arXiv:2601.01675v1 [cs.RO] 4 Jan 2026

利用视觉和触觉传感器数据进行手持物体的视触觉六维姿态估计

Snehal Dikhale[1], Karankumar Patel[1], Daksh Dhingra[3],Itoshi Naramura[2],Akinobu Hayashi[2],Soshi Iba[1] 和 Nawid Jamali[1]

摘要——**了解物体的六维位姿有助于手持物体操作。现有的六自由度位姿估计方法使用视觉数据。由于机器人的夹爪会造成严重遮挡，手持物体的六维位姿估计颇具挑战性，这可能会对仅依赖视觉数据的方法产生不利影响。许多机器人在指尖处配备了触觉传感器，可用于补充视觉数据。本文提出一种方法，结合触觉数据和视觉数据，估计机器人手中所抓取物体的位姿。这项研究的主要挑战包括：1）缺乏触觉传感器数据的标准表示形式；2）如何融合来自视觉和触觉这两种异构来源的传感器数据；3）需要大规模训练数据集。为应对这些挑战，首先，我们提出使用点云来表示与触觉传感器接触的物体表面。其次，我们提出一种基于逐像素密集融合的网络架构，用于融合视觉和触觉数据，以估计物体的六维位姿。第三，我们扩展了英伟达的深度学习数据集合成器，使其能够在虚幻引擎4中为YCB物体与模型集中的11种物体生成合成照片级真实感视觉数据及相应的触觉点云。我们给出了仿真实验结果，表明在视觉数据之外加入触觉数据，可以提升手持物体的六维位姿估计精度。我们还展示了实验的定性结果：将网络部署到真实实体机器人上，成功验证了在合成数据上训练的网络向真实系统的迁移能力。**

索引词—**抓取与操作感知、抓取与操作中的深度学习、力觉与触觉感知**

I. 引言

ACCURATE 物体的六维位姿估计——包括三维旋转和三维平移——可用于操作与抓取、规划以及虚拟现实等应用。近期的位姿估计研究采用深度学习，利用彩色与深度图像估计物体的六维位姿 [1], [2], [3]。彩色图像能够很好地捕捉纹理 [3]，而深度数据包含几何信息，可提升不同光照条件下的预测精度 [1]。

大多数关于六维姿态估计的研究都聚焦于被其他物体部分遮挡的场景 [3], [2], [1]。

稿件收到日期：2021年9月9日；修订日期：2021年11月25日；录用日期：2021年12月21日。经副编辑评估审稿人意见后，Markus Vincze 推荐本文发表。[1]本田研究院美国公司{sdikhale, karankumar_patel, siba, njamali}@honda-ri.com[2]本田研发有限公司创新研究卓越中心，日本埼玉{itoshi_naramura,akinobu_hayashi}@jp.honda[3]美国华盛顿州西雅图，华盛顿大学机械工程系。本研究是在他于本田研究院美国公司实习期间完成的。dd292@uw.edu数字对象标识符（DOI）：见本页顶部。

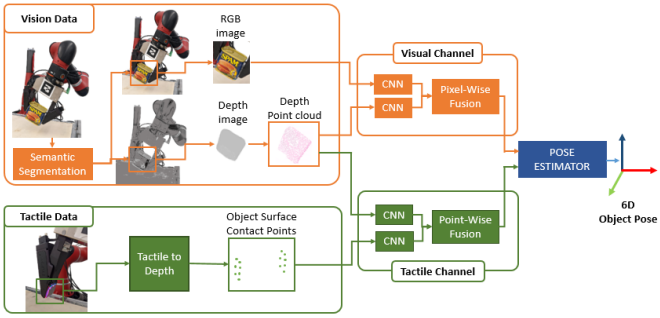


图 1：所提出方法的示意图。模型的输入是相机采集的彩色与深度图像，以及以点云形式表示的触觉接触点。输出是对感兴趣对象的六维位姿估计。

然而，手持物体的六维位姿鲜少受到关注 [4]。手内物体位姿估计面临的挑战包括机器人的夹爪和环境中的其他物体造成的严重遮挡。在这种情况下，机器人可以利用触觉传感器来改进其估计结果。此外，触觉数据能够捕捉被物体本身遮挡的表面几何结构 [5]。本文提出一种方法，将触觉数据与视觉数据融合，以估计机器人手中所抓取物体的六维位姿。我们推测，同时使用触觉数据和视觉数据，效果将优于单独使用视觉数据。

图 1 展示了所提出方法的示意图。模型的输入是相机采集的彩色与深度图像，以及以点云形式表示的触觉接触点。我们采用 Wang 等人提出的逐像素密集融合方法 [1] 来融合传感器数据。为实现视觉和触觉等异构传感器数据的有效融合，网络分为两个通道：视觉通道和触觉通道。在视觉通道中，彩色图像的特征与从深度图像得到的点云特征在像素级进行融合。在触觉通道中，从深度图像得到的点云特征与触觉传感器获取的物体表面点云特征在点级进行融合。这种融合方式的依据是彩色图像与深度图像彼此相关。类似地，触觉传感器获取的物体表面点云与深度图像中的点云也彼此相关。因此，深度图像中的点云将彩色图像与触觉传感器数据连接起来，从而考虑到被机器人夹爪遮挡的部分。

为机器人操作收集训练数据既耗时，又因机器人磨损而成本高昂。

和损耗。一种方法是使用合成数据训练模型。利用视觉传感器生成的合成数据已成功应用于抓取 [6], 和六维位姿估计 [2]。然而，包含触觉数据的操作合成数据生成仍较少受到探索。与视觉传感器不同，触觉传感器的发展还不够成熟[7]。目前尚无触觉数据的标准表示格式。表示方式取决于底层换能方法，例如，基于光学的触觉传感器会输出 RGB 图像 [8],，而其他传感器可能呈随机分布 [9]。因此，本文提出一种适用于生成六维位姿估计合成触觉传感器数据的表示格式。我们将触觉数据表示为手指与物体接触位置处的物体表面点云。我们扩展了英伟达的深度学习数据集合成器（NDDS） [10], ，以生成合成照片级真实感视觉数据、手指与物体各接触位置处对应的物体表面点云，以及来自 YCB物体与模型集 [11]中 11 个物体的六自由度位姿真值。我们的数据集为每个物体包含 20000 个样本。

本文的贡献包括：1）一种触觉传感器不变表示，用于捕捉手指与物体的接触表面；2）一种网络架构，用于融合视觉和触觉数据，以估计严重遮挡情况下手持物体的六维位姿；3）一种生成适用于手持物体六维姿态估计的合成视触觉数据集的方法。

我们展示了实验结果：使用所提出的合成视触觉数据集训练的模型，性能优于仅使用同一数据集中的彩色图像和深度图像训练的基线模型。我们还展示了定性实验结果：将使用合成数据集训练的网络和基线网络部署到真实实体机器人上后，网络能够泛化到不同的物理设置。

II. 背景

现有的六自由度位姿估计文献可分为三大类：几何匹配 [4], 概率方法 [12], [13], [14] 和机器学习方法 [15],[5]。几何方法利用触觉-几何形状一致性描述符来估计物体的位姿 [4]。这类技术的运行时性能较差。概率方法利用粒子滤波器 [12], 同步定位与建图（SLAM） [13], 和蒙特卡洛方法来估计物体位姿。这些方法已相当成熟，但可能需要与物体进行多次接触，也可能存在运行时性能较差的问题 [16]。

近年来，基于深度学习的六维姿态估计方法已展现出令人瞩目的成果。不同方法之间的一个区别在于网络的输入。Xiang 等人 [3]提出了一种新型网络 PoseCNN，用于估计物体的六维位姿。PoseCNN 使用 RGB 图像获得六维位姿的初始估计。随后，使用深度传感器数据，通过迭代最近点（ICP）算法改进物体位姿的初始估计。

与 PoseCNN 类似，Tremblay 等人[2] 使用在 RGB 图像上训练的卷积神经网络，来获得物体的六维位姿估计。作者使用合成数据集训练网络。为了弥合仿真到现实差距，作者开发了一种独特方法，将域随机化图像与照片级真实感数据结合起来训练卷积神经网络。作者展示的结果表明，在合成数据集上训练的网络能够泛化，并在 YCB-Video数据集上超越最先进方法。

Wang 等人 [1] 提出了一种结合 RGB 图像和深度数据来估计物体六维位姿的方法。作者提出了一种融合 RGB 与深度数据的新方法。该方法并非在全局层面融合数据，而是采用逐像素融合。逐像素融合的优点在于，网络可以基于物体的可见部分进行训练，从而最大限度地减小遮挡的影响。

大多数关于六维姿态估计的研究都没有使用触觉数据。近期，Bauza 等人 [17] 提出了一种使用触觉传感器的位姿估计方法。与 Bimbo 等人 [18], 类似，作者利用触觉数据估计物体在接触点处的局部形状。随后，将局部形状的估计结果与一组模拟接触进行比对排序，从而确定可能产生该接触表面的物体位姿的概率分布。

Watkins 等人 [5] 提出了一种用于确定未知物体几何形状的框架。在该框架中，将深度和触觉数据作为卷积神经网络的输入，以估计物体形状。作者证明，加入触觉数据能够提高形状估计的准确性。

相关研究在形状和位姿估计方面取得了显著进展。Watkins 等人 [5] 表明，使用触觉数据能够显著提高对受遮挡物体进行形状估计的准确性；然而，他们并未考虑估计物体的位姿。大多数最先进方法 $6D$ 中的位姿估计算法使用 RGB 数据，或同时使用 RGB 和深度数据，但没有一种利用触觉数据，因此不适用于手持物体操作，因为机器人夹爪会造成严重的自遮挡。本文旨在解决手持物体操作中的 $6D$ 位姿估计问题。

III. 网络架构

本节介绍一种方法，用于估计机器人抓持的已知物体的六维位姿。模型的输入是相机拍摄的彩色图像和深度图像，以及以物体表面点云形式表示的触觉接触点。网络的输出是物体在相机坐标系中的六维位姿。

A. 视觉-触觉传感器融合

图 2 展示了我们网络的整体架构。如第 I 节所述，颜色、深度和触觉输入以保持特征对应关系的方式进行融合。传感器融合分为两个通道：视觉通道和触觉通道。

1）视觉通道：视觉通道的输入由摄像头拍摄的彩色图像和深度图像组成。首先，对彩色图像进行语义对象分割，以分割出感兴趣对象。在语义分割中，每个像素都与一个对象类别相关联。

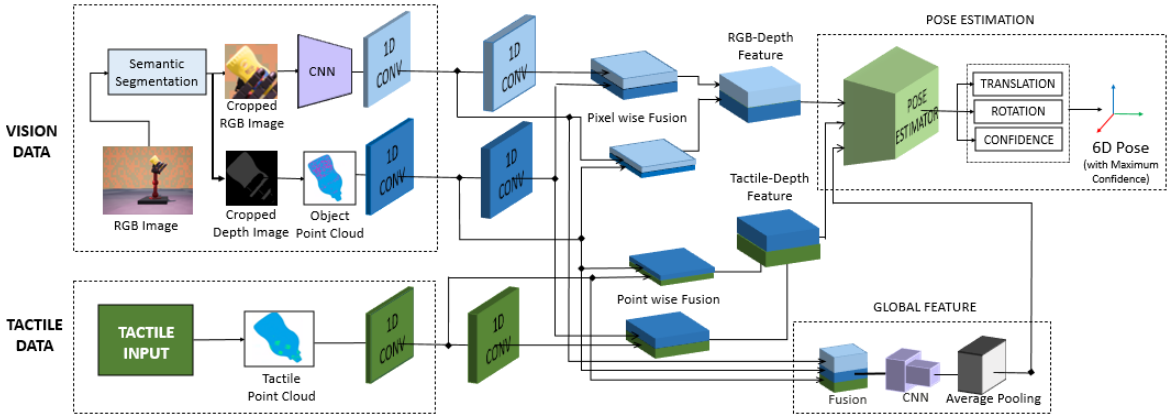


图 2：所提出的网络架构：给定彩色图像、深度图像以及手指与物体接触位置处的物体表面点云，网络估计物体的 $6D$ 位姿。位姿估计网络的详情见图 3。

在我们的实验中，我们采用了 Xian 等人提出的方法 [3], ，该方法使用编码器-解码器架构。

使用物体的语义分割结果对深度图像进行掩膜处理，并将其转换为相机坐标系中的三维点云。也就是说，对于彩色图像中可见的每个物体像素，我们将其与相机坐标系中物体对应的三维位置关联起来。

接下来，使用语义分割的边界框裁剪彩色图像。随后，将裁剪后的图像输入卷积神经网络（CNN），该网络处理彩色图像，并将裁剪图像中的每个像素映射为颜色特征嵌入。由于颜色嵌入对应整个裁剪图像，因此使用语义分割掩膜仅保留 n 个随机嵌入及其对应的三维点。这确保后续网络只考虑位于物体上的点。在我们的实验中， n 设为 1000。

颜色嵌入和深度嵌入分别经过一个两阶段卷积神经网络。颜色嵌入卷积神经网络和深度嵌入卷积神经网络的输出通过逐像素密集融合进行融合。

2) 触觉通道：触觉通道的输入由相机深度图像生成的点云和触觉传感器生成的物体表面点云组成。所有手指的触觉数据都以单个三维点云的形式输入网络。这种设计使网络不受机器人手指数量的影响，即实现夹爪不变性。与视觉通道类似，我们只保留 n 个随机点。随后，触觉数据经过一个两阶段卷积神经网络，生成触觉嵌入。触觉嵌入和深度嵌入通过逐点密集融合进行融合。

3) 全局特征：为了向网络提供全局上下文，我们还通过融合颜色、深度和触觉嵌入来生成全局特征，然后将其输入卷积神经网络和平均池化层。

B. 位姿估计器

上一节融合的传感器数据作为位姿估计器的输入，即来自

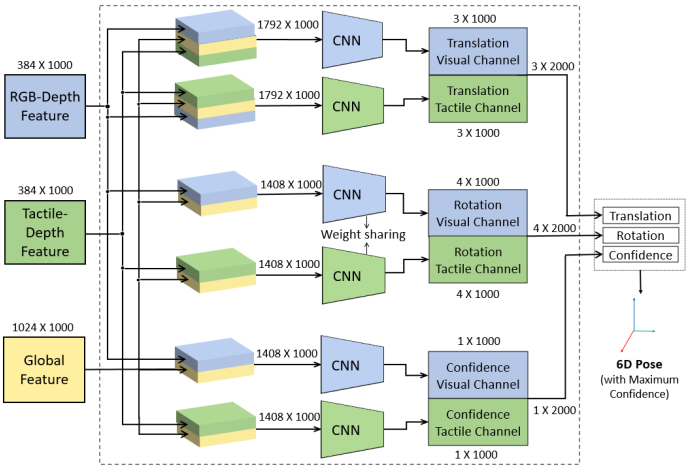


图 3：位姿估计器的网络架构。蓝色、绿色和黄色分别表示视觉通道特征、触觉通道特征和全局特征。

视觉通道、来自触觉通道的触觉-深度特征以及全局特征。图 3 展示了位姿估计器的网络架构。对于视觉通道和触觉通道中的每种特征，网络都会估计一个平移向量、一个旋转向量和一个置信度，分别以蓝色和绿色表示。为了估计平移向量，RGB-深度特征、触觉-深度特征和全局特征分别输入视觉通道卷积神经网络（蓝色）和触觉通道卷积神经网络（绿色）。为了估计旋转向量，在孪生网络架构中，将触觉-深度特征与全局特征结合后输入触觉通道卷积神经网络，而将 RGB-深度特征与全局特征结合后输入视觉通道卷积神经网络；两个网络共享权重。置信度则通过将触觉-深度特征与全局特征结合后输入触觉通道卷积神经网络进行估计，而将 RGB-深度特征与全局特征结合后输入视觉通道卷积神经网络。在我们的实验中，视觉通道和触觉通道各自输出 1000 个姿态估计值及其对应的置信度。由于这些点是随机选取的，其位姿估计的准确性会因位置和 RGB 颜色而异。通过学习置信度，网络会根据最具信息量的特征点选择位姿估计

因此，置信度最高的六维位姿将被选中。

C. 损失函数

我们使用式（3）中的损失函数来训练网络。如上一节所述，对于输入中的每个点 i，网络都会估计一个平移向量 T_i、一个旋转向量 R_i 和一个置信度 C_i。给定包含 M 个点的物体模型，逐点损失 L_i^p 的计算方法是：将模型点变换到估计位姿 $\hat{}p = [R_i, T_i]$ 下，再与变换到真实位姿 $p = [R, T]$ 下的相同模型点进行比较，如式（1）所定义。

逐点置信度损失 L_i^c 是估计置信度 C_i 与期望置信度 C_i^e 之间的差值，再加上正则化项 $wlog(C_i)$，如式（2）所定义。正则化项会对置信度较高的点施加惩罚。期望置信度是像素损失 $C_i^e = 0.5^{L_i^p}$ 的指数函数。网络损失 L 定义于式（3），即将逐点损失与置信度损失相加，并对所有点取平均。

$$L_i^p = \frac{1}{M} \sum_{j \epsilon M} ||(Rx_j + t) - (\hat{R}_i x_j + \hat{t}_i)|| \quad (1)$$

$$L_i^c = (\hat{C}_i - C_i^e - w\ log(C_i)) \quad (2)$$

$$L = \frac{1}{N} \sum_{i \epsilon N} (L_i^p + L_i^c) \quad (3)$$

IV. 数据集生成

本节将介绍触觉数据的生成方法，随后介绍数据集采集设置及其特征。

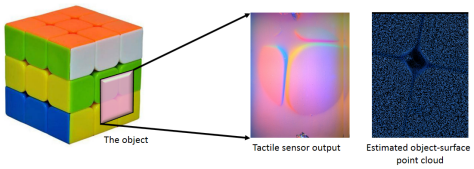
A. 触觉传感器不变位姿估计

如前所述，不同触觉传感器的空间分辨率可能差异很大。为了使位姿估计算法不受触觉传感器类型的影响，我们提出将触觉数据以物体表面点云的形式输入算法。也就是说，当触觉传感器与物体接触时，我们根据触觉传感器数据估计与触觉传感器接触的物体表面。可利用正向运动学获取与物体接触的每个触觉像素的三维位置。图4展示了如何将这一方法应用于基于光学的触觉传感器和基于压力的触觉传感器。除传感器外，这种表示方式还使网络不受所用夹爪类型的影响。

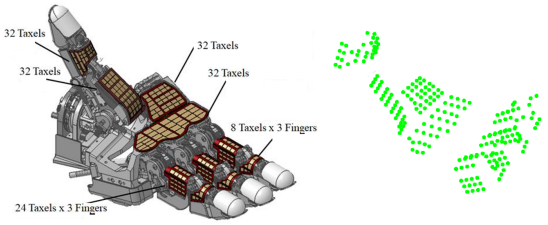
B. 数据采集设置

我们采用了Tremblay等人提出的方法 [10]，该方法使用虚幻引擎4生成照片级真实感仿真数据。我们扩展了作者开发的插件，以生成触觉点云以及颜色和深度数据。

仿真环境由一台主摄像头和一台配备四指夹爪（Allegro Hand）及触觉传感器的6自由度机械臂组成。如图5b所示，我们在夹爪的每个指骨上安装

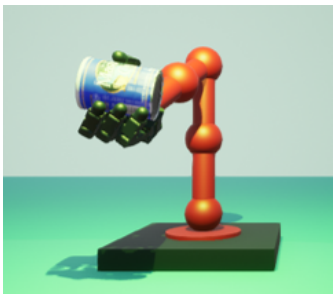


(a) GelSight [8] 触觉传感器

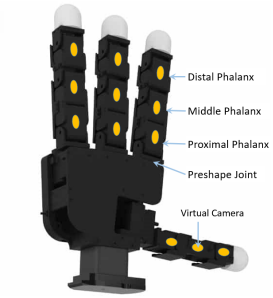


(b) 配备触觉传感器的多指手

图4：触觉数据生成物体表面点云的示意图：a）基于光学的触觉传感器及其对应的点云；b）基于压力的触觉传感器及其对应的触觉点云。



(a)



(b)

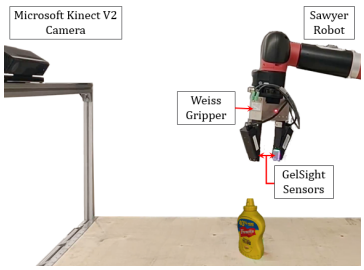
图5：a）采用域随机化的虚幻引擎数据采集设置；b）配备触觉传感器的 Allegro Hand。黄色椭圆表示用于在仿真中生成触觉数据的虚拟深度相机的位置。

一个虚拟相机。每个虚拟相机的原点位于相应指骨的表面上，z轴指向手指外侧。当夹爪与物体接触时，这些虚拟相机会捕获物体表面的点云。主摄像头位于机器人前方，用于生成彩色与深度图像。数据集按照算法1所述的流程生成。

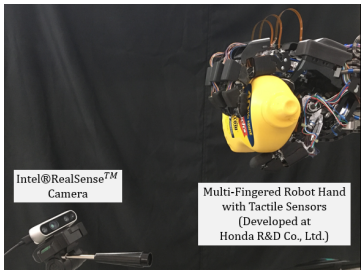
C. 数据集特征

我们根据可抓取性和三维模型的质量，选取了图6c所示的11个 YCB物体 [11]。该数据集每个物体包含20000个样本，每个样本由放置在机器人夹爪每个指骨上的12个虚拟触觉相机所采集的深度图像（图5b）以及主摄像头采集的 RGB-D 图像组成。放置在手指上的相机的高度和深度分辨率为32×32，并与触觉传感器的形状成比例。主摄像头采集整个场景的640×480图像。

为了在数据集中提供足够的变化以提升网络泛化能力，收集的数据集包含



(a) 机器人设置 I



(b) 机器人设置-II



(c) YCB物体

图6：用于硬件部署的实验装置，以及我们数据集中的 YCB物体。

算法1 数据采集流程

对每个： 物体
1：**当**随机轨迹 $<$ 200**执行时**
2：　将物体随机放置在机器人工作空间内
3：　随机设置物体的朝向
4：　随机设置夹爪预形状关节（图5b）
5：　抓取物体
6：　**当**实例 $<$ 100**时**
7：　　记录 RGB-D图像，
8：　　记录触觉物体表面点云
9：　　记录物体在相机坐标系中的位姿
10：　　将机械臂移动到新位置
11：　**结束 while 循环**
12：　释放物体
13：**结束循环**

经过域随机化处理的样本。数据集中有 20% 的样本未应用域随机化，其余 80% 为经过域随机化处理的样本。我们使用英伟达的深度学习数据集合成器的虚幻引擎4域随机化插件，对背景和地板表面进行随机化，并添加位置和颜色随机变化的聚光灯，如图 5a 所示。

V. 实验

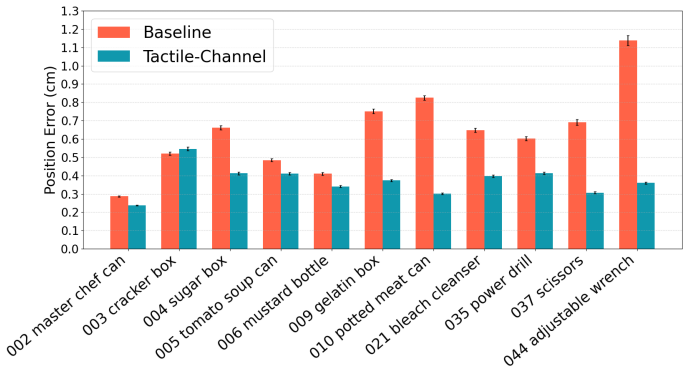
网络在第 IV 节所述的数据集上进行训练。

A. 网络训练设置

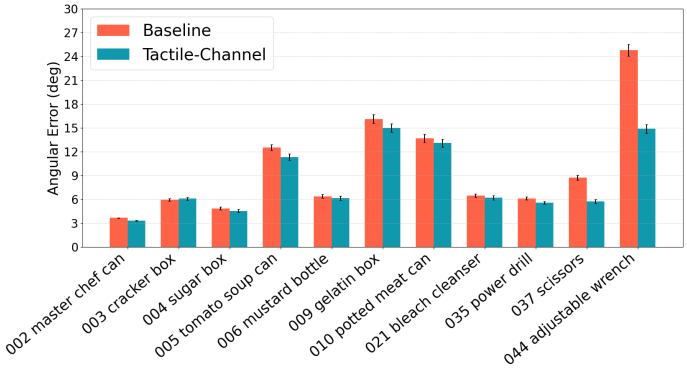
数据集按 4:1 的比例划分，其中每个物体的训练集包含 16000 个样本，测试集包含 4000 个样本。网络训练 1500 轮，批量大小为 16。网络的学习率 lr 设为 0.0002。式（2）中的超参数 w 设为 0.015。网络使用 PyTorch 中的 Adam优化器进行训练。

B. 硬件部署

我们在两种配备不同夹爪和触觉传感器的设置中部署了网络。第一种设置（设置一）由一个安装在Sawyer机器人上的Weiss夹爪和两个Gelsight触觉传感器组成。Kinect2 RGB-D相机安装在距机器人底座 $4\,m$ 处， $45°$ 朝向工作空间（见图6a）。工作空间内布置了



(a) 位置误差



(b) 角度误差

图 7：数据集中每个物体的位置误差和角度误差。

OptiTrack动作捕捉系统，用于记录感兴趣对象的真实六自由度位姿。

第二种设置（设置二），如图 6b 所示，由一台 Intel ®RealSense™ 相机和一只由本田研发有限公司开发的多指机器人手组成，如图 4b 所示。它有 4 根手指、16 个关节（其中 11 个为驱动关节），还配备了一套包含 224 个触觉像素的触觉传感器套件。

该网络部署在配备 NVIDIA Quadro RTX 6000 的工作站上。仅网络本身的推理时间为 10.1 ms，而包括语义分割网络和 ROS 开销在内的整个处理流程耗时 109.5 ms。

VI. 结果

为验证我们的假设——同时使用触觉与视觉数据能够提升物体的六维位姿估计效果——我们展示了实验结果，将我们的方法与一种基线方法进行比较。

该方法仅使用视觉数据来估计物体的六维位姿。我们通过修改自身网络构建基线方法：移除触觉通道，仅保留视觉通道。随后，我们使用来自同一数据集的彩色与深度图像训练这个纯视觉网络。在接下来的章节中，我们将对该方法在合成数据集上的性能进行定量分析，随后对部署在实体机器人上的网络进行定性分析。除非另有说明，所有报告的误差均为标准误差。

A. 定量评估

为了评估所提出方法的性能，我们使用两个指标：位置误差和角度误差。位置误差是估计位置$\hat{T}$与真实位置 T 之间的欧几里得距离。我们使用两个单位四元数之间的夹角来表示角度误差 [19]，如式（4）所示。

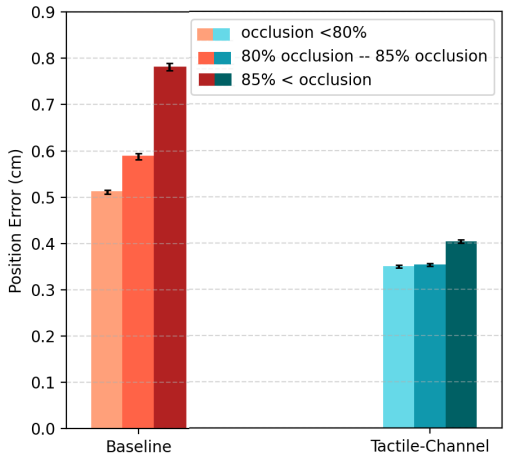
$$\theta = \cos^{-1}(2\langle \hat{q}, q \rangle^2 - 1) \tag{4}$$

其中，θ 为角度误差，$\hat{q}$ 为网络给出的估计四元数，q 为真实四元数，$\langle \hat{q},\ q \rangle$ 为两个四元数的内积。

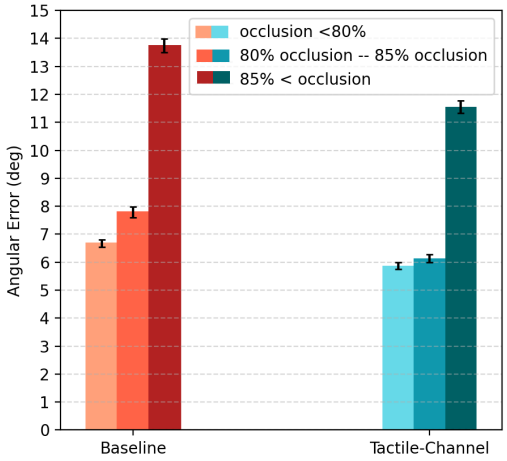
1）位置和角度指标上的表现：图7展示了数据集中每个物体的位置误差和角度误差。结果表明，除饼干盒外，对于所有物体，我们的网络都以具有统计显著性的优势胜过基线方法。对于某些物体，例如活动扳手，我们方法的位置误差为$0.36\,cm$，而基线方法为 $1.14\,cm$。其他相较基线有显著改进的物体包括剪刀、罐装肉罐头和明胶盒。分析角度数据时，我们发现活动扳手和剪刀的表现有显著改进。我们还注意到，在角度估计方面，基线方法与我们方法之间的差距没有位置估计方面那么明显。这些结果表明，网络依赖从相机图像中提取的颜色特征来推断物体的朝向。这可能导致饼干盒这类视觉上纹理密集的物体获得较高置信度，使其朝向估计与基线方法相近，而位置估计的准确度相对较低。

2) 遮挡程度对性能的影响：为分析遮挡程度对网络性能的影响，我们将网络性能按照遮挡程度划分为三个级别进行分析：低于 80%、介于 80% 和 85% 之间，以及高于 85%。给定物体的三维模型点云，以及从主摄像头深度图像获取的点云，遮挡程度按模型中对主摄像头可见的物体点所占百分比计算。首先，我们将真实变换应用于模型；随后，使用 K近邻算法，如果主摄像头点云中存在距离模型点云中某个点不超过阈值 ϵ 的点，则将该点标记为可见。遮挡程度是模型中被标记为不可见的点所占的百分比。

图 8 展示了该分析的结果。我们可以看到，随着遮挡程度的增加，基线网络的性能显著下降，



(a) 位置误差

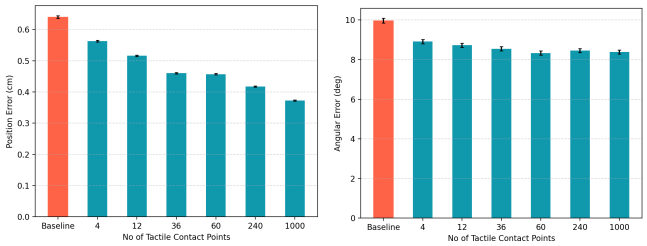


(b) 角度误差

图 8：基线和 Tactile- Channel 网络在不同遮挡级别下的位置误差和角度误差。较深的色调表示更高的遮挡程度。

在严重遮挡的情况下，我们的网络的位置误差（0.4 ± 0.003 厘米）几乎只有基线（0.78 ± 0.008 厘米）的一半。我们还看到角度误差降低了 2.3 度，分别为我们的网络和基线的 11.5 ± 0.22 度和 13.8 ± 0.24 度。我们注意到，与我们网络能够保持性能的位置误差相比，严重遮挡下的角度误差虽然优于基线，却出现了大幅下降。这表明，物体纹理对于推断物体的朝向非常重要。例如，在没有任何视觉线索的情况下，很难判断芥末瓶的背面还是正面朝向相机，这是预料之中的。事实上，通过查看每个物体的角度误差（图 7b），我们注意到，在加入触觉数据后角度性能提升较大的物体，通常是那些纹理较少的物体；例如，活动扳手的纹理是均匀的。要区分物体的一侧和另一侧，很大程度上取决于点云数据所捕获的物体形状。

3）触觉点数量对性能的影响：我们进一步开展分析，以研究触觉接触点数量对性能的影响。我们修改了所提出的网络，使其能够接受 4、12、36、60、240 个触觉接触



(a) 位置误差 (b) 角度误差

图9：触觉接触点数量变化对网络性能的影响，其中橙色和蓝色柱分别表示基线网络和触觉通道网络。

点。图 9 展示了我们数据集中所有物体的这项实验结果。我们还绘制了基线网络的性能作为参考。正如预期的那样，触觉接触点数量减少会增大位置误差和角度误差。结果表明，即使触觉接触点数量减少，加入触觉数据仍优于纯视觉基线。

B. 定性评估

本节将定性分析在 V-B 节所述的硬件设置上实时同时部署我们的网络和基线网络的情况。请注意，这些网络仅使用合成数据集进行训练。部署期间，我们的网络和基线网络接收相同的输入。随附视频展示了本节所述结果的可视化。对于机器人设置 I（图6a），视频所示数据的平均位置误差和角度误差分别为 36.88 ± 1.27 度和 2.69 ± 0.13 厘米；相比之下，基线网络的相应误差为 64.57 ± 1.40 度和 4.76 ± 0.32 厘米。

1) 位置和角度指标上的性能： 在部署过程中，与基线网络相比，我们的网络展现出了稳定的逐帧输出。图 10 展示了一些示例：在这些示例中，基线网络（蓝色）的估计结果与真实位姿（白色）存在显著偏差，而我们的网络（绿色）仍能保持合理的估计结果。这表明，该网络能够利用触觉信息，尤其是来自物体后方手指的触觉信息，从而更准确地估计物体的六维位姿。

2) 遮挡程度对性能的影响： 我们还在真实硬件设置上进行了一项实验，通过使用一张纸遮挡主摄像头的视野，逐渐增加遮挡程度。图 11 展示了该实验的结果，表明与基线网络（蓝色）相比，我们的网络（绿色）能够应对更高程度的遮挡。真实位姿以白色显示。结果表明，在严重遮挡的情况下，同时使用触觉和视觉数据的网络优于仅使用视觉数据的网络。

C. 消融研究

为评估我们的设计选择对网络性能的影响，我们进行了三项消融研究，以量化以下因素带来的提升：1）孪生网络；2）全局

表 I：消融研究结果

模型	位置误差（cm）	角度误差（度）
触觉通道（孪生）	0.299 ± 0.002	8.074 ± 0.105
非孪生	0.305 ± 0.002	8.269 ± 0.102
无全局特征	0.876 ± 0.004	11.385 ± 0.123
仅触觉	1.849 ± 0.009	59.268 ± 0.282

表 II：与最先进方法的比较

模型	位置误差（厘米）	角度误差（度）
PoseCNN	6.146 ± 0.023	10.896 ± 0.082
PoseCNN+ICP	6.158 ± 0.023	10.897 ± 0.083
触觉通道	0.299 ± 0.002	8.074 ± 0.105

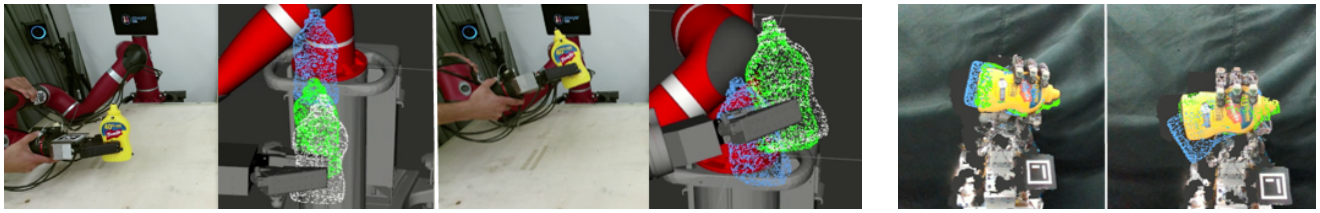
特征；3）视觉特征，即移除视觉通道，仅使用触觉数据进行训练的纯触觉网络。表 I 展示了全部 11 个物体的平均网络性能。结果表明，使用孪生网络可使位置精度提升 $0.006\,cm$ ，角度精度提升 0.2 度。全局特征可使位置精度提升 $0.57\,cm$，角度精度提升 3.31 度。视觉数据可使位置精度提升 $1.55\,cm$ ，角度精度提升 51.19 度，这与我们的观察一致：对物体朝向的推断取决于 RGB 图像中的颜色特征。

D. 与最先进方法的比较

为了评估我们数据集的难度，我们在该数据集上训练了 PoseCNN [3], ，这是一种基于视觉的最先进六维位姿估计算法。表 II 展示了分析结果。我们注意到，位置估计的性能显著下降；对于一种仅使用 RGB图像、且受到机器人夹爪严重遮挡影响的方法而言，这一结果在预期之中。与我们的方法相比，角度误差也增加了 4.78 度。这些结果表明，我们的数据集为位姿估计问题带来了挑战。

VII. 结论与未来工作

我们提出了一种新方法，用于解决估计手持物体 $6D$ 位姿的问题。手持位姿估计面临的一项挑战，是如何补偿物体大面积被遮挡部分周围的几何信息。在所提出的方法中，我们利用机器人手指上的触觉感知数据，因为这些手指会遮挡物体。我们提出了一种传感器无关表示，以手指—物体接触面点云表示触觉数据。我们还引入了一种视触觉网络架构，以有意义的方式融合视觉与触觉传感器的数据。为高效训练模型，我们提出了一种合成数据采集流程，用于生成适合手持 $6D$ 位姿估计的大规模视触觉数据集。我们展示了实验结果，并将所提出的方法与纯视觉基线进行了比较。在合成数据上的结果表明，我们的网络性能优于纯视觉基线，且差异具有统计显著性，尤其是在严重遮挡的情况下。我们还展示了一项研究的结果，其中将触觉输入的数量在 4 至 240 之间进行了调整。结果表明，尽管估计精度有所下降，



(a) 设置一 (b) 设置二

图 10：在 V-B 节所述的设置中进行部署：绿色和蓝色点云分别表示触觉通道和基线。设置一中，左图为主相机的视角；右图中的白色点云表示动作捕捉系统提供的真实位姿。请注意，设置二没有动作捕捉系统。

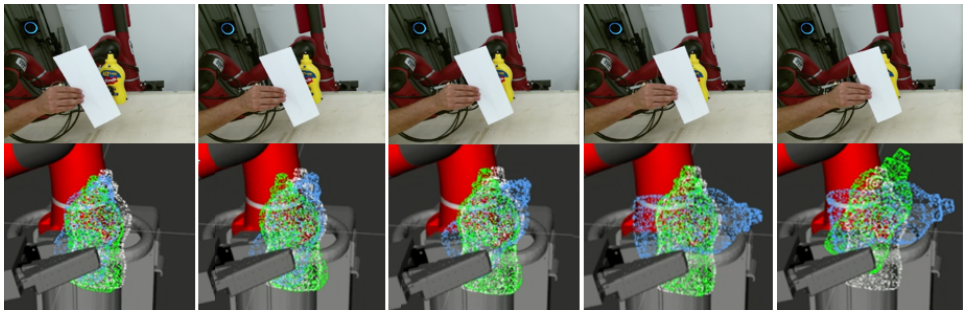


图 11：在 V-B 节所述的机器人设置 I 中，不同遮挡级别下的位姿估计，其中白色、绿色和蓝色点云分别表示真实位姿、我们的网络和基线。

即使触觉输入的数量减少到 4，我们的方法仍优于基线。此外，我们还展示了定性结果：我们将网络部署在采用不同硬件配置的实体机器人上。定性结果表明，仅使用合成数据训练的网络能够成功迁移到真实世界环境中。

我们的方法存在的局限之一是，方向推断依赖于视觉传感器获取的物体颜色特征。尽管在中度至重度遮挡下，我们的方法仍能保持性能，但在极重度遮挡下，性能会退化至基线水平。为缓解这一局限，未来可以探索多种方法。一种方法是采用长短期记忆网络等方法，利用时间连贯性过滤错误估计。另一个值得探索的方向是利用夹爪几何形状和手指位置，帮助网络滤除物理上不合理的位姿估计值。我们的数据集中大多数物体至少沿一个轴对称。未来，我们希望扩展数据集，纳入 3D打印非对称物体，以评估我们的方法在此类物体上的有效性。我们还注意到，我们的方法能够抵御环境遮挡。未来，我们希望开展研究，测试我们的方法在不同环境遮挡程度下的表现。

参考文献

[1] C. Wang、D. Xu、Y. Zhu、R. Martin-Martin、C. Lu、L. Fei-Fei 和 S. Savarese，《DenseFusion：基于迭代密集融合的6D物体姿态估计》，载于计算机视觉与模式识别，2019年。[2] J. Tremblay、T. To、B. Sundaralingam、Y. Xiang、D. Fox 和 S. Birchfield，《用于家庭物体语义机器人抓取的深度物体姿态估计》，载于机器人学习会议（*CoRL*），2018年。[3] Y. Xiang、T. Schmidt、V. Narayanan 和 D. Fox，《PoseCNN：用于杂乱场景中6D物体姿态估计的卷积神经网络》，载于机器人学：科学与系统（*RSS*），2018年。

[4] J. Bimbo、S. Luo、K. Althoefer 和 H. Liu，“基于协方差的触觉-几何匹配进行手内物体位姿估计”，*IEEE*机器人与自动化快报，2016年。[5] D. Watkins-Valls、J. Varley 和 P. Allen，“用于抓取与操作的多模态几何学习”，载于*IEEE*机器人与自动化国际会议 *(ICRA)*，2019年。[6] J. Mahler、J. Liang、S. Niyaz、M. Laskey、R. Doan、X. Liu、J. Aparicio Ojea 和 K. Goldberg，“Dex-Net 2.0：利用合成点云和解析抓取指标，通过深度学习规划稳健抓取”，载于机器人学：科学与系统 *(RSS)*，2017年。[7] A. Yamaguchi 和 C. G. Atkeson，“机器人操作触觉感知与传感器的最新进展：我们能否将触觉感知转化为视觉？”先进机器人学，2019年。[8] M. K. Johnson 和 E. H. Adelson，“用于测量表面纹理和形状的逆向成像传感”，载于*IEEE*计算机视觉与模式识别会议 *(CVPR)*，2009年。[9] N. Wettels、J. A. Fishel 和 G. E. Loeb，“多模态触觉传感器”，载于人手对机器人手开发的启发，2014年。[10] T. To、J. Tremblay、D. McKay、Y. Yamaguchi、K. Leung、A. Balanon、J. Cheng、W. Hodge 和 S. Birchfield，“NDDS：NVIDIA深度学习数据集合成器”，2018年。[11] B. Calli、A. Singh、A. Walsman、S. Srinivasa、P. Abbeel 和 A. M. Dollar，“YCB物体与模型集：迈向操作研究的通用基准测试”，载于国际先进机器人学会议，*(ICAR)*，2015年。[12] A. Aggarwal 和 F. Kirchner，“物体识别与定位：触觉传感器的作用”，传感器，2014年。[13] C. Fox、M. Evans、M. Pearson 和 T. Prescott，“利用仿生触须机器人实现触觉SLAM”，载于*IEEE*机器人与自动化国际会议 *(ICRA)*，2012年。[14] A. S. Vazquez、R. Fernandez、A. Lopez、E. Valero、I. Payo 和 A. Adan，“手内物体定位：简单与复杂触觉传感器的比较”，*IEEE*传感器，2014年。[15] J. Varley、C. DeChant、A. Richardson、J. Ruales 和 P. Allen，“形状补全助力机器人抓取”，载于*IEEE*智能机器人与系统国际会议 *(IROS)*，2017年。[16] S. Luo、J. Bimbo、R. Dahiya 和 H. Liu，“物体属性的机器人触觉感知：综述”，机电一体化，2017年。[17] M. Bauza、E. Valls、B. Lim、T. Sechopoulos 和 A. Rodriguez，“通过几何接触渲染，从首次触碰开始进行触觉物体位姿估计”，*arXiv*，2020年。[18] J. Bimbo、S. Rodr´ıguez-Jim´enez、H. Liu、X. Song、N. Burrus、L. D.

Senerivatne、M. Abderrahim 和 K. Althoefer，“融合视觉与触觉信息进行物体姿态估计与跟踪”，载于*IEEE*智能系统多传感器融合与集成国际会议，2012年。[19] D. Q. Huynh，“三维旋转度量：比较与分析”，数学成像与视觉杂志，第35卷，第2期，2009年。