Any6D：无模型新物体 6D 姿态估计

我们推出了 Any6D，这是一个无模型的 6D 物体姿态估计框架，仅需一张 RGB-D 锚图像即可估计新型场景中未知物体的 6D 姿态和大小。与依赖纹理 3D 模型或多视点的现有方法不同，Any6D 利用联合物体对齐过程来增强 2D-3D 对齐和度量尺度估计，从而提高姿态精度。我们的方法集成了渲染-比较策略来生成和优化姿态假设，在遮挡、非重叠视图、光照条件多样以及跨环境变化较大的场景下均能实现稳健的性能。我们在五个具有挑战性的数据集上评估了我们的方法：REAL275、ToyotaLight、HO3D、YCBINEOAT 和 LM-O，证明了其在新型物体姿态估计方面的有效性，显著优于最先进的方法。

1. 引言

物体 6D 姿态估计是计算机视觉和机器人技术中的关键问题，其重点是确定参考坐标系和相机坐标系之间的刚性 6D 变换，包括 3D 方向和 3D 平移。该任务具有广泛的实际应用，例如机器人操控 [4, 13, 64, 69, 73, 78] 和增强现实 [43, 44, 55]。近年来，该领域取得了显著进展 [3, 13, 20, 33]，相关研究仍在持续进行中。

6D 物体姿态估计的研究大致可分为三种方法：实例级 [22, 51, 62, 68]、类别级 [26–28, 32, 61, 65] 和类别无关方法 [47, 48, 71, 77]。实例级方法精度较高，但也存在明显的局限性。它们依赖于精确的RGB纹理CAD模型来估计物体姿态，这使得它们的有效性仅限于训练过程中看到的物体。它们的主要缺点是无法在不进行额外微调的情况下处理新物体。类别级方法通过使用特定于类别的先验知识部分缓解了这些限制，但它们仍然局限于预定义的物体类别。此外，由于对齐典型姿态的复杂性，它们在获取全面的训练数据集方面面临着巨大的挑战。相比之下，类别无​​关方法旨在推广到不同的物体，而不局限于特定的实例或类别，从而为6D姿态估计提供了一种更灵活的方法。

最近的研究已转向类别无关方法[25, 30, 47, 47, 52, 58, 70, 71]，以解决类别级和实例级姿态估计的局限性。这些努力大致可以分为两个方向：基于模型的方法 [30, 47, 52]，在测试时需要带纹理的 RGB 3D CAD 模型；以及无模型方法 [16, 39, 59, 77]，在推理过程中利用目标物体的多视图参考图或视频序列。尽管这两种方法都表现出良好的效果，但在处理无法物理接触的未见物体时，它们仍然存在显著的实际局限性。例如，在机器人操控场景中，当机器人在新环境中遇到意外物体而没有可用的 3D 模型或多视图图像时，这些方法就会面临困难。这种对先前物体数据的依赖极大地限制了它们在实际应用中的有效性。

为了应对这些挑战，Oryon [11] 提出了一种新颖的无模型方法，用于从单个参考 RGBD 图像估计物体的 6D 姿态。与早期在测试时严重依赖详细物体数据的方法不同，Oryon 使用语言引导，仅使用单个 RGBD 参考即可进行姿态估计。该方法展现出卓越的适应性，即使参考图像和目标图像来自完全不同的环境，也能有效运行。Oryon 利用语言线索建立图像之间的对应关系，并通过对齐参考图像中的可见点云来估计姿态。然而，当参考图像和目标物体之间存在遮挡或极小的重叠区域时，其性能会下降，从而限制匹配点的数量。因此，Oryon 在涉及人类或机械臂的物体操控场景中表现不佳，尤其是在目标被遮挡、出现在非重叠视图中或缺乏足够纹理的情况下，正如我们的实验所示 [15, 67]。

为了克服这些局限性，我们提出了 Any6DPose，一种新颖的无模型方法，使用单个锚点 RGBD 图像进行物体姿态估计。受图像到 3D 模型的最新进展 [21, 36, 41, 72, 74, 76] 的启发，我们的方法估计度量尺度的物体形状以进行物体姿态估计。尽管现有的 3D 生成方法在归一化空间中实现了输入图像和生成的 3D 形状之间良好的照片级真实感一致性，但它们往往忽略了 2D-3D 对齐的关键方面，尤其是度量尺度，这对于准确的姿态估计和后续的下游任务至关重要。因此，我们引入了一种简单而有效的物体对齐方法，通过在 2D 和 3D 空间中进行对齐，共同改进物体大小和姿态估计。在先前研究 [71] 的基础上，我们生成多个姿态假设，并使用渲染和比较策略来选择最优假设。这使得我们的方法能够有效地处理查询对象和锚点对象之间不相交的视图，以及遮挡、不同的光照条件和巨大的跨环境变化。我们在五个公共数据集（HO3D、YCBINEOAT、REAL275、ToyotaLight 和 LM-O）上验证了我们的方法在不同场景下的性能。我们的 Any6DPose 在新型物体姿态估计方面显著优于最先进的方法。

我们的主要贡献如下：

• 我们引入了 Any6DPose，这是一个新颖的框架，仅凭一张参考图像即可对不同场景中的新型物体进行 6D 姿态和大小估计。

• 我们提出了一种简单而有效的物体对齐技术，该技术解决了现有 3D 生成模型的挑战，特别是改进了 2D-3D 对齐和大小估计，从而实现了精确的姿态估计。

• 我们通过大量实验验证了我们的方法，并在五个基准数据集上证明了其优于最先进的方法的性能。

2. 相关研究

2.1. 基于 CAD 模型的物体姿态估计

实例级姿态估计方法 [17, 18, 24, 50, 64] 依赖于特定物体的带纹理 CAD 模型，并在相同实例上进行训练和测试。类别级方法 [8, 10, 31, 65, 79] 可以推广到已知类别中的新实例，但需要昂贵的类别标注，并且仍然局限于预定义的物体类别。类别无关方法 [6, 9, 30] 通过估计不受类别约束的任意新物体的姿态来解决这些限制。然而，大多数方法在推理过程中仍然使用真实 CAD 模型，这限制了它们在无法获得此类模型的实际场景中的实际应用。一些近期的研究 [35, 54] 尝试从现有数据库中检索 CAD。

2.2.无模型物体姿态估计

为了克服 CAD 依赖性的局限性，人们开发了无模型方法，这些方法无需显式纹理模型，而是依赖于目标物体的一组参考图像。Gen6D [39]、OnePose [60] 和 OnePose++[16] 使用运动结构恢复 (SfM) [57] 从视频或多个视图生成 3D 点云，并通过 2D-3D 匹配利用这些点云进行姿态估计 [23, 29]。FS6D [77] 和 FoundationPose [71] 将这种方法扩展到 RGB-D 图像，并取得了良好的效果。然而，这些方法仍然依赖于多视图图像，大多数方法需要相机姿态来合并这些视图，这在实际场景中通常是不切实际的。

受这些挑战的启发，最近的研究致力于减少这些依赖性。例如，NOPE [46] 仅使用单个锚图像来估计相对方向。 LoFTR [59] 也通过使用基于变换器的方法进行特征匹配做出了贡献，有效地增强了单幅图像姿态估计的鲁棒性。同样，Oryon [11] 通过加入语言引导，降低了对单幅图像的要求。然而，这些部分匹配方法 [11、12、34、38、53、59] 在目标图像中存在非重叠区域或严重遮挡的场景下效果不佳。GigaPose [47] 与我们的方法密切相关，它利用图像到 3D（Wonder3D [41]）来估计物体姿态，但仍然需要一个新的初始物体姿态来确定物体大小。同时，HIPPo [40]、OmniManip [49] 和 SceneComplete [1] 正在探索使用图像到 3D 方法从图像中进行无模型物体姿态估计。

3. 方法

给定一张 RGB-D 锚图像 (IA) 和一张 RGB-D 查询图像 (IQ)，我们的任务是估计它们之间的相对姿态。查询图像可能从与单个锚图像截然不同的视点和场景捕捉到同一物体。我们将该问题表述为相对姿态估计任务 [2, 11, 53, 59]。我们的方法旨在估计 IA 和 IQ 之间的相对六维姿态 TA→Q ∈ SE(3)，其中 TA→Q 定义为刚性变换 [R | t]，由旋转 R ∈ SO(3) 和平移 t ∈ R3 组成。

先前的方法尝试使用可见 RGB 图像 [56, 59] 或点云 [2, 53] 进行匹配。虽然这些方法在重叠较大的情况下有效，但在遮挡或视点变化较大的场景下会遇到困难，因为部分到部分匹配缺乏足够的共享特征。为了应对这些挑战，我们的方法采用了一种从全到部分的匹配策略，通过重建完整的三维物体形状，即使在重叠度较小的情况下也能确保稳健的配准。通过渲染和比较流程 [71] 实现准确的姿态估计，有效处理遮挡或部分可见的物体。我们方法的优势已通过大量实验得到证实（第 4.3 节），在低重叠度场景下表现出色。

我们引入了 Any6D 框架，用于估计锚图像 IA 和查询图像 IQ 之间的相对姿态 TA→Q。该框架包含两个部分：首先，我们使用图像到 3D 模型，从锚图像重建归一化物体形状 ON，不考虑真实世界尺度或姿态；然后，通过确定实际物体尺寸 s ∈ R3 和姿态 TOM→A 来估计度量尺度的物体形状 OM，并在二维和三维空间中正确配准（第 3.1 节）。接下来，我们使用重建的度量尺度物体形状和查询图像来估计姿态，通过结合 TOM→A 和 TOM→Q 推导出相对变换 TA→Q（见 3.2 节）。Any6D 框架的完整工作流程如图 2 所示。

图 2. Any6D 框架用于无模型物体姿态估计的概览。首先，我们从图像到 3D 模型重建归一化的物体形状 ON。然后，我们使用提出的物体对齐方法，从锚图像 IA 估计准确的物体姿态和尺寸（见 3.1 节）。接下来，我们使用查询图像 IQ，结合重建的度量尺度物体形状 OM 来估计姿态（见 3.2 节）。

3.1.粗略物体对齐

据我们所知，目前尚无可靠的 RGB-D 单视图度量尺度重建解决方案，能够有效处理各种物体。鉴于基于 RGB 的单视图重建技术 [21, 37, 76] 的最新进展，我们采用 InstantMesh [76]，该方法已在各种物体上展现出良好的效果。然而，一个关键的限制是，3D 物体重建仅生成一个标准化尺度 ON 的形状，每个 XYZ 轴的尺度范围为 [-1, 1]，这意味着生成的网格相对于实际场景的缩放或定位不正确。这一限制阻碍了我们进一步获得精确的姿态对齐，因此我们提出了物体对齐步骤：首先估计物体形状 OM 的粗略尺寸，然后通过联合求解精确姿态 TOM→A 来优化该尺寸。我们的方法涉及在 IA 和 ON 之间的 3D 和 2D 空间中估计和对齐物体形状，包括 Ta ∈ SE(3) 和尺寸 s ∈ R3。

具体来说，我们使用 IA 以由粗到精的方式估计物体尺寸 s。首先，我们通过比较 IA 和 ON 之间从各自物体中心点云数据来初始化粗略的物体尺寸。虽然点均值是一种简单的中心估计方法，但由于部分视点和锚图像中存在噪声点，该方法变得不可靠，如图 3-(a, b) 所示。使用简单的轴对齐边界框也会导致中心估计不准确，因为存在部分可观测性，如图 3-(c) 所示。因此，我们建议使用有向边界框来确定物体中心，如图 3-(d) 所示，这可以为 IA 提供更可靠的粗略中心估计。对于轴对齐，我们将有向边界框与 XYZ 轴对齐。然后，我们对各种旋转角度进行采样，并计算不同角度的 IA 和 ON 旋转边界框之间的 IoU。导致最高 IoU 的旋转和缩放组合用于将 ON 转换为粗略对齐的物体形状，将其更新为初始物体形状 OM′，随后用于精确的姿势和尺寸估计。

3.2. 精细物体对齐

我们的框架旨在确定锚图像 IA 和查询图像 IQ 之间的相对姿态 TA→Q。此过程包含多个步骤，首先重建粗略的物体形状 OM′，然后对其进行细化以获得最终的度量尺度物体形状 OM。给定粗略缩放的初始形状 OM′，我们通过物体和尺寸细化来联合细化姿态和物体尺寸。

为了细化物体形状和姿态，我们借鉴了 FoundationPose [71] 的高效姿态生成、细化和选择能力。然而，它在基于模型的设置中需要真实的度量尺度物体 CAD，在无模型的设置中则需要多张姿态参考图。这阻碍了它在我们考虑的设置中的直接应用，因为在这种设置中只提供了一张 RGBD 参考图。因此，我们开发了一个联合模块，将尺寸估计任务注入到姿态细化过程中。这使我们能够同时可靠地估计度量尺度尺寸和姿态。

我们的细化流程涉及三个主要模块：姿态估计、尺寸估计和轴对齐——它们在一个统一的过程中协同工作，交替执行尺寸细化和姿态细化任务。我们首先使用 OM′ 估计初始姿态，同时细化物体尺寸。在 FoundationPose [71] 中，姿态假设生成的采样仅在 SO(3) 中进行，并未考虑尺寸变化。因此，我们除了 SO(3) 采样外，还额外采样了不同的尺寸。具体而言，尺寸样本沿每个轴在 ∆s ∈[s0, s1] 的范围内绘制（我们根据经验设定 s0 = 0.6，s1 = 1.4）。然后，我们使用 [71] 提供的细化模块细化采样的姿态假设，并对其进行渲染以与查询图像观测进行比较。最佳姿态假设是根据 [71] 中的姿态选择模块指示的比较分数来选择的。确定最佳尺寸后，我们会缩放物体形状并进入姿态优化阶段，其中包括轴对齐步骤，以提高精度。更新后的对齐方式利用了尺寸和姿态的联合估计，从而比传统的基于 IoU 的方法具有更高的精度。

优化物体参数后，我们确定最终的物体姿态 TOM→A，从而实现重建物体形状的精确对齐。利用锚图像 IA、查询图像 IQ 和度量尺度的物体形状 OM，我们通过组合两个变换来估计相对姿态 TA→Q：从物体到锚图像 (TOM→A) 和从物体到查询图像 (TOM→Q)。相对姿态可以表示如下：

对于姿态选择，我们采用两级渲染与比较策略。首先，姿态排序网络通过将渲染视图与裁剪后的观测值进行比较来评估每个假设，并生成一个嵌入向量来量化对齐质量。然后，我们将自我注意力应用于所有假设的连接嵌入，结合全局背景来生成选择最佳姿势的最终分数。