arXiv:2304.01192v1 [cs.CV] 3 Apr 2023

导航至图像指定的物体

Jacob Krantz[1*] Theophile Gervet[2] KarmeshYadav[3] Austin Wang[3]Chris Paxton[3] Roozbeh Mottaghi[3,4] Dhruv Batra[3,5] JitendraMalik[3,6]Stefan Lee[1] DevendraSinghChaplot[3]1俄勒冈州立大学 [2]卡内基梅隆大学 [3]Meta AI [4]华盛顿大学 [5]佐治亚理工学院 [6]加州大学伯克利分校 项目页面与视频:
jacobkrantz.github.io/modular iin

摘要

图像是指定具身智能体应导航至哪个特定对象实例的一种便捷方式。解决此任务需要语义视觉推理，并在未知环境中进行探索。我们提出了一套可在仿真环境和真实世界中执行此任务的系统。我们的模块化方法解决了探索、目标实例重识别、目标定位和局部导航等子任务。我们利用自中心视觉中的特征匹配重新识别目标实例，并将匹配到的特征投影到地图上以定位目标实例。每个子任务均由现成组件完成，无需微调。在 *HM3D* 实例图像导航基准测试上，该系统的表现分别是基线端到端强化学习策略的 *7* 倍、最先进 *ImageNav* 模型的 *2.3* 倍（成功率为 *56%* 相比 *25%*）。我们将该系统部署到移动机器人平台上，并展示了其在真实世界中的有效表现：在家庭和办公环境中的成功率达到 *88%*。

1. 引言

假设你希望机器人助手检查一下你是否把笔记本电脑落在床上了。可以用自然语言描述这项任务："检查我的笔记本电脑是否在床上"，但如果家里有好几张床呢？另一种方法是向机器人展示床的图像。图像是指定特定对象实例的一种便捷方式。例如，在图1 （上）中，机器人从厨房附近出发，看到一张床的图像，并必须导航到那张特定的床。图像目标导航任务需要推理场景中物体之间的关系（例如，区分外观相似的实例），并通过高效探索来发现目标所在位置（例如，在寻找床的过程中进入卧室）。



图 1：**上：InstanceImageNav**要求智能体导航至目标图像所描述的对象实例。**下：真实世界部署**。我们的方法在仿真环境中表现领先，并能迁移到真实环境中。在这里，我们找到了 10 米外的一把椅子。视频请见项目页面。

*在 Meta AI 的 FAIR 实验室实习期间完成。联系邮箱：krantzja@oregonstate.edu

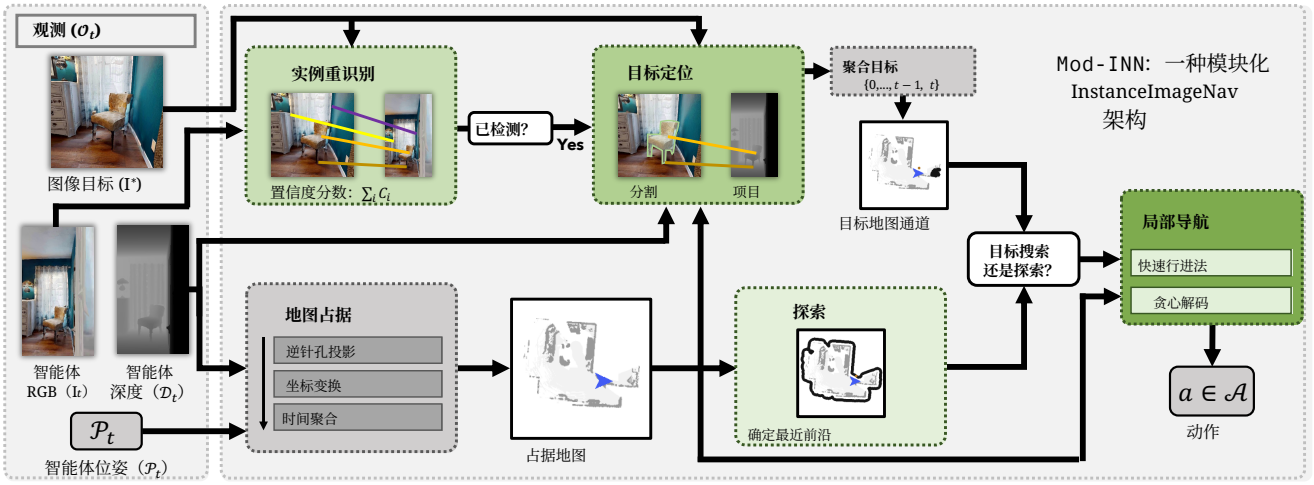


图 2: **模型概览**。 我们采用 探索 基于前沿的探索，采用 实例重识别 特征匹配，采用 目标定位 基于掩码的特征投影，并采用 局部导航 解析规划。子任务模块以 绿色 表示，支持组件以灰色表示。

本文提出一种导航系统，能够在真实环境中可靠地执行这种基于实例的图像目标导航任务。具体而言，我们提出一个由探索、实例重识别、目标定位和局部导航组成的模块化框架。我们通过简单组合现成组件来实现该框架，无需进行零微调。图2所示，我们的系统采用基于前沿的探索策略，通过特征匹配重新识别目标，利用投影后的特征匹配结果定位目标，并使用解析式规划器规划路径。在具有挑战性的 Habitat-Matterport3D (HM3D) [32] 实例图像导航基准 [24], 上，我们取得了56% 对比 25% 的成功率，优于表现最佳的基线。我们将系统部署在移动机器人平台上，并在两个真实环境中进行测试，取得了88% 的成功率（例如，图1 下）。

大多数以往研究在处理图像目标导航（ImageNav）时[53, 10, 19, 29, 1, 28, 49]都假设目标图像是在环境中的随机位姿下采集的，并且始终与智能体的相机参数相匹配。正如 Krantz，等人所 [24], 指出的，这种表述可能导致图像目标含糊不清（例如，拍摄到不起眼的墙面），且与潜在的用户应用脱节。为解决这些问题，Krantz，等人提出的 [24]基于实例的图像目标导航任务（InstanceImageNav）具有两个关键特性：（1）目标图像呈现一个对象实例；（2）目标图像不受智能体实体形态影响。一个端到端强化学习策略基线仅取得 8.3% 的成功率。我们的系统表现比它高 7 倍。为进一步与以往工作进行比较，我们在 InstanceImageNav 上评估了一种先进的图像目标导航方法 [48] 。尽管该方法优于基线，我们的系统仍比它高 2.3 倍。

许多先前的导航方法，例如上面提到的两种方法，都涉及通过强化学习或模仿学习训练端到端的传感器到动作策略。尽管端到端方法易于实现且适用于多种任务，但也存在样本复杂度高、 [24], 过拟合 [46], 以及仿真到现实迁移效果不佳 [18]等局限性。另一种方法是将导航分解为多个子任务，并通过约束更强的技能来解决 [8, 9, 17, 18]。这种模块化范式能够提高样本效率，并改善现实世界执行效果 [31, 18, 14]。我们的方法展现了这些优势：样本复杂度低（零微调），且在现实世界中的表现出色（图1 下），同时在仿真基准测试中取得了顶尖表现。

总而言之，我们致力于解决实例特定图像目标导航这一重要且具有挑战性的任务。我们提出了一种模块化方法，在 HM3D InstanceImageNav 基准上取得最佳成绩，并优于最先进的 ImageNav 模型。我们将系统部署在移动机器人平台上，并展示了它能够在室内真实环境中可靠运行。

2. 相关工作

图像目标导航。 在具身导航中，目标可以通过坐标指定 [2]，也可以通过各种形式的语义描述指定 [6, 53, 11, 3, 25]。图像目标导航（图像导航）是后一种形式之一；智能体根据图像提供的视觉描述进行导航。图像导航通常在此前未见过的环境中进行研究 [53]，其中目标图像从场景各处随机采样 [10]。许多研究关注室内环境，也有一些研究关注室外环境 [38, 39, 40]。许多解决图像导航问题的方法采用深度强化

学习（DRL），以学习将自我中心视觉映射为动作的端到端策略 [53, 1, 28, 49]。然而，与视觉场景理解、语义探索和长期记忆相关的技能往往难以通过端到端方式学习。因此，这些方法通常结合精心设计的奖励塑形 [12], 预训练流程 [49],以及先进的记忆模块 [34, 29, 46]。与端到端 DRL 不同，有些方法将任务拆分为可通过监督方式学习的子任务，例如通过拓扑同步定位与建图进行图预测 [10], 、基于图的距离学习 [19, 39], 以及用于最后一公里导航的相机位姿估计 [44]。在这项工作中，我们研究目标图像呈现对象实例的图像导航任务（实例图像导航 [24]）。我们探究在不进行微调的情况下，采用纯模块化方法能将该基准测试的性能提升到何种程度，并证明其性能优于最先进的端到端策略。

语义导航的模块化方法。在此前未见过的环境中进行导航的经典方法包括构建几何地图并对智能体进行定位（ SLAM [16]）。语义导航的模块化方法将高层任务分解为多个组件，这些组件可以利用经典导航流程，也可以借助现代视觉系统（例如物体检测器）来解决。[20]一项相关任务是物体目标导航（ObjectNav [6]）：智能体会收到一个物体类别（例如，盆栽植物），并必须导航至该类别的任意一个实例。Chaplot 等人 [9]将 ObjectNav 任务分解为探索、物体检测和局部导航。在此基础上，CLIP on Wheels（CoW [17]）采用了探索与物体定位的分解方式，以应对开放集物体词汇表。模块化方法在将系统从仿真环境迁移到真实环境（仿真到现实迁移）方面也大有可为。Gervet 等人 [18] 对模块化和端到端 ObjectNav 系统进行了仿真到现实迁移，发现模块化避免了导致端到端策略性能下降的视觉仿真到现实差距。在本研究中，我们扩展了模块化导航智能体的能力，使其能够处理由图像指定的导航目标。我们提出了一种可通过模块化组件解决的问题因子分解方法，并在仿真和真实环境中验证了其有效性。

实例重识别。对象实例重识别（OIRe-ID）旨在判断给定图像中的对象是否与锚点图像中的对象相同。OIRe-ID 通常通过图像检索任务来实现 [5, 4]。在深度学习革命之前，图像检索及相关视觉识别任务主要基于局部特征描述子 [26] ，例如尺度不变特征变换 [27] 或方向梯度直方图 [13]。一种奠基性物体检索方法是将文本检索方法应用于这些局部特征，从而形成视觉词袋模型（BoVW [41]）。依赖局部特征的模型受到特征表示表达能力的限制。因此，一些现代方法利用深度网络更新这一技术体系，以生成描述特征（例如 SuperPoint [15]）并进行匹配（例如 SuperGlue [33]）。此外，研究人员还在 OIRe-ID 领域提出了使用 Transformer [43] 和极线引导 Transformer [7]的重排序方法。在这项工作中，我们提出解决具身 OIRe-ID 问题，通过自我运动获取新的实例视图。我们证明，基于 SuperGlue 的关键点方法不仅能够实现重新识别，还能帮助定位该实例。

3. InstanceImageNav 任务设置

我们采用 Krantz 等 [24]提出的任务设置。实例特定图像目标导航任务（InstanceImageNav）将智能体随机放置在一个尚未探索的室内环境中的某个位姿，并要求智能体导航至 RGB 图像中作为主要主体呈现的特定对象实例。我们的任务设置与 2023 年 Habitat 导航挑战赛的 ImageNav赛道[1] [47] 保持一致，以便进行清晰的比较，并实现从仿真环境到真实环境的平滑迁移。

观测空间。 在每个时间步 t，智能体的观测 $\mathcal{O}_t$ 由 RGB 目标图像 $\mathcal{I}^*$、自中心 RGB 图像 $\mathcal{I}_t$、自中心深度图像 $\mathcal{D}_t$以及智能体相对于起始位姿$\mathcal{P}_0 = (0, 0, 0)$ 的位姿 $\mathcal{P}_t = (x, y, \theta)$ 组成。总体而言， $\mathcal{O}_t = (\mathcal{I}^*, \mathcal{I}_t, \mathcal{D}_t, \mathcal{P}_t)$。采集 $\mathcal{I}^*$ 的相机与采集$\mathcal{I}_t$的相机相互独立。$\mathcal{I}^*$ 的相机参数经过采样，以反映真实部署场景，包括外参（高度、俯仰角、与对象的距离）和内参（视场角）。更多详情请参见 Krantz 等 [24]。$\mathcal{I}^*$ 是一张 512×512 RGB 图像。智能体的自中心感知与 Intel RealSense 相机相匹配： $\mathcal{I}_t$ 和 $\mathcal{D}_t$ 均与 640×360 对齐，且水平视场角为 $42°$。

动作空间。 我们采用离散动作空间 $\mathcal{A}$ ={前进、左转、右转、停止}，其中前进会使智能体移动。25米，转向指令会使智能体旋转 30°。

成功判据。 如果智能体在距离 $\mathcal{I}^*$ 所描绘的对象实例 1。0 米的欧几里得距离内执行停止动作，则该 InstanceImageNav 回合成功。通过转动智能体并向上或向下看，目标实例还必须能够从预言机视角看到。

实体化。我们使用 Hello Robot 的 Stretch 执行现实世界任务[2]。在仿真实验中，我们将该实体化建模为一个刚体零转弯半径圆柱体，高度为1.41米，半径为0.17米。前向 RGB-D 相机安装在1.31米高处。

[1] aihabitat.org/challenge/2023
[2] hello-robot.com/stretch-2

4. 方法

我们提出将 InstanceImageNav 任务分解为可分别处理的子任务。具体而言，我们划分出探索、目标实例重新识别、目标定位和局部导航，通过综合完成这些子任务来解决 InstanceImageNav。我们将这些子任务描述如下。

探索。 在此前未知的环境中寻找对象实例需要进行探索——目标的位置和环境的地图均未知。在 InstanceImageNav 中，目标由 $\mathcal{I}^*$ 描述，成功导航意味着在某个时间t 观察到一个 $\mathcal{I}_t$，且其与 $\mathcal{I}^*$具有较高的语义相似度。高效访问环境中可能提供这一视角的位置，即最大化可观测空间的覆盖率，有助于实现成功导航。

目标实例重识别。 如果探索策略实现了 100% 的可观测空间覆盖率，那么至少存在一个时间步 t，使得与之对应的图像$\mathcal{I}_t$ 描绘了由 $\mathcal{I}^*$描述的目标实例。因此，目标实例重识别就是二分类器 f_{ID}，用于回答“图像中描绘的对象是否在 $\mathcal{I}_t$中可见？”具体而言，

$$\hat{y}_t = f_{ID}(\mathcal{I}^*, \mathcal{I}_t). \tag{1}$$

这项任务需要利用前景和背景，从新颖视角重新识别对象，并已在对象实例重识别（OIRe-ID）文献中得到广泛研究。

目标定位。 智能体识别出目标实例时，可能距离它很远，因此仅调用停止动作不足以成功。因此，必须利用 $\mathcal{I}^*$ 与 $\mathcal{I}_t$ 之间的配对关系来定位目标。目标定位，f_{GL}，将配对的 RGB 图像和自中心深度映射为目标实例相对于智能体当前位姿的位置：

$$\left(\mathcal{P}_G^{(x,y,\cdot)} - \mathcal{P}_t^{(x,y,\theta)}\right) = f_{GL}(\mathcal{I}^*, \mathcal{I}_t, \mathcal{D}_t). \tag{2}$$

局部导航。 局部导航是实现探索和前往目标实例的基础。我们考虑一种局部导航策略 π，它将相对极坐标目标（r，θ）映射为动作序列$\{a_0, a_1, \ldots, a_n\} \in \mathcal{A}$。 π可以基于地图和/或自中心视觉进行条件化。

4.1. 提出的模块

我们实例化了一个系统，将上述因子分解付诸实践：`Mod-IIN`。具体而言，我们提出了一种纯模块化方法，无需任何重新训练或微调组件即可执行 InstanceImageNav。这种方法与当前主流的端到端学习式图像导航策略截然不同，却在仿真环境和真实环境中展现出令人瞩目的效果。我们在图2中对该模型进行了可视化。

探索。我们采用一种基于前沿的探索（FBE [51]）策略，该策略在自上而下的二维地图上运行，追踪占用状态、自由空间和前沿区域；前沿区域标示已探索区域与未探索区域之间的边界。每个时间步都会利用自中心深度（$\mathcal{D}_t$）和位姿（$\mathcal{P}_t$）的逆透视投影更新地图。FBE 会贪心地选择地图中最近的前沿区域作为导航目标。到达该前沿区域后，便重复这一过程，选择下一个最近的前沿区域。该策略能够高效扩大覆盖率，并已证明在仿真环境和真实环境中均有效 [18]。

目标实例重识别。 我们采用一种基于关键点的重识别方法，该方法以目标图像和自我中心图像为条件进行二元分类。首先，我们提取目标图像中关键点$K^* \in \mathbb{R}^{n\times 2}$ 的逐像素（x，y）位置及其对应的向量描述 $V^* \in \mathbb{R}^{n\times 256}$。我们使用 SuperPoint [15], 提取这些特征。SuperPoint 是一种单次前向传播的卷积神经网络，采用单应性自适应这一自监督一致性方法进行训练。我们对自我中心图像重复这一过程，得到 $K_t \in \mathbb{R}^{m\times 2}$ 和 $V_t \in \mathbb{R}^{m\times 256}$。具体而言：

$$(K^*, V^*) = \text{SuperPoint}(\mathcal{I}^*) \tag{3}$$

$$(K_t, V_t) = \text{SuperPoint}(\mathcal{I}_t) \tag{4}$$

随后，我们计算 (K^*，V^*) 与 (K_t V_t $K^* \in \mathbb{R}^{w\times 2}$ $K_t \in$,) 之间 $\mathbb{R}^{w\times 2}$ 的对应关系，得到i^{th}匹配子集 (ˆ , ˆ)，其中 ˆ 的关键点与 ˆ K_t 的 K^* i^{th} 关键点对应，且 $w \leq \min(n,\ m)$。每对关键点都有一个匹配置信度分数 $0 \leq C \in \mathbb{R}^w \leq 1$。我们使用 SuperGlue [33], , 这是一种通过最优传输优化部分匹配分配的图神经网络（GNN）：

$$(\hat{K}^*, \hat{K}_t), C = \text{SuperGlue}((K^*, V^*), (K_t, V_t)) \tag{5}$$

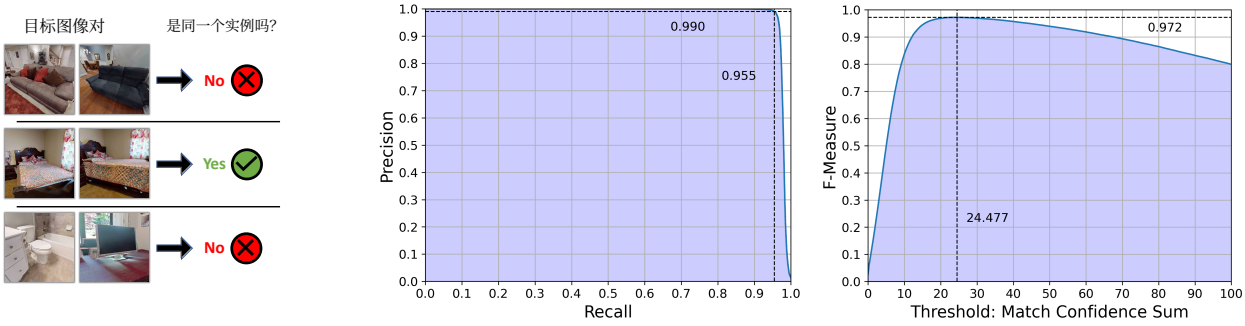
然后，我们可以通过对置信度分数之和进行阈值化，将此输出转换为二分类器：

$$\text{sum}(C) \geq \tau \tag{6}$$

其中，正向判定表示已重新识别目标实例，而 τ 是选定的阈值。该分类器在每个步骤 t 对 ($\mathcal{I}^*$，$\mathcal{I}_t$) 进行评估。对于特征提取（SuperPoint）和特征匹配（SuperGlue），我们均使用未经下游微调的现成模型。

确定检测阈值 τ。如果 τ过低，未观察到目标的自我中心图像会被错误地传递给目标定位（假阳性过多）。如果τ过高，探索过程中观察到目标的步骤就会被漏掉（假阴性过多）。为了在这两种错误之间取得平衡，我们收集目标图像对组成的数据集，并选择能使分类器 F 值最大的阈值。具体方法如下。

我们从 InstanceImageNav 回合数据集的训练集划分中抽取一组对象实例。



(a) 分类数据集　(b) 精确率-召回率曲线　(c) F度量曲线

图 3：为了确定实例重识别（Re-ID）阈值 τ，我们收集了一个目标图像对数据集（a）。我们的 Re-ID 方法计算每对图像的特征匹配分数总和，进而据此绘制精确率-召回率曲线（b）和 F 值曲线（c）。我们选择 F 值最高时对应的 τ（$\tau = 24.5$）。

每个物体都有 1 到 50 张展示它的目标图像。我们提取 3 % 的物体实例及其对应的目标图像，得到 121 个物体和 2270 张目标图像（平均每个物体 18.8 张图像）。我们使用这些数据构建成对数据集，其中输入由两张图像 $(\mathcal{I}_a, \mathcal{I}_b)$ 组成，输出 y 则表示这些图像是否观测到同一个物体：$((\mathcal{I}_a, \mathcal{I}_b)\ y) \in D$。随后，我们对所有图像对运行分类器，生成一组置信度分数。这样便可计算分类器的精确率-召回率（PR）曲线和 F度量曲线（图 3）。最后，我们选择使 F 值达到最大的阈值（即精确率与召回率的调和平均数）。我们在补充材料 A. 中验证了该阈值在经验上是最优的。

目标定位。观察到目标实例后，我们需要在世界中对其进行定位（式2）。根据之前的ˆ ˆ重新识别步骤，我们得到了目标图像与自我中心图像之间的对应关系$(K^*,\ K_t)$。利用与自我中心图像对齐的深度图 $\mathcal{D}_t$，我们可以通过逆透视投影将这些点投影到世界中。然而，并非所有匹配关键点都位于我们要导航到的目标实例上；与背景和相邻物体相关联的视觉特征可能导致智能体停在错误的位置。因此，问题在于“哪些匹配关键点应作为目标靶点进行投影？”。

我们对目标图像进行实例分割，以确定目标实例的掩码——掩码内的关键点可映射到自我中心坐标系，然后进行投影。我们使用现成的预训练网络（Detic [52]）进行实例分割。我们选择 Detic，是因为它是一款高性能检测器，可通过 CLIP [30]对概念进行编码，从而支持开放词汇。使用 Detic 后，我们的方法不受固定物体类别的限制，还可扩展到描绘任意物体的图像目标。为确定目标实例掩码，我们选择包含图像中心点且置信度最高的掩码。

ˆ K^* 中所有位于此掩码内的关键点都会映射ˆ到 K_t 中的对应关键点，并投影到世界坐标系。目标点的点云沿高度维度合并，并体素化为一个二维地图通道。随后，将此通道与探索过程中构建的地图拼接，并将其作为导航目标。

局部导航。局部导航器必须前往前沿点（探索）和目标点（目标定位）。这两种情况都可以按如下方式处理：给定部分占据地图和一组目标点，在智能体动作空间 $\mathcal{A}$ 中规划一条路径，使智能体到达最近的可达点。我们采用基于快速行进法（FMM [37]）的增量式路径规划器来解决这一问题，该方法由 [8] 提出，并在后续工作 [9, 19, 23]中使用。

5. 实验

我们在仿真环境和真实环境中评估模型。在仿真环境中，我们将其与已有方法及不同的子任务模块进行比较（第5.2节）。我们分析失败模式（第5.3节），并讨论一个定性示例（第5.4节）。最后，我们展示成功的仿真到现实迁移（第5.5节）。

5.1. 实验设置

仿真实验。我们的仿真实验遵循为 2023 Habitat 导航挑战赛 ImageNav赛道提出的任务定义和数据集 [24]。我们基于 Habitat 模拟器构建并评估智能体 [35, 42]。回合数据集遵循 Krantz 提出的生成流程，等人。 [24] 针对 InstanceImageNav，并额外要求任何回合都不需要进行多楼层导航。该数据集所用场景来自带有语义标注的 Habitat-Matterport3D数据集 [32]（HM3D-SEM [50]）。HM3D-SEM 中的 216 个场景按 145/36/35 划分为训练集/验证集/测试集；InstanceImageNav 回合则按以下数量划分为训练集/验证集/标准测试集/挑战测试集：

7,056K/1K/1K/1K。目标图像中描绘的物体实例所属类别包括以下之一：{椅子、沙发、植物、床、马桶、电视}。我们使用由 795 个独特对象实例组成的验证集划分。

评估指标。 我们从成功率和效率两方面评估模型。具体而言，我们报告成功率（SR）、按路径长度倒数加权的成功率（SPL）、导航误差（NE，以及在部分分析中报告最大步数（最大成功轨迹长度）。若智能体调用停止时距离目标的欧几里得距离不超过 1.0 米，且通过转身或向上/向下看，目标物体处于预言机可见状态，则判定为成功。按路径长度倒数加权的成功率 是一种效率指标，其定义见 [2]，并针对目标视点作了修改，见 [6]。导航误差 是智能体停止位置与最近目标视点之间的测地距离。最大成功轨迹长度 是使用完整步数预算（1000）的回合所占百分比。

基线。 我们将智能体与以下模型进行比较：

- IIN 强化学习基线：InstanceImageNav（IIN）基线模型 [24]是一种端到端的传感器到动作循环策略，由用于目标图像、自我中心图像、自中心深度和位姿的单模态编码器组成。该模型从头开始训练，使用近端策略优化（PPO [36]）和可变经验展开（VER [45]），通过强化学习利用 35 亿步经验进行训练。

- OVRL-v2：离线视觉表征学习 V2（OVRL-v2 [48]）是一种无模型、端到端的语义导航策略，采用 ViT+LSTM 架构和自监督视觉预训练。截至本文撰写时，OVRL-v2 在此前的 ImageNav 基准上取得了最先进的性能，在 ObjectNav 基准上的表现也接近最先进水平。我们在 InstanceImageNav（OVRL-v2-ImageNav）上对预训练模型进行零样本评估。为进行公平比较，我们随后采用 [48]在其 ImageNav 实验中使用的相同训练流程和参数，对该预训练模型进行微调，使其适用于 InstanceImageNav（OVRL-v2-IIN）。

消融实验：目标实例重识别。我们将目标实例重识别方法与基线方法进行比较。我们考虑以下方法：

- 关键点置信度：这是我们在第 4.1 节中描述的主要方法。简而言之，我们从目标图像和自我中心视角图像中提取关键点并进行匹配。随后，对应关系置信度分数之和经阈值处理后转换为二元分类。

- 关键点匹配：此方法与上述方法相同，但采用了不同的分类策略；它对匹配的关键点对数量设定阈值，而不是对置信度总和设定阈值。这使我们能够测试匹配置信度是否能提供额外的判别价值。

#	模型	验证	
		导航误差↓	路径长度加权成功率↑
1	IIN 强化学习基线	6.3	.035
2	OVRL-v2-ImageNav	6.9 0.0	02
3	OVRL-v2-IIN	5.0	.118
4	Mod-IIN（我们的方法）	**3.1**	**.233**

表 1：我们将模块化实例图像导航方法（Mod-IIN）与先前的 InstanceImageNav 方法进行了比较。Mod-IIN 的表现优于基线模型（第 1 行），成功率（SR）提高了 6.8 倍；同时也优于当前最先进的图像导航模型（OVRL-v2，第 3 行），提高了 2.3 倍。

- *ResNet*：此方法完全不使用关键点，而是通过 ResNet 对目标和自我中心视角图像进行编码，ResNet50 [21] 在 ImageNet 上预训练。我们计算所得特征向量之间的余弦相似度并设定阈值，使得高于τ 该阈值的余弦相似度意味着重新识别出目标实例。

- *CLIP*：该方法与 ResNet 相同，但使用经过对比学习训练的模型对图像进行编码：CLIP [30]。

- 预言机：该方法通过在智能体的自我中心坐标系中向仿真环境查询预言机实例掩码，为实例重识别子任务提供上界。如果掩码中的任意像素与目标物体的 ID 匹配，则视为检测为正。

上述所有方法（预言机除外）都会对某个标量值应用阈值来执行检测。每个阈值均通过第 4.1 节所述的最大 F 值方法确定。我们在补充材料 B 中给出了每种方法的 PR 曲线和 F值曲线。

消融实验：目标定位。 我们将目标定位方法与基线方法进行比较。我们考虑：

- *Detic* 投影：这是我们在第 4.1 节中介绍的主要方法。简而言之，位于目标实例掩码（由 Detic 计算）内且相匹配的目标图像关键点会被投影到目标地图通道中。

- *Crop-Projected*：该方法通过 Detic 消融实例分割。将位于目标图像静态中心裁剪区域内的匹配关键点进行投影。我们在补充材料中进一步介绍该方法。C。

- *Detic-ObjectNav*：该方法消融了关键点。Detic 推断目标物体的物体类别。重识别成功后，Detic 会识别自我中心图像中属于推断类别的所有像素。这些像素会被投影，智能体随后导航至该类别中最近的物体。该方法采用模块化目标导航的语义分割形式 [9, 18]。

#	模型变体 实例重识别	目标定位	验证 NE ↓	最大成功轨迹长度 ↓	SR ↑	路径长度加权成功率 ↑
1	关键点置信度	Detic 投影	**3.096**	**0.310**	**0.561**	**0.233**
2	关键点匹配	Detic 投影	3.261	0.339	0.539	0.221
3	关键点置信度	Crop-Projected	3.128	0.301	0.523	0.224
4	关键点置信度	Detic-ObjectNav	3.235	0.327	0.488	0.205
5	ResNet	Detic-ObjectNav	6.264	0.857	0.138	0.044
6	CLIP	Detic-ObjectNav	6.570	0.917	0.097	0.029
7	预言机	预言机	1.300	0.089	0.845	0.453
8	预言机	Detic 投影	2.551	0.116	0.498	0.250
9	关键点置信度	预言机	2.914	0.348	0.647	0.266

表 2：我们的模块化 InstanceImageNav 方法（Mod-IIN：第 1 行），以及在实例重新识别和目标定位方面的不同变体。

- 预言机：该方法为目标定位子任务提供了一个上限。检测成功后，所有观测到目标物体实例的自我中心像素（依据预言机实例掩码）都会被投影。

5.2. 仿真结果

我们围绕关键观察来讨论结果。

Mod-IIN优于已有方法。我们的方法在性能上优于基线 InstanceImageNav 方法（表1 第 1 行与第 4 行），成功率提高了近 7 倍（0.561 成功率与0.083 成功率）。Mod-IIN 无需微调，而端到端的IIN 强化学习基线则在 64 个图形处理器上，经过 35 亿步的训练。

在 InstanceImageNav 上对 OVRL-v2 进行零样本评估（表 1 第 2 行）时，性能很低（0.006 成功率）。造成这一结果的因素包括场景数据集不同（Gib- son 与 HM3D）、实体化不同（LocoBot 与 Stretch），以及目标目的地不同（图像来源 与 图像主体）。针对 InstanceImageNav 训练的 OVRL-v2（OVRL-v2-IIN）表现更好，0.248 成功率（表 1第 3 行），然而尽管进行了视觉预训练，仍表现出明显的过拟合，集中于训练集划分（0.850 成功率）。我们的方法在验证集上的表现提升了 2.3 倍（表 1 第 4 行与第 3 行）。

关键点置信度比匹配数量更具判别力。对匹配置信度分数之和进行阈值筛选（关键点置信度），相比对匹配关键点数量进行阈值筛选（），成功率提高了 0.022（关键点匹配）（表 2 第 1 行与第 2 行）。通过比较两个分类器在成对目标图像数据集上的最大 F 值，也能验证这一结果；关键点匹配的最大 F 值为 0.972，而 关键点置信度的最大 F 值为 0.962。

目标定位受益于实例分割。替换实例掩码 (

Detic 投影）采用中心裁剪（*Crop-Projected*）会使成功率降低 0.038（表 2 第 1 行 与第 3 行相比）。除了性能更差之外，中心裁剪方法还利用了数据集中的尺寸偏差，而这一偏差可能无法泛化。

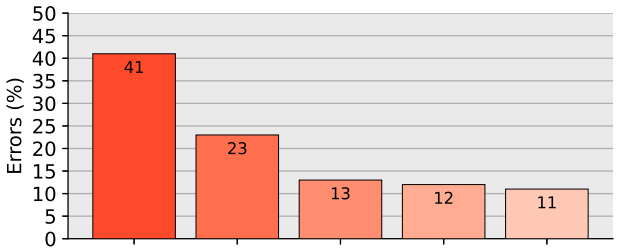


图 4：Mod-IIN 失败模式的分布。

基于关键点的定位优于基于类别的定位。两种匹配关键点投影的方法——*Detic* 投影和 *Crop-Projected*——均优于基于类别的定位（*Detic-ObjectNav*）（表 2第 1、3 行 与第 4 行相比）。ObjectNav 中常用于目标定位的语义分割方法无法直接应用于 InstanceImage-Nav。这验证了：在室内环境中导航至对象实例，从根本上要求能够区分同一类别的不同物体实例。

现成图像编码器无法实现实例重识别。ResNet 和 CLIP 均无法区分物体实例的不同视角。这一点在表 2 的第 5、6 行中显而易见，其中 最大成功轨迹长度分别为 86% 和 92%，因此成功率分别为 14% 和 10%。我们发现，降低 ResNet 的检测阈值并未提升性能。

5.3. 失败分析。

在表 2的第 7 行中，我们使用完美的实例重识别和目标定位来评估智能体，这意味着任何失败都可归因于通过 FBE 进行探索和局部导航。由此得到的 0.845 成功率 是系统中处理实例重识别和目标定位时的性能上限。在第 8 行中，我们采用预言机实例重识别和预测定位。性能并未优于完整模型。我们认为，这可能是因为检测与定位系统都依赖关键点匹配；假设检测的匹配置信度较低，就会导致定位效果不佳。最后，第 9 行显示，我们的目标定位方法与完美目标定位之间存在 0.086 的 成功率 差距（第 9 行与第 1 行）。

我们对Mod-IIN 在图4中的失败模式进行了定性分析。我们将随机抽取的100个失败回合的失败原因标记为以下类别之一：

实例重识别：假阴性（41%）。最常见的情况是，智能体观察到了目标，却未能检测到它。改进重识别方法或许能够缓解这一问题，但并非所有新颖的实例视图都同样具有挑战性；如果探索产生的视图与目标图像更相似，重识别就会更容易。

实例重识别：假阳性（23%）。错误检测

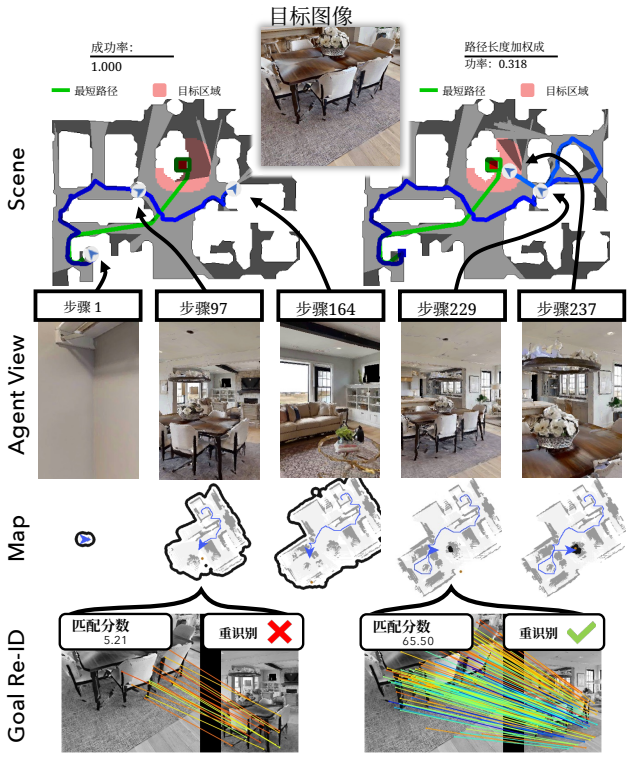


图 5：我们的 Mod-IIN 智能体在 Habitat 模拟器中执行 InstanceImageNav 任务的定性示例。

通常源于与目标图像背景正确匹配的视觉特征。通过进一步推理背景与前景，可以缓解这种失败模式。

探索错误（13%）。探索未能在规定的时间预算内生成目标实例的视图。

定位错误（12%）。目标定位错误是由于投影了不属于目标实例的点所致，原因可能是对应关系错误或目标掩膜不准确。

局部导航错误（11%）。局部导航偶尔无法规划出通往目标的路径。

5.4. 定性示例

我们在图中展示了智能体在仿真环境中执行一个回合的定性示例。5。在步骤1、97和164，智能体正在探索环境。步骤229时检测到目标，并将关键点投影为目标靶点。智能体在步骤237到达目标（一把餐椅）后，调用停止。这个回合展示了较长的探索过程和具有挑战性的目标实例；场景中存在多个相同的物体，但智能体仍导航到了正确的物体。这得益于我们的实例重识别方法匹配帧中的所有关键点，以及我们的定位方法仅投影那些描绘目标的关键点。我们在项目页面提供了智能体在仿真环境中的视频。



图 6：我们将 Mod-INN 部署到 Hello Robot Stretch 上，并在办公室（环境 A）和公寓（环境 B）中进行评估。我们的系统在 8 个回合中实现了 88% 的成功率。

5.5. 真实世界部署

我们通过将 Hello Robot Stretch [22]部署到真实世界中来评估我们的方法。我们在两个不同的环境中评估，共进行八个回合，使用五种不同的图像目标（图6）。环境 A（环境 A）是一个带家具的办公空间，设有走廊和休息室。场景中有多盆盆栽植物、沙发、椅子以及其他干扰物体。我们使用描绘沙发、椅子和盆栽植物的图像目标进行实验。环境 B（环境 B），如图1 所示，是一套带家具的公寓，设有厨房、起居区、卧室、浴室和办公室。我们使用描绘床和盆栽植物的图像目标进行实验。图像目标通过手机拍摄。所有回合开始时都不会看到目标，因此需要进行探索。各回合的最短路径测地距离为3至10米。总的来说，我们的智能体在 $7/8 = 88\%$ 的回合中成功完成任务。每个回合的视频都发布在项目页面。

6. 讨论

我们将 InstanceImageNav 分解为探索、目标重识别、目标定位和局部导航。我们基于这一框架构建了一个无需微调即可执行 InstanceImageNav 的系统。实验表明，该系统优于现有最先进的端到端学习策略，并能迁移到现实世界执行。

局限性与影响。我们系统的一个局限是检测和定位不具备记忆能力；在持久环境中执行序列任务时，若能将目标图像与压缩记忆关联起来，将会有所助益。我们的方法可作为评估训练所得导航策略的强大且稳健的基线，而我们的框架则可推动模块化语义导航器的开发。

致谢俄勒冈州立大学的相关工作部分得到了 DARPA机器常识项目的支持。本文所述观点和结论仅代表作者本人，不应被解读为美国政府或任何赞助方明示或暗示的官方政策或认可。

参考文献

[1] Ziad Al-Halah、Santhosh Kumar Ramakrishnan 和 Kristen Grauman。零经验要求：面向语义视觉导航的即插即用模块化迁移学习。载于计算机视觉与模式识别（CVPR），2022 年。2, 3[2] Peter Anderson、Angel Chang、Devendra Singh Chaplot、Alexey Dosovitskiy、Saurabh Gupta、Vladlen Koltun、Jana Kosecka、Jitendra Malik、Roozbeh Mottaghi、Manolis Savva 等。具身导航智能体的评估。arXiv 预印本 arXiv:1807.06757，2018 年。2, 6[3] Peter Anderson、Qi Wu、Damien Teney、Jake Bruce、Mark Johnson、Niko S¨underhauf、Ian Reid、Stephen Gould 和 Anton Van Den Hengel。视觉语言导航：在真实环境中解读视觉 grounding 导航指令。载于计算机视觉与模式识别（CVPR），2018 年。2[4] Vaibhav Bansal、Gian Luca Foresti 和 Niki Martinel。我在哪里见过它？结合注意力的物体实例重识别。载于国际计算机视觉会议（ICCV），2021 年。3[5] Vaibhav Bansal、Stuart James 和 Alessio Del Bue。re-obj：联合学习物体实例重识别的前景与背景。载于国际图像分析与处理会议（ICIAP），2019 年。3[6] Dhruv Batra、Aaron Gokaslan、Aniruddha Kembhavi、Olek- sandr Maksymets、Roozbeh Mottaghi、Manolis Savva、Alexander Toshev 和 Erik Wijmans。重新审视物体导航：评估前往物体的具身智能体。arXiv 预印本 arXiv:2006.13171，2020 年。2, 3, 6[7] Yash Bhalgat、Joao F Henriques 和 Andrew Zisserman。以轻量方式教会 Transformer 多视图几何。arXiv 预印本 arXiv:2211.15107，2022 年。3[8] Devendra Singh Chaplot、Dhiraj Gandhi、Saurabh Gupta、Ab- hinav Gupta 和 Ruslan Salakhutdinov。使用主动神经 SLAM 学习探索。载于国际学习表征会议（ICLR），2020 年。2, 5[9] Devendra Singh Chaplot、Dhiraj Prakashchand Gandhi、Abhi- nav Gupta 和 Russ R Salakhutdinov。使用目标导向语义探索进行物体目标导航。载于神经信息处理系统大会（NeurIPS），2020 年。2, 3, 5, 6[10] Devendra Singh Chaplot、Ruslan Salakhutdinov、Abhinav Gupta 和 Saurabh Gupta。用于视觉导航的神经拓扑 SLAM。载于计算机视觉与模式识别（CVPR），2020 年。2, 3[11] Changan Chen、Unnat Jain、Carl Schissler、Sebastia Vi- cenc Amengual Gari、Ziad Al-Halah、Vamsi Krishna Ithapu、Philip Robinson 和 Kristen Grauman。SoundSpaces：三维环境中的视听导航。载于欧洲计算机视觉会议（ECCV），2020 年。2[12] Yunho Choi 和 Songhwai Oh。基于关键点的强化学习实现图像目标导航。载于国际普适机器人会议（UR），2021 年。3[13] Navneet Dalal 和 Bill Triggs。用于人体检测的方向梯度直方图。载于计算机视觉与模式识别（CVPR），2005 年。3[14] Matt Deitke、Dhruv Batra、Yonatan Bisk、Tommaso Ca mpari、Angel X Chang、Devendra Singh Chaplot、Changan Chen, Claudia P´erez D'Arpino、Kiana Ehsani、Ali Farhadi 等。具身人工智能研讨会回顾。arXiv 预印本 arXiv:2210.06849，2022。2[15] Daniel DeTone、Tomasz Malisiewicz 和 Andrew Rabi-novich。SuperPoint：自监督兴趣点检测与描述。载于计算机视觉与模式识别研讨会（CVPRW），2018。3、4[16] Hugh Durrant-Whyte 和 Tim Bailey。同时定位与地图构建：第一部分。《IEEE机器人与自动化杂志》，2006。3[17] Samir Yitzhak Gadre、Mitchell Wortsman、Gabriel Ilharco、Ludwig Schmidt 和 Shuran Song。CLIP on Wheels：将零样本物体导航作为物体定位与探索。arXiv 预印本 arXiv:2203.10421，2022。2、3[18] Theophile Gervet、Soumith Chintala、Dhruv Batra、Jitendra Malik 和 Devendra Singh Chaplot。在真实世界中导航至物体。arXiv 预印本 arXiv:2212.00922，2022。2、3、4、6[19] Meera Hahn、Devendra Singh Chaplot、Shubham Tulsiani、Mustafa Mukadam、James M Rehg 和 Abhinav Gupta。无强化学习，无仿真：学习导航而不进行导航。载于神经信息处理系统大会（NeurIPS，2021）。2、3、5[20] 何恺明、Georgia Gkioxari、Piotr Doll´ar 和 Ross Gir-shick。掩码 R-CNN。载于国际计算机视觉会议（ICCV），2017。3[21] 何恺明、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。载于计算机视觉与模式识别（CVPR），2016。6[22] Charles C Kemp、Aaron Edsinger、Henry M Clever 和 Blaine Matulevich。Stretch 的设计：面向室内人类环境的紧凑型轻量移动操作机器人。载于国际机器人与自动化会议（ICRA），2022。8[23] Jacob Krantz 和 Stefan Lee。连续环境中的视觉语言导航仿真到仿真迁移。载于欧洲计算机视觉会议（ECCV），2022。5[24] Jacob Krantz、Stefan Lee、Jitendra Malik、Dhruv Batra 和 Devendra Singh Chaplot。实例特定的图像目标导航：训练具身智能体寻找物体实例。arXiv 预印本 arXiv:2211.15876，2022。2、3、5、6[25] Jacob Krantz、Erik Wijmans、Arjun Majumdar、Dhruv Batra 和 Stefan Lee。超越导航图：连续环境中的视觉语言导航。载于欧洲计算机视觉会议（ECCV），2020。2[26] Li Liu、Wanli Ouyang、Xiaogang Wang、Paul Fieguth、Jie Chen、Xinwang Liu 和 Matti Pietik¨ainen。通用目标检测的深度学习：综述。《国际计算机视觉杂志》，2020。3[27] David G Lowe。基于局部尺度不变特征的目标识别。载于国际计算机视觉会议（ICCV），1999。3[28] Arjun Majumdar、Gunjan Aggarwal、Bhavika Devnani、Judy Hoffman 和 Dhruv Batra。ZSON：使用多模态目标嵌入的零样本物体目标导航。arXiv 预印本 arXiv:2206.12403，2022。2、3

[29] Lina Mezghan、Sainbayar Sukhbaatar、Thibaut Lavril、Olek-sandr Maksymets、Dhruv Batra、Piotr Bojanowski 和 Kar-teek Alahari。用于图像目标导航的记忆增强强化学习。载于智能机器人与系统国际会议（IROS），2022。2、3[30] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议（ICML），2021。5、6[31] Santhosh K. Ramakrishnan、Devendra Singh Chaplot、Ziad Al-Halah、Jitendra Malik 和 Kristen Grauman。PONI：通过无交互学习实现物体目标导航的势函数。载于计算机视觉与模式识别（CVPR），2022。2[32] Santhosh Kumar Ramakrishnan、Aaron Gokaslan、Erik Wi-jmans、Oleksandr Maksymets、Alexander Clegg、John M Turner、Eric Undersander、Wojciech Galuba、Andrew West-bury、Angel X Chang、Manolis Savva、Yili Zhao 和 Dhruv Batra。Habitat-Matterport 3D 数据集（HM3D）：用于具身人工智能的 1000 个大规模三维环境。载于神经信息处理系统大会（NeurIPS）数据集与基准赛道，2021。2、5[33] Paul-Edouard Sarlin、Daniel DeTone、Tomasz Malisiewicz 和 Andrew Rabinovich。SuperGlue：使用图神经网络学习特征匹配。载于计算机视觉与模式识别（CVPR），2020。3、4[34] Nikolay Savinov、Alexey Dosovitskiy 和 Vladlen Koltun。用于导航的半参数拓扑记忆。载于国际学习表征会议（ICLR），2018。3[35] Manolis Savva、Abhishek Kadian、Oleksandr Maksymets、Yili Zhao、Erik Wijmans、Bhavana Jain、Julian Straub、Jia Liu、Vladlen Koltun、Jitendra Malik 等。Habitat：具身人工智能研究平台。载于国际计算机视觉会议（ICCV），2019。5[36] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。近端策略优化算法。arXiv 预印本 arXiv:1707.06347，2017。6[37] James A Sethian。用于单调推进前沿的快速行进水平集方法。《美国国家科学院院刊》（PNAS），1996。5[38] Dhruv Shah、Benjamin Eysenbach、Gregory Kahn、Nicholas Rhinehart 和 Sergey Levine。ViNG：使用视觉目标学习开放世界导航。载于国际机器人与自动化会议（ICRA），2021。2[39] Dhruv Shah、Benjamin Eysenbach、Nicholas Rhinehart 和 Sergey Levine。开放世界导航的快速探索与潜在目标模型。载于机器人学习会议（CoRL），2021。2、3[40] Dhruv Shah 和 Sergey Levine。ViKiNG：带有地理提示的基于视觉的公里级导航。载于机器人学：科学与系统（RSS），2022。2[41] Josef Sivic 和 Andrew Zisserman。Video Google：一种用于视频中物体匹配的文本检索方法。载于国际计算机视觉会议（ICCV），2003。3

[42] Andrew Szot、Alex Clegg、Eric Undersander、Erik Wijmans、Yili Zhao、John Turner、Noah Maestre、Mustafa Mukadam、Devendra Chaplot、Oleksandr Maksymets、Aaron Gokaslan、Vladimir Vondrus、Sameer Dharur、Franziska Meier、Woj- ciech Galuba、Angel Chang、Zsolt Kira、Vladlen Koltun、Ji- tendra Malik、Manolis Savva 和 Dhruv Batra。《Habitat 2.0：训练家居助手重新布置其栖居环境》。载于神经信息处理系统大会（NeurIPS），2021。5[43] Fuwen Tan、Jiangbo Yuan 和 Vicente Ordonez。《使用重排序Transformer进行实例级图像检索》。载于国际计算机视觉会议（ICCV），2021。3[44] Justin Wasserman、Karmesh Yadav、Girish Chowdhary、Ab- hinav Gupta 和 Unnat Jain。《最后一公里具身视觉导航》。载于机器人学习会议（CoRL），2022。3[45] Erik Wijmans、Irfan Essa 和 Dhruv Batra。《VER：扩大按策略强化学习规模，促使具身重排中涌现导航能力》。载于神经信息处理系统大会（NeurIPS），2022。6[46] Yi Wu、Yuxin Wu、Aviv Tamar、Stuart Russell、Georgia Gkioxari 和 Yuandong Tian。《用于语义视觉导航的贝叶斯关系记忆》。载于国际计算机视觉会议（ICCV），2019。2，3[47]Karmesh Yadav、Jacob Krantz、Ram Ramrakhya、San- thosh Kumar Ramakrishnan、Jimmy Yang、Austin Wang、John Turner、Aaron Gokaslan、Oleksandr Maksymets、Angel X Chang、Manolis Savva、Devendra Singh Chaplot、Alexander Clegg 和 Dhruv Batra。《Habitat挑战赛2023》。https: //aihabitat.org/challenge/2023/，2023。3[48] Karmesh Yadav、Arjun Majumdar、Ram Ramrakhya、Naoki Yokoyama、Alexei Baevski、Zsolt Kira、Oleksandr Maksymets 和 Dhruv Batra。《OVRL-v2：用于图像导航和物体导航的简易最先进基线》。arXiv预印本 arXiv:2303.07798，2023。2，6[49] Karmesh Yadav、Ram Ramrakhya、Arjun Majumdar、Vincent- Pierre Berges、Sachit Kuhar、Dhruv Batra、Alexei Baevski 和 Oleksandr Maksymets。《用于具身导航的离线视觉表征学习》。arXiv预印本 arXiv:2204.13226，2022。2，3[50] Karmesh Yadav、Ram Ramrakhya、Santhosh Kumar Ramakr- ishnan、Theo Gervet、John Turner、Aaron Gokaslan、Noah Maestre、Angel Xuan Chang、Dhruv Batra、Manolis Savva 等。《Habitat-Matterport三维语义数据集》。arXiv预印本 arXiv:2210.05633，2022。5[51] Brian Yamauchi。《一种用于自主探索的基于前沿的方法》。载于机器人与自动化计算智能国际研讨会（CIRA），1997。4[52] Xingyi Zhou、Rohit Girdhar、Armand Joulin、Philipp Kr¨ahenb¨uhl 和 Ishan Misra。《使用图像级监督检测两万类》。载于欧洲计算机视觉会议（ECCV），2022。5[53] Yuke Zhu、Roozbeh Mottaghi、Eric Kolve、Joseph J Lim、Ab- hinav Gupta、Li Fei-Fei 和 Ali Farhadi。《使用深度强化学习在室内场景中进行目标驱动视觉导航》。载于国际机器人与自动化会议（ICRA）。IEEE，2017。2，3

补充材料

A. 验证阈值的最优性

我们评估了实例重识别阈值确定方法（第4.1节）的有效性。我们从训练集划分中抽取400个回合作为子集，并通过从20个不同场景中采样、每个场景选取40个物体实例、每个物体选取1个回合，确保场景和物体的多样性。随后，我们运行完整的`Mod-IIN`系统，并使用不同的检测阈值搜索最佳检测阈值。如图7所示，搜索结果以成功率随阈值变化的曲线呈现。通过最大F值方法确定的阈值，恰好与下游性能的最优阈值一致。

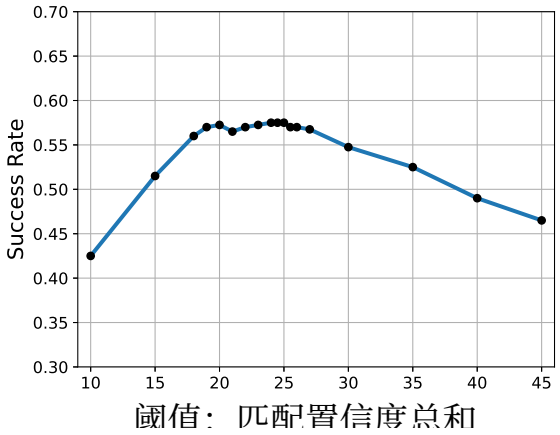


图 7：训练集划分的一个子集上实例重识别阈值的实测表现。

B. 其他方法的重识别阈值

在图 9中，我们展示了主论文中实验的每种目标实例重识别方法的精确率-召回率曲线和F值曲线： *ResNet*、*CLIP*、关键点匹配和关键点置信度。与 *ResNet* 和 *CLIP*相比，两种基于关键点的方法的精确率-召回率曲线下面积和最大 F 值都显著更高。尽管 关键点匹配 与 关键点置信度 的最大 F 值非常接近（0.962 与0.972），但这一差异导致下游 InstanceImageNav 性能相差 0.022成功率（表 2）。值得注意的是，最大 F 值的排名与下游性能的排名一致。

C. Crop-Projected 目标定位详情

我们的主要目标定位方法 *Detic* 投影会生成目标实例的实例分割掩码，以确定将哪些特征对应关系投影到世界坐标。我们使用 *Crop-Projected* 方法对该实例掩码进行消融：该方法不使用针对特定实例的掩码，而是 *Crop-Projected* 为目标图像采用一个静态中心裁剪，对所有回合均相同。位于此掩码内的匹配关键点会被投影到目标地图通道。我们发现，将目标图像中央三分之一区域向下延伸至八分之一高度作为裁剪区域，效果较好。图中展示了该掩码的两个示例。8。

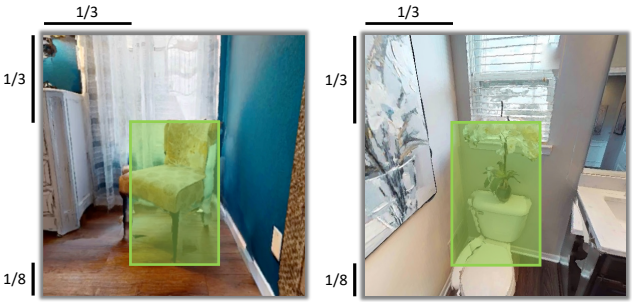


图 8：在验证集划分中对目标图像应用静态中心裁剪的示例。虽然裁剪范围较合理地框住了椅子（左），但植物（右）比裁剪区域小得多。

D. 实现细节

我们使用 PyTorch 实现模型和向量化映射，参考 Home-Robot[3] 和 SemExp ObjectNav[4] 代码库。采用 ResNet 的模型变体使用在 ImageNet-1K 上预训练的 ResNet50。采用 CLIP 的模型变体使用 *ViT-B/32* 权重。我们使用预训练的 SuperGlue 模型[5]，其室内训练权重来自 ScanNet。对于 Detic 模型，我们使用官方代码库[6] 及其默认的 21K 训练模型。我们采用以下自定义词汇表：{ 椅子、沙发、床、马桶、盆栽植物、电视显示器 }。在真实环境实验中，我们将系统部署在 Hello Robot 的 Stretch V1 上，并配备默认的 Intel RealSense D435i RGB-D 相机。

[3] github.com/facebookresearch/home-robot [4] github.com/devendrachaplot/Object-Goal-Navigation [5] github.com/magicleap/SuperGluePretrainedNetwork

[6] github.com/facebookresearch/Detic [7] wiki.ros.org/hector_slam

E. 真实环境定性示例

图 1 展示了我们的智能体在真实环境中表现的定性示例。在此示例中，智能体在环境 A（带家具的办公空间）中运行，并获得了一张椅子的图片。智能体沿走廊经过一张沙发（步骤 1、97），进入放有椅子的房间（步骤164），重新识别并定位椅子（步骤229），最终成功导航至椅子处（步骤237）。该示例展示了我们的方法对于仿真环境与真实环境之间领域差距的鲁棒性——Stretch 上的 RGB-D 观测包含真实世界噪声，位姿则使用 ROS 上的 Hector SLAM 进行估计[7]。该图像

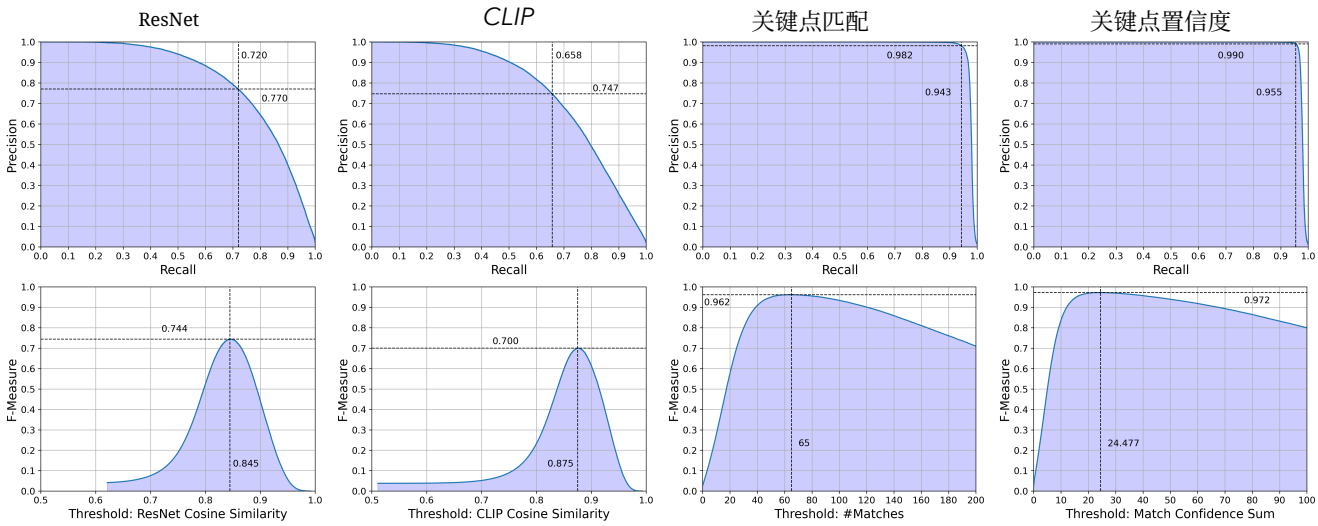


图9：每种目标实例重识别方法的精确率-召回率曲线（上）和 F 值曲线（下）。

目标取自手机摄像头，并且拍摄位姿能够根据发出任务的用户的判断，为其提供目标实例的直观视角。无需依赖机器人硬件，这展示了 InstanceImageNav 中目标指定的便捷性以及我们方法的鲁棒性。

F. 完整真实世界结果

表。3 展示了 Mod-INN 在真实环境中执行的每个回合的描述和结果。表 3 中的回合编号对应于项目页面上视频的编号。

回合编号	环境	目标编号	**成功?**
1	A	椅子 0_	True
2	A	植物 0_	True
3	A	沙发 0_	否
4	A	沙发 0_	True
5	A	沙发 0_	True
6	A	沙发 0_	True
7	B	植物 1_	True
8	B	床 0_	True

表 3：Mod-INN 在真实环境中的回合及结果描述。