SRPose:双视图相对姿态估计
稀疏关键点

Rui Yin [1], Yulun Zhang [2], Zherong Pan [3], Jianjun Zhu [1],
Cheng Wang [1], and Biao Jia [1]

[1] Hanglok-Tech,中国
[2] 上海交通大学人工智能教育部重点实验室
[3] 光速工作室、腾讯美国

摘要。双视角姿态估计对于无地图视觉重新定位和物体姿态跟踪任务至关重要。然而，传统的匹配方法存在耗时的鲁棒估计器问题,而深度

基于学习的姿态回归器仅适用于相机到世界的姿态估计,缺乏对不同图像尺寸和相机的通用性
内在函数。在本文中,我们提出了 SRPose,一种基于稀疏关键点的
相机到世界的双视图相对姿态估计框架和
物体到相机场景。SRPose 由稀疏关键点检测器、固有校准位置编码器和可提示的先验组成
知识引导的注意层。给定一个固定的
场景或移动物体,SRPose 估计相对相机或 6D
物体姿态变换。大量实验表明,SR-Pose 在准确度和速度方面与最先进的方法相比具有竞争力或更优异的性能,表现出普遍性

适用于两种场景。它对不同的图像大小和相机固有特性具有鲁棒性,并且可以用较少的计算资源进行部署。项目页面：
https://frickyinn.github.io/srpose。

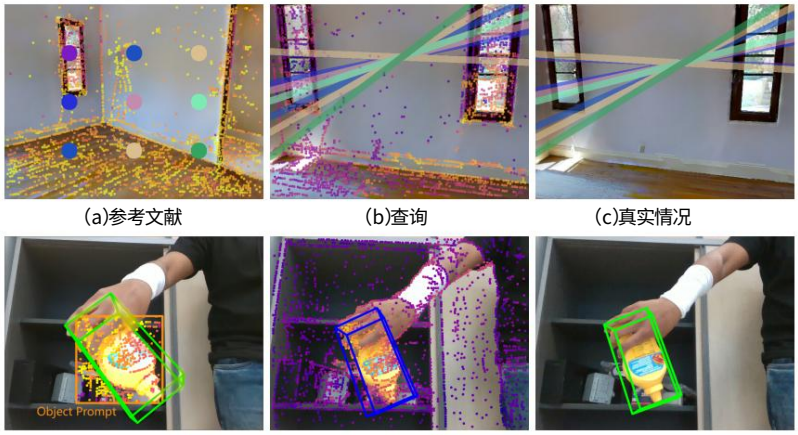
关键词:相对姿态估计·6D物体姿态估计

1 简介

两幅图像之间的相对姿态估计在许多 3D
视觉任务,包括视觉里程计和无地图视觉重定位[3],6D
物体姿态跟踪 [67,68] 等。这些任务可以分为两种场景:相机到世界和物体到相机的估计。在相机到世界
场景,例如在无地图视觉重定位中,我们估计从参考图到查询图像的姿态变换,预测变化

两个视图之间的相机外部特性。在物体到相机的场景中，
以物体姿态跟踪为例,我们的目标是通过估计目标物体在相机中的相对变换来跟踪视频中目标物体的 6D 姿态
两个相邻帧之间的坐标系统。虽然这两种情况可能

表示通讯作者。



(a)参考文献 (b)查询 (c)真实情况

(d)参考文献 (e)查询图 (f)真实情况

1:SRPose 的相对姿势估计。图中绘制的点表示
两幅图像中稀疏关键点的交叉注意力得分,其中较亮的点
代表更高的注意力。相机到世界:(a)、(b)、(c)可视化极线
线,代表两个九个对应点之间的连接
视图。场景的重叠部分更受关注。物体到相机:
(d)、(e)、(f)显示只有一个
参考图中的可访问对象提示。对目标的关注度更高
对象。SRPose 建立隐式对应关系。

看起来不同,它们都需要从两个向量中预测姿势矩阵
RGB 图像描绘场景或物体,进而表示相对
从第一张图像到第二张图像的姿势变换。

传统的姿态估计方法通常涉及检测和匹配两幅图像中的关键点或局部特征以建立点对点
对应关系,然后使用稳健估计器对异常值进行去噪并恢复

通过求解二视图几何问题,根据基本矩阵得出相对位姿,
即对极约束方程[48]。这些方法可以得到准确的
无论图像大小和相机本质如何,结果都是如此,这要归功于相机的本质
在求解方程之前进行校准。鉴于分割
两幅图像中的目标物体,这些方法也可以得到可接受的
预测相对物体姿态变换。然而,不利的是,稳健估计器在计算上可能很昂贵。事实上,

消除不匹配的对应关系,并使用稳健方法求解方程
与检测和匹配关键点相比,估计器的速度明显较慢
或特征,使其在实时应用中受到限制。此外,实时对象姿态跟踪需要视频对象分割模型来消除

偏离目标的关键点 [67, 68],这也会带来额外的开销。

深度学习的成熟提供了直接回归的优势
来自两个 RGB 输入的相对姿态变换,显著提升
运行时性能。然而,现有的回归方法缺乏对不同尺寸和相机本征图像的通用性或精度

[3, 38, 53, 65, 75, 78]。他们通常使用只接受

在批量训练期间使用固定大小的输入。与传统方法不同，
大多数基于深度学习的回归器缺乏对原始相机固有参数的认识，无法利用双视图几何来实现更高的精度
在姿态估计中。此外，最先进的深度回归器仅适用于
相机到世界的姿态估计，尽管两者共享一个共同的解决方案
相机到世界和物体到相机的任务。

为了解决上述挑战，我们提出了 SRPose：一种稀疏
基于关键点的相对姿势估计框架。SRPose 估计
通过隐式求解对极约束，基于双视图几何构建相对位姿矩阵，如图1所示。首先，SRPose采用点检测器
提取稀疏关键点和相关描述符以形成候选集
隐式对应关系。然后，我们使用内在校准 (IC) 位置
编码器在计算位置嵌入之前对关键点进行调制，
考虑到不同的图像尺寸和相机固有特性。关键点描述符和位置嵌入随后被融合到先验知识引导的

注意层，利用关键点相似性的先验知识来建立隐式跨视图对应关系。借助这些层中可提示的交叉注意机制，SRPose 只需要一个可访问的对象

在两个视图之一中提示计算相对 6D 姿势变换
从而无需进行难以捉摸的对象分割。

本文的主要贡献可以概括如下：

- 我们提出了一种使用稀疏关键点 SRPose 的新框架，用于双视图
  相对姿态估计。据我们所知，这是第一次尝试
  直接从稀疏关键点回归此任务的相对姿势。
- 我们引入了内在校准位置编码器，以使 SRPose 适应
  不同的图像尺寸和相机内在特性。
- 我们通过利用一个可用的对象提示，实现对象到相机的估计以及相机到世界的估计。

- 我们的 SRPose 通过替换稳健估计器大大减少了估计时间
  通过直接回归，实现了最先进的准确率
  以及相机到世界和物体到相机场景中的速度。

2 相关作品

相对姿势估计：在相对姿势估计中，传统方法
从本质矩阵中恢复旋转和平移[27,42]。利用对极约束[48]，基于

通过稀疏关键点匹配建立跨视图对应关系
或密集特征。使用此管道，基于稀疏关键点的方法通常从检测稀疏关键点及其相关描述符开始，包括经典的手工制作的检测器 [5, 44, 54] 和基于深度学习的检测器 [6,15–17,24,35,40,45,52,64,73,76,77]。最近的方法进一步采用了深度学习

基于学习的匹配器[12, 39, 55, 57, 71]建立关键点匹配，而
经典方法依赖于最近邻搜索。与使用稀疏
关键点，密集匹配器 [11, 13, 18, 19, 36, 47, 60, 63, 66, 72, 74, 80] 无需检测器，可执行像素级密集特征匹配。一旦对应关系

通过匹配建立,可以使用诸如 RANSAC [23] 之类的稳健估计器
用于估计本质矩阵并恢复相对姿态。总体而言,
虽然传统的基于匹配器的方法已经显示出良好的效果,
它们仍然面临着耗时的稳健估计器的挑战。此外,密集匹配器虽然提供了出色的准确性,但可能会受到以下问题的影响
由于其资源密集型设计,导致性能缓慢。

相对姿势回归:深度学习回归器通过使用专用神经网络直接预测相对姿势,提供了一种替代方法。

无需明确建立对应关系。双视图相对姿态估计侧重于帧间姿态变换,旨在实现视觉里程计

以及无地图重定位 [1, 3, 4, 21, 34, 46, 69]。利用先验知识
这是双视图估计中的一个重要方面。例如,SparsePlanes [32],
PlaneFormer [8] 和 NOPE-SAC [62] 利用平面
表面来提高性能。不幸的是,这些方法依赖于图像
编码器专为固定大小和内在特征的图像而设计,缺乏通用性
在不同设置下。稀疏视图相对姿态估计强调全局
姿态优化或涉及多视图的端到端 SfM [38,58,65,75]。虽然这些稀疏视图方法可以扩展到双视图估计,并且
甚至可以预测未知的内在特性,但它们对不同相机设置的适应性取决于多视角信息。相比之下,
SRPose 利用稀疏
关键点和双视图几何,以实现更高的精度和通用性
跨越不同图像尺寸和相机本质。

物体姿态估计:虽然上述大多数方法都侧重于固定场景中的相机姿态估计,但有些任务,如 6D 物体

姿势跟踪,也需要相对物体姿势估计。大多数框架
使用相同的关键点对应管道预测相对物体姿势。BundleTrack [67] 和 BundleSDF [68] 通过以下方式跟踪视频中的物体姿势:
匹配两帧之间的关键点,并进行全局姿态图优化。POPE [22] 结合了分割的高级基础模型 [33]

并匹配 [49] 以实现零样本物体姿态估计。这些方法鲁棒性
经典方法可能存在一些缺点,它们需要专用的对象掩码来聚焦对象,而这些掩码由其他资源消耗型分割模型提供。
摆脱物体分割,OnePose [61]、OnePose++ [29] 和 Gen6D [41]
依赖于对象的多个视图,而 ZeroShot [25] 建立语义对应关系来消除背景,并使用深度图来实现姿势
估计。尽管如此,这些方法需要额外的高质量信息,并且它们限制查询视图仅容纳单个对象。

作为补救措施,SRPose 采用可访问的对象提示来聚焦目标对象,从而实现有效的无深度和无掩模双视图相对对象
在物体到相机场景中具有卓越性能的估计。

3 方法

给定两幅图像 I_1、I_2,SRPose 基于双视图几何估计由旋转 R\in \mathcal {SO}(3) 和平移 t\in \mathbb {R}^3 组成的相对姿势。 ,
详细的问题定义可以在补充材料中找到。传统方法使用极线约束 [48] 来估计以下

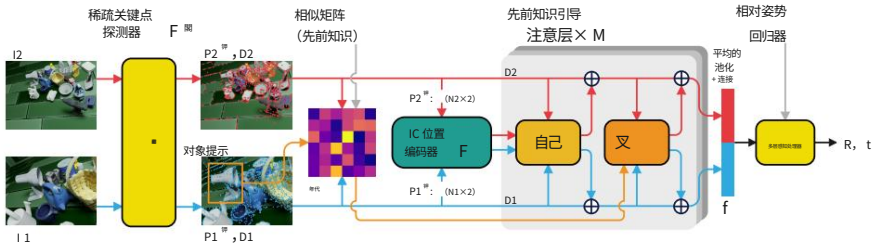


图 2:概览。SRPose 包含四个主要部分:1)稀疏关键点
检测器分别从两幅图像中检测与描述符相关的关键点;
2)本征校准（IC)位置编码器调节关键点的坐标
结合相机本征函数,并对其位置信息进行编码;3)在先验
关键点相似性的认知,以及物体提示,注意力层
建立隐式的跨视图对应关系;4)回归器估计相对姿态
R,t 在隐式对应的约束下。

本质矩阵 E,从中恢复 R,t:

(2)

其中 q_1, q_2 是 是 I_1, I_2 中的任意两个对应点 ，和 K_1, K_2
两幅图像的相机本征。SRPose 构建并解决约束
方程（1)使用神经网络进行隐式求解。SRPose 表示为 \protect \mathcal {SR},
我们的框架的功能表示为:

(3)

图 2 总结了 SRPose,涉及四个组件:稀疏关键点检测器、内在校准位置编码器、可提示的先验知识
引导的注意层和 MLP 回归器。

3.1 稀疏关键点检测器

我们没有使用图像编码器来提取图像特征,而是使用检测器从两幅图像中提取稀疏关键点和相关描
述符。
为此,我们可以使用多年来开发的现有技术之一，
其中包括经典方法,如 SIFT [44],以及基于深度学习的方法,如 SuperPoint [16]、ALIKE [77]、
DISK [64] 等。给定图像 I,我们的
的关键点,表示为 \protect ，检测由 N 个坐标 P^k \subset \mathbb {R}^{2} 检测器组成
在图像空间中,以及图像中的相关描述符 {d_k}_{k=1}^{N} 描述局部特征
\mathbb {R}^{d},其中 d 表示关键点描述符在图像空间中的维度
网络。总而言之,我们的检测器定义为以下函数:

(4)

实际上,我们将所有图像调整为标准尺寸,以实现并行批处理
检测,然后将关键点的坐标 P^k 重新调整到其原始位置。
稀疏关键点检测器产生两组关键点 P_1^k、P_2^k,从中可以隐式地得到一组对应关系 从 I_1 开 ，
\ifmmode \lbrace \else \textbraceleft \fi (q_1, 始,I_2 将成立
q_2)\} \subset P_1^k \times P_2^k,用于求解约束方程 (1)。

3.2 内在校准 (IC) 位置编码器

根据公式（1），对于从图像 I 中检测到的关键点，应首先通过相机本征 K 对其坐标进行校准或归一化：

(5)

执行内在校准 (IC) 可以大大改善
准确性和稳健性 [48]。IC 允许 SRPose 调整由
具有不同内在参数的不同相机。以前的回归器[3, 53]
使用在固定大小图像上训练的图像编码器提取隐式图像特征，
假设所有输入的相机本征默认都是相同的。
SRPose 中的稀疏关键点通过内在校准进行调节，
提供统一相机中所有关键点的精确位置，即 P^c ，
跨不同内在函数的坐标系。

在进行本征校准后，对关键点的坐标进行编码，以根据其位置对关键点进行寻址。由于相对位置编码可能造成位置失真 [39, 60]，我们采用绝对位置

编码保持精确的坐标值，这对于
双视图几何问题。使用单层全连接网络，表示为以下函数 \protect \mathcal {F}^{pe} 位置编码器映射 2D 校准
，
坐标到高维位置嵌入：

(6)

得到的编码位置嵌入表示为 P^{e}，其与关键点描述符和描述符嵌入一起表示为 ，然后incor-

D \oplus P^{e} 使用一组注意层提供经过校准和规范化的位置信息，以解决约束方程 (1)。

3.3 可提示的先验知识引导注意力

给定位置嵌入和关键点描述符，我们使用多层
注意力网络利用语义信息并建立隐式对应关系。我们的网络由 M 个连续的注意力层组成，每个

层由一个自注意力模块和一个可提示的先验知识引导的交叉注意力模块组成。我们将 X^m\subset \mathbb {R}^{d} 表示为关键点描述符
嵌入在第 m 层进行处理，其中 m \in M，X^m 为 ，
来自图像 I_1 或 I_2 的 X_1^m 或 X_2^m。每一层都以 X^{m-1} 作为输入，然后
计算更深的嵌入 X^m 。总之，整个多层网络在 M 层之后输入关键点描述符 D_1\, D_2，
，并输出 X_1^M, X_2^M
即，我们将 D 定义为嵌入 X^0 的初始状态。在每次注意力之前
层，X_1^{m-1} 与相关位置嵌入合并
， . 按照 [28] 的方法，残差连接被应用到每一个自和交叉 P_1^{e}
P_2^{e} 注意力模块。注意力层的架构如图 3 所示。

自注意力：自注意力模块捕获整个图像中所有关键点的综合语义信息。在此模块中，首先通过线性变换将两幅图像的 X^{m-1} \oplus P^{e} 映射到三个向量，分别表示为查询 q、键 k 和值 v，其中

化 W_s^m ，

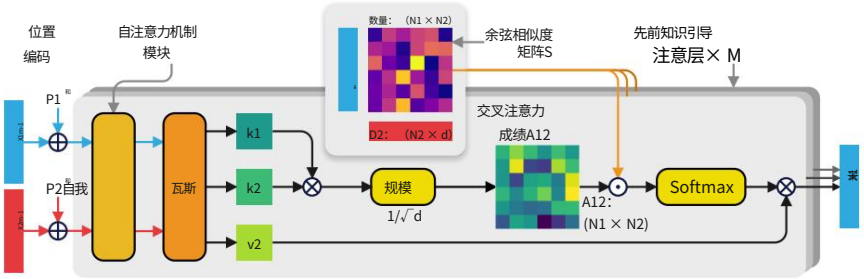


图 3:先验知识引导的注意力层概览。每一层
包含一个自注意力模块和一个交叉注意力模块。
利用关键点描述符作为先验知识来重新调整交叉注意力
分数,引导更多注意力到具有隐含对应关系的跨视图关键点对。图中省略了对象提示和残差连接。

q, k, v 。然后使用 k 和 q 来计算自注意力分数 A_s ,
\subset \mathbb {R}^{d} 然后根据描述符和嵌入 d 的维度进行缩放:

(8)

经过 \protect \mathrm {Softmax} 函数归一化后,自注意力得分 A_s 为
然后用于从 v 中提取高度关注的信息:

(9)

在自注意力模块中,关键点位置嵌入和描述符嵌入一起计算,以总结语义信息

建立更深层次的跨视图对应关系之前,我们需要先绘制整个图像。

先验知识引导的交叉注意力:以先验知识为引导
关键点相似性,交叉注意模块利用两幅图像之间的相互信息来建立隐式对应关系。类似于

自注意力、交叉注意力将输入嵌入 X_s^{m-1} 映射到两个键
k_1, k_2 以及两幅图像的值 v_1, v_2 分别通过另一线性
变换 W_c^m 层。按照 [50],我们从前一个图像的嵌入结果中计算出两个图像之间的双向注意力分数

自注意力模块,使用跨视图键 k_1、k_2 :

(12)

该模块识别场景或物体之间的重叠区域
通过从计算出的交叉注意力得分中识别两组关键点嵌入之间的相互性,可以区分两种观点。为了促进这
一过程,事先

通过使用关键点相似性矩阵调整交叉注意力得分,利用知识来引导和增强对高度相关关键点的注意力

矩阵的每个元素 s_{ij} 都计算为余弦相似度
每个跨视图关键点对的描述符之间:

(13)

具有相似局部特征的关键点往往具有相似的描述符,描述相似视图的相关语义信息。因此,余弦相似度

捕捉两个关键点之间的语义相似性,并指导框架识别图像之间的重叠部分。

我们将矩阵 S 归一化为 [0,1],然后逐元素将其与
交叉注意力得分。最后,这些得分用于提取和交换关键
从彼此的值 v_1, v_2 交叉查看信息 :

(16)

通过捕捉每幅图像的相关信息并将其整合到
另外,在两幅图像之间建立了隐式对应关系。
如图 1a 和 1b 所示,其中更亮的点表示更高的交叉注意力
分数,SRPose在两幅图像的重叠区域中建立隐式对应点,用于构建和解决

对极约束方程 (1) 的隐式版本。

对象提示:在对象到相机的估计中,我们采用可访问的
仅在一张图像中根据用户提供的对象提示来识别目标对象。在
相比之下,传统的匹配方法需要在两个图像中进行高质量的对象分割。对象提示 b 是一个边界框,由

左上角和右下角的两个坐标,界定了对象的
在其中一个图像 I_1 中,以 I_1 为参考视图。一旦关键点
从 I_1 中提取对象提示删除了 ,
参考视图中的边界框,消除大部分不相关的关键点
位于背景或其他物体上。因此,
边界内的关键点,重点关注目标对象 o,稍后将被处理
在相似度矩阵和其他后续模块中。借助引导式交叉注意模块,SRPose 可以学习识别位于

从整个查询视图 I_2 中筛选出相同的对象 o,排除不相关的背景
关键点,如图 1d 和 1e 所示。在实际实施中,为了方便
并行计算,批次内的所有关键点集都填充到相同的
从不同图像中移除不同数量的关键点后的维度。
通过对象提示,SRPose 搜索隐式对应点
目标物体并计算相对 6D 物体姿态变换
与相机到世界的场景过程相同。

3.4 相对姿态回归器

通过前面的模块计算出嵌入 X_1^M、X_2^M 之后,采用多层感知器 (MLP) 来回归相对姿势作为最后一步。

先前的注意层 X_1^M、X_2^M 的输出首先经过二维平均
池化,然后连接在一起,得到联合嵌入
f \in \mathbb {R}^{2d} 承载着深度特征和隐式对应关系
两幅图像。最后,嵌入 f 输入到 MLP,回归器
隐式求解极线约束方程（1）,并回归相对位姿,
包括旋转 R 和平移 t。根据 [79],我们设置旋转
回归器的输出为 6d 向量 r \in \mathbb {R}^6,然后转换为 ,
通过部分 Gram-Schmidt 过程对旋转矩阵 R \in \mathbb {R}^{3\times 3} 进行旋转。这
技术,如[79]中所述,可以实现旋转的连续表示。

3.5 损失函数

SRPose 采用监督方式进行训练。利用真实旋转 R_{gt},我们计算旋转角度误差的 Huber Loss [31],从而最小 ,
化预测旋转和真实旋转之间的角度误差:

(17)

其中 \protect \mathcal {H} 表示 Huber 损失函数。同样,给定真实值
翻译 t_{gt} 的误差是在标准化和非标准化中计算的
形式。为了进一步提高精度,我们还考虑了
翻译,导致以下三个损失项:

(20)

上述三个损失函数分别监督平移的尺度和方向。我们通过三个函数的加权和来计算最终的损失:

(21)

其中\lambda_1 \lambda_2 \lambda_3 是用于平衡四种不同损失权重的标量。

3.6 实现细节

SRPose 首先在 ScanNet [14] 上进行训练,然后在其他数据集上进行微调,以
实现更好的性能。对于所有数据集,我们将图像大小调整为 640\times 480,
并使用 SuperPoint [16] 提取 1,024 个稀疏关键点进行训练,并使用 2,048
进行评估。其他稀疏关键点检测器也可用于此步骤。
我们利用官方提供的预训练的 SuperPoint 并冻结
训练阶段的权重。其余可学习模块被初始化
具有随机权重。实际上,SRPose 由 6 个引导注意层组成,

每个注意力模块有 4 个注意力头。关键点的维度 d
描述符和嵌入设置为 256。旋转和平移使用两个单独的 3 层 MLP 进行估计。所有三个平衡标量 \lambda _t、\lambda _{t_n}、\lambda _{t_a}
设置为 1。在 ScanNet 上,我们的框架使用 AdamW 优化器进行训练 [43]
并遵循 1cycle 学习率策略 [59],最大学习率为
1 \times 10^{-4},500 个 epoch。我们将训练的批处理大小固定为 32。预训练需要 120 小时,使用 8 个 RTX 3090 GPU。更多实施

详细信息请参阅补充材料。

4 实验

我们评估了 SRPose 在双视图相对姿态估计方面的表现
在相机到世界和物体到相机的场景中。此外,我们
评估其在无地图重定位应用中的有效性。最后,我们
进行消融研究来分析我们提出的组件的作用。

4.1 相机到世界姿态估计

设置:在相机到世界的场景中,我们评估了我们的框架在 Matterport [10] 和 ScanNet [14] 上的性能。
按照 [32] 的说法,对于 Matterport,我们
分别报告平移和旋转误差的中值和平均值,以及阈值为 1m/30◦ 时的平移/旋转精度。对于 ScanNet,

我们采用[55]中的指标来计算5处的姿态误差的AUC ◦,
10◦,和 20◦ 。这里的位姿误差是旋转中的最大角度误差
和翻译。我们还报告了对
ScanNet-1500 测试集。对于所有基于匹配器的方法,我们使用 OpenCV [7] 中的 RANSAC [23] 实现来恢复相对姿势。所有方法
在具有 RTX 4090、i9-13900K 和 128GB 内存的设备上进行比较。

基线:我们选择三类方法作为基线,包括
基于稀疏关键点的匹配器、密集匹配器和基于深度学习的相对姿态回归器。所有基线的结果均按照

官方论文或实现。所有稀疏匹配器和 SRPose 都使用 Su-perPoint [16] 作为稀疏关键点检测器。所有回归器都在
在评估之前,需要使用特定的数据集,例如 Matterport 或 ScanNet。LightGlue [39] 是
在 MegaDepth [37] 上进行训练,而所有其他匹配器均在 ScanNet 上进行训练。

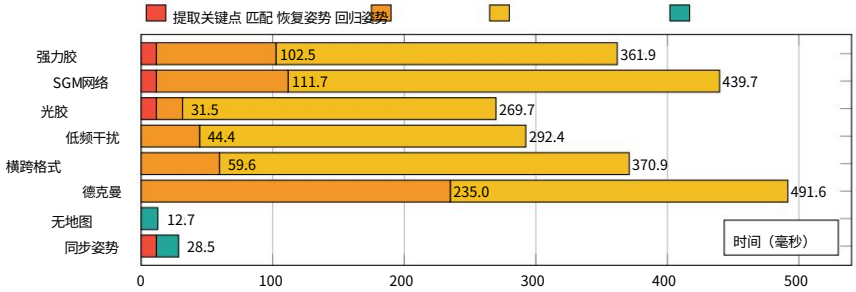


图 4:ScanNet 上的时间消耗比较 [14]。回归器包括
SRPose 比所有匹配器具有更高的计算效率。

表 1:Matterport 上的相对姿态估计 [10]。如果没有深度,基于匹配器的方法就无法进行缩放平移估计,而 SRPose 实现了
仅使用 RGB 输入即可实现较高的翻译精度。

分类 方法		破碎的(^\循环)	传输 (米)		
疏	强力胶 [55]	3.88 24.17 77.8 1.06 21.32 80.9	不适用/	不适用/	无
	SGM网 [12]	1.32 22.45 80.0	不适用/	不适用/	无
	光胶 [39]		不适用	不适用	无
稠密	洛夫特 [60]	0.71 11.11 90.5 75.7 89.0	不适用/	不适用/	无
	ASpanFormer[13] 3.73 31.45 DKM[18] 0.46 12.89		不适用/	不适用/	无
			不适用	不适用	无
回归器	稀疏平面 [32]	7.33 22.78 1.25 66.6 83.4	0.63		
	PlaneFormers [2] 8点	5.96 22.20 1.19 66.8 83.8	0.66		
	[53]	8.01 19.13 1.01 67.4 85.4	0.64		
	NOPE-SAC [62]	2.77 14.37 0.94 73.2 89.0	0.52		
	同步姿势	2.65 11.12 91.6 0.27 0.61 83.7			

表 2:ScanNet 上的相对姿态估计 [14]。SRPose 在各种指标方面都优于最先进的回归器和稀疏匹配器,并且由于消除了稳健估计器,大大减少了计算时间。

类别	方法	姿势估计 AUC @5 °	@10 ◦	@20 ◦	比赛总数	时间（毫秒）
疏	强力胶 [55]	16.2	33.8	51.8	102.5	361.9
	SGM网 [12]	15.4	32.1	48.3	111.7	439.7
	光胶 [39]	16.4	33.6	50.2	31.5	269.7
稠密	洛夫特 [60]	22.0	40.8	57.6	44.4	292.4
	ASpanFormer [13]	25.6	46.0	63.3	59.6	370.9
	德克曼 [18]	29.4	50.7	68.3	235.0	491.6
回归器	无地图 [3]	2.70	11.5	29.0	不适用/	12.7
	同步姿势	13.3	34.3	56.8	不适用	28.5

结果:表 1 表明 SRPose 的表现远远优于现有的
Matterport 上所有指标的基于深度学习的回归器。我们的框架
实现了最准确的缩放翻译估计,这是无法实现的
对于缺乏深度信息的基于匹配器的方法。表 2 显示了 SRPose
与所有稀疏匹配器相比,实现了具有竞争力或更优异的性能
和回归器在 ScanNet 上的表现。尽管密集匹配器的表现优于我们的
与框架在准确性方面相比,SRPose 在计算方面具有显著优势
通过使用直接回归代替稳健估计器,与所有基于匹配器的方法相比,速度更快。如图 4 所示,基于匹配器的方法花费
大部分时间都花在利用稳健估计器恢复姿势上
匹配。相比之下,SRPose 消耗的时间甚至比匹配阶段还要少
所有匹配器,总时间至少节省 200 毫秒。

4.2 物体到相机姿态估计

设置:在物体到相机的场景中,我们评估 HO3D [26] 上的 SRPose。
HO3D 包含来自 YCB 的 21 个类别的手持物体视频
数据集 [9]。还提供了立体深度相机捕获的深度和视频中每帧的物体分割。我们随机选择具有小物体旋转变换的图像对进行训练和评估。

表 3:HO3D 上的相对物体姿态估计 [26]。SRPose 无需额外的深度信息,在物体到相机的估计中实现了最先进的性能,仅从 RGB 输入即可进行尺度姿态估计。

方法	破碎的(^\循环)			对象 (厘米)			
	Med.\定界 "3223379					添加(-S)delimiter 3222378	ADD-S\delimiter 3222378
SGMNet [12] 24.1 LightGlue [39] 24.4		29.0	62.7	13.9	17.5	36.8	16.8	31.3
LoFTR [60] 29.2 DKM [18] 23.3		28.7	61.7	13.5	16.3	38.2	17.2	31.2
		57.9	50.9	13.9	16.4	35.6	16.7	30.9
		25.6	64.8	13.6	15.4	39.8	17.8	33.3
同步姿势	8.9	11.4	95.4	5.9	8.0	73.9	36.4	56.8

更多细节可以在补充材料中找到。继[30,70]之后，
我们采用广泛使用的平均距离（ADD,ADD-S)度量进行评估。为了测量物体姿态变换中较小的尺度,我们设置

度量阈值为 10 厘米,与 1 米阈值不同
相机到世界场景。具体来说,我们使用 ADD 计算 AUC
和 ADD-S,阈值设置为 10cm。我们还报告了中位数、平均值
平移/旋转误差,以及10cm/30 ◦的精度。

基线:我们仅选择第 4.1 节中列出的匹配器作为基线,例如
此双视图相关对象中没有可用的基于深度学习的回归器
姿势估计任务。我们利用数据集中提供的立体深度,通过正交方法启用基于匹配器的方法估计缩放相对姿势

Procrustes [20],而 SRPose 只需要 RGB 图像作为输入。数据集中的分割用于基于匹配器的方法识别

两幅图像中的目标物体,而 SRPose 利用的是
在第一幅图像中进行分割作为对象提示。

结果:如表 3 所示,SRPose 的表现明显优于最先进的基于匹配器的方法。SRPose 在相对 6D 方面表现出更高的准确率
使用对象提示在两幅图像之间进行对象姿态估计
参考视图。立体深度质量低导致性能受限
基于匹配器的方法的缩放姿势估计,而 SRPose 可以学习
缩放信息以实现更低的估计误差。

4.3 无地图视觉重定位

我们在 Niantic [3] 上评估了 SRPose 的应用,这是一个无地图视觉重定位任务的数据集。对于验证集和测试集中的每个场景，
提供参考图像和一系列查询图像来重新定位
查询位置相对于参考位置。根据 [3],我们采用虚拟
对应重投影误差 (VCRE) 和平移中值误差
和旋转作为指标。具体来说,我们计算 VCRE 的精度
90像素,重定位精度为0.25m和5 ◦。

基线:我们仍然选择三类方法作为基线,包括稀疏匹配器 SIFT [44]、SuperGlue [55];密集匹配器 LoFTR [60]，

RoMa [19];回归器 Map-free [3]。对于基于匹配器的方法,我们
通过从基本矩阵中恢复来估计姿势,这通常会产生

比使用正交普洛克鲁斯特法在这个任务中略有更高的精度,根据
然后我们使用深度预测模型 DPT [51] 估计的深度,
为匹配者提供尺度信息。

表 4:Niantic 上的无地图视觉重定位结果 [3]。SRPose 实现
与最先进的方法相比,无需依赖
一个额外的深度预测模型来协助缩放估计。

分类 方法		VCRE Prec.\分隔符“3222378 / Med.\分隔符”3223379	中。(m / 翻译。)delimiter 3223379 / 旋转。)delimiter 3223379	Prec.\delimiter 3222378
疏	SIFT[44]	25.0 / 222.8 36.1 /	2.93 / 61.4 1.88 /	10.3
+DPT [51]	强力胶 [55]	160.3	25.4	16.8
稠密	洛夫特 [60]	34.7 / 167.6 45.6 /	1.98 / 30.5	15.4
+DPT [51]	罗马 [19]	128.8	1.23 / 11.1	22.8
回归器	无地图 [3]	40.2 / 147.1 46.4 /	1.68 / 22.9	6.0
	同步姿势	127.7	1.37 / 17.2	16.9

结果:表 4 显示 SRPose 优于其他稀疏匹配器,并且
回归方法在无地图视觉重定位中的应用。尽管最先进的密集匹配器 RoMa 在使用尺度的重定位中实现了更高的精度
通过额外的深度预测模型估计,SRPose 仍然优于
VCRE. 与无地图视觉重定位任务中最先进的方法相比,我们的框架表现出了极具竞争力的性能。

4.4 洞察

消融研究:我们评估
Matterport 的 SRPose 中不同组件的有效性 [10]

和 MegaDepth [37] 数据集。具体来说,我们训练了四种不同的变体
使用随机权重初始化
Matterport 通过删除我们提出的三种不同的设计来验证
其有效性。如表5所示:1)去除关键点相似度的先验知识指导

表 6:消融研究
MegaDepth[37]。内在校准
位置编码器可以适应
不同的图像尺寸和相机内在特性。

方法	铺设是。曲线下面积 @5 。 @10。 @20。
无地图 [3]	2.6 9.3 22.9
SRPose(完整版)1)	16.6 36.0 58.0
无内在校准 1.5 2)无位置编码 1.0	8.5 24.3
	6.2 18.9

交叉注意模块会导致准确率略有下降,因为
它协助SRPose寻找跨视图对应关系。2) 替换所有
交叉注意模块与自注意模块的结合会导致准确率显著下降,因为建立隐式跨视图对应关系的能力

消除了交叉注意的影响。3)移除整个位置编码器
由于缺少关键点位置,导致准确率略低
信息。实验强调了三种设计的关键作用
SRPose 进行准确的相对姿势估计。

此外,我们训练了三种用随机权重初始化的不同变体
MegaDepth 是一个由不同尺寸和相机拍摄的图像组成的数据集

表 5:Matterport 上的消融研究 [10]。先验知识引导的注意力层有助于建立隐式跨视图对应关系。

方法	传输 (米)		破碎的(^{\circ})	
8分 [53]	1.01	67.4	19.13	85.4
SRPose（完整） 0.84 74.2 14.32 88.9				
1)无事先知识指导 0.97 69.6 16.91 86.8				
2) 用自我关注取代所有交叉关注 1.90 26.1 48.0 42.0				
3)无位置编码 1.17 61.0 21.39 81.3				

内在因素，进一步验证内在校准位置的有效性
编码器，结果如表 6 所示。1)如果不对关键点坐标进行内在校准，SRPose 仅能实现相当的性能

无地图，既不能适应不同的图像尺寸，也不能适应相机的内在特性
MegaDepth。2) 移除整个 IC 位置编码器会导致显著
由于消除了关键位置信息，准确率下降的情况正如预期的那样。通过将完整框架与缺少 IC 和

整个位置编码器，我们展示了两种设计的必要作用
适应不同的图像尺寸和相机内在特性。

可视化注意力：在图 5 中，我们可视化了交叉注意力得分 A_{12} ，并将没有先验知识指导的 ，
SRPose 与其完整设计进行了比较。我们
还绘制了高度关注的跨视图关键点对的连接，
表示SRPose建立的隐式对应关系。

图 5：隐式跨视图对应关系的可视化。点和线
用更亮的颜色绘制的表示更高的跨视图注意力分数。先前的
知识引导增强了对重叠区域的注意力，消除了不相关的跨视图连接，并协助建立隐含的对应关系。

5 结论

本文提出了一种基于深度学习的新型回归器 SRPose,利用
相机到世界双视图相对姿态估计的稀疏关键点和
物体到相机场景。作为我们的主要创新,SRPose 提取关键点
具有稀疏关键点检测器,并采用内在校准位置编码器,以适应不同的图像大小和相机内在特性。此外,我们提出的可提示先验知识引导的注意层建立了

隐式对应关系来估计两种情况下极线约束下的相对姿态。通过直接回归旋转和平移,SR-Pose 显著减少了计算时间,时间最短

减少 200ms。大量实验表明 SRPose 实现了
在两种场景中准确度和速度方面均达到最佳表现,
以及无地图视觉重定位任务。

致谢

该研究得到上海市科技重大专项（2021SHZDZX0102)和基础科研基金项目资助。

中央大学。

参考

1. Abouelnaga, Y.、Bui, M.、Ilic, S.:Distillpose:使用轻量级相机定位
   辅助学习。在:IROS（2021）
2. Agarwala, S.、Jin, L.、Rockwell, C.、Fouhey, DF:平面形成者:从稀疏角度看
   平面到 3D 重建。在:ECCV（2022 年）
3. 阿诺德,E.,韦恩,J.,维森特,S.,加西亚-埃尔南多,G.,蒙斯帕特,A.,Prisacariu,
   V.、Turmukhambetov、D.、Brachmann、E.:无地图视觉重定位:度量
   相对于单个图像的姿势。在:ECCV（2022）
4. Balntas, V.、Li, S.、Prisacariu, V.:Relocnet:连续度量学习迁移
   使用神经网络进行训练。见:ECCV（2018）
5. Bay, H.、Tuytelaars, T.、Van Gool, L.:Surf:加速稳健特征。在:ECCV
   (2006)
6. Bhowmik, A.、Gumhold, S.、Rother, C.、Brachmann, E.:强化特征点:
   优化高级任务的特征检测和描述。在:CVPR（2020）
7. Bradski, G.: OpenCV 库。Dr. Dobb s 软件工具杂志 (2000)
8. Cai, R.、Hariharan, B.、Snavely, N. 和 Averbuch-Elor, H.:极端旋转估计
   使用密集相关体积。 出处:CVPR（2021）
9. Calli, B.、Singh, A.、Walsman, A.、Srinivasa, S.、Abbeel, P.、Dollar, A. M.:ycb
   对象和模型集:面向操纵研究的共同基准。在:
   ICAR（2015）
10. Chang, A.、Dai, A.、Funkhouser, T.、Halber, M.、Niebner, M.、Savva, M.、Song, S.、
    Zeng, A.,Zhang, Y.:Matterport3d:从室内环境中的 rgb-d 数据中学习。在:3DV（2017 年）

11. Chang, J.、Yu, J.、Zhang, T.:用于局部特征匹配的结构化极线匹配器。在:CVPR（2023）

12. Chen, H.、Luo, Z.、Zhang, J.、Zhou, L.、Bai, X.、Hu, Z.、Tai, CL、Quan, L.：学习使用种子图匹配网络匹配特征。在：ICCV (2021)
13. Chen, H.、Luo, Z.、Zhou, L.、Tian, Y.、Zhen, M.、Fang, T.、Mckinnon, D.、Tsin, Y.、Quan, L.：Aspanformer：使用自适应跨度变换器的无检测器图像匹配。在：ECCV（2022 年）

14. Dai, A.、Chang, AX、Savva, M.、Halber, M.、Funkhouser, T. 和 Nießner, M.：Scannet：室内场景的富注释 3D 重建。刊于：CVPR（2017 年）
15. DeTone, D.、Malisiewicz, T.、Rabinovich, A.：走向几何深度满贯。arXiv preprint arXiv:1707.07410 (2017)

16. DeTone, D.、Malisiewicz, T.、Rabinovich, A.：Superpoint：自监督兴趣点检测和描述。刊于：CVPR（2018 年）

17. Dusmanu, M.、Rocco, I.、Pajdla, T.、Pollefeys, M.、Sivic, J.、Torii, A.、Sattler, T.：D2-net：一种用于联合描述和检测局部特征的可训练 cnn。刊于：CVPR（2019 年）

18. Edstedt, J.、Athanasiadis, I.、Wadenbäck, M.、Felsberg, M.：Dkm：用于几何估计的密集核化特征匹配。在：CVPR （2023 年）
19. Edstedt, J.、Sun, Q.、Bökman, G.、Wadenbäck, M.、Felsberg, M.：RoMa：稳健密集特征匹配。arXiv preprint arXiv:2305.15404 (2023)
20. Eggert, DW、Lorusso, A.、Fisher, RB：估算三维刚体变换：四种主要算法的比较。机器视觉与应用 (1997)

21. En, S.、Lechervy, A.、Jurie, F.：Rpnet：用于相对相机姿势估计的端到端网络。出处：ECCV（2018 年）

22. 范哲、潘鹏、王鹏、姜宇、徐丹、姜华、王哲：Pope：在任何场景中对任何物体进行 6 自由度可提示姿势估计，仅提供参考。arXiv 预印本 arXiv:2305.15727 (2023)

23. Fischler, MA、Bolles, RC: 随机样本共识：一种应用于图像分析和自动制图的模型拟合范例。COMMUN ACM (1981)

24. Gleize, P.、Wang, W.、Feiszli, M.：Silk：简单学习的关键点。出处：ICCV（2023 年）
25. Goodwin, W.、Vaze, S.、Havoutis, I.、Posner, I.：零样本类别级物体姿势估计。出处：ECCV（2022 年）

26. Hampali, S.、Rad, M.、Oberweger, M.、Lepetit, V.：Honnotate：一种 3d 的方法手部和物体姿势的注释。 出处：CVPR（2020）
27. Hartley, RI: 为八点算法辩护。TPAMI（1997）
28. 何凯、张鑫、任绍兴、孙建军：深度残差学习在图像识别中的应用。出处：CVPR（2016）
29. 何鑫、孙建、王颖、黄丹、鲍华、周鑫：Onepose++：无需 CAD 模型的关键点一次性物体姿态估计。NeurIPS（2022 年）

30. Hinterstoisser, S.、Lepetit, V.、Ilic, S.、Holzer, S.、Bradski, G.、Konolige, K.、Navab, N.：基于模型的训练、检测和无纹理 3D 物体在杂乱场景中的姿态估计。出处：ACCV（2013 年）

31. Huber，PJ：位置参数的稳健估计。数学年鉴统计年鉴 (1964)
32. Jin, L.、Qian, S.、Owens, A.、Fouhey, DF: 从稀疏视图。出处：ICCV（2021 年）
33. Kirillov, A.、Mintun, E.、Ravi, N.、Mao, H.、Rolland, C.、Gustafson, L.、Xiao, T.、Whitehead, S.、Berg, AC、Lo, WY,等人：分割任何东西。 arXiv 预印本 arXiv:2304.02643 (2023)

34. Laskar, Z.、Melekhov, I.、Kalia, S.、Kannala, J.：通过使用卷积神经网络计算成对相对姿势来实现相机重新定位。引自：ICCVW（2017 年）

35. 李柯、王玲、刘玲、冉倩、徐凯、郭燕:脱钩使
更好地监督局部特征。 出处:CVPR（2022）
36. Li, X.、Han, K.、Li, S. 和 Prisacariu, V.:双分辨率对应网络。
NeurIPS（2020）
37. Li, Z.、Snavely, N.:Megadepth:从互联网照片中学习单视图深度预测。出处:CVPR（2018 年）

38. Lin, A.、Zhang, JY、Ramanan, D.、Tulsiani, S.:Relpose++:从稀疏视图观测中恢复 6d 姿势。见:2024 年 3D 视觉国际会议 (3DV)。第 106-115 页。IEEE（2024 年）

39. Lindenberger, P.、Sarlin, PE、Pollefeys, M.:LightGlue:光速局部特征匹配。出处:ICCV（2023 年）

40. Liu, Y.、Shen, Z.、Lin, Z.、Peng, S.、Bao, H.、Zhou, X.:Gift:通过群组 CNN 学习变换不变的密集视觉描述符。
NeurIPS（2019 年）
41. Liu, Y.、Wen, Y.、Peng, S.、Lin, C.、Long, X.、Komura, T.、Wang, W.:Gen6d:基于 RGB 图像的可推广无模型 6 自由度物体姿态估计。引自:ECCV (2022)

42. Longuet-Higgins, HC:一种从两个图像重建场景的计算机算法
投影。《自然》（1981 年）
43. Loshchilov, I.、Hutter, F.:解耦权重衰减正则化。引自:ICLR（2018 年）
44. Lowe, DG: 基于尺度不变关键点的独特图像特征。IJCV（2004）
45. Luo, Z.、Zhou, L.、Bai, X.、Chen, H.、Zhang, J.、Yao, Y.、Li, S.、Fang, T.、Quan, L.:Aslfeat:学习精确形状和定位的局部特征。在:CVPR（2020）

46.Melekhov, I., Ylioinas, J., Kannala, J., Rahtu, E.:相对相机位姿估计
使用卷积神经网络。出处:ACIVS（2017）
47. Ni, J.、Li, Y.、Huang, Z.、Li, H.、Bao, H.、Cui, Z.、Zhang, G.:Pats:用于局部特征匹配的细分块区域传输。在:CVPR（2023）
48. Nistér,D.:五点相对姿态问题的有效解决方案。TPAMI
（2004 年）
49. Oquab, M.、Darcet, T.、Moutakanni, T.、Vo, H.、Szafraniec, M.、Khalidov, V.、Fernandez, P.、Haziza, D.、Massa, F.、El-Nouby, A. 等人:Dinov2:无需监督即可学习稳健的视觉特征。arXiv 预印本 arXiv:2304.07193 (2023)
50. Phil, W.:双向交叉注意力。https ://github.com/lucidrains/
双向交叉注意力（2022）
51. Ranftl、R.、Lasinger、K.、Hafner、D.、Schindler、K.、Koltun、V.:实现稳健的单目深度估计:混合数据集以实现零样本跨数据集传输。
TPAMI (2020)
52. Revaud, J.、Weinzaepfel, P.、de Souza, CR、Humenberger, M.:R2D2:可重复
和可靠的检测器和描述符。 引自:NeurIPS（2019）
53. Rockwell, C.、Johnson, J.、Fouhey, DF: 8 点算法作为归纳
通过 vits 预测相对姿势的偏差。在:3DV（2022 年）
54. Rublee, E.、Rabaud, V.、Konolige, K.、Bradski, G.:Orb:一种有效的替代方案
筛选或浏览。出处:ICCV（2011）
55. Sarlin, PE、DeTone, D.、Malisiewicz, T.、Rabinovich, A.:SuperGlue:使用图神经网络学习特征匹配。出处:CVPR（2020 年）
56. Savva, M.、Kadian, A.、Maksymets, O.、Zhao, Y.、Wijmans, E.、Jain, B.、Straub, J.、Liu, J.、Koltun, V.、Malik, J. ., et al.:栖息地:具体人工智能研究的平台。
出处:ICCV（2019）
57. Shi, Y.、Cai, JX、Shavit, Y.、Mu, TJ、Feng, W.、Zhang, K.:Clustergnn:基于聚类的粗到细图神经网络,用于高效特征匹配。在:CVPR（2022 年）

58. Sinha, S.、Zhang, JY、Tagliasacchi, A.、Gilitschenski, I.、Lindell, DB：稀疏姿势：稀疏视图相机姿势回归和细化。在：
IEEE/CVF 计算机视觉和模式识别会议。第 21349 页-
21359（2023年）
59. Smith, LN、Topin, N.：超收敛：神经网络的极快速训练
使用较大的学习率。在：人工智能和机器学习在多领域操作应用（2019 年）

60. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR：无检测器局部特征
与 transformers 匹配。CVPR（2021 年）
61. 孙建军，王志军，张绍刚，何鑫，赵华，张光国，周鑫：Onepose：
无需 CAD 模型的一次性物体姿态估计。出处：CVPR（2022 年）
62. Tan, B.、Xue, N.、Wu, T.、Xia, GS：Nope-sac：用于稀疏视图平面 3d 重建的神经单平面搜索。IEEE 模式分析与机器智能
汇刊 (2023)

63. 唐胜、张建、朱胜、谭平：用于视觉变换器的四叉树注意力机制。
ICLR（2022）
64. Tyszkiewicz, M.、Fua, P.、Trulls, E.：磁盘：使用策略梯度学习局部特征。NeurIPS（2020 年）

65. Wang, J.、Rupprecht, C.、Novotny, D.：姿势扩散：通过以下方式解决姿势估计
扩散辅助束调整。在：IEEE/CVF 国际会议论文集
计算机视觉会议。第 9773–9783 页（2023 年）
66. Wang, Q.、Zhang, J.、Yang, K.、Peng, K.、Stiefelhagen, R.：Matchformer：在 transformer 中交叉使用注意力机制
进行特征匹配。引自：ACCV (2022)
67. Wen, B.、Bekris, K.：Bundletrack：针对没有实例或类别级 3d 模型的新物体的 6d 姿势跟踪。在：IROS (2021)

68. Wen, B.、Tremblay, J.、Blukis, V.、Tyree, S.、Müller, T.、Evans, A.、Fox, D.、Kautz,
J.、Birchfield、S.：Bundlesdf：神经 6 自由度跟踪和未知物体的 3D 重建。在：CVPR（2023 年）

69. Winkelbauer, D.、Denninger, M.、Triebel, R.：通过合成训练数据学习在新环境中进行定位。引自：ICRA（2021 年）

70. Xiang, Y.、Schmidt, T.、Narayanan, V.、Fox, D.：Posecnn：卷积神经网络
网络用于在杂乱场景中进行 6d 物体姿态估计。机器人技术：科学与
系统 XIV（2018）
71. Xue, F.、Budvytis, I.、Cipolla, R.：Imp：迭代匹配和姿势估计
具有自适应池化。出处：CVPR（2023）
72. Xue, F.、Budvytis, I.、Cipolla, R.：Sfd2：语义引导特征检测和
描述。出处：CVPR（2023）
73. Yi, KM, Trulls, E., Lepetit, V., Fua, P.：提升：学习不变特征变换。
见：ECCV (2016)
74. Yu, J., Chang, J., He, J., Zhang, T., Yu, J., Wu, F.: 自适应点引导传输
前者用于一致的局部特征匹配。 出处：CVPR（2023）
75. Zhang, JY、Ramanan, D.、Tulsiani, S.：Relpose：预测概率相对
野外单个物体的旋转。欧洲计算机会议
Vision。第 592–611 页。Springer（2022 年）
76. Zhao, X., Wu, X., Chen, W., Chen, PCY, Xu, Q., Li, Z.: Aliked: 更轻的关键点
以及通过可变形变换实现的描述符提取网络。TIM（2023）
77. Zhao, X.、Wu, X.、Miao, J.、Chen, W.、Chen, PC、Li, Z.：相似：准确和
轻量级关键点检测和描述符提取。TMM（2022）
78. Zhou, Q.、Sattler, T.、Pollefeys, M.、Leal-Taixe, L.：学习还是不学习：视觉
来自基本矩阵的定位。在：ICRA（2020）

79. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: 论旋转表示的连续性
神经网络中的句子。 出处:CVPR（2019）
80. 朱胜,刘晓玲:Pmatch:用于密集几何的配对掩码图像建模
匹配。出处:CVPR（2023）

概述

这是补充材料SRPose:使用稀疏关键点的双视图相对姿势估计。在附录B中,我们提供了详细的定义
两种场景中解决的问题。在附录 C 中,我们进一步阐述了为什么以及如何在 SRPose 中强制执行内在校准。附录 D
在我们的实验中包含了更多的实现细节。附录 F 和附录 G 提供了更多我们实验结果和机制的可视化。

框架。在附录 H 中,我们讨论了 SRPose 的局限性,并提出了
未来研究的几个方向。SRPose 的代码可以在
https://github.com/frickyinn/SRPose/tree/main。

B 问题定义

我们的目的是估计两个视图之间的相对姿势变换
相机到世界和物体到相机的场景。估计的相对姿态
由旋转矩阵 R\in \mathcal {SO}(3) 和平移向量 t\in \mathbb 组成， ，
{R}^3 将 3D 世界点集 P_{\mathrm {w}} \subseteq \mathbb {R}^3 从相机坐标系映射到
第一幅图像 I_1 至第二幅图像 I_2：

(23)

其中 P_1 (分别为 P_2) \subseteq \mathbb {R}^3 表示投影到相机上的点 P_{\mathrm {w}}
空间中的 I_1 (分别为 I_2) 和 K_1, K_2 \in \mathbb {R}^{3\times 3} 表示相机本征
两幅图像。在相机到世界的场景中,当给定两幅重叠图像时
对于静态场景,SRPose 可以估计相机的姿态变换
I_1 至 I_2 其中,P_{\mathrm {w}} 表示场景中的静态 3D 点集。
另一方面,在物体到相机的场景中,当给定两张包含
I_1 中的多个对象和一个对象提示 b,用于标识目标对象 o
重点介绍一下,SRPose 估计了 o 之间的 6D 物体姿态变换
两个视图。在这种情况下,我们假设拍摄图像的相机是固定的,并且
P_{\mathrm {w}}表示运动目标物体o上的点的集合。

C 内在校准

图 6 展示了场景中同一 3D 世界点 p_\mathrm {w} 到
两个相机空间 I_1, I_2 由两个具有不同内在函数的相机组成。尽管
两个相机具有相同的外部参数T,即相同的相机位置和
方向,得到的坐标是不同的：

(25)

差异源于图像尺寸和相机内在特性的变化。
以前的回归器将所有输入调整为固定大小,从而改变像素或点
图像中的坐标,导致位置信息不准确。虽然 SR-Pose 也首先将所有图像调整为相同大小以进行批量关键点提取，

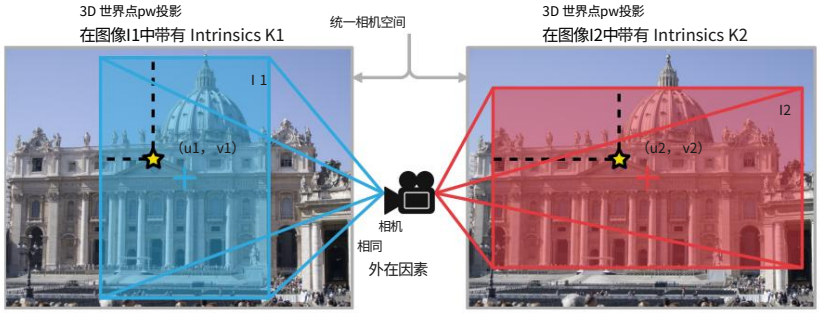


图 6:3D 世界点 p_\mathrm {w} \in P_\mathrm {w} 由两个具有相同外部参数的相机拍摄
T 但内在函数 K_1, K_2 不同 ,导致两幅图像的坐标不同

进行内在校准,以使关键点的坐标不失真,
将它们标准化到统一的相机空间。

然后将检测到的关键点的坐标重新调整到其原始位置,
这样就可以消除图像大小变化的影响。现有回归器使用的图像编码器仅处理相机中的位置图像

坐标系,即 P_1,P_2,不考虑不同的内在函数。然而,
相机本征矩阵 K_1、K_2 中的不同值是由两个不同的设备（例如手机和数码单镜头相机)引起的,

将 P_w 投影到相机空间中的不同位置。这种差异阻碍了
由于由此产生的不准确的
位置关系。为此,SRPose 根据以下条件强制执行内在校准:
至 [48]。利用内在函数对关键点坐标进行去扭曲,可以实现精确的
对应设立职位信息:

(27)

内在校准将不同相机空间中的点归一化到统一的相机坐标系,从而确保相机不变、可推广、

并在相对姿态估计方面表现准确。

D 实施细节

D.1 微调

为了获得更好的性能,SRPose首先在ScanNet [14]上训练了500
epoch 作为预训练阶段。然后我们在其他
数据集,包括 Matterport [10]、Linemod [30]、HO3D [26]、Niantic [3] 和
MegaDepth [37]。由于只有 Matterport 和 Niantic 拥有验证集,
我们选择模型检查点和在验证集上表现最好的超参数。对于其他数据集,我们选择 2\times 10^{-5} 作为

1cycle 策略的最大学习率（LR)[59]用于微调。如 ScanNet
和 MegaDepth 的训练集中包含超过 100 万张图像对,我们
将训练周期设置为 500,以充分学习信息,而其他人

200. 由于内存限制,我们在所有数据集上提取了 1,024 个关键点
在训练阶段,除了 Linemod。因为 Linemod 中的目标对象只
占用一小部分像素,我们提取了 1,200 个关键点,以方便
训练。值得注意的是,我们训练我们的框架时,使用随机权重初始化
不同稀疏关键点的消融研究及对比实验
检测器。表7列出了不同数据集上的超参数选择。

表 7:不同数据集上的超参数选择。

数据集	最大 LR 周期	
马特波特 [10]	—	200
扫描网 [14]	—	500
巨深 [37]	—	500
HO3D [26]	—	200
線型 [30]	—	200
奈安蒂克 [3]	—	200

D.2 实验设置

在本节中,我们将详细介绍我们在正文和补充材料中使用的数据集和实验设置。然后，

我们解释了实验中使用的一些复杂指标。

相机到世界的姿态估计:在相机到世界的场景中,我们在 Matterport [10]、ScanNet [14] 和 MegaDepth [37] 上评估了我们的 SRPose。Mat-terport 是一个使用 Habitat [56] 系统从真实场景中重新渲染的数据集。
我们采用[32]以下的预处理数据集进行训练和评估。
该数据集包含 31,932/4,707/7,996 张训练/验证/测试图像
测试集的平均旋转角度为 53.5 °, 和
平均翻译长度为 2.31 米。ScanNet 是一个由真实
场景,包含 1613 个单目序列。遵循 [55] 中的指导原则，
我们抽取了 230M 幅图像对进行训练,并使用 ScanNet-1500 测试集进行
评估。ScanNet-1500 测试集的平均旋转角度为 29.6 °和
平均翻译长度为 0.88 米。MegaDepth 是一个包含以下内容的数据集：
196 个不同户外场景的 100 万张互联网图像。我们使用 MegaDepth-1500
测试集遵循[60,64],其中包括从训练中排除的两个场景
集合:圣心大教堂和圣彼得广场。测试集的平均旋转
12.7 度 °平均平移长度为2.46米。

物体到相机的姿势估计:在物体到相机的场景中,我们
在 HO3D [26] 和 Linemod [30] 上评估 SpaRelPose。HO3D 包含视频
YCB 数据集中 21 个类别的手持物体的检测结果 [9]。每个视频
描绘了人类手持单个物体并改变其姿势的过程，
固定相机和背景。Linemod 包含 15 个类别的对象，
整个训练集是合成数据,测试集是真实场景
数据。Linemod 中的每个图像包含多个任意分散的对象
各种场景。Linemod 提供数据集中每个视图的深度图,这些深度图
由 LiDAR 深度相机高精度捕捉。边界框和

我们还提供了目标物体的对象分割。对于两个数据集,我们从训练集中随机选择相对旋转角度小于 45° 的图像对进行训练,旋转角度小于 30° 的图像对进行评估。对于 HO3D,我们从训练集中排除的五个视频中选择了 3000 个帧对作为新的测试集,平均旋转角度和平均平移长度为 17.2 °和 0.12 m。测试集包含训练集中出现的三类无纹理物体,以及规则形状、看不见的和有纹理的物体两类。对于 Linemod,我们从真实场景数据中随机挑选了 1500 个图像对作为新的测试集,平均旋转角度和平移长度为 20.7°和0.88 m。

对于基于匹配器的基线,我们首先从两个视图中裁剪出目标对象。
然后将得到的两幅裁剪图像调整大小,使其较大的尺寸为 640 像素。我们利用对象分割屏蔽掉背景上的匹配特征,最后从基本矩阵中恢复相对姿势,就像在相机到世界的场景中一样。

无地图视觉重定位:我们在 Niantic [3] 数据集上评估了无地图重定位任务中的 SRPose。Niantic 是一个专门为无地图视觉定位任务构建的数据集,由训练集、验证集和测试集分别包含 460/65/130 个场景组成。对于验证集和测试集,提供了单个参考图像和一系列查询图像。目标是根据参考重新定位查询位置。测试集的地面实况不公开,该数据集的评估是通过在项目页面上提交我们的 SRPose 估计的相对姿势来执行的。

不同指标的评估分数将由服务器衡量。

度量:我们采用平均距离度量,即 ADD 和 ADD-S,按照 [30] 来评估物体到相机场景下的性能。给定真实旋转和平移 R_{gt}, t_{gt} 和估计的 R, t,ADD 计算根据真实值和估计值变换的 3D 物体模型点之间的成对距离的平均值:

,

—— —— (28)

其中 \protect \mathcal {M} 为 3D 模型点集,即物体点云,m 为点数。为了评估对称物体,例如 Mug 和 Banana,ADD-S 也使用最近点距离计算:

—— —— (29)

按照[70]的方法,我们计算并报告了ADD和ADD-S的曲线下面积(AUC),阈值设置为10cm。

我们按照 [3] 中的要求,使用虚拟对应重投影误差 (VCRE) 度量来评估无地图视觉重定位中的 SRPose。地面实况和估计的相对姿态变换用于将虚拟

表 8:Linemod 上的相对物体姿态估计 [30]。SRPose 在
翻译估计,同时在轮换比较中取得竞争性表现
基于匹配器的方法加上 LiDAR 深度图和物体分割。

方法	对象 (厘米)		破碎的(^\循环)				
SGM网 [12]	不适用/a/a/a/a/	不适	24.16 40.1 60.1 22.24 40.46			不适	无
光胶 [39]	a/a/a	用/—	62.7 21.17 41.38 64.1 12.14			用/—	无
洛夫特 [60]		个/—	23.24 78.3			个/—	无
德克曼 [18]		个/一个				个/一个	无
SGMNet [12] + 深度 11.8 24.4 LightGlue [39] + 深度 13.0 24.3 LoFTR		45.2	9.7	17.2	86.0 20.3 83.5		35.1
[60] + 深度 10.1 19.2 DKM [18] + 深度 6.0 10.2		43.3	9.7	20.4	23.6 85.6 25.8		38.1
		49.9	7.8	22.6	96.5 37.8		41.7
		67.7	4.6	7.7			56.5
同步姿势	13.8 16.1	29.7	9.1	10.5 98.7 10.2			27.8

3D 点,位于查询相机的局部坐标系中。VCRE 是
重新投影误差的平均欧几里得距离:

$$\text{——} \tag{30}$$

其中 \pi 是图像投影函数,\protect \mathcal {V} 是相机中的一组 3D 点
空间代表虚拟物体。该指标提供了直观的衡量标准
无地图重定位任务中的 AR 内容错位。

E 附加结果

在本节中,我们将展示更多实验结果,以进一步评估我们框架的优势和局限性。本节包括 Linemod [30]
和 MegaDepth [37] 数据集上的两个额外评估结果,这些结果被排除在外

从正文来看,关于 HO3D 的进一步报道 [26],以及
使用不同稀疏关键点检测器的 SRPose 变体。

E.1 物体到相机姿态估计

Linemod 上的结果:如表 8 所示,SRPose 在 Linemod 上表现不佳
与其他基于匹配器的方法相比,ADD 和 ADD-S 中存在差异。尽管我们的
框架在旋转估计方面表现出了有竞争力的结果,但翻译的失败阻碍了整体性能的更高准确度。首先,

高精度 LiDAR 深度为基线方法提供了更多关于 6D 姿态变换的信息。其次,Linemod 图像中的目标对象
仅
在场景中的多个物体中占据一小部分像素。SRPose
只能从每幅图像中的物体提示中提取大约 100 个关键点,
这使得任务变得困难。虽然基于匹配器的方法需要在两个图像中进行对象分割,并且它们将分割后的对象
调整为更大的尺寸,以便
获得更高的准确率。尽管如此,SRPose 仍然取得了有竞争力的表现
旋转,其中一个视图中只有一个对象提示。

HO3D 上的附加结果:表 9 显示了不同类别的结果
在 HO3D 中。通常,所有方法在具有丰富
纹理,因为它有助于建立对应关系。在表 10 中,我们还比较了
使用正交 Procrustes [20] 的基线,无需深度图
使用基本矩阵。我们只评估旋转的表现,因为
传统方法无法进行缩放翻译估计,而无需
深度。虽然立体深度与 LiDAR 深度相比精度较低,但
仍然有助于匹配器的相对 6D 物体姿势估计。

表 9:HO3D [26] 上的分类相对姿态估计结果。003:
饼干盒;006:芥末瓶;011:香蕉;025:马克杯;037:剪刀。

方法	添加/添加-S				
	003	006	011	025	037
SGMNet [12] 16.5 / 36.7 12.4 / 24.1 17.6 / 30.1 17.1 / 32.3 22.7 / 37.5					
光胶 [39] 16.6 / 36.9 12.6 / 23.9 17.1 / 29.7 17.9 / 31.8 22.4 / 34.6					
低利率周转率 [60] 16.8 / 36.9 12.2 / 23.9 18.2 / 30.8 15.2 / 29.6 20.6 / 35.6					
德克曼 [18] 16.6 / 36.7 12.8 / 24.9 18.6 / 32.0 18.0 / 34.4 24.0 / 38.7					
同步姿势	44.9 / 68.2 55.6 / 75.1 21.7 / 35.8 36.2 / 60.1 22.5 / 44.9				

表 10:HO3D 上有深度和无深度的比较 [26]。深度图
协助基于匹配器的方法估计相对物体姿势变换。

方法	旋转(^\circ)			
	Med.误界 "3223379			
SGM网 [12]	27.7	36.0	55.1	23.4
SGMNet[12]+深度	24.1	29.0	62.7	29.1
LightGlue [39] 28.0 LightGlue [39] + 深度 24.4		38.2	54.1	23.9
		28.7	61.7	29.1
洛夫特 [60]	33.5	66.8	45.7	20.5
LoFTR [60] + 深度	24.4	28.7	61.7	29.1
德克曼 [18]	27.9	34.9	54.1	24.0
DKM [18] + 深度	23.3	25.6	64.8	29.9
同步姿势	8.9	11.4 95.4		73.3

E.2 物体与世界姿态估计

MegaDepth 上的结果:我们进一步评估了 MegaDepth 上的室外相对姿态估计 [37]。表 11 显示
SRPose 在 MegaDepth 上的表现不佳。
我们讨论了我们的框架在这种表现方面的局限性
附录 H。

E.3 不同的稀疏关键点检测器

SRPose 可以采用不同类型的方法作为其稀疏关键点检测器,
包括经典方法 SIFT [44] 和基于深度学习的检测器,
例如 DISK [64]、ALIKED [76]、SuperPoint [16] 等。表 12 展示了使用不同方法作为稀疏关键点的
Matterport 上的 SRPose 性能
探测器。因此,SuperPoint 优于其他探测器,这是
我们选择默认检测器在其他数据集上评估我们的框架。

表 11:MegaDepth 上的相对姿态估计 [37]。SRPose 并不擅长
估计小的相对姿势变换。

类别	方法	姿势估计 AUC @5 °	@10 。	@20 。
疏	强力胶 [55]	49.7	67.1	80.6
	SGM网 [12]	43.2	61.6	75.6
	光胶 [39]	49.4	67.0	80.1
稠密	洛夫特 [60]	52.8	69.2	81.2
	ASpanFormer [13]	55.3	71.5	83.1
	德克曼 [18]	60.4	74.9	85.1
回归器	无地图 [3]	2.6	9.3	22.9
	同步姿势	20.5	43.1	65.1

表 12:SRPose 在 Matterport [10] 上进行的相对姿态估计
不同的稀疏关键点检测器。SuperPoint [16] 的表现优于其他检测器。

方法	破碎的(^\循环)			传输 (米)		
	Med.\定界 °3223379					
8分 [53]	8.01	19.13	85.4	0.64	1.01	67.4
SRPose + SIFT [44] 7.04 18.19 84.8 SRPose + DISK [64] 7.73 19.96 84.0 SRPose + ALIKED [77] 5.65 15.92 88.1 SRPose + SuperPoint [16] 5.57 14.32 88.9				0.57 1.00 68.9 0.60 1.03 66.6 0.50 0.89 73.3 0.47 0.84 74.2		

F 定性结果

图 8 展示了 SR-Pose 在 Matterport [10] 和 ScanNet [14] 上进行相机到世界姿态估计的定性结果。图 9 展示了 SR-Pose 在 Matterport [10] 和 ScanNet [14] 上进行相机到世界姿态估计的定性结果。
我们的框架在 HO3D 上对物体到相机姿态估计的结果 [26]。

G 可视化

我们展示了相机到世界和物体到相机场景中交叉注意力得分的更多可视化结果。图 10 可视化了更多交叉注意力
ScanNet [14] 中两种视图的得分。此外,我们还展示了
每种情况下的 SRPose 内的相似度矩阵。如图所示并在正文中讨论,交叉注意力得分在

场景的重叠区域。值得注意的是,相似性矩阵通常在图像的边缘或角落附近显示高值,从而引导交叉注意力
模块专注于这些信息关键点,并促进建立隐式对应关系。图 11 可视化了交叉注意力得分

在 Linemod [30] 和 HO3D [26] 上对物体到相机场景的检测中。
对象提示在参考图中（用橙色矩形表示）,SR-Pose 会自动在同一张图像上搜索隐式对应的关键点
查询视图中的目标对象。使用对象提示可以实现相对姿势
在没有物体分割的情况下,物体到相机场景中的估计。

H 局限性和未来研究

小姿态变换:SRPose 在估计
小的姿势变换,如附录 E.2 中所述。值得提醒的是

我们的框架建立对应关系并使用神经网络隐式求解极约束方程。SRPose 接收编码的位置信息,这些信息用于通过多个

层。然而,在这个过程中,精确的位置信息可能会被低估,导致姿态估计精度有所下降。相比之下,传统的基于匹配器的方法在匹配局部特征方面表现出色。

高度重叠的图像对,通常对应于微小的变换。
通过明确解决具有最小噪声和异常值的约束,这些方法可以产生高度准确的结果。这解释了为什么基于匹配器的
基线模型通常优于包括 SRPose 在内的神经网络回归模型,
MegaDepth 是一个由小姿态变换组成的数据集,如表 11 所示。进一步的定量分析表明,性能不佳也是由于
在较小的姿势误差阈值下,SRPose 的精度相对较低。MegaDepth
平均姿势变换比 ScanNet 要小,表明视角相似
基于匹配器的方法更容易匹配关键点。通过直接使用
极线约束,匹配器可以在小姿势变换上产生较低的误差。而回归方法(包括 SRPose)近似解

通过神经网络,导致小阈值下的精度较低。
然而,如图 7 所示,与匹配器相比,SRPose 在更大的阈值下实现了具有竞争力的精度或更高的精度。尽管 SRPose
在精度曲线下的累积面积上,MegaDepth 表现不佳
(AUC),进一步的分析仍然显示出其在精度方面的有效性。SR-Pose 利用语义信息和连接来隐式去噪

在这些困难的情况下,异常值可以提高准确性。一个需要进一步研究的领域是
研究可以最大限度地减少位置信息精度的损失
通过神经网络层传播。

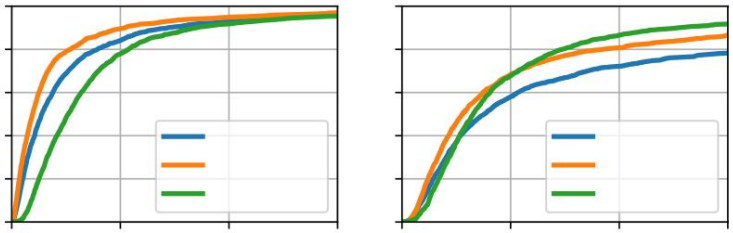


图 7:MegaDepth [37]和 ScanNet [38] 上的精度曲线,其中 Area
计算曲线下面积(AUC)。

相对物体姿态估计:估计相对 6D 物体姿态变换一直是基于匹配器和基于回归器的方法的挑战。当前的视频物体姿态跟踪框架 [67,68] 解决了

挑战首先通过估计两个相邻帧之间的粗略姿势来
借助视频对象分割模型,然后估计粗
使用全局姿势图优化来优化姿势,从而大大提高