定向万物：从渲染 3D 模型中学习鲁棒的物体方向估计

图 1. 我们介绍了一种用于估计图像中物体方向的新方法，红色轴表示方向，蓝色和绿色轴表示物体的上方和左侧。附录中提供了更多示例。最佳屏幕缩放观看效果。

摘要

方向是物体的一个关键属性，对于理解其在图像中的空间姿态和排列至关重要。然而，从单幅图像中精确估计方向的实用解决方案仍未得到充分探索。在本研究中，我们介绍了“定向万物”，这是第一个旨在在单幅自由视角图像中估计物体方向的专家级基础模型。由于标注数据的稀缺，我们提出从 3D 世界中提取知识。通过开发一个流程来注释 3D 物体的正面并从随机视角渲染图像，我们收集了 200 万张带有精确方向注释的图像。为了充分利用数据集，我们设计了一个稳健的训练目标，将 3D 方向建模为三个角度的概率分布，并通过拟合这些分布来预测物体的方向。此外，我们采用了多种策略来改进从合成到真实的迁移。我们的模型在渲染图像和真实图像中均达到了最佳的方向估计精度，并在各种场景中展现出令人印象深刻的零样本能力（图 1）。更重要的是，我们的模型增强了许多应用，例如复杂空间概念的理解和生成以及 3D 物体姿态调整。

1. 引言

感知单幅图像中的物体属性是计算机视觉的核心问题。当前的视觉基础模型和大型视觉语言模型 (VLM) 在物体识别 [24, 49]、定位 [22, 25]、跟踪 [33, 43] 和分割 [20, 34] 等任务中表现出色。

然而，由于缺乏标注数据，物体方向（这对于理解物体姿态和排列至关重要）的研究仍未得到充分重视。Omni3D [6] 通过整合 3D 物体检测数据实现了 3D 方向预测，但其应用范围仍然局限于特定领域，主要是房间和街道场景，难以推广到各种现实世界场景。

此外，即使是最先进的通用视觉理解系统，例如 GPT-4o [18] 和 Gemini [41, 42]，也难以理解基本的物体方向。因此，它们在基于方向的问题上表现不佳，例如想象物体运动趋势或理解物体的空间关系，如图 2 所示。

从猎鹰（图中左侧人物）的角度来看，美国队长（图中右侧人物）是在他的左边还是右边？

图 2. 理解物体方向对于空间推理至关重要。然而，即使是像 GPT-4o 和 Gemini-1.5-pro 这样的先进 VLM 也尚无法解决基本的方向问题。

本文提出通过渲染 3D 模型来学习各种物体在不同方向下的外观。通过注释这些 3D 物体的正面，我们可以轻松且经济地为每个渲染视图获取精确的方向标签。这一想法提供了可扩展、多样化且易于获取的数据，从而能够开发准确且可推广的方向估计模型。

为此，我们开发了一个数据收集流程，用于自动筛选、注释和渲染 3D 资源 [9]，从而能够以任何所需比例生成可扩展的数据。具体而言，我们利用先进的 VLM [42] 从正交视图中识别 3D 物体的正面，并辅以规范的姿态检测和对称性分析，以简化任务并提高准确性。然后，我们从随机视角渲染图像，使用相对于物体方向向量的方位角和极角，结合相机旋转角度来表示三维方向。

尽管现在已经有了可扩展的方向数据，但训练一个可靠的方向预测模型仍然并非易事。三个角度的直接回归难以收敛，导致性能不佳。为了克服这一挑战，我们将单个角度值重新表述为概率分布，以更好地捕捉相邻角度之间的相关性。通过驱动模型拟合这些角度概率分布，我们简化了学习过程并显著增强了模型的鲁棒性。此外，考虑到渲染图像和真实图像之间的领域差距，我们研究了各种结合现实世界先验知识的模型初始化方法，以及数据增强策略，以改进从合成到真实的迁移。

我们的贡献可以概括为：

• 我们开发了一套可靠且自动化的 3D 物体方向标注流程，并强调了渲染 3D 物体的价值，它能够生成具有精确方向标签且经济高效、多样化且可扩展的图像数据集。

• 我们引入方向概率分布拟合任务作为学习目标，以稳定训练过程并提升泛化能力。

• 我们研究了各种模型初始化和数据增强策略，以提升从合成到真实的迁移。

• 与专业模型 (Cube RCNN) 和领先的 VLM (GPT-4o 和 Gemini) 相比，我们的模型展现出更强大的方向估计能力。

2. 相关工作

2.1. 基于方向的理解

物体方向提供了物体相对于彼此以及相对于观察者（或相机）的位置关系，这对于理解物体姿态和关系至关重要。准确的方向理解在许多高级应用中发挥着关键作用。

在 3D 场景理解中，许多研究 [1, 3, 8] 强调了基于物体方向的空间关系的重要性。 SQA3D [28] 首先描述代理在 3D 场景中的位置和方向，然后让模型根据给定的空间上下文回答问题。EmboidedScan [44] 手动注释 3D 物体的方向，并利用姿态信息描述 3D 空间中物体之间的空间关系。

在 2D 图像领域，理解物体方向对于准确解释 [13, 46] 或生成 [17, 38, 45] 空间关系和属性也至关重要。Goral 等人 [13] 提出了视觉透视取景任务，以评估 2D VLM 理解图像中人物方向和视角的能力，并重点介绍了基于此能力的各种应用。此外，物体相对于相机的方向决定了其在图像中的姿态，这对于区分空间属性（例如汽车前轮和人物左肩）以及复杂的空间关系至关重要。此外，生成具有给定姿态条件的物体对于可控图像生成至关重要 [17, 45]。

尽管物体方向与众多问题和应用密切相关，但在图像中估计物体方向的实用解决方案仍未得到充分探索。我们的工作填补了这一空白，提出了第一个用于物体方向估计的基础模型，该模型在实际场景中表现出了强大的零样本性能。

2.2. 图像中的物体方向识别

一些任务试图在特定条件下或借助额外信息识别图像中的物体方向。

6DoF 物体姿态估计 [14] 专注于检测图像中物体的位置和方向。然而，现有方法需要目标物体的 CAD 模型 [31, 40] 或同一物体的其他参考视图 [10, 12, 29]，这意味着这些方法无法从单个图像推断物体的方向。另一方面，旋转物体检测 [6, 15, 47] 专注于为物体生成可旋转的 2D 或 3D 边界框。Omni3D [6] 统一了多个 3D 物体检测数据集并训练 Cube R-CNN 从单个图像中识别物体的 3D 位置和方向。虽然 Cube R-CNN 在检测 3D 空间中的物体方面表现出一定的能力，但其性能受到 Omni3D 数据范围的限制，Omni3D 的数据范围主要以室内场景和街道环境为主。此外，Cube R-CNN 预测的方向角主要用于旋转 3D 边界框，并不总是与物体的正面对齐。

与上述任务不同，我们的工作侧重于单视角和自由视角图像中物体的 3D 方向估计，并且方向严格与物体有意义的正面对齐。

3. 二维视觉语言模型 (VLM) 中的方向理解

在提出我们的物体方向估计方法之前，我们首先研究在拥有数十亿参数的网络规模图像数据集上训练的二维视觉语言模型 (VLM) 是否具备理解物体方向的能力。

为此，我们推出了 Ori-Bench，这是首个专门设计用于评估二维视觉语言模型 (VQA) 理解物体方向和解决相关问题能力的基准测试。我们手动筛选了总共 200 幅图像，其中 100 幅来自 COCO [23]，100 幅由 DALL-E 3 [5] 生成。为了实质性地评估对物体方向的理解能力，我们将每幅图像水平翻转，生成一对镜像版本，并相应地调整答案。只有当模型在两个版本上都正确回答了问题时，样本才会被标记为已解决。任务分为三类：(1) 物体方向识别（73+73 个样本）：识别图像中物体的方向； (2) 空间局部推理（39+39 个样本）：区分物体具有特定空间含义的部分，例如人的左手和右手；以及 (3) 空间关系推理（88+88 个样本）：从另一个物体的角度想象一个物体的相对位置。

表 1 展示了 GPT-4o、Gemini-1.5-Pro 以及我们的 Orient Anything+LLM 的准确率（详见第 7.1 节）。在基本方向识别任务中，高级 VLM 只能正确解析约 60% 的样本。这种局限性在空间推理和关系任务中尤为明显，强大的 GPT-4o 和 Gemini-1.5-Pro 的表现与随机猜测相似。这项初步研究强调了对精确估计图像中物体方向的基础工具的需求。所有示例均在补充材料中提供。

4. 方向数据收集

方向标注的匮乏是学习通用方向估计的主要障碍。现有的图像标注，通常是标题 [36, 37]、边界框 [23] 或分割蒙版 [20, 50]，很少包含物体方向信息，而手动标注图像中的物体方向极其耗时且成本高昂。