arXiv:2303.00423v1 [cs.HC] 1 Mar 2023

基于共享注视的多模态人机交互实现多视角未知物体教学

Daniel Weber Wolfgang Fuhl 蒂宾根大学 蒂宾根大学
daniel.weber@uni-tuebingen.de wolfgang.fuhl@uni-tuebingen.de
EnkelejdaKasneci Andreas Zell 慕尼黑工业大学 蒂宾根大学
enkelejda.kasneci@tum.de andreas.zell@uni-tuebingen.de

2023年3月2日

摘要

为了使机器人能够成功部署于复杂多样的情境中，机器人必须理解其所处的环境。随着最先进的目标检测器性能不断提升，机器人检测其交互范围内物体的能力也随之增强。然而，这使机器人受限于少数经过训练的类别，无法适应预定义场景之外的陌生环境。在此类场景中，面对数量庞大的交互实体，人类可以协助机器人，并以教师的身份传授必要的专业知识。我们提出了一种新颖的流程，在人机协作情境中有效利用人类注视和增强现实，教机器人识别其周围环境中的新颖物体。通过将视线（引导机器人关注感兴趣对象）与增强现实（传达相应的类别信息）相结合，我们使机器人能够快速自主获取大量自动标注的训练数据。我们采用迁移学习方式进行训练，展示机器人检测近期学会物体的能力，并评估不同机器学习模型、学习过程以及训练数据量对检测效果的影响。事实证明，我们的多模态方法能够基于少量实例，以高效且自然的方式教机器人识别新颖物体，并使其能够检测没有可用训练数据集的类别。此外，我们向研究界公开提供数据集，其中包含RGB和深度数据、相机内参和外参，以及感兴趣区域。

1 引言

随着技术进步，越来越多的机器人被应用于工业领域，其应用范围也日益多样化。汽车、电子、橡胶和塑料、化妆品、制药以及食品和饮料等众多行业，都受益于机器人卓越的精度、效率、工作能力，以及对艰苦和危险环境的耐受力 [52]。在人类周围的环境中，机器人也越来越多地以服务助理的形式投入使用，例如在 Walmart 等超市 [2, 19] ，或在博物馆担任导游 [41, 46]。这一成功也得益于机器学习领域，尤其是计算机视觉领域的最新进展，这些技术使机器人能够理解其环境并检测其中的物体和人员。然而，一个几乎始终成立的核心假设是：存在大量带有高质量标签的训练数据。因此，许多大型汽车制造商或 Google、Tesla 和 Uber 等公司都建立了自己的图像和视频数据库，其中大多数是通过 Amazon Mechanical Turk 等众包平台外包完成的 [39]。对于在仓库或办公环境中运行的服务机器人而言，往往没有公开可访问的数据集能够针对相应环境量身定制、涵盖所有相关物体类别并且具备完整标注。因此，在有充足训练数据的情况下，最先进的目标检测器表现出色，但其应用仍局限于训练数据预先定义的特定场景 [51]。一旦某个物体未被纳入其中，就会超出

目标检测器的能力范围，并将机器人逼至其能力极限。事实上，与现实中存在的物体数量相比，数据集所涵盖的物体比例微乎其微。例如，ImageNet [6], 作为最大的公开可用图像数据集之一，仅包含 1000 个类别，而现实世界中存在的类别数量显然远远超过这一数字。这一事实阻碍了机器人在未知环境中的部署，也妨碍其动态适应陌生条件。

在这项工作中，我们旨在提升移动机器人的任务执行能力，同时减轻其对数据的依赖：通过利用其对物体已有的通用认知，并添加新的类别，扩展机器人的基础知识。具体而言，我们向机器人教授新颖的未知物体，使其能够在学习过程中所在的环境中重新检测这些物体。为此，我们采用流畅、智能的人机交互方式，结合计算机视觉、眼动追踪和增强现实（AR）等研究领域。借助增强现实，我们实现了多模态通信通道：利用人类注视将机器人的注意力引向某个物体，并通过语音或手势向机器人传达相关类别信息。随后，机器人通过视觉对感兴趣对象进行分割，并从略有不同的角度拍摄一系列图像。在这一便捷、易用的过程中获取的数据包含丰富的信息，除了 RGB 图像和深度图像外，还包括相机的内参和外参以及感兴趣区域。结合人类提供的类别信息，机器人据此学习相应物体。

总之，我们的主要贡献如下：

1. 我们提出了一种全新的流程，用于教机器人认识新的未知物体。2. 为实现这一目标，我们在人机交互场景中结合视线与增强现实，从而能够高效、快速地获取大量带标注的训练数据。3. 我们使用多种模型、不同的学习方法以及不同数量的数据，对学习过程进行了详细评估。4. 我们推出了多视角细节物体（OMD），这是一个用途广泛的数据集，并通过 `https://cloud.cs.uni-tuebingen.de/index.php/s/2oRPs2o3FZkdBHW` 向研究界公开。5. 我们通过 `https://github.com/dnlwbr/Multiperspective-Teaching` 向研究界公开了我们的系统及其完整代码库。

2 相关工作

要通过多模态人机交互教机器人认识未知物体，需要开展大量多学科协作，本节将对此进行讨论。内容涵盖：1）机器人学中的增强现实；2）计算机视觉领域中眼动追踪的既有应用；3）未知物体检测；4）机器人如何学习；以及 5）机器人如何与人类协作。

2.1 机器人学中的增强现实

随着增强现实设备日益普及，工业应用和相关研究也不断拓展 [25]。尤其是增强现实与机器人辅助手术的结合，展现出巨大潜力 [32],：人类仍可通过增强现实进行控制，同时还能利用机器人的精确性和一致性 [44]。在机器人控制和路径规划方面， [18]使用增强现实设备控制工业机械臂执行拾放任务，而 [34] 则设计了一种增强现实界面，用于规划、预览和执行机械臂的轨迹。在多智能体系统中，增强现实也被用于提供视觉反馈 [47]以及远程控制 [35] 机器人集群。因此，通过增强现实增强对现实世界的感知，已被证明是一种颇具吸引力的机器人交互方式。

2.2 眼动追踪与计算机视觉

眼动追踪也已成为研究和工业领域中的重要工具 [16]。因此，[16] 开发了一款可在手机或平板电脑等移动设备上运行的眼动追踪软件

且除相机外无需任何其他传感器。 [31], 的作者利用计算机视觉（和增强现实）分析了移动式三维眼动追踪数据。其中包括追踪三维标记，并将其与虚拟代理对齐。在 [43],中，通过将已知物体的特征与人类注视点邻域中的特征进行匹配，识别出了物体类别。然而，这要求物体此前已出现在专门为此目的创建的数据库中，而且也无法说明物体的位置。尽管如此，眼动追踪数据仍有助于提升分割算法的性能 [28]。因此，在 [30],中，利用注视训练模型来标注物体位置；在 [50]中，则将视线融入点云的分割过程。不过，在这两种情况下，都无法说明物体类别。

2.3 未知物体检测

在 [51] 中，也使用眼动追踪来解决未知物体检测问题。在这项工作中，区域提议方法生成的候选边界框数量大幅减少，但无法进行分类。 [49] 使用热图而非场景帧，根据视频片段中人物是否正在注视某个物体进行分类，并确定相应边界框的参数。检测到的物体均为未知物体，但同样未能确定其类别。 [13] 的作者使用单类支持向量机解决未知物体检测问题。由于学习过程是增量式的，因此多个机器人参与其中，并通过云端站点相互连接，所有处理都在该站点进行。在这种方法中，未知物体仅被标记为未知，并未将类别信息纳入学习过程。其目的是从已分类物体中筛选出未知物体，只转发已知物体，以避免向机器人发送错误信息。在另一项近期工作中， [20] 提出利用语义分割模型的额外预测结果，并量化所提出分割的不确定性。同样，分类任务是二元的，只区分已知和未知类别，最终并未学习这些物体。

2.4 机器人与机器的教学

机器人广泛应用于各种领域，尤其是工业领域。其中包括教授装配任务 [48],，让机器人从人类示范中学习：机器人先观察人类，然后模仿人类的动作。 [10]提出了另一种通过用户交互进行学习的方法。在这种方法中，人们使用自然语言向机器人说明它应该执行哪些任务。抓取物体就是此类典型任务之一。解决这一挑战是当前研究的一部分，例如 [14], 采用物体检测方法来学习良好的抓取姿态。正如 [3], 所述，数据驱动方法通常通过提供带标签样本形式的训练数据、试错或人类示范来实现。因此，为了在没有事先进行离线数据生成的情况下实现灵活的学习过程，人类与机器人之间的交流在这里也很重要。

2.5 协作场景中的联合注意

除其他研究外， [17] 和 [8],也对人类与机器人在协作场景中的教学进行了研究。在 [17], 中，研究者利用自然语言上下文对视觉物体进行单样本学习，使机器人能够识别被描述的物体。然而，在这一概念验证中，物体必须能够通过颜色或空间关系明确区分，其组成部分也必须能够用语言表达唯一描述。在 [8], 中，研究者提出了一种用于物体分类的教学系统。该系统允许用户查看分类的中间状态，也就是机器人会将物体归入哪个类别。通过交互，可以改进和纠正分类结果，但所有物体都必须标记增强现实标记，才能被识别。在结合抓取任务方面， [45] 和 [5] 也教会机器人识别新物体。然而，在这两种情况下，每个类别中用于训练和测试的物体完全相同，这会使结果产生正向偏差。我们为每个类别使用多个不同的物体，这更符合真实世界条件。

在本研究中，我们结合眼动追踪和增强现实这两个不同的研究领域，使机器人与人类之间能够进行简单而自然的交互，共同解决一个计算机视觉问题。

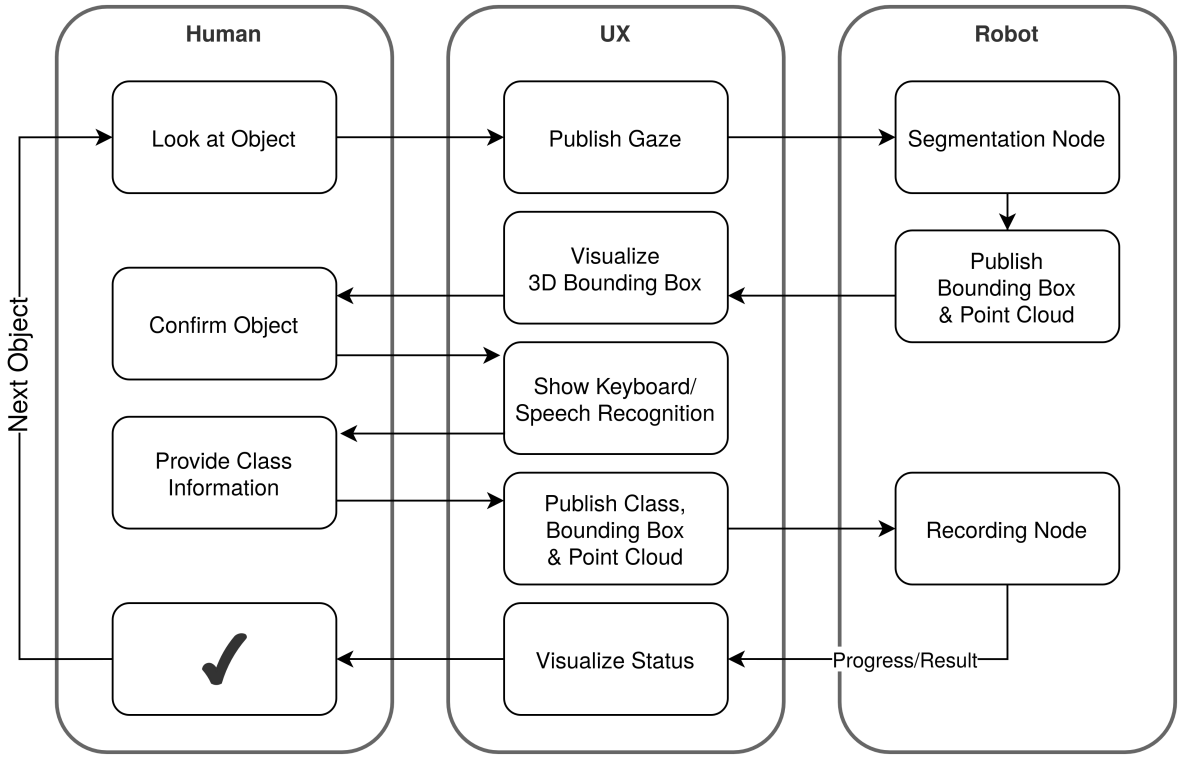


图 1：整个教学流程概览。增强现实用户界面（UX）充当人类与机器人之间的桥梁。

3 方法

本研究的目标是教会机器人识别其环境中的未知物体，使机器人之后能够在同一环境中检测这些物体。为此，下面将说明：1）我们如何通过视觉、视线、语音和手势等模态与机器人进行交流；2）机器人如何识别感兴趣对象；3）人类如何教会机器人；以及4）机器人最终如何实现学习。

3.1 增强现实界面

为了让人类能够教机器人任何东西，通信渠道必不可少。正如 [50], 所提出的，我们通过增强现实界面来满足这一需求。人类与机器人之间的所有通信都通过该界面进行，也就是说，可以控制机器人、传输人类的注视信息，并传达物体各自的类别信息。在人类一侧，我们使用 HoloLens 2——这是一款由微软制造、内置眼动追踪器的头戴式增强现实眼镜。我们使用游戏引擎和实时开发平台 Unity 2019.4.36 版开发了该界面，即 HoloLens 应用程序。此外，我们还使用了微软专为此用途提供的 Microsoft Reality Toolkit（MRTK）2.7.2 版资源。HoloLens 上的通用 Windows 平台（UWP）应用与机器人操作系统 ROS [33],之间的数据互连通过 ROS# [1]实现。开源软件库 ROS# 在 Unity 应用程序中通过 rosbridge_套件与 ROS 交换基于 JSON 的命令。HoloLens 和机器人通过 WiFi 持续相互连接，因而可以实时发送和接收数据，例如人类的注视信息。最后，手势和语音用于操作增强现实界面并与机器人交互。图 1 展示了增强现实界面如何充当人类与机器人之间的桥梁，并呈现了教学流程各个组成部分之间的协同关系；我们将在下文中进一步详细介绍这些内容。

3.2 识别感兴趣的未知物体

为了让机器人学会一种新物体，它首先必须能够将其识别为物体。这是一个相当根本的问题，因为从某种意义上说，这就像先有鸡还是先有蛋的问题。对机器人而言，检测感兴趣对象很困难，因为它此时还不认识该对象，也尚未接受相关教学。因此，由于机器人必须先识别物体才能学会它，此时无法部署神经网络，而且确定物体从哪里开始、到哪里结束也并非易事。为此，我们希望结合人类的注视信息，帮助机器人定位目标对象。这意味着人类注视物体后，机器人便能将其与环境中的其他部分区分开来。为此，我们以 [50]的方法为基础，该方法利用人类注视以及从机器人场景相机的深度传感器获取的点云，对观察到的物体进行分割。通过标定确定机器人与 HoloLens 各自的位置，并由 HoloLens 的运动传感器在人类移动期间持续追踪二者的相对位置。此外，HoloLens 会持续追踪注视点，即人类正在注视的位置，并通过我们的用户界面使用 ROS# 将其作为三维空间中的一个点发布。因此，机器人的 ROS 系统可以订阅相应的 ROS话题，这意味着机器人始终能够获知注视点，并将其用于分割。在 [50] 所述分割流程的第一步，应用直通滤波器和体素网格滤波器，以缩减点云规模。随后，使用随机抽样一致提取地面，最终通过注视点和欧几里得聚类分离出物体。我们对该方法的最后一步作了少量调整。我们不再只将距离注视点最近的簇归为物体，而是考虑一定距离范围内且具有一定大小的所有簇。我们将簇与注视点之间的最大距离阈值设为 2 厘米，并将簇的最小大小设为五个点。这样，即使是没有明显高出地面的扁平物体，也能够被分割。根据人当前注视的物体，相应物体会被实时分割。上述所有点云处理均使用开源点云库（PCL） [37]完成。

由于在开始时进行了标定，因此无论从佩戴 HoloLens 的人类的位置，还是从机器人的位置，都能确定感兴趣对象的位置。前者使机器人能够通过订阅器，将有关分割物体的反馈以三维边界框的形式显示在 HoloLens 上，也就是显示在人类的视野中。该订阅器附加在虚拟边界框上，并根据分割结果更新边界框的位置。我们可以在下文所述的教学过程中利用后者。

3.3 通过联合注意进行教学

现在，机器人和人类的注意力共同指向同一个物体。下一步是向机器人确认框选的物体就是感兴趣对象，并提供类别信息。人类用双眼注视该物体时，在 HoloLens 的视野内用食指和拇指做出捏合手势，即可通过增强现实界面选择该物体。或者，也可以直接说“选择”。随后，人类的视野中会出现一个虚拟键盘，供其输入该物体的类别名称。同样，也可以直接使用语音输入。图 2 展示了从人类视角看到的已实现键盘。

我们的下一个目标是让机器人自主捕获感兴趣对象的图像，随后将这些图像用作训练数据。为了尽可能从多个角度获取图像，我们在机器人机械臂的腕部安装了第二台相机。现在，机器人需要让这台相机绕物体做圆周运动。这意味着，当人类通过 ROS动作将类别名称传输给机器人后，机器人会计算一条由可达点组成的圆形轨迹。由于机械臂长度等物理限制，这通常只是圆形轨迹中的部分圆弧。在运动过程中，相机会以这样的方式对准：其视线与先前确定的三维边界框中心呈45度角。对于中心与相机之间的距离，我们采用边界框对角线长度的两倍，同时确保该距离至少为相机与机器人机械臂末端执行器之间距离的两倍。记录过程如图3所示。

借助上一节所述的标定，机器人不仅能够获知感兴趣对象的位置，还能够计算其所有

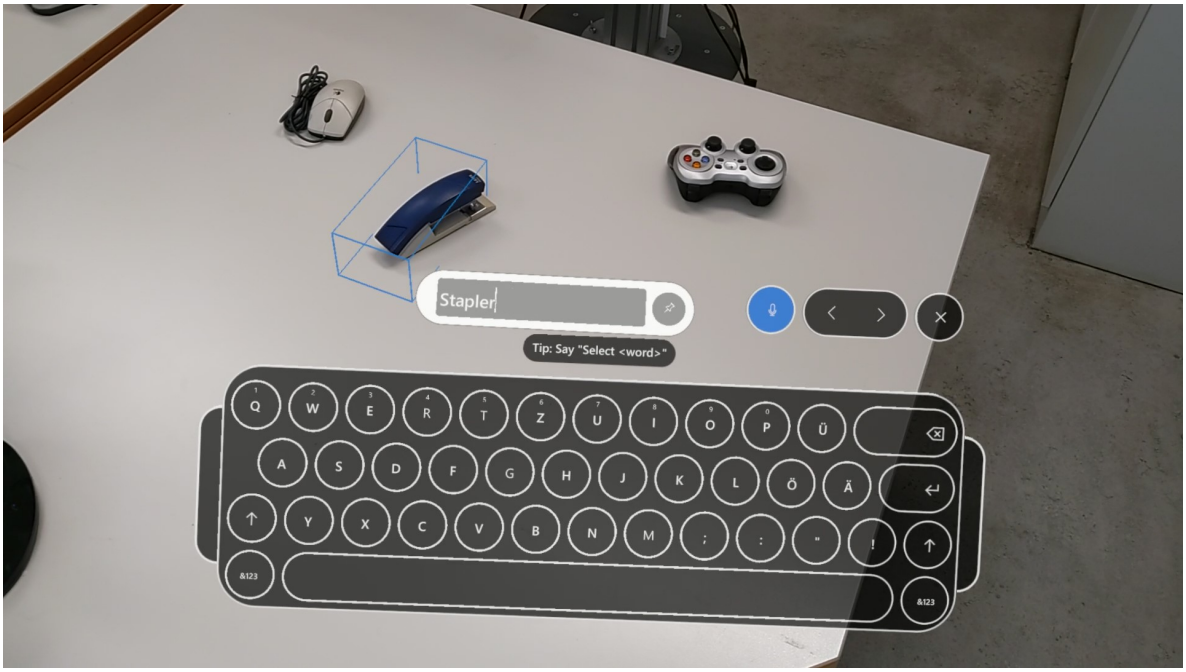


图 2：在人类的视野中，显示了机器人分割出的物体边界框。选中后，可以使用虚拟键盘或语音指定相应物体的类别名称。

坐标系，即机器人坐标系之间的变换。这其中包括安装在机器人机械臂上的相机。在此步骤中，关键是持续将分割物体的点云变换到相机的坐标系中。通过将点云投影到相机的二维图像平面上，我们可以根据边界点得出感兴趣区域。因此，对于每幅采集的图像，我们还可以存储以此方式计算出的二维边界框。总体而言，存储的同步数据包括 RGB 图像和深度图像，以及带有二维边界框的感兴趣区域。此外，我们还会存储相机相对于所捕获物体的位置。最终，机器人能够在极短时间内自动生成数百张带标签的训练图像。更确切地说，训练一个物体大约需要一分钟，并可生成约 300 张图像。记录过程的进度和完成情况随后会通过 ROS动作传输至 HoloLens，以向人类显示机器人何时准备好处理新的物体。

3.4 迁移学习

在教学部分之后，机器人便进入学习过程。为了使机器人今后能够自主检测此前见过的物体，它必须利用手头的信息，也就是它自行创建的训练数据。换言之，机器人，或者更确切地说，其基于神经网络的物体检测器，将使用获取的 RGB图像进行训练。这将通过迁移学习实现。因此，我们假设机器人对目标性已有一定了解，因为我们的方法应被视为对使用常见大型数据集（如 ImageNet [6], PASCAL VOC [9]或 MS COCO [23]进行训练的补充，而非替代。这种做法是合理的，因为大多数现有物体并不属于这些数据集。我们旨在通过所提出的方法扩展现有的知识。因此，我们采用 Faster R-CNN [36] 和 FCOS [42]等最先进的目标检测器。分别以这些预训练模型之一为基础，我们删除最后的分类层，然后根据应用场景所需的输出神经元数量重新初始化这些层。最后，使用我们的数据重新训练这些层，同时冻结前面特征层中的所有其他神经元。通过冻结负责理解物体一般特征的特征层，仅微调最后几层，我们可以防止过拟合 [53, 55]。事实上，由于我们只重新训练模型的分类头，因此无需依赖高性能图形处理器，也可以直接在机器人上进行训练。当然，训练可能会耗时更久。在第5节中，我们将评估这些

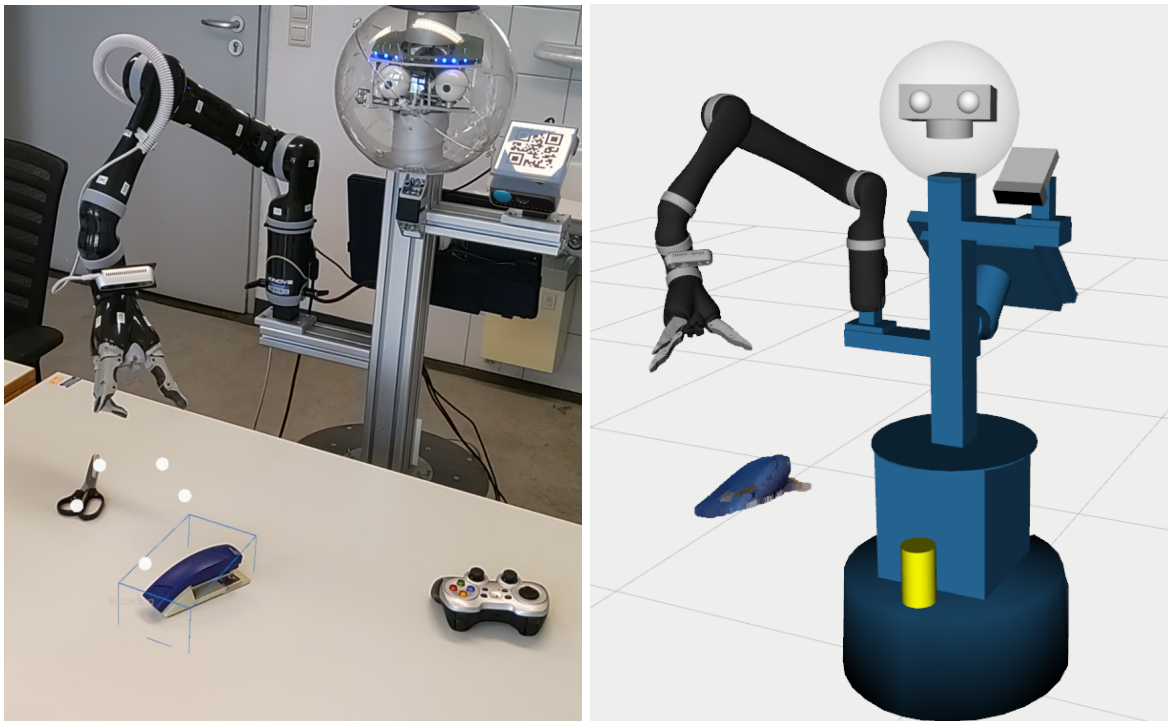


图3：左侧展示了人类增强现实视角下的记录过程，右侧则是在 RVIZ 中对分割物体点云的可视化。点云既用于向人类显示边界框，也用于标注由机械臂腕部相机拍摄的图像。

模型以及其他模型的性能。

4 数据集：多视角物体细节

上一节介绍的方法使我们能够创建一种新型数据集。我们发布验证集和测试集，主要是为了确保研究结果具有可复现性；而我们认为，训练集或许对进一步研究尤其有价值。接下来，我们将更详细地介绍训练数据，并提供一些统计信息。至于验证集和测试集，我们会在第5节的评估部分进行详细说明。

我们的训练集有一个显著特点：它提供了有关单个物体的大量细节。大多数物体检测数据集通常包含许多展示不同物体的图像，而我们的数据则从许多不同角度呈现物体。对于那些正面和背面外观不同的物体，例如游戏手柄，这种数据尤其有价值。

该数据集总共包含3113个视角，涵盖叉子、飞盘、游戏手柄、打孔器、刀、剪刀、羽毛球、订书机、乒乓球和牙刷等类别。每个类别在数据集中都有两个不同的实体，它们在颜色、形状或两者上有所差异。图4展示了一些样本图像，图5则展示了视点的分布。每个物体都单独放置在桌面上，并且针对每个相机视角都包含多项信息。对于每幅RGB图像，除感兴趣区域外，还提供了一幅与颜色图像对齐的相应深度图像。所有RGB图像和深度图像的分辨率均为 1920×1080。此外，还提供了所有相机位姿。这些位姿以与机器人无关的方式表示为一种变换，由平移向量和旋转四元数组成，描述从相机坐标系到物体坐标系的变换。最后一项是相机的元信息，其中包含相机标定的内参。

总而言之，我们认为，我们数据中的高信息密度对于其他需要精确相机位置的研究领域也很有价值，例如神经辐射场 [27, 54, 21, 7]。在这些领域中，要么必须使用合成数据，要么在使用真实数据时，只能通过运动恢复结构对相机位置（和深度）进行

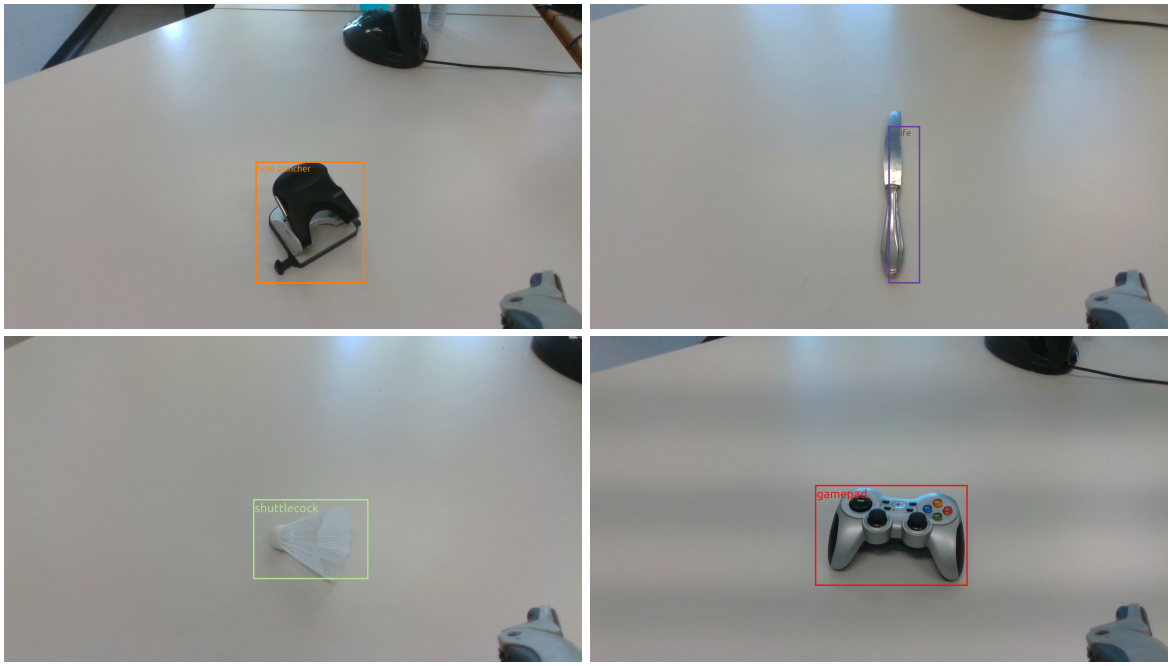


图4：数据集中打孔器、刀、羽毛球和游戏手柄的示例图像。边界框的质量可能因视角而异，有时会略大、略小或发生偏移。然而，在所有图像中，边界框的大部分始终覆盖相应的物体。这些物体在平坦度和颜色方面与背景形成了不同程度的对比。

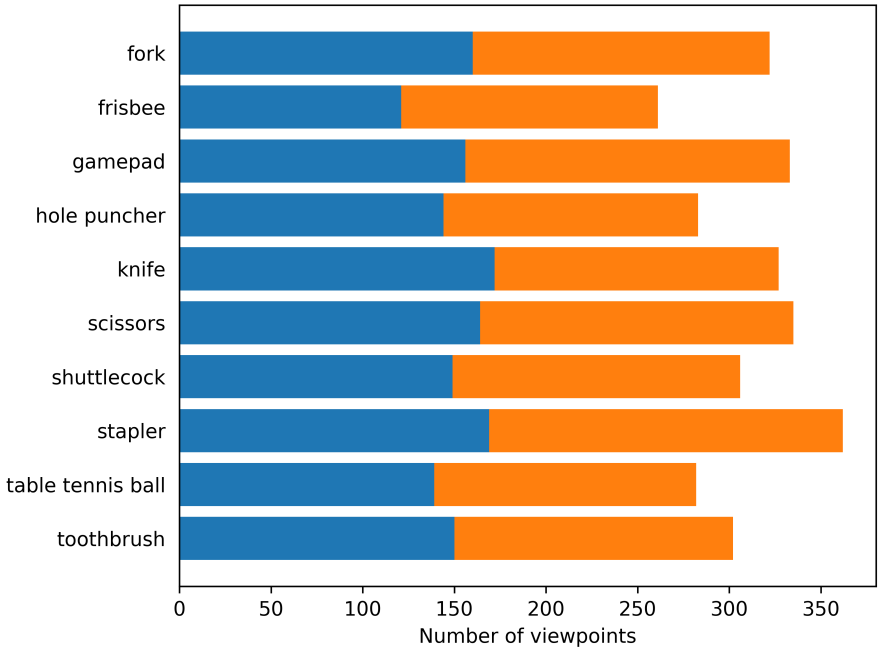


图5：各类别中视点的分布。颜色表示类别中的两个不同物品。

粗略估算。因此，我们将数据集 Objects in Multiperspective Detail（多视角细节物体（OMD））公开给研究界。

5 评估

在我们的所有实验中，除人类佩戴的前述 HoloLens 2 外，我们还使用了 MetraLabs 的 Scitos G5 [26] 作为机器人。机器人通过其观察场景并执行分割的机身相机，是微软的 Azure Kinect DK。安装在 Scitos 上的机械臂是 Kinova Jaco2 [15] ，具有 6 个自由度。为拍摄物体图像而连接在机械臂腕部的相机，是 Intel RealSense D435。

训练时，我们使用自己的训练集，该训练集已在第 4 节中进行了详细说明。由于本工作的目标是教会机器人了解其环境，因此我们还必须记录一个验证集和测试集，这些数据集位于学习过程实际进行的环境中。如前所述，除训练集外，我们也将发布验证集和测试集，以确保结果的可复现性。验证集和测试集分别包含 1051 张和 1410 张常规图像，这些图像均使用 DarkLabel [4]手动标注。其所表示的类别与训练集中的十个类别相同。每个类别均有四个不同的相应类别物体可供使用。验证集包含训练集中每个类别的相同两个物体，而测试集包含另外两个物体。也就是说，这些物体在形状、颜色或二者方面均不同于训练时使用的物体。物体在机器人办公室环境中以随机分组的方式（在各自数据集内）进行拍摄。我们还确保构建了具有挑战性的场景，例如物品堆叠在一起，或牙刷仍处于包装内，如图6所示。

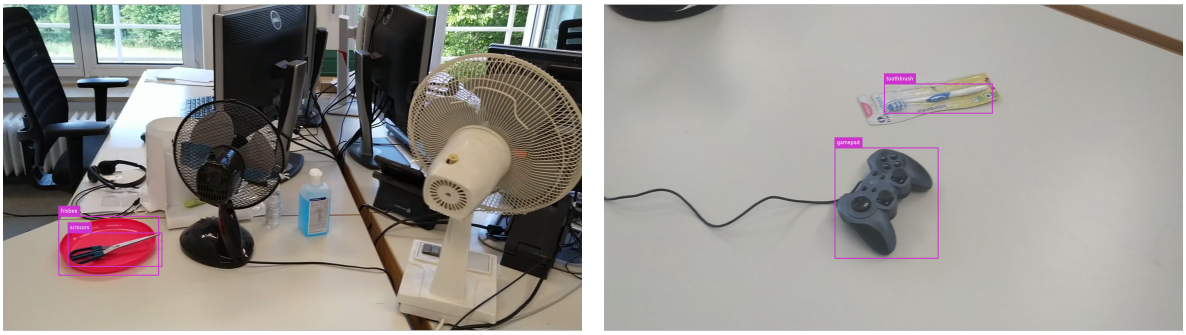


图6：来自测试集的示例图像。物体集合与训练集中的物体互不重叠（参见游戏手柄）。此外，这些图像在背景杂乱程度和物体距离方面也各不相同。

接下来，我们使用多种最先进的目标检测器评估我们的学习流程。因此，我们以 Faster R-CNN [36], RetinaNet [22], FCOS [42], 和 SSD300 [24]作为基础，并搭配多种骨干网络，例如 ResNet-50-FPN [11],VGG16 [40] 和 MobileNetV3 Large [38, 12]。所有骨干网络均在 ImageNet 上训练过，并且可以保持原样，因为我们特意为评估挑选了与该数据集完全没有交集的物体类别。我们选择这十种测试物体的原因如下：一方面，考虑到骨干网络，它们不能出现在 ImageNet 中；另一方面，其中至少一部分应包含在 MS COCO 中，以便我们稍后进行比较。此外，每个类别中都必须有若干外观各异的物体。综合这些条件，选择范围因此受到限制，尤其是我们还试图避免食物等易腐类别。因此，就实际使用的目标检测器而言，为确保这些物体对模型确实是未知的，我们必须在 MS COCO 的一个子集上训练模型。具体来说，我们使用 Fiftyone [29] 工具从数据集中提取了叉子、飞盘、刀、剪刀、运动球和牙刷这些类别，然后在剩余部分上训练上述检测器。为此，我们遵循了原始实现各自的训练方案，并为保持一致，在后续所有使用自有训练流程进行的实验中也都遵照该方案。迁移学习方法中唯一的例外是所采用的数据增强类型。在这种情况下，我们对所有模型（不只是 SSD300）都使用了随机光度失真、随机缩小、随机裁剪和随机水平翻转，以防止过拟合。随后，我们按照第 3 节所述的方法进行训练。

在我们的所有实验中，我们均按照 MS COCO 指标 [23], 进行评估，即针对不同交并比阈值（IoU）计算平均精度。在此语境下，我们使用缩写 AP $=\mathrm{AP}^{\mathrm{IoU}=0.5:0.05:0.95}$、$\mathrm{AP}_{50}=\mathrm{AP}^{\mathrm{IoU}=0.5}$ 和 $\mathrm{AP}_{75}=\mathrm{AP}^{\mathrm{IoU}=0.75}$ 表示单个类别内的平均精度，并以 mAP 表示

表1：以迁移学习方式训练的所有机器学习模型的比较。最佳值以粗体标出。

模型	骨干网络	mAP	mAP_{50}	mAP_{75}	mAR_1	mAR_{10}	mAR_{100}
Faster R-CNN [36]	ResNet-50 [11]	**33.6**	**66.9**	31.4	43.7	50.1	50.4
Faster R-CNN [36]	MobileNetV3 [38, 12]	15.5	38.1	6.1	23.7	27.4	27.7
Faster R-CNN [36]	MobileNetV3 [38, 12]†	13.0	38.4	3.1	22.2	25.5	25.5
FCOS [42]	ResNet-50 [11]	30.6	47.6	**35.9**	44.7	53.8	55.0
RetinaNet [22]	ResNet-50 [11]	31.2	52.4	34.3	**46.2**	**57.6**	**59.1**
SSD300 [24]	VGG16 [40]	8.0	19.1	5.0	21.0	31.6	34.0

†针对移动端使用场景进行了调优

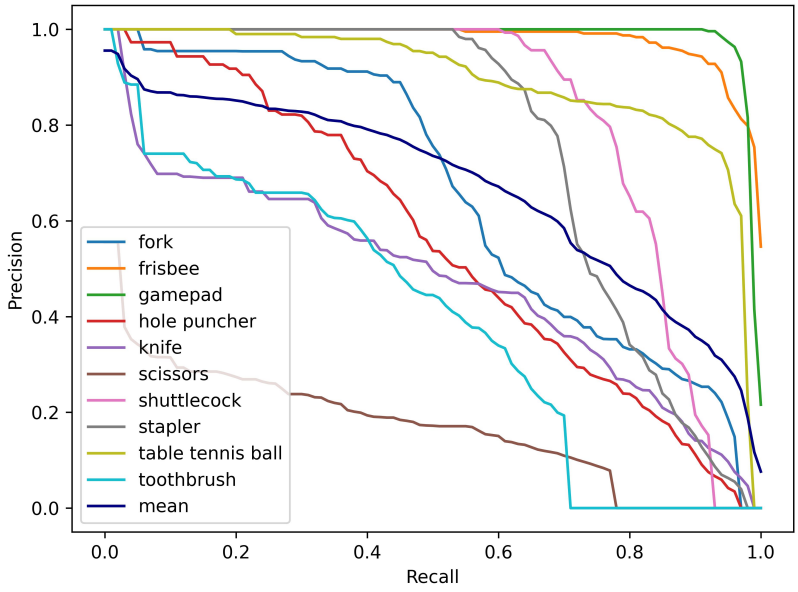


图7：交并比为0.5时 Faster R-CNN 的精确率-召回率曲线。超过此值时，物体可视为已检测到 [9, 56]。

所有类别的平均值。类似地，对于平均召回率，我们分别考察每张图像最多检测 1、10 和 100 个目标时的召回率，并使用缩写 $AR_1 = AR^{max=1}$、$AR_{10} = AR^{max=10}$ 和 $AR_{100} = AR^{max=100}$。同样，mAR 表示所有类别的平均值。除非另有说明，平均精度和平均召回率分别指 AP 和 AR。

所有受测模型的比较结果见表1。采用 ResNet-50 骨干网络的 Faster R-CNN 在平均精度方面总体表现最佳。采用 MobileNetV3 骨干网络时，其精确率和召回率均显著较差。FCOS 和 RetinaNet 的平均准确率略低于 Faster R-CNN。后者的召回率最高，而当交并比为 0.75 时，FCOS 的平均精度最高。SSD300 的精确率明显落后于所有其他模型。

接下来，我们将继续使用 Faster R-CNN 进行进一步分析，因为 MS COCO [23]将平均精度均值视为最重要的单一指标。总体而言，模型似乎能够很好地检测物体，但某些类别比其他类别更难检测。从图 7 的曲线中也可以明显看出这一点。即使在召回率较低时，剪刀的精确率也低于 0。5。牙刷类别的精确率也会较早下降。另一方面，游戏手柄和飞盘类别的精确率始终非常出色，即使召回率很高时也是如此。

在表2中，我们比较了不同的训练变体。暂且不考虑基线变体（COCO），我们发现，仅训练最后几层的迁移学习方法（TL-F）在大多数类别上取得了最佳平均精度。尤其是，mAP_{75} 显著更高。

值得一提的是，基线（COCO）的数值自然更高。完整的 MS COCO 训练集与我们用于预训练的部分之间，仍相差 25 713 个物体和 14 296 张图像；前者的图像数量是我们所用图像的五倍。此外，我们的图像涵盖全部十个类别，而 MS COCO 不包含其中四类，因此基线完全无法识别它们。这展示了我们流程的优势：该流程旨在支持学习额外的、迄今未知的类别，也就是扩展现有知识。与基线的比较代表一种理想情况：1）存在合适的数据集；2）该数据集可供使用；3）物体属于该数据集。此比较主要是为了更好地将我们的结果置于整体图景中，而非旨在超越基于如此庞大数据集训练的模型；相反，它用于确定一个上界，并检验我们的方法能达到多接近这一上界。考虑到这一点，我们的流程对某些类别的学习效果之好令人瞩目，尤其是游戏手柄和羽毛球，其平均精度甚至高于基线的平均精度均值。此外，由于乒乓球在 MS COCO 中仅作为运动球类别的一个子集出现，我们可以看到系统如何变得更加关注乒乓球。相比之下，飞盘类别的表现略逊于基线，因为 MS COCO 中对应的类别只包含飞盘。对于打孔器类别，即使没有预先的基础知识（S），结果也令人满意。

考虑到各方案所需的训练时间，差异十分明显。完整的 MS COCO 训练集耗时最长：在两块并行部署的 NVIDIA RTX A4000 上训练超过两天。其余三种方案均使用我们自己的数据集，并且仅在一块图形处理器上训练；完整模型（S、TL-U）耗时 4 小时，而冻结变体（TL-F）耗时 2.5 小时。如上所述，尽管为了保持一致，我们按照原始方案训练了 26 个训练轮次，但最终用于测试的最佳验证准确率对应的权重往往更早就已得到。通过 TL-F 训练的 Faster R-CNN 甚至在三个训练轮次后就达到了这一状态（训练前的平均精度均值为 $= 0.0$ ），在我们配备 NVIDIA GeForce GTX 1050 Ti 的 Scitos 上，训练时间约为 40 分钟。因此，整个流程也适用于直接在机器人上进行独立学习。

最后，我们分析了用于训练的图像数量所产生的影响。表3列出了使用 TL-F 训练的 Faster R-CNN 在不同数据集大小下的结果。图像按照等间距从视角序列中移除。除使用 75% 数据的情况外，结果呈现出一种趋势：随着图像数量增加，平均精度和平均召回率也随之提高。使用全部数据时取得了最佳结果。

表2：我们测试集上 Faster R-CNN 不同训练类型的平均精度（AP）比较。具体而言，除骨干网络外，分别从头开始训练（S）、采用冻结非分类层的迁移学习方式训练（TL-F），或完全解冻后训练（TL-U）。这三种方式均使用机器人采集的数据。最佳结果以粗体标出。最后一列（COCO）供参考，报告了在整个 MS COCO 训练集上训练的 Faster R-CNN 的结果。

类别	S	TL-U	TL-F	COCO
fork	**21.7**	**21.7**	18.6	63.8
飞盘	19.5	47.6	**58.8**	65.9
游戏手柄	38.4	24.8	**62.6**	-
打孔器	**42.5**	26.3	23.0	-
刀	20.4	19.1	**27.6**	50.0
剪刀	6.5	**7.3**	5.1	70.9
羽毛球	24.9	27.8	**51.5**	-
订书机	29.3	24.8	**38.9**	-
乒乓球	3.4	27.6	**44.0**	17.3
牙刷	**21.6**	8.8	6.2	37.4
mAP_{50}	62.8	**68.8**	66.9	84.4
平均精度均值$_{75}$	9.8	7.5	**31.4**	55.4
mAP	22.8	23.6	**33.6**	50.9

表3：通过 TL-F 在我们数据集的不同规模子集上训练的 Faster R-CNN 的结果。最佳数值以粗体标出。

	25%	50%	75%	100%
mAP	31.5	32.9	31.7	**33.6**
mAP_{50}	64.7	66.7	66.7	**66.9**
平均精度均值$_{75}$	28.6	28.4	26.0	**31.4**
mAR_1	41.1	42.9	41.4	**43.7**
mAR_{10}	46.9	49.9	46.9	**50.1**
mAR_{100}	47.2	50.2	47.1	**50.4**

6 局限性与讨论

与所有监督式机器学习方法类似，我们依赖训练数据的质量。在我们的案例中，这一质量会因点云预先进行的分割而有所不同，而分割质量自然取决于机器人场景相机的深度传感器所获取数据的质量。我们注意到，深度传感器在处理非常暗或有光泽的表面时，难以准确测定深度。因此，边界框的准确性会受到影响，最终影响性能。不过，增加物体数量可以弥补这一问题；此外，我们的测试表明，尽管存在这些困难，机器人仍能够检测其环境中已学习的物体。

另一个需要指出的问题是，尽管我们的方法对已学习类别中的其他物体也具有良好的泛化能力，但在 MS COCO 等常用数据集上的测试效果较差。这是由于图像的多样性以及其中呈现的各种情境。例如，香蕉和苹果等水果既可能以天然形态出现，也可能被切成小块放在水果沙拉中。要解决这项任务，需要大量训练数据。另一方面，尽管我们的方法无法达到在大型数据集上训练所能取得的性能，但其主要目的是教会机器人识别尚无现成数据的物体。在这类情境中，语义场景理解必不可少，而人类可以通过我们的方法协助获得这种理解。虽然扩充现有数据集也是一种可行方案，但与之相比，我们的方法无需大量标注资源，并且可以在相应情境中即时使用。我们的评估（表2）表明，我们的方法能够学习基线方法未能检测到的未知物体。因此，无需预先了解情境，也不必依赖适用数据集的存在，就能更灵活地使用该方法。此外，通过双向交互进行教学非常自然，尤其是增强现实系统能够直接连接两个世界——人类所处的模拟世界和机器人所处的数字世界——从而实现人类与机器人之间的实时通信。人类可以向机器人提供信息（视线、类别），由此启动记录过程；机器人则可以通过视觉方式反馈信息。通过视线指向物体，进一步完善了这种协作方式：这种方式直观、比手指指向更少歧义，而且不同于语音，在机器人尚不认识物体之前就可以使用。总而言之，我们认为，我们的方法与其说是替代方案，不如说是补充方案。

7 结论

在本研究中，我们提出了一种新颖的流程，旨在将机器人部署于非预设场景中。为此，我们利用机器人与人类交互中的人类注视和增强现实，成功地教会机器人识别其环境中的新物体（此前未知的物体）。

为了让配备基于机器学习的物体检测器的机器人能够检测其环境及其中的物体，通常需要大量训练数据。然而，在实际应用中，受不可预测的条件和现有物体种类繁多的影响，并不总能保证获得合适的数据集。在这类情况下，我们的方法可以补充常用数据集，以简便易用的方式在短时间内生成大量自动标注的非合成训练数据。基于此类数据，我们通过多种不同方式训练了最先进的目标检测器，并证明了以这种方式学习和检测新物体是可行的。事实上，得益于迁移学习，训练甚至可以独立在机器人上进行

无需耗费大量计算资源。此外，只需少量样本，机器人就能将所学泛化到给定类别中的未见物体，并检测出由于底层训练数据集不适用而无法被基线方法检测的类别。这使我们的教学流程成为仅使用标准数据集进行训练的有益补充。总体而言，得益于其多模态性（包括增强现实以及人类与机器人之间的共享注视），我们的方法极其自然直观。我们在评估过程中记录的数据集也已公开，其信息密度很高，因为其中包含了针对相应物体的多种不同视角，以及在此过程中收集的数据。因此，除我们的研究用途外，该数据集还有望适用于多种其他用途。

不过，未来仍有大量工作需要完成：我们计划通过用户研究调查系统的可用性，并扩展我们的方法，使机器人能够成功检测其训练环境之外的物体，同时通过主动学习，在另一种人机交互场景中进一步利用所获取的知识。

致谢

本研究由德国研究联合会（DFG，德国研究基金会）通过德国卓越战略资助，项目编号为 EXC 2064/1，项目编号 390727645。我们还感谢 Julia Dietl 在标注工作中付出的宝贵努力。

参考文献

[1] Martin Bischoff。ROS#，2019年6月。

[2] Robert Bogue。机器人在零售业的广阔前景。《*Industrial Robot*》：机器人研究与应用国际期刊，2019年。[3] Jeannette Bohg、Antonio Morales、Tamim Asfour 和 Danica Kragic。数据驱动的抓取合成——综述。《*IEEE*机器人学汇刊》，30(2):289–309，2013年。[4] DarkLabel。 `https://github.com/darkpgmr/DarkLabel`，2021年。访问日期：2021-12-07。[5] Masood Dehghan、Zichen Zhang、Mennatullah Siam、Jun Jin、Laura Petrich 和 Martin Jagersand。通过人机交互进行在线物体与任务学习。载于*2019*年国际机器人与自动化会议（*ICRA*），第2132–2138页。IEEE，2019年。[6] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：大型分层图像数据库。载于*IEEE*计算机视觉与模式识别会议（*CVPR*）论文集，第248–255页。IEEE，2009年。[7] Kangle Deng、Andrew Liu、Jun-Yan Zhu 和 Deva Ramanan。深度监督神经辐射场：以更少视图实现更快训练。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第12882–12891页，2022年。[8] Lotfi El Hafi、Hitoshi Nakamura、Akira Taniguchi、Yoshinobu Hagiwara 和 Tadahiro Taniguchi。通过混合现实中的人机交互实现多模态物体分类的教学系统。载于*2021*年*IEEE/SICE*系统集成国际研讨会（*SII*），第320–324页。IEEE，2021年。[9] Mark Everingham、Luc Van Gool、Christopher KI Williams、John Winn 和 Andrew Zisserman。PASCAL视觉物体类别（VOC）挑战赛。《国际计算机视觉杂志》，88(2):303–338，2010年。[10] Guglielmo Gemignani、Emanuele Bastianelli 和 Daniele Nardi。通过语音交互向机器人传授参数化可执行计划。载于*2015*年自主智能体与多智能体系统国际会议论文集，第851–859页。自主智能体与多智能体系统国际基金会，2015年。

[11] Kaiming He、张祥雨、Shaoqing Ren 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。[12] Andrew Howard、Mark Sandler、Grace Chu、Liang-Chieh Chen、Bo Chen、Mingxing Tan、Weijun Wang、Yukun Zhu、Ruoming Pang、Vijay Vasudevan 等。探索 MobileNetV3。载于*IEEE/CVF*国际计算机视觉会议论文集，第1314–1324页，2019年。[13] Raihan Kabir、Yutaka Watanobe、Md Rashedul Islam、Keitaro Naruse 和 Md Mostafizer Rahman。面向云机器人系统的基于单类支持向量机的未知物体检测。*Sensors*，22(4):1352，2022年。[14] Hakan Karaoguz 和 Patric Jensfelt。用于机器人抓取检测的物体检测方法。载于*2019*年国际机器人与自动化会议（*ICRA*），第4953–4959页。IEEE，2019年。[15] Kinova。 https://www.kinovarobotics.com/product/gen2-robots，2022年。访问日期：2022年8月14日。[16] Kyle Krafka、Aditya Khosla、Petr Kellnhofer、Harini Kannan、Suchendra Bhandarkar、Wojciech Matusik 和 Antonio Torralba。人人可用的眼动追踪。载于*IEEE*计算机视觉与模式识别会议论文集，第2176–2184页，2016年。[17] Evan Krause、Michael Zillich、Thomas Williams 和 Matthias Scheutz。通过自然语言对话中的人机交互，单次学习识别新颖物体。载于*AAAI*人工智能会议论文集，第28卷，2014年。[18] Dennis Krupke、Frank Steinicke、Paul Lubos、Yannick Jonetzko、Michael Görner 和张建伟。共址混合现实人机交互中多模态转向与指向手势的比较。载于*2018*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第1–9页。IEEE，2018年。[19] 李毅和王崇立。顾客感知对服务机器人接受度的影响。*InternationalJournal of Consumer Studies*，46(4):1241–1261，2022年。[20] 李一萌和 Jana Košecká。面向未知物体检测的感知不确定性提议分割。载于*IEEE/CVF*冬季计算机视觉应用会议论文集，第241–250页，2022年。[21] Chen-Hsuan Lin、Wei-Chiu Ma、Antonio Torralba 和 Si mon Lucey。BARF：神经辐射场的束调整。载于*IEEE/CVF*国际计算机视觉会议论文集，第5741–5751页，2021年。[22] Tsung-Yi Lin、Priya Goyal、Ross Girshick、Kaiming He 和 Piotr Dollár。用于密集目标检测的焦点损失。载于*IEEE*国际计算机视觉会议论文集，第2980–2988页，2017年。[23] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。Microsoft COCO：常见物体上下文。载于欧洲计算机视觉会议，第740–755页。Springer，2014年。[24] 刘伟、Dragomir Anguelov、Dumitru Erhan、Christian Szegedy、Scott Reed、Cheng-Yang Fu 和 Alexander C Berg。SSD：单次多框检测器。载于欧洲计算机视觉会议，第21–37页。Springer，2016年。[25] Zhanat Makhataeva 和 Huseyin Atakan Varol。机器人学中的增强现实：综述。机器人学，9(2):21，2020年。[26] MetraLabs。 https://www.metralabs.com/mobiler-roboter-scitos-g5/，2022年。访问日期：2022年8月14日。[27] Ben Mildenhall、Pratul P Srinivasan、Matthew Tancik、Jonathan T Barron、Ravi Ramamoorthi 和 Ren Ng。NeRF：将场景表示为用于视图合成的神经辐射场。载于欧洲计算机视觉会议，第405–421页。Springer，2020年。

[28] Ajay Mishra、Yiannis Aloimonos 和 Cheong Loong Fah。基于注视的主动分割。载于*2009*年*IEEE*第*12*届国际计算机视觉会议，第468–475页。IEEE，2009年。[29] B. E. Moore 和 J. J. Corso。FiftyOne。*GitHub*。注：*https://github.com/voxel51/fiftyone*，2020年。[30] Dim P Papadopoulos、Alasdair DF Clarke、Frank Keller 和 Vittorio Ferrari。利用眼动追踪数据训练目标类别检测器。载于欧洲计算机视觉会议，第361–376页。Springer，2014年。[31] Thies Pfeiffer 和 Patrick Renner。Eyesee3D：一种利用计算机视觉和增强现实技术分析移动式三维眼动追踪数据的低成本方法。载于眼动追踪研究与应用研讨会论文集，第195–202页，2014年。[32] Long Qian、Jie Ying Wu、Simon P DiMaio、Nassir Navab 和 Peter Kazanzides。机器人辅助手术中的增强现实综述。*IEEE*医学机器人与仿生学汇刊，2(1):1–16，2019年。[33] Morgan Quigley、Ken Conley、Brian Gerkey、Josh Faust、Tully Foote、Jeremy Leibs、Rob Wheeler 和 Andrew Y Ng。ROS：一个开源机器人操作系统。载于*ICRA*开源软件研讨会，第3卷，第5页。日本神户，2009年。[34] Camilo Perez Quintero、Sarah Li、Matthew KXJ Pan、Wesley P Chan、HF Machiel Van der Loos 和 Elizabeth Croft。通过增强现实中的增强轨迹进行机器人编程。载于*2018*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第1838–1844页。IEEE，2018年。[35] Andreagiovanni Reina、Alex J Cope、Eleftherios Nikolaidis、James AR Marshall 和 Chelsea Sabo。Ark：面向千机器人的增强现实。*IEEE*机器人与自动化快报，2(3):1755–1761，2017年。[36] Shaoqing Ren、Kaiming He、Ross Girshick 和 Jian Sun。Faster R-CNN：利用区域提议网络实现实时物体检测。神经信息处理系统进展，28:91–99，2015年。[37] Radu Bogdan Rusu 和 Steve Cousins。三维时代已来：点云库（PCL）。载于*2011*年*IEEE*机器人与自动化国际会议，第1–4页。IEEE，2011年。[38] Mark Sandler、Andrew Howard、Menglong Zhu、Andrey Zhmoginov 和 Liang-Chieh Chen。MobileNetV2：倒残差与线性瓶颈。载于 *IEEE* 计算机视觉与模式识别会议论文集，第4510–4520页，2018年。[39] Florian Alexander Schmidt。人工智能训练数据的众包生产：人工劳动者如何教会自动驾驶汽车如何“看”。技术报告，Working Paper Forschungsförderung，2019年。[40] Karen Simonyan 和 Andrew Zisserman。用于大规模图像识别的超深卷积网络。*arXiv*预印本 *arXiv:1409.1556*，2014年。[41] Sebastian Thrun、Maren Bennewitz、Wolfram Burgard、Armin B Cremers、Frank Dellaert、Dieter Fox、Dirk Hahnel、Charles Rosenberg、Nicholas Roy、Jamieson Schulte 等。Minerva：第二代博物馆导览机器人。载于*1999*年*IEEE*机器人与自动化国际会议论文集（目录号*99CH36288C*），第3卷。IEEE，1999年。[42] Zhi Tian、Chunhua Shen、Hao Chen 和 Tong He。FCOS：全卷积单阶段物体检测。载于*IEEE*国际计算机视觉会议（*ICCV*），2019年。[43] Takumi Toyama、Thomas Kieninger、Faisal Shafait 和 Andreas Dengel。使用头戴式眼动仪进行注视引导的物体识别。载于*2012*年*ACM*眼动追踪研究与应用研讨会论文集，第91–98页。ACM，2012年。[44] Gianluca Vadalà、Sergio De Salvatore、Luca Ambrosio、Fabrizio Russo、Rocco Papalia 和 Vincenzo Denaro。机器人脊柱手术与增强现实系统：最新进展。*Neurospine*，17(1):88，2020年。

[45] Sepehr Valipour、Camilo Perez 和 Martin Jagersand。通过人机交互进行机器人感知的增量学习。载于*2017*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第2772–2777页。IEEE，2017年。[46] Anna-Maria Velentza、Dietmar Heinke 和 Jeremy Wyatt。通过协作导览机器人实现人机交互并增进知识。载于*2019*年第*28*届*IEEE*机器人与人类交互通信国际会议（*RO-MAN*），第1–7页。IEEE，2019年。[47] Sebastian von Mammen、Heiko Hamann 和 Michael Heider。机器人花园：面向植物—机器人生物混合系统的增强现实原型。载于第*22*届*ACM*虚拟现实软件与技术会议论文集，第139–142页，2016年。[48] Weiwei Wan、Feng Lu、Zepei Wu 和 Kensuke Harada。利用多模态三维视觉教机器人进行物体组装。*Neurocomputing*，259:85–93，2017年。[49] Daniel Weber、Wolfgang Fuhl、Andreas Zell 和 Enkelejda Kasneci。真实环境中的基于注视的物体检测。*arXiv*预印本 *arXiv:2203.15651*，2022年。[50] Daniel Weber、Enkelejda Kasneci 和 Andreas Zell。利用增强现实进行机器人外部标定和基于眼动的人机协作。载于*2022*年*ACM/IEEE*人机交互国际会议论文集，2022年。[51] Daniel Weber、Thiago Santini、Andreas Zell 和 Enkelejda Kasneci。利用注视信息提取未知物体的位置候选区域，以实现人机交互。载于*2020*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第11086–11093页。IEEE，2020年。[52] Li Xiao 和 Vikas Kumar。客户服务机器人：有益的补充还是最终的替代品？*Journal of Service Research*，24(1):9–29，2021年。[53] Jason Yosinski、Jeff Clune、Yoshua Bengio 和 Hod Lipson。深度神经网络中的特征有多强的可迁移性？神经信息处理系统进展，27，2014年。[54] Alex Yu、Vickie Ye、Matthew Tancik 和 Angjoo Kanazawa。pixelNeRF：由一张或少数几张图像构建神经辐射场。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第4578–4587页，2021年。[55] Fuzhen Zhuang、Zhiyuan Qi、Keyu Duan、Dongbo Xi、Yongchun Zhu、Hengshu Zhu、Hui Xiong 和 Qing He。迁移学习综合综述。*Proceedings of the IEEE*，109(1):43–76，2020年。[56] C Lawrence Zitnick 和 Piotr Dollár。Edge Boxes：从边缘定位物体候选区域。载于欧洲计算机视觉会议，第391–405页。Springer，2014年。