arXiv:2302.03573v2 [cs.RO] 6 Mar 2023

局部神经描述符场：用于操作的局部条件化物体表示

Ethan Chun[1], Yilun Du[1], Anthony Simeonov[1],Tomas Lozano-Perez[1],Leslie Kaelbling[1]1麻省理工学院计算机科学与人工智能实验室，美国



图 1：仅凭少量（5–10 次）对两种不同直立马克杯进行抓取和拿起的现实世界示范，局部神经描述符场便能成功抓取并拿起一组几何形状各异、位于任意 SE(3) 位姿的物体。

摘要——在家庭环境中运行的机器人会遇到各种各样独特而陌生的物体。虽然系统可以用其中许多物体进行训练，但要预测机器人将会遇到的所有物体并不可行。本文提出一种方法，将从有限数量的示范中习得的物体操作技能，泛化到来自未见形状类别的新物体。我们的方法——局部神经描述符场（L-NDF）——利用定义在物体局部几何上的神经描述符，在测试时有效地将操作示范迁移到新物体上。为此，我们利用物体之间共有的局部几何，构建出更具泛化能力的操作框架。我们通过在仿真和现实世界中操作位姿新颖的新物体，展示了该方法的有效性。项目网站、视频和代码： https://elchun.github.io/lndf/。

I. 引言

在家庭环境中自主运行的机器人会遇到各种各样的未见物体。虽然个别物体的形状可能是新颖的，但许多物体都可以分解为一组此前见过的组成部分。请看图1中所示的新物体——尽管带把手的瓶子可能是未见过的，但瓶子和马克杯各自都是已知的。因此，可以设想机器人通过瓶子和马克杯上习得的技能来操作这个新物体。本文探讨如何借助模仿学习范式实现此类泛化。具体而言，我们希望构建一个系统：给定在单一类别物体上进行的少量（5–10 次）操作示范后，它便能在任意 SE(3) 方向下，成功对新型物体执行该技能。

为实现高效学习，我们基于神经描述符场（NDF）系统 [23]。NDF 为形状中的每个点分配一个稠密描述符，并为同一类别中不同物体的相似点分配相似的描述符。通过在新物体中找到一组对应的稠密描述符，可以将物体操作推广到同一类别中的新物体。然而，NDF 的一个局限是，它依赖单个全局潜变量来编码给定类别中形状的所有几何特征。当输入一个属于新类别的物体时，这种表示无法捕捉由此产生的几何结构，从而阻碍 NDF 将物体操作迁移到新的物体类别。

我们通过使用潜变量体素网格来局部捕获形状的几何信息和描述符，从而规避这一问题（见图 2）；其中，每个潜在表示对一个局部空间区域进行编码。采用这种编码方案后，新类别形状的描述符能够得到更准确的编码，因为新形状中的各个局部块分别对应于不同训练对象类别中的局部块。我们展示了这种编码如何使物体操作能够泛化到新类别，并将我们的方法称为局部神经描述符场（L-NDF）。

在进行局部描述符编码时，会出现一个问题：物体某一部分的描述符可能会随着物体的变换而发生变化。例如，马克杯的把手在发生平移和旋转时，会由不同的体素潜变量表示。为确保描述符在刚体变换下保持一致，我们提出了一种对比损失，在物体发生变换时显式约束描述符的一致性。

要将物体操作示范从一个物体迁移到另一个物体，我们必须找到两个物体之间对应的描述符集合。在 NDF 中，使用全局梯度优化过程来最小化描述符

通信地址：yilundu@mit.edu

距离。对于 L-NDF，由于物体的描述符仅在局部进行编码，缺乏描述符在形状中变化的统一全局方向，因此难以运行类似的全局优化过程。为克服这一困难，我们提出在形状中的一组不同位置初始化优化，并运行局部优化，选择描述符距离最小的描述符作为最终匹配描述符。

我们证明，L-NDF 可可靠地用于将物体操作泛化到新物体，以及处于未见过的 SE(3) 位姿的物体。仅需 (5-10) 次示范，我们的框架便能够在仿真和真实机器人上操作新物体（例如茶杯，或带把手的碗）。

II. 相关工作

A. 可泛化操作

我们的工作延续了利用模仿学习进行操作的长期研究脉络。当物体模型已知时，可以使用位姿估计来执行操作 [22, 30, 31]。当物体的精确几何未知时，可以使用基于粗略三维基元的模板匹配 [10, 16, 27] 或非刚性配准 [22]；但当物体与模板相差较大时，这些方法仍会遇到困难。近期研究探索了更灵活的模仿学习表征，例如关键点 [8, 9, 14]或稠密描述符 [6,23, 26]。与我们的工作最为相似的是 DON [6] 和 NDF，它们分别探索了用于物体操作的二维和三维稠密描述符——但二者都只展示了在同一物体类别内的泛化能力。相比之下，我们的方法能够在测试时对形状类别全新的物体进行操作。

B. 机器人学中的神经隐式表示

神经隐式表示 [15, 19] 已成为机器人学中表示三维几何的一种前景广阔的方法。不同的研究探索了隐式表示在导航 [1]，定位 [7, 17, 28],同步定位与建图 [18, 25, 32], 以及操作 [12, 13, 21, 23, 24,29]中的应用。在操作领域，[12, 29] 将神经辐射场作为一种提取场景底层三维几何的方法。相比之下，[21, 23, 24] 基于神经描述符场框架学习操作技能，并利用底层高维神经描述符迁移和泛化示范。我们的工作将神经描述符场扩展至局部条件化隐式表示。

III. 背景：使用神经描述符场进行操作

神经描述符场（NDF）[23] 通过一个函数 f 对物体形状进行编码，该函数将一个三维点$\mathbf{x} \in \mathbb{R}^3$和一个部分物体点云$\mathbf{P} \in \mathbb{R}^{3\times N}$ 映射为 $\mathbb{R}^d$中的空间描述符：

$$f(\mathbf{x}|\mathbf{P}) : \mathbb{R}^3 \times \mathbb{R}^{3\times N} \to \mathbb{R}^d. \tag{1}$$

NDF 也经过训练，能够学习同一类别物体之间的对应关系，使不同实例中靠近相似几何特征的点（例如，两个不同瓶子瓶颈附近的点）映射为相似的描述符值。

NDF 可以推广到为完整的 SE(3) 位姿而非单个点分配描述符。这是通过拼接一组刚性查询点 $\mathcal{X} \in \mathbb{R}^{3\times N_q}$ 中各个点的描述符来实现的；也就是说，这组点由三个或更多不共线的点$\mathbf{x}_i,\ i = 1...N_q$组成，并受到约束，必须作为一个整体进行刚性变换。这种构造使 NDF 能够通过 SE(3) 位姿 $\mathbf{T}$ 对 $\mathcal{X}$ 的作用来表示该位姿；也就是说，通过d $\mathbf{T}\mathcal{X}$中的点来表示： the *transformed query point clou*

$$\mathcal{Z} = F(\mathbf{T}|\mathbf{P}) = \bigoplus_{\mathbf{x}_i \in \mathcal{X}} f(\mathbf{T}\mathbf{x}_i|\mathbf{P}) \tag{2}$$

因此，F 将点云映射为 E 3d位P姿T d对应S的类别级位姿描述符 $\mathcal{Z} \in \mathbb{R}^{d\times N_q}$。

使用神经描述符场进行少样本操作学习。接下来，我们讨论如何利用 NDF 进行物体操作技能的少样本学习。考虑一组 K 示范，$\{\mathcal{D}_i\}_{i=1}^K$，其中每个示范 $\mathcal{D}_i$ =(P^i、T $^i_{pick}$、$\mathbf{T}^i_{place}$)由一个物体 P^{i}以及两个位姿组成：抓取前的末端执行器位姿 $\mathbf{T}^i_{pick}$，和放置表面的相对位姿 $\mathbf{T}^i_{place}$。我们定义一组查询点 $\mathcal{X}_{pick}$ 和 $\mathcal{X}_{place}$，分别用于表示夹爪和放置表面。随后，我们利用式（2）将每个位姿 $\mathbf{T}^i_*$ 编码为其描述符向量 $\mathcal{Z}^i_*$，并以相应的物体点云 P^{i}为条件，从而得到一组空间描述符元组 $\{(\mathcal{Z}^i_{pick}、\ \mathcal{Z}^i_{rel})\}_{i=1}^K$。对 K 示范中的描述符进行平均，以获得单一的抓取 和放置描述符 $\mathcal{Z}_{pick}$ 以及 $\mathcal{Z}_{rel}$

在测试时将新物体放置到场景中后，我们获得一个点云 $\mathbf{P}^{test}$，并利用3）恢复 $\mathbf{T}^{test}_{pick}$ 和 $\mathbf{T}^{test}_{rel}$，其方法是最小化与空间描述符之间的距离 $\mathcal{Z}_{pick}$ 以及 $\mathcal{Z}_{rel}$。

$$\bar{\mathbf{T}} = \underset{\mathbf{T}}{\operatorname{argmin}} \|F(\mathbf{T}|\mathbf{P}) - F(\hat{\mathbf{T}}|\hat{\mathbf{P}})\| \tag{3}$$

我们依靠现成的逆运动学和运动规划算法来执行最终预测的位姿。

IV. 局部神经描述符场

给定一组 K、单一物体类别的抓取与放置示范，$\{\mathcal{D}_i\}_{i=1}^K$，其中每个示范 $\mathcal{D}_i = (\mathbf{P}^i、\mathbf{T}^i_{pick}、\mathbf{T}^i_{place})$ 由部分物体点云 $\mathbf{P}^i$、末端执行器抓取位姿 $\mathbf{T}^i_{pick}$ 和放置位姿 $\mathbf{T}^i_{place}$组成，我们希望将任务推广到一组来自未见物体类别的新物体 $\mathbf{P}'$。为解决这一问题，我们开发了一种使用局部定义的描述符的方法，并提出了对神经描述符场流程（第 III节）的适当修改，以利用这类描述符。

具体而言，在第IV-A节中，我们介绍局部神经描述符场，并说明如何利用它们对物体的几何进行局部编码。在第 IV-B节中，我们讨论如何将 SE(3) 等变性引入 L-NDF 的底层描述符。最后，在第 IV-C节中，我们讨论如何修改底层优化过程，使其能够在局部描述符场景观中搜索理想位姿。

A. 局部描述子场

Aglobal NDF 模型无法有效泛化到新的物体类别。为了解决这个问题，我们为物体使用局部描述符场：体素网格中的每个元素

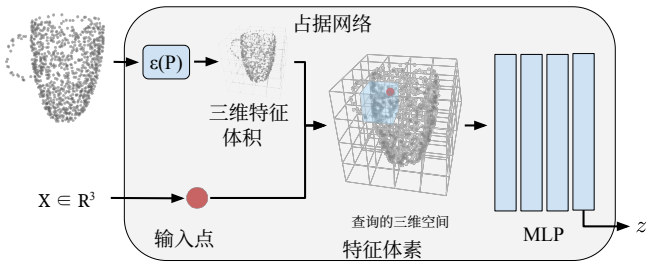


图 2：**局部神经描述场架构**– L-NDF 接受三维空间中的任意坐标 $\mathbf{x}$ 和条件点云 $\mathbf{P}$。随后，它使用编码器 $\epsilon(\mathbf{P})$ 将 $\mathbf{P}$ 编码为一个三维特征体，并查询包含$\mathbf{x}$ 的体素。接着，将这些特征输入多层感知机解码器，提取解码器最后一层的激活值，以生成空间描述符 z。

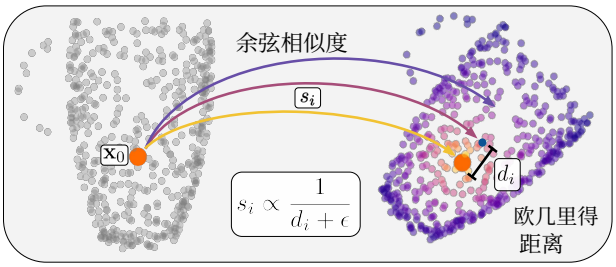


图 3：**L-NDF 的对比损失项**– 对于任意变换，三维坐标 $\mathbf{x}$ 相对于观测点云 $\mathbf{P}$ 的空间描述符都相似，$\mathbf{T} \in \mathrm{SE}(3)$。此外，几何距离越远的点，其描述符的相似度越低。

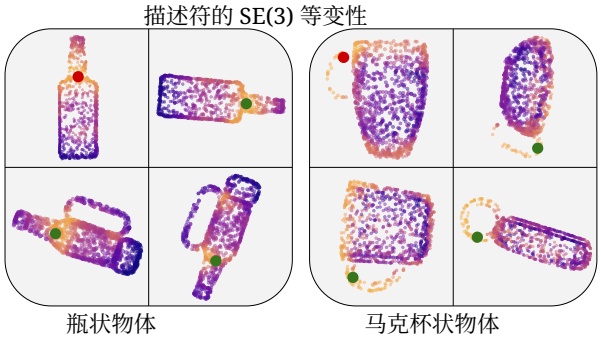


图 4：**物体编码的 SE(3) 等变性**– 所选点（红色）对应的余弦描述符差异热图。对于不同物体在任意 SE(3) 变换下，描述符场始终保持一致。

都包含该体素附近物体局部形状的潜向量表示。

在 L-NDF 中，我们使用卷积占据网络编码器 [20], $\epsilon(\mathbf{P})$，将部分点云编码$\mathbf{P}$ 为潜变量体素网格（如图 2所示）。查询特定点 $\mathbf{x}$时，会提取潜特征 $\epsilon(\mathbf{P})$中对应的体素，并通过多层感知机处理。随后，将多层感知机激活值拼接起来，生成潜代码 z。形式上，该编码器定义为（4）。

$$z = f(\mathbf{x}|\mathbf{P}) = \Phi(\mathbf{x}|\epsilon(\mathbf{P})_{|\mathbf{x}|}). \tag{4}$$

随后 [23], 我们利用占据重建来训练和学习神经描述符场的特征。

B. 训练和学习 *SE(3)* 等变性

为确保我们的模型能够泛化到目标物体的刚体变换，我们设计了一种训练方案，强制物体在同一局部坐标系中的同一点处的描述符，在物体进行 SE(3) 变换后保持不变。

强制实现 SE(3) 等变性。与 [23]，不同，我们的系统并非固有地具备 SE(3) 等变性。相反，我们利用对比损失项来塑造网络激活，使其表现出 SE(3) 等变性。形式上，如果对于任意刚体变换$\mathbf{T} \in \mathrm{SE}(3)$，编码器$f(\mathbf{x}|\mathbf{P})$满足以下条件，则称其具有 SE(3) 等变性：

$$f(\mathbf{x}|\mathbf{P}) \equiv f(\mathbf{T}\mathbf{x}|\mathbf{T}\mathbf{P}) \tag{5}$$

一种强制实现等变性的简单方法，是直接强制要求对应点的编码在 SE(3) 变换前后保持不变。然而，直接施加这一约束存在问题，因为我们发现 f 会将所有输入映射到相同的编码。因此，我们考虑直接施加另一项约束，即要求不同的输入点产生不同的编码，但发现所得描述符在不同形状之间不再具有语义一致性。

我们发现，构造同时具备 SE(3) 等变性和语义一致性的描述符，一种稳健的替代方法是施加约束（6）：在不同刚体变换$\mathbf{T}$ 下，两点之间的描述符相似度大致与它们的反距离成正比（如图所示。 3).

$$\mathrm{sim}(f(\mathbf{x}_1|P), f(\mathbf{T}\mathbf{x}_2|\mathbf{T}P)) \propto \frac{1}{\|\mathbf{x}_1 - \mathbf{x}_2\| + \epsilon}, \tag{6}$$

这一约束不仅确保描述符在形状的刚体变换下保持等变，同时也确保它们随点的小幅欧几里得扰动而平滑变化。

为施加此损失函数，我们在物体的包围盒内采样 k 个点。我们将第一个点 $\mathbf{x}_0$, 指定为参照点，并计算其与剩余 $k-1$ 个点的描述符相似度。对于每个点，我们计算 $f(\mathbf{x}_0|\mathbf{P})$ 与 $f(\mathbf{T}\mathbf{x}_i|\mathbf{T}\mathbf{P})$ 之间的余弦相似度 s_i,

$$s_i = \frac{f(\mathbf{x}_0|\mathbf{P}) \cdot f(\mathbf{T}\mathbf{x}_i|\mathbf{T}\mathbf{P})}{\max(||f(\mathbf{x}_0|\mathbf{P})|| \cdot ||f(\mathbf{T}\mathbf{x}_i|\mathbf{T}\mathbf{P})||, \epsilon)} \tag{7}$$

我们针对每个$\mathbf{x}_i$，计算其相对于第一个点$\mathbf{x}_0$的对应目标相似度值

$$t_i = \frac{1}{d(\mathbf{x}_0, \mathbf{x}_i) + \beta}, \tag{8}$$

并施加约束，使相似度大致与距离倒数成正比。如图4所示，此损失函数成功实现了物体间的 SE(3)等变性。

C. 位姿优化

使用 L-NDF 进行少样本任务学习时，我们必须在新的点云 $\mathbf{T}$ 上优化一个位姿 $\mathbf{P}$，使其与参考点云上的目标参考位姿 $\mathbf{T}^*$ 相匹配，参考点云为 $\mathbf{P}$。此优化过程详见（3）。传统 NDF 在一组查询点上进行全局优化，以获得最优位姿 $\mathbf{T}$；优化从物体原点处的随机方向开始。

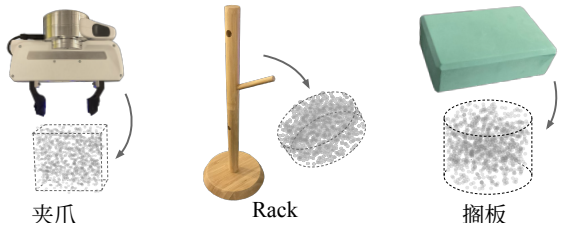


图 5：**查询点选择**– 每个执行任务中查询点的相对大小。对于抓取和置物架放置任务，我们使用大小与已知物体的接触几何（夹爪和销钉的大小）相近的查询点。对于放置表面，我们发现选择较大的查询点能够取得良好效果。

然而，将此方法用于局部 NDF 时会失效。由于 L-NDF 仅聚合局部几何中的信息，因此与相距较远的几何特征相关的信息很少。为缓解这些问题，我们引入了两种技术：初始平移和查询点选择。

初始平移。与传统 NDF 不同，我们在观测点云的包围盒内以随机旋转和平移初始化查询点。当使用足够数量的查询点实例时（我们发现 20 个已足够），至少有一组经过平移的查询点会初始化在接近目标几何特征的位置。随后，位姿优化会将查询点云调整到正确的目标位置。

查询点调优。我们发现，查询点选择对 L-NDF 的性能至关重要。如果查询点云过大，就会编码混淆几何和空空间。如果查询点云过小，则无法捕获足够的局部几何。我们发现，对于精确操作，可以在已知物体的预期接触几何附近采样查询点。对于更一般的位姿（例如放置在表面上），选择一种查询点云，使其在最小化所包含空空间体积的同时，最大化点云中所包含观测物体的预期体积，能够产生稳健的结果。有关更多细节，请参见图 5。

V. 实验：设计与设置

我们设计实验来检验以下问题：（1）L-NDF 对未见物体类别的泛化效果如何？（2）L-NDF 能否应用于真实机器人，仅凭少量单一物体类别的示范实现泛化？

A. 机器人环境设置

我们的环境由一台安装在桌面上的 Franka Panda 机械臂组成。桌子的每个角落都放置了深度相机，并经过统一标定，以获取机器人可达范围内物体的融合点云。我们在仿真中使用四台深度相机，在现实环境中使用两台。仿真相机生成完整点云，而现实环境中的相机生成部分点云。根据任务需要，桌面上会放置置物架或搁板。对于定量数据，我们在 Pybullet 中对这一设置进行仿真 [4]。关于仿真设置，请参阅图 6。关于现实世界中的设置，请参阅图7。

B. 任务设置

我们测试了四项任务：（1）抓取类马克杯物体的杯沿，并通过把手将其悬挂在置物架上。（2）抓取类碗物体的杯沿，并将其直立放置在搁板上。（3）抓取类瓶状物体的瓶颈，并将其直立放置在搁板上。（4）抓取放置在任意物体类别物体上的把手。任务 1、2 和 3 分别使用包含普通马克杯、碗和瓶子的演示数据。任务 4 使用普通马克杯的演示数据。

我们将类马克杯物体定义为标准马克杯和装有把手的碗；将类碗物体定义为标准碗、标准马克杯和装有把手的碗；将类瓶状物体定义为标准瓶子和装有把手的瓶子。

每项任务我们提供 10 个示范，并在随机生成位姿、方向和均匀缩放的 200 个未见物体上进行测试。我们假设示范与测试期间环境保持静态，并且能够获取物体的（可能不完整的）点云。在仿真中，我们为每个分布内类别使用 ShapeNet [3]物体，并筛除与任务不兼容的物体。对于分布外物体，我们会根据需要修改 ShapeNet 物体。示例见图。6 。

C. 训练细节

我们使用各系统的占据网络从部分深度图像重建物体，以此预训练 NDF 和语言条件神经距离场。我们在一个联合数据集上对每个系统训练 300,000 次迭代，该数据集包含来自全部三种物体类别的物体，并对其施加随机旋转和平移。对于每个物体，我们将其放入 PyBullet 仿真中并采集深度图像，从而获取点云数据。

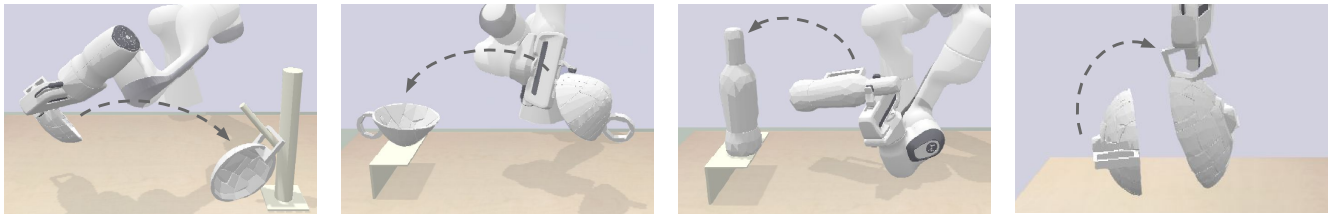
测试时，我们使用单一物体类别，通过少量特定任务示范（仿真中为 10 次，现实环境中为 4–6 次）进行训练。随后，系统利用这些示范，在演示物体类别以及未见物体类别上执行指定任务。

D. 评估指标

在仿真中，我们通过测量抓取成功率（物体与末端执行器稳定接触）、放置成功率（物体以正确朝向与放置表面稳定接触）以及总体任务成功率来评估各方法；总体任务成功要求抓取成功和放置成功均已实现。在实体机器人上，人工评估员判断物体是否已被抓取并放置在正确位置。

E. 基线方法

对于每项任务，我们将 L-NDF 的性能与传统 NDF 和几何方法进行比较。L-NDF 的查询点采用上述启发式方法选取。NDF 的查询点从 [23]提供的代码库中提取。几何方法使用 ICP 和 RANSAC [2, 5]进行位姿估计。



置物架上的“马克杯” 搁板上的“碗” 搁板上的“瓶子” “马克杯”把手抓取

图6：**实验设置** – 我们为每项任务提供十次仿真演示，然后在一组包含 200 个未见物体的集合上分别执行。我们测量抓取成功率、放置成功率和总体成功率。抓取和放置成功率用于检查仿真中的物体是否处于稳定状态。总体成功率用于检查抓取和放置是否都成功。

直立姿态		马克杯演示			碗演示			瓶子演示		马克杯把手演示			
		Mug	碗*	瓶子*	Bowl	碗*	Mug	瓶子	瓶子*	Mug	碗*	瓶子*	Bowl
Geom	抓取	0.945	0.245	0.170	0.670	0.605	0.305	0.675	0.605	0.350	0.375	0.510	0.215
	放置	0.360	0.455	0.330	0.890	0.870	0.625	0.875	0.885	-	-	-	-
	总体 0.3		0.160	0.060	0.590	0.530	0.205	0.625	0.545	0.350	0.375	0.510	**0.215**
NDF	抓取	1.000	0.615	0.010	0.925	0.725	0.265	0.805	0.695	0.805	0.305	0.235	0.000
	放置	0.925	0.620	0.225	0.910	0.730	0.145	0.935	0.870	-	-	-	-
	总体 0.9		0.450	0.000	0.885	0.670	0.125	0.805	**0.665**	0.805	0.305	0.235	0.000
L-NDF	抓取	1.000	0.950	0.160	0.990	0.985	0.970	0.875	0.760	0.980	0.730	0.915	0.190
	放置	0.995	0.830	0.900	0.990	0.990	0.865	0.975	0.670	-	-	-	-
	总体 **0.9**		**0.800**	**0.135**	**0.985**	**0.975**	**0.845**	**0.850**	0.590	**0.980**	**0.730**	**0.915**	0.190
任意姿态		马克杯演示			碗演示			瓶子演示		马克杯把手演示			
		Mug	碗*	瓶子*	Bowl	碗*	Mug	瓶子	瓶子*	Mug	碗*	瓶子*	Bowl
Geom	抓取	0.570	0.170	0.150	0.730	0.690	0.555	0.660	0.570	0.345	0.420	0.440	0.250
	放置	0.345	0.380	0.330	0.905	0.880	0.665	0.850	0.860	-	-	-	-
	总体 0.2		0.075	0.065	0.665	0.615	0.400	0.600	0.525	0.345	0.420	0.440	**0.250**
NDF	抓取	0.900	0.460	0.045	0.675	0.575	0.150	0.575	0.385	0.555	0.105	0.190	0.070
	放置	0.735	0.370	0.235	0.840	0.800	0.565	0.955	0.955	-	-	-	-
	总体 0.6		0.250	0.010	0.655	0.565	0.120	0.570	0.365	0.555	0.105	0.190	0.070
L-NDF	抓取	0.770	0.755	0.110	0.910	0.960	0.880	0.790	0.720	0.930	0.540	0.815	0.130
	放置	0.960	0.635	0.850	0.985	0.940	0.885	0.970	0.820	-	-	-	-
	总体 **0.7**		**0.470**	**0.095**	**0.905**	**0.820**	**0.795**	**0.775**	**0.635**	**0.930**	**0.540**	**0.815**	0.130

表 I： **仿真中未见实例的拾取与放置成功率**。 给定使用单一物体类别的示范，我们测试了模型在多种其他物体类别上的性能。NDF 在来自演示物体类别的未见物体上表现良好，但在新的物体类别上表现不佳。L-NDF 在来自演示物体类别和类似物体类别的未见物体上，无论直立状态还是任意旋转下，都表现良好。这里提供了 ICP 和 RANSAC 的实现（几何）作为参考。绿色表示测试物体与示范所用物体属于同一类别；蓝色表示测试物体来自与示范所用物体类似的类别；红色表示测试物体来自与示范所用物体有显著差异的类别。 *物体经过修改以包含一个把手。各项任务的示意图见图 6。

第六节 实验：结果

我们在仿真中开展实验，以比较几何方法、神经描述符场和语言条件神经距离场在四项任务中的性能（如图6所示），并使用相关的分布内和分布外物体。随后，我们进行消融实验，以考察不同损失函数和不同 3D 特征体积对 L-NDF 性能的影响。最后，我们将 L-NDF 应用于实体机器人，并验证所提出的方法能够在现实世界中泛化到分布外的位姿和物体类别。

A. 仿真实验

分布内物体。我们首先考察如何将技能从示范类别迁移到未见物体上，使其能够适用于新颖的直立位姿或任意旋转的位姿。参见表I的绿色列，我们发现，在所有拾取与放置任务中，L-NDF 的表现都优于传统 NDF，有时成功率甚至提高了 0.25 以上。此外，我们发现，在把手抓取方面，L-NDF 在任意旋转的马克杯把手上取得的任务成功率比 NDF 高出 0.38（表I，最后一列绿色数据）。我们注意到，几何方法确实能够取得一定的任务成功率；然而，其表现不及 NDF 和 L-NDF。我们观察到，NDF 在抓取把手时的失败情况是：虽然找到了靠近目标位置的抓取点，但与预期位置相比存在轻微的偏移或旋转。鉴于 NDF 采用全连接结构，我们推测，观测到的物体显著特征附近的描述子场可能受到物体自身无关几何结构的干扰，而局部场正是针对这一问题提出的。

相似的分布外物体。接下来，我们考虑一个更困难的任务。我们仍希望将在示范中获得的技能迁移到处于新颖直立或任意旋转位姿的测试物体上。然而，现在测试物体具有与示范物体相似的几何形状，但排列不同，或带有混淆特征。参照表I中最前和最后的蓝色列，我们发现，在重新排列的几何形状对任务成功至关重要的任务中，

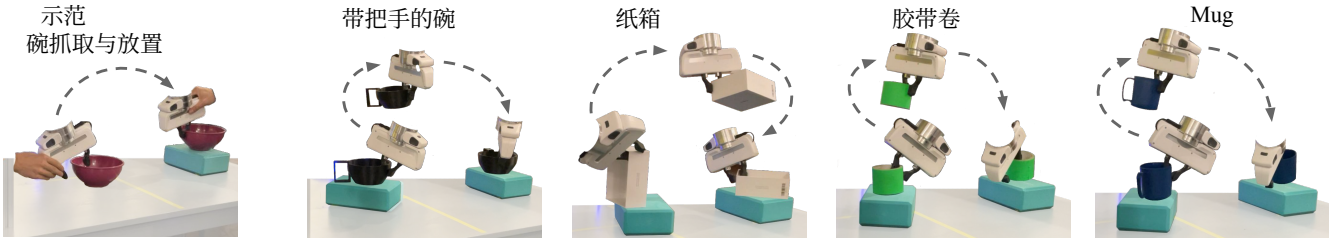


图7：**真实世界执行** – 我们展示了四次在现实世界中抓取和放置两种不同碗的示范。随后，我们使用 Franka Panda 机械臂成功抓取并放置了各种未见物体。更多结果请参阅我们的补充视频。

仅随机			L-NDF占用			硬对比			距离对比		
G	P	O	G	P	O	G	P	O	G	P	O
0.02	0.73	0.02	0.70	0.66	0.47	0.64	0.63	0.39	0.79	0.97	0.78

表 II：**损失函数的影响**。我们测试了一个随机初始化的系统，以及分别采用纯三维重建、简单对比损失和我们提出的基于距离的对比损失进行训练的系统。

32^3			64^3			128^3		
抓取	放置	总体抓取		放置	总体抓取		放置	总体
0.63	0.90	0.56	0.77	0.96	0.75	0.79	0.97	0.78

表 III：**三维特征体积尺寸的影响**。我们研究了三维特征体积尺寸（以体素计）对 L-NDF 性能的影响。所有系统均使用我们的基于距离的对比损失函数进行训练

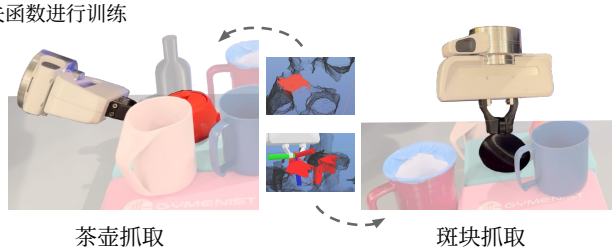


图 8：**在杂乱环境中操作** – 我们展示了四个在整洁场景中抓取马克杯的现实世界演示。随后，我们使用部分点云在杂乱环境中抓取茶壶和 blob。我们使用 Mask R-CNN [11] 进行场景分割。更多结果请参阅我们的补充视频。

NDF 的性能大幅下降。相比之下，L-NDF 的性能也有所下降，但降幅明显小于 NDF。在许多情况下，我们观察到，将 NDF 的查询点调整为与 L-NDF 的查询点更加匹配，可以挽回部分性能损失。然而，我们仍然发现，NDF 的性能落后于 L-NDF 的任务成功率。

在表 I中间的三列蓝色数据中，我们发现，在额外特征起混淆特征作用的任务中，NDF 和 L-NDF 的总体任务成功率下降幅度相近。值得注意的是，对于 NDF，这种性能下降同时归因于抓取成功率和放置成功率的下降。然而，对于 L-NDF，这种下降主要归因于放置成功率降低。我们推测，抓取是一项高度局部化的任务，只关注机械臂手指的位置；而放置则反映了一项全局任务，物体的朝向在很大程度上由其整体几何形状决定。因此，在放置任务中，使用局部场的优势会减弱，混淆特征仍会影响性能。

差异显著的分布外物体。最后，我们通过测试与演示物体类别差异显著的物体，考察 L-NDF 的泛化能力极限。特别值得关注的是：在给定马克杯示范的情况下，将带把手的瓶子放置到置物架上；以及抓取没有把手的碗的“把手”（见表 I 中红色列）。在这些极端情况下，我们发现 NDF 完全失效，成功率微乎其微。L-NDF 的表现稍好，成功率在 10% 至 20% 之间。几何方法也取得了一些成功，在最右侧的任务中表现优于 L-NDF。有趣的是，对于带把手的瓶子，L-NDF 的放置成功率超过了 80%。正如预期的那样，这些总体成功率尚不足以支持通用机器人

操作，但这表明局部场可能是值得探索的方向，有望实现更通用的机器人操作。

B. 消融分析

接下来，我们使用任意旋转瓶子放置任务对 L-NDF 进行消融研究。

损失函数。首先，我们分析损失函数对 L-NDF 性能的影响。在表 II 中，我们发现，随机网络的抓取成功率可以忽略不计，放置成功率也较差。这表明，对 L-NDF 进行预训练非常重要。当相似点的真实相似度为 1、不同点的真实相似度为 0 时，这种简单的对比损失函数同样表现不佳。我们认为，施加这种形式的损失函数无法正确描述我们对网络的目标，因为直观而言，不同的示例点应具有不同的代价。仅在重建任务上进行训练的表现优于简单的对比损失，但在任意旋转下的性能较差。然而，我们基于距离的对比损失显著优于这两种方法，在保持重建质量的同时实现了 SE(3) 等变性。

三维特征体积尺寸。接下来，我们分析体素尺寸对 L-NDF 性能的影响。参见表 III，我们发现任务成功率随三维特征体积尺寸增大而单调提升。将特征体从 32^3个体素增至 64^3 个体素会带来显著提升，而从 64^3 个体素增至 128^3 个体素则能进一步提高成功率，但增幅逐渐减小。我们选择使用 128^3体素系统，因为它的运行时间与 64^3 体素系统相近，同时成功率略高。

C. 现实世界

最后，我们在真实环境中评估系统。我们收集了 5–10 个把手抓取任务示范以及

使用直立物体进行碗抓取与放置的示范，然后在各种任意姿态的未见物体上评估系统。此外，我们在杂乱环境中使用 Mask R-CNN [11] 进行场景分割，并使用 L-NDF 进行位姿估计。正如图 8所示，场景分割得到的点云通常是不完整且含噪的，但 LNDF 仍能成功推断物体姿态。请参见图 1 和图 7了解我们的单物体试验，参见图 8 了解我们在杂乱环境中的评估结果；更多细节和定性结果请访问我们的网站。

VII. 结论

我们提出了局部神经描述符场，这是一种物体表示，能够在测试时通过少样本模仿学习，在可能属于新类别的形状上完成操作任务。我们展示了该方法强大的泛化能力：仅给出抓取马克杯把手的示例，我们便能泛化到茶杯或瓶子等形状，无论是在仿真中还是现实世界中。

VIII. 致谢

我们衷心感谢美国国家科学基金会资助（2214177）；美国空军科学研究办公室资助（FA9550-22-1-0249）；美国海军研究办公室 MURI 资助（N00014-22-1-2740）；美国陆军研究办公室资助（W911NF- 23-1-0034）；MIT-IBM Watson 实验室；以及 MIT 智能研究计划。

参考文献

[1] Michal Adamkiewicz 等。“神经辐射世界中的纯视觉机器人导航”。载于：*IEEE Robotics and Automation Letters*。2022 年。[2] P.J. Besl 和 Neil D. McKay。“一种三维形状配准方法”。载于：*IEEE* 模式分析与机器智能汇刊 14.2 (1992)，第 239–256 页。[3] Angel X Chang 等。“ShapeNet：一个信息丰富的三维模型库”。载于：*arXiv* 预印本 *arXiv:1512.03012* (2015)。[4] Erwin Coumans 和 Yunfei Bai。“PyBullet：用于游戏、机器人学和机器学习物理仿真的 Python 模块”。载于：*GitHub* 仓库 (2016)。[5] Martin A. Fischler 和 Robert C. Bolles。“随机采样一致性：一种用于模型拟合的方法及其在图像分析和自动制图中的应用”。载于：*Commun. ACM* 24.6 (1981)，381–395。[6] Peter R Florence、Lucas Manuelli 和 Russ Tedrake。“稠密物体网络：为机器人操作学习稠密视觉物体描述符”。载于：机器人学习会议。机器学习研究论文集。2018 年，第 373–385 页。[7] Jiahui Fu 等。“基于神经描述符场的鲁棒变化检测”。载于：*arXiv* 预印本 *arXiv:2208.01014* (2022)。[8] Wei Gao 和 Russ Tedrake。“kPAM 2.0：类别级机器人操作的反馈控制”。载于：*IEEE* 机器人与自动化快报 6.2 (2021)，第 2962–2969 页。

[9] Wei Gao 和 Russ Tedrake. “kPAM-SC：利用关键点可供性和形状补全实现可泛化的操作规划”。载于：*arXiv*预印本 *arXiv:1909.06980* (2019).[10] Kensuke Harada 等. “通过圆柱体近似的物体分拣概率方法”。载于：*2013*年*IEEE*机器人与自动化国际会议. IEEE. 2013, 第3742–3747页.[11] Kaiming He 等. “Mask R-CNN”。载于：*CoRR*abs/1703.06870 (2017). arXiv: 1703.06870.[12] Jeffrey Ichnowski* 等. “Dex-NeRF：利用神经辐射场抓取透明物体”。载于：*CoRL*. 2020.[13] Zhenyu Jiang 等. “可供性与几何的协同作用：通过隐式表示实现六自由度抓取检测”。载于：*RSS*. 2021.[14] Lucas Manuelli 等. “kpam：用于类别级机器人操作的关键点可供性”。载于：*arXiv*预印本 *arXiv:1903.06684* (2019).[15] Lars Mescheder 等. “占据网络：在函数空间中学习三维重建”。载于：*CVPR*论文集. 2019.[16] Andrew T Miller 等. “使用形状基元的自动抓取规划”。载于：*2003*年*IEEE*机器人与自动化国际会议（目录号：*03CH37422*）. 第2卷. IEEE. 2003, 第1824–1829页.[17] Arthur Moreau 等. “LENS：通过NeRF合成增强定位”。载于：机器人学习会议. 2022. [18] Joseph Ortiz 等. “iSDF：用于机器人感知的实时神经有符号距离场”。载于：*RSS*. 2022.[19] JeongJoonPark 等. “DeepSDF：学习用于形状表示的连续有符号距离函数”。载于：*CVPR*论文集. 2019.[20] Songyou Peng 等. “卷积占据网络”。载于：*ECCV*论文集. 2020.[21] Hyunwoo Ryu 等. “等变描述子场：用于端到端视觉机器人操作学习的SE (3)等变基于能量的模型”。载于：*arXiv*预印本 *arXiv:2206.08321* (2022). [22] John Schulman 等. “通过非刚性配准从示范中学习”。载于：机器人学研究：第*16*届国际机器人学研究研讨会. Masayuki Inaba 和 Peter Corke 编 . Cham：Springer International Publishing，2016，第339–354页.[23] Anthony Simeonov 等. “神经描述子场：用于操作的SE (3)等变物体表示”。载于：*arXiv*预印本 *arXiv:2112.05124*(2021).[24] Anthony Simeonov 等. “利用神经描述子场进行SE(3)等变关系重排”。载于：机器人学习会议 *(CoRL)*(2022).[25] Edgar Sucar 等. “iMAP：实时隐式映射与定位”。载于：*ICCV*. 2021.[26] Priya Sundaresan 等. “利用在合成深度数据上训练的稠密物体描述子学习绳索操作策略”。载于：*arXiv*预印本 *arXiv:2003.01835* （2020）。

[27] Skye Thompson、Leslie Pack Kaelbling 和 Tomas Lozano-Perez. “通用技能的基于形状的迁移”。载于：国际机器人与自动化会议（*ICRA*）论文集. 2021.[28] Lin Yen-Chen 等. “iNeRF：反演神经辐射场以进行位姿估计”。载于：*IEEE/RSJ*智能机器人与系统国际会议（*IROS*）. 2021.[29] Lin Yen-Chen 等. “NeRF监督：从神经辐射场中学习稠密物体描述符”。载于：*ICRA*. 2022.[30] Youngrock Yoon、Guilherme N DeSouza 和 Avinash C Kak. “利用几何特征对工业物体进行实时跟踪和位姿估计”。载于：*2003*年*IEEE*机器人与自动化国际会议（目录号：*03CH37422*）. 第3卷. IEEE. 2003, 第3473–3478页.[31] Menglong Zhu 等. “用于抓取的单图像三维物体检测与位姿估计”。载于：*2014*年*IEEE*机器人与自动化国际会议（*ICRA*）. IEEE. 2014, 第3936–3943页.[32] Zihan Zhu 等. “Nice-slam：用于同步定位与建图的神经隐式可扩展编码”。载于：*IEEE/CVF*计算机视觉与模式识别会议论文集. 2022, 第12786–12796页。