arXiv:2303.02615v2 [cs.CV] 8 Mar 2024

使用级联注意力估计极端三维图像旋转

Shay Dekel 巴伊兰大学 拉马特甘，以色列
shaydekel@gmail.com

Yosi Keller 巴伊兰大学 拉马特甘，以色列yosi.keller@gmail.com

ˇMartin Cad´ık FIT，布尔诺理工大学 布尔诺，捷克共和国
cadik@fit.vut.cz

2024年3月12日

摘要

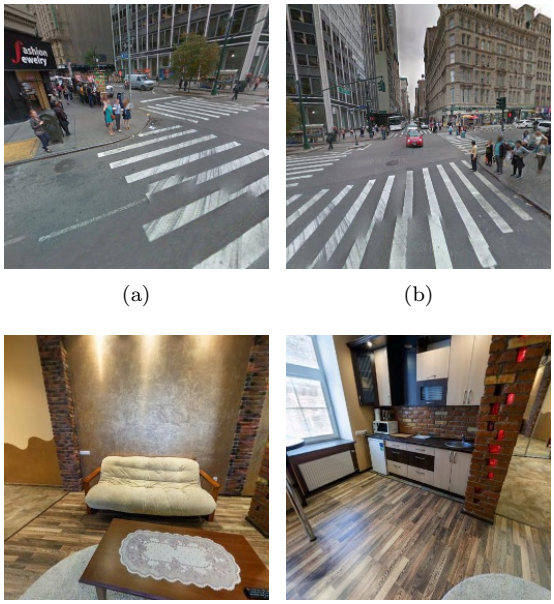
估计大幅度、极端的图像间旋转，对于众多涉及视场有限或不重叠图像的计算机视觉领域至关重要。在本文中，我们提出了一种基于注意力的方法，其处理流程包含一系列新颖的算法组件。首先，由于旋转估计针对的是图像对，我们引入了一种利用解码器的图像间蒸馏方案，以改进嵌入。其次，尽管当前方法会计算用于编码图像间关系的四维相关体（4DCV），我们提出了一种基于编码器的激活图交叉注意力方法，以计算增强版的等效四维相关体。最后，我们提出了一种基于级联解码器的技术，用于交替细化交叉注意力和旋转查询。我们的方法在极端旋转估计方面优于当前最先进的方法。我们公开了代码[1]。

1 引言

估计一对图像之间的相对位姿是计算机视觉中的一项关键任务，可用于室内导航、增强现实、自动驾驶、三维重建 [44, 40]，相机定位 [5, 45, 47]，同步定位与建图 [12, 38]，以及新视角合成 [35, 42]等多种应用。当前的图像配准方法包括提取特征、匹配特征，并建立它们之间的对应关系。然而，对于重叠程度很低或完全没有重叠的输入图像对，这种方法效果不佳，因此难以建立足够的特征对应关系来进行匹配，例如图中所示的图像1。

众多应用 [49, 32, 1] 都需要精确估计图像间旋转。如图1所示，对于重叠程度有限或没有重叠的图像之间的极端三维旋转估计，普遍采用的方法源自 Coughlan and Yuille 的开创性工作 [10]。他们提出了一种以图像中的线性结构为基础的技术，这些结构主要源自三个相互正交的方向——一个垂直方向（建筑墙面）和两个水平方向（地面铺装、道路等）。类似地，Criminisi 等人的“单视图测量” [11] 及其扩展方法[61, 26, 41] 利用平行图像线及其对应的消失点 [19]进行相机标定。此外，还可以通过分析光照和阴影方向，利用光照线索 [2], 估计相对相机旋转。

在这项工作中，我们提出了一种深度学习方法，用于估计图像间显著且极端的旋转。不同于经典方法 [10, 11]显式检测直线、阴影和消失点等人工设计的线索，我们的方法通过深度神经网络直接从输入图像回归相对旋转。受 Transformer 在 [53] 目标检测等计算机视觉任务 [8] 和图像识别 [24]，中的近期成功应用启发，我们将 Transformer 应用于所提出的处理流程中的多项任务，如图2所示，进一步拓展了此前的应用范围



(a) (b) (c) (d)

图1：极端三维图像旋转的估计。第一行：小重叠区域的图像对。第二行：不重叠的图像对。所提出的方法估计图像对之间的相对旋转。

Transformer。

首先，我们通过交叉解码，在图像之间提炼图像间信息，从而改进输入图像嵌入；其中，每个嵌入都将另一个嵌入用作查询。这样可以更好地对图像进行相互关联的编码。其次，Transformer编码器计算堆叠的多头注意力，以编码图像对潜在表示之间的交叉注意力。因此，它改进了先前工作中使用的四维相关体（四维相关体） [51, 39, 30, 21, 15]，后者通过内积计算四维相关体。不同于四维相关体中使用的单层 N^2 内积，所提出的基于 Transformer 编码器的方法利用多头注意力的先进架构，更好地编码激活图条目之间的交互。第三，我们进一步改进采用由两个解码器和一个学习得到的旋转查询组成的级联结构进行交叉注意力编码，从而联合优化交叉注意力编码和旋转查询。所提出的方法是一种通用的注意力架构，可用于估计与两幅输入图像相关的属性，例如光流、配准、相对姿态回归等。本研究的动机是极端旋转估计；其他应用留待未来研究，因为这些应用需要进行额外的任务特定修改。

有趣的是，我们的方法计算出的注意力图（见第3.2节）表明，Transformer编码器会对包含旋转信息图像线索的区域分配较高的注意力分数，突出垂直线和水平线。我们还发现，所提出的方法能够以最先进（SOTA）的准确率预测不重叠图像对的旋转。我们的框架可进行端到端训练，并通过回归损失进行优化。我们在三个数据集基准上对其进行了评估：StreetLearn [36], SUN360 [54] 和 InteriorNet [29]，涵盖室内和室外场景中的不同重叠类别以及不同光照条件。第4 节的实验结果表明，我们的模型达到了最先进（SOTA）的准确率。

总之，我们的贡献如下。

- 我们提出了一种用于估计极端旋转的新方案，适用于图像重叠程度极低的场景。

- 通过交叉解码增强图像嵌入，从而提炼图像间信息。

- 我们提出了一种 Transformer编码器交叉注意力机制，用于编码图像对之间的潜在空间交互。

- 解码器-解码器模块通过学习并应用四元数旋转查询，从交叉注意力编码中推断相对旋转。

- 定量评估表明，与当前先进的旋转估计技术相比，该方法在室内和室外数据集上均表现出色。

2 相关工作

我们提出的旋转估计方法是相对位姿估计这一更一般问题的一个特例，尤其是相对位姿回归（RPR）。常见的相对旋转估计技术会检测并匹配图像中的二维特征点 [，例如尺度不变特征变换 [34], SURF [4]]。对于位姿定位任务 [50], ，PnP方案会估计相对三维旋转；给定锚定图像的三维坐标和位姿后，即可确定查询图像的相机位姿。其他三维旋转估计方案则利用三维傅里叶变换 [22, 23], ，其幅值具有平移不变性。近期方法将可端到端训练的深度网络应用于两幅图像 [3, 14]。图神经网络（GNN）通过聚合视频帧中的定位线索，实现了多图像相对位姿回归 [55, 52]。神经辐射场（NeRF）也被探索用于相对位姿回归编码，作为传统图像或特征点存储方式的替代方案。有些方案采用针对旋转设计的参数化表示，尤其是四元数和欧拉角，来估计相对三维旋转 [59]。这类参数化表示，尤其是四元数，能够解决旋转表示中固有的不连续性问题，这种不连续性源于双覆盖性质。Levinson等人 [27] 研究了通过神经网络进行三维旋转估计时采用的SVD正交化方法。他们将推断出的旋转矩阵通过SVD投影到旋转群上，并表明这种方法优于传统表示，在多种深度学习范式中推动了技术发展。此外，Mohlin等人 [37] 提出了一种基于神经网络的方法，用于估计表示三维旋转概率分布的Fisher分布矩阵参数。通过优化该分布的负对数似然损失，他们在多个真实世界数据集上超越了先前的基准。类似地，Liu等人也采用了这种优化方法来估计头部姿态 [33]。作为一个值得关注的基线方法，Rockwell等人 [43] 设计了一种视觉Transformer（ViT），用于近似八点算法，以直接估计两幅图像之间的相对位姿，并在多种场景中展现出具有竞争力的性能。上述方法主要依赖于子输入图像对之间有显著的重叠程度。明显的旋转会导致重叠程度有限，从而影响这些估计的准确性。具体而言，这类技术无法有效对齐互不重叠的图像。Caspi和Irani [9] 证明了对齐两个在时间和空间维度上视场均不重叠的图像序列是可行的，前提是相机彼此接近。这种对齐利用了序列间共享的时间变化。类似地，Shakil [46] 证实，由未经标定的视频摄像机拍摄的多个互不重叠的视频序列，可以通过序列中固有的时间波动和帧间运动实现同步。超越单纯图像范畴，配准互不重叠的RGB-D扫描[48, 20]也是一个值得关注的衍生问题。在这种情况下，通常会推断出场景的整体表示。在本研究中，我们采用Cai等人 [7], 提出的框架，任务是根据一对输入图像估计极端相对三维旋转。Cai等人提出了一种方案，先使用卷积神经网络对图像进行嵌入，随后根据嵌入计算四维相关体（4DCV）。然后将多层感知机应用于该相关体，并通过交叉熵损失对其进行优化，从而在不重叠图像上取得最先进（SOTA）的准确率。从本质上说，4DCV是双线性池化[31, 25], 的一种扩展，用于编码与图像对对应的二维嵌入图中所有实体之间的成对相关性。凭借这种编码能力，4DCV已应用于需要远距离空间对应关系的任务，例如RAFT这一最先进的光流模型 [51] 以及其他光流模型 [21, 15, 56]。在相关领域，三维相关体已用于深度立体匹配任务 [39, 30, 18, 28], ，其中一幅图像中的像素与另一幅图像中受限空间支持域内的像素进行匹配。与这些方法不同，我们提出通过多头Transformer编码器计算图像对激活图之间的交叉注意力，并结合相应的激活掩码，从而构造类似的四维相关体。具体而言，该Transformer编码器通过多头注意力有效实现多个聚合

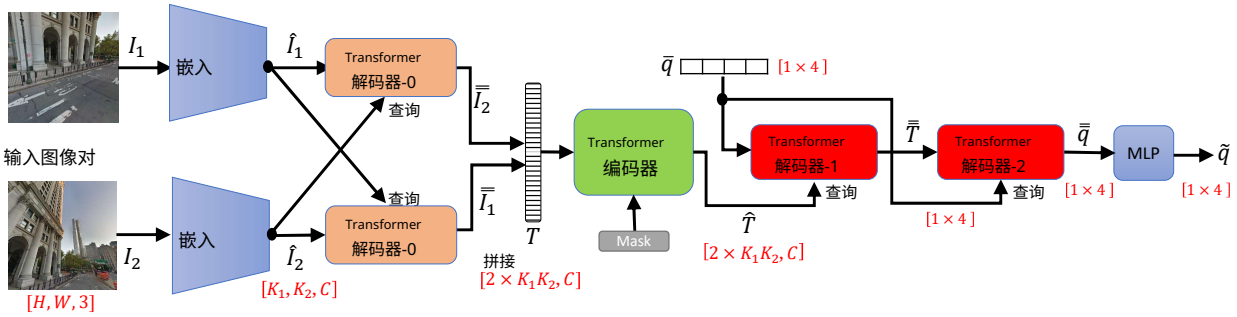


图2：**所提出的架构**利用权重共享的孪生卷积神经网络对输入图像对（I_1，I_2）$\in \mathbb{R}^{H\times W}$进行编码，得到特征图（$\hat{I}_1$，$\hat{I}_2$）。随后，权重共享的Transformer 解码器-0 层对这些特征图进行交叉解码，将（I_1，I_2）交叉蒸馏为表征 I_1 和 I_2。拼接后的精炼嵌入 T 与注意力掩码 M 一同输入Transformer编码器，以推导出交叉注意力编码 T。T 进入由两个Transformer解码器构成的级联结构，其中第一个解码器Transformer 解码器-1在学习得到的四元数旋转查询 $\bar{q}$的引导下，将交叉注意力增强为 T。第二个解码器Transformer 解码器-2将旋转编码为 $\bar{\bar{q}}$，再通过多层感知机（MLP）进行变换，以预测相对四元数旋转 $\tilde{q}$。

门控相关体积 [53] 操作。此外，我们提出使用Transformer解码器进行图像间嵌入蒸馏，并通过级联交替旋转解码改进旋转推理。

3 使用级联注意力进行旋转估计

所提出的方法估计输入图像对$\mathbf{R} \in \mathbb{R}^{3\times3}$ 之间的相对三维旋转$\mathbf{I}_1, \mathbf{I}_2 \in \mathbb{R}^{H\times W}$，如图2所示。采用权重共享的孪生残差 U-Net [57] 将输入编码为激活图 $\hat{\mathbf{I}}_1$，$\hat{\mathbf{I}}_2 \in \mathbb{R}^{c\times K_1\times K_2}$，其中 c 表示通道数，K_1和 K_2 表示空间维度。为通过交叉$\mathbf{I}_1$,$\mathbf{I}_2$ 解码改善嵌入，$\mathbf{I}_1$,$\mathbf{I}_2$ 被交叉传播至权重共享的 Transformer 解码器-0 单元中。每个输入嵌入提取与任务相关的表征$\bar{\mathbf{I}}_1$和$\bar{\mathbf{I}}_2$。

为了进一步关联两幅输入图像，我们计算交叉注意力 $\mathbf{T}$，它是对先前工作中使用的四维相关体（4DCV）[51, 39, 30, 21, 15]的增强等价形式。因此，我们将其向量化这些行 $\bar{I}_1$ 和 $\bar{I}_2$ 作为两个序列 $\in \mathbb{R}^{c\times K_1K_2}$，并将它们拼接为一个张量 $T\in \mathbb{R}^{c\times 2K_1K_2}$，然后如第3.2节所述，应用 Transformer编码器。旋转通过一种新颖的基于注意力的架构进行解码，该架构采用了由两个 Transformer 解码器构成的级联结构，其依据是交叉注意力T。**首先，使用 Transformer解码器-1**将T**作为查询，并在可学习的四元数** $\bar{\mathbf{q}} \in \mathbb{R}^4$ 的引导下对其进行增强，四元数作为输入。随后，将 Transformer解码器-1的输出（记为$\mathbf{T}$）作为后续 Transformer 解码器-2的输入，以进一步优化查询 $\bar{\mathbf{q}}$。最后，通过全连接多层感知机层预测以四元数 $\tilde{q}$表示的相对旋转。按照第**3.4**节的讨论，使用其四元数表示q对三维旋转进行回归。

3.1 通过交叉解码进行图像嵌入蒸馏

给定输入图像的嵌入 $\hat{\mathbf{I}}_1$，$\hat{\mathbf{I}}_2 \in \mathbb{R}^{c\times K_1\times K_2}$，我们的目标是通过蒸馏输入图像之间的信息来细化这些嵌入。为此，我们将 Transformer-解码器应用于每个嵌入，其中另一个嵌入被用作

解码器的查询。为了将激活图转换为兼容 Transformer 的输入，我们遵循与 [8]相同的序列准备过程。激活图 $\widehat{\mathbf{I}}_1$，$\widehat{\mathbf{I}}_2 \in \mathbb{R}^{c\times K_1\times K_2}$ 首先被展平为序列表示 $\mathbf{I}_1$,$\mathbf{I}_2 \in \mathbb{R}^{c\times K_1 K_2}$。随后，为激活图中的每个位置分配一个学习得到的编码，以保留每个位置的空间信息。为了减少参数数量，我们分别为 X、Y 轴学习两个一维编码。具体而言，对于激活图 $\mathbf{I}$ ，我们定义位置嵌入向量集合 $\mathbf{E}_u \in \mathbb{R}^{K_1\times C/2}$ 和$\mathbf{E}_v \in \mathbb{R}^{K_2\times C/2}$，使得空间位置（$i$，$j$），$i \in 1..K_1$，$j \in 1..K_2$通过拼接两个对应的嵌入向量进行编码：

$$\mathbf{E}^{i,j}_{pos} = \begin{bmatrix}\mathbf{E}^i_u \\ \mathbf{E}^j_v\end{bmatrix} \in \mathbb{R}^C. \tag{1}$$

经过处理的序列作为 Transformer 的输入，因此可表示为：

$$\widehat{\mathbf{I}} = \widehat{\mathbf{I}} + \mathbf{E}_{\mathbf{A}} \in \mathbb{R}^{K_1 K_2\times C}, \tag{2}$$

其中 $\mathbf{E}_{\mathbf{A}}$ 是$\mathbf{I}$的位置编码。

3.2 使用 Transformer编码器计算交叉注意力

输入图像表示细化过程之间的交叉注意力 $\bar{I}_1$ 与 $\bar{I}_2$ 通过一个包含 $l = 2$ 层和 $h = 4$ 每层 个注意力头的 Transformer编码器计算。第 4.6 节4.6给出了对这一配置的消融研究。Transformer编码器计算得到的交叉注意力图是对 4D 相关体的一种改进等价表示 [51, 7], ，用于编码激活图中所有图像线索之间的交互（内积）。默认情况下，Transformer编码器会计算输入序列的自注意力图。因此，通过应用式 ˆ3 中给出的注意力掩码 $\mathbf{T}$ ，即可计算向量化并拼接后的激活图 T 的交叉注意力 $\mathbf{M}$ 。掩码 3M **会将整个 Transformer编码器中计算得到的注意力图里的自注意力项置零，同时保留交叉注意力项**

[1] https:/anonymous.4open.science/r/AttExtremeRotation-A467/

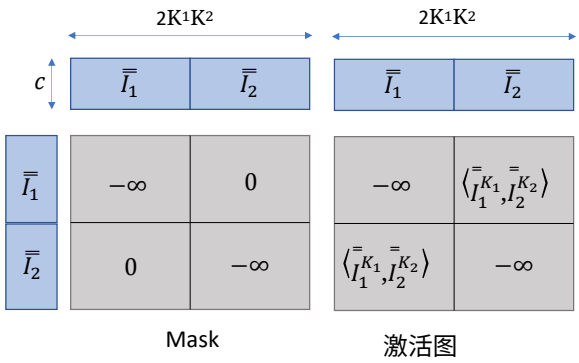


图 3：使用 Transformer编码器和输入掩码$\mathbf{M}$，计算交叉注意力。掩码$\mathbf{M}$将自注意力项置零，仅保留交叉注意力项。

Transformer编码器，同时保留交叉注意力项，

$$\mathbf{M} = \begin{bmatrix} -\infty & -\infty & \cdots & 0 & 0 \\ -\infty & -\infty & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \cdots & -\infty & -\infty \\ 0 & 0 & \cdots & -\infty & -\infty \end{bmatrix} \tag{3}$$

掩码的使用$\mathbf{M}$ 以及注意力图的相应结构如图3所示。任意一对图像块都可能包含有关图像整体几何关系的宝贵信息。Transformer编码器能够隐式地发掘这些线索。如补充材料第 2 节所示，输入图像中包含与旋转相关线索的区域，无论这些线索是显式还是隐式的，都会获得更高的注意力分数。这为蒸馏和后续的多层感知机层提供了更有意义且精简的输入，最终提高估计精度。与一样，即使图像对不重叠，Transformer编码器也能够仅利用单幅图像中存在的直线来预测旋转，这与人类的认知能力相同。例如，在图像对不重叠这一极端情形下，可以通过隐含地假设建筑物及其边缘与地面高度垂直，从单幅图像估计滚转角。

同样，可以假设街道和人行道与地平面平行，或计算相应的消失点，从而估计相对仰角。该领域的大多数训练数据集和测试数据集都描绘城市场景，符合这些假设。

3.3 基于级联注意力的解码

ˆ给定编码成对输入图像之间相互关系的交叉注意力张量 $\mathbf{T}$ ，我们的目标是推断 3D 相对旋转四元数。为此，我们提出了一种创新的级联解码方案，交替细化查询旋转和交叉注意力。最初，Transformer解码器-1 根据学习得到的四元数 $\mathbf{q}$ 增强 $\mathbf{T}$，以计算 $\mathbf{T}$。随后，使用 $\mathbf{q}$ 查询细化后的交叉注意力 $\mathbf{T}$，通过 Transformer 解码器-2 推导旋转编码 $\bar{\mathbf{q}}$。这种级联推理方法可以通过增加额外的双单元进行扩展。然而，我们观察到，增加级联结构并未带来性能提升。由于解码器输入是语义表示，因此不使用位置编码。

3.4 相对旋转回归

编码后的四元数向量 $\bar{\mathbf{q}}$ 随后被输入多层感知机（MLP）回归器，计算以 $\tilde{\mathbf{q}}$ 表示的四元数输出。该四元数表示为 $\mathbf{q} = [q_w, q_x, q_y, q_z]$。训练损失定义为：

$$\mathbf{L} = \mid q_0 - \tilde{\mathbf{q}}/||\tilde{\mathbf{q}}|| \mid_2, \tag{4}$$

其中，q_0 和 $\tilde{q}$ 分别为真实四元数和预测四元数。归一化可确保四元数是有效的三维旋转表示。

4 实验结果

所提出的方案通过将其应用于当代基准数据集进行了实验验证，其中包含重叠和非重叠图像对。我们的实验设置严格遵循了 Cai等人建立的范式 [7], ，使用相同的数据集和图像重叠类别。利用他们提供的源代码[2]，从全景图像生成透视视图，确保两项研究中的输入图像相同，从而能够与先前的最先进结果及其他当代方案进行公平比较。为此，我们还使用了与 [58] 相同的[7]。第4.1节根据 Cai等人详细介绍了我们所使用的图像数据集及其处理方法 [7], ，以构建训练数据集和测试数据集。训练细节见第4.2节。我们使用先前工作中采用的测地误差度量，与第4.3节列出的最新最先进方案进行比较 [7]

$$\mathbf{E} = \arccos\left(\frac{tr(\mathbf{R}^T\mathbf{R}^*) - 1}{2}\right), \tag{5}$$

其中，$\mathbf{R}$ 是预测旋转矩阵，$\mathbf{R}^*$ 是每对图像对应的真实相对旋转矩阵。实验比较结果见第4.4 节；注意力图则在补充材料第 2 节中可视化，以直观地展示Transformer编码器计算出的交叉注意力分数。我们在第4.5节研究了所提方案的跨数据集泛化特性，并在第4.6节报告了针对不同参数和设计选择的消融研究。

4.1 图像数据集及其处理

我们使用了以下数据集以及先前工作中采用的训练集/测试集划分：

InteriorNet [29]是一个用于理解和绘制室内场景的合成数据集。我们使用了来自 112 栋不同房屋的 10,050 张全景图子集，其中 82 栋房屋的图像用于训练，30 栋房屋的图像用于测试。

StreetLearn [36] 是一个室外数据集，包含约 140,000 张匹兹堡和曼哈顿的全景视图。我们使用了 56K 张全景

[2] https://github.com/RuojinCai/ExtremeRotation_code

视图，这些视图来自曼哈顿；我们从中随机选取了1000 张全景视图用于测试。

SUN360 [54] 是一个室内数据集，包含高分辨率全景图像，覆盖了 $360° \times 180°$ 多种从互联网下载的环境场景。它还提供位置类别标签。我们分别使用 7K 和 2K 张全景图像进行训练和测试。

由于这些数据集包含全景图像，我们在每张全景图像中随机裁取 200 个不同位置，生成 200 张透视 128×128 图像。这种采样策略确保真实图像对的分布一致，其俯仰角分辨率覆盖 $[-45°, 45°]$，偏航角分辨率覆盖 $[-180°, 180°]$。我们仅估计偏航角和俯仰角，并假设成对图像之间的滚转角为零。为避免生成无纹理图像对（即主要包含房屋天花板或地板，或室外场景天空的图像），我们将室外数据集的俯仰角范围限制为$[-45°, 45°]$，将室内数据集的俯仰角范围限制为 $[-30°, 30°]$。训练数据集与测试数据集之间没有重叠。为将我们的结果与先前研究进行比较，并分析相机平移对我们旋转估计方法的影响，我们将 InteriorNet 和 StreetLearn 数据集分为两组：包含相机平移的图像和不包含相机平移的图像。未发生平移的图像对，是从同一张全景图像中随机选取的裁剪图像对。相比之下，包含平移的数据集（称为 StreetLearn-T 和 InteriorNet-T）是从不同全景图像中随机选取裁剪图像对生成的，其中平移距离小于 $3m$。不过，我们的方法并未用于估计这些平移。我们在重叠图像对和非重叠图像对上评估性能，并采用 Cai等人 [7]的设置，将数据集划分为三个重叠类别：

大规模，包含重叠程度很高的图像对，其相对旋转角度最大可达 $45°$

小重叠，包含部分重叠的图像对，其相对旋转角度为 $\in [45°, 90°]$

无重叠，包含相对旋转为 $> 90°$ 的图像对。

4.2 训练细节

我们使用预训练的 Residual-Unet [58] （与 Cai等人 [7] 中使用的相同）作为骨干网络，计算两幅输入图像（$\hat{I}_1$、$\hat{I}_2$）$\in \mathbb{R}^{c \times k_1 \times k_2} = \mathbb{R}^{128 \times 32 \times 32}$的特征图。根据图 2，随后将这些特征图交叉传播到专用解码器单元中，从而细化表征 $\bar{I}_1$ 和 $\bar{I}_2$。表征 $\bar{I}_1$ 和 $\bar{I}_2$ 的细化结果沿样本轴重塑并拼接，形成张量 T $\in \mathbb{R}^{(2 \cdot 32 \cdot 32) \times 128} = \mathbb{R}^{2048 \times 128}$。T 作为 Transformer编码器的输入，该编码器由 $l = 2$ 层组成，采用 ReLU非线性函数和 $p = 0.1$的丢弃法。每个编码器层使用 $h = 4$ 个多头注意力头，隐藏维度为 $C_h = 768$。第 4.6 节给出了 Transformer编码器参数的消融研究。随后，Transformer编码器的输出 T 被输入一个由两个串联解码器组成的双路径结构。主解码器接收一个由白色高斯噪声初始化的学习型四元数向量 $\bar{q}$ 作为输入，并以交叉注意力 T 作为$\bar{}$查询，生成 T，增强上下文细微差异$\bar{}$；后续解码器则将 T 作为输入，并以相同的空四元数向量 $\bar{q}$ 作为查询，生成 $\bar{\bar{q}}$，封装关键的旋转属性。这两个序列注意力解码器使用 $l = 2$ 层、 $h = 2$ 个多头注意力头，隐藏维度为 $C_h = 768$。最后，MLP回归器计算用于公式 4 中回归损失的四元数表征。MLP回归器包含两个全连接层。在所有实验中，我们使用 Adam优化器对模型进行优化，初始学习率为 $\lambda = 5e-4$，并设置 $\beta_1 = 0.9$、 $\beta_2 = 0.999$、$\epsilon = 10^{-10}$，批大小为 20。我们的模型使用 PyTorch 实现，可进行端到端训练，所有实验均在配备 8GB NVIDIA GeForce GTX 2080 GPU 的设备上完成。

4.3 对比基线

按照Cai等人的做法 [7], ，我们使用第4.1节中的数据集与当代方案进行比较：

一种基于 SIFT 的方法 [6]。一种用于

重叠方法		InteriorNet			InteriorNet-T			SUN360			StreetLearn			StreetLearn-T		
		平均值(°↓)	中位数(°↓)	10°(%↑)	平均值(°↓)	中位数(°↓)	10°(%↑)	平均值(°↓)	中位数(°↓)	10°(%↑)	平均值(°↓)	中位数(°↓)	10°(%↑)	平均值(°↓)	中位数(°↓)	10°(%↑)
大规模	尺度不变特征变换* [34]	6.09	4.00	84.86	7.78	2.95	55.52	5.46	3.88	93.10	5.84	3.16	91.18	18.86	3.13	22.37
	SuperPoint* [13]	5.40	3.53	87.10	5.46	2.79	65.97	4.69	3.18	92.12	6.23	3.61	91.18	6.38	1.79	16.45
	Reg6D [59]	9.05	5.90	68.49	17.00	11.95	41.79	16.51	12.43	40.39	11.70	8.87	58.24	36.71	24.79	23.03
	DenseCorrVo [7]	1.53	1.10	99.26	2.89	1.10	97.61	1.00	0.94	**100.00**	1.19	1.02	99.41	9.12	2.91	87.50
	8PointViT [43]	0.48	0.40	**100.00**	2.90	1.83	97.91	-	-	-	0.62	0.52	**100.00**	4.08	2.43	**90.13**
	Ours	**0.43**	**0.38**	99.65	**1.75**	**0.95**	**98.8**	**0.85**	**0.45**	99.95	**0.58**	**0.48**	99.31	**3.88**	**1.69**	87.20
小重叠	尺度不变特征变换* [34]	24.18	8.57	39.73	18.16	10.01	18.52	13.71	6.33	56.77	16.22	7.35	55.81	38.78	13.81	5.68
	SuperPoint* [13]	16.72	8.43	21.58	11.61	5.82	11.73	17.63	7.70	26.69	19.29	7.60	24.58	6.80	6.85	0.95
	Reg6D [59]	25.71	15.56	33.56	42.93	28.92	23.15	42.55	32.11	9.40	24.77	15.11	30.56	46.61	34.33	13.88
	DenseCorrVol [7]	6.45	1.61	95.89	10.24	1.38	89.81	3.09	1.41	98.50	2.32	1.41	98.67	13.04	3.49	84.23
	8PointViT [43]	1.84	0.94	99.32	4.48	2.38	96.30	-	-	-	1.46	1.09	**100.00**	9.19	3.25	87.7
	Ours	**1.55**	**0.872**	**99.85**	**4.25**	**0.777**	**97.55**	**2.109**	**0.831**	**98.99**	**1.21**	**0.718**	99.122	**7.48**	**1.8666**	**88.996**
None	尺度不变特征变换* [34]	109.30	92.86	0.00	93.79	113.86	0.00	127.61	129.07	0.00	83.49	90.00	0.38	85.90	106.84	0.38
	SuperPoint* [13]	120.28	120.28	0.00	–	–	0.00	149.80	165.24	0.00	–	–	0.00	–	–	0.00
	Reg6D [59]	48.36	32.93	10.82	60.91	51.26	11.14	64.74	56.55	3.77	28.48	18.86	24.39	49.23	35.66	11.86
	8PointViT [43]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DenseCorrVol [7]	37.69	3.15	61.97	49.44	4.17	58.36	34.92	4.43	61.39	5.77	1.53	**96.41**	30.98	3.50	**72.69**
	Ours	**35.13**	**2.814**	**65.20**	**45.32**	**4.05**	**59.56**	**32.46**	**4.19**	**63.17**	**5.33**	**1.20**	96.22	**28.13**	**3.25**	72.43

表1：相对旋转估计结果。我们使用了InteriorNet、SUN360和StreetLearn数据集，并展示测地误差的平均值和中位数。我们还给出了第 10°节中各重叠类别下，相对旋转误差低于4.1的图像对所占百分比。灰色数字表示误差超过50%。星号 $*$ 表示平均误差和中位误差未能实现位姿估计，且仅针对成功的图像对进行计算。

匹配 SIFT特征 [34] ，使用 RANSAC [17] 在同一全景图的图像对中进行匹配，并利用单应性方程或本质矩阵估计相对旋转矩阵。

基于CNN的方法 [13]。采用SuperPointNet [13] 和 D2-Net [16]检测并编码局部图像特征的深度学习方案。

自监督兴趣点 [59]。Zhou等人提出的方案 [59]（Reg6D）使用卷积神经网络近似各种旋转表示之间的映射，并拟合连续的5D和6D旋转表示，而非常用的欧拉角和四元数表示。

极端旋转估计 [7]。这是一种深度学习技术，用于在极端情况下估计图像对之间的相对三维旋转 [7] ，此时图像之间几乎没有重叠或完全没有重叠。他们提出了一种网络，通过计算四维相关体来自动学习隐式视觉线索。

基于注意力的方法。我们还与 Rockwell等人 [43] (8PointVit)近期的工作进行了比较，该方法使用视觉 Transformer（ViT）来估计相对位姿。尽管 Rockwell等人在多种设置下取得了有竞争力的结果，但他们的方法不太适用于极端视角变化。

4.4 实验比较 s

表1报告了所提出方案与基线方案及最先进方案的比较结果。我们报告了公式5 中给出的测地线误差的均值和中位数，以及估计的相对旋转误差小于 10°的图像对所占的百分比。我们将所提出模型的准确率与第4.3节中详细介绍的方案进行了比较。结果表明，所提出的方法在室内和室外场景中均具有较高的准确率，并且在所有重叠程度类别中都显著优于基线方案。对于非重叠图像对，SIFT [6], SuperPointNet [13],Reg6D [59] 和 8PointViT [43] 等基于对应关系的方法无法提供任何估计结果，因为它们需要特征对应关系。DenseCorrVol 方法 [7] 在极端情况下能够提供准确结果，但我们的方法优于该方法。定性实验结果见补充材料第 1 节。旋转估计的定性结果如图4 所示，图中展示了 StreetLearn 和 SUN360



图4：旋转估计结果。全景真实图像和裁剪的真实图像分别用绿色线和黄色点线标出。其中一张裁剪图像的预测足迹用红色点线标出。第一行展示了大重叠区域图像的匹配结果。第二行和最后一行展示了小重叠区域图像和不重叠图像的匹配结果。

数据集在大规模、小重叠和非重叠情况下的结果。我们展示了完整全景图、作为所提出方案输入的裁剪图像的轮廓，以及根据估计旋转角度得到的估计图像裁剪区域轮廓。在所有情况下，我们都实现了较高的估计精度。

4.5 跨数据集泛化

重叠程度	本文方法 [°]	Cai等人 [7] [°]
大规模	**9.55**	11.23
小重叠	**16.33**	20.87
None	**38.48**	40.82

表2：跨数据集泛化。我们在曼哈顿数据集上训练模型，并在伦敦数据集上测试。报告的是平均测地线误差

我们使用Holicity数据集评估了所提方法的跨数据集泛化性能[60]。使用曼哈顿数据集训练模型，并使用伦敦数据集进行测试。根据第4.1节，将测试图像划分为三个重叠类别。我们将所提方法的泛化能力与Cai等人的方法进行了比较 [7]。

表2中的结果表明，我们的方法在所有重叠类别中均优于Cai等人的方法。

4.6 消融研究

Transformer编码器参数。表3总结了针对各个重叠类别的多种Transformer编码器配置。Transformer编码器的表达能力取决于注意力头和层的数量。使用的数量越多，表达能力越强。然而，数量过多可能导致过拟合；就表3中的准确率而言，最优配置为 $h = 4$， $l = 2$。具体而言，这种配置是一个最佳平衡点，增加注意力头或层的数量都会降低准确率。

骨干网络消融。在表4中，我们通过改变 Residual-Unet 中残差块的数量来考察其 [58] 骨干网络的深度。增加残差块的数量可以增强骨干网络的表达能力，但架构过深可能会导致过拟合。我们发现，使用三个残差块是最优选择，这与我们最初的决定一致。

三维旋转编码和训练损失。

重叠程度	注意力头	层	旋转误差 平均值 [°]	中位数 [°]	10° [%]
大规模	1	1	1.21	0.97	99.1
	4	1	0.82	0.65	99.1
	2	2	0.87	0.71	99.12
	4	2	**0.58**	**0.48**	**99.31**
	4	4	0.65	0.59	99.2
小重叠	1	1	7.13	3.26	94.99
	4	1	6.42	2.46	95.32
	2	2	5.44	2.05	96.55
	4	2	**1.21**	**0.718**	**99.122**
	4	4	4.61	0.98	95.43
None	1	1	7.43	2.88	91.55
	4	1	6.15	2.55	92.55
	2	2	6.23	3.02	91.35
	4	2	**5.33**	**1.20**	**96.22**
	4	4	5.78	1.87	95.22

表3：使用 StreeLearn 数据集对 Transformer编码器参数进行消融研究。对于每个重叠类别，都有一种最优配置，能够平衡 Transformer编码器的表达能力与过拟合。

对不同旋转编码及其相应训练损失进行了消融实验评估，结果如表5所示。评估通过应用 Residual-Unet 骨干网络 [58]以及所提出的基于 Transformer编码器的交叉注意力方法，在来自具有大重叠区域的 StreeLearn 数据集的图像对上进行。对于离散形式化方法，按照 Cai等人 [7], 的方法，将俯仰角和偏航角离散为 360 个分箱 $\in [-180°, 180°]$，并使用交叉熵损失训练网络。这些结果与我们方案中公式 L_2 4所述使用回归损失获得的结果进行了比较。表5 中的结果表明，我们的L_2 回归方法优于 Cai等人提出的离散欧拉角方法 [7]。

架构消融 为了评估所提出的架构并考察各个组件的贡献，特别是不同 Transformer解码器的使用效果，我们针对第 3 节和图 2中介绍的所提出架构，采用多种架构变体进行了一系列实验。结果见表 6，相应的架构则见补充材料。在每项实

骨干网络	层	旋转误差 平均值 [°]	中位数 [°]	10° [%]
1	卷积(k=7, s=2, d=64) 1× 残差块 卷积(k=3, s=1, d=512) 卷积(k=3, s=1, d=256)	2.95	1.53	94.13
2	卷积(k=7, s=2, d=64) 2× 残差块 卷积(k=3, s=1, d=512) 卷积(k=3, s=1, d=256)	1.75	1.05	96.13
3	卷积(k=7, s=2, d=64) **3 × 残差块** 卷积(k=3, s=1, d=512) 卷积(k=3, s=1, d=256)	**1.21**	**0.7818**	**99.122**
4	卷积(k=7, s=2, d=64) 4× 残差块 卷积(k=3, s=1, d=512) 卷积(k=3, s=1, d=256)	1.45	0.86	97.12

表4：骨干网络消融。我们通过改变残差块的数量，评估方案中所用 Residual-Unet 骨干网络的深度。[58],

表示损失函数		旋转误差 平均值 [°]	中位数 [°]	10° [%]
四元数	L_2 回归	**1.21**	**0.78**	**99.122**
欧拉角	交叉熵	2.37	1.46	98.13
欧拉角	L_2 回归	2.22	1.32	98.99

表 5：三维旋转编码和训练损失的消融实验。我们比较了离散域和连续域中的欧拉角与四元数三维旋转编码，以及相应的训练损失。

验中，我们采用所提出的 Transformer解码器的一种特定部分配置，并使用输入图像之间具有大重叠区域的 StreetLearn数据集评估所得的估计误差。表 6 中的结果表明，所提出的配置优于所有其他配置。具体而言，配置 #1 表明，依次使用基于注意力的解码器 TD1 和 TD2 可显著提高准确率。TD0 进行的交叉解码也带来了额外提升，但提升幅度并不显著。

	TD0	TD1	TD2	平均值 [°]	中位数 [°]	10°[%]
0	+	+	+	**0.58**	**0.48**	**99.3**
1	-	+	+	0.98	0.81	95.15
2	-	-	-	3.35	2.44	88.16
3	-	+	-	1.76	1.55	93.12
4	+	+	-	0.86	0.72	95.64
5	+	-	-	1.97	1.65	92.82

表 6：架构消融研究。我们比较不同配置的估计精度 Transformer解码器（TD）。相应的架构见补充材-料中，第一种配置见第 3和图 2。

5 结论

我们提出了一种新颖的方法，用于估计一对图像之间的相对旋转。具体而言，我们研究了小重叠和无重叠图像之间的旋转估计。我们提出了一种基于注意力的方法，使用Transformer编码器计算图像对嵌入图之间的交叉注意力，其性能优于以往使用四维相关体 [51, 7] 以及解码器-解码器机制来估计输出四元数的方法。我们的框架可以端到端训练，并优化回归损失。实验结果表明，该框架在多个当代研究所用的数据集上优于以往最先进的方案 [7] ，尤其是在具有挑战性的小重叠和无重叠情形下。

参考文献

[1] Hadar Averbuch-Elor 和 Daniel Cohen-Or。《Rin- git：对时间事件的随拍照片进行环排序》。*ACM Trans. Graph.*, 34(3), 2015。1[2]Alexandru O. Balan、Michael J. Black、Horst Haussecker 和 Leonid Sigal。《照亮人体姿态：关于阴影、明暗处理以及姿态和形状估计》。载于 *IEEE International Conference on Computer Vision (ICCV)*，第 1–8 页，2007。1[3] Vassileios Balntas、Shuda Li 和 Victor Prisacariu。《Relocnet：使用神经网络进行相机重定位的连续度量学习》。载于 *European Conference on Com-puter Vision (ECCV)*，2018年9月。3[4] Herbert Bay、Tinne Tuytelaars 和 Luc Van Gool。《Surf：加速稳健特征》。载于 *European Conference on Computer Vision (ECCV)*，第 404–417 页。施普林格，2006。3[5]Eric Brachmann、Alexander Krull、Sebastian Nowozin、Jamie Shotton、Frank Michel、Stefan Gumhold 和 Carsten Rother。《DSAC——用于相机定位的可微 RANSAC》。载于 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*，第 6684–6692 页，2017。1[6]Matthew Brown、Richard I Hartley 和 David Nist´er。《全景拼接的最简解》。载于 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*，第 1–8 页。IEEE，2007。7, 8[7] Ruojin Cai、Bharath Hariharan、Noah Snavely 和 Hadar Averbuch-Elor。《利用稠密相关体进行极端旋转估计》。载于 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*，2021。3, 5, 6, 7, 8, 9, 10, 11[8] Nicolas Carion、Francisco Massa、Gabriel Synnaeve、Nicolas Usunier、Alexander Kirillov 和 Sergey Zagoruyko。《使用变换器进行端到端目标检测》。载于 Andrea Vedaldi、Horst Bischof、Thomas Brox 和 Jan-Michael Frahm 编，*EuropeanConference on Computer Vision (ECCV)*，第 213–229 页，Cham，2020。施普林格国际出版公司。1, 5[9] Yaron Caspi 和 Michal Irani。《对齐不重叠的序列》。*International Journal ofComputer Vision*, 48(1):39–51, 2002。3[10] J.M. Coughlan 和 A.L. Yuille。《曼哈顿世界：通过贝叶斯推断从单幅图像确定罗盘方向》。载于 *IEEE International Conference on Com-*

puter Vision (ICCV) 2, 941 947 2, , 卷、1999 1
页码 – 卷。[11] Antonio Criminisi、Ian Reid 和 Andrew Zisserman。《单视图测量》。载于 *IEEE International Conference on Computer Vision (ICCV)* 434 441 , 第 – 页。
IEEE计算机学会，1999。1[12] Andrew J Davison、Ian D Reid、Nicholas D Molton 和 Olivier Stasse。《MonoSLAM：实时单目相机 SLAM》。*IEEE Transactions on Pattern Analysis and Machine Intelligence* 29 6 1052 1067, 2007 1 , ( ): – 。
[13] Daniel DeTone、Tomasz Malisiewicz 和 Andrew Rabinovich。《SuperPoint：自监督兴趣点检测与描述》。载于 *IEEE Conf.* , 。[14] *Comput. Vis. Pattern Recog. Worksh.* 2018 8
Mingyu Ding、Zhe Wang、Jiankai Sun、Jianping Shi 和 Ping Luo。《CamNet：用于相机重定位的由粗到细检索》。载于 *IEEE International Conference on Computer Vision (ICCV)*，2019年10月。3[15] Alexey Dosovitskiy、Philipp Fischer、Eddy Ilg、Philip Hausser、Caner Hazirbas、Vladimir Golkov、Patrick Van Der Smagt、Daniel Cremers 和 Thomas Brox。《FlowNet：利用卷积网络学习光流》。载于 *IEEE International Conference on Computer Vision (ICCV)* 2758 2766, 2015 , 第 – 页。
2, 3, 4[16] Mihai Dusmanu、Ignacio Rocco、Tomas Pajdla、Marc Pollefeys、Josef Sivic、Akihiko Torii 和 Torsten Sattler。《D2-Net：用于局部特征联合检测与描述的可训练卷积神经网络》。载于 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2019 8 , 。[17]
Martin A Fischler 和 Robert C Bolles。《随机样本一致性：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》。*Communications of the ACM*, 24(6):381–395, 1981。8[18] Xiaodong Gu、Zhiwen Fan、Siyu Zhu、Zuozhuo Dai、Feitong Tan 和 Ping Tan。《用于高分辨率多视图立体视觉和立体匹配的级联代价体》。载于 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2495 2504, , 第 – 页。2020 3
[19] R. I. Hartley 和 A. Zisserman。《*Multiple View Geometry in Computer Vision* 》。剑桥大学出版社，ISBN：0521540518，第二版，2004。1[20] Shengyu Huang、Zan Gojcic、Mikhail Usvyatsov、Konrad Schindler 和 Andreas Wieser。《Predator：低重叠程度三维点云配准》。*IEEE/CVF Conference on Computer Vision and* 在 *Pattern Recognition (CVPR)* 2021 3 , 。
[21] Di Jia、Kai Wang、ShunLi Luo、TianYu Liu 和 Ying Liu。Braft：基于相关性块的光流循环全对场变换。*IEEE Signal Processing Letters* 28 1575 1579, 2021 2, 3, , : –。
4[22]Y. Keller、A. Averbuch 和 Y. Shkolnisky。利用欧拉定理和三维伪极坐标快速傅里叶变换实现代数精确的体积配准。载于*IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*，第2卷，第795–800页，第2卷，2005年。3[23] Y. Keller、Y. Shkolnisky 和 A. Averbuch。利用三维伪极坐标傅里叶变换进行体积配准。*IEEE Transactions on Signal Processing*，54 11 4323 4331，2006年 3 ( ): –。[24]Alexander Kolesnikov、Alexey Dosovitskiy、Dirk Weissenborn、Georg Heigold、Jakob Uszkoreit、Lucas Beyer、Matthias Minderer、Mostafa Dehghani、Neil Houlsby、Sylvain Gelly、Thomas Unterthiner 和 Xiaohua Zhai。一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer。载于 *International Conference on Learning Representations*，2021年1。[25] S. Kong 和 C. Fowlkes。用于细粒度分类的低秩双线性池化。载于 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*，第7025–7034页，美国加利福尼亚州洛斯阿拉米托斯，2017年7月。IEEE计算机学会。3[26] Jinwoo Lee、Minhyuk Sung、Hyunjoon Lee 和 Junho Kim。用于单幅图像相机标定的神经几何解析器。载于 Andrea Vedaldi、Horst Bischof、Thomas Brox 和 Jan-Michael Frahm（编），*European Conference on Computer Vision (ECCV)* 541 557, 2020 , 第–页，Cham，。施普林格国际出版公司。1[27] Jake Levinson、Carlos Esteves、Kefan Chen、Noah Snavely、Angjoo Kanazawa、Afshin Rostamizadeh 和 Ameesh Makadia。深度旋转估计中的奇异值分解分析。*Advances in Neural Information Processing Systems (NIPS)* 33 2020 3 , , 。
[28] Minhua Li、Qingling Chang、Yuhan Wang、Xinglin Liu、Shiting Xu 和 Yan Cui。采用多尺度混合代价体的立体匹配。 *IEEE Access*，10 100128 100136，2022年3：–。[29] Wenbin Li 、 Sajad Saeedi 、 John McCormac、Ronald Clark、Dimos Tzoumanikas、Qing Ye、Yuzhong Huang、Rui Tang 和 Stefan Leutenegger。Interior-net：超大规模多传感器照片级真实室内

场景数据集。载于 *British Machine Vision Conference* 2018 2, 6 ，。[30]

Zhengfa Liang、Yulan Guo、Yiliu Feng、Wei Chen、Linbo Qiao、Li Zhou、Jianfeng Zhang 和 Hengzhu Liu。利用多层代价体和多尺度特征恒定性的立体匹配。 *IEEE Transactions on Pattern Analysis and Machine Intelligence*，2019年。2、3、4[31] Tsung-Yu Lin、Aruni RoyChowdhury 和 Subhransu Maji。用于细粒度视觉识别的双线性卷积神经网络模型。载于 *IEEE International Conference on Computer Vision (ICCV)* 2015 3 ，。

[32] Chenxi Liu、Alex Schwing、Kaustav Kundu、Raquel Urtasun 和 Sanja Fidler。Rent3D：用于单目布局估计的平面图先验。载于 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2015 1 ，。[33]

Hai Liu、Shuai Fang、Zhaoli Zhang、Duantengchuan Li、Ke Lin 和 Jiazhang Wang。MFDNet：用于头部姿态估计的协作式姿态感知与矩阵Fisher分布。 *IEEE Transactions on Multimedia* 24 2449 2460, 2022 3 ，：–。[34]

David G Lowe。来自尺度不变关键点的显著图像特征。 *International Journal of Computer Vision* 60 2 91 110, 2004 3, 8 ，():–。

[35] Ben Mildenhall、Pratul P. Srinivasan、Matthew Tan-cik、Jonathan T. Barron、Ravi Ramamoorthi 和 Ren Ng。NeRF：将场景表示为用于视图合成的神经辐射场。载于 *European Conference on Computer Vision (ECCV)* 2020 1 ，。[36]

Piotr Mirowski、Andras Banki-Horvath、Keith An-derson、Denis Teplyashin、Karl Moritz Hermann、Mateusz Malinowski、Matthew Koichi Grimes、Karen Simonyan、Koray Kavukcuoglu、Andrew Zis-serman 等。StreetLearn环境与数据集。 *arXiv preprint arXiv:1903.01292*，2019年。2、6[37] David Mohlin、Josephine Sullivan 和 G´erald Bianchi。基于矩阵Fisher分布的概率方向估计。 *Advances in Neural Information Processing Systems (NIPS)*，33，2020年。3[38] Raul Mur-Artal、Jose Maria Martinez Montiel 和 Juan D Tardos。ORB-SLAM：一种通用且准确的单目SLAM系统。 *IEEE Transactions on Robotics* 31 5 1147 1163, 2015 1 ，():–。

[39] Guang-Yu Nie、Ming-Ming Cheng、Yun Liu、Zhengfa Liang、Deng-Ping Fan、Yue Liu 和 Yongtian Wang。用于立体匹配的多层上下文超聚合。载于 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*，第3283–3291页，2019年。2、3、4[40] Onur Ozyesil、Vladislav Voroninski、Ronen Basri 和 Amit Singer。结构运动综述。 *arXiv preprint arXiv:1701.08493*，2017年。1[41] Yiming Qian 和 James H. Elder。联合估计焦距与相机旋转的可靠在线方法。载于 *European Conference on Computer Vision (ECCV)*，第249–265页，柏林、海德堡，2022年。施普林格出版社。1[42] Gernot Riegler 和 Vladlen Koltun。自由视点合成。载于 *European Conference on Computer Vision (ECCV)* 2020 1 ，。

[43] Chris Rockwell、Justin Johnson 和 David F. Fouhey。八点算法作为视觉变换器相对位姿预测的归纳偏置。载于 *International Conference on 3D Vision (3DV)* 2022 3, 8 ，。[44]

Johannes L Schonberger 和 Jan-Michael Frahm。重新审视结构运动。载于 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2016 1 ，。

[45] Johannes L Sch¨onberger、Marc Pollefeys、Andreas Geiger 和 Torsten Sattler。语义视觉定位。载于 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*，第6896–6906页，2018年。1[46] O. Shakil。针对具有自由相机运动的不重叠序列的高效视频对齐方法。载于 *ICASSP*，第2卷，第II–II页，2006年。3[47] Yoli Shavit 和 Yosi Keller。用于改进位姿回归的相机位姿自动编码器。载于 *European Conference on Computer Vision (ECCV)*，第140–157页，Cham，2022年。瑞士施普林格自然出版社。1[48] Che Sun、Yunde Jia、Yi Guo 和 Yuwei Wu。低重叠RGB-D扫描的全局感知配准。载于 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*，第6357–6366页，2022年6月。3[49] Dekel (Basha) T、Moses Y 和 Avidan S。照片排序。 *International Journal of Computer Vision* 110 3 275 289, 2014 8 1 ，():–。

引用数：。[50] Hajime Taira、Masatoshi Okutomi、Torsten Sattler、Mircea Cimpoi、Marc Pollefeys、Josef Sivic、Tomas Pajdla 和 Akihiko Torii。InLoc：通过稠密匹配与视点合成实现室内视觉定位。载于 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 7199 7209, ，第–页。2018 3

[51] Zachary Teed 和 Jia Deng。RAFT：用于光流的循环全对场变换。载于 *European Conference on Computer Vision (ECCV)*，第402–419页，柏林、海德堡，2020年。施普林格出版社。[52] M. Turkoglu、E. Brachmann、K. Schindler、G. J. Brostow 和 A. Monszpart。利用图神经网络和相对位姿监督进行视觉相机重定位。载于 *International Conferenceon 3D Vision (3DV)*，第145–155页，美国加利福尼亚州洛斯阿拉米托斯，2021年12月。[53] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、ukasz Kaiser 和 Illia Polosukhin。注意力就是一切。载于 I. Guyon、U. Von Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett 编，*Advances in Neural Information Processing Systems (NIPS)*，第30卷。Curran Associates, Inc.，2017年。[54] Jianxiong Xiao、Krista A Ehinger、Aude Oliva 和 Antonio Torralba。使用全景地点表示识别场景视点。载于 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*，第–页。IEEE，。[55] Fei Xue、Xin Wu、Shaojun Cai 和 Junqiu Wang。利用图神经网络学习多视图相机重定位。载于 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*，第–页。[56] Gengshan Yang 和 Deva Ramanan。用于光流的体积对应网络。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alch´e-Buc、E. Fox 和 R. Garnett 编，*Advances in Neural Information Processing Systems (NIPS)*，第32卷。Curran Associates, Inc.，。[57] Zhengxin Zhang、Qingjie Liu 和 Yunhong Wang。通过深度残差U-Net进行道路提取。*IEEE Geoscience and Remote Sensing Letters* 15 749 753, ，：–。2018 [58] Zhengxin Zhang、Qingjie Liu 和 Yunhong Wang。通过深度残差U-Net进行道路提取。*IEEE Geoscience and Remote Sensing Letters* 15 5 749 753, ，（）：–。2018 [59] Yi Zhou、Connelly Barnes、Jingwan Lu、Jimei Yang 和 Hao Li。神经网络中旋转表示的连续性。载于 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*，第–页。[60] Yichao Zhou、Jingwei Huang、Xili Dai、Linjie Luo、Zhili Chen 和 Yi Ma。HoliCity：城市尺度数据用于学习整体三维结构的平台。2020年。arXiv:2008。03286 [计算机视觉与模式识别]。[61] Rui Zhu、Xingyi Yang、Yannick Hold-Geoffroy、Federico Perazzi、Jonathan Eisenmann、Kalyan Sunkavalli 和 Manmohan Chandraker。在野外进行单视图测量。载于 Andrea Vedaldi、Horst Bischof、Thomas Brox 和 Jan-Michael Frahm（编），*European Conference on Computer Vision (ECCV)* 316 333, 2020，第 – 页，Cham，。施普林格国际出版。