arXiv:2301.01441v1 [cs.CV] 4 Jan 2023

利用机器人手持观测与合成自动准备 YOLO 训练数据

Hao Chen[1], Weiwei Wan[1*]、Masaki Matsushita[2],Takeyuki Kotaka[2] 和 Kensuke Harada[13]

摘要——近年来，深度学习方法在目标检测方面展现出令人瞩目的性能。然而，要达到较高的识别准确率，这类方法需要大量训练数据，而收集数据既耗时，又需要进行图像标注等大量人工工作。本文利用机器人自动准备训练数据。考虑到机器人运动效率低、能耗高，我们提出结合机器人手持观测与数据合成，以扩充机器人采集的有限数据集。我们首先使用配备深度传感器的机器人采集物体被机器人手持时的图像，并分割出物体图像。然后，采用复制粘贴方法，将分割出的物体与试管架背景合成。我们将采集图像与合成图像组合起来，用于训练深度检测神经网络。我们进行了实验，将使用所提出的方法采集的图像训练出的 YOLOv5x 检测器与使用其他几种方法训练的检测器进行了比较。结果表明，组合观测图像与合成图像的性能与人工数据准备相当。这些结果为优化数据配置和训练检测器的参数设置提供了良好参考。所提出的方法只需单次处理，便能以低成本生成组合数据。感兴趣的读者可以在以下 GitHub 仓库中找到数据集和训练好的模型：
github.com/wrslab/tubedet

致实践者——本研究源于实验室自动化中的一项需求：利用机器人自动整理随机放置的试管。在将试管送入检测设备进行梯度测试之前，人工需要将试管分类并整理成特定排列，以适配设备的内部设计。由于试管排列要求会随时间变化，依靠人工操作并不容易。较理想的解决方案是用机器人替代人工。机器人应配备视觉系统以检测试管，并配备操作系统来执行实际整理操作。机器人会在繁忙时段投入使用，闲暇时则用于执行其他任务。YOLO 等深度神经网络在试管检测任务中表现出色。然而，训练数据的准备过程颇具挑战，也不适合实验室终端用户。预训练神经网络虽可作为一种选择，但其试管检测能力有限，且无法处理新加入的试管类型。本文提出的方法有助于解决训练数据准备问题。在该方法的支持下，机器人能够以单流程、低成本的方式自动准备训练数据，且数据质量可与人工标注的数据相媲美。

索引词—**机器人数据准备、数据合成、试管检测**

I. 引言

近年来，深度学习的进展推动了目标检测领域的一场变革。基于深度学习的方法利用深度神经网络从训练数据中学习特征。它们的表现优于传统的手工设计特征，取得了令人瞩目的成果。尽管具有这些优势，基于深度学习的目标检测仍需收集大量标注数据用于训练，这既耗时又费力，并且严重阻碍了基于深度学习的应用在可扩展性和灵活性方面的发展。

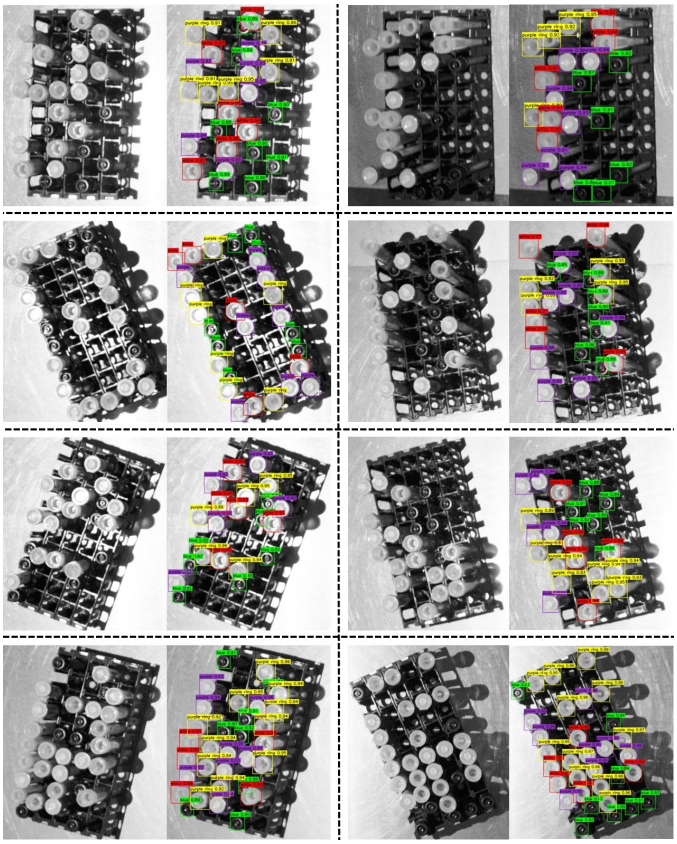


图 1：货架内试管检测的若干示例。每个网格包含两张图像。左侧图像由视觉传感器采集。右侧图像为识别结果。用于训练检测神经网络的数据采用所提出的方法进行准备。

此前，研究人员已开发出多种降低数据收集成本的方法。例如，数据增强[1] 通过应用图像旋转或缩放等随机变换，丰富现有的训练数据集。数据合成[2][3]则通过仿真或对抗神经网络生成此前未见过的数据。增强或合成方法的主要挑战是“域差距” [4][5]:

[1]大阪大学工学研究科系统创新系，日本大阪丰中市。[2]H.U. Group Research Inst. G. K.，日本。[3]National Inst. of AIST，日本。

*联系人：Weiwei Wan，wan@sys.es.osaka-u.ac.jp

增强数据的视觉上下文变化较少。合成数据则容易与真实世界存在差异。近期，研究人员重新关注使用复制粘贴方法 [6]来增加数据。该方法有效弥补了“域差距”问题，展现出令人瞩目的性能。使用复制粘贴方法生成数据时，增强与合成之间并没有明确的界限。它主要被归类为合成方法 [7][8], ，尽管有些研究认为它属于增强 [9]。为避免与变换和基于尺度的数据生成混淆，本文将其称为合成方法。

复制粘贴方法中最费力的部分，是如何将各种各样的目标区域整齐地裁剪下来，并粘贴到新的背景上。此前，从事机器人操作研究的学者已经开发出机器人方法，可将新物体从背景中分割出来。例如，Florence 等人 [10], Boerdijk 等人 [11]，和 Pathak 等人 [12]分别利用机器人手内操作或非抓取操作改变物体的观测视角，并依据机器人运动对物体进行分割。这类系统可以在各种条件下代替人工分割目标物体区域。最近的研究 [11][13] 注意到了这一优势，并通过将机器人系统分割出的物体粘贴到随机背景上，增加了数据量和上下文多样性。尽管这些开创性工作提出了相关方法，但复制粘贴合成的必要性，以及数据量和比例的影响，仍未得到讨论。

基于当前的研究现状，本文进一步探讨如何利用机器人自动收集训练数据。考虑到机器人数据收集效率低、能耗高，我们提出结合机器人观测与复制粘贴合成以降低成本。我们以图 1 所示的试管检测任务为例，使用配备深度传感器的机器人移动并观测试管。机器人收集观测图像，同时从图像中分割出试管，用于复制粘贴合成。观测图像和合成图像用作深度检测神经网络的训练数据。尤其是在合成流程中，我们重视试管与试管架的共现，并将试管粘贴到试管架区域内，以确保上下文一致性。此外，我们还考虑试管间遮挡，以及环境或视觉差异导致的前景变化等因素，以减少不真实的合成结果。所提出的方法有助于丰富数据集并解决“域差距”问题，而且不需要大量机器人操作。

在实验中，我们训练了多个 YOLOv5x 网络，以了解所提出方法的性能。训练数据采用所提出的方法和其他几种方法收集。结果证实，使用所提出的方法收集的数据确实具有所宣称的优势。我们还进行了多项消融研究，考察使用所提出的方法收集的数据训练检测神经网络时，数据量和比例所产生的影响。结果为优化数据配置和训练检测器的参数设置提供了良好指导。

本研究的贡献如下。(1) 我们开发了一种自动数据收集方法，让机器人抓持目标物体并对其进行观测。该方法可生成观测图像以及从图像中分割出的目标区域。(2) 我们开发了一种复制粘贴图像合成方法，以丰富训练数据。该方法将物体区域粘贴到各种试管架背景上，从而平衡“域随机化”和“域差距”。试管架背景也由机器人自动收集。(3) 我们研究了观测图像与合成图像的不同组合，并将其与其他数据集进行比较，以了解数据量和比例的影响。

本文其余部分安排如下：第II节回顾相关工作。第III节介绍硬件系统和所提出方法的工作流程。第IV节详述技术细节。第V 节展示实验与分析。第VI 节总结全文。

II. 相关工作

我们分别回顾机器人数据收集和数据合成方面的相关工作。

A. 使用机器人进行自动数据收集

从图片中分割出物体区域是自动数据收集的基础。传统方法使用简单背景 [14], 已知环境 [15],或设计易于识别的小器件 [16][17] 来简化物体提取。这些方法需要仔细准备场景和物体。

基于机器人的方法利用驱动机器人来简化物体分割。这类方法可以追溯到早期关于物体识别和三维物体建模的研究 [18][19][20][21]。这些研究利用机器人操作序列，从不同视点感知物体，并将物体从背景中分割出来。从机器人操作的角度来看，这类分割可分为两类：手持物体分割和基于交互的分割。

1）手持物体分割：先前的手持物体分割研究利用已知的机器人模型和手工设计的视觉特征，将手持物体与背景环境及机器人手部分离。例如，Krainin 等人[21] 通过检查点云与机器人模型之间的欧几里得距离，分离出手持物体的点云。Welke 等人[19] 基于特征背景减除、视差图和手部定位，从图像中分割手持物体。考虑到不同目标的视觉特征，这些方法需要人工为各种目标准备检测器。

近期研究利用深度学习，降低手持物体分割对人工设计视觉特征的依赖。例如，Florence 等人 [10] 提出了一种用于分割手持物体的自监督框架。该框架包含两个步骤，且两步采用相同的训练和学习流程。第一步中，作者结合深度和 RGB 信息为机器人生成掩膜，并根据这些掩膜训练神经网络模型，以区分机器人与背景。第二步中，作者对抓取物体进行掩膜处理，并训练神经网络模型，使其能够将物体与机器人手部分离。Boerdijk 等人 [13] 使用光流来

分别分割抓着物体和未抓着物体的机械臂。随后，利用分割后的数据集训练神经网络，以分离机械臂和抓取物体。

2）机器人-物体交互：另一方面，一些研究人员利用推搡等非抓取式机器人操作来改变物体视角，并根据机器人运动线索对物体进行分割 [12][22][23]。例如，Pathak 等人 [12] 设计了一种框架，用于持续改进通过机器人交互生成物体分割掩码的神经网络模型。该模型最初会为物体生成候选分割掩码，再根据机器人交互前后拍摄图像的像素差异对掩码进行改进。生成模型也会随改进后的掩码一同更新。Singh 等人 [24]提出了一种方法：在杂乱场景中分割未知物体，同时反复通过机器人轻推动作与物体交互，以引入几何约束。机器人交互式分割通常要求场景或表面保持静止，以便机器人与物体进行交互 [25][26]。与手持物体分割相比，这种方法更加复杂，因为需要通过机器人操作来控制和改变物体姿态。

机器人方法的一个关键问题在于不适合用于准备大量训练数据，因为机器人执行物理运动会耗费大量时间和能源。让机器人执行数千条运动轨迹来收集数据并不切实际。此外，系统中的机器人位置固定、视野有限，只能在狭窄的场景范围内收集数据。使用所收集数据训练的神经网络可能会受到上下文（背景）偏差的影响，且泛化能力较差 [27][28]。

本研究聚焦于机器人数据收集，同时考虑利用数据合成来减少机器人的使用。我们首先让机器人握住一支试管，并根据机器人的工具中心点（TCP）和手部模型，提取手持点云，为试管标注包围多面体。然后，我们将标注的包围多面体映射到机器人相机视野中的二维图像区域，以提取试管区域。机器人将试管移动到不同的位置并进行旋转，从而获得各种各样的二维图像和试管区域。这些图像和试管区域将分别用于训练和后续阶段的新数据合成。

B. 数据增强与合成

数据增强与合成是丰富训练数据最常用的两种方法。数据增强通过特定规则或基于学习的方法变换现有训练数据，从而生成新数据。数据合成则通过将现有数据与其他数据融合，或利用计算机仿真来生成合成数据。同期发表的论文往往混用这些术语。因此，下面我们将对相关文献进行统一综述，不加区分。

复制粘贴方法广泛用于生成合成数据。它会从现有图像中分割出前景物体（也可能对其进行修改），然后将其粘贴到新的背景上 [8][7][9]。复制粘贴方法易于实现，且与仅使用纯真实数据相比表现出显著优势。先前的研究表明，粘贴物体时仔细选择背景非常重要。例如，Divvala 等人 [29] 通过实验表明，视觉上下文有助于提升目标检测性能并减少检测错误。Dvornik 等人 [30] 表明，粘贴物体时使用恰当的视觉上下文可以提升预测性能，而不恰当的视觉上下文则会导致负面结果。Wang 等人 [31] 将不同图像中属于同一类别的物体互换，以确保物体与背景之间的上下文一致性，并表明使用现有背景比使用随机背景效果更好。此外，复制粘贴方法需要一个包含物体多种可能视角且易于抠出的数据集。由人工准备这些数据集十分费力。

图形仿真是生成训练数据的另一种常用方法。仿真的优点在于可以自由改变光照条件和材质，以增加多样性；还可以通过简单地变换虚拟相机的姿态，采集物体的多个视角。例如，Hodaˇn 等人 [32] 和 Richter 等人 [33] 分别利用照片级真实感渲染，对三维物体模型和场景进行图像合成。这些方法需要大量计算资源，才能缩小合成数据与真实数据之间的域差距。Tobin 等人 [34] 提出了域随机化（DR）的概念。他们对仿真器进行随机化，使模型接触范围广泛的环境，从而获得多样化的训练数据。与照片级真实感渲染不同，该方法只需低保真渲染结果，就能使中等尺寸物体达到令人满意的准确率。Carlson 等人 [35], 、Hinterstoisser 等人 [4], 、Prakash 等人 [36], 和 Tremblay 等人 [37] 分别利用域随机化来缩小域差距。作者在仿真中随机改变上下文，使得“真实数据就像另一种仿真” [38]。Yang 等人 [39] 和 Sundermeyer 等人 [40] 分别通过仿真对三维物体模型的视点进行采样，并将样本与真实背景混合，以减少人工准备具有丰富域随机性的场景所需的工作量。除域随机化外，生成对抗网络（GANs）也有望缩小域差距。例如，Chatterjee 等人 [41] 设计了一种轻量级生成对抗网络，用于合成训练塑料瓶检测器的数据。

在本研究中，我们利用数据合成来丰富训练数据。我们开发了一种基于复制粘贴的方法，将从机器人观测图像中分离出的试管帽区域贴到试管架背景上，从而合成新图像。合成过程中可以考虑试管架尺寸和试管遮挡等多种约束，以缩小域差距。合成数据与真实世界数据混合，以提升基于 YOLO 的试管识别神经网络的性能。我们还将其与其他数据收集方法进行比较，以了解数据量和数据组合比例的影响。

III. 机器人系统与工作流程

A. 机器人系统配置

图 2(a) 展示了我们用于准备训练数据的机器人系统。我们使用 Photoneo Phoxi M 3D扫描仪来

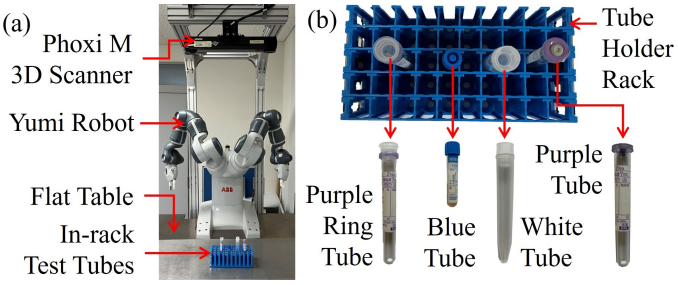


图 2：（a）系统配置。（b）Phoxi M3D 扫描仪视角下的试管和试管架。

采集平面桌面上的物体。系统采用配有两指夹爪的 ABB Yumi 双臂机器人来操纵物体。在 Yumi机器人前方设置了一张平面桌面。待识别的试管架内试管放置在桌面上。Phoxi 扫描仪是一种基于结构光的深度传感器，可同时采集灰度图像和点云。Phoxi 扫描仪采集的点云中，每个数据点都与灰度图像中的一个像素一一对应。我们可以利用物体的点云，在灰度图像中分割出该物体。

具体而言，我们将 Phoxi 扫描仪安装在机器人顶部，以俯视试管架和试管。识别试管架中的试管时，我们选用试管盖作为主要标识符。我们倾向于使用试管盖进行识别，原因有二。首先，受基于结构光的深度传感器的限制，难以获取半透明或透明试管的点云。其次，试管放在试管架中并从上方观察时，试管主体会被试管盖遮挡，也会被周围的试管遮挡，因此不太容易看见。不过，尽管使用试管盖有上述原因和优势，仍存在一个问题：不同类型的试管可能共用同一种试管盖。本文假设，盖子相同的试管可根据其在试管架中的高度加以区分，并通过分析点云来识别它们。

B. 数据准备工作流程

我们按照图中所示的工作流程，使用机器人系统准备训练数据。3图中有四个虚线框，其中 (a.1) 和 (a.2) 具有蓝色背景，表示数据采集组件；(b) 具有橙色背景，表示数据合成组件；(c) 具有灰色背景，表示生成的数据。

第一个蓝色虚线框（图3(a.1)）包含三个步骤。首先，人将一支未知的试管递给机器人。假设交接后机器人垂直抓住试管，管帽位于机器人指尖上方。其次，机器人将试管移动到离线预先设定的观测位姿，同时尽量避免自遮挡。Phoxi传感器会在每个观测位姿下采集试管的灰度图像和点云。第三，系统根据试管对应点云的映射，从采集的图像中分割出管帽区域。分割结果仅包含管帽。得益于点云映射，背景将被去除。该虚线框的输出包括许多管帽区域图像。这些图像从不同视角观测，因此具有不同的光照和视觉条件。

第二个蓝色虚线框（图3(a.2)）与第一个类似，也包含三个步骤。首先，有人将一个试管架放置在环境中。然后，机器人将试管架推至随机位姿，并在每个位姿处采集试管架的灰度图像和点云。第三，系统根据试管架对应点云的映射，从采集到的图像中分割出架子区域。此虚线框的结果包括许多架子区域图像。与帽子一样，由于数据是在不同观察位置采集的，架子区域图像也具有不同的光照和视觉条件。

橙色虚线框展示了数据合成过程：将第一个“数据收集”虚线框中获得的帽子区域图像复制粘贴到第二个“数据收集”虚线框中获得的架子区域图像上，以合成新图像。合成过程中会考虑试管架边界以及透视投影造成的重叠等约束。该虚线框的输出将是装有许多试管盖的试管架。“复制粘贴数据合成”子模块展示了若干输出示例。

最终的数据准备结果包括收集管帽数据时获得的图像（观测图像）和合成图像。这些结果展示在灰色虚线框中（图3（c））。

需要注意的是，上述工作流程并非完全自动化。文本以绿色突出显示的子模块涉及人工干预。此外，在数据收集之前，我们需要准备相机标定矩阵和试管观测位姿。相机标定矩阵将相机局部坐标系中采集的点云转换到机器人坐标系中。已有许多用于获取标定矩阵的方法 [42]。为避免重复，本文不讨论其细节。试管观测位姿是一组试管位置和旋转，用于让机器人持握试管并采集观测图像。所开发的方法将综合考虑机器人抓取和试管观测位姿，生成机器人关节配置。第IV节将详细介绍其生成算法。

IV. 实现细节

A. 采集管帽的观测位姿

采集管帽数据时，机器人会将手中抓着的试管移动到不同的观测位姿。生成观测位姿时会考虑两个约束：(1) 所采集管帽数据的多样性；(2) 机器人连杆遮挡。考虑这两个约束，可以从多个视点采集试管盖，从而覆盖多种光照和视觉条件。同时，这些约束还有助于避免机器人连杆遮挡被抓取的试管，并确保视觉传感器能够看到试管。

图 4 展示了观测姿态的生成过程，以及如何在其中考虑这两项约束。首先，我们在 Phoxi深度传感器的可见范围内，均匀采样由机器人手部持握的试管的位置和旋转。试管

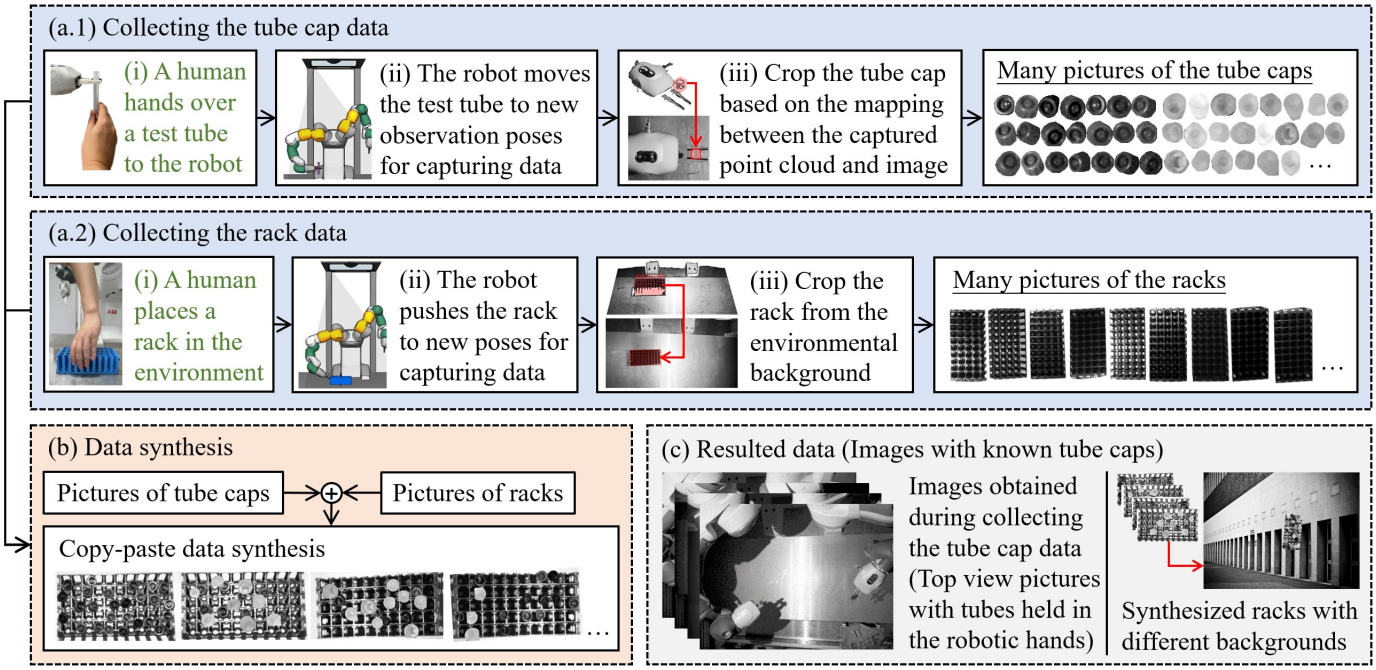


图 3：所提出的方法的工作流程。（a.1、2）数据采集组件。（b）数据合成组件。（c）生成的数据。

在采样姿态下采集的数据将具有丰富的光照条件，以及多种多样的可见试管边缘，有利于训练识别神经网络。特别是，试管旋转是根据四级二十面体球的顶点进行采样的[43]。二十面体球是一种顶点均匀分布的球面多面体。指向二十面体球顶点的向量有助于定义试管的旋转[1]。四级二十面体球有 642 个顶点，因此对应642 个向量和试管旋转姿态。得益于可见性约束，我们不会将试管移动到所有旋转姿态来采集数据，因为朝下的试管盖无法被 Phoxi传感器看到。我们根据这些向量与放置试管架的桌面法线之间的角度，对642 个向量进行筛选。与表面法线夹角较大的向量不可见，因此不会纳入考虑。图 4(b.1) 中的球面多面体展示了四级二十面体球。指向红色顶点的向量与表面法线的夹角大于θ 角度，因此会被剔除。绿色顶点是剩余的候选顶点。图4(b.1) 右侧的紫色试管花束展示了指向剩余候选顶点的向量所对应的试管姿态。

接下来，我们规划机器人运动，将机器人手部握持的试管移动到采样得到的试管位置和旋转角度。我们假设试管在手指中心处被垂直抓取。由于试管具有中心对称性，许多抓取姿态都符合这一假设。如图所示，抓取试管的手可以绕试管的对称轴自由旋转。4(b.2)。这种旋转是紧致的，并构成一个 SO(2) 群。为进行数值分析，我们以名为ω 的旋转间隔超参数对 SO(2) 群中的旋转进行采样，从而得到一系列离散化抓取姿态。机器人手部

[1]试管具有中心对称性。因此，无需考虑其绕中心轴的旋转。这样，指向二十面体球顶点的向量即可定义试管位姿。

图中的示意图4(b.2)展示了使用$\omega = 60°$得到的抓取姿态。这些采样得到的抓取姿态为机器人运动规划提供了许多候选目标，从而提高了成功移动和观察试管的可能性。

在确定要移动到的确切候选目标时，我们会检查机械臂连杆造成的遮挡，并避免选择会导致试管不可见的抓取姿态。具体而言，通过检查视觉多面体与机器人手部模型连杆之间的碰撞来判断遮挡情况。如图4（c.1）所示，视觉多面体由相机原点和机器人手部模型的顶点计算得到。当视觉多面体与机械臂连杆发生碰撞时，机械臂可能会遮挡试管，导致视觉传感器无法捕捉到试管。图 4（c.2）展示了这种情况。

B. 使用标注掩膜分割瓶盖图像

由于试管由人手递交，且 Phoxi 传感器从多个不同视角采集管帽数据，因此采集到的试管点云会动态变化并带有噪声。通过自主检测来提取管帽点云并不稳定。因此，我们不采用自主检测，而是在机器人手部的局部坐标系中准备一个标注掩膜，以帮助提取试管盖的点云。随后将提取出的点云反投影到对应的二维灰度图像中，以分割管帽区域图像。图 5 展示了该掩膜的细节，以及它如何帮助分割瓶盖区域。掩膜和反投影使我们能够精确分割瓶盖区域，同时避免将背景包含在内。

为了生成标注掩膜，我们将夹持试管的机器人手部移动到 Phoxi 传感器下方的固定位置，并触发传感器采集点云。我们可以

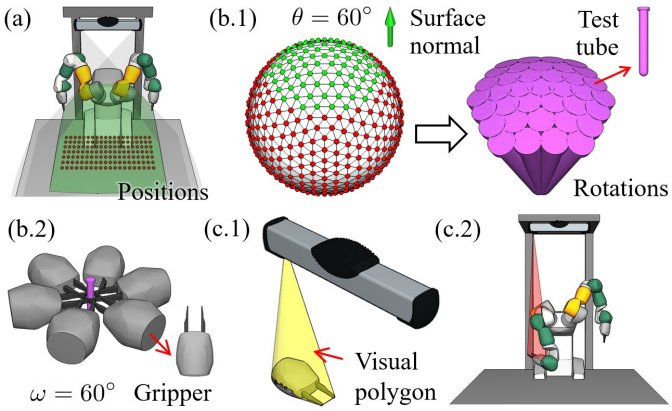


图4：(a) 观测位置采样。绿色区域是 Phoxi 扫描仪的可见区域。红点是采样位置。(b.1) 基于四级二十面体球进行旋转采样。左侧的球面多面体展示了二十面体球。绿色顶点是可行向量的端点，这些向量与表面法线的夹角小于 $\theta = 60°$ 角度。它们对应于右侧所示的管体旋转姿态。(b.2) 每个采样管体姿态对应的抓取姿态构成一个 SO(2) 群。为进行数值分析，按间隔 ω 对其进行采样。(c.1) 使用相机原点和机器人手部模型的顶点计算得到的视觉多面体。(c.2) 当视觉多面体与机械臂连杆发生碰撞时，抓取物体有被机械臂遮挡的风险。

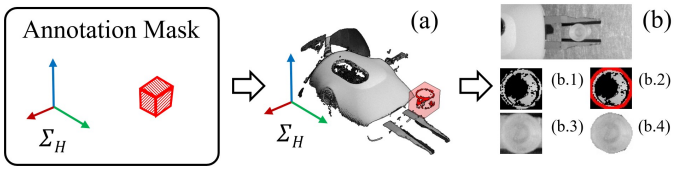


图5：使用标注掩膜提取管帽图像的工作流程。(a) 将以持物机器人手部的局部坐标系描述的掩膜应用于采集的点云。(b) 将提取出的点云反向投影至二维灰度图像，以分割管帽区域的图像。(b.1) 反向投影的结果可能包含不连通的像素。(b.2) 计算不连通像素的包围凸包。(b.3,4) 根据包围凸包分割管帽区域。

通过检查夹持手指上方的区域，轻松获取瓶盖的点云数据，并通过根据这些数据构建包围多面体来生成标注掩膜。然而，单个包围多面体可能不适用于其他情况，因为采集到的点云容易受到光反射或透视投影（自遮挡）的影响。因此，我们不再使用单个点云和多面体，而是从多个视角采集点云，在机器人手部的局部坐标系下将其合并，并将合并结果的边界框计算为标注掩膜。图6 展示了一个示例。多个视角的采样方式与上一小节中提到的观测位姿相同。不过，我们无需改变观测位置，因为我们的目标是

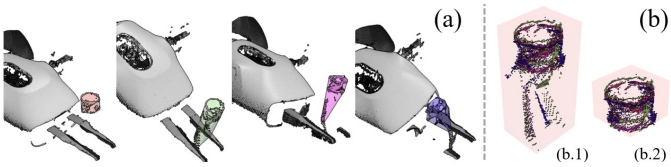


图 6：(a) 从不同视角采集数据。通过检查夹持手指上方的区域，获取管帽的点云。点云以彩色多面体突出显示。(b) 在机器人手部的局部坐标系下合并 (a) 中的管帽点云，并为合并结果计算一个边界框，作为标注掩膜。(b.1) 原始边界框。(b.2) 如有需要，可交互式调整边界框。

在手部坐标系中获取一个边界框掩膜。不同旋转角度下的视角能够提供足够的表面点云数据，以满足要求。请注意，合并结果可能包含由透明管体反射造成的噪声点数据，从而导致掩膜范围大于管帽。我们提供了一个交互式用户界面，可手动调整边界框尺寸，以尽量减少噪声造成的不利影响。此调整为可选操作，可在需要精确分割管帽区域时进行。

C. 复制粘贴合成

我们对分割后的试管盖进行随机缩放、模糊、亮度和对比度处理，然后将其粘贴到分割后的试管架背景上进行数据合成。粘贴时，我们允许管帽区域相互重叠，以模拟试管间遮挡。粘贴完成后，我们随机更换环境背景（试管架的背景），进一步缩小合成图像与真实世界中拍摄的图像之间的域差距。

这里的一个关键操作是，我们考虑试管与试管架共同出现的情况，并将试管盖图像粘贴到试管架上，而不是粘贴到随机背景上，例如[11]。我们从试管架图像中随机采样粘贴位置来粘贴试管盖，并使用粘贴数量 T 来控制画面杂乱程度。请注意，无需将试管盖精确粘贴在试管架孔的中心附近，因为试管在试管架孔内会随机倾斜。可见盖帽区域与孔边界或其他孔适度重叠也是合理的。

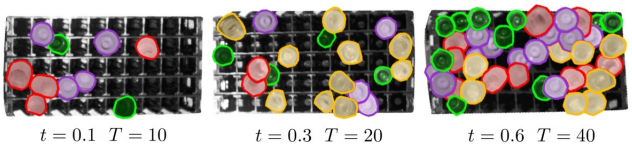
对于试管间遮挡，我们考虑视觉传感器的透视投影，并定义一个遮挡阈值t ，以允许可见盖帽区域之间存在重叠。视觉传感器的透视投影会导致试管架在某些视点出现相互遮挡。遮挡阈值有助于模拟遮挡，并定义分割后的瓶盖图像能够重叠或相互遮挡的最大百分比。图 7展示了 t 阈值的作用。它为粘贴操作增加了一项约束：先前粘贴的管帽图像“A”与之后粘贴的盖帽并集之间的重叠百分比必须低于 t 。图 $B \cup C \cup ...$ 中分子里的这一项7 表示在“A”之后粘贴的盖帽并集。随机生成新的管帽时，必须将其与这一项求并集

以确保不会违反对所有先前“A”的 t 约束。关于 t，有两点值得注意。第一，可以根据特定类型试管的高度分别设定其值。第二，其值与粘贴数量T相关。满足 t 阈值的试管架中，粘贴的试管盖最大数量可能小于给定的 T。在这种情况下，我们将粘贴的试管盖最大数量限制为较小值，以确保 t不会失效。

(a)

$$\text{Overlap}(A,B,C,...) = \frac{|A \cap (B \cup C \cup ...)|}{|A|} = \frac{\quad}{\quad} < t$$

(b)



$t = 0.1 \quad T = 10$ $t = 0.3 \quad T = 20$ $t = 0.6 \quad T = 40$

图7：(a) 使用阈值 t 模拟瓶盖遮挡。“A”表示先前粘贴的管帽区域。“B”、“C”……表示在“A”之后粘贴的瓶盖。(b) 不同t 值下的结果。

对于环境背景，我们使用 BG-20k数据集 [44] 获取高分辨率随机背景图像，并以 0.5 的概率更换合成图像的背景。

V. 实验与分析

我们开展了实验，对比使用所提出的方法及其他几种方法收集的数据集训练的 YOLOv5x [45]检测器，以了解其性能。表II 展示了这些方法。SR（粘贴到试管架合成）方法将随机选取的瓶盖图像粘贴到货架背景上，以合成训练数据。这代表了本文使用的合成方法。SB（粘贴到BG-20k合成）方法是一种替代合成方法。与粘贴到试管架上不同，该方法将随机选取的瓶盖图像粘贴到从BG-20k数据集中选取的随机背景上。RO（机器人观测）方法是机器人瓶盖分割的副产品，在该过程中，机器人手持试管采集数据。我们将 RO 视为一种独立方法，因为我们想了解手持观测是否足以用于训练。我们还组合了 RO、SR 和 SB 方法（** 行），以观察它们能否取得令人满意的性能。RO+SR 组合正是本文提出的方法。我们特别提出这种组合，是因为 RO 是机器人瓶盖分割的预处理步骤。采用 RO+SR 不会增加工作量。组合 RO+SB 或 RO+SR+SB 也是可选方案。它们的成本与使用独立的 SR 或 SB 数据[2]相同。最后，CL（众包标注）是一种传统方法，需要人工将装有试管的试管架放在机器人下方，并手动标注采集图像。图8 展示了使用不同方法采集的示例图像。

表I：数据收集方法汇总

缩写	全称	描述
SR	粘贴到试管架上进行合成	架子上的瓶盖
SB	通过粘贴到 BG-20k 进行合成	随机背景上的瓶盖
RO	机器人观测	由机器人手持的试管
**	上述方法的组合	SR+SB 是所提出的方法
CL	众包标注	桌面上的试管架里的试管

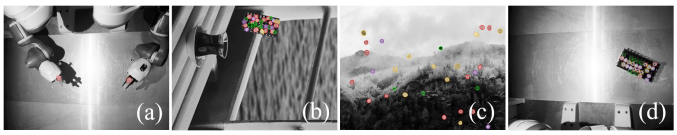


图 8：使用各种方法采集的示例图像。(a) RO。(b) SR。(c) SB。(d) CL。

A. 各数据集的性能

我们使用这些方法及其组合收集了各种数据集，用这些数据集训练 YOLOv5x 检测器，并使用测试数据集评估训练所得检测器的性能，以进行比较。

第一个数据集是 CL^{200}，作为比较的基准。采集该数据集时，我们在试管和试管架状态随机的情况下采集了 200 张图像，并使用 LabelImg[3] 手动标注了试管区域。这 200 张图像中总共包含 5916 个已标注实例。

第二个数据集是 SR^{1600}。为了收集该数据集，我们首先通过机器人观测准备了许多管帽图像。如图 2(b) 所示，我们选取了四种不同的试管，并利用 Yumi机器人的双臂快速收集管帽数据。对于每种试管，我们将两个相同的试管分别交给两条机器人手臂进行观测。每条手臂将其握持的试管移动到 400 个观测位姿进行数据收集。示例见图 9(a)。在这里，我们将超参数 θ 和ω 设为 $30°$ 和 $360°$（单个抓取位姿），并将位置设置为在桌面上以 $0.1m$ 的粒度均匀采样，从而生成观测位姿。在该参数设置下，每条手臂总共获得了 400 多个观测位姿，我们使用前 400 个位姿采集图像。因此，对于单种试管，我们获得了 400 张观测图像（每张图像中有两支试管，因此共 800 张管帽图像，示例见图 9(b)）；所有试管类型共获得 1600 张观测图像。考虑到点云映射，我们从观测图像中分割出 3200 张管帽区域图像。图 9(c) 展示了收集到的点云，其中管帽以绿色突出显示。图 9(d) 展示了分割后的管帽区域。除了管帽区域，我们还收集了 15 张带有试管架的图像（每张图像中有一个试管架），并分割出 15 张试管架图像。我们通过将从 3200 张管帽图像中随机选取的管帽粘贴到从 15 张试管架图像中随机选取的试管架上，合成了一个包含 1600 张图像的数据集（SR 方法）。合成过程中，我们将粘贴数量设为 $T = 30$，并将“Blue Tube”的遮挡阈值设为 t_{blue} =0.4，将其他试管的遮挡阈值设为 $t_{others} = 0.15$。我们选择这些参数设置，是因为“Blue Tube”较短，容易受到遮挡。我们提高了它的遮挡阈值

[2]合成数据通常被视为免费的，因为它只需要计算资源。因此，SR 和 SB 的成本取决于 RO 流程。

[3]https://github.com/heartexlabs/labelImg

以模拟其他试管造成的频繁视觉遮挡。此外，我们通过使用 Albumentations[4]$0.9 \sim 1.1$ 库应用随机缩放（相对于原始图像尺寸，0.5 的概率）、随机模糊（3×3 卷积核，0.5 的概率）、随机亮度（相对于原始亮度，0.5的概率）和随机对比度（相对于原始对比度值，0.5 的概率），增加了分割后的瓶盖图像的多样性。试管架的背景从BG-20k数据集中随机选取，变化概率为0.5 。

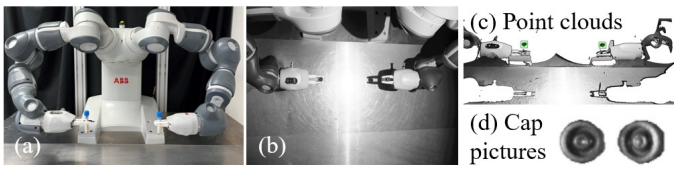


图 9：（a）机器人移动试管以进行观测。两个机械臂均投入使用。（b）观测图像。（c）由 Phoxi传感器采集的点云。（d）从观测图像中分割出的瓶盖图像。

第三个数据集是 RO^{1600}。它是机器人瓶盖分割的副产品，包含机器人观测期间获取的 1600 张观测图像。

第四个数据集是 SB^{1600}。与 SR^{1600}数据集相比，我们直接将随机选取的瓶盖粘贴到 BG-20k数据集的图像上以获取数据。粘贴的瓶盖可以自由分布在图像背景上。未使用分割后的试管架。粘贴数量 T 和遮挡阈值 t 分别为 35 和 0。15。不同试管的 t 没有差异。随机化方式与获取 SR^{1600}时相同。

我们还采用组合方法收集数据集，并研究组合是否能带来更好的结果。组合数据集包括 RO^{1600}+SR^{800}，、RO^{1600}+SB^{800}，、RO^{1600}+SR^{400}+SB^{400}，和 SR^{800}+SB^{800}。这里，方法名称右上角的上标数字表示使用该方法收集的图像数量。“+”符号表示数据集包含使用不同方法收集的数据。RO^{1600}+SR^{800} 数据集表示使用所提出的方法收集的数据。

表格II 左侧总结了各数据集。这些数据集用于训练 YOLOv5x 检测器，以进行比较。训练前，所有数据集对应的 YOLOv5x 检测器均使用 COCO数据集预训练得到的权重进行初始化。所有数据集中的图像均调整为 1376×1376分辨率。每个数据集按照4 比1 的比例划分为训练子集和验证子集。训练期间，训练子集以批量大小为2的设置输入训练程序，训练程序每个周期都会进行验证。当所有目标的平均精度均值（mAP） [46] 在99.0% 的 0.5 交并比（IoU）阈值下均超过该数值时，训练过程便会停止。这里，我们将检测到的边界框与真实标注中的瓶盖边界框的交并比大于0.5时，定义为检测正确。

[4]https://albumentations.ai/

为了评估使用各数据集训练的 YOLOv5x 检测器的性能，我们收集了一个包含 100 张图像的测试数据集，并采用与 CL 相同的方法标注真实标注。我们使用训练好的检测器检测测试数据集中的试管。与验证时一样，当检测到的边界框与真实标注中的瓶盖边界框的交并比大于 0.5 时，我们将其视为正确检测。我们使用平均精度（AP）指标衡量单个目标类别的检测性能，并使用平均精度均值（mAP）衡量所有目标的检测性能。由于通过一次验证取得满意结果的检测器未必是最佳检测器，因此我们对每个检测器训练两次，并将其在测试数据集上获得的较高精确率作为最终评估结果。

表 II 展示了评估结果。我们从中得出了以下观察结果和推测。

i) 如第 2 行（RO^{1600}）所示，使用通过机器人观测收集的数据集进行训练，表现最差。推测：数据集中的所有图像都具有相似的机器人背景，因此受到域偏移的影响。~~ii)~~ 合成数据集不一定能带来较高的平均精度，如第 3 行（SR^{1600}）和第 4 行（SB^{1600}）所示。SR^{1600} 数据集的表现优于 SB^{1600} 数据集。

推测：复制粘贴合成未能涵盖某些视觉上下文；将对象粘贴到试管架上（SR）能提供更多有效的视觉上下文，相比粘贴到随机背景上（SB），对神经网络的帮助更大。

iii) 将合成数据集与机器人观测相结合是有效的。比较第 5、6、7 行（RO^{1600}+SR^{800},RO^{1600}+SB^{800},RO^{1600}+SR^{400}+SB^{400}）与第 2、3、4 行（RO^{1600}，SR^{1600}，和 SB^{1600}）即可得出这一结论。前几行的平均精度均值高于后几行。

推测：机器人观测数据集还提供了有用的视觉上下文。

iv) 第5行（RO^{1600}+SR^{800}）的平均精度均值比第6行（RO^{1600}+SB^{800}）高2.4%。尤其是，第5行中“Blue Tube”的平均精度比第6行高8.7%。其他试管的平均精度也提升了0.1% $\sim$ 0.7%。

推测：将试管架视为局部上下文有助于提升领域特定性能；较短的“Blue Tube”很容易被遮挡。使用 SR 方法收集的数据集包含更多模拟遮挡。这些遮挡对于识别较短的“Blue Tube”很重要。

v) 第7行（RO^{1600}+SR^{400}+SB^{400}）的平均精度均值略高于第5行（RO^{1600}+SR^{800}）（高0.3%）。推测：将图像贴到~~试管架上~~（SR）提供了更好的领域特定特征。试管的随机背景对神经网络略有帮助；如果目标上下文有限，则不那么必要。

vi) 第 5 行（RO^{1600}+SR^{800}）与第 1 行（CL^{200}）相比具有竞争力。平均精度均值低.7%。BlueTube 和 WhiteTube 的平均精度分别低.5% 和 1.1%。紫色管的平均精度

表 II：使用不同数据集训练的检测器比较

ID	数据集名称	# 帽子	备注	AP				mAP
				Blue	紫色	白色	紫色环	
1	CL^{200}	5916	每幅图像多根试管	**0.993**	**0.995**	**0.989**	0.984	**0.990**
2	RO^{1600}	3200	每幅图像两根试管	0.380	0.923	0.695	0.630	0.657
3	SR^{1600}	40000	t_{blue}=0.4 & t_{others}=0.15, $T=25$	0.955	0.979	0.871	0.953	0.940
4	SB^{1600}	56000	t=0.15 （所有试管均相同）， $T=35$	0.808	0.978	0.812	0.897	0.874
5	RO^{1600}+SR^{800}	23200	参见注释2	0.968	**0.995**	0.978	0.992	0.983
6	RO^{1600}+SB^{800}	31200	参见注释2	0.881	0.994	0.971	0.992	0.959
7	RO^{1600}+SR^{400}+SB^{400}	27200	参见注释2	0.969	0.994	0.986	**0.993**	0.986
8	SR^{800}+SB^{800}	48000	参见注释2	0.973	0.993	0.969	0.985	0.980

注释1：最大平均精度和平均精度均值以粗体突出显示。注释2：组合数据集采用与各自数据集相同的参数采集。

相同。紫色环试管的平均精度高.8%。

推测：机器人观测和粘贴到试管架合成相互弥补了各自的不足；仍有一些极端情况可以通过人工标注，但机器人观测或合成未能涵盖，尤其是“Blue Tube”。

图10展示了若干失败案例，以便读者深入了解我们的观测和推测。图10(a) 和 (b) 展示了使用第 5 个数据集（RO^{1600}+SR^{800}）和第 6 个数据集（RO^{1600}+SR^{800}）训练的检测器的识别结果。后者未能识别被遮挡的试管，因为训练数据集中的模拟遮挡较少。该示例与 iv) 中的观测和推测一致。图10(c) 和 (d) 展示了使用第 5 个数据集（RO^{1600}+SB^{800}）训练的检测器未能处理的情况。第一种情况下，其他试管投下的阴影落在蓝色试管盖上，检测器未能识别该试管。第二种情况下，由于试管主体与试管架重叠造成的视觉错觉，检测器将水晶试管管身误识别为“Blue Tube”管帽。这两个失败案例与 vi) 中的观测和推测一致。合成数据集不包含阴影或试管主体。使用这些数据集训练的检测器在这些情况下的表现不如使用众包真实世界数据训练的检测器。

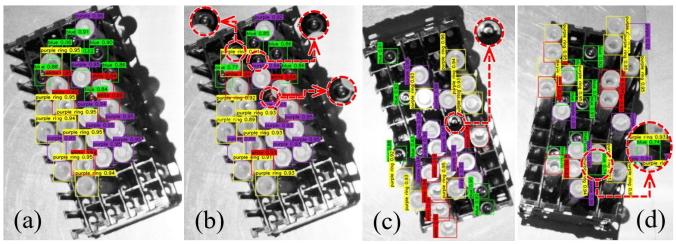


图10：(a) 使用第5个数据集训练的检测器成功识别了所有试管。(b) 使用第6个数据集训练的检测器未能识别红圈中的遮挡试管。(c) 使用第5个数据集训练的检测器未能识别红圈中的阴影试管。(d) 使用第5个数据集训练的检测器将红圈中的试管管身误识别为“Blue Tube”。

总之，各种训练数据集的结果表明，将使用 RO 和 SR 方法收集的数据结合起来是有效的。这个结论令人满意，因为 RO 方法是 SR 方法的子集。收集这些数据的工作流程简单明了。不过，我们想知道能否减少 RO 数据集中的图像数量，因为收集这些图像需要大量人工操作。这一疑问促使我们在下一小节开展相关研究。

B. 消融研究

在本小节中，我们针对组合后的 RO+SR 数据集开展多项消融研究，以进一步了解：1）数据组合比例的影响；2）生成合成数据时所用的粘贴数量 T 和遮挡阈值 t 的影响。

1）数据组合比例的影响：研究数据组合比例影响的实验分为两部分。第一部分中，我们将使用 RO 方法采集的图像数量设为800 ，并以200 至 1600按2倍比例变化，以了解 SR 数据的重要性。表III的上半部分展示了使用不同数据训练的检测器的精确率。结果表明，当 SR 图像数量从200 增加到1600时，平均精度均值有所提高。第二部分与第一部分类似。在这一部分中，我们将使用 SR 方法采集的图像数量固定为800 ，并将使用 RO 方法采集的图像数量从 200 变化至 1600 ，按两倍比例变化，以了解 RO 数据的重要性。表III 的下半部分展示了使用不同数据训练的检测器的精确率。结果表明，当 RO 图像数量从 200 增加到1600时，平均精度均值有所提高。

2）超参数的影响：除了数据组合比例之外，我们还研究了粘贴数量T 和遮挡阈值 t 对 SR 方法的影响。我们将 RO 和 SR 图像数量都设为800 ，并观察使用不同 T和 t 值采集的数据集训练出的检测器的性能。虽然我们之前对“Blue Tube”使用了不同的 t 值，但此处没有区分这些试管。与不同数据组合比例的研究类似，本研究也分为两个部分。在第一部分中，我们将 t 设为0。1 ，并将 T 从10 增至40 ，步长为 10。表IV 上半部分展示了参数变化对精确率的影响。

表 III：数据组合比例的影响

数据集名称	AP				mAP
	Blue	紫色	白色	紫色环	
RO^{800}+SR^{200}	0.958	0.994	0.973	0.987	0.978
RO^{800}+SR^{400}	0.964	0.992	0.975	0.985	0.979
RO^{800}+SR^{800}	0.966	**0.995**	0.979	**0.986**	**0.981**
RO^{800}+SR^{1600}	**0.970**	0.994	**0.987**	0.987	0.985
RO^{200}+SR^{800}	0.962	0.992	0.952	0.978	0.971
RO^{400}+SR^{800}	0.965	0.992	**0.979**	0.978	0.979
RO^{800}+SR^{800}	0.966	**0.995**	**0.979**	0.986	0.981
RO^{1600}+SR^{800}	**0.968**	**0.995**	0.978	**0.992**	**0.983**

注释 1：最大的平均精度和平均精度均值以粗体标出。注释2：我们使用以下超参数设置 $t_{blue} = 0.4$ & $t_{others} = 0.15$, $T = 30$ 来收集SB数据集。数值与第五节 V.A节中的实验相同。

结果显示，精确率从10到30显著提升。然而，更大的 T 对识别性能影响甚微。在第二部分中，我们将 T 设为30，并将 t从0。20 增至0。80，步长为0。2。表IV 下半部分展示了参数变化对精确率的影响。结果显示，“Blue Tube”的精确率明显提高。我们推测，这是因为“Blue Tube”较短，容易受到遮挡。较大的t 有助于在训练数据集中加入更多遮挡情况，从而提高检测率。结果还表明，随着 t 增加，“White Tube”和“Purple Ring Tube”的精确率变化不规律。它们较高，不会受到遮挡。为它们添加遮挡会造成意料之外的错误。为了全面了解情况，我们建议感兴趣的读者对照表上半部分第三行，观察从 $t = 0.1$ 开始的变化。上半部分第三行的T 值与下半部分各行相同。

表 IV：用于合成的参数的影响

参数 (T, t)	AP				mAP
	Blue	紫色	白色	紫色环	
(10, 0.10)	0.904	**0.995**	0.971	0.973	0.961
(20, 0.10)	0.915	**0.995**	0.976	0.985	0.968
(30, 0.10)	**0.939**	**0.995**	**0.987**	**0.992**	**0.978**
(40, 0.10)	0.934	0.994	0.972	0.983	0.970
(30, 0.20)	0.945	**0.995**	**0.985**	0.989	0.978
(30, 0.40)	0.969	**0.995**	**0.985**	**0.994**	0.986
(30, 0.60)	0.985	**0.995**	0.965	0.967	0.978
(30, 0.80)	**0.987**	**0.995**	0.984	0.988	**0.988**

*最大平均精度和平均精度均值以粗体突出显示。

C. 关于合成数据的进一步分析

我们还研究了帽子变化和组合比例对合成数据集的影响（使用 SR、SB 或 SR+SB 方法收集的数据集）。我们的目标是了解通过合成所能达到的最佳性能。

首先，我们将 SR 和 SB 方法收集的图像数量分别固定为 800 张。为研究帽子变化的影响，我们将用于合成的帽子区域图像数量（等于观测图像数量的两倍）从 400 张增加到 3200 张。以 2 倍的比例变化。使用不同数据集的 YOLOv5x 检测器的精确率见表 V。结果显示，400 这一行的精确率与 1600 行或 3200 行相比具有竞争力。这个数量足以训练出令人满意的检测器。因此，我们认为帽子变化对学习的影响较小。

其次，我们将帽子区域图像的数量固定为3200，并分别调整通过 SR 和 SB 方法采集的图像数量，以研究组合比例的影响。与第V-B1节中的消融研究类似，我们将此处的实验分为两部分。第一部分中，我们将通过 SR 方法采集的图像数量设为800，并以 2 倍的比例将通过 SB 方法采集的图像数量从200 调整至1600，以了解 SB 数据的重要性。表VI上半部分展示了使用不同数据训练的检测器的精确率。SB 图像数量似乎与最终检测器的精确率并无正相关，尽管 SB 图像数量为800时，观察到的平均精度均值最高。第二部分中，我们将通过 SB 方法采集的图像数量固定为800，并调整通过 SR 方法采集的图像数量，以了解 SR 数据的重要性。表VI下半部分展示了使用不同数据训练的检测器的精确率。结果表明，随着 SR 图像数量增加至800，平均精度均值有所提升。图像数量从800增加至1600时，性能没有显著差异。

表 V：#帽子对合成的影响

#帽子	AP				mAP
	Blue	紫色	白色	紫色环	
400	0.970	0.994	0.969	0.984	0.979
800	0.971	0.993	0.954	0.976	0.973
1600	0.971	0.992	0.980	0.985	0.982
3200	0.973	0.993	0.969	0.985	0.980

表 VI：SR 和 SB 比例的影响

数据集名称	AP				mAP
	Blue	紫色	白色	紫色环	
SB^{200} +SR^{800}	**0.975**	**0.994**	0.943	0.957	0.967
SB^{400} +SR^{800}	0.973	0.990	0.960	0.970	0.973
SB^{600} +SR^{800}	0.967	0.988	0.951	0.968	0.969
SB^{800} +SR^{800}	0.973	0.993	**0.969**	**0.985**	**0.980**
SB^{1600} +SR^{800}	0.952	0.978	0.926	0.972	0.957
SB^{800} +SR^{200}	0.951	0.982	0.925	0.963	0.955
SB^{800} +SR^{400}	0.932	0.969	0.914	0.952	0.942
SB^{800} +SR^{600}	0.966	0.990	0.930	0.893	0.945
SB^{800} +SR^{800}	0.973	**0.993**	**0.969**	0.985	**0.980**
SB^{800} +SR^{1600}	**0.975**	**0.993**	0.967	**0.986**	**0.980**

注1 最大的平均精度和平均精度均值以粗体突出显示。注2 我们使用以下超参数设置 $t_{blue} = 0.4$ & $t_{others} = 0.15$, $T = 30$ 来收集SB数据集，并使用以下超参数设置 $T = 30$, $t = 0.15$（对所有试管均相同）来收集SR数据集。这些数值与V.A节实验中的数值相同。注2 对于两种方法，我们均使用了3200张分割后的管帽区域图像。

VI. 结论

本文提出了一种用于数据准备的机器人观测与数据合成一体化框架。该框架能够显著减少数据准备所需的人力。它只需执行一次流程，便能以低成本生成组合数据。实验结果表明，组合观测图像与合成图像的性能与人工数据准备相当。消融研究为优化使用组合数据训练检测器时的数据配置和参数设置提供了良好指导。

参考文献

[1] C. Shorten 和 T. M. Khoshgoftaar，《深度学习图像数据增强综述》，大数据期刊，第 6 卷，48，2019年 1，1第。页 – 。[2] A. Gupta、A. Vedaldi 和 A. Zisserman，《用于自然图像中文本定位的合成数据》，载于*IEEE*计算机视觉与模式识别会议论文集，2016年，第 2315–2324 页。[3] X. Yi、E. Walia 和 P. Babyn，《医学成像中的生成对抗网络：综述》，医学图像分析，第 58 卷，101552，2019年 页。[4] S. Hinterstoisser、O. Pauly、H. Heibel、M. Martina 和 M. Bokeloh，《省下一个标注，就多赚一个标注：使用完全合成训练进行目标检测》，载于*IEEE/CVF*国际计算机视觉会议研讨会，2019年，第 0–0 页。[5] K. Karsch、V. Hedau、D. Forsyth 和 D. Hoiem，《将合成物体渲染到旧照片中》，*ACM*图形学汇刊，第 30 卷，第 6 期，第 1–12 页，2011年。[6] P. P´erez、M. Gangnet 和 A. Blake，《泊松图像编辑》，*ACM*图形学汇刊，第 22 卷，第 3 期，第 313–318 页，2003年。[7] D. Dwibedi 、 I. Misra 和 M. Hebert，《剪切、粘贴并学习：令人惊讶地简单的实例检测合成方法》，载于*IEEE*国际计算机视觉会议，2017年，第 1301–1310 页。[8] G. Georgakis、A. Mousavian、A. Berg 和 J. Ko˘seck´a，《为室内场景中的目标检测合成训练数据》，载于机器人学：科学与系统，07 2017年。[9] G. Ghiasi、Y. Cui、A. Srinivas、R. Qian、T.-Y. Lin、E. D. Cubuk、Q. V. Le 和 B. Zoph，《简单复制粘贴是实例分割的强大数据增强方法》，载于*IEEE/CVF*计算机视觉与模式识别会议，2021年，第 2918–2928 页。[10] V. Florence、J. J. Corso 和 B. Griffin，《用于目标分割的机器人监督学习》，载于*IEEE*国际机器人与自动化会议，2020年，第 1343–1349 页。[11] W. Boerdijk、M. Sundermeyer、M. Durner 和 R. Triebel，《“这是什么？”——从操作序列中学习分割未知物体》，载于*IEEE*国际机器人与自动化会议，2021年，第 10 160–10 167 页。[12] D. Pathak、Y. Shentu、D. Chen、P. Agrawal、T. Darrell、S. Levine 和 J. Malik，《通过交互学习实例分割》，载于*IEEE*计算机视觉与模式识别会议研讨会，2018年，第 2042–2045 页。[13] W. Boerdijk、M. Sundermeyer、M. Durner 和 R. Triebel，《从机器人运动中自监督分割夹爪中的物体》，载于机器人学习会议，2020年。[14] B. Sapp、A. Saxena 和 A. Y. Ng，《一种用于物体识别的快速数据收集与增强流程》，载于*AAAI*，2008年，第 1402–1408 页。[15] M. Suchi、T. Patten、D. Fischinger 和 M. Vincze，《EasyLabel：用于创建机器人RGB-D数据集的半自动逐像素目标标注工具》，载于*IEEE*国际机器人与自动化会议，2019年，第 6678–6684 页。[16] T. Kiyokawa、K. Tomochika、J. Takamatsu 和 T. Ogasawara，《利用噪声屏蔽视觉标记实现完全自动标注，用于基于深度学习的目标检测》，*IEEE*机器人与自动化快报，第 4 卷，第 2 期，第 1972–1977 页，2019年。[17] D. De Gregorio、A. Tonioni、G. Palli 和 L. Di Stefano，《面向机器人深度学习的半自动标注》，*IEEE*自动化科学与工程汇刊，第 17 卷，第 2 期，第 611–620 页，2020年。

[18] P. Fitzpatrick 和 G. Metta，《通过实验性操作实现视觉落地》，《伦敦皇家学会哲学汇刊》A辑：数学、物理与工程科学，第 361 卷，第 1811 期，第 2165–2185 页，2003年。[19] K. Welke、J. Issac、D. Schiebener、T. Asfour 和 R. Dillmann，《在类人机器人上自主获取用于物体识别的视觉多视角物体表征》，载于 IEEE国际机器人与自动化会议。IEEE，2010年，第 2012–2019 页。[20] B. Browatzki、V. Tikhanoff、G. Metta、H. H. B¨ulthoff 和 C. Wallraven，《类人机器人上的主动式物体识别》，载于 IEEE国际机器人与自动化会议，2012年，第 2021–2028 页。[21] M. Krainin、P. Henry、X. Ren 和 D. Fox，《用于手持式三维物体建模的机械臂与物体跟踪》，《国际机器人研究杂志》，第 30 卷，第 11 期，第 1311–1327 页，2011年。[22] M. Bj¨orkman 和 D. Kragic，《主动式三维场景分割与未知物体检测》，载于 IEEE国际机器人与自动化会议，2010年，第 3114–3120 页。[23] T. F¨aulhammer、R. Ambrus、C. Burbridge、M. Zillich、J. Folkesson、N. Hawes、P. Jensfelt 和 M. Vincze，《移动机器人上的物体模型自主学习》，IEEE机器人与自动化快报，第 2 卷，第 1 期，第 26–33 页，2016年。[24] C. D. Singh、N. J. Sanket、C. M. Parameshwara、C. Ferm¨uller 和 Y. Aloimonos，《NudgeSeg：通过反复物理交互实现零样本物体分割》，载于 IEEE/RSJ智能机器人与系统国际会议，2021年，第 2714–2712 页。[25] D. Schiebener、A. Ude 和 T. Asfour，《通过物理交互分割未知的有纹理和无纹理刚性物体》，载于 IEEE国际机器人与自动化会议，2014年，第 4959–4966 页。[26] A. Eitel、N. Hauff 和 W. Burgard，《通过物理交互进行实例分割的自监督迁移学习》，载于 IEEE/RSJ智能机器人与系统国际会议，2019年，第 4020–4026 页。[27] S. K. Divvala、D. Hoiem、J. H. Hays、A. A. Efros 和 M. Hebert，《目标检测中上下文的实证研究》，载于 IEEE计算机视觉与模式识别会议，2009年，第 1271–1278 页。[28] E. Barnea 和 O. Ben-Shahar，《探索上下文对目标检测的效用边界》，载于 IEEE/CVF计算机视觉与模式识别会议，2019年，第 7412–7420 页。[29] S. K. Divvala、D. Hoiem、J. H. Hays、A. A. Efros 和 M. Hebert，《目标检测中上下文的实证研究》，载于 IEEE计算机视觉与模式识别会议，2009年，第 1271–1278 页。[30] N. Dvornik、J. Mairal 和 C. Schmid，《视觉上下文建模是增强目标检测数据集的关键》，载于 欧洲计算机视觉会议，2018年，第 364–380 页。[31] H. Wang、Q. Wang、H. Zhang、J. Yang 和 W. Zuo，《用于目标检测的受约束在线剪切粘贴》，IEEE视频技术电路与系统汇刊，第 31 卷，第 10 期，第 4071–4083 页，2020年。[32] T. Hoda˘n 、 V. Vineet 、 R. Gal 、 E. Shalev 、 J. Hanzelka 、 T. Connell、P. Urbina、S. N. Sinha 和 B. Guenter，《用于物体实例检测的照片级真实感图像合成》，载于 IEEE图像处理国际会议，2019年，第 66–70 页。[33] S. R. Richter、V. Vineet、S. Roth 和 V. Koltun，《为数据而战：从电脑游戏中获取真实标注》，载于 欧洲计算机视觉会议，2016年，第 102–118 页。[34] J. Tobin、R. Fong、A. Ray、J. Schneider、W. Zaremba 和 P. Abbeel，《通过域随机化将深度神经网络从仿真迁移到真实世界》，载于 IEEE/RSJ智能机器人与系统国际会议，2017年，第 23–30 页。[35] A. Carlson、K. A. Skinner、R. Vasudevan 和 M. Johnson-Roberson，《对相机效应建模以改善基于合成数据的视觉学习》，载于 欧洲计算机视觉会议研讨会，2018年，第 0–0 页。[36] A. Prakash、S. Boochoon、M. Brophy、D. Acuna、E. Cameracci、G. State、O. Shapira 和 S. Birchfield，《结构化域随机化：通过上下文感知合成数据弥合现实差距》，载于 IEEE国际机器人与自动化会议，2019年，第 7249–7255 页。[37] J. Tremblay、A. Prakash、D. Acuna、M. Brophy、V. Jampani、C. Anil、T. To、E. Cameracci、S. Boochoon 和 S. Birchfield，《使用合成数据训练深度网络：通过域随机化弥合现实差距》，载于 IEEE计算机视觉与模式识别会议研讨会，2018年，第 969–977 页。[38] A. Tsirikoglou、G. Eilertsen 和 J. Unger，《面向视觉机器学习的图像合成方法综述》，载于 《计算机图形论坛》，第 39 卷，第 6 期，2020，第 426–451 页。

[39] X. Yang、X. Fan、J. Wang 和 K. Lee，“用于工业目标检测和姿态估计的基于图像转换的合成数据生成”，IEEE机器人与自动化快报，2022年。[40] M. Sundermeyer、Z.-C. Marton、M. Durner、M. Brucker 和 R. Triebel，“基于RGB图像的六维目标检测隐式三维方向学习”，载于欧洲计算机视觉会议，2018年，第699–715页。[41] S. Chatterjee、D. Hazra、Y.-C. Byun 和 Y.-W. Kim，“利用迁移学习和基于GAN的合成数据增强提升图像分类”，Mathematics，第10卷，第9期，第1541页，2022年。[42] I. Lundberg、M. Bj¨orkman 和 P. ¨Ogren，“使用点标记进行机器人视觉系统的内参标定与手眼标定”，载于IEEE-RAS国际人形机器人会议，2014年，第59–66页。[43] W. Wan 和 K. Harada，“单臂与双臂重新抓取的开发和比较”，IEEE机器人与自动化快报，第1卷，第1期，第243–250页，2016年。[44] J. Li、J. Zhang、S. J. Maybank 和 D. Tao，“连接合成图像与真实图像：迈向端到端深度图像抠图”，国际计算机视觉期刊，第130卷，第2期，第246–266页，2022年。[45] G. Jocher 等，“ultralytics/yolov5：v3.1 - 错误修复与性能改进”，2020年10月。 [在线]。获取地址：https: //doi.org/10 5281 4154370 。/zenodo。Z. Zou、Z. Shi、Y. Guo [46] 和 J. Ye，“20年来的目标检测：综述”，arXiv预印本 arXiv:1905.05055，2019年。