arXiv:2305.01191v2 [cs.RO] 7 Nov 2023

EasyHeC：通过可微渲染与空间探索实现精准且自动化的手眼标定

Linghao Chen[1,2] Yuzhe Qin[2] Xiaowei Zhou[1] Hao Su[2†]
浙江大学 [2]加利福尼亚大学圣迭戈分校

摘要—手眼标定是机器人学中的一项关键任务，因为它直接影响操作和抓取等关键操作的效果。实现这一目标的传统方法需要仔细设计关节位姿并使用专用的标定标记，而近期大多数仅使用位姿回归的学习方法在诊断误差方面的能力有限。在本文中，我们提出了一种新的手眼标定方法，称为 EasyHeC。该方法无需标记、具有白盒特性，并能提供更高的精度和鲁棒性。我们提出采用两项关键技术：基于可微渲染的相机位姿优化和基于一致性的关节空间探索，从而实现对标定过程的精准端到端优化，并免除繁琐的机器人关节位姿手动设计。我们的评估表明，该方法在合成数据集和真实世界数据集上均表现出优越性能，并通过为定位和交互物体提供精确的相机位姿，提升了下游操作任务的效果。代码可在项目页面获取：https://ootts.github.io/easyhec。

I. 引言

手眼标定是机器人学中一项历史悠久的基础性问题，旨在确定相机与机器人之间的变换关系。它将机器人感知的两种最重要模态——本体感知与外部感知——进行对齐。因此，手眼标定是各种视觉机器人应用至关重要的第一步。

传统标定方法依赖标记（例如 ArUco 标记 [1] 或棋盘格图案 [2, 3],）来建立机器人和相机坐标系之间的关联。尽管这些方法具有坚实的理论基础，但在实际机器人系统中实施时仍会遇到诸多挑战。研究表明，传统方法的精度会受到标记固定质量、标记定位精度以及关节位姿选择等因素的显著影响。至关重要的是，针对不同任务改变相机布置需要创建不同的关节位姿轨迹，而这项工作极具挑战性。

对人类而言颇具挑战。尽管行业用户可以通过针对特定任务进行大量调试来获得令人满意的精度，但传统的基于标记的方法难以满足在日常环境中开发精度高且只需极少人工干预的通用机器人的需求。因此，亟需一种精确、自动化的手眼标定方案，既易于实施，又便于用户使用。随着

本工作完成于 Linghao Chen 在加州大学圣地亚哥分校访学期间。通讯作者：Hao Su。

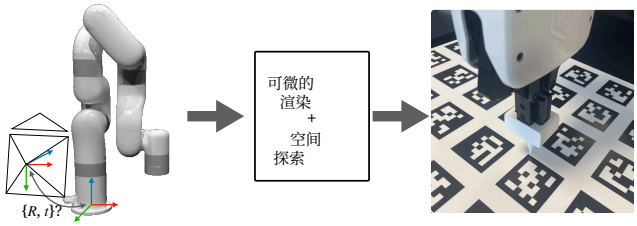


图 1：**我们提出了用于手眼标定的 EasyHeC**。左：手眼标定问题。我们的目标是估计相机与机器人基座之间的相对位姿。中：我们的方法 EasyHeC 的主要技术是可微渲染（III-C）和空间探索（III-D）。右：使用真实机械臂进行高精度目标定位。

计算机视觉算法在各类机器人学任务中的应用日益广泛，这一需求显得尤为迫切。

近来，研究者开始开发无标记手眼标定技术 [4]–[6],，利用基于学习的方法降低标定所需的成本和人力。然而，这些方法的性能在很大程度上取决于训练数据的规模和质量。无论是传统方法还是基于学习的方法，在机器人不同构型下，相机空间中的手部定位误差分布都不均匀。与采样构型相近的构型能够达到更高的精度，而在标定过程中未探索的构型空间区域则会出现更大的偏差。此外，大多数基于学习的方法依赖直接回归，缺乏透明度，因此在实际应用中难以诊断标定误差。

在这项工作中，我们旨在简化机器人学研究人员的手眼标定。为此，我们提出了一个名为 EasyHeC 的系统，它具有以下几项关键特性：

- 无需在机器人上安装标记；
- 采用白盒设计，便于诊断；
- 无需为标定费力手动设计机器人关节位姿；
- 精度高且鲁棒性强；

为借助 EasyHeC 实现这一目标，我们提出采用两项关键技术：(i) 基于可微渲染的相机位姿优化，以及 (ii) 基于一致性的关节空间探索。

首先，我们将基于变换等价性构建的广泛使用的目标函数（例如 $AX = XB$ [3,）替换为

7, 8], 基于机械臂逐像素分割的损失函数。我们假设机械臂的三维模型可直接获得；鉴于大多数常用机械臂都可以在网上找到，这一假设是合理的。在我们的公式中，通过最小化三维机械臂模型的投影与观测到的分割掩码之间的差异，来优化相机与机器人基座坐标系之间的变换。这种白盒方法在相机观测结果与相机位姿之间建立了直接联系，从而能够对标定过程进行端到端优化，并使我们的系统具有很强的可诊断性。此外，通过利用机械臂本身的形状，我们无需使用标定标记，并降低了手眼标定过程的经济成本。由于机械臂的尺寸远大于附着在其上的标记，机械臂的掩膜为估计变换提供的证据，明显多于传统的标记点。

其次，我们开发了一种基于一致性的联合空间探索模块，用于帮助选择信息量丰富的关节位姿。该模块通过评估机器人关节位姿与一组相机位姿候选的一致性，找出信息量最高的机器人关节位姿。这一自动化选择过程无需手动设计机器人关节位姿。此外，由于该过程在仿真中进行，因此无需成本，并且可以反复执行，以提高标定精度。

我们在多个合成数据集和真实世界数据集上评估了我们的方法，并对所提出系统的不同设计选择进行了消融研究。在合成数据集上的评估结果表明，**EasyHeC在单个视角下的精度是传统基线方法的 4 倍**，并且在 5 个视角下的精度约为 0.2 厘米。在一个公开可用的真实世界数据集上进行评估时，EasyHeC 在所有指标上均持续展现出优异性能。此外，我们还展示了所提出的手眼标定系统能够通过提供精确的相机位姿，帮助定位物体并与其交互，从而提升下游操作任务的表现。我们的系统易于部署且便于使用，并已开源，以造福机器人学界。

II. 相关工作

传统手眼标定。 手眼标定（HeC）问题涵盖两种设置：（i）眼在手外，即“眼睛”相机固定在世界坐标系中；（ii）眼在手上，即相机固定在“手”上，通常是末端执行器。传统的手眼标定方法需要对特定标记 [1]进行显式跟踪。根据手眼标定的设置不同，标记在眼在手外的情况下附着于机器人末端执行器上，或在眼在手上的情况下固定于地面上。随后，通过求解线性系统 [2, 3, 7, 8] 获得相机位姿。然而，这些方法要求标记始终完全可见，这使得标定轨迹的设计更加复杂。为了解决这一局限性，研究人员提出了无标记方法 [9]–[11] ，这些方法使用运动恢复结构（SfM）用于估计眼在手上的相机运动。然而，这类方法依赖丰富的视觉特征进行匹配，因此不适用于眼在手外场景下的机器人运动跟踪。与上述方法不同，我们的工作无需使用标记即可解决眼在手外的 HeC 问题。我们的方法并非求解线性系统，而是将 HeC 构建为一个迭代优化问题，通过最小化机器人掩码的逐像素差异来求解。

Learning-based Hand-eye Calibration. Several 近期有若干方法提出采用基于学习的方法解决手眼标定问题。例如，DREAM [12] 使用深度神经网络预测机器人关节关键点，然后利用透视-点（PnP）算法 [13] 恢复相机位姿。Labb´e 等人 [4] 将神经网络用作细化模块，根据渲染图像和观测结果迭代预测机器人的状态增量。然而，这些基于回归的方法未能建立相机位姿与观测之间的直接联系，从而限制了其在实际应用中的诊断能力。Valassakis 等人 [14] 直接从图像回归相机位姿，但在实际场景中缺乏鲁棒性。Sefercik 等人 [5]使用神经网络进行末端执行器分割，并采用迭代最近点（ICP）算法 [15]估计相机位姿，但依赖深度传感器，且在缺乏高质量深度数据时效果可能不佳。Lu 等人 [16] 提出先通过迭代过程检测一组最优关键点，然后利用 PnP 求解相机与机器人基座之间的变换。我们的方法与其主要区别在于，我们使用可微渲染器优化位姿，无需定义关键点。Bahadir 等人 [6]跟踪机器人上的一个参考点，并回归相机位姿。近期一项工作 [17] 提出了与我们类似的思路，即利用可微渲染解决手眼标定问题。然而，我们的方法与其有两点主要区别。首先，他们尝试使用单张图像同时优化相机位姿和机器人形状。这种方法会给优化过程带来歧义，因为对机器人形状或相机与机器人距离的微小调整，都可能生成相似的图像。最终，他们未能实现高精度。相比之下，我们作出合理假设，即机器人形状已知，并且只将相机位姿作为变量进行优化。这一假设显著提高了我们方法的精度。其次，他们将机械臂视为静态刚体，没有充分利用机器人的关节结构。相比之下，我们使用空间探索模块，在保持流程完全自动化的同时，显著提高了标定结果的精度。

下一最佳视角选择。 下一最佳视角选择问题已在机器人学领域得到广泛应用。Zhang 等人 [18]引入了离散视角质量场，以寻找最具代表性的相机视角。Yang 等人 [19] 提出将标定不确定性估计作为确定能够提供

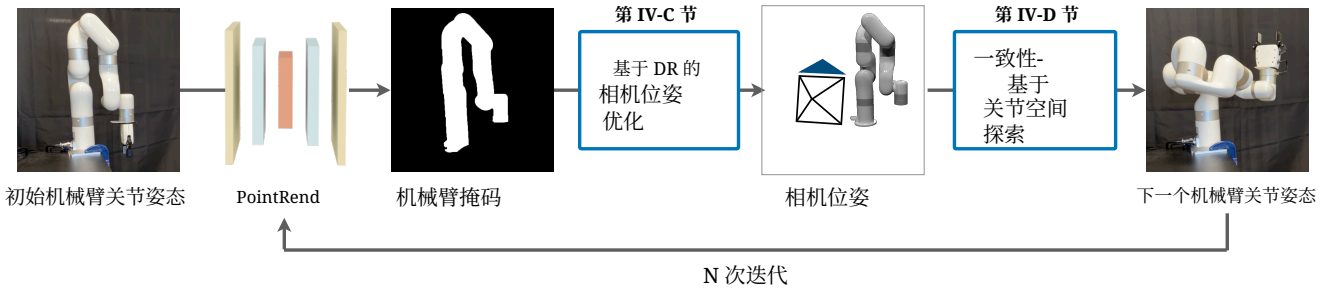


图 2：**EasyHeC 架构**。该架构由三个主要组件组成，并以迭代方式执行。首先，将 RGB图像输入掩码预测网络（PointRend），以获取观测到的机械臂掩膜。其次，使用基于可微渲染（DR）的相机位姿优化模块，通过最小化渲染掩膜与观测掩膜之间的差异来生成相机位姿。接下来，机械臂移动至由基于一致性的关节空间探索模块生成的下一个机械臂关节姿态。随后，我们重复上述步骤 N 次。

最高信息增益的下一视点的一种手段。我们在工作中采用了类似的方法，通过选择下一个最佳关节位姿来提高标定精度。然而，他们的方法是针对使用标记的眼在手上配置设计的，而我们的方法则是针对无需任何标记的眼在手外配置设计的。此外，我们的方法在掩膜空间中计算相机位姿不确定性，这一点在现有工作中尚未得到探索。

III. 方法

A. 问题表述

给定具有已知三维模型的机械臂和内参已知的相机，目标是求解眼在手外标定问题，即估计相机 c 与机器人基座 b之间的变换。在本文中，该变换表示为 $T_{cb} \in SE(3)$，并满足：$P^c = T_{cb}P^b$，其中 P^c 和 P^b 分别是相机坐标系和机器人基坐标系中的三维点。

B. 概述

如图 2 所示，我们以迭代方式优化相机位姿。在每次迭代中，给定特定关节位姿下的机械臂图像，我们首先使用 PointRend [20]预测观测掩码。随后，我们利用可微渲染（第 III-C 节），通过最小化渲染掩码与观测掩码之间的差异来优化相机位姿。接下来，为了更高效地优化 T_{cb}，我们利用基于一致性的关节空间探索模块（第 III-D 节）选择下一次迭代中信息量最大的关节位姿。然后，机械臂被驱动至该信息量最大的关节位姿，所提出的方法进入下一次迭代。需要注意的是，在后续迭代中，相机位姿会利用所有已观测关节位姿对应的掩码进行优化。

C. 基于可微渲染的相机位姿估计

不同于在基于标记的手眼标定中广泛使用的方程 $AX = XB$ [3]，我们的方法采用了一种新颖的像素级目标函数。假设已知机械臂的三维模型，我们可以通过最小化渲染出的机械臂与相机观测之间的差异来获得相机位姿，这与 [21] 中使用的渲染与比较技术类似。与基于标记的手眼标定方法和基于回归的方法不同，逐像素目标函数在相机位姿与观测图像之间建立了直接联系。这一特性使我们的方法更易于诊断。

具体而言，我们使用目标函数来最小化渲染掩膜与观测掩膜之间的差异：

$$L(T_{cb}) = (\min(1, \sum_l \pi(T_{cb}T_{bl}l)) - M)^2, \qquad (1)$$

其中，π 是可微掩码渲染器，l 表示机器人连杆的三维形状，M 是机器人的掩膜，T_{cb} 是机器人基座 b 与相机 c之间的变换，T_{bl} 是连杆 l 与机器人基座 b之间的变换，可通过关节位姿的正向运动学计算得到。由于我们为每个独立连杆生成掩膜，并将其组合以形成整个机械臂的掩膜，因此使用最小值函数将所有连杆掩膜之和限制为最大值 1。

由于优化目标 T_{cb} 属于$SE(3)$空间，我们将变换矩阵转换到李代数空间 [22]，然后使用以下目标函数进行优化：

$$L(\xi_{cb}) = (\min(1, \sum_l \pi(\exp(\xi_{cb})T_{bl}l)) - M)^2, \qquad (2)$$

上述损失函数完全可微，因此我们可以使用梯度下降法，借助 PyTorch [23] Autograd 执行优化。

掩码预测。为了获取机械臂掩码 M，作为基于可微渲染的优化过程的监督信号，我们利用现成的分割网络 PointRend [20]，从 RGB图像中预测掩膜。我们首先使用

由 SAPIEN [24] 机器人模拟器生成的合成数据训练 PointRend 模型，然后在真实数据上对其进行微调，以增强其泛化能力。

相机位姿初始化。在我们的工作中，我们采用 PVNet [25] 来进行位姿初始化。给定机械臂在初始关节位姿下的 RGB图像，PVNet 通过逐像素投票估计预定义的二维关键点位置，随后使用透视-点（PnP） [13]求解变换。

D. 基于一致性的关节空间探索

在本小节中，我们提出一种新颖的方法：交替进行两项操作，一是找出对相机位姿估计最具信息量的补充关节位姿，二是纳入这一新的关节位姿并重新计算相机位姿。

请回顾一下，我们通过求解式（2）实现相机位姿估计，其目标是使机械臂投影三维模型的掩膜与观测到的机械臂掩膜对齐。我们可以用更简洁的形式重写目标函数:

$$L(t; J) = \|\pi(J, t) - M_J\|^2, \tag{3}$$

其中， π 是掩膜渲染器， J 是关节位姿， t 是相机位姿， M_J 是关节位姿 J 下观测到的机械臂掩膜。在关节位姿观测有限的情况下，这种二维匹配约束通常不足以约束六维相机位姿的解空间。例如，仅凭一次关节位姿观测，很难准确估计相机与机器人基座之间的距离。换句话说，可能存在一组合理的相机位姿候选 T，满足 $L(t, J) < \epsilon$，其中 ϵ 是一个较小的值。

目标是通过获得一个新的 $T' \subseteq T$，使其成为 $L(t; J') < \epsilon$ 在新的关节位姿 J' 下的解集，从而尽可能缩小 T。接下来的问题是，如何选择 J'，才能尽快缩小 T？ 等价地，如何选择 J'，才能让 $L(t, J') > \epsilon$ 尽可能多的 $t \in T$?

我们的方法是对 $\{t_i\}_{i=1}^K \in T$采样，基于此，我们使用 $\sum_i L(t_i, J')$ 作为替代指标，帮助我们选择 J'。事实上，可以证明:

$$\sum_i L(t_i, J') \geq \mathrm{Var}(m_1, m_2, \ldots, m_K), \tag{4}$$

其中 $m_i = \pi(J', t_i)$。该不等式表明，如果选择 J' 以最大化掩膜方差，就能尽快缩小 T。

基于这一观察，我们提出了一种联合空间探索策略，该策略完全在仿真中运行，使我们能够在向机器人执行运动指令之前生成具有信息量的关节位姿。算法总结于算法1中：我们首先在仿真器中采样大量关节位姿，然后评估渲染掩膜在相机位姿候选上的方差。接着，我们将机器人移动到方差最大的关节位姿，并执行下一轮基于DR的优化。不同于需要人工设计关节位姿的传统基于标记的方法，

算法 1： 关节空间探索。

输入： 相机位姿候选 $T = \{t_i\}_{i=1}^K$;
关节位姿采样数量 N_s;
输出： 下一个关节位姿 J'

随机采样 N_s 个关节位姿 J_1， J_2，.. . ， J_{N_s}。
对于每个关节位姿 J_i **执行**
  对于每个相机位姿 t_k **执行**
    // 使用该关节位姿渲染机器人掩码
      J_i 和相机位姿 t_k。
    $m_{ij} = \pi$（J_i， t_k）
  V_i =方差（m_{i1}、 m_{i2}、...、 m_{iK}）
$J' = J_{argmax(V)}$

我们的方法完全自动化，无需人工干预。

在后续的基于 DR 的优化迭代中，我们纳入所有先前观测到的掩码，而不是仅使用最近一次观测。最终的优化目标定义如下:

$$L(\xi_{cb}) = \frac{1}{K}\sum_{i=1}^{K}(\min(1, \sum_l \pi(\exp(\xi_{cb})T_{bl,i}l)) - M_i)^2, \tag{5}$$

其中， $K = \{1, 2, \ldots, N\}$ 是图像的数量， $T_{bl,\ i}$ 是第 i个关节位姿下连杆 l 与机器人基座 b 之间的变换， M_i 是第i张图像中机器人的掩膜。

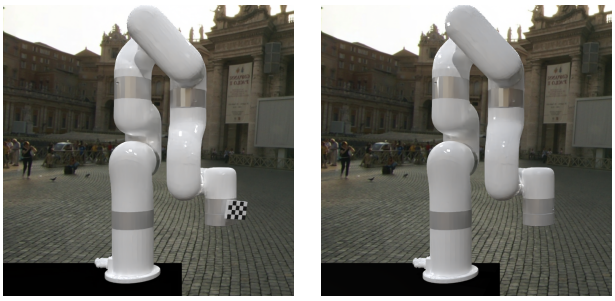
E. 实现细节

在我们的实现中，我们采用 nvdiffrast [26] 作为可微掩码渲染器。在优化过程中，我们使用学习率为 3e-3 的 Adam优化器，并执行 1000 次迭代。对于空间探索模块，我们从第 200 步到第 1000 步的范围内随机选择 50 个相机位姿，作为相机位姿候选。实际应用中，采样的相机位姿候选数量及范围可能会根据优化迭代次数而变化。接下来，我们随机采样 2000 个关节位姿，并剔除所有无效位姿，包括导致机械臂自相交、连杆与机械臂基座之间的距离超过预定义距离阈值，或与周围环境发生碰撞的位姿。随后，我们选取方差最大的关节位姿，作为下一步要探索的关节位姿。停止准则可根据所需的精度水平设定。通常，经过足够次数的迭代后，标定误差会收敛。在无法获得标定误差的真实场景中，可以将两次空间探索迭代之间的方差变化作为替代指标，用于确定停止准则。

IV. 实验

A. 合成数据集上的评估

数据生成。在我们的第一个实验中，我们使用 SAPIEN [24] 合成了 100 个照片级真实感场景，



(a) (b)

图3：**xArm合成数据集中的示例图像**。（a）用于基于标记的方法，（b）用于我们的方法。

这些场景采用不同的相机位姿，并使用 xArm机械臂。我们将我们的方法与传统的基于标记的方法进行比较。对于基于标记的方法，我们为每个场景手动设计了 20 个关节位姿。

我们的方法只需要初始关节位姿对应的图像，后续视角则在线生成。合成数据集示例如图3所示，其中我们在机械臂上安装了棋盘格标记，用于基于标记的方法。标记的尺寸为4cm×5cm，每个网格为1cm×1cm。我们渲染了10,000张图像用于PointRend [20] 训练，并另外渲染了10,000张图像用于PVNet [25] 训练，以获取我们的方法所需的观测分割掩码和初始相机位姿。

手眼标定评估。我们根据真实相机姿态与求解得到的相机位姿之间的平移误差和旋转误差进行评估。请注意，手动设计的关节位姿并不总能确保标记在所有场景中都可见，尤其是在存在遮挡的情况下。为确保公平，我们还考虑针对每个场景随机采样大量关节位姿，并仅选择标记可见的位姿。然而，我们发现，采样的关节姿态中只有1.9%能使标记清晰可见，因此这种方法并不实用。因此，我们对所有场景使用相同的关节位姿，但只分析检测到超过6个标记的场景。

结果如表I和表II所示，我们的方法以显著优势优于基于标记的方法 [3]。具体而言，基于标记的方法的平均旋转误差和平移误差分别为 0.9 度和 2.1 厘米。对于所提出的方法，我们采用 1 至 5 个不同视角进行评估。随着视角数量增加，误差持续降低。具体而言，仅使用 1 个视角时，我们的方法误差为 0.32 度和 0.5 厘米；使用 5 个视角时，误差则降至 0.08 度和 0.2 厘米。

DREAM [12]使用网络检测机器人机械臂上的关键点，并使用透视-点（PnP）算法估计相机位姿。由于这些关键点缺乏区分度，网络往往无法准确检测关键点。此外，关键点检测对遮挡和截断不够鲁棒，因此该方法的结果精度较低，尤其是

方法	1	2	3	4	5
基于标记的	0.870	-	-	-	-
DREAM [12]	1.924	1.240	0.981	0.764	0.704
我们的方法（Rand）	0.318	0.214	0.151	0.112	0.098
我们的方法（SE）	**0.322**	**0.128**	**0.109**	**0.097**	**0.081**

表 I：**xArm合成数据集上的旋转误差评估结果**。对于我们的方法，报告视图数量为 1 至 5 时的误差，单位为度。我们的方法（SE）和我们的方法（Rand）分别表示采用基于一致性的空间探索和随机采样进行后续关节配置生成的我们的方法。请注意，基于标记的方法使用 20 个视角计算手眼标定结果。

方法	1	2	3	4	5
基于标记的方法	2.000	-	-	-	-
DREAM	0.529	0.473	0.374	0.347	0.303
我们的方法（Rand）	0.496	0.388	0.361	0.322	0.312
我们的方法（SE）	**0.488**	**0.298**	**0.252**	**0.206**	**0.206**

表 II：**xArm合成数据集上的平移误差评估**。我们报告以厘米为单位的误差，视图数量为 1 至 5。

在旋转误差方面。即使通过采用多个视角提升了DREAM方法的性能，其表现仍不如我们的方法。相比之下，我们的方法无需使用任何关键点，并且在存在遮挡和截断的情况下依然表现出鲁棒性。

B. 在真实世界 *Baxter* 数据集上的评估

数据集介绍。在本节中，我们在真实世界的 Baxter 数据集上进行实验 [16] ，并将我们的方法与当前最先进的基于学习的方法进行比较。Baxter 数据集包含真实 Baxter 机器人的 100 张图像、单个相机位姿和 20 个不同的关节位姿。鉴于只有一个相机位姿，为简便起见，我们对相机位姿进行手动初始化。

实现细节和评估指标。 为了训练分割网络，我们首先在合成 DREAM 数据集 [12] 上使用域随机化训练 PointRend，然后在使用 SAPIEN [24]渲染的一些真实图像上进行微调。实验结果见表 III 和表 IV。我们报告二维和三维的百分比。

方法	20px	30px	40px	50px	100px	150px	200px
DREAM [12]	0.16	0.23	0.29	0.33	0.52	0.62	0.64
确定 [16]	0.34	0.54	0.66	0.69	0.88	0.93	0.95
IPE [17] (盒子)	-	-	-	0.65	0.94	0.95	0.95
IPE [17] (圆柱体)	-	-	-	0.80	0.91	0.93	0.95
IPE [17] (计算机辅助设计)	-	-	-	0.74	0.90	0.94	**1.00**
我们的方法（1个视角）	0.35	0.55	0.75	0.90	0.95	0.95	**1.00**
我们的方法（2个视角）	0.40	0.75	0.95	**1.00**	**1.00**	**1.00**	**1.00**
我们的方法（3个视角）	**0.55**	**0.85**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

表 III：**Baxter 数据集上的二维 PCK 评估结果**。我们提供了不同阈值下的二维 PCK 分数，阈值范围为 20 至 200 像素。



(a) (b)

图4：**Baxter 数据集上的定性结果**。我们使用求解得到的相机位姿渲染 Baxter 机械臂的连杆，并将掩膜叠加到 RGB图像上。

方法	2cm	5cm	10cm	20cm	30cm	40cm
DREAM [12]	0.01	0.08	0.32	0.43	0.54	0.66
OK [16]	0.10	0.34	0.54	0.66	0.69	0.88
IPE [17](方框)	-	-	0.8	0.95	0.95	0.95
IPE [17] (圆柱体)	-	-	0.71	0.93	0.94	0.95
IPE [17](计算机辅助设计)	-	-	0.78	0.93	0.97	**1.00**
我们的方法（1个视角）	0.10	0.65	0.90	**1.00**	**1.00**	**1.00**
我们的方法（2个视角）	**0.15**	**0.80**	**0.95**	**1.00**	**1.00**	**1.00**
我们的方法（3个视角）	**0.15**	**0.80**	0.90	**1.00**	**1.00**	**1.00**

表 IV：**Baxter 数据集上的三维关键点正确率评估结果**。我们提供了不同阈值下的三维关键点正确率得分，范围为 2 至 40 厘米。

正确关键点（PCK），即距离真实值在特定阈值范围内的二维/三维关键点所占的百分比。我们还在图 4 中提供了一些结果的可视化。除了[16, 17],中采用的阈值外，我们还报告了使用更小阈值时的结果，以进一步阐明我们方法的准确性，因为在以往工作所使用的阈值下，我们的方法几乎达到了 100%。遵循 [16, 17], 中的惯例，我们在表 III 和表 IV 中展示了训练图像上的评估结果；而在图5中，我们展示了在整个数据集上评估的结果，因为在其他关节位姿下评估性能更具代表性。

手眼标定评估。 对于我们的方法，我们使用 1 至 3 个视角进行评估。结果表明，我们的方法大幅优于以往最先进的方法，证明了我们方法的有效性和鲁棒性。

C. 在真实世界 *xArm* 数据上的评估

为评估我们提出的手眼标定方法在真实场景中的有效性，我们使用 ArUco 标记板和安装在机器人末端执行器上的指针进行了高精度目标定位实验。

我们使用 OpenCV [27]检测标记板的角点，并通过手眼标定结果将其位置转换到机械臂基坐标系中。接着，我们用指针触碰标记板角点，并手动测量指针与实际角点位置之间的距离。图 1 展示了一个代表性示意图。

方法	DREAM	Ours	我们的方法（SAM）
误差（厘米）	1.5	0.4	0.3

TABLE V：**真实世界机器人目标定位实验的评估结果**。我们的方法使用 PointRend 进行分割，而我们的方法（SAM）使用 Segment Anything 模型，并以手动标注的边界框作为提示。

结果及与 DREAM 的比较见表 V。由于 DREAM 估计二维关键点，并使用 PnP 算法估计相机位姿，因此无法稳健应对遮挡和截断，且会受到仿真到现实的域差异影响。相比之下，我们的方法使用稠密像素级误差作为目标函数，因此能够天然地稳健应对截断和自遮挡。尽管人工测量的可靠性不高，但它初步反映了系统的准确性，并有力地证明了我们方法的有效性。为进一步提高准确性，我们可以使用更精确的分割网络，例如 SAM [28] ，以达到表 V 所示的更高准确性。实际应用中，这两种分割方法的选择取决于具体需求。PointRend 需要预先进行预训练，但可实现全自动标定过程；而 SAM 只需进行简单的掩膜标注，但需要在过程中进行人工干预。

D. 消融研究

关节空间探索。 为展示关节空间探索模块的有效性，我们通过与随机选择方法进行比较，开展了一项消融研究。实验在 xArm 合成数据集和 Baxter 数据集上进行。对于 xArm 合成数据集，关节姿态选择来自采样得到的关节位姿；对于 Baxter 数据集，关节姿态选择则来自数据集提供的关节位姿。结果见表I、表II和图5。xArm 合成数据集上的结果表明，使用空间探索来选择下一个关节位姿，始终能够加快收敛速度，并且误差小于随机选择。Baxter 数据集上的结果表明，基于空间探索的方法比随机选择方法取得了更高的 PCK，尤其是在 PCK2D 方面。

采样的关节位姿数量不同。 为评估关节空间探索过程中关节位姿数量变化对我们方法性能的影响，我们在 xArm合成数据集上通过采样不同数量的关节位姿进行了实验。实验结果如图6所示，结果表明，随着采样的关节姿态数量增加，误差会随之减小，尤其是在关节位姿数量低于1000时。当关节

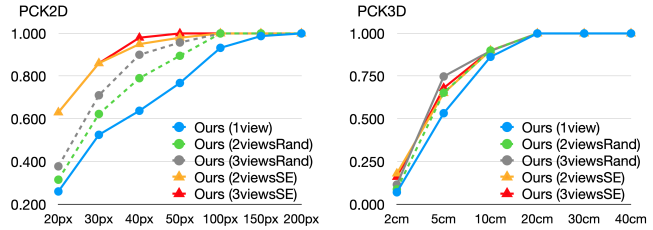


图5：**Baxter 数据集上的消融研究**。我们报告了在不同阈值下，Baxter 数据集上的 PCK2D（左）和 PCK3D（右）。我们的方法（SE）和我们的方法（Rand）分别表示采用基于一致性的空间探索和随机采样来生成下一个关节位姿的我们的方法。

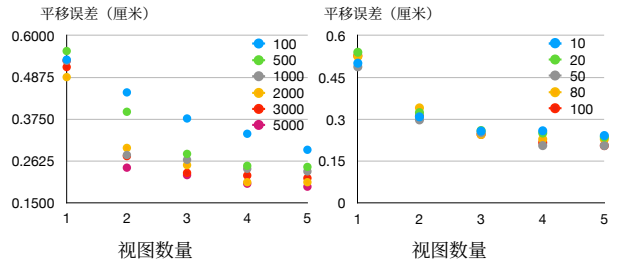


图6：**在关节空间探索过程中，使用不同数量的采样关节位姿（左）和相机位姿候选（右）时，xArm合成数据集上的位姿估计结果**。我们报告了以厘米为单位的平移误差。

位姿数量超过1000后，误差几乎不再变化。因此，在其他实验中，我们将采样的关节位姿数量设为2000年。

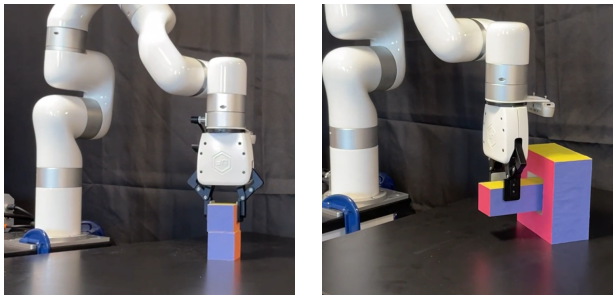
不同数量的相机候选采样。为验证关节空间探索过程中采样的相机候选数量对我们方法性能的影响，我们在xArm合成数据集上进行消融研究，改变采样的相机候选数量。图6中的结果表明，随着采样的相机候选数量变化，误差大致保持不变。这说明采样的相机候选数量不会影响我们方法的性能，因此在实践中可以使用较少的相机位姿候选，以节省计算时间。在其他实验中，我们将采样的相机候选数量设为 50。

E. 运行时间

通常，系统的总运行时间由三个主要部分组成：基于DR的优化、空间探索，以及在真实世界应用中机器人在不同关节位姿之间的移动。实际上，运行时间取决于优化步骤的数量、空间探索迭代次数、相机位姿候选的数量以及采样的关节姿态数量。在真实实验中，运行时间还取决于诸如以下因素：

总体	基于可微渲染（DR）的优化	空间探索	机器人运动
250-370秒	40-50秒	150-200秒	60-120秒

表 VI：**在 RTX 4090 GPU 上测试的 EasyHeC 运行时间**。我们对基于DR的优化使用 1000 个优化步骤，对空间探索使用 5 次迭代；每次迭代包含 2000 个采样的关节姿态和 50 个相机位姿候选项。



(a) (b)

图 7：**由我们的方法支持的现实应用：**堆叠方块 (a) 和插销插入 (b) 实验的截图。完整过程请参阅随附的补充视频。

机器人的速度、关节位姿之间的距离以及规划轨迹。我们方法的运行时间见表 VI。与耗时的基于标记的标定相比，我们的方法运行时间可以接受；后者可能需要数小时进行微调。需要注意的是，标定完成后，无需重复标定，除非更换相机或机器人。

F. 真实世界应用

EasyHeC 可用于各种现实应用 [29, 30]。本节将介绍一些现实应用，以展示我们所提出方法的有效性。假设我们打算采用仅使用视觉输入的基于状态的策略，那么就必须将物体状态从相机坐标系转换到机器人基座坐标系。在这种情况下，可以利用所提出的 EasyHeC 执行这一变换。在我们的实验中，我们选择 CoTPC [29] 作为基于状态的算法，该算法以状态作为输入，并输出下一步机器人动作。我们使用 CoTPC 算法进行了两项实验：堆叠方块和插销插入。具体而言，我们使用 PVNet [25] 估计相机坐标系中的物体位姿，然后利用手眼标定结果，将物体位姿从相机坐标系变换到基座坐标系，并将其作为输入状态提供给 CoTPC 网络。图 7 展示了执行过程中的截图，结果表明方块实现了精确对齐，插销也成功插入，从而验证了我们的方法在基于状态的算法中的有效性。

V. 局限性

我们的工作存在一些局限性。**(1)** 我们的方法需要精确的机器人模型，但由于某些机器人带有保护外壳，该方法可能不具备普适性。不过，在多视角标定场景下，机器人连杆的物体形状可以得到有效且无歧义的优化。或者，也可以采用诸如 [31]之类的先进技术，在使用我们的方法之前重建高保真度的连杆形状。**(2)** 我们的方法仅适用于眼到手标定场景。然而，对于眼在手上的配置，同样的流程也可用于标定。尽管该场景下相机的视野受限，使用 PVNet [25] 进行初始化时可能会引入误差，但眼在手上的配置会大幅限制初始相机位姿的分布范围，因此可以进行手动初始化，或从若干预定义的相机位姿中选择一个进行初始化。**(3)** 对于新的机械臂，需要重新训练 PointRend。不过，这只需进行一次训练。训练完成后，该模型可用于所有同类型的机械臂。此外，我们还可以使用现成的 Segment Anything 模型进行分割，只需对若干手动标注的边界框提示进行标注。

VI. 结论

本文提出了一种新颖的手眼标定方法。我们的方法利用可微渲染技术，通过最小化渲染掩码与观测掩码之间的差异来优化相机位姿。我们采用关节空间探索策略，寻找有助于相机位姿优化的信息丰富的关节位姿，从而减少获得精确结果所需的样本数量。实验结果表明，在合成数据集和真实世界数据集上，我们的方法均优于传统方法以及几种最先进的方法。总体而言，我们的方法为机器人学中的相机位姿估计提供了一种实用解决方案，并可应用于广泛的场景，例如物体操作、抓取以及抓取与放置任务。我们相信，我们的工作能够缩小仿真到现实的差距，并促进更高效、更准确的机器人系统的发展。

参考文献

[1] S. Garrido-Jurado、R. Mu˜noz-Salinas、F. J. Madrid-Cuevas 和 M. J. Mar´ın-Jim´enez，“遮挡条件下高可靠性基准标记的自动生成与检测”，模式识别，第 47 卷，第 6 期，第 2280–2292 页，2014 年。[2] J. Ilonen 和 V. Kyrki，“鲁棒的机器人-相机标定”，载于*2011*年第*15*届国际先进机器人会议（*ICAR*）。IEEE，2011 年，第 67–74 页。[3] R. Y. Tsai、R. K. Lenz 等，“一种用于全自主高效 3D 机器人手眼标定的新技术”，*IEEE*机器人与自动化汇刊，第 5 卷，第 3 期，第 345–358 页，1989 年。[4] Y. Labb´e、J. Carpentier、M. Aubry 和 J. Sivic，“通过渲染与比较进行单视图机器人位姿和关节角度估计”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021 年，第 1654–1663 页。

[5] B. C. Sefercik 和 B. Akgun，“学习无标记机器人深度相机标定与末端执行器位姿估计”，载于机器人学习会议。PMLR，2023年，第 1586–1595 页。[6] O. Bahadir、J. P. Siebert 和 G. Aragon-Camarasa，“一种使用机器人操作臂上单个参考点的深度学习手眼标定方法”，载于2022年IEEE机器人与仿生学国际会议（ROBIO）。IEEE，2022年，第 1109–1114 页。[7] F. C. Park 和 B. J. Martin，“机器人传感器标定：在欧几里得群上求解 ax=xb”，《IEEE机器人与自动化汇刊》，第 10 卷，第 5 期，第 717–721 页，1994年。[8] K. Daniilidis，“使用双四元数进行手眼标定”，《国际机器人研究杂志》，第 18 卷，第 3 期，第 286–298 页，1999年。[9] N. Andreff、R. Horaud 和 B. Espiau，“使用运动恢复结构进行机器人手眼标定”，《国际机器人研究杂志》，第 20 卷，第 3 期，第 228–248 页，2001年。[10] J. Heller、M. Havlena、A. Sugimoto 和 T. Pajdla，“基于运动恢复结构并使用 l_∞ 最小化的手眼标定”，载于CVPR 2011。IEEE，2011年，第 3497–3503 页。[11] X. Zhi 和 S. Schwertfeger，“手眼标定与重建同步进行”，载于2017年IEEE/RSJ智能机器人与系统国际会议（IROS）。IEEE，2017年，第 1470–1477 页。[12] T. E. Lee、J. Tremblay、T. To、J. Cheng、T. Mosier、O. Kroemer、D. Fox 和 S. Birchfield，“基于单幅图像的相机到机器人位姿估计”，载于2020年IEEE机器人与自动化国际会议（ICRA）。IEEE，2020年，第 9426–9432 页。[13] V. Lepetit、F. Moreno-Noguer 和 P. Fua，“EPnP：一种针对透视-点（PnP）问题的精确 O(n) 解法”，《国际计算机视觉杂志》，第 81 卷，第 2 期，第 155–166 页，2009年。[14] E. Valassakis、K. Dreczkowski 和 E. Johns，“从单幅图像学习眼在手上相机标定”，载于机器人学习会议。PMLR，2022年，第 1336–1346 页。[15] P. J. Besl 和 N. D. McKay，“三维形状配准方法”，载于《传感器融合IV：控制范式与数据结构》，第 1611 卷。SPIE，1992年，第 586–606 页。[16] J. Lu、F. Richter 和 M. C. Yip，“通过关键点优化和仿真到现实迁移进行机器人操作臂位姿估计”，《IEEE机器人与自动化快报》，第 7 卷，第 2 期，第 4622–4629 页，2022年。[17] J. Lu、F. Liu、C. Girerd 和 M. C. Yip，“通过可微渲染对机器人操作臂及柔性连续体机器人进行基于图像的位姿估计与形状重建”，IEEE国际机器人与自动化会议（ICRA），2023年。[18] X. Zhang、Y. Xi、Z. Huang、L. Zheng、H. Huang、Y. Xiong 和 K. Xu，“通过在线精度驱动的下一最佳视点选择进行主动手眼标定”，《视觉计算机》，第 39 卷，第 1 期，第 381–391 页，2023年。[19] J. Yang、J. Rebello 和 S. L. Waslander，“用于机器人眼在手上标定的下一最佳视点选择”，arXiv预印本 arXiv:2303 06766，2023年。.[20] A. Kirillov、Y. Wu、K. He 和 R. Girshick，“PointRend：将图像分割视为渲染”，载于IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第 9799–9808 页。[21] A. Kundu 、 Y. Li 和 J. M. Rehg，“3D-RCNN：通过渲染与比较进行实例级三维物体重建”，载于IEEE计算机视觉与模式识别会议论文集，2018年，第 3559–3568 页。[22] D. Gabay，“在微分流形上最小化可微函数”，《优化理论与应用杂志》，第 37 卷，第 177–219 页，1982年。[23] A. Paszke、S. Gross、S. Chintala、G. Chanan、E. Yang、Z. DeVito、Z. Lin、A. Desmaison、L. Antiga 和 A. Lerer，“PyTorch中的自动微分”，2017年。[24] F. Xiang、Y. Qin、K. Mo、Y. Xia、H. Zhu、F. Liu、M. Liu、H. Jiang、Y. Yuan、H. Wang、L. Yi、A. X. Chang、L. J. Guibas 和 H. Su，“SAPIEN：基于仿真的部件交互环境”，载于IEEE计算机视觉与模式识别会议（CVPR），2020年6月。[25] S. Peng、Y. Liu、Q. Huang、X. Zhou 和 H. Bao，“PVNet：用于6自由度位姿估计的逐像素投票网络”，载于IEEE/CVF计算机视觉与模式识别会议论文集，2019年，第 4561–4570 页。[26] S. Laine、J. Hellsten、T. Karras、Y. Seol、J. Lehtinen 和 T. Aila，“用于高性能可微渲染的模块化基元”，《ACM图形学汇刊》（TOG），第 39 卷，第 6 期，第 1–14 页，2020年。[27] G. Bradski，“OpenCV库”，《Dr. Dobb’s期刊：专业程序员的软件工具》，第 25 卷，第 11 期，第 120–123 页，2000年。

[28] A. Kirillov、E. Mintun、N. Ravi、H. Mao、C. Rolland、L. Gustafson、T. Xiao、S. Whitehead、A. C. Berg、W.-Y. Lo、P. Doll´ ar 和 R. Gir- shick，“万物分割”，*arXiv:2304.02643*，2023年。[29] Z. Jia、F. Liu、V. Thumuluri、L. Chen、Z. Huang 和 H. Su，“结合行为克隆的思维链预测控制”，载于国际学习表征会议 *2023*年重生强化学习研讨会。[30] L. Chen、Y. Song、H. Bao 和 X. Zhou，“通过戳碰物体感知未见过的三维物体”，载于*2023*年 *IEEE*国际机器人与自动化会议（*ICRA*），2023年，第4834–4841页。[31] K. Yang、X. Zhang、Z. Huang、X. Chen、Z. Xu 和 H. Su，“Mov- ingparts：动态辐射场中的基于运动的三维部件发现”，*IEEE*计算机视觉与模式识别会议论文集（*CVPR*），2023年。