arXiv:2303.05308v2 [cs.CV] 3 Jan 2024

SpyroPose：用于物体位姿分布估计的 SE(3) 金字塔

Rasmus L aurvigHaugaard Frederik Hagelskjær Thorbjørn Mosekjær Iversen
SDU Robotics，南丹麦大学{rlha,frhag,thmi}
@mmmi.sdu.dk

摘要

物体姿态估计是许多机器人系统中的一项重要计算机视觉问题。通常的做法是估计单个位姿并为其赋予一个分数。然而，分数只能提供很少的不确定性信息，因此下游操作任务难以据此评估风险。与位姿分数相比，位姿分布可用于概率框架，使下游任务能够做出更明智的决策，并最终提高系统可靠性。位姿分布可能具有任意复杂性，因此需要无参数化的分布；然而，迄今为止，由于难以在 *SE(3)* 上进行训练和归一化，这类方法仅限于 *SO(3)* 上的旋转估计。我们提出了一种新方法 *SpyroPose*，利用 *SE(3)*金字塔进行位姿分布估计：这是一种分层网格，其分辨率随着层级加深而提高。该金字塔通过重要性采样实现高效训练，并通过稀疏评估实现实时推理。*SpyroPose* 在 *SO(3)* 分布估计方面达到了最先进水平；据我们所知，我们还首次给出了 *SE(3)* 分布估计的定量结果。位姿分布也为传感器融合带来了新的机会，我们展示了 *SpyroPose* 的一种简单多视图扩展。项目页面：*spyropose.github.io*

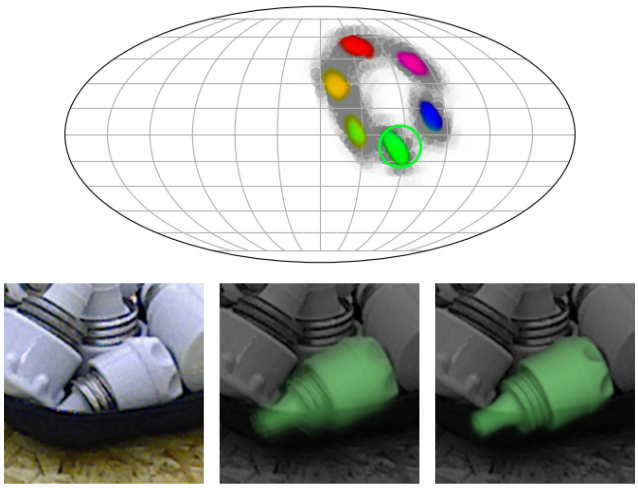


图 1. 金字塔中不同分辨率层级下 SE(3) 分布的可视化。下：输入图像（左），以及金字塔第三级（中）和第五级（右）中按估计概率加权的姿态绿色渲染图。上：边缘化后的 SO(3) 分布，其中两个维度通过莫尔威德投影显示，最后一个维度用色调表示。为在同一图中展示两个分辨率层级，第三级以灰度显示。真实旋转以圆圈标出。

1. 引言

机器人学中的许多任务涉及刚性物体的操作，并要求已知物体的旋转和平移，即物体的姿态。当环境相对不受控时，例如物体杂乱地堆放在桌面或箱子中，视觉系统通常会用于估计物体姿态。

大多数位姿估计文献都致力于开发能够给出单一最佳姿态估计的算法。如果估计结果始终足够准确，能确保下游任务成功，或者任务允许偶尔失败，那么估计单一姿态或许就足够了。然而，即使采用理想的方法，图像中固有的视觉歧义也可能过多，以至于无法可靠地完成任务；而对于某些任务，失败会造成不利影响。

表示视觉带来的不确定性，不仅有助于传感器融合，也能促成计算机视觉与机器人学之间更有依据、更合理的交互。例如， [7]，中将假设为正态分布的视觉不确定性与抓取任务的成功阈值结合起来，以估计受限环境中的抓取成功率。然而，视觉不确定性并非总是相互独立且服从正态分布；作出这些假设，会限制机器人控制与计算机视觉交叉领域中的概率框架。

对概率分布进行建模的方法有很多，包括参数化函数的混合模型、直方图、加权粒子集成以及隐式

函数。隐式函数尤其值得关注，因为它们能够表示任意分布；在训练过程中，已有研究成功地使用对比损失和均匀采样的负样本来估计 SO(3) 上的分布，并在推理过程中使用均匀分布的网格。由于维度灾难，将这种方法扩展到 SE(3) 面临两个问题。首先，随着空间从 SO(3) 扩展到 SE(3)，均匀采样的负样本所提供的信息会减少；其次，在 SE(3) 上评估均匀分布的网格，在任何实际可用的分辨率下都将变得成本高昂。

我们提出了 SpyroPose，这是一种新颖的姿态分布估计方法，采用 SE(3)金字塔（一种分层网格），并在金字塔更深层级提供更高的分辨率。图 1 展示了六重对称歧义下估计得到的姿态分布示例。

我们的方法基于三个关键思想：

- 在训练过程中进行重要性采样，这得益于金字塔结构；它能够提供更难的负样本，并降低配分函数估计的方差。

- 在推理时对金字塔进行稀疏评估，将所需评估次数减少数个数量级，从而实现实时姿态分布估计，即使是在 CPU 上也能做到。

- 采用关键点特征提取，在模型中引入相机模型偏置，从而实现平移等变性，并避免依赖单一潜在嵌入来表示复杂的高分辨率分布。

我们在 SYMSOL [21] 和 TLESS [12], 上取得了最先进的旋转分布估计结果；据我们所知，我们还首次给出了 SE(3)分布估计的定量结果。

2. 相关工作

处理视觉歧义是物体位姿估计中的一项挑战。大多数位姿估计工作 [30, 26, 24, 19] 都使用人工标注来显式处理视觉歧义的一种成因——对称性，并在不表达不确定性的情况下推断一个“最佳猜测”位姿估计。事实上，位姿估计基准 BOP [13], 也以点估计和已知物体对称性为中心。

然而，有些提供点估计的方法会以有原则的方式处理视觉歧义。EPOS [11] 对物体表面区域进行密集的逐像素直方图回归，以获得2D-3D对应关系，并在PnP-RANSAC框架中使用这些对应关系，从而在无需预先了解对称性的情况下提供姿态估计。SurfEmb [8] 拓展了这一思路，并估计密集的完整2D-3D对应关系分布。然而，密集的对应关系直方图和分布尚未被证明有助于对姿态分布进行建模。

此外，还有一些处理姿态不确定性的位姿估计方法。在 [28], 中，模型集成被用于估计认知不确定性，该不确定性表示由训练数据不足或领域差异导致的泛化不确定性程度。然而，这并无助于表示位姿估计中固有的偶然歧义。

对偶然不确定性进行建模，传统上采用参数化分布的方法。最常见的不确定性参数化模型是多元正态分布。早期研究聚焦于将对应关系不确定性传播到位姿空间，以用于经典位姿估计方法，例如用于 ICP 的方法 [1] 和用于 PnP 的方法 [4]。然而，这些方法假设对应关系歧义相互独立且服从高斯分布；即便是简单的对称性和遮挡，也可能违背这一假设。

旋转和位置分别定义在不同的流形 SO(3) 和 $\mathbb{R}^3$上，人们通常假设二者相互独立并分别处理，且通常假设位置服从正态分布。对于旋转不确定性，Bingham 是最常用的参数化分布模型。估计 Bingham 混合模型的参数有多种方法，包括基于深度学习的回归 [22, 2, 5] ，以及将 Bingham 混合模型拟合到一组位姿假设 [20]。此外， [25] 表明可以估计 von Mises 混合模型的参数。

最后，还有一些研究探讨了 SO(3) 上的非参数分布。[22, 17]直接回归旋转直方图，能够以较粗的分辨率表示任意分布，但跨区间的泛化能力有限，因此要求训练数据充分覆盖所有区间。这使得将直方图扩展到更高分辨率变得困难，更不用说 SE(3) 了。 [29, 3] 在图像裁剪块上训练去噪自编码器，并利用学习到的潜在嵌入之间的余弦相似度来估计视觉歧义。然而，由于其损失是重建损失而非概率损失，且分布是人为设计而非学习得到的，因此最终得到的分布只是启发式的。

近期，已有许多方法被提出，用于隐式建模非参数分布。在 [15], 中，方法对图像上的关键点投影边缘分布进行估计，从而得到保守的未归一化位姿似然估计；然而，由于没有对关键点的联合分布建模，模型的表达能力有限。该方法也可用于姿态分布估计，但由于缺少在 SE(3) 上进行归一化的方法，他们只展示了旋转分布估计的结果。受本研究启发的 ImplicitPDF [21],通过训练多层感知机（MLP），将图像嵌入和旋转映射为未归一化似然。他们在补充材料中展示了一个基于简单合成数据进行 SE(3) 分布估计的定性结果，但没有给出定量结果，也没有展示真实数据上的结果。

Hyper-PosePDF [14] 与之类似，但他们并非从图像估计潜在嵌入并将其输入 MLP，而是学习从图像到 MLP 权重的映射，由该 MLP 将旋转映射为未归一化似然。I2S [18]也与之类似，但它将图像域的特征映射到 SO(3)，再经过 SO(3) 等变层，之后映射为未归一化旋转似然。他们展现了良好的泛化能力，但分辨率低于相关方法。

与 [21, 14, 18], 类似，我们也采用隐式形式来估计未归一化的对数似然，但我们使用 SE(3)金字塔（一种分层网格），从而能够在训练期间进行重要性采样，高效学习无参数化的 SE(3) 分布。在推理阶段，我们使用同一金字塔进行稀疏评估，以实现实时姿态分布估计。此外，我们还在潜在姿态分布嵌入中利用图像的空间维度，减轻单个图像嵌入表征复杂高分辨率分布的负担。

3. 方法

我们的方法核心是学习不同分辨率下的姿态分布。给定一个姿态假设，我们将物体关键点投影到图像中，在投影点处提取图像特征，并将采样到的特征输入针对不同分辨率设计的多层感知机。在推理阶段，这使我们能够对姿态金字塔进行稀疏的自顶向下评估，只将最有可能的姿态扩展到下一个更高分辨率的层级。训练期间，不同分辨率的模型使我们能够从一种比均匀分布更接近模型分布的姿态分布中，以已知概率采样负样本，从而实现重要性采样。

3.1. 问题定义

给定一张感兴趣物体的图像裁剪 $I \in \mathbb{R}^{H\times W\times C}$，我们的目标是估计该物体六维姿态 $x \in \mathrm{SE(3)}$ 的非参数化分布 $p(x|I)$。

3.2. SE(3) 金字塔定义

我们在 SE(3) 中使用等体积分层网格，并将其称为 SE(3)金字塔。金字塔的每一层都是位置网格与旋转网格的笛卡尔积，分别位于 $\mathbb{R}^3$ 和 SO(3) 中。

对于金字塔的旋转部分，我们使用 HealPix 网格 [6] ，并通过 [31]将其扩展到 SO(3)。与以往工作[21, 14, 15], 一样，我们利用该网格的等体积特性，同时也利用其分层结构。令 $R^{(r)} \subset \mathrm{SO(3)}$ 表示递归层级为 r 的旋转网格，并令 $R_i^{(r)} \in R^{(r)}$ 表示网格中的一个单元，以其中心点表示。最粗层级（第 0 层）由 72 个单元组成，$|R^{(0)}| = 72$；每次递归时，每个单元都会分成八个单元，$|R^{(r)}| = 72 \cdot 8^r$。网格的体积为 $V(R^{(r)}) = \pi^2$，由于该网格是等体积的，因此一个旋转单元的体积为 $V(R_i^{(r)}) = \pi^2/(72 \cdot 8^r)$。

对于位置部分，由于 $\mathbb{R}^3$ 无界，因此需要定义网格的边界。我们分两步定义边界。第一步，使用对视觉歧义的保守估计来建模位置误差，以确保真实姿态位于估计边界内。第二步，定义一个分层网格，使其完全覆盖保守估计的边界。设 $\hat{t} \in \mathbb{R}^{3\times 1}$ 为物体位置 t 的粗略估计。根据具体应用，该估计可以来自检测器、先验已知，或通过其他方式获得。本文中，我们假设 $\hat{t}$ 来自检测器。随后，我们根据 t 为 t 定义一个保守边界。具体而言，我们假定感知到的位置歧义最大值等于物体的半径。设 d 为感兴趣物体的直径。在平行于像平面的方向上，我们允许误差最高达 $d/2$。沿观察方向，我们允许物体与相机之间的距离缩短至一半；假设采用针孔相机模型，这会使物体的成像尺寸大约增大一倍。

形式上，我们通过引入一个截断多元正态随机变量来定义一个满足上述条件的界，$\tilde{e} = \mathcal{N}(0,\ \sigma I,\ 1/2)$，其中“1/2”表示其在 $||\tilde{e}||_2 = 1/2$处被截断，然后将随机误差变量 e定义为

$$e = A\tilde{e}, \quad A = \begin{bmatrix} d & 0 & \hat{t}_x \\ 0 & d & \hat{t}_y \\ 0 & 0 & \hat{t}_z \end{bmatrix}. \tag{1}$$

$\tilde{t} = \hat{t} + e$ 得到的界是一个以 t 为中心、沿观察方向拉长的球体，因此涵盖的深度歧义多于平面内歧义。我们用位置单元的中心表示该位置单元，并将递归 r 时的$p^{(r)} \in \mathbb{R}^{3\times N}$位置网格定义为

$$p^{(r)} = \hat{t} + Ag^{(r)}, \tag{2}$$

其中，$g^{(r)} \in \mathbb{R}^{3\times N}$ 由原点居中的单位立方体内一个 2^r 乘 2^r 乘 2^r 的规则网格中各立方体的中心组成。注意，$p^{(r)}$ 涵盖 $\tilde{t}$，因为 $g^{(r)}$ 涵盖$\tilde{e}$。另请注意，$p^{(r)}$ 具有层次结构且等体积，因为 $g^{(r)}$ 具有层次结构且等体积，并且 $g^{(r)} \mapsto \hat{t} + Ag^{(r)}$ 是一个仿射变换。网格的体积为 $V(p^{(r)}) = \det(A)$，因为 $g^{(r)}$ 的体积为 1。因此，$p^{(r)}$ 中一个单元的体积为 $V(p_i^{(r)}) = \det(A)/8^r$。

SE(3)金字塔就是位置网格和旋转网格的笛卡尔积，但必须选择在哪一级递归中将二者对齐。在 $R^{(0)}$ 处，最近邻之间的角距离约为 $\phi = 1\mathrm{rad}$，导致视觉距离最高约为 $d/2$。由于$p^{(1)}$ 具有相同的视觉分辨率，我们将递归 0 定义为

SE(3)金字塔 $x^{(0)}$，即$R^{(0)}$ 和 $p^{(1)}$的笛卡尔积：$x^{(0)} = R^{(0)} \times p^{(1)}$。由于 $x^{(r)}$ 也是一个等体积网格，因此

$$V(x_i^{(r)}) = \frac{V(x^{(r)})}{|x^{(r)}|} = \frac{\det(A)\pi^2}{(72+8)64^r}. \tag{3}$$

为防止我们的模型学习固定 SE(3) 网格的结构，我们在训练期间随机偏移并旋转 $p^{(1)}$ 并旋转 $R^{(0)}$。

3.3. 对比损失

受噪声对比估计启发的 InfoNCE 损失最初在 [23], 中提出，

$$L_{\text{InfoNCE}} = -\mathop{\mathbb{E}}_{x,I,X}\left[\log \frac{f_\theta(x,I)}{f_\theta(x,I)+\sum_{x_j\in X} f_\theta(x_j,I)}\right], \tag{4}$$

其中，(x, I)从数据分布 $p(x, I)$中采样，X 是一组从噪声分布 $x_j \sim p_n(x)$中采样的 N，而 θ 是模型的参数。他们证明，对于任意 N，该损失会使模型近似于 $f_\theta(x, I) \propto p(x|I)p_n(x)^{-1}$。由此可知，若噪声分布设为均匀分布，$f_\theta$ 便近似于一个未归一化分布，即 $f_\theta(x|I) = \tilde{p}(x|I;\theta) \propto p(x|I)$。

请注意，公式（4）分母中的最后一项$\hat{Z} = \sum_{x_j\in X} f_\theta(x_j, I)$，与配分函数的无偏估计$\hat{\mathbb{E}}_X \hat{Z} \propto \int f_\theta(x, I)dx$成正比。将基于噪声对比损失的方法通过均匀分布采样负样本扩展到更高维度时，存在一个固有问题：由于维度灾难，配分函数估计的方差会增大。

重要性采样可用于降低配分函数估计的方差，但这需要一种接近 $p(x|I;\theta)$ 的重尾分布，能够从中采样且其样本似然度已知，而通常并不存在这样的分布。

3.4. 金字塔模型与重要性采样

我们使用式（4）中的损失，并将涵盖真实姿态 x 的单元设为正样本，该单元位于 $x^{(r)}$ 中。我们不再以单一的高分辨率学习 $\tilde{p}(x|I;\theta)$，而是在不同分辨率下学习分布，金字塔 $\tilde{}$ 的每一层对应一种分辨率，即 $\mathrm{p}(x_i^{(r)}|I;\theta^{(r)})$。对于递归层级为 r 的模型，可以利用较粗层级的模型估计 $\mathrm{p}(x_i^{(r)}|I;\theta^{(r)})$，进而用于重要性采样。

设 $P(x_i^{(r)}) = x_{i\backslash 64}^{(r-1)}$ 为前一次递归中的父节点，其中 $\backslash$ 表示整数除法；设 $C(x_i^{(r)}) = \left\{x_{64i+0}^{(r+1)}、\ldots\ldots、x_{64i+63}^{(r+1)}\right\}$ 为下一次递归中 $x_i^{(r)}$ 的子节点集合。因此，$x_i^{(r)}$ 及其自身在内的兄弟节点为 $S(x_i^{(r)}) = C(P(x_i^{(r)}))$，并设 $S(x_i^{(0)}) = x^{(0)}$。

在以下记法中，为简洁起见，假设参数 $\theta^{(r)}$，并省略以 I 为条件这一设定。我们将兄弟节点之间的相对概率表示为

$$\mathrm{q}(x_i^{(r)}) = \frac{\tilde{\mathrm{p}}(x_i^{(r)})}{\sum_{x_j^{(r)}\in S(x_i^{(r)})}\tilde{\mathrm{p}}(x_j^{(r)})}, \tag{5}$$

并在不同分辨率间递归应用这些模型，便得到一种生成式由粗到细马尔可夫链模型。它类似于自回归模型，但模型针对的是不同分辨率，而非不同维度，

$$\mathrm{p}(x_i^{(r)}) \approx \bar{\mathrm{p}}(x_i^{(r)}) = \mathrm{q}(x_{i\backslash 64^r}^{(0)})\mathrm{q}(x_{i\backslash 64^{r-1}}^{(1)})\cdots\mathrm{q}(x_i^{(r)}). \tag{6}$$

因此，在式（4）的 InfoNCE 损失中，$\bar{\mathrm{p}}$ 可用作重要性采样分布，

$$\hat{Z}_{\text{IS}} = \sum_{x_i^{(r)}\in X}\frac{\tilde{\mathrm{p}}(x_i^{(r)})}{\bar{\mathrm{p}}(x_i^{(r)})}, \quad x_i^{(r)} \sim \bar{\mathrm{p}}(x_i^{(r)}), \tag{7}$$

它与配分函数的无偏估计成正比，如 Z，但方差更低。

请注意，我们可以直接最大化对数似然，并用重要性采样配分函数估计值进行归一化，但初步实验表明，式（4）分母中的额外项使训练更加稳定。直观地说，如果负样本很容易，且 $f_\theta(x, I)$ 是式（4）分母中的主导项，那么梯度将接近于零，$\nabla_\theta L_{\text{InfoNCE}} \approx 0$。如果分母中没有 $f_\theta(x, I)$，情况就不会如此。

3.5. 使用金字塔模型进行推理

推理时，只对金字塔中的一棵稀疏树进行评估，并在概率最高的位置获得最高分辨率。首先，最粗模型会得到所有最粗网格单元上的分布 $\mathrm{p}(x^{(0)}|I;\theta^{(0)})$。令$x_k^{(0)}$ 表示按估计概率排序的前 k 个单元，并令 p_k^0 表示 $x_k^{(0)}$ 的概率之和。然后，将前 k 个单元扩展为其子节点 $C(x_k^{(0)})$，并由下一个模型进行评估，以更高的分辨率重新分配概率 p_k^0：

$$\hat{\mathrm{p}}(x_i^{(r)}) = p_k^{(r-1)}\frac{\tilde{\mathrm{p}}(x_i^{(r)})}{\sum_{x_j^{(r)}\in C(x_k^{(r-1)})}\tilde{\mathrm{p}}(x_j^{(r)})}. \tag{8}$$

重复此过程，直至最后一次递归。所有未被扩展的单元（包括最后一次递归中的所有单元）都是稀疏树中的叶节点。叶节点的概率之和为一，构成估计分布。连续姿态 x 的估计似然由包含 x 的叶节点 $x_i^{(r)}$ 决定，

$$\hat{p}(x) = \hat{\mathrm{p}}(x_i^{(r)})V(x_i^{(r)})^{-1}. \tag{9}$$

表 1. SYMSOL 上的旋转分布估计结果。表中数值为每个物体的 5 k 张测试图像上真实旋转的平均估计对数似然↑。灰线以下的结果基于我们对 SYMSOL I 的实现。为进行验证，我们展示了原始数据集和我们实现的数据集上不含 KP 的结果。IS：重要性采样。KP：关键点。

	SYMSOL I						SYMSOL II			
方法	avg.	cone	cyl.	tet.	cube	ico.	avg.	sphX	cylO	tetX
Prokudin 等人 [25](2018年)	-1.87	-3.34	-1.28	-1.86	-0.50	-2.39	0.48	-4.19	4.16	1.48
Gilitschenski 等人 [5](2019)	-0.43	3.84	0.88	-2.29	-2.29	-2.29	3.70	3.32	4.88	2.90
Deng 等人 [2](2020)	-1.48	0.16	-0.95	0.27	-4.44	-2.45	2.57	1.12	2.99	3.61
ImplicitPDF [21] (2021)	4.10	4.45	4.26	5.70	4.81	1.28	7.57	7.30	6.91	8.49
I2S [18] (2023)	3.41	3.75	3.10	4.78	3.27	2.15	4.84	3.74	5.18	5.61
HyperPosePDF [14] (2023)	5.78	5.74	4.73	7.04	6.77	5.10	7.72	7.73	7.12	8.53
我们的方法（不含 KP）和重要性采样	5.65	6.77	6.07	6.04	6.23	3.12	7.16	6.96	7.59	6.92
我们的方法（不含 KP）	**7.33**	**7.62**	**6.46**	**8.69**	**8.63**	**5.23**	**9.27**	**9.07**	**9.32**	**9.41**
我们的方法（不含 KP）	7.12	7.37	6.54	8.39	8.62	4.70				
我们的方法（不使用重要性采样）	8.19	7.40	6.69	10.04	8.82	7.99				
我们的方法（使用立方体关键点）	8.86	7.55	**7.06**	9.60	**10.58**	**9.50**				
Ours	**8.97**	**7.67**	6.96	**11.04**	10.10	9.06				

请注意，与公式（6）中的重要性采样分布不同，后者用于训练；在推理期间，给定递归层级中各单元之间的相对概率完全由该层级的模型决定，不受前面递归层级概率的影响。这使得低分辨率下的单元边界歧义能够在更高分辨率下得到消解。

3.6. 网络架构

我们使用带有 ResNet 骨干网络的 UNet [27]，以获得一个 64 维特征图，其空间分辨率与图像相同。从物体的三维网格中，通过最远点采样选取 16 个近似均匀分布的关键点。给定一个位于姿态单元中心的姿态，这些关键点会被投影到图像中，并通过双线性插值从特征图中提取投影点处的特征。投影到图像外的关键点会获得一个学习得到的图像外嵌入。采样得到的关键点特征经过拼接后，输入一个具有 256 个隐藏神经元的三层多层感知机。网络的输出是一个标量，表示对姿态单元未归一化对数似然的估计。

4. 实验

我们展示了 SO(3) 分布估计的结果，并与先前的工作进行比较；此外，据我们所知，我们首次展示了 SE(3) 分布估计的定量结果。我们在 SYMSOL I、SYMSOL II 和 TLESS 上展示了 SO(3) 的结果，并在 TLESS 和 HB 上展示了 SE(3) 的结果。最后，我们展示了 SpyroPose 的一种直接的多视角扩展，与单视角结果相比，该方法取得了显著提升，表明使用未参数化分布进行传感器融合具有潜力。

4.1. SO(3) 结果

SYMSOL。 ImplicitPDF [21] 引入了合成数据集 Symmetric Solids（SYMSOL），用于评估 SO(3) 分布估计。该数据集包含多种具有不同对称性的几何基元。该数据集分为两部分：SYMSOL I 包含无纹理物体，SYMSOL II 则包含标记，这些标记仅在特定旋转下可见，从而造成动态歧义。

该数据集不包含相机内参或物体的三维模型。虽然物体的平移是固定的，但其平移量未知。由于我们的方法基于关键点投影，因此无法直接应用于原始 SYMSOL 数据集。为此，我们在原始数据集上使用不含关键点的修改版方法，并实现了相机内参、平移量和三维模型已知的 SYMSOL I，以评估完整方法。

对于不使用关键点的方法，我们采用与 ImplicitPDF [21], 类似的架构，使用 ResNet50 [10] 获取潜在图像嵌入，对旋转进行位置编码，并将这两种嵌入输入一个小型多层感知机。详情请参见[21]。尽管不使用关键点的方法与 ImplicitPDF 的架构相似，但我们的方法仍为 SO(3) 金字塔的每一层都设置了一个多层感知机，并展示了使用和不使用重要性采样的结果。对于 SYMSOL I 的实现，我们尽可能匹配原始数据集的透视效果、尺度和着色器。由于 SYMSOL II 包含纹理，因此不易实现公平比较。

在完整方法中，我们从三维模型中通过最远点采样选取关键点；但在某些应用中可能无法获取三维模型，因此我们也展示了将关键点选在两个

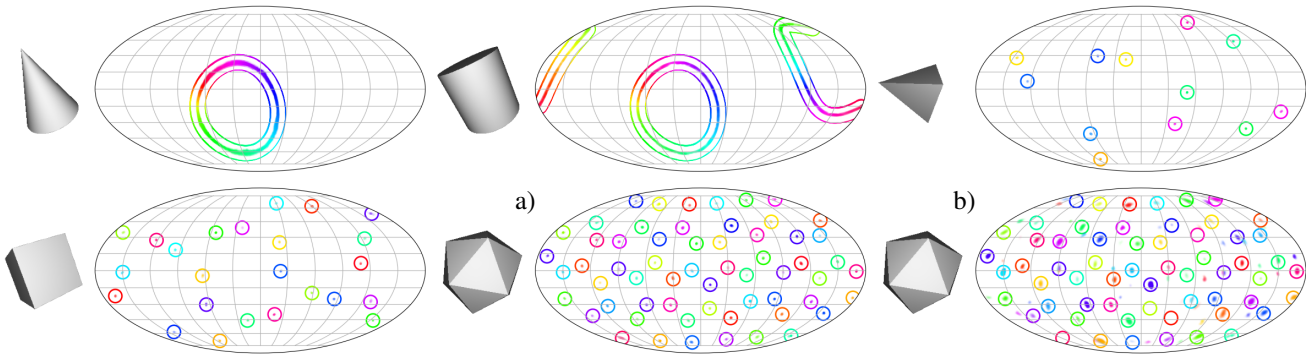


图 2. SYMSOL I 的定性结果。我们将金字塔最后一层（第 6 层）的旋转及其似然度可视化为 alpha 值，并进行归一化以便观察。圆形（对于连续对称性则为环形）表示在对称性意义下正确的旋转。a) 和 b) 来自同一张图像，但 b) 展示的是不使用关键点的我们的方法。我们的方法准确捕捉到了二十面体的全部 60 个模态。

边长分别为物体直径的 1 倍和 0.5 倍的立方体顶点处时的结果。

我们使用七个多层感知机训练特定物体模型，其范围从 $R^{(0)}$ 到 $R^{(6)}$，并且每张图像在每次递归中使用 1024 个负样本。对于采用重要性采样的模型，这通过 128 条采样轨迹实现；由于旋转金字塔的分支因子为 8，每次递归中还包括兄弟节点，因此每个多层感知机对应 1024 个负样本。参见第 3.2 节和第 3.4 节。请注意，尽管训练集中因对称性包含多个旋转标注，但训练时每张图像仅使用所提供标注中的第一个，并假设我们不了解对称性。对于完整方法，我们使用 4 的批量大小。由于不使用关键点的模型计算成本较低，我们将其批量大小设为 16，以获得相近的训练时间。我们所有的 SO(3) 模型均训练 50 k 次迭代；在单块 RTX 2080 上，每个物体约需 2 小时。

结果见表 1，定性示例见图 2。我们的方法在 SYMSOL I 和 SYMSOL II 的所有物体上均取得了最先进的结果。对于真实旋转，我们的方法预测的平均似然度分别是 HyperPosePDF [14] 和 ImplicitPDF [21]，的 24 倍和 130 倍。从物体表面采样关键点需要假设三维模型可用，但我们使用立方体关键点取得的结果几乎同样出色，消除了这一假设。

使用关键点带来了显著提升。对于不使用关键点、且架构与 ImplicitPDF 类似的模型，视觉模型产生的非空间潜在嵌入必须表达复杂且高分辨率的分布。提取关键点特征后，模型可以利用图像空间作为分布的中间表示，并实现平移等变性。

由于 SO(3) 金字塔的评估是稀疏的，因此即使分布细化到递归层级 6、包含 18.8 M 个旋转，使用 $k = 512$ 也只需进行 21 k 次评估，比评估完整网格所需的评估次数少近三个数量级。这使我们的方法能够实时运行，即使是在 CPU 上，

表 2. 单张图像的推理时间比较。对于我们的方法，使用 Intel i9-9820X CPU 和 Nvidia RTX 2080 GPU。批处理可进一步提高帧每秒，在 SO(3) 上使 GPU 上不使用关键点的我们的方法达到 241 帧每秒。评估次数：函数评估次数。

空间	方法	网格尺寸	评估次数	dev.	fps
SO(3)	ImplicitPDF	2.3 M	2.3 M	gpu	2.4
	我们的方法（不含 KP）	18.9 M	21 k	cpu	16.9
				gpu	53.5
	Ours			cpu	3.3
				gpu	48.3
SE(3)	Ours	618 B	164 k	cpu	0.5
				gpu	16.1

其运行速度也快于在图形处理器上运行的 ImplicitPDF。参见表 2。

虽然金字塔使我们能够高效地评估递归层数超过 6 的情况，但即使在递归层数为 5 时，我们的方法也达到了最先进水平，与 ImplicitPDF 相当；在 sphX、cylO 和 tetX 上的平均对数似然分别为 8.58、8.42 和 8.80。

关键点和重要性采样都有助于在更深的递归层数下进行学习。见图 3。不过，不使用关键点和重要性采样时，对数似然在递归层数约为 5 时趋于平稳；而我们的完整方法或许还能从更多递归中获益。

ImplicitPDF 和 HyperPosePDF 在所有 SYMSOL I 物体上训练一个模型，而我们为每个物体分别训练一个模型。不过，我们在训练期间为每个物体进行的函数求值次数更少；在 SYMSOL II 上，我们也取得了类似的改进，而他们在该数据集上同样为每个物体分别训练一个模型。

泛化。 I2S [18] 在低数据量设置下应用了他们的方法和 ImplicitPDF 的方法，仅使用 1 万张图像而非 4.5 万张图像在 SYMSOL 上训练。I2S 的 ImplicitPDF 模型表现不佳，似乎严重过拟合，因此我们不再重复报告 I2S 的 ImplicitPDF 结果，而是指出，我们的方法（不含 KP）和重要性采样与 ImplicitPDF 非常相似，在完整数据集上的表现也相近。我们使用

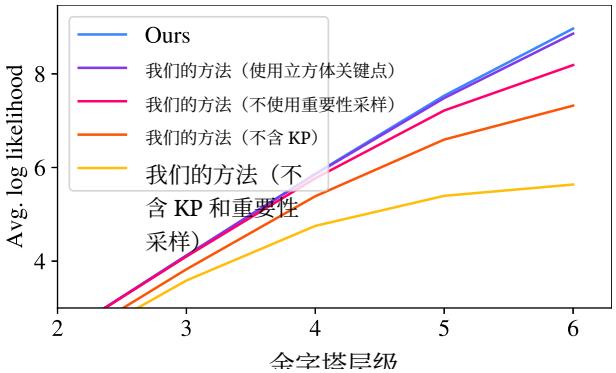


图 3. SYMSOL I 上的对数似然，按物体取平均，并展示金字塔不同递归层级的结果。关键点和重要性采样均能改善较深层级的学习效果。

表 3. 在低数据量场景下，使用 1 万张训练图像（而非 4.5 万张）对 SYMSOL II 进行旋转分布估计。对数似然 ↑。

方法	avg.	sphX	cylO	tetX
I2S [18] (2023)	3.61	3.12	3.87	3.84
我们的方法（不含 KP）和重要性采样	3.95	3.56	4.68	3.60
我们的方法（不含 KP）	**5.47**	**5.38**	**6.82**	**4.19**

表 4. 在 TLESS 上进行旋转分布估计。

方法	LL ↑
Prokudin 等人 [25] (2018年)	8.8
Gilitschenski 等人 [5] (2019)	6.9
ImplicitPDF [21] (2021)	9.8
我们的方法（不含 KP）和重要性采样	10.3
我们的方法（不含 KP）	**11.9**

训练图像中的 9500 张训练我们的模型（分别采用和不采用重要性采样），剩余 500 张用于早停。结果见表 3。我们的方法优于 I2S。

TLESS. 我们遵循 ImplicitPDF 的做法，为全部 30 个物体训练一个单一模型。结果见表 4。我们的方法得到的似然度几乎比 ImplicitPDF 高出一个数量级。值得注意的是，我们的方法在不使用 KP 和重要性采样（IS）的情况下，递归 5 的平均对数似然为 9.9，与 ImplicitPDF 相近。这项基准测试使用 Kinect 训练图像中单独物体的紧密裁剪图像，背景均匀且没有遮挡；训练集和测试集均通过随机划分得到。因此，该基准测试不存在领域差异，也无需在平移方面进行泛化，或处理遮挡和背景杂乱。这与下一节中的实验形成对比。

4.2. TLESS 和 HB 上的 SE(3) 结果

我们在 TLESS [12] 和 HB [16]上训练 SE(3)金字塔模型，递归层级最高达到 5，并从中选取四个物体其中，TLESS 的物体涵盖不同程度的对称性，HB 的物体则是位姿估计领域中常见的四种物体。我们使用来自 [13] 的基于物理的渲染（PBR）进行训练，并展示模型在留出的 PBR 场景和真实图像上的结果。对于 TLESS，我们展示真实测试图像上的结果；对于 HB，由于测试标注未公开，我们展示真实验证图像上的结果。结果见表 5，TLESS 上的定性示例见图 4。

所选 TLESS 物体的训练图像裁剪块数量分别为 16 k、19 k、23 k 和 24 k；HB 物体则分别为 25 k、25 k、24 k 和 23 k。请注意，与 SYMSOL 相比，每个物体的图像数量更少，而且存在仿真到现实差距，我们还试图学习 SE(3) 上的分布。鉴于上述原因，我们在多层感知机中使用丢弃，并在训练期间进行大量数据增强，以此对模型进行正则化。我们根据式（6），每次递归使用 2048 个负样本，对应于沿金字塔进行的 32 条轨迹。

由于我们是首个展示 SE(3) 分布定量结果的工作，因此没有可直接比较的基线；不过，我们展示了使用和不使用重要性采样的结果，以及 SE(3) 网格上均匀分布的结果。我们考虑过与点估计器进行比较，即在点估计周围添加高斯分布，但这很容易比均匀分布表现更差，因为如果估计到了错误的模态，就会受到严重惩罚。

我们尝试将联合分布与解耦版本进行比较，但未能成功训练出不使用关键点的旋转模型；这类模型类似于 ImplicitPDF，在这些更具挑战性的图像上难以训练。事实上，对完整方法得到的 SE(3) 分布中的 SO(3) 进行边缘化后，在四个 T-LESS 物体上的真实旋转平均对数似然为 5.8；而不使用关键点以及不使用关键点和重要性采样（类似于 ImplicitPDF）时，结果分别为 0.5 和 0.1。作为对比，均匀分布对应的数值为 -2.3。因此，学习联合分布使得我们能够采用一种架构，显著改善了这一更具挑战性的基准上的边缘分布甚至。

在留出的基于物理渲染图像上，与未使用重要性采样相比，使用重要性采样时，我们的模型对真实姿态的似然预测高出 220 倍；而在 SYMSOL I 上仅高出 2 倍。这与图 3 一致，图中表明，当概率分布更加集中时，重要性采样会更有益，因为此时对均匀样本进行评估所提供的信息较少。

4.3. 多视图

最后，我们展示一个直接的应用：多视图位姿估计。由于不确定性是在位姿空间而非图像空间中表示的，因此这种方法非常适合进行有理论依据的传感器融合，结合多个来源的信息。利用多视角裁剪和已知外参，我们令 $A = dI$，使得 $p^{(r)}$ 是一个立方网格，并使用相同的

表 5。对 TLESS 和 HB 各自的四种代表性物体进行 SE(3) 分布估计。模型在合成数据（PBR）上训练，并在留出的基于物理渲染图像和真实图像上报告结果。表中数值为真实姿态的平均对数似然。

		TLESS					HB				
Data	方法	avg.	1	14	25	27	avg.	2	7	9	21
PBR	均匀分布	2.7	3.5	3.2	2.4	1.7	0.8	0.4	0.5	1.4	0.9
	我们的方法（不使用重要性采样）	16.7	16.5	16.6	17.8	16.0	15.9	16.3	15.1	16.4	15.8
	Ours	21.6	22.4	20.5	22.6	20.9	21.7	21.7	21.2	21.8	21.9
Real	均匀分布	2.7	3.5	3.3	2.4	1.7	0.9	0.6	0.6	1.5	1.0
	我们的方法（不使用重要性采样）	16.6	14.1	15.9	19.2	17.2	16.0	18.6	16.5	13.2	15.8
	Ours	18.8	16.9	17.5	20.8	20.0	18.9	21.3	20.6	16.4	17.3

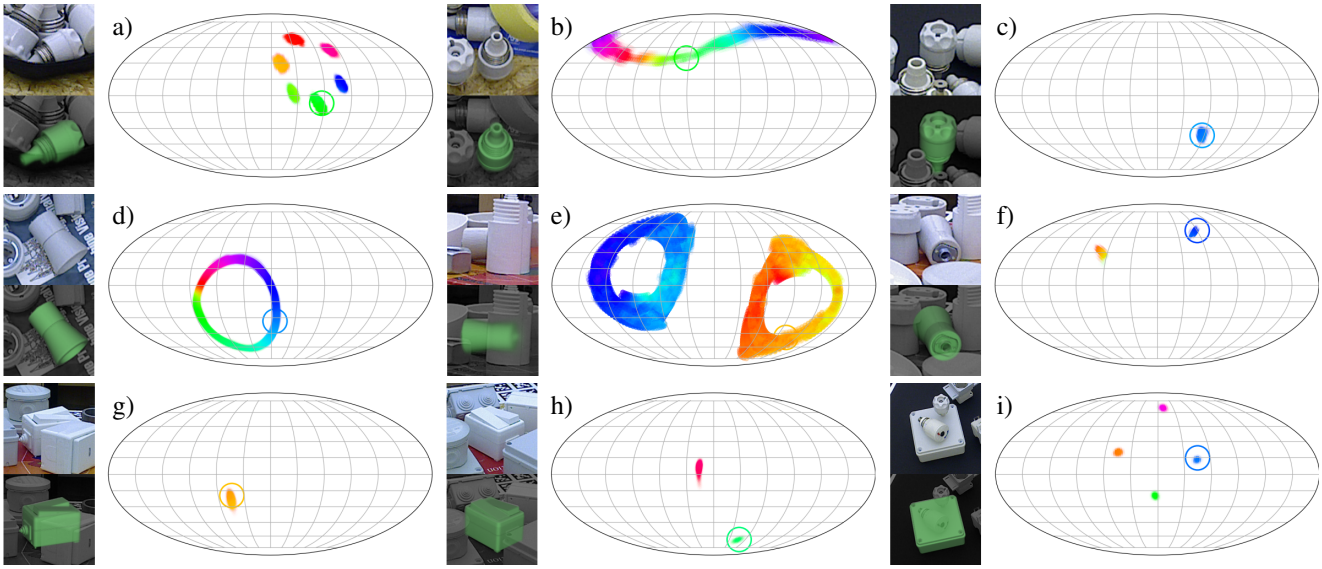


图 4. TLESS 上的 SE(3) 分布。第一行展示物体 1 的分布。a）六重旋转对称性。b）连续旋转对称性。c）无对称性。第二行展示物体 14 的分布。d）连续旋转对称性。e）目标物体位于前景物体后方。二重旋转对称性和连续旋转对称性。请注意，这两个离散模态的深度不同，只有联合分布才能表示这种情况。f）物体末端现在可见的特征消除了连续旋转对称性的歧义，仅保留沿同一轴的二重旋转对称性。g）和 h）分别展示物体 25 的无对称性和二重旋转对称性。i）展示物体 27 的四重旋转对称性。

表 6. TLESS 上的 SE(3) 分布估计结果。LL↑。

方法	avg.	1	14	25	27
Ours	18.8	16.9	17.5	20.8	20.0
我们的方法（使用 $A = dI$）	18.4	16.4	16.6	20.9	19.9
我们的方法（使用多视图）	25.2	23.7	23.6	28.2	25.2

网格，在各视图间的公共坐标系中进行推理。在推理时，对于金字塔中的每次递归，只需对各视图的未归一化对数似然取平均。表 6 展示了 TLESS 上的多视图姿态分布估计结果，使用与 [19, 9] 中相同的视图组合，最多包含四个视图。我们注意到，仅使用 $A = dI$ 并不能提升性能。事实上，推理时深度方向的分辨率过高可能有害，因为大多数节点都会被用于表示深度歧义。这个简单的扩展使真实姿态的似然度提高了近三个数量级。

5. 结论

本文提出了 SpyroPose，一种用于 SE(3) 上姿态分布估计的新方法。我们的方法基于分层 SE(3) 网格（金字塔），通过在不同分辨率层级上学习姿态分布，实现了训练时的高效学习重要性采样和推理时的稀疏评估，从而支持实时姿态分布估计。我们的方法在 SYMSOL 和 TLESS 数据集上的 SO(3) 旋转分布估计方面优于最先进的方法；据我们所知，我们还首次给出了 SE(3) 上姿态分布估计的定量结果。最后，通过对 SpyroPose 进行简单的多视图扩展，我们展示了姿态分布如何轻松融合来自多个来源的信息，表明我们的方法作为未来工作的核心组件具有巨大潜力。

参考文献

[1] Andrea Censi。ICP 协方差的精确闭式估计。载于 *2007*年 *IEEE* 国际机器人与自动化会议，第 3167–3172 页，2007 年 4 月。[2] Haowen Deng、Mai Bui、Nassir Navab、Leonidas Guibas、Slobodan Ilic 和 Tolga Birdal。深度 Bingham 网络：处理姿态估计中的不确定性与歧义。国际计算机视觉期刊，第 1–28 页，2022 年。[3] Xinke Deng、Arsalan Mousavian、Yu Xiang、Fei Xia、Tim- othy Bretl 和 Dieter Fox。PoseRBPF：用于六自由度物体姿态跟踪的 Rao–Blackwell化粒子滤波器。*IEEE*机器人学汇刊，37(5):1328–1342，2021 年。[4] Marek Franaszek 和 Geraldine S Cheok。三维刚体方向不确定性向其点的传播。载于*IEEE*国际计算机视觉会议研讨会论文集，第 2183–2191 页，2017 年。[5] Igor Gilitschenski、Roshni Sahoo、Wilko Schwarting、Alexander Amini、Sertac Karaman 和 Daniela Rus。基于 Bingham 损失的深度方向不确定性学习。载于国际学习表征会议，2019 年。[6] Krzysztof M Gorski、Eric Hivon、Anthony J Banday、Ben- jamin D Wandelt、Frode K Hansen、Mstvos Reinecke 和 Matthia Bartelmann。HEALPix：一种用于球面分布数据的高分辨率离散化与快速分析框架。《天体物理学杂志》，622(2):759，2005 年。[7] Frederik Hagelskjær、Aljaˇz Kramberger、Adam Wolni- akowski、Thiusius Rajeeth Savarimuthu 和 Norbert Kr¨uger。面向先进工业装配的夹爪手指设计与姿态估计流程联合优化。载于*2019*年 *IEEE/RSJ* 智能机器人与系统国际会议（*IROS*），第 2022–2029 页。IEEE，2019 年。[8] Rasmus Laurvig Haugaard 和 Anders Glent Buch。Surfemb：利用学习得到的表面嵌入实现物体姿态估计的稠密连续对应关系分布。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 6749–6758 页，2022 年。[9] Rasmus Laurvig Haugaard 和 Thorbjørn Mosekjær Iversen。基于对应关系分布与极线几何的多视图物体姿态估计。*arXiv* 预印本 *arXiv:2210.00924*，2022 年。[10] 何恺明、张祥雨、任少卿和孙剑。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。[11] Tomas Hodan、Daniel Barath 和 Jiri Matas。EPOS：估计具有对称性的物体的六自由度姿态。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 11703–11712 页，2020 年。ˇ[12] Tom´aˇs Hodan、Pavel Haluza、Step´an Obdrˇz´alek、Jiri Matas、Manolis Lourakis 和 Xenophon Zabulis。T-LESS：用于无纹理物体六自由度姿态估计的 RGB-D 数据集。载于*2017*年 *IEEE*冬季计算机视觉应用会议（*WACV*），第 880–888 页。IEEE，2017 年。[13] Tom´aˇs Hodaˇn、Martin Sundermeyer、Bertram Drost、Yann Labb´e、Eric Brachmann、Frank Michel、Carsten Rother 和 Jiˇr´ı Matas。BOP挑战赛2020：六自由度物体定位。载于欧洲计算机视觉会议，第577–594页。施普林格，2020年。[14] Timon H¨ofer、Benjamin Kiefer、Martin Messmer和Andreas Zell。Hyperposepdf-hypernetworks：预测SO(3)上的概率分布。收录于*IEEE/CVF*冬季计算机视觉应用会议，第2369–2379页，2023年。[15] Thorbjørn Mosekjær Iversen、Rasmus Laurvig Haugaard和Anders Glent Buch。Ki-pode：基于关键点的刚体隐式姿态分布估计。收录于第*33*届英国机器视觉会议（*BMVC 2022*），英国伦敦，*2022*年*11*月*21*日至*24*日，第222页。英国机器视觉协会出版社，2022年。[16] Roman Kaskman、Sergey Zakharov、Ivan Shugurov和Slobodan Ilic。HomebrewedDB：用于三维物体6D姿态估计的RGB-D数据集。收录于*IEEE/CVF*国际计算机视觉会议研讨会论文集，第0–0页，2019年。[17] Wadim Kehl、Fabian Manhardt、Federico Tombari、Slobodan Ilic和Nassir Navab。SSD-6D：让基于RGB的三维检测和六自由度姿态估计再度出色。收录于*IEEE*国际计算机视觉会议论文集，第1521–1529页，2017年。[18] David M Klee、Ondrej Biza、Robert Platt和Robin Walters。图像到球面：学习等变特征以实现高效姿态预测。*arXiv*预印本 *arXiv:2302.13926*，2023年。[19] Yann Labb´e、Justin Carpentier、Mathieu Aubry和Josef Sivic。CosyPose：一致的多视图多物体6D姿态估计。收录于欧洲计算机视觉会议，第574–591页。施普林格，2020年。[20] Fabian Manhardt、Diego Martin Arroyo、Christian Rupprecht、Benjamin Busam、Tolga Birdal、Nassir Navab和Federico Tombari。解释视觉数据中物体检测和六自由度姿态的歧义。收录于*IEEE/CVF*国际计算机视觉会议论文集，第6841–6850页，2019年。[21] Kieran Murphy、Carlos Esteves、Varun Jampani、Srikumar Ramalingam和Ameesh Makadia。Implicit-PDF：旋转流形上概率分布的非参数表示。*arXiv*预印本 *arXiv:2106.05965*，2021年。[22] Brian Okorn、Mengyun Xu、Martial Hebert和David Held。学习用于物体姿态估计的方向分布。收录于*2020*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第10580–10587页。IEEE，2020年。[23] Aaron van den Oord、Yazhe Li和Oriol Vinyals。利用对比预测编码进行表征学习。*arXiv*预印本 *arXiv:1807.03748*，2018年。[24] Sida Peng、Yuan Liu、Qixing Huang、Xiaowei Zhou和Hujun Bao。PVNet：用于六自由度姿态估计的逐像素投票网络。收录于*IEEE/CVF*计算机视觉与模式识别会议论文集，第4561–4570页，2019年。[25] Sergey Prokudin、Peter Gehler和Sebastian Nowozin。深度方向统计：带不确定性量化的姿态估计。收录于欧洲计算机视觉会议（*ECCV*）论文集，第534–551页，2018年。[26] Mahdi Rad和Vincent Lepetit。BB8：一种可扩展、精确且能应对部分遮挡的方法，用于预测三维

在不使用深度信息的情况下估计具有挑战性的物体姿态。收录于*IEEE*国际计算机视觉会议论文集，第3828–3836页，2017年。[27] Olaf Ronneberger、Philipp Fischer和Thomas Brox。U-Net：用于生物医学图像分割的卷积网络。收录于医学图像计算与计算机辅助干预——*MICCAI 2015*：第*18*届国际会议，德国慕尼黑，*2015*年*10*月*5*日至*9*日，论文集，第*III*部分第*18*卷，第234–241页。施普林格，2015年。[28] Guanya Shi、Yifeng Zhu、Jonathan Tremblay、Stan Birchfield、Fabio Ramos、Animashree Anandkumar和Yuke Zhu。深度物体姿态估计的快速不确定性量化。收录于*2021*年*IEEE*国际机器人与自动化会议（*ICRA*），第5200–5207页。IEEE，2021年。[29] Martin Sundermeyer、Zoltan-Csaba Marton、Maximilian Durner、Manuel Brucker和Rudolph Triebel。通过RGB图像进行六自由度物体检测的隐式三维方向学习。收录于欧洲计算机视觉会议（*ECCV*）论文集，第699–715页，2018年。[30] Yu Xiang、Tanner Schmidt、Venkatraman Narayanan和Dieter Fox。PoseCNN：用于杂乱场景中六自由度物体姿态估计的卷积神经网络。2018年。[31] Anna Yershova、Swati Jain、Steven M Lavalle和Julie C Mitchell。利用Hopf纤维化生成SO(3)上的均匀增量网格。《国际机器人研究杂志》，29(7):801–812，2010年。