arXiv:2305.01910v1 [cs.CV] 3 May 2023

分布式实例分割：使用 Latent-MaskRCNN 对不确定性和高置信度预测进行建模

YuXuan Liu[1,2], Nikhil Mishra[1,2],Pieter Abbeel[1,2],Xi Chen[1]

摘要——目标识别和实例分割是任何机器人或自主系统的基本技能。现有最先进的方法往往无法在复杂或模糊场景中捕捉有意义的不确定性，因此可能导致高性能应用中出现严重错误。本文探讨了一类使用潜变量编码的分布式实例分割模型，该模型能够对物体掩码的合理假设进行不确定性建模。针对机器人抓取应用，我们提出一种置信度掩码方法，以达到工业应用场景所需的高精确率。我们证明，该方法能够显著减少机器人系统中的严重错误，其中包括我们新近发布的、涵盖机器人应用中模糊场景的数据集。在真实世界的服装抓取机器人上，该方法在保持高性能的同时，显著减少了双重抓取错误。

I. 引言

实例分割是许多现实世界机器人系统中的一个基本问题。实例分割的目标是枚举图像中出现的物体（或实例），并指出图像中的哪些像素属于每个物体。

过去几年，近期的研究主要致力于开发专用架构，使实例分割任务更适用于深度学习。例如，先检测后分割 方法 [1], [2], [3], [4], [5],[6]，依靠分类、回归和过滤级联，首先为每个实例确定一个边界框 （这一相关问题称为目标检测），随后根据其边界框预测每个实例的掩码。另一个例子是像素嵌入 方法[7], [8], ，该方法优化像素级辅助任务，然后使用专门的聚类过程，从密集像素表示中提取实例预测。

我们观察到，现有方法并不擅长处理现实世界中固有的歧义。我们认为，这源于一种我们称之为分布表达能力有限的现象，即大多数实例分割模型都被设计为只预测一个 可能的分割假设（一组单一的物体）。仅做出单一预测在以下方面存在局限

高性能自主系统所能达到的精度：机器人抓取应用可能只能容忍 $< 1\%$ 的错误分割所导致的错误。

为克服这些局限性，我们提出分布式实例分割 方法，该方法对物体的合理假设进行分布建模。本工作的主要贡献如下：

[1]Covariant.ai， [2]加利福尼亚大学伯克利分校。联系邮箱：yuxuanliu@berkeley.edu。发表于 ICRA 2023。

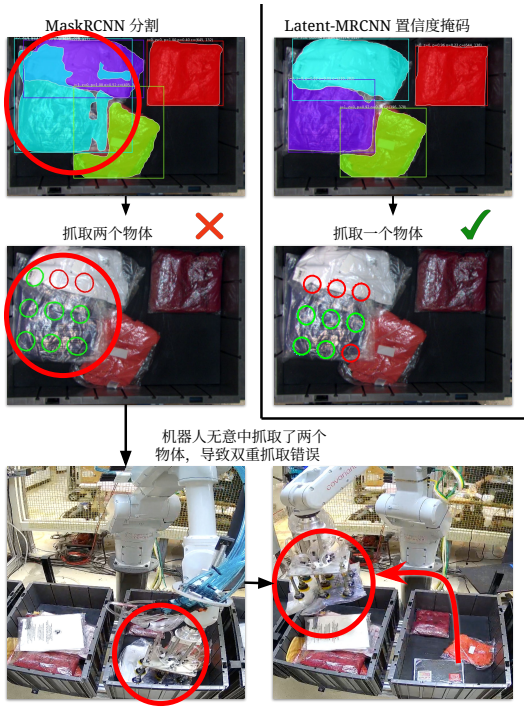


图 1. MaskRCNN 等传统实例分割模型无法对物体掩码的不确定性进行建模。对于机器人技术，这可能导致严重错误，例如意外抓取两个物体。我们的 Latent-MRCNN 可以预测物体掩码的多个假设，并利用这些假设做出高置信度预测，从而降低双重抓取错误的发生率。

1. 我们介绍了一种使用潜变量编码的分布式实例分割模型 Latent-MaskRCNN，它能够预测多个物体掩码假设。 2. 我们提出了利用分布式实例分割模型输出的新方法。针对机器人应用，我们提出了使用置信度掩码生成高精度预测的方法，并通过 Union-NMS 实现了高召回率。 3. 我们将发布一个数据集，其中包含来自真实世界机器人应用的 5000 多张标注图像，突显了实例分割中的模糊性。我们展示了我们的方法在该数据集以及常用的驾驶和实例分割数据集上均取得了优异表现。 4. 在真实世界的服装抓取机器人上，我们的方法能够显著减少严重错误，同时保持高水平的性能（图 1）。

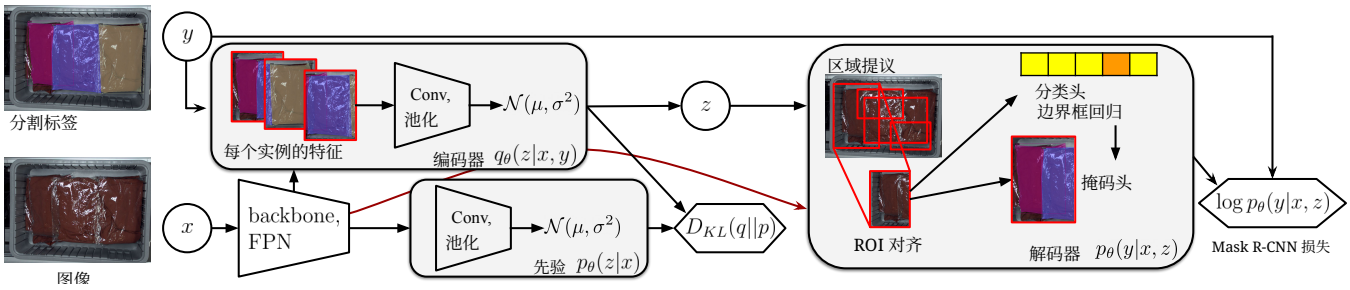


图 2. Latent-MaskRCNN 概览：在训练时，编码器 q_θ 使用从图像 x 和标签 y 中提取的特征来采样潜在编码 z，并将其传入解码器。解码器以 z 为条件，并使用典型的 MaskRCNN 架构来预测掩码，其头包括区域提议、分类器、边界框和掩码头。在推理时，通过先验 $p_\theta(z|x)$采样 z，该先验仅将图像 x 作为输入。 $D_{KL}(q||p_\theta(z|x))$确保先验能够充分覆盖潜在空间。

II. 相关工作

检测后分割 方法是最流行的实例分割方法，MaskRCNN 就属于这一类。尽管这些方法都会先执行目标检测，然后根据每个实例的边界框进行分割，但具体做法有所不同。例如，YOLACT [9] 采用与 MaskRCNN 相同的结构，但使用 YOLO [10] 而非 Faster R-CNN [2]作为目标检测器。YOLO 与 Faster R-CNN 非常相似，其架构上的改动牺牲了一些准确率，以换取实时推理速度。因此，我们预计 YOLACT 会受到与 MaskRCNN 相同的分布局限性影响。其他方法 [11], [12],[13]则探索如何在检测步骤中表达不确定性，但它们考虑的是单个边界框的分布，而不是物体掩码集合的分布。

掩码提议方法 [14], [15], [16] 旨在避免将边界框用作中间表示。它们的结构类似于 Faster R-CNN [2]，，但直接提出掩码。从经验上看，它们的表现与 MaskRCNN 并无太大差异。从分布角度看，它们面临着与 MaskRCNN 许多相同的局限：每个提议仍然独立地对每个像素进行建模，而且仍依赖非极大值抑制来筛选提议。

像素嵌入 [7], [8], [17], [18]方法的工作方式与上述两类方法截然不同。它们通常会优化某种辅助任务，以促使同一实例中的像素具有相似的表示。随后，它们依赖基于聚类的推理过程，从其逐像素表示中提取实例预测。然而，与先检测后分割方法相比，这类方法的性能一直落后很多，因此相对不受欢迎。它们可以采用与朴素语义分割方法类似的方式对逐像素不确定性进行建模，但这很可能不足以实现分布表达能力。

许多方法探索了如何在其他结构化预测任务中表达不确定性。然而，其中许多方法通过训练整个模型或部分参数的多个副本，并修改训练目标以鼓励副本之间的多样性来实现这一点 [19], [20], [21], [22]。这会导致计算成本和内存占用成倍增加而这些成本在训练时可能高得令人难以承受，尤其对于大型模型而言。其他潜变量方法 [23],[24], [25]则在医学语义分割和视频分割任务上取得了改进。然而，我们发现，实例分割面临着更多样的挑战，并且有不同的特定应用用途。

III. 基于潜变量的分布式实例分割

A. 潜变量公式

如何在保留现有模型架构归纳偏置的同时，将实例分割转化为具有分布表达能力的模型？借鉴变分推断领域的先前工作 [23], [26] ，我们考虑一种潜变量形式，在其中融入变分自编码器风格的潜变量编码。采用这一框架后，实例分割模型便成为一个条件变分自编码器，并通过最大化证据下界进行训练：

$$\log p(y|x) \geq \mathbb{E}_{z\sim q}[\log p(y|x, z)] - D_{KL}(q||p(z|x)) \quad (1)$$

通常， $q(z|y,\ x)$ 被称为编码器， $p(y|x,\ z)$ 被称为解码器，而 $p(z|x)$ 被称为先验，这些组件均通过学习来最大化下界。从本质上讲，解码器是一个传统意义上的实例分割模型，只不过经过扩展，还能够额外接收潜在编码 z。这种通用技术使我们能够重用任何现有的实例分割模型来实现解码器（并以相同的方式对其进行训练），只需稍作修改，将 z 纳入输入即可。

在推理过程中，我们可以通过对不同的潜变量编码 $z^{(k)} \sim p(z|x)$ 进行采样，并将其解码为不同的实例预测 $y^{(k)} \sim p(y|x,\ z)$，从 $p(y|x)$ 中进行采样。这样做非常强大，因为我们现在可以针对给定图像采样多个结构丰富且表达力强的假设。

B. Latent-MaskRCNN

原则上，潜变量方法可以应用于任何现有的实例分割模型。本节将探讨如何将其应用于 MaskRCNN。我们将由此得到的模型称为*Latent-MaskRCNN* （图 2）。我们选择 MaskRCNN，是因为它是最受欢迎的实例分割模型之一，近年来也一直是大多数最先进方法的基础。

Latent-MaskRCNN 的解码器采用与 MaskRCNN 相同的架构和训练目标，主要变化在于需要融入潜变量编码。为了尽可能让它们影响预测过程，我们希望在模型中相对靠前的位置注入这些编码。我们选择在区域提议之前直接注入潜变量编码，使其能够影响区域提议网络、目标检测头和掩码头。我们将潜变量编码平铺到图像的空间维度上，并与特征金字塔网络（FPN）[27]的特征图拼接。随后，我们使用若干卷积层，将组合后的特征图投影回原有的通道维度。

Latent-MaskRCNN 的编码器接收一幅图像 x 以及一组真实实例 y，并生成潜变量编码分布 $q_\theta(z|y,\ x)$ $=\mathcal{N}(\mu_\theta(y,\ x)\ \sigma^2_\theta(y,\ x))$。 $\mu_\theta(y,\ x)$ 和 $\sigma_\theta(y,\ x)$ 的架构借鉴了 MaskRCNN 的掩码头：它类似一个“反向掩码头”，对每个真实标注实例进行处理，然后汇集所有实例的特征。对于每个真实标注实例 y_i，我们从 FPN 特征图中提取 ROI对齐特征。接着，我们使用一个小型卷积神经网络，将每个特征嵌入为一个特征向量。此时，我们采用图神经网络 [28]，从逐实例特征中汇总信息，因为我们需要为整幅图像生成一个潜在编码。经过若干图网络层后，我们对节点特征进行均值池化，并使用一个全连接层为潜在分布生成均值和对数方差。编码器仅在训练时使用，因为它可以访问真实标注掩码标签。

在推理时，我们必须从先验分布中采样潜变量编码，以生成掩码样本。该先验接收一幅图像 x ，并生成一个潜变量编码分布$p_\theta(z|x)\ =\mathcal{N}(\mu_\theta(x)\ \sigma^2_\theta(x)$。我们对 FPN 特征图应用若干卷积层，在空间维度上进行均值池化，然后使用一个小型多层感知机预测均值和对数方差。对于所有潜在分布，我们均采用固定的 64 维高斯分布，并使用对角协方差。

在训练中，我们使用编码器 q_θ 采样潜变量编码 z，并将其传递给 Mask-RCNN 解码器。我们最大化证据下界（公式 1）目标，其中 $-\log\ p(y|x,z)\ =\ \mathcal{L}_M(x,y,z)$ 是常规的 Mask-RCNN损失：

$$\mathcal{L}_M(x,y,z)=\mathcal{L}_{RPN}+\mathcal{L}_{cls}+\mathcal{L}_{box}+\mathcal{L}_{mask} \quad (2)$$

$D_{KL}(q||p(z|x))$ 确保先验能够充分覆盖编码器分布。在推理过程中，我们从先验中采样潜变量编码（而非从编码器中采样），但解码器以与训练期间相同的方式处理它们。

我们发现，采用 KL 预热很有帮助，这是训练变分自编码器时的常见做法 [29]。因此，Latent-MaskRCNN 的总训练损失变为：

$$\mathcal{L}(x,y)=\mathbb{E}_{z\sim q}[\mathcal{L}_M(x,y,z)]+\beta D_{KL}(q||p(z|x)) \quad (3)$$

在训练的前半部分，我们使用 $\beta\ =\ 0$ ，并在训练接近结束时增大 β。这使得潜在编码能够

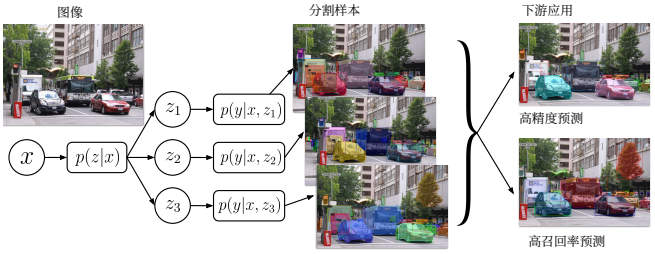


图3. 在推理时，编码器 q_θ 会被丢弃，并从以图像 x 为条件的先验 p_θ $(z|x)$中采样潜变量 z_i 。每个潜变量都通过 $p_\theta(y|x,\ z_i)$解码为一组掩码，具体可根据应用场景用于生成高精确率或高召回率的预测。

在模型其余部分仍在学习时，尽早编码有用信息；在训练后期，较高的 β 会推动先验覆盖潜在空间，从而获得更好的样本。有关我们模型和代码的更多详情，请访问我们的网站 segm.yuxuanliu.com。

IV. 分布式实例分割的应用

对于一个具有分布表达能力的分割模型，一个自然的问题是：下游应用如何使用它的分布输出？实例分割通常位于感知流水线的起始阶段，因此如何将分布式分割模型生成的样本用于下游，并非一目了然。此外，不同应用对错误不对称性的容忍程度可能各不相同：在自动驾驶中，漏检物体可能造成灾难性后果；而在机器人抓取中，这种情况或许可以接受。相反，在机器人抓取中，将两个物体合并为一个属于严重错误，但在驾驶应用中则相对可以接受。本节将展示，如何将单个 Latent-MaskRCNN 模型灵活地用于具有不同需求的多种应用（图3）。

A. 使用 p置信度掩码实现高精度

在某些应用中，出现欠分割 错误的代价可能很高，即实例的掩码预测范围大于其实际范围。例如，考虑一种机器人操作应用：机器人必须一次抓取一个物体，并将其送入分拣流程。如果模型对某个实例欠分割，就可能无意中抓取多个物体，从而给下游应用造成高昂代价。我们如何确保不发生此类错误?

假设我们从 Latent-MaskRCNN 中抽取多个样本。如果两个像素在许多样本中都属于同一个实例掩码，那么我们就可以相当有置信度地认为，它们实际上确实属于同一个真实标注实例。基于这一直觉，我们可以进一步计算一个p-置信度掩码，其中包含的所有像素都很可能属于同一个真实标注实例。

对于给定的置信度要求 p，我们将置信度掩码 c_p 定义为：以至少 p 的概率完全包含在真实标注掩码 m 中的掩码：$\mathbb{P}(c_p\ \subseteq$

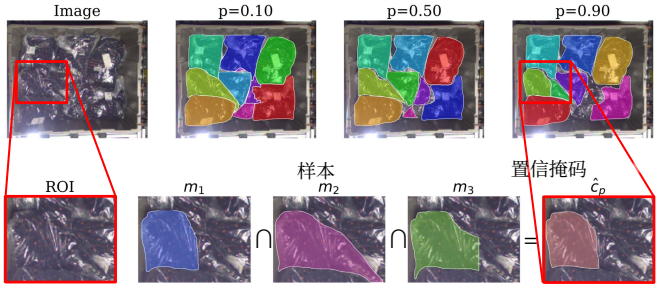


图 4。上：不同 p 下的置信度掩码预测。请注意，随着置信度要求 p 增加，单个对象可能会分裂成两个，对象范围的模糊性会降低，而不确定对象则会被完全消除。下：通过对样本 m_1，m_2，m_3 取交集，构建经验置信度掩码 $\hat{c}_p$。当对象范围不确定时，高置信度掩码预测将仅包含根据样本判定极有可能属于某个对象的像素。

$m) \geq p$。使用 Latent-MaskRCNN，我们可以将该概率近似为：

$$\mathbb{P}(c_p \subseteq m) = \mathbb{E}_{p(m|x)}[\mathbb{1}\{c_p \subseteq m\}] \approx \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}\{c_p \subseteq m_i\}$$

在有限样本情形下，如果 $\hat{c}_p$ 在 p 比例的样本中包含于一个采样掩码内，则它是经验置信度掩码。

现在考虑由至少 kp 个不同样本中的一个掩码 m_j 组成的任意掩码子集 I。如果我们取 I中所有掩码的交集 $\hat{c}_p = \bigcap_{m_j \in I} m_j$，那么该交集掩码必须包含于用于交集 $\hat{c}_p \subseteq m_j$ 的每个掩码中。因此我们有：

$$\frac{1}{k}\sum_{m_j \in I} \mathbb{1}\{\hat{c}_p \subseteq m_j\} = \frac{|I|}{k} \geq p$$

并且，$\hat{c}_p$ 根据构造即为经验置信度掩码。

图 4 展示了针对不同 p 的置信度掩码预测。请注意，随着置信度要求提高，这些掩码中不确定的范围会缩小，一些不确定的掩码也会被消除。我们还可以看到，通过交集构造置信度掩码能够得到高置信度区域预测。

B. 置信度掩码评分

由于每个置信度掩码都是掩码$c_p = \bigcap_I m_j$的交集，我们应如何为置信度掩码预测分配分数？一种直观的方法是，取交集中的每个掩码的分数 s_i 的平均值：$\frac{1}{|I|}\sum s_j$。然而，置信度掩码并不是掩码的平均值，而是它们的交集。

为了更好地定义置信度掩码预测的分数，我们考虑两种情形。在第一种情形中，模型对某个物体的掩码具有很高的置信度，因此它在每个样本中预测的掩码都大致相同。由此得到的置信度掩码 c_p 与样本 m_j中使用的每个掩码都具有较高的交并比。另一方面，考虑一种置信度较低的预测，即该物体的掩码在不同样本之间存在显著差异。在这种情况下，置信度掩码 c_p 表示一个较小的但这是物体中置信度较高的区域，其范围却高度不确定。当模型有把握且掩码在不同样本间保持一致时，c_p 与每个 m_j 之间的交并比会更小。

我们将每个置信度掩码的分数定义为预测掩码 c_p 与交集中所使用的每个掩码之间的分数加权交并比的均值：

$$s_{c_p} = \frac{1}{|I|}\sum_{m_j \in I} s_j \frac{|c_p \cap m_j|}{|c_p \cup m_j|}$$

当 s_{c_p} 较大时，这表示 c_p 是交并比非常相似的掩码的高置信度交集。另一方面，较小的 s_{c_p} 表示 c_p 与其样本的分数或交并比较低，可能无法很好地捕捉物体的完整范围。

为了预测图像的一组置信度掩码，我们迭代地选择得分最高的置信度掩码，同时排除迄今为止预测的所有置信度掩码所覆盖的像素。该算法通过贪心方式近似求解得分最高的置信度掩码选择优化问题。

C. 使用 *Union-NMS* 实现高召回率

其他应用可能会关注过度分割，即分割不足的反面。例如，在自动驾驶中，未能识别行人，或预测行人的尺寸小于实际尺寸，都可能导致灾难性错误。

为了使用 Latent-MaskRCNN 生成高召回率预测，我们采用一种称为 *Union-NMS* 的过程。我们首先从模型中采样多个分割结果，然后对预测掩码运行非极大值抑制。它会检查任意两个掩码 m_i, m_j的交并比是否大于某个阈值，随后丢弃得分较低的那个。假设执行非极大值抑制后仍有某个掩码 m_i 保留。那么，Union-NMS 会返回 m_i与其抑制的每个掩码的并集，通过纳入原本会被忽略的掩码来实现更高的召回率。

D. 使用先验均值进行普通预测

有些应用可能没有特定的性能要求，也可能对推理时长有严格要求。在这些情况下，点估计就足够了。使用 Latent-MaskRCNN，我们可以通过始终解码先验的均值来实现这一点：$z = \mu_\theta(x)$，其中 $\mu_\theta(x)$ 是先验 $p_\theta(z|x)$ 的均值（对于高斯分布 $p_\theta(z|x)$，它也是该分布的众数）。我们发现，这种方案的表现通常与 MaskRCNN 的预测相当，或略有提升，这表明 Latent-MaskRCNN 严格增强了 MaskRCNN 的表达能力，而且使用表达能力更强的分布不会损失性能。

V. 实验

我们开展实验，旨在回答以下问题：

1. 带有置信度掩码的 Latent-MaskRCNN 能否在各种数据集上做出高精度预测？2. Union-NMS 能否实现高召回率预测？3. Latent-MaskRCNN 能否减少机器人抓取应用中的关键双重抓取错误？

表 I MaskRCNN 和 Latent-MaskRCNN 在不同数据集和指标上的评估。

	COCO			Cityscapes			Apparel-5k		
方法	高精度平均召回率	AR	平均精度均值	高精度平均召回率	AR	平均精度均值	高精度平均召回率	AR	mAP
MaskRCNN	20.0	66.0	35.0	25.3	55.1	**35.8**	23.6	39.9	26.9
潜在 并集非极大值抑制	7.4	**72.3**	26.5	17.7	**57.5**	33.6	13.6	**61.4**	34.1
潜在置信度掩码	**22.0**	48.1	30.5	**28.0**	48.6	34.1	**42.4**	41.8	**35.1**
潜在先验均值	19.5	65.8	**35.3**	25.7	53.8	35.0	26.9	49.4	34.3

A. 数据集

为了帮助我们回答这些问题，我们在多个数据集上比较了 MaskR- CNN 和 Latent-MaskRCNN，每个数据集都有各自的挑战。

COCO [30]：这个大型数据集是实例分割的标准基准。它包含许多物体类别，图像构成也极为多样。

Cityscapes [31]：一个来自自动驾驶应用的真实世界数据集。虽然它比 COCO 更小且专业性更强，但仍是实例分割领域广受欢迎的基准数据集。一项显著的挑战是，许多背景实例（例如行人）仍需要进行分割，但有限的图像分辨率可能会带来一些不确定性。

Apparel-5k：我们从机器人抓取应用中收集了这个约含 5000 张图像的数据集。我们使用训练集中的 4198 张图像和验证集中的 463 张图像。该数据集仅包含一个物体类别，但由于复杂的遮挡、光照、透明度等因素，图像存在大量固有歧义。我们将在网站 segm.yuxuanliu.com 上发布该数据集，供更广泛的社区在我们的工作基础上进一步研究。

对于每个数据集，我们都使用 MaskRCNN 发布的超参数和训练方案，在 8 个 GPU 上训练了 MaskRCNN 和 Latent-MaskRCNN。所有实验和数据集均使用相同的公开训练集/验证集划分。我们使用从预训练 ImageNet [33]权重初始化的 Resnet-50 骨干网络 [32]，（用于 COCO），或从预训练 COCO 权重初始化的 Resnet-50 骨干网络（用于其他数据集）。MaskRCNN 的推理耗时为 80–100 毫秒，具体取决于物体数量；Latent-MaskRCNN 的推理耗时为 500–1000 毫秒，具体取决于样本数量和物体数量。

B. 评估 p置信度掩码

在第 IV-A 节中，我们介绍了使用 p置信度掩码进行高精度预测，以解决分割不足的问题。在这些情况下，我们关注预测具有较高的交集占比：IoP(m_i, g) $= \frac{|m_i \cap g|}{|m_i|}$。交集占比越高，因分割不足而产生错误的可能性就越低。

在此评估设定下的模型时，我们需要权衡精确率（以 IoP 衡量）与召回率（以避免退化解）。为此，我们考虑高精度下的最大召回率（MR@HP）：

$$\text{MR@HP} = \frac{1}{|p| \cdot |\tau|} \sum_{p_i \in p, \tau_j \in \tau} \max_{t:\text{Precision}(\tau_j) \geq p_i} \text{Recall}(t, \tau_j)$$

给定精确率阈值 p_i 和 IoP 阈值 τ_j，我们可以计算每个模型达到的最大召回率（如果模型始终无法达到精确率 p_i，则记为零）。MR@HP 指标是这些召回率在一系列精确率阈值 p 和 IoP 阈值 τ上的平均值。对于要求高精确率的应用场景，我们关注模型在这些阈值较高时的表现，因此使用 $p = \tau = [0.75, 0.8, 0.85, 0.9, 0.95]$。

在表 I 中，我们使用第 IV-D 节中的先验均值方案以及置信度水平为 $p = 0.9$ 的置信度掩码，对 Latent-MaskRCNN 进行了评估。在全部三个数据集上，我们发现潜在置信度掩码在高精度平均召回率方面表现最佳。至于平均精度均值，我们发现潜在先验均值在全部三个数据集上的表现不逊于 MaskRCNN，甚至有所超越。在具有挑战性的 Apparel-5k 数据集上，我们发现，潜在置信度掩码和潜在先验均值在高精度平均召回率和平均精度均值方面均显著优于 MaskRCNN。总体而言，我们发现，Latent-MaskRCNN 严格优于 MaskRCNN：它在平均精度均值方面保持了整体检测性能，并在高精度平均召回率方面表现最佳。

C. 并集非极大值抑制的评估

针对过度分割问题，我们在第 IV-C 节介绍了并集非极大值抑制方法。在这类情况下，我们关注的是高召回率（即检测到每个存在的实例），以及每个掩码预测都具有较高的 交并比 （与真实标注的交集比）：IoG(m_i, g) $= \frac{|m_i \cap g|}{|g|}$。为兼顾这两方面，我们采用基于交并比的平均召回率（AR） [30] 。该指标既衡量召回率，也会惩罚过度分割（是否存在过小的预测掩码）。

我们使用其先验均值（第 IV-D 节）和并集非极大值抑制评估了 Latent-MaskRCNN。表 I 显示，在全部三个数据集上，先验均值预测与 MaskRCNN 的结果相近，而并集非极大值抑制取得了显著更高的平均召回率（使用交并比）。这表明，对于高召回率应用，采用并集非极大值抑制的 Latent-MaskRCNN 能够更有效地覆盖不同的不确定性模式。

D. 置信度掩码能否减少 *Apparel-5K* 上的双重抓取错误？

在机器人抓取应用中，机器人误以为只抓取了一件物品，却意外抓起两件，这会造成高昂代价，因为这会影响库存计数和下游订单。对于 Apparel-5k 数据集，我们可以通过以下方式估算模型分割预测的双重抓取率：

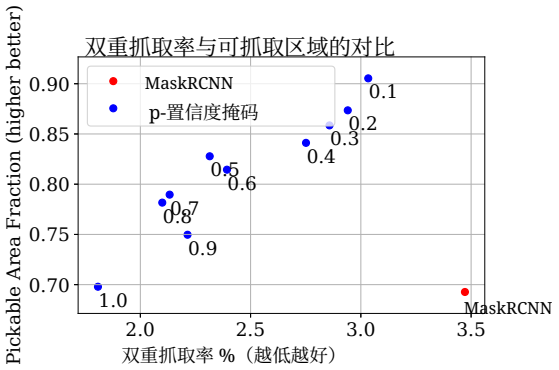


图 5. 与 MaskRCNN 相比，潜在置信度掩码实现了更低的双重抓取率，并且通常具有更大的可抓取区域。

将机器人的夹爪近似为像素空间中半径固定的圆。然后，我们在图像上随机采样圆，并统计落在一个预测掩码内、但落在多个真实标注掩码内的圆的数量 D。再将这个数量除以落在一个预测掩码内的圆的数量 N，从而得到估算的双重抓取率 $R = \frac{D}{N}$。通过实证，我们发现这种模拟得出的双重抓取率与真实机器人上的双重抓取率相关。

工业机器人抓取中另一个值得关注的指标是可抓取面积，即机器人能够抓取的可见表面总面积。能够预测更大且更准确的可抓取面积的模型，可以让机器人在抓取策略上拥有更大的灵活性。为此，我们计算所有预测掩码的面积与所有真实掩码面积之比，作为可用可抓取面积的占比。

我们在图 5 中比较了使用不同 p 置信度掩码的 MaskRCNN 和 Latent-MaskRCNN。我们发现，在所有情况下，Latent-MaskRCNN 的可抓取面积占比和双重抓取率都优于 MaskRCNN。此外，Latent 置信度掩码中的可调参数 p 可以根据具体应用，在双重抓取率和可抓取面积之间进行权衡。随着每项预测的置信度要求提高，p 的值越高，通常对应越低的双重抓取率和越小的可抓取面积。传统的 MaskRCNN 没有可调节参数，因此只能实现一种双重抓取率和可抓取面积占比。

E. 置信度掩码能否减少真实世界服装抓取机器人中的双重抓取错误?

为了评估数据集评估结果是否能反映真实机器人的性能，我们在服装抓取机器人上比较了 MaskRCNN 和 Latent-MaskRCNN。我们使用配备 9 杯吸盘式夹爪的 ABB1300，在两个料箱之间抓取装在塑料袋中的服装（图 1）。机器人使用两套顶部相机系统进行实例分割，然后生成抓取点。抓取点经过优化，尽可能让多个吸盘都落在分割模型检测到的同一个物体上。

在我们的评估中，我们只更换所使用的分割模型，同时保持系统的其他部分不变，

表 II 服装抓取机器人评估。 * 表示差异具有统计显著性

方法	双重抓取率	吸附成功的吸盘平均数量
MaskRCNN	4.40%*	4.76
Latent-MRCNN	**0.82%***	**4.91**

包括硬件、物体集合和抓取点生成。每个分割模型都使用相同的 Apparel-5k 数据集进行训练。我们让每个模型执行数百次抓取，并记录双重抓取的次数，即意外抓取两个物体的情况。在工业仓库应用中，这类双重抓取会造成高昂的错误成本，因为它们会导致库存计数不准确，并给下游分拣和订单履行系统带来错误。典型的高自动化仓库最多只能容忍 1% 的双重抓取率；超过这一比例，机器人造成的问题就会多于它所解决的问题。

我们还测量被抓取物体上密封吸盘的平均数量，因为吸附保持力与密封吸盘的数量成正比。使用较少密封吸盘的抓取更容易导致物体掉落，进而造成卡滞、库存损失以及高昂的人工干预成本。机器人系统可以通过缩小物体掩码、将较大的掩码切分成较小的掩码，或仅使用一个小型吸盘来降低双重抓取率。然而，这些方法都会不加区分地降低所有物品的吸附保持力，而我们的方法只会在存在歧义时采取保守策略。

表 II 报告了我们的服饰抓取实验结果。我们发现，采用 0.9 置信度掩码的 Latent-MaskRCNN 显著降低了双重抓取率。这验证了我们在第 V-D 节中对 Apparel-5K 的模拟结果。此外，Latent-MaskRCNN 达到的密封吸盘平均数量略高，表明吸附稳定性并未受到牺牲。这说明我们的方法能够生成高置信度预测，并在面临不确定性时作出恰当的权衡。

VI. 讨论

我们提出了一系列新模型，通过使用潜变量增强现有实例分割模型，从而提高其分布表达能力。Latent-MaskRCNN 能够表达广泛的不确定性，而现有实例分割模型往往难以做到这一点。我们可以利用模型通过置信度掩码和 Union-NMS 表达的不确定性，分别实现高精确率和高召回率。这些方法在机器人技术、自动驾驶和通用目标数据集上都展现出出色的性能。在一台真实的服装抓取机器人上，我们发现，该模型在保持高性能的同时，能够显著降低严重错误的发生率。最后，我们强调了分布表达能力的重要性，并希望未来的实例分割研究能够在我们的工作基础上继续推进，同时利用本文分享的数据集。

参考文献

[1] R. Girshick，“Fast R-CNN”，发表于《IEEE国际计算机视觉会议论文集》，2015年，第1440–1448页。[2] S. Ren、K. He、R. Girshick 和 J. Sun，“Faster r-cnn：借助区域提议网络实现实时目标检测”，发表于《神经信息处理系统进展论文集》，2015年，第91–99页。[3] K. He、G. Gkioxari、P. Doll´ar 和 R. Girshick，“Mask R-CNN”，发表于《IEEE国际计算机视觉会议论文集》，2017年，第2961–2969页。[4] R. Girshick、J. Donahue、T. Darrell 和 J. Malik，“用于精准目标检测和语义分割的丰富特征层次结构”，发表于《IEEE计算机视觉与模式识别会议论文集》，2014年，第580–587页。[5] N. Carion、F. Massa、G. Synnaeve、N. Usunier、A. Kirillov 和 S. Zagoruyko，“使用Transformer进行端到端目标检测”，发表于欧洲计算机视觉会议，2020年。[6] H. Zhang、F. Li、S. Liu、L. Zhang、H. Su、J. Zhu、L. M. Ni 和 H.-Y. Shum，“Dino：采用改进去噪锚框的 DETR 端到端目标检测”，2022年。[7] A. Arnab 和 P. H. S. Torr，“通过动态实例化网络实现像素级实例分割”，发表于 IEEE 计算机视觉与模式识别会议论文集（CVPR），2017年7月。[8] M. Bai 和 R. Urtasun，“用于实例分割的深度分水岭变换”，发表于《IEEE计算机视觉与模式识别会议论文集》，2017年，第5221–5229页。[9] D. Bolya、C. Zhou、F. Xiao 和 Y. J. Lee，“Yolact：实时实例分割”，发表于《IEEE国际计算机视觉会议论文集》，2019年。[10] J. Redmon、S. Divvala、R. Girshick 和 A. Farhadi，“You Only Look Once：统一的实时目标检测”，发表于《IEEE计算机视觉与模式识别会议论文集》，2016年，第779–788页。[11] D. Hall、F. Dayoub、J. Skinner、H. Zhang、D. Miller、P. Corke、G. Carneiro、A. Angelova 和 N. Sunderhauf，“概率目标检测：定义与评估”，2020年3月，第1020–1029页。[12] A. Harakeh、M. Smart 和 S. L. Waslander，“BayesOD：用于深度目标检测器不确定性估计的贝叶斯方法”，《2020 IEEE国际机器人与自动化会议》，第87–93页，2020年。[13] D. Miller、N. Sunderhauf、H. Zhang、D. Hall 和 F. Dayoub，“基于采样的概率目标检测器基准评测”，发表于《IEEE/CVF计算机视觉与模式识别会议论文集（CVPR）研讨会》，2019年6月。[14] X. Chen、R. Girshick、K. He 和 P. Doll´ar，“TensorMask：密集目标分割的基础”，发表于《IEEE国际计算机视觉会议论文集》，2019年。[15] E. Xie、P. Sun、X. Song、W. Wang、X. Liu、D. Liang、C. Shen 和 P. Luo，“PolarMask：采用极坐标表示的单阶段实例分割”，发表于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），2020年6月。[16] B. Cheng、I. Misra、A. G. Schwing、A. Kirillov 和 R. Girdhar，“用于通用图像分割的掩码注意力掩码Transformer”，2022年。[17] D. Neven、B. D. Brabandere、M. Proesmans 和 L. V. Gool，“通过联合优化空间嵌入和聚类带宽进行实例分割”，发表于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），2019年6月。[18] L. Rumberger、L. Mais 和 D. Kainmueller，《用于实例分割的概率深度学习》，2020年1月，第445–457页。[19] B. Lakshminarayanan、A. Pritzel 和 C. Blundell，“使用深度集成进行简单且可扩展的预测不确定性估计”，arXiv 预印本 arXiv:1612.01474，2016年。[20] A. Guzman-Rivera、D. Batra 和 P. Kohli，“多选学习：学习生成多个结构化输出”，发表于 NIPS，第1卷，第2期。Citeseer，2012年，第3页。[21] C. Rupprecht、I. Laina、R. DiPietro、M. Baust、F. Tombari、N. Navab 和 G. D. Hager，“在不确定的世界中学习：通过多个假设表示歧义”，发表于《IEEE国际计算机视觉会议论文集》，2017年，第3591–3600页。[22] B.-B. Gao、C. Xing、C.-W. Xie、J. Wu 和 X. Geng，“具有标签歧义的深度标签分布学习”，《IEEE图像处理汇刊》，第26卷，第6期，第2825–2838页，2017年。[23] S. A. A. Kohl、B. Romera-Paredes、C. Meyer、J. D. Fauw、J. R. Ledsam、K. H. Maier-Hein、S. M. A. Eslami、D. J. Rezende 和 O. Ronneberger，“用于歧义图像分割的概率U-Net”，CoRR，第 abs/1806.05034卷，2018年。 [在线]。可访问：http://arxiv.org/abs/1806.05034

[24] W. Jang、D. Wei、X. Zhang、B. D. Leahy、H. Y. Yang、J. Tompkin、D. Ben-Yosef、D. Needleman 和 H. Pfister，“学习用于无模态卵裂球实例分割的向量量化形状编码”，*CoRR*，第 abs/2012.00985 卷，2020年。 [在线]。地址：https://arxiv.org/abs/2012 00985 。C.-C. Lin、Y. Hung、R. Feris 和 L. He，“采用改进 VAE 架构的视频实例分割跟踪”，载于*2020 IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*），2020年，第 13 144–13 154 页。[26] D. P. Kingma 和 M. Welling，“自编码变分贝叶斯”，载于第二届国际学习表征会议（*ICLR 2014*），加拿大艾伯塔省班夫，*2014*年*4*月*14*日至*16*日，会议论文集，2014年。[27] T.-Y. Lin、P. Dollar、R. Girshick、K. He、B. Hariharan 和 S. Belongie，“用于目标检测的特征金字塔网络”，载于*IEEE*计算机视觉与模式识别会议论文集，2017年7月。[28] Z. Wu、S. Pan、F. Chen、G. Long、C. Zhang 和 P. S. Yu，“图神经网络综合综述”，2019年，引用 arXiv:1901.00596。注：已更新表格和参考文献。 [在线]。地址：http://arxiv.org/abs/1901 00596 。[25] [29] I. Higgins、L. Matthey、A. Pal、C. Burgess、X. Glorot、M. Botvinick、S. Mohamed 和 A. Lerchner，“β-VAE：通过受约束变分框架学习基本视觉概念”，载于第五届国际学习表征会议（*ICLR 2017*），法国土伦，*2017*年*4*月*24*日至*26*日，会议论文集。OpenReview.net，2017年。 [在线]。地址：https://openreview.net/forum?id=Sy2fzU9gl [30] T.-Y. Lin、M. Maire、S. Belongie、J. Hays、P. Perona、D. Ramanan、P. Doll´ar 和 C. L. Zitnick，“Microsoft COCO：上下文中的常见物体”，载于欧洲计算机视觉会议论文集施普林格，2014年，第 740–755 页。[31] M. Cordts、M. Omran、S. Ramos、T. Rehfeld、M. Enzweiler、R. Be-nenson、U. Franke、S. Roth 和 B. Schiele，“用于城市场景语义理解的 Cityscapes 数据集”，载于*IEEE*计算机视觉与模式识别会议论文集，2016年，第 3213–3223 页。[32] K. He、X. Zhang、S. Ren 和 J. 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，2016年6月。[33] J. Deng、W. Dong、R. Socher、L.-J. Li、K. Li 和 L. Fei-Fei，“ImageNet：大规模层级图像数据库”，载于*IEEE*计算机视觉与模式识别会议论文集，IEEE，2009年，第 248–255 页。

附录

A. *Latent-MaskRCNN* 架构与训练

本节将更详细地介绍 Latent-MaskRCNN 的架构（第 III-B 节）。图 6 展示了整体示意图。

在我们的实验中，我们使用维度固定为 $z \in \mathbb{R}^d$ 的潜变量编码。我们发现，$d = 64$ 是一个合理的选择，适用于所有数据集。

在解码器中（图 6 (a)），我们将潜变量编码沿图像的空间维度平铺，并与特征金字塔网络输出的特征图拼接。然后，我们使用若干卷积层将组合后的特征图投影回原始通道维度。模型的其余部分（区域提议、分类头、掩码头）使用这些经过潜变量增强的特征图，但除此之外与 MaskRCNN 保持不变。正如我们在 III-B 节中所讨论的，这种方案使潜变量编码能够影响预测的每个阶段。

编码器（图 6（b））基于图像和真实标注实例生成潜变量编码上的高斯分布。对于每个实例 y_i，我们从 FPN 特征图中提取 RoI 对齐特征，以及真实标注实例掩码的特征。然后，我们使用一个小型卷积神经网络将每个特征嵌入为单个特征向量。此时，我们希望从逐实例特征中累积信息，并生成固定大小的潜在表示。为此，我们采用图神经网络（GNN，[28]），其中每个实例构成全连接图中的一个节点。经过若干图网络层后，我们对节点特征进行均值池化，并使用全连接层为潜变量编码生成均值和对数方差。

先验（图 6 (c)）同样会预测潜在编码上的高斯分布，但它无法访问真实实例。相反，我们对 FPN 特征图应用几层卷积层，在空间维度上进行均值池化，然后使用一个小型多层感知机预测均值和对数方差。

考虑以下训练目标，并注意到当 $\beta = 1$ 时，它等价于标准的变分自编码器目标。

$$\mathcal{L}(x, y) = \mathbb{E}_{z \sim q(z|y,x)}[\log p(y|x, z)] - \beta D_{KL}(q(z|y, x)||p(z|x))$$

训练变分自编码器时，通常会采用 KL预热，即在训练开始后的前几个训练轮次中，将 β 从 0 逐步提高。这有助于促使潜变量编码变得更具信息量，因为训练初期（此时编码器尚未训练，潜变量编码也没有信息量）使用它们不会受到惩罚。

我们在内部集群的一台 8-GPU 机器（1080Ti）上训练了所有模型。根据数据集的不同，每个模型的训练时间约为 20–30 小时（与 MaskRCNN 相当）。我们的代码可在 https: //segm.yuxuanliu.com 获取。

B. 样本的定性评估

在本节中，我们定性地探讨 Latent-MaskRCNN 能够表达哪些类型的不确定性。在图7中，我们

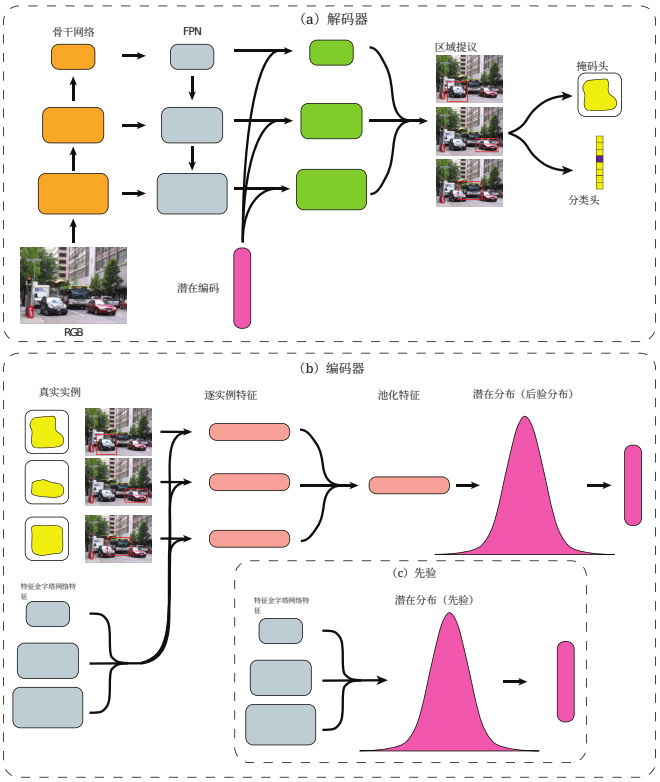


图 6. Latent-MaskRCNN 的架构。(a) 解码器与 MaskRCNN 完全相同，只是在区域提议之前，我们将潜在编码加入 FPN 特征图中。(b) 编码器接收（未增强的）FPN 特征图和真实标注实例列表，并预测潜变量编码的对角高斯分布。(c) 先验接收（未增强的）FPN 特征图，并预测潜变量编码的对角高斯分布。

将该模型在来自各种数据集的图像上的样本进行可视化，并观察到它确实捕捉到了几种不同类型的歧义：

类别混淆： Latent-MaskRCNN 能够在分类头中表达有意义的不确定性。在第 2 行中，MaskRCNN 自信地将这个水槽分类为马桶，而 Latent-MaskRCNN 的不同样本则将其分类为马桶、水槽和碗。

类别不精确性：即使某个实例的类别显而易见，其类别的定义方式仍可能存在歧义。例如，在第 1 行中，MaskRCNN 和 Latent-MaskRCNN 都（正确地）确信该实例是交通信号灯。然而，具体如何界定交通信号灯的范围，可能会导致实例掩码看起来截然不同。Latent-MaskRCNN 对各种合理的可能结果进行了广泛采样，而 MaskRCNN 则选取了单一模式（在此例中是不正确的模式）。

对象与掩码歧义： Latent-MaskRCNN 让我们能够从对象集合分布中进行采样。例如，在第 3 行中，我们可以看到样本包含数量不同的物体、各异的边界框以及各异的实例掩码。尽管这些样本都不完美，但它们都合理可信，并且都明显优于 MaskRCNN。

在第 2 行中，我们看到 Latent-MaskRCNN 的假设在掩码和边界框两方面都表达了不确定性。

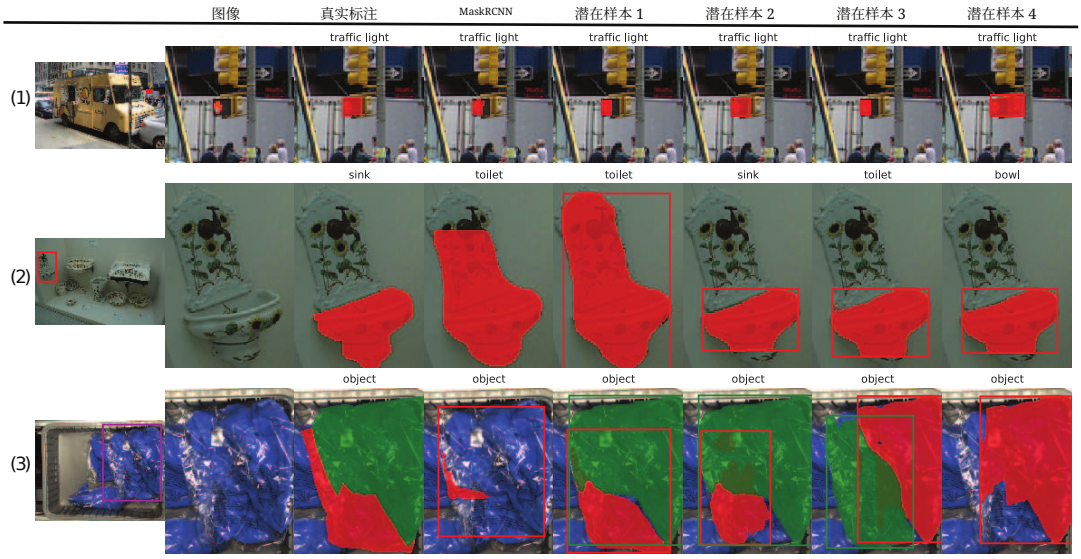


图 7. 每一行对应一张图像。第 1-2 列显示原始图像及其放大版本。其余各列显示实例掩码，图像上方的文字表示该实例的类别。第 3 列为真实标注，第 4 列为 MaskRCNN 的预测结果，第 5-8 列为 Latent-MaskRCNN 生成的样本。

边界框，而 MaskRCNN 选择单一众数。

C. 指标

本节将详细介绍我们在第 IV 节中使用的特定于应用的指标。

首先，我们简要介绍一下通常用于评估实例分割预测的平均精度均值（mAP）指标。如果预测实例与真实标注具有相同的类别，且它们的掩码之间的交并比达到某个 IoU 阈值，则称二者匹配。我们按照置信度从高到低的顺序遍历预测实例，并确定哪些实例能够与真实标注实例匹配（注意，每个真实标注实例只能参与一次匹配）。出现在匹配中的预测实例被视为真正例，其余预测实例则被视为假正例。随后，我们可以利用这些信息绘制精确率-召回率曲线，并计算曲线下面积。通常，这一过程针对每个类别分别进行，然后通过对各类别的平均精度求平均值，得到平均精度 （AP）。需要注意的是，AP 仍然取决于我们选择的 IoU 阈值。通常报告的平均精度均值 （mAP）是多个 IoU 阈值（0.5，0.55，...，0.9，0.95）下 AP 的平均值。

mAP是一种均衡性良好的指标，因为它对过度分割和分割不足施加同等惩罚，同时兼顾精确率和召回率。然而，对于我们在第五-B节和第四-C节讨论的具体应用，可能需要使用能够更好反映不同类型错误的非对称代价的其他指标。

对于第 V-B 节讨论的高精度应用场景，我们考察了高精度下的最大召回率（MR@HP）。我们采用与平均精度均值类似的方式进行匹配，只是使用交集占比而非交并比。接下来，对于给定的精确率阈值 p_i 和交集占预测区域比例阈值 τ_j，我们可以计算每个模型达到的召回率（如果】【。它始终无法达到精确率 p_i）。高精度平均召回率指标是这些召回率的平均值，计算范围涵盖一系列精确率阈值p 和交集占预测区域比例阈值 τ:

$$\text{MR@HP} = \frac{1}{|p| \cdot |\tau|} \sum_{p_i \in p, \tau_j \in \tau} \max_{t:\text{Precision}(\tau_j) \geq p_i} \text{Recall}(t, \tau_j)$$

对于第五-B节中的评估，我们使用 $p = \tau = [0.75, 0.8, 0.85, 0.9, 0.95]$。

AR在 [30] 中被提出，如今在实例分割领域越来越受欢迎，成为mAP的补充指标。它同样采用mAP使用的匹配过程，但随后只是对每种方法在标准IoU阈值范围内的召回率取平均值。

对于第四-C节讨论的高召回率应用场景，我们考虑了平均召回率（AR），但使用IoG而不是IoU。该指标既衡量召回率（模型是否预测出了所有实例？），同时也会惩罚过度分割（是否存在过小的预测掩码）。

D. 算法

在第四节中，我们提出了置信度掩码和并集非极大值抑制，作为 Latent-MaskRCNN 的两种应用。算法 1 详细介绍了迭代贪心置信度掩码预测，算法 2 则详细介绍了并集非极大值抑制流程。

E. 不确定性

在第四节 B 中，我们提出了一种为置信度掩码预测评分的方法。MaskR-CNN 等现有模型通常置信度极高，即使预测错误也是如此。然而，具有分布表达能力的模型应当能够给出校准得更好的置信度分数。

为了评估这些分数是否与掩码质量相关，我们将计算得到的分数与预测掩码和最接近的真实标注掩码之间的交并比进行比较

算法 1： 置信度掩码

给定： 置信度要求 p, k 个 y的样本
每个样本的掩码为 $m_j \in y$
初始化 $M \leftarrow \{\}$ 为一组预测掩码
当 M 中的掩码具有分数 > 0.1 **时**
  对 $m_h \in y_i$ **执行**
    计算交集的面积
    $I_{jg} = |m_h \cap m_j \setminus M|$ 与所有其他掩码
    $m_j \in y_g$
    令 C 为具有最高 kp 的掩码集合
    I_{jg} 其中每个掩码都必须来自一个
    唯一 y_g
    计算交集 $m_h^* = \bigcap_{m \in C} m \setminus M$
    忽略预测掩码 M
  添加分数最高的掩码 $\arg\max^{s_{m_h^*}}$ to M

算法 2： Union-NMS

给定： 掩码 $M = \{m_i\}_{i=1}^{n}$ 按置信度排序，
IoU阈值 τ
将 $S \leftarrow \{\}$ 初始化为空映射
对于 $i = 1 : n$ **执行**
  如果 $i \in K$ **则**
    继续
  $S[i] \leftarrow \{\}$
  对于 $j = i+1 : n$ **执行**
    如果 交并比$(m_i, m_j) > \tau$ **则**
      添加 j 至 $S[i]$
初始化 $U \leftarrow \{\}$
对于 $i \in$ 键(S) **执行**
  $m \leftarrow m_i$
  对于 $j \in S[i]$ **执行**
    $m \leftarrow m \cup m_j$
  将 m 添加到 U
结果： U

（该交并比反映了掩码的正确程度）。图 8 绘出了 Apparel-5k 数据集上的这两个量。我们看到，MaskRCNN 几乎总是具有满置信度（无论交并比如何），而 Latent-MaskRCNN 产生的置信度分数与掩码准确度具有较强的相关性。

表 III 还展示了 ROC AUC，其中我们使用各模型的分数来预测真实标注交并比超过某个阈值交并比$(c_p, G) \geq 0.5$ 的情况。在所有三个数据集上，与 MaskR- CNN 相比，Latent-MaskRCNN 的 AUC 均更高，这表明其分数更能预测掩码质量。这说明，像 Latent-MaskRCNN 这样具有较强分布表达能力的模型，结合所提出的置信度掩码评分方法，可以更有效地衡量不确定性。

当我们在 V-B 节提出 p-置信度掩码时，我们希望每个置信度掩码以 p 的概率包含在某个物体内。为了评估这一点，我们可以比较每个置信度掩码所做预测中 IoP 为 > 0.95 的比例与置信度要求 p。图 8 显示了不同置信度对应的 IoP 分位数。

表 III 使用分数预测真实标注交并比的 ROC AUC > 0.5

方法	COCO	Cityscapes	Apparel-5k
MaskRCNN	0.7372	0.8443	0.7756
置信度掩码	**0.7941**	**0.8666**	**0.9435**

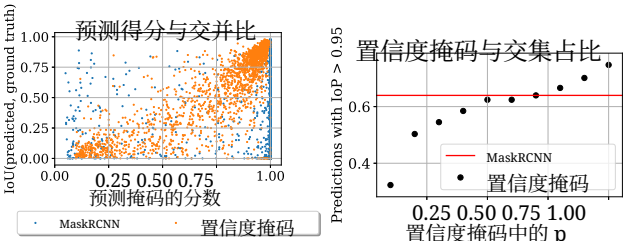


图 8。左：在 Apparel-5k 数据集上，预测得分与每个掩码的真实交并比之间的相关性。潜在置信度掩码的分数往往相关性更强，而 MaskRCNN 则过度自信。右：提高置信度要求 p 会提高预测的交集占比，而 MaskRCNN 只能得到一种交集占比分布。

Apparel-5k 上的掩码 p。我们通常发现，提高置信度要求 p 也会提高预测结果的交集占比（IoP）。与此同时，由于 MaskRCNN 没有控制输出置信度的参数，因此只能得到一种交集占比分布。