编码器（例如 ViT-MAE-Base）时，这样做可以使训练更加稳定 [13]。

3.3. 损失

我们使用多实例学习损失 $\mathcal{L}_{\text{mil}}$ 和条件随机场损失 $\mathcal{L}_{\text{crf}}$ 作为框监督损失：

$$\mathcal{L} = \alpha_{\text{mil}}\mathcal{L}_{\text{mil}} + \alpha_{\text{crf}}\mathcal{L}_{\text{crf}} \tag{1}$$

多实例学习损失。多实例分割旨在利用紧致边界框标注所蕴含的先验信息。

学生网络生成输出 $\boldsymbol{m}$后，我们对输出掩码 $\boldsymbol{m}$应用多实例学习（MIL）损失。我们在图3中展示了这一过程。

我们将$\boldsymbol{m}_{i,\ j}$ 定义为图像 $\boldsymbol{I}^c$中位置 i、j 处的掩码分数。在 MIL 损失中，我们将每个像素定义为一个实例。受 BBTP [4], 启发，我们将每行或每列像素视为一个包。根据一个包是否经过真实标注框，我们判定其为正包或负包。我们将这些包定义为 $\boldsymbol{B}$，每个包 $\boldsymbol{B}_i$ 包含一行或一列像素。此外，我们将每个包的标签定义为 $\boldsymbol{g}$，每个标签 $\boldsymbol{g}_i$ 对应一个包 $\boldsymbol{B}_i$。

因此，我们使用最大池化作为归约函数，并采用 Dice 损失 [55]:

$$\mathcal{L}_{mil} = 1 - \frac{2\sum_i \boldsymbol{g}_i \cdot \max\{\boldsymbol{B}_i\}^2}{\sum_i \max\{\boldsymbol{B}_i\}^2 + \sum_i \boldsymbol{g}_i^2} \tag{2}$$

条件随机场损失。CRF 损失旨在通过能量最小化施加平滑先验，从而优化掩码预测。随后，我们利用优化后的掩码作为伪标签，以在线教师的方式对掩码预测进行自训练。我们使用平均掩码预测 $\boldsymbol{m}^a = \frac{1}{2}(\boldsymbol{m} + \boldsymbol{m}^t)$ 作为待优化的掩码预测，_以实现更稳定的训练。

接下来，我们定义一个随机场 $\boldsymbol{X} = \{\boldsymbol{X}_1, ..., \boldsymbol{X}_N\}$，其中 $N = H^c \times W^c$ 是裁剪图像的大小，每个 $\boldsymbol{X}_i$ 表示与 $\boldsymbol{I}^c$ 中某个像素对应的标签，因此我们有 $\boldsymbol{X} \in \{0,1\}^N$，表示背景或前景。我们用 $\boldsymbol{l} \in \{0,1\}^N$ 表示 $\boldsymbol{X}$ 的一个标记，该标记使以下条件随机场能量最小化：

$$E(\boldsymbol{l}|\boldsymbol{m}^a, \boldsymbol{X}^c) = \mu(\boldsymbol{X}|\boldsymbol{m}^a, \boldsymbol{I}^c) + \psi(\boldsymbol{X}|\boldsymbol{I}^c), \tag{3}$$

其中，$\mu(\boldsymbol{X}|\boldsymbol{m}^a,\ \boldsymbol{I}^c)$ 表示一元势，用于对齐 $\boldsymbol{X}_i$ 和 $\boldsymbol{m}_i^a$，因为我们假设大多数掩码预测都是正确的。同时，$\psi(\boldsymbol{X}|\boldsymbol{I}^c)$ 表示成对势，可使优化后的掩码更加清晰。具体而言，我们将成对势定义为：

$$\psi(\boldsymbol{X}|\boldsymbol{I}^c) = \sum_{\substack{i\in\{0..N-1\},\\ j\in\mathcal{N}(i)}} \omega \exp\left(\frac{-|\boldsymbol{I}_i^c - \boldsymbol{I}_j^c|^2}{2\zeta^2}\right)[\boldsymbol{X}_i \neq \boldsymbol{X}_j], \tag{4}$$

其中 $\mathcal{N}(i)$ 表示$\boldsymbol{X}_i$ 的 8 个直接邻居的集合，如图 3所示。随后，我们使用 MeanField 算法 [7, 56] 高效地近似最优解，记作 $\boldsymbol{l} = MeanField(\boldsymbol{I}^c, \boldsymbol{m}^a)$。推导过程和 PyTorch 代码见补充材料。最后

我们应用 Dice损失，利用细化后的掩码 $\boldsymbol{l}$ 对模型进行自训练，具体如下：

$$\mathcal{L}_{crf} = 1 - \frac{2\sum_i \boldsymbol{l}_i \boldsymbol{m}_i}{\sum_i \boldsymbol{l}_i^2 + \boldsymbol{m}_i^2} \tag{5}$$

4. 实验

我们在 COCO 数据集 [1], 和 LVIS [57]上评估 MAL。在 COCO 和 LVIS 上的主要结果如表1和2所示。定性结果见图1 和图5。

4.1. 数据集

COCO 数据集。包含 80 个语义类别。我们遵循标准划分，包括 train2017 （ 115K 张图像）、 val2017 （ 5K 张图像）和 test-dev（20k 张图像）。

LVIS 数据集。包含 1200+ 个类别和 164K 张图像。我们遵循训练集和验证集的标准划分。

4.2. 实现细节

我们使用 8 张 NVIDIA Tesla V100 运行实验。**阶段一（掩码自动标注）**。我们使用 AdamW [58]作为网络优化器，并将两个动量参数设为 0.9、0.9。我们使用余弦退火调度器调整学习率，该学习率设为 $1.5 \cdot 10^{-6}$ 每张图像。多实例学习损失权重 α_{mil}、CRF 损失权重 α_{crf}、ζ和 ω 在 CRF 成对势中的取值分别设为 4、0.5、0.5、2。我们在图 8中分析了损失权重和 CRF 超参数的敏感性。我们使用 512 ×512的输入分辨率，批量大小为 32（每张 GPU 处理 4 个样本）。对于 EMA，我们使用 0.996 的动量。对于任务网络和教师网络，我们采用随机翻转数据增强。此外，我们还对任务网络应用随机颜色抖动、随机灰度转换和随机高斯模糊。我们训练 MAL 共 10 轮。使用标准 ViT-Base 在 COCO 和 LVIS 数据集上训练 MAL 分别约需 23 小时和 35 小时。

阶段二（训练实例分割模型）。我们选取了几种性能优异的全监督实例分割模型：ConvNeXt [44]搭配 Cascade R-CNN [28], Swin Transformer [12] 搭配 Mask2Former [41], ResNets [59] 以及 ResNeXts [60] 搭配 SOLOv2 [31]。MAL 与这些架构配合得非常出色，这从准确率和泛化能力的角度展现了掩码自动标注器的强大能力。阶段二使用 MMDetection [61]代码库。同样，我们只需将真实标注掩码替换为 MAL生成的掩码伪标签，即可将这些全监督模型调整为框监督学习。

4.3. 实例分割结果

保留率。我们认为，仅用实例分割的平均精度均值来评估框监督实例分割并不够公平，因为性能提升可能

方法	标注器骨干网络	实例分割骨干网络	实例分割模型	Sup	(%)掩码 AP_{val}	(%)掩码 AP_{test}	(%)保留率$_{val}$	(%)保留率$_{test}$
Mask R-CNN* [24]	-	ResNet-101	Mask R-CNN	Mask	38.6	38.8	-	-
Mask R-CNN* [24]	-	ResNeXt-101	Mask R-CNN	Mask	39.5	39.9	-	-
CondInst [33]	-	ResNet-101	CondInst	Mask	38.6	39.1	-	-
SOLOv2 [31]	-	ResNet-50	SOLOv2	Mask	37.5	38.4	-	-
SOLOv2 [31]	-	ResNet-101-DCN	SOLOv2	Mask	41.7	41.8	-	-
SOLOv2 [31]	-	ResNeXt-101-DCN	SOLOv2	Mask	42.4	42.7	-	-
ConvNeXt [44]	-	ConvNeXt-Small [44]	Cascade R-CNN	Mask	44.8	45.5	-	-
ConvNeXt [44]	-	ConvNeXt-Base [44]	Cascade R-CNN	Mask	45.4	46.1	-	-
Mask2Former [41]	-	Swin-Small	Mask2Former	Mask	**46.1**	**47.0**	-	-
BBTP† [4]	-	ResNet-101	Mask R-CNN	Box	-	21.1	-	59.1
BoxInst [5]	-	ResNet-101	CondInst	Box	33.0	33.2	85.5	84.9
BoxLevelSet [6]	-	ResNet-101-DCN	SOLOv2	Box	35.0	35.4	83.9	83.5
DiscoBox [7]	-	ResNet-50	SOLOv2	Box	30.7	32.0	81.9	83.3
DiscoBox [7]	-	ResNet-101-DCN	SOLOv2	Box	35.3	35.8	84.7	85.9
DiscoBox [7]	-	ResNeXt-101-DCN	SOLOv2	Box	37.3	37.9	88.0	88.8
BoxTeacher [8]	-	Swin-Base	CondInst	Box	-	40.0	-	-
掩码自动标注器	ViT-MAE-Base [13]	ResNet-50	SOLOv2	Box	35.0	35.7	93.3	93.0
掩码自动标注器	ViT-MAE-Base [13]	ResNet-101-DCN	SOLOv2	Box	38.2	38.7	91.6	92.6
掩码自动标注器	ViT-MAE-Base [13]	ResNeXt-101-DCN	SOLOv2	Box	38.9	39.1	91.7	91.6
掩码自动标注器	ViT-MAE-Base [13]	ConvNeXt-Small [44]	Cascade R-CNN	Box	42.3	43.0	94.4	**94.5**
掩码自动标注器	ViT-MAE-Base [13]	ConvNeXt-Base [44]	Cascade R-CNN	Box	42.9	43.3	**94.5**	93.9
掩码自动标注器	ViT-MAE-Base [13]	Swin-Small [12]	Mask2Former [41]	Box	**43.3**	**44.1**	93.9	93.8

表 1. COCO 上的主要结果。Ret 表示边界框监督掩码平均精度监督掩码 AP 的保留率。使用 SOLOv2/ResNeXt-101 的 MAL 在 val2017 上比使用 SOLOv2/ResNeXt-101 的 DiscoBox 高 1.6%，在 test-dev 上高 1.3%。我们的最佳模型（Mask2former/Swin-Small）在 val 上达到 43.3% AP，在 test-dev 上达到 44.1% AP。

来自与分割质量无关的边界框质量提升。然而，保留率能更好地反映真实的掩码质量，因为全监督方法也会受益于更好的边界框结果。

COCO 上的结果。在表1中，我们展示了多种现代实例分割模型使用全监督预言模型的伪标签后，性能最高可达 94.5%。我们的最佳结果是在 COCO test-dev 上达到 43.3% mAP，在 COCO val 上达到 44.1% mAP；具体做法是阶段一使用 MAL（由 MAE 预训练的标准 ViT-Base [11] ），阶段二使用 Mask2Former（Swin-Small） [12,41] 。使用掩码伪标签训练更强大的实例分割模型时，保留率并未显著下降。相反，较重型的实例分割模型在 COCO 上取得了更高的保留率，例如使用 ConvNeXt 的 Cascade R-CNN 和使用 Swin-Small 的 Mask2Former。然而，与 MAL 相比，其他方法的保留率显著较低。实验结果定量表明，掩码质量远远优于其他方法。**LVIS 上的结果**。在表2中，我们还观察到，所有实例分割模型都能很好地利用 MAL 生成的掩码伪标签（保留率 $=93\%\sim98\%$）。我们在图5中展示了部分结果。我们还通过在 COCO 数据集上训练 MAL、在 LVIS 上生成掩码伪标签，并使用这些伪标签训练实例分割模型，评估了 MAL 的开放词汇能力。

4.4. 图像编码器变体

为支持我们关于视觉变换器是优秀自动标注器的主张，我们比较了三种常见网络作为 MAL 的图像编码器：标准视觉变换器 [11, 13, 16], Swin Transformer [12], ConvNeXt [44] 在表 4。

首先，我们比较这三种模型经过全监督预训练的权重。我们选用 ConvNeXt 和 Swin Transformer 的官方全监督预训练权重。对于标准视觉变换器，我们采用一种常见的全监督方法 DeiT [16]。我们观察到，作为掩码自动标注器图像编码器的全监督标准视觉变换器（DeiT）优于 Swin Transformer 和 ConvNeXt，尽管 Swin Transformer 和 Conv NeXt 在 imagenet-1k 上的表现优于 DeiT。我们认为，标准视觉变换器的成功可能归功于其自发涌现特性 [9, 11] （如图 6所示），以及全局多头自注意力层带来的更大感受野。

其次，通过掩码自编码器（MAE）预训练，可以进一步改进掩码伪标签 [13]。潜在原因可能是，MAE 预训练通过学习像素级信息增强了标准 ViT，而这对于分割等密集预测任务非常重要。

4.5. 掩码解码器变体

我们在表 [62], 3中比较了四种不同的现代掩码解码器设计：全连接解码器 [24,63], 、全卷积解码器 [30,31],、基于注意力的解码器 [41] 和基于查询的解码器。我们在图4中展示了不同的掩码解码器设计。对于全连接解码器，我们使用两个隐藏维度为 2048 的全连接层，然后为每个像素输出一张置信度图。我们将该输出向量重塑为二维置信度图。我们在第3.2节介绍基于注意力的解码器。对于全卷积解码器，我们采用基于注意力的解码器中的逐像素头 V。



图 5. 掩码自动标注器在 LVIS v1 上生成的掩码伪标签的定性结果。

方法	自动标注器骨干网络	实例分割骨干网络	实例分割模型	训练数据	Sup	(%) 掩码 AP_{val}	(%) 保留率$_{val}$
Mask R-CNN [24]	-	ResNet-50-DCN	Mask R-CNN [24]	-	Mask	21.7	-
Mask R-CNN [24]	-	ResNet-101-DCN	Mask R-CNN [24]	-	Mask	23.6	-
Mask R-CNN [24]	-	ResNeXt-101-32x4d-FPN	Mask R-CNN [24]	-	Mask	25.5	-
Mask R-CNN [24]	-	ResNeXt-101-64x4d-FPN	Mask R-CNN [24]	-	Mask	25.8	-
掩码自动标注器	ViT-MAE-Base [13]	ResNet-50-DCN	Mask R-CNN [24]	LVIS v1	Box	20.7	95.4
掩码自动标注器	ViT-MAE-Base [13]	ResNet-101-DCN	Mask R-CNN [24]	LVIS v1	Box	23.0	**97.4**
掩码自动标注器	ViT-MAE-Base [13]	ResNeXt-101-32x4d-FPN	Mask R-CNN [24]	LVIS v1	Box	23.7	92.9
掩码自动标注器	ViT-MAE-Base [13]	ResNeXt-101-64x4d-FPN	Mask R-CNN [24]	LVIS v1	Box	**24.5**	95.0
掩码自动标注器	ViT-MAE-Base [13]	ResNeXt-101-32x4d-FPN	Mask R-CNN [24]	COCO	Box	23.3	91.8
掩码自动标注器	ViT-MAE-Base [13]	ResNeXt-101-64x4d-FPN	Mask R-CNN [24]	COCO	Box	**24.2**	**93.8**

表 2. LVIS v1 上的主要结果。训练数据指我们用于训练 MAL 的数据集。我们还在 COCO 上对其进行微调，然后生成 LVIS v1 的伪标签。与直接在 LVIS v1 上训练相比，仅在 COCO 上微调的 MAL 在最终结果中的平均精度均值下降了约 0.35%，这表明 MAL 的开放集能力具有巨大潜力。Ret 表示边界框监督掩码平均精度监督掩码 AP 的保留率。

掩码解码器	(%)掩码 AP_{val}	(%)保留率$_{val}$
全连接解码器	35.5	79.2
全卷积解码器	36.1	80.5
基于注意力的解码器	**42.3**	**94.4**
基于查询的解码器	-	-

表3. 框扩展的消融研究。阶段一中，我们使用 Standard ViT- 掩码自编码器-Base 作为 MAL 的图像编码器；阶段二中，我们使用以 ConvNext-Small 为骨干网络的 Cascade RCNN 作为实例分割模型。结果以 % 掩码平均精度均值表示。在不同设计中，基于注意力的解码器表现最佳。使用基于查询的解码器无法获得合理的结果。

骨干网络	IN-1k Top-1准确率	掩码 AP_{val}	保留率$_{val}$
ConvNeXt-Base [44]	**83.8**	39.6	88.4
Swin-Base [12]	83.5	40.2	89.7
ViT-DeiT-Small [64]	79.9	40.8	91.0
ViT-DeiT-Base [64]	81.8	**41.1**	**91.7**
ViT-MAE-Base [13]	83.6	**42.3**	**94.4**
ViT-MAE-Large [13]	85.9	**42.3**	**94.4**

表 4. 不同骨干网络的消融研究。所有模型均在 ImageNet-1k 上进行预训练。ConvNeXt 和 Swin Transformer 在图像分类任务上的表现优于 DeiT，但标准 ViT-Small [16]（ViT-DeiT-Small）在掩码自动标注任务上的表现优于 ConvNeXt-base 和 Swin-Base。通过 MAE 预训练的标准 ViT-Base（ViT-MAE-Base）和标准 ViT-Large（ViT-MAE-Large）在掩码自动标注任务上取得了最佳表现。

对于基于查询的解码器，我们遵循 Mask2Former [41] 中的设计。由于基于查询的解码器在全监督实例分割中表现极佳，我们花了大量精力探索其在 MAL 上的应用。然而，结果却出乎意料地不理想。我们怀疑，在框监督损失下，稍重一些的层可能会导致优化问题。

实验表明，框监督实例分割更适合采用基于注意力的解码器。然而，最先进的实例分割和目标检测方法通常采用全卷积解码器 [15, 43]或基于查询的解码器 [41]。我们提出的两阶段框架解决了这一矛盾，使网络能够兼具基于注意力的解码器和非注意力解码器的优势。

4.6. 聚类分析

如表 4所示，我们想知道为什么标准视觉 Transformer在自动标注方面优于其他现代图像编码器。由于分类能力的比较似乎无法反映自动标注的实际能力，我们尝试使用能力聚类来评估图像编码器，因为前景（前景）/背景分割与二元聚类问题非常相似。

具体而言，我们提取 Swin Transformer [12], ConvNeXt [44], 标准 ViT [11]最后几层输出的特征图。随后，我们使用真实标注掩码将特征向量划分为前景和背景特征集。通过评估前景/背景特征向量到其聚类中心的平均距离，我们可以从经验上揭示网络区分前景和背景像素的能力。

形式上，我们将骨干网络 E 生成的 token i 的特征向量定义为 $\boldsymbol{f}_i^E$。我们将前景/背景聚类中心 $\boldsymbol{f}'_1$, $\boldsymbol{f}'_0$ 定义为前景/背景特征向量的均值。然后，我们使用以下指标作为聚类得分：

$$S = \frac{1}{N}\sum_{i}^{N}\left(\frac{\boldsymbol{f}_i^E}{|\boldsymbol{f}_i^E|} - \frac{\boldsymbol{f}'_{\gamma(i)}}{|\boldsymbol{f}'_{\gamma(i)}|}\right)^2, \tag{6}$$

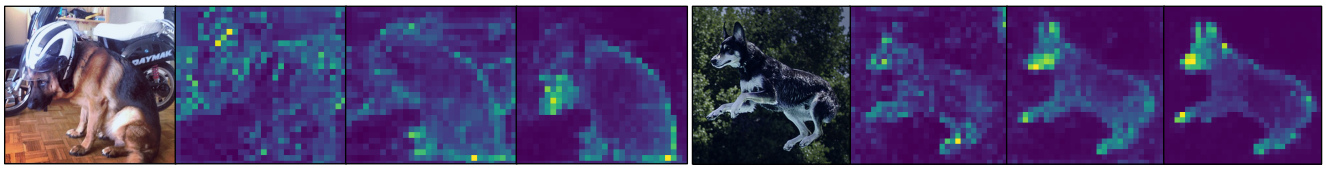
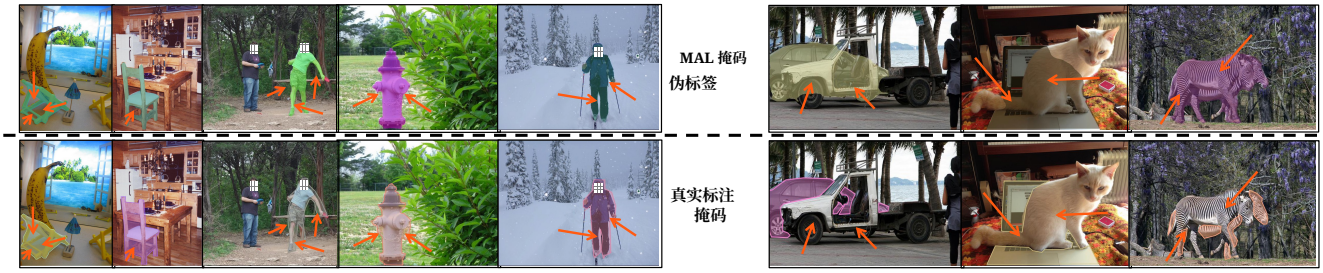


图 6. MAL 生成的两个 RoI 图像的注意力可视化。在每组图像中，最左侧的图像是原始图像。我们将 MAL 中标准视觉 Transformer 的 4^{th}, 8^{th}, 12^{th} MHSA层输出的注意力图可视化。



(a) MAL 生成的掩码更清晰，边界贴合度更高　　(b) 遮挡问题

图 7. COCO val2017 上 MAL 生成的伪标签（上方）与真实标注掩码（下方）的并排比较。左侧可以看到，在某些情况下，MAL 生成的伪标签比真实标注掩码边缘更清晰、贴合度更高。右侧可以看到，在遮挡严重的情况下，人工标注掩码仍然更好。

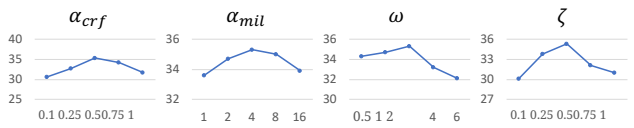


图 8. 损失权重和条件随机场超参数的敏感性分析。第一阶段使用经掩码自编码器 [11] 预训练的 ViT-Base [13] 作为图像编码器，第二阶段使用 SOLOv2 (ResNet-50)。x 轴和 y 轴分别表示超参数值和（%）掩码 AP。

θ	掩码 AP_{val}	保留率。val
0.6	41.3	92.2
0.8	41.7	93.1
1.0	42.2	94.2
1.2	**42.3**	**94.4**
1.4	42.0	93.8
1.6	41.8	93.3

表 5. 边界框扩展比例的消融实验。阶段一和阶段二分别使用通过掩码自编码器（ViT-MAE-Base）预训练的标准 ViT-Base 和 Cascade R-CNN（ConvNeXt-Small）。

骨干网络	得分（↓）
ConvNeXt-Base [44]	0.459
Swin-Base [12]	0.425
ViT-DeiT-Small [64]	0.431
ViT-DeiT-Base [64]	0.398
ViT-MAE-Base [13]	0.324
ViT-MAE-Large [13]	**0.301**

表 6. 不同图像编码器的聚类得分。聚类得分越低，区分前景和背景特征的能力越强。

其中，如果像素 i 是前景，则为 $\gamma(i) = 1$；否则为 $\gamma(i) = 0$。

我们在表 6 中展示了 COCO 2017 年验证集上的聚类评估。结果与我们的结论一致：标准视觉变换器更擅长掩码自动标注。

4.7. MAL 掩码与真实标注掩码

我们在图中展示了苹果对苹果的定性比较。7并得出以下观察结果。首先，MAL 生成的掩码伪标签比人工标注的标签边缘更清晰，也更贴合边界，因为人类难以与真实边界对齐。其次，严重遮挡也是一个棘手的问题。

5. 结论

在这项工作中，我们提出了一种用于框监督实例分割的新型两阶段框架，以及一种新颖的基于 Transformer 的架构——掩码自动标注器（MAL），用于在阶段一生成高质量的掩码伪标签。我们揭示了标准视觉变换器是出色的掩码自动标注器。此外，我们发现，随机使用框扩展 RoI 输入、采用基于注意力的解码器以及进行类别无关训练，对于实现强大的掩码自动标注性能至关重要。此外，得益于两阶段框架设计和 MAL，我们几乎可以将所有类型的全监督实例分割模型调整为框监督学习，且性能下降很小，这体现了 MAL 出色的泛化能力。

局限性。 尽管我们的方法显著提升了掩码自动标注效果，但在遮挡情况下仍有许多失败案例，此时人工标注明显优于 MAL 生成的掩码。此外，将模型从标准 ViT-Base 扩展到标准 ViT-Large 时，我们遇到了性能饱和问题。我们将这些问题留待未来工作解决。

更广泛的影响。 我们提出的基于 Transformer 的掩码自动标注器和两阶段架构，为高质量的框监督实例分割提供了一种标准范式。如果后续工作能够在我们提出的范式下发现并解决相关问题，那么未来实例分割很有可能不再需要大量人工标注的掩码。

参考文献

[1] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar 和 C Lawrence Zitnick。Microsoft COCO：情境中的常见物体。在 欧洲计算机视觉会议，第 740–755 页。Springer，2014 年。1，5

[2] Alina Kuznetsova、Hassan Rom、Neil Alldrin、Jasper Ui-jlings、Ivan Krasin、Jordi Pont-Tuset、Shahab Kamali、Stefan Popov、Matteo Malloci、Alexander Kolesnikov、Tom Duerig 和 Vittorio Ferrari。《Open Images 数据集 V4：大规模统一图像分类、目标检测和视觉关系检测》。国际计算机视觉杂志，2020年。1

[3] Shuai Shao、Zeming Li、Tianyuan Zhang、Chao Peng、Gang Yu、Xiangyu Zhang、Jing Li 和 Jian Sun。《Objects365：用于目标检测的大规模高质量数据集》。发表于*IEEE/CVF*国际计算机视觉会议论文集，第8430–8439页，2019年。1

[4] Cheng-Chun Hsu、Kuang-Jui Hsu、Chung-Chi Tsai、Yen-Yu Lin 和 Yung-Yu Chuang。《利用边界框紧致度先验进行弱监督实例分割》。神经信息处理系统进展，第32卷，2019年。1，3，4，5，6

[5] Zhi Tian、Chunhua Shen、Xinlong Wang 和 Hao Chen Boxinst：基于框标注的高性能实例分割。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第5443–5452 页，2021年。1，3，4，6

[6] Wentong Li、Wenyu Liu、Jianke Zhu、Miaomiao Cui、Xi- ansheng Hua 和 Lei Zhang。基于水平集演化的框监督实例分割。*arXiv*预印本 *arXiv:2207.09055*，2022年。1，3，4，6

[7] Shiyi Lan、Zhiding Yu、Christopher Choy、Subhashree Rad- hakrishnan、Guilin Liu、Yuke Zhu、Larry S Davis 和 An- ima Anandkumar。DiscoBox：通过框监督实现弱监督实例分割和语义对应。发表于 *IEEE/CVF*国际计算机视觉会议论文集，第 3406–3416 页，2021年。1，3，4，5，6

[8] Tianheng Cheng、Xinggang Wang、Shaoyu Chen、Qian Zhang 和 Wenyu Liu。Boxteacher：探索用于弱监督实例分割的高质量伪标签。*arXiv* 预印本 *arXiv:2210.05174*，2022年。1，3，6

[9] Mathilde Caron、Hugo Touvron、Ishan Misra、Herv´e J´egou、Julien Mairal、Piotr Bojanowski 和 Armand Joulin。自监督视觉变换器中涌现出的特性。载于*IEEE/CVF*国际计算机视觉会议论文集，第 9650–9660 页，2021年。2，3，6

[10] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszko- reit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。Attention Is All You Need。神经信息处理系统进展，第 30 卷，2017年。2

[11] Alexey Dosovitskiy、Lucas Beyer、AlexanderKolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner, Mostafa Dehghani, Matthias Mind$_{erer}$, Georg Heigold, Syl - vain Gelly 等。《一张图像相当于 16×16 个词：用于大规模图像识别的 Transformer》。arXiv 预印本 arXiv:2010.11929，2020年。2、4、5、6、7、8、12[12] Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin 和 Baining Guo。《Swin Transformer：使用移位窗口的层次化视觉Transformer》。载于 IEEE/CVF国际计算机视觉会议论文集，第 10012–10022 页，2021年。2、3、5、6、7、8、12[13] Kaiming He、Xinlei Chen、Saining Xie、Yanghao Li、Piotr Doll´ar 和 Ross Girshick。《掩码自编码器是可扩展视觉学习器》。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 16000–16009 页，2022年。2、4、6、7、8、12

[14] Hangbo Bao、Li Dong、Furu Wei。Beit：用于图像变换器的 BERT 预训练。*arXiv* 预印本 *arXiv:2106 08254*，2021年2，4。。

[15] Wenhui Wang、Hangbo Bao、Li Dong、Johan Bjorck、Zhil- iang Peng、Qiang Liu、Kriti Aggarwal、Owais Khan Mo- hammed、Saksham Singhal、Subhojit Som 等。《图像作为外语：面向所有视觉及视觉语言任务的 Beit 预训练》。*arXiv* 预印本 *arXiv:2208.10442*，2022年。2，7

[16] Hugo Touvron、Matthieu Cord、Matthijs Douze、Francisco Massa、Alexandre Sablayrolles 和 Herv´eJ´egou。《训练数据高效的图像变换器：通过注意力机制进行蒸馏》。发表于 国际机器学习会议，第 10347–10357 页。PMLR，2021年。2，4，6，7，12

[17] Hugo Touvron、Matthieu Cord 和 Alexandre Sablayrolles。Deit iii：Revenge of the vit 的复仇。*arXiv* 预印本 *arXiv:2204.07118*，2022年。2

[18] Hugo Touvron、Matthieu Cord、Alexandre Sablayrolles、Gabriel Synnaeve 和 Herv´eJ´egou。使用图像变换器深入探索。在*IEEE/CVF*国际计算机视觉会议（*ICCV*）论文集，第32–42页，2021年10月。2

[19] Lingchen Meng、Hengduo Li、Bor-Chun Chen、Shiyi Lan、Zuxuan Wu、Yu-Gang Jiang 和 Ser-Nam Lim。AdaViT：用于高效图像识别的自适应视觉变换器。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第12309–12318页，2022年。2

[20] Wenhai Wang、Enze Xie、Xiang Li、Deng-Ping Fan、Kaitao Song、Ding Liang、Tong Lu、Ping Luo 和 Ling Shao。金字塔视觉变换器：一种无需卷积、适用于密集预测的通用骨干网络。发表于*IEEE/CVF*国际计算机视觉会议论文集，第568–578页，2021年。2

[21] Wenhai Wang、Enze Xie、Xiang Li、Deng-Ping Fan、Kaitao Song、Ding Liang、Tong Lu、Ping Luo 和 Ling Shao。PVT V2：采用金字塔视觉变换器的改进基线。《计算视觉媒体》，8(3)：415–424，2022年。2

[22] Ze Liu、Han Hu、Yutong Lin、Zhuliang Yao、Zhenda Xie、Yixuan Wei、Jia Ning、Yue Cao、Zheng Zhang、Li Dong 等人。

Swin Transformer V2：扩展容量与分辨率 . 发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第12009–12019页，2022年。2，3

[23] Daquan Zhou、Bingyi Kang、Xiaojie Jin、Linjie Yang、Xiaochen Lian、Zihang Jiang、Qibin Hou 和 Jiashi Feng。DeepViT：迈向更深层的视觉Transformer。*arXiv*预印本 *arXiv:2103.11886*，2021年。2

[24] Kaiming He、Georgia Gkioxari、Piotr Doll´ar 和 Ross Gir-shick。Mask R-CNN。发表于*IEEE*国际计算机视觉会议论文集，第2961–2969页，2017年。2，3，6，7，12

[25] Yi Li、Haozhi Qi、Jifeng Dai、Xiangyang Ji 和 Yichen Wei . 全卷积实例感知语义分割。发表于*IEEE*计算机视觉与模式识别会议论文集，第2359–2367页，2017年。2

[26] Ke Xu、Kaiyu Guan、Jian Peng、Yunan Luo 和 Sibo Wang。DeepMask：一种使用深度残差网络检测光学卫星遥感图像中云及云影的算法。arXiv 预印本 arXiv:1911.03607，2019年。2[27] Hexiang Hu、Shiyi Lan、Yuning Jiang、Zhimin Cao 和 Fei Sha。Fast Mask：一次性分割多尺度候选目标。载于 IEEE计算机视觉与模式识别会议论文集，第991–999页，2017年。2[28] Zhaowei Cai 和 Nuno Vasconcelos。Cascade R-CNN：深入探索高质量目标检测。载于 IEEE计算机视觉与模式识别会议论文集，第6154–6162页，2018年。2, 3, 5[29] Kai Chen、Jiangmiao Pang、Jiaqi Wang、Yu Xiong、Xiaoxiao Li、Shuyang Sun、Wansen Feng、Ziwei Liu、Jianping Shi、Wanli Ouyang 等。用于实例分割的混合任务级联。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第4974–4983页，2019年。2[30] Daniel Bolya、Chong Zhou、Fanyi Xiao 和 Yong Jae Lee。YOLACT：实时实例分割。载于 IEEE/CVF国际计算机视觉会议论文集，第9157–9166页，2019年。2, 4, 6[31] Xinlong Wang、Rufeng Zhang、Tao Kong、Lei Li 和 Chunhua Shen。SOLOv2：动态快速的实例分割。神经信息处理系统进展，33：17721–17732，2020年。2, 5, 6[32] Xinlong Wang、Rufeng Zhang、Chunhua Shen、Tao Kong 和 Lei Li。SOLO：一种简单的实例分割框架。IEEE Transactions on Pattern Analysis and Machine Intelligence，2021年。2[33] Zhi Tian、Chunhua Shen 和 Hao Chen。用于实例分割的条件卷积。载于欧洲计算机视觉会议论文集，第282–298页。Springer，2020年。2, 6[34] Enze Xie、Peize Sun、Xiaoge Song、Wenhai Wang、Xuebo Liu、Ding Liang、Chunhua Shen 和 Ping Luo。PolarMask：采用极坐标表示的单阶段实例分割。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第12193–12202页，2020年。2

[35] Enze Xie、Wenhai Wang、Mingyu Ding、Ruimao Zhang 和 Ping Luo。PolarMask++：用于单阶段实例分割及更多任务的增强型极坐标表示。*IEEE Transactions on Pattern Analysis and Machine Intelligence*，2021年。2

[36] Nicolas Carion、Francisco Massa、Gabriel Synnaeve、Nicolas Usunier、Alexander Kirillov 和 Sergey Zagoruyko。使用变换器进行端到端目标检测。载于 欧洲计算机视觉会议，第 213–229 页。Springer，2020年。 2

[37] Xizhou Zhu、Weijie Su、Lewei Lu、Bin Li、Xiaogang Wang 和 Jifeng Dai。《可变形 DETR：用于端到端目标检测的可变形变换器》。载于《国际学习表征会议论文集》，2021年。2[38] Huiyu Wang、Yukun Zhu、Hartwig Adam、Alan Yuille 和 Liang-Chieh Chen。《Max-deeplab：利用掩码变换器进行端到端全景分割》。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第5463–5474页，2021年。2[39] Bowen Cheng、Alex Schwing 和 Alexander Kirillov。《像素级分类并非语义分割所需的一切》。神经信息处理系统进展，34：17864–17875，2021年。2[40] Zhiqi Li、Wenhai Wang、Enze Xie、Zhiding Yu、Anima Anandkumar、Jose M Alvarez、Ping Luo 和 Tong Lu。《全景 SegFormer：深入探索利用变换器进行全景分割》。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第1280–1289页，2022年。3 [41] Bowen Cheng、Anwesa Choudhuri、Ishan Misra、Alexander Kirillov、Rohit Girdhar 和 Alexander G Schwing。《用于视频实例分割的 Mask2Former》。arXiv预印本 arXiv:2112.10764，2021年。3, 5, 6, 7[42] Feng Li、Hao Zhang、Shilong Liu、Lei Zhang、Lionel M Ni、Heung-Yeung Shum 等。《Mask DINO：迈向用于目标检测与分割的统一基于变换器的框架》。arXiv预印本 arXiv:2206.02777，2022年。3[43] Yanghao Li、Hanzi Mao、Ross Girshick 和 Kaiming He。《探索用于目标检测的普通视觉变换器骨干网络》。arXiv预印本 arXiv:2203.16527，2022年。3, 4, 7, 12[44] Zhuang Liu、Hanzi Mao、Chao-Yuan Wu、Christoph Feichtenhofer、Trevor Darrell 和 Saining Xie。《面向2020年代的卷积网络》。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第11976–11986页，2022年。3, 5, 6, 7, 8, 12, 13 – 。[45] Thibaut Durand、Taylor Mordan、Nicolas Thome 和 Matthieu Cord。《Wildcat：用于图像分类、点级定位与分割的深度卷积网络弱监督学习》。载于《IEEE计算机视觉与模式识别会议论文集》，第642–651页，2017年。3[46] Bin Jin、Maria V Ortiz Segovia 和 Sabine Susstrunk。《弱监督语义分割》。载于《IEEE计算机视觉与模式识别会议论文集》，第3626–3635页，2017年。3

[47] Jiwoon Ahn 和 Suha Kwak。学习像素级语义亲和性，以实现弱监督语义分割。载于 *IEEE*计算机视觉与模式识别会议论文集，第 4981–4990 页，2018年。 3

[48] Ruochen Fan、Qibin Hou、Ming-Ming Cheng、Gang Yu、Ralph R Martin 和 Shi-Min Hu。关联图像间显著实例，以实现弱监督语义分割。载于 欧洲计算机视觉会议（*ECCV*）论文集，第 367–383 页，2018年。 3

[49] Guolei Sun、Wenguan Wang、Jifeng Dai 和 Luc Van Gool。挖掘跨图像语义，以实现弱监督语义分割。载于 欧洲计算机视觉会议，第 347–365 页。Springer，2020年。 3

[50] Bolei Zhou、Aditya Khosla、Agata Lapedriza、Aude Oliva 和 Antonio Torralba。学习用于判别性定位的深度特征。载于 *IEEE*计算机视觉与模式识别会议论文集，第 2921–2929 页，2016。

[51] Ramprasaath R Selvaraju、Michael Cogswell、Abhishek Das、Ramakrishna Vedantam、Devi Parikh 和 Dhruv Batra。Grad-CAM：通过基于梯度的定位，对深度网络进行视觉解释。收录于 *IEEE*计算机视觉国际会议论文集，第 618–626 页，2017年。

[52] Daquan Zhou、Zhiding Yu、Enze Xie、Chaowei Xiao、An- imashree Anandkumar、Jiashi Feng 和 Jose M Alvarez。理解视觉变换器的鲁棒性。收录于 国际机器学习会议，第 27378–27394 页。PMLR，2022年。 3

[53] Tsung-Yi Lin、Piotr Doll´ar、Ross Girshick、Kai ming He、Bharath Hariharan 和 Serge Belongie。用于目标检测的特征金字塔网络。收录于 *IEEE*计算机视觉与模式识别会议论文集，第 2117–2125 页，2017年。 4，12，13

[54] Kaiming He、Haoqi Fan、Yuxin Wu、Saining Xie 和 Ross Girshick。用于无监督视觉表征学习的动量对比。载于 *IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*）论文集，2020年6月。 4

[55] Carole H Sudre、Wenqi Li、Tom Vercauteren、Sebastien Ourselin 和 M Jorge Cardoso。广义Dice重叠：一种用于高度不平衡分割的深度学习损失函数。载于 医学图像分析中的深度学习与临床决策支持多模态学习，第240–248页。Springer，2017年。 5

[56] Philipp Krähenbühl 和 Vladlen Koltun。利用高斯边势在全连接条件随机场中进行高效推理。神经信息处理系统进展，第 24 卷，2011 年。5， 12

[57] Agrim Gupta、Piotr Dollar 和 Ross Girshick。LVIS：用于大词汇表实例分割的数据集。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 5356–5364 页，2019年。 5

[58] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。*arXiv* 预印本 *arXiv:1711.05101*，2017 年。 5

[59] Kaiming He、Xiangyu Zhang、任少卿和 Jian Sun。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第770–778页，2016年。5，13

[60] Saining Xie、Ross Girshick、Piotr Doll´ar、Zhuowen Tu 和 Kaiming He。面向深度神经网络的聚合残差变换。载于*IEEE*计算机视觉与模式识别会议论文集，第1492–1500页，2017年5。

[61]Kai Chen、Jiaqi Wang、Jiangmiao Pang、曹宇航、Yu Xiong、Xiaoxiao Li、Shuyang Sun、Wansen Feng、Ziwei Liu、徐嘉瑞、Zheng Zhang、程大志、朱晨晨、Tianheng Cheng、赵启杰、李步宇、陆鑫、朱锐、吴越、Jifeng Dai、王井东、Jianping Shi、Wanli Ouyang、陈昌来和林达华。MMDetection：开放式 mmlab 检测工具箱和基准测试。*arXiv* 预印本 *arXiv:1906.07155*，2019年。5

[62] Jifeng Dai、Kaiming He 和 Jian Sun。通过多任务网络级联实现实例感知语义分割。在*IEEE*计算机视觉与模式识别会议论文集，第 3150–3158 页，2016 年。6

[63] Jonathan Long、Evan Shelhamer 和 Trevor Darrell。用于语义分割的全卷积网络。在*IEEE*计算机视觉与模式识别会议论文集，第 3431–3440 页，2015 年。6

[64] Hugo Touvron、Matthieu Cord、Matthijs Douze、Francisco Massa、Alexandre Sablayrolles 和 Herve Jegou。《训练数据高效的图像变换器：通过注意力进行蒸馏》。载于国际机器学习会议，第 139 卷，第 10347–10357 页，2021年 7 月。7，8

A. 附录

A.1. 条件随机场的更多细节

在论文正文中，我们定义了条件随机场（CRF）的能量项，但省略了如何使用均值场算法来最小化能量的细节。这里，我们将更详细地介绍如何使用均值场算法 [56]。

我们将 $\boldsymbol{l} = \{\boldsymbol{l}_1, ..., \boldsymbol{l}_N\}$ 定义为待推断的标签，其中 $N = H \times W$ 是输入图像的尺寸，$\boldsymbol{x}_i$是 i中第 $\boldsymbol{I}$个像素的标签。我们还假设网络预测的掩码 $\boldsymbol{m} = \{\boldsymbol{m}_1, ..., \boldsymbol{m}_N\}$ 满足 $\boldsymbol{m}_i$是 i中第 $\boldsymbol{I}$个像素的一元掩码分数。使用均值场算法求得 $\boldsymbol{l}$ 的伪代码见算法1：

算法 1条件随机场的均值场算法。

过程 MEANFIELD($\boldsymbol{m}, \boldsymbol{I}$)
  $\boldsymbol{K}_{i,j} \leftarrow \omega \exp(-\frac{|\boldsymbol{I}_i - \boldsymbol{I}_j|}{2\zeta^2})$
    ▷ 初始化高斯核
  $\boldsymbol{l} \leftarrow \boldsymbol{m}$ ▷ 初始化 $\boldsymbol{l}$ 使用 $\boldsymbol{m}$
  当尚未收敛 **时** ▷ 迭代直至收敛
    对于 $i \leftarrow 1$**到** $|\boldsymbol{l}|$ **执行**
      $\hat{\boldsymbol{l}}_i \leftarrow \boldsymbol{l}_i$
      对于 $j \in \mathcal{N}(i)$ **执行**
        $\hat{\boldsymbol{l}}_i \leftarrow \hat{\boldsymbol{l}}_i + \boldsymbol{K}_j * \boldsymbol{l}_j$
        ▷ 消息传递
      循环结束
    循环结束
    $\boldsymbol{l} \leftarrow \varphi(\hat{\boldsymbol{l}})$ ▷ φ是一个钳位函数
  while 循环结束
  返回 $\lambda(\boldsymbol{l})$ ▷ λ 是一个阈值函数
结束过程

A.2. 更多实现细节

对于所有基准测试中的所有图像编码器（标准 ViT [11, 13, 16], 、Swin Transformer [12], 和 Conv NeXt [44]）以及掩码解码器（全连接解码器、全卷积解码器、基于注意力的解码器），我们都使用相同的超参数，包括批量大小和优化超参数。当我们在图像编码器（标准 ViT、Swin Transformer、ConvNeXt）与掩码解码器（基于注意力的解码器）之间加入参数化层或多尺度横向/跳跃连接 [43, 53] 时，我们观察到性能下降。我们插入若干双线性插值层，以调整图像编码器与掩码解码器之间特征图的大小，并调整分割得分图的大小。具体而言，我们根据物体的大小，将图像编码器生成的特征图调整为原始输入的 1/16（小）、1/8（中）或 1/4（大）尺寸。我们根据边界框的面积将物体分为三个尺度。我们使用以下面积范围 $[0, 32^2)$、$[32^2, 96^2)$、$[96^2, \infty)$ 来涵盖分别对应小型、中型和大型物体。我们将掩码预测图调整为 512×512，使其达到输入图像的原始分辨率。

此外，我们还尝试了三种简单的方法来添加分类损失，但它们与 MAL 配合时效果不佳。首先，我们添加另一个全连接层作为分类解码器，该层以实例感知头 K 中第一个全连接层的特征图为输入。采用这种设计后，分类会导致性能显著下降。其次，我们使用两个额外的全连接层，或使用标准视觉Transformer的原始分类解码器作为分类解码器，直接以图像编码器的特征图为输入。然而，在这种情况下，分类损失既没有带来性能提升，也没有造成性能损失。

A.3. 目标检测的优势

有监督目标检测模型能够从额外的掩码监督中受益 [24], ，从而提升检测结果。具体而言，我们遵循 Mask R-CNN 中的设置 [24]。首先，我们使用 RoI Align、边界框分支和边界框监督，但不使用掩码监督。其次，在第一个基线的基础上，我们加入掩码分支和真实掩码监督。第二个基线即原始的 Mask R-CNN。第三，我们在第二个基线的基础上，将真实标注掩码替换为 MAL 生成的掩码伪标签。结果表明，使用 MAL生成的掩码伪标签进行掩码监督，在检测方面带来的提升与使用真实标注掩码相近。结果见表7。

A.4. 更多定性结果

我们还在图中展示了使用真实标注掩码和掩码伪标签训练的实例分割模型所产生的预测结果。9。在大多数情况下，我们认为人们无法分辨哪些结果是由使用人工标注标签进行监督的模型生成的。

实例分割骨干网络	数据集	掩码标签	(%)平均精度	(%)平均精度$_{50}$	(%)平均精度$_{75}$	(%)平均精度$_{S}$	(%)平均精度$_{M}$	(%)平均精度$_{L}$
ResNet-50-DCN [59]	LVIS v1	None	22.0	36.4	22.9	16.8	29.1	33.4
ResNet-50-DCN [59]	LVIS v1	真实标注掩码	22.5	36.9	23.8	16.8	29.7	35.0
ResNet-50-DCN [59]	LVIS v1	MAL 掩码	22.6	37.2	23.8	17.3	29.8	34.6
ResNet-101-DCN [59]	LVIS v1	None	24.4	39.5	26.1	17.9	32.2	36.7
ResNet-101-DCN [59]	LVIS v1	真实标注掩码	24.6	39.7	26.1	18.3	32.1	38.3
ResNet-101-DCN [59]	LVIS v1	MAL 掩码	25.1	40.0	26.7	18.4	32.5	37.8
ResNeXt-101-32x4d-FPN [53,59]	LVIS v1	None	25.5	41.0	27.1	18.8	33.7	38.0
ResNeXt-101-32x4d-FPN [53,59]	LVIS v1	真实标注掩码	26.7	42.1	28.6	19.7	34.7	39.4
ResNeXt-101-32x4d-FPN [53,59]	LVIS v1	MAL 掩码	26.3	41.5	28.3	19.5	34.5	39.6
ResNeXt-101-64x4d-FPN [53,59]	LVIS v1	None	26.6	42.0	28.3	19.8	34.7	39.9
ResNeXt-101-64x4d-FPN [53,59]	LVIS v1	真实标注掩码	27.2	42.8	29.2	20.2	35.7	41.0
ResNeXt-101-64x4d-FPN [53,59]	LVIS v1	MAL 掩码	27.2	42.7	29.1	19.8	35.9	40.7
ConvNeXt-Small [44]	COCO	None	51.5	70.6	56.1	34.8	55.2	66.9
ConvNeXt-Small [44]	COCO	真实标注掩码	51.8	70.6	56.3	34.5	55.9	66.6
ConvNeXt-Small [44]	COCO	MAL 掩码	51.7	70.5	56.2	35.2	55.7	66.8

表7. 添加不同掩码监督后的检测结果。模型在 COCO val2017 和 LVIS v1 上进行评估。通过使用真实标注掩码或掩码伪标签添加掩码监督，我们在 LVIS v1 的不同 AP指标上均可获得约 1% 的提升。在 COCO val2017 上，检测性能也受益于掩码伪标签。尽管提升幅度低于 COCO 上的结果，但在不同随机种子下，这一提升始终稳定。

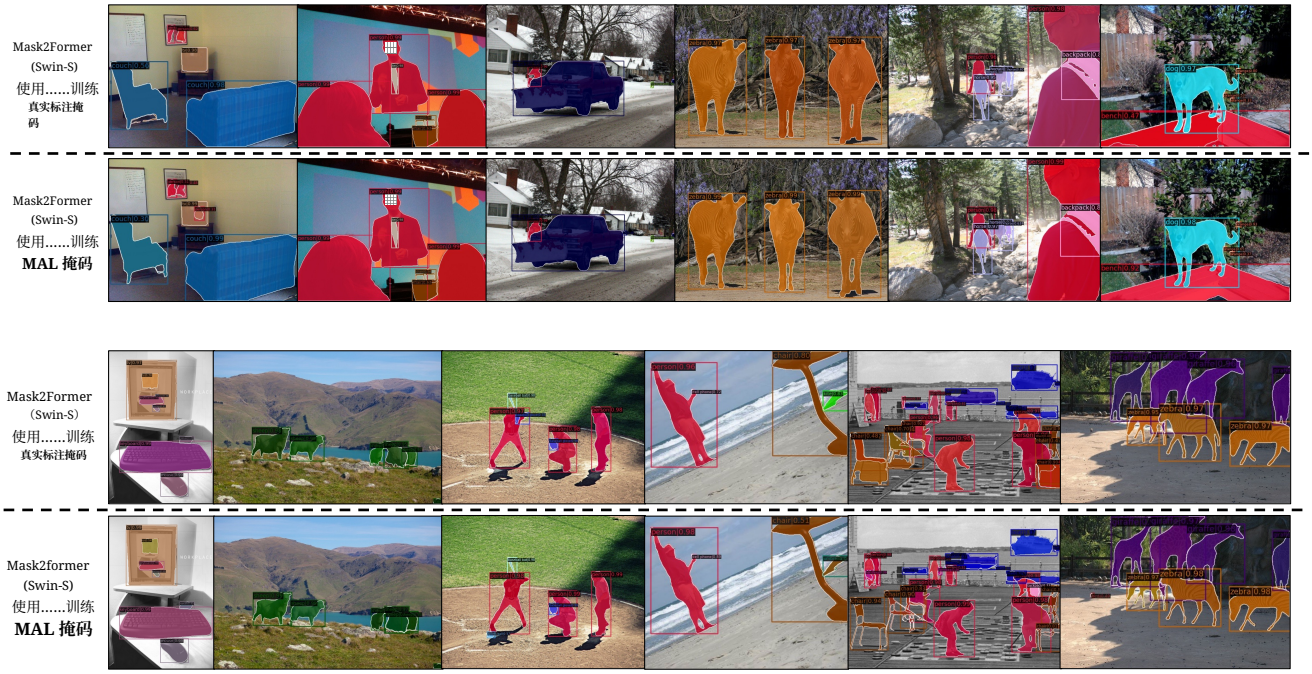


图9. 使用真实标注掩码训练的 Mask2Former 与使用 MAL生成的掩码伪标签训练的 Mask2Former 之间的定性比较。请注意，我们使用 ViT-MAE-Base 作为 MAL 的图像编码器，并使用 Swin-Small 作为 Mask2Former 的骨干网络。