arXiv:2303.02610v2 [cs.CV] 4 Mar 2025

HyperPose：超网络增强的相机位姿定位与扩展剑桥地标数据集

Ron Ferens Yosi Keller 以色列拉马特甘巴伊兰大学
工程学院{ronferens, yosi.keller}@gmail.com

摘要

在这项工作中，我们提出了 *HyperPose*，它在绝对相机位姿回归器中采用超网络。自然场景固有的外观变化源于环境条件、透视和光照，会导致训练集与测试集之间出现显著的域差异。这种差异降低了当代定位网络的精度。为缓解这一问题，我们主张将超网络融入单场景和多场景相机姿态回归模型。在推理过程中，超网络会根据具体的输入图像动态计算定位回归头的自适应权重，从而有效缩小域差距。我们使用室内和室外数据集，在多个成熟的绝对姿态回归架构上评估了 *HyperPose* 方法。我们还推出并分享了扩展剑桥地标（*ECL*）这一新型定位数据集。该数据集基于剑桥地标数据集，展示了多个季节中外观条件差异显著的场景。我们的实证实验表明，*HyperPose* 显著提升了单场景和多场景架构的性能。我们已公开提供源代码、预训练模型 [1]，以及 *ECL* 数据集 [2]。

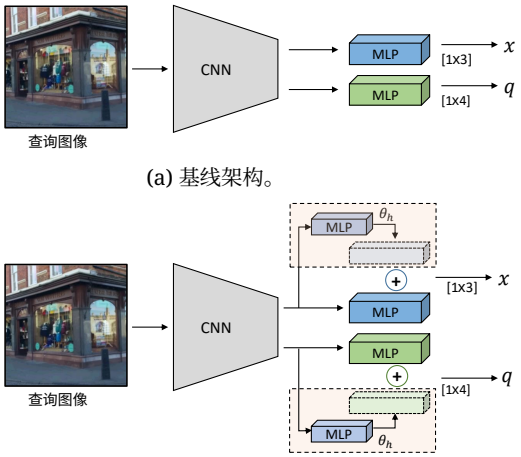


图 1. 图1a展示了基线绝对位姿回归架构。它由一个卷积神经网络主干网络组成，用于生成编码输入查询图像的特征向量，随后由回归头估计平移和方向。图 1b 展示了这一基线架构的扩展版本，其中超网络在推理过程中输出回归头（平移和方向）的权重。

1. 引言

在许多计算机视觉应用中，例如室内导航、增强现实和自动驾驶，利用相机拍摄的图像确定该相机的位置和方向至关重要。当前的相机定位技术在精度、运行时间和内存使用量之间存在各种权衡。特征匹配（FM）和场景坐标回归（SCR）被认为是精度最高的定位方法。基于特征匹配的方法 [15, 27, 37–41, 51, 54] 采用分层定位流程，通常使用由粗到细的方处理流程首先通过图像检索，根据与查询图像相似的图像执行初始定位。随后，通过局部特征提取与匹配对初始估计进行细化。为了准确估计相机的六自由度（6DoF），会将所得的二维—二维匹配映射为场景点云中的三维对应关系，然后通过透视-n-点（PnP）和RANSAC确定相机位姿。基于SCR的方法 [4–8, 47]直接回归场景中二维像素与3D坐标之间的对应关系。相对姿态回归（RPR）方案 [2, 14, 17, 20, 26, 57, 62] 直接估计查询图像与一组具有明确真实位姿的参考图像之间的相对姿态，从而无需传统流程。然而，图像检索方案对于检索一组最近邻图像必不可少。另一种方法是绝对姿态

[1] https://ronferens.github.io/hyperpose/ [2] https://ronferens.github.io/extcambridgelandmarks/

回归器（绝对位姿回归，APR）仅使用查询图像，通过一次前向传播即可估计相机位姿，因此延迟显著降低，但准确性也明显较低。此外，由于内存占用较小，APR可以作为独立应用部署在计算资源有限的边缘设备上。关于视觉相机位姿定位的综述见 [43, 65]。

Kendall 等人提出的 PoseNet [25]是首个绝对位姿回归方法，它采用卷积主干网络和多层感知机（MLP）头来回归相机位置和方向。这种简单且计算高效的方法可在 5ms 内执行，由此催生了众多绝对位姿回归方法，它们通过改进主干网络和 MLP 架构提升了精度 [11, 32, 33, 44, 49, 60, 61,63]。此外，研究人员还探索了其他损失函数和优化策略 [23, 24, 43]。室内和室外环境在光照、透视和物体运动方面都存在动态变化。尽管计算效率较高，绝对位姿回归模型往往难以在这些多样化条件下保持精度，而且对新场景的泛化能力有限，因此需要针对特定场景进行训练。因此，统一网络虽然架构简单，但在适应异构的真实世界视觉输入时，性能可能不尽理想。近期研究提出将绝对位姿回归从单场景扩展到多场景。Blanton 等人 [3] 采用卷积神经网络（CNN）主干网络，为输入图像计算全局描述符。他们没有采用针对特定场景的多层感知机（MLP），而是优化了一组全连接（FC）层，并根据预测的场景标识符为每个场景分配一层。尽管该方法提出了一个可在多个场景上优化单个模型的框架，但其绝对位姿回归精度未达到最先进水平。 [46] 提出了另一种采用双分支 Transformer 的方法 [58]。编码器专注于提取与位姿相关的特征，而解码器则将编码后的场景标识符转换为潜在位姿表示。解码器的输出用于场景分类，并选择对应的位置和朝向嵌入向量，再据此回归位置和朝向向量。

超网络 [22] 是一种深度学习架构，其中辅助网络（即超网络）会根据当前输入和下游任务为主网络生成权重。主网络和超网络通过反向传播进行端到端训练。超网络能够动态、自适应地调制主网络的权重，从而提升其应对条件变化时的灵活性[16, 18, 34, 55]。

在这项工作中，我们提出*HyperPose*，一种用于绝对相机位姿回归的超网络方法。该方法扩展了常见的 APR 方法 [11, 23, 25, 44, 45, 60]，超网络会为回归头计算权重。这使其能够自适应地分析图像内容，让回归权重聚焦于信息丰富的特征，而不是依赖一组固定的预训练权重。我们通过将超网络集成到单场景和多场景 APR 架构中，验证了所提出的方法。对于单场景网络，我们将 HyperPose 与现有的最先进 APR 模型 [61] 结合，同时也提出了一种受 [25] 所述原则启发、采用现成主干网络[53]和浅层回归头的简易基线 APR。对于多场景网络，我们通过集成超网络架构，扩展了一种领先的多场景 APR 模型 [46]。在这两种情况下，经过超网络增强的模型在本研究所用的基准数据集上评估时，均优于相应的原始模型。为评估定位鲁棒性，我们提出了扩展剑桥地标（ECL）数据集——在原始基准数据集[25] 的基础上，加入算法生成的环境变化。ECL 融合傍晚、冬季和夏季场景，涵盖了光照变化和季节转换等动态因素。

因此，我们提出以下贡献：

• 我们提出了一种通过集成超网络架构来增强现有绝对相机姿态回归模型的方法。

•我们在单场景和多场景位姿回归模型上全面评估了所提出的 HyperPose，结果表明，超网络增强模型的准确度优于原始版本。

• 所提出的方案在涵盖多种室内外定位任务的多场景 APR 基准测试上实现了新的最先进准确度。

• 我们推出了扩展剑桥地标（ECL）数据集，该数据集以原始剑桥地标数据集为基础 [25]。ECL 数据集通过纳入各种外观条件，扩展了现有测试场景。

2. 相关工作

相机位姿估计技术在推理阶段的输入和算法属性方面各不相同。

基于 3D 或基于结构的定位方法被认为是相机位姿估计中最准确的方法，并在领先的基准测试中展现出最先进的性能。这些方法利用图像中的 2D 特征与 3D 世界坐标之间的对应关系。[15, 27, 35, 37–41, 51, 54]提出了一种适用于分层姿态估计流水线的两阶段方法。因此，首先使用针对图像检索（IR）训练的卷积神经网络，对每张查询图像进行编码；该网络基于预先映射的图像数据集。随后，通过将局部图像

特征与 3D匹配相匹配，估计初步对应关系。接着，PnP-RANSAC 利用这些匹配来估计相机位姿。DSAC [8]及其后续方案 DSAC++ [4]采用卷积神经网络架构，直接从图像中的 2D 位置估计 3D坐标，并且在推理时只需要查询图像。与绝对姿态回归（APR）方法一样，每个场景都必须单独训练模型，从而达到最先进的精度。最近， [15] 提出了一种通过引入场景无关架构来缩小这一差距的解决方案。

相对姿态回归 (RPR) 算法根据输入查询图像与一组已知位置的参考图像，估计二者之间的相对姿态。计算相对姿态需要估计查询图像与参考图像之间的相对位置和方向 [1, 6, 68]。该方法利用图像检索（IR）确定一组最近的邻近图像，然后估计查询图像与每张检索到的图像之间的相对运动。结合估计出的相对姿态和参考图像已知的位置，便可估计相机的绝对姿态。与 3D 和基于结构的定位方法类似，RPR 需要经过多个步骤，并且在推理阶段需要访问一个包含带标签姿态图像的数据库。

绝对位姿回归（APR）根据输入查询图像直接估计相机位姿。Kendall [25], 首次提出了 APR，采用经过修改的 GoogLeNet 架构，将 Softmax 分类层替换为一系列输出姿态估计的全连接层。尽管 APR 的准确度低于经典的基于结构的方法，但它具有多项优势，例如推理速度更快（毫秒级）、内存占用更低（兆字节级），且不需要特征工程。研究人员提出了多种编码器和 MLP 架构，以提高 APR 的准确度 [11, 23, 33, 33, 44, 49]。此外，也有研究建议修改损失函数，以改善方向与位置目标之间的平衡 [24] ，或纳入其他类型的信息 [9]。有关相机姿态回归中采用的各类架构设计和损失函数的深入分析，请参阅 [43]。尽管绝对位姿回归（APR）模型推理速度更快且内存需求较低，但通常需要针对每个场景单独训练，才能达到与其他方法相比具有竞争力的准确度。这一局限引发了人们对 APR 相较于 [4, 6, 8, 36]等准确度显著更高的技术究竟有何价值的质疑。为解决这一问题，近期的研究提出将 APR 扩展至多场景范式。 [3] 提出了多场景PoseNet（MSPN），利用 CNN 回归图像特征向量。随后，两个分支分别处理该特征向量。第一个分支通过一个带有 softmax 运算符的全连接层对相关场景进行分类。估计出的场景索引从数据库中选取场景特定权重。第二个分支将这些权重与特征向量结合起来，以回归位姿。模型通过对每个场景同时应用二元交叉熵（用于分类）和相机位姿损失，进行端到端训练。受 Transformer 在计算机视觉任务中近期取得的成功 [12, 19, 29] 启发， [46] 提出采用共享的卷积网络骨干，后接由 Transformer 和回归头组成的两个网络分支。一个分支回归输入查询图像的位姿，另一个则回归其方向。采用 Transformer 编码器的主要动机是关注位姿相关特征，并使用解码器将编码后的场景标识符转换为潜在位姿表示。

超网络 最早由 Ha 等人 [22] 提出。这类神经网络旨在预测主网络的权重。主网络的权重可以根据特定输入进行调整，从而构建出表达能力更强、适应性更高的模型。超网络已应用于多种任务，例如学习隐式神经表示 [16], 、语义分割 [34], 3、3D 场景表示与建模 [28, 48, 50] 以及持续学习 [59], , 等等。

3. 使用超网络的绝对位姿回归

HyperPose 是一种通用方法，适用于单场景和多场景情形，可用于增强绝对位姿回归方案。它由两个关键组件构成：用于估计相机绝对位置的主网络（绝对位姿回归），以及根据输入查询图像计算主网络回归权重的超网络。主网络对输入图像进行前向传播，以定位拍摄相机。相机位姿表示为元组 $< \mathbf{x}, \mathbf{q} >$，其中 $\mathbf{x} \in \mathbb{R}^3$ 是相机在世界坐标中的位置，而 $\mathbf{q} \in \mathbb{R}^4$ 是其方向的四元数编码。按照 [13],，我们将超网络 H 表示为参数化函数$\theta_h = H(\theta, \mathbf{x})$，其中 θ 表示超网络的参数，$\mathbf{x}$ 表示网络的输入，而 θ_h 表示针对输入查询图像的回归头层参数 I_q。

3.1. 网络架构

3.1.1. 用于单场景APR的 HyperPose

我们将所提出的基于 HyperPose 的方法应用于单场景绝对位姿回归，采用基于 [25], 的基线绝对位姿回归器以及最先进的 *AtLoc* 绝对位姿回归器 [61]。

采用超网络的基线绝对位姿回归。 如图1a所示，基线模型的主网络由一个 CNN 主干网络组成，用于生成编码输入查询图像的特征向量。随后，该向量由两个回归头处理。第一个回归头用于估计相机位置，另一个用于估计方向。图1b 展示了如何将超网络集成到该架构中。输入

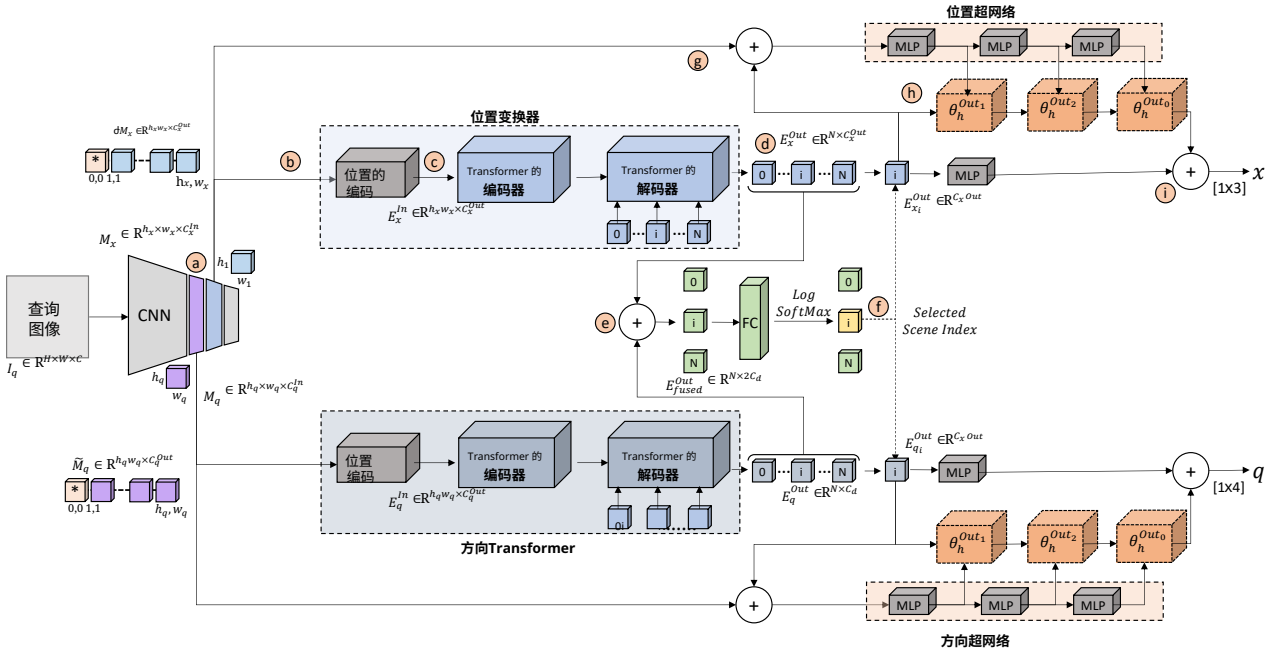


图 2. **带超网络的 MS-Transformer（MS-HyperPose）** - 所提出的采用超网络的多场景绝对位姿回归架构。主网络采用位置和方向 Transformer 的双分支架构，从底层卷积主干网络中提取激活图。超网络利用输入查询图像生成主网络中回归头的权重。这些自适应权重与 Transformer 生成的潜在向量相结合，用于估计相机位姿，包括空间位置（x）和角度方向（q）。

超网络的输入是由主网络的 CNN 主干网络生成的特征向量。该向量经过一个全连接（FC）层和一个 Swish 激活函数 [36]处理。处理后的向量作为多层感知机（MLP）的输入，由其输出主网络回归头中单个层的权重。随后，主网络的位置回归头和方向回归头估计相机位姿。

带超网络的 AtLoc。 [61] 介绍了注意力引导的相机定位（AtLoc），这是一种基于自注意力的绝对位姿回归模型。在该架构中，视觉编码器将图像编码为潜在向量。基于提取的特征，注意力模块计算自注意力图，将该表示重新加权到新的特征空间中。随后，位姿回归器将注意力算子处理后的特征映射为相机位姿估计。我们在 AtLoc 中加入超网络，与基线绝对位姿回归模型中的做法相同。在这种情况下，超网络的输入是由视觉编码器计算得到的重新加权表示特征向量。

3.1.2. 用于多场景绝对位姿回归方法的 HyperPose

所提出的多场景模型名为*MS-HyperPose* 是对 [45]中介绍的架构的扩展。如图 2所示，输入查询图像 I_q 由一个卷积主干网络处理，该网络从两个不同层计算中间激活图。随后，这些激活图被用作主网络和超网络的输入。在主网络中，位置和方向 Transformer 对激活图进行处理，并将序列表示整合为单个潜在向量。将激活图转换为表示向量后，这些向量与潜在向量一同输入超网络中的多层感知机（MLP）层，以生成用于平移和方向回归头的权重 θ_h 。利用超网络生成的权重 θ_h H 以及主网络中的回归头，即可估计输入查询图像的位置和方向。

主网络 由一个卷积主干网络和两个独立分支组成，分别用于回归相机的位置和方向。每个分支都包含专用的 Transformer 和一个多层感知机（MLP）头。研究表明，在视觉 Transformer（ViT）的早期视觉处理中，用卷积干网络替换 patchify 干网络能够提升优化稳定性，并进一步提高性能 [64]。因此，给定输入查询图像 $I_q \in \mathbb{R}^{H \times W \times C}$，共享的卷积主干网络会生成激活图。根据特定的学习任务，从卷积神经网络中选取两组不同的激活图（图 2a）

。激活图表示为 $M \in \mathbb{R}^{h\times w\times C^{In}}$，并通过卷积层和展平操作处理，转换为 $\tilde{M} \in \mathbb{R}^{h\times w\times C^{Out}} 1\times 1$（图 2b）。

接下来 [45]，我们采用了 [12]，中的 Transformer 架构，并对标准编码器和解码器进行修改，在每个注意力层添加位置编码。每个 Transformer 包含 L 个相同的模块，每个模块均包含多头注意力（MHA）机制和一个带有 Swish 激活函数的两层多层感知机 [36]。与 [58]，中的做法相同，层归一化（LN）在 MHA 和多层感知机处理之前应用，输出通过残差连接和随机失活进行合并。为保留空间信息，每个激活图位置都会获得一个可学习的编码。处理后的序列作为 Transformer 的输入，如下所示（图 2c）：

$$E^{In} = [T, \tilde{M}] \in \mathbb{R}^{(hw)\times C^{Out}}, T \in \mathbb{R}^{C_M^{Out}}. \quad (1)$$

其中，T是激活图的位置编码。

给定一个包含 N 个场景的数据集，Transformer 解码器会生成嵌入向量 $E^{Out} =\in \mathbb{R}^{N\times C^{Out}}$（图 2d）。对于一张查询图像，只有一个位置与其场景相匹配。组合后的 Transformer 输出（图 2e）经过一个全连接层和对数 SoftMax，选出概率最高的向量（图 2f）。

选定的向量 $E_{x_i}^{Out}$和 $E_{q_i}^{Out}$通过一个多层感知机进行处理。最后，利用超网络预测的权重，通过对主网络和超网络的输出求和，获得估计的相机位置和方向（图 2h 和 2i）。

超网络由多个多层感知机组成，每个多层感知机负责回归主网络中相应回归头（位置或方向）的权重。其输入为任务特定编码激活图 M_x 和 M_q的~ ~和，以及全局潜向量 $E_{x_i}^{Out}$和 $E_{q_i}^{Out}$（图 2g）。为确保尺寸兼容性，编码激活图在求和前会通过一个全连接层。随后，该向量经由采用 Swish 激活函数的多层感知机处理，生成另一组多层感知机的输入，其中每个多层感知机估计主网络回归头中特定层的权重：

$$\theta_h^{Out_j} = H(\theta_h^{Out_j}, E_i^{Out}) \in \mathbb{R}^{C_{E_i}^{Out}\times C_H^{Out_j}}, \quad (2)$$

其中，j 是回归头层的索引。

3.2. 优化准则

定位误差通过比较真实位姿$\mathbf{p}_{gt} =< \mathbf{x}^{gt},\mathbf{q}_{gt} >$的平移位置和方向来计算，其中$\mathbf{x} \in \mathbb{R}^3$ 表示相机在世界坐标系中的位置，$\mathbf{q} \in \mathbb{R}^4$ 表示其编码为四元数的方向，估计位姿为$\mathbf{p}_{est} =< \mathbf{x}_{est},\mathbf{q}_{est} >$。平移误差记为 L_x，通过真实相机位置与估计相机位置之间的欧氏距离计算得出。方向误差 L_q定义为真实单位四元数与估计单位四元数之间的欧氏距离

[43]:

$$L_x = ||\mathbf{x}_0 - \mathbf{x}||_2, \quad (3)$$

$$L_q = \left\| \mathbf{q_0} - \frac{\mathbf{q}}{||\mathbf{q}||} \right\|_2 \quad (4)$$

其中，$\mathbf{q}$ 是归一化单位四元数，以确保其为有效的方向编码。

作为一个多回归优化问题，我们采用 [24]中概述的方法，根据每项任务的不确定性使用不同的加权方案，以平衡为不同目标分别定义的损失函数。模型优化的是聚合损失函数，即各加权单项损失的组合。

$$L_p = L_x \exp(-s_x) + s_x + L_q \exp(-s_q) + s_q, \quad (5)$$

其中，s_x 和 s_q 是学习参数。作为进一步的优化，我们采用 [44]中详述的方法，并遵循三步训练过程。首先，使用公式5中的聚合损失训练整个网络。在接下来的两个阶段中，每个MLP头分别使用其对应的损失函数进行微调。

3.3. 实现细节

共享主干网络使用 EfficientNet-B0 [31, 52]。我们将两个端点（降采样级别）的激活图作为位置分支和方向分支的输入：$m_{rdct4} \in \mathbb{R}^{14\times 14\times 112}$ 和 $m_{rdct3} \in \mathbb{R}^{28\times 28\times 40}$，分别对应两个分支。根据 [45]，，我们对每个激活图应用线性投影，将其映射到公共深度维度 $C_M^{Out} = 256$，并学习相同深度的位置编码。主网络中的回归头由带有输入层和输出层的两层MLP组成。位置和方向的输入层维度均为1024，而输出维度分别为3和4。

单场景APR。对于基线绝对位姿回归以及 AtLoc [61] 模型的 HyperPose 扩展，位置分支的超网络会为一个维度为256的多层感知机生成权重，而方向超网络则会生成维度为512的权重。

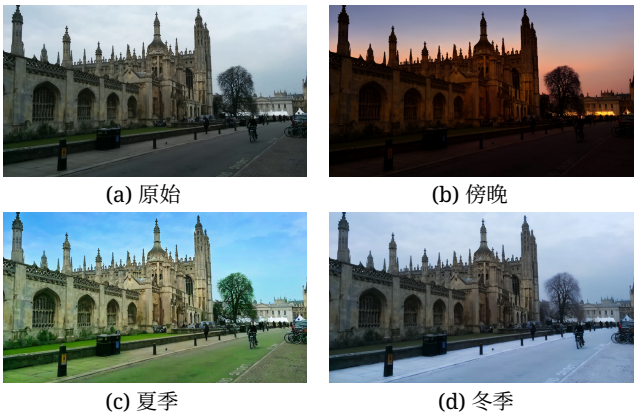
多场景APR。与 [45]，相比，我们在主网络架构的每个分支（位置和方向）中采用了四层（而非六层）Transformer。每个模块由一个包含四个头的多头注意力层组成，后接一个保持输入维度不变的多层感知机。每个多层感知机层的丢弃率设为 $p = 0.1$。主网络中每个Transformer的输出维度为 $C_M^{Out} = 256$。超网络的位置分支和方向分支生成三个回归层（输入层、隐藏层和输出层）的权重。位置分支和方向分支生成的权重数量不同。超网络的位置分支生成一个

表 1. **使用剑桥地标数据集对添加和未添加所提出的超网络的 APR 架构进行比较分析**。我们报告位置/方向误差中位数，单位分别为米/度。最佳性能以**粗体**标出。(*) AtLoc 的作者未报告其在剑桥数据集上的结果，因此我们在该数据集上训练了 AtLoc 模型。

方法	国王学院	旧医院	商店立面	圣玛丽	Avg.
AtLoc* [61]	1.53，**3.21**	2.17，3.99	0.93，4.61	2.14，5.76	1.69，4.39
AtLoc 搭配超网络	**0.96**,3.43	**2.04**,**3.61**	**0.88**,**4.22**	**1.72**,**5.44**	**1.40**,**4.18**
基线绝对位姿回归	0.89,2.29	1.49,3.32	0.74,4.79	1.40,4.95	1.13,3.84
使用超网络的基线 APR	**0.77**,**2.07**	**1.47**，**3.29**	**0.72**，**4.01**	**1.37**,**4.90**	**1.08**,**3.57**

维度为256、128和3的多层感知机的权重，而方向分支生成一个维度为512、512和4的多层感知机的权重。随后，超网络生成的权重被输入主网络的回归头，以估计输入查询图像的位置和方向。回归头的输入层和隐藏层采用Swish 激活函数。

4. 扩展剑桥地标数据集



(a) 原始 (b) 傍晚 (c) 夏季 (d) 冬季

图 3. 所提出的扩展剑桥地标（ECL）数据集中，国王学院场景的一个样例

为了进一步评估 HyperPose 方法的局限性，我们构建了剑桥地标 [25] 基准数据集的扩展版本，称为扩展剑桥地标（ECL）数据集。ECL 数据集利用原始场景，通过为每个地点创建合成的傍晚、冬季和夏季变体，引入了由外观变化带来的新定位挑战。这些场景变体通过引入光照、天气模式、植被密度等方面的变化，模拟真实世界的环境变化。分析模型在 ECL 变化版本及其对应的原始剑桥地标场景中的表现，可以评估模型对真实世界变化的鲁棒性。这些增强是借助新颖的 *InstructPix2Pix* 方法 [10]完成的。ECL 数据集的视觉样例见补充材料。图 3 展示了扩展数据集中国王学院场景的样例。更多 ECL 数据集中各场景的样例见补充材料。对这些更具挑战性的条件与未经修改的序列进行定量位姿回归分析，有助于科学地评估模型的适应能力。我们推测，揭示传统测试集中未显现的局限性，将推动开发出更具不变性的解决方案。因此，作为对既有基准测试的可控扩展，程序生成的数据集能够引导开发适用于真实世界的鲁棒模型。

5. 实验结果

我们使用 7 Scenes [21]和剑桥地标 [25] 数据集开展实验评估，以评估绝对位姿回归。剑桥地标数据集涵盖室外城市环境，面积范围为 900 至 5500 平方米。我们报告了其中六个场景中的四个场景的结果。由于先前研究的覆盖范围有限，我们未纳入其余两个场景。为比较鲁棒性，我们还报告了扩展剑桥地标数据集（ECL）上的结果（补充材料第 1 节）。7 Scenes 数据集由七个小规模室内场景组成，空间范围约为 1 至 10 平方米。这些数据集带来了各种定位挑战，包括尺度差异、室内与室外环境的区别、重复元素、无纹理特征、显著的视点变化，以及训练集与测试集之间轨迹的变化。为了更全面地分析所提出的方法，我们还在 Oxford RobotCar 数据集 [30]上进行了额外评估。这些动态环境因素可用于评估 HyperPose 在多样且具有挑战性的视觉环境中提升定位精度的能力。我们的实验流程遵循与 [9, 61]相同的训练和评估协议。有关这些序列的更多信息，以及模型训练方法和超参数选择的全面说明，见补充材料。

我们评估了在单场景绝对位姿回归器中集成超网络的影响。为此，我们比较了两种模型变体（分别带有和不带有超网络）的准确性，并报告中位位置误差和方向误差。表1展示了剑桥数据集上的结果。超网络提升了所提出的基线绝对位姿回归

表2。**与剑桥地标单场景最先进方法的比较：**我们报告以米/度为单位的位置/方向误差中位数。最有效的单场景和多场景绝对位姿回归方法以**粗体**标示。

		学院	医院	商店立面	圣玛丽	Avg.
	DSAC* [6]	0.18,0.3°	0.21,0.4°	0.05,0.3°	0.15,0.5°	0.15,0.4°
Seq.	MapNet [9]	1.08,1.9°	1.94,3.9°	1.49,4.2°	2.00,4.5°	1.63,3.6°
	GL-Net [67]	0.59,0.7°	1.88,2.8°	0.50,2.9°	1.90,3.3°	1.22,2.4°
IR	VLAD [56]	2.80,5.7°	4.01,7.1°	1.11,7.6°	2.31,8.0°	2.56,7.1°
	VLAD+Inter [42]	1.48,4.5°	2.68,4.6°	0.90,4.3°	1.62,6.1°	1.67,4.9°
RPR	EssNet [68]	0.76,1.9°	1.39,2.8°	0.84,4.3°	1.32,4.7°	1.08,3.4°
	NC-EssNet [68]	0.61,1.6°	0.95,2.7°	0.7,3.4°	1.12,3.6°	0.85,2.8°
	RelocGNN[21]	0.48,1.0°	1.14,2.5°	0.48,2.5°	1.52,3.2°	0.91,2.3°
APR	PoseNet [25]	1.92,5.40°	2.31,5.38°	1.46,8.08°	2.65,8.48°	2.08,6.83°
	BayesianPN [23]	1.74,4.06°	2.57,5.14°	1.25,7.54°	2.11,8.38°	1.91,6.28°
	LSTM-PN [60]	0.99,3.65°	1.51,4.29°	1.18,7.44°	1.52,6.68°	1.30,5.57°
	SVS-Pose [33]	1.06,2.81°	1.50,4.03°	**0.63**,5.73°	2.11,8.11°	1.32,5.17°
	GPoseNet [11]	1.61,2.29°	2.62,3.89°	1.14,5.73°	2.93,6.46°	2.07,4.59°
	PoseNetLearn [24]	0.99,1.06°	2.17,**2.94**°	1.05,3.97°	1.49,3.43°	1.42,2.85°
	GeoPoseNet [24]	0.88, **1.04**°	3.20, 3.29°	0.88, 3.78°	1.57, **3.32**°	1.63, 2.86°
	IRPNet [44]	1.18,2.19°	1.87,3.38°	0.72,**3.47**°	1.87,4.94°	1.41,3.50°
	使用超网络的基线 APR	**0.61**, 1。84°	**1.44**, 3。03°	0.70, 3。62°	**1.37**, 4。85°	**1.03**, **2**。**67**°
MS-APR	MSPN [3]	1.73,3.65°	2.55,4.05°	2.92,7.49°	2.67,6.18°	2.47,5.34°
	MS-Trans[46]	0.83,1.47°	**1.81**,2.39°	0.86,**3.07**°	1.62,3.99°	1.28,2.73°
	MS-HyperPose	**0.78**,**1.18**°	1.84,**2.11**°	**0.83**, 3。13°	**1.61**, **3.22**°	**1.27**,**2.41**°

和 AtLoc [61]模型的准确性。基线绝对位姿回归模型的大小为 26.7MB，推理时间为 9.32ms；加入超网络后，模型大小增至 40.9MB，推理时间增至 9.45ms。原始发表的 AtLoc 模型大小为 97.9MB，推理时间为 5.14ms。增强超网络的 AtLoc 版本，模型大小为 116.8MB，推理时间为 5.56ms。请注意，报告的误差是指模型训练第一阶段后的模型表现，未应用训练细节中所述的另外两个优化步骤（见补充材料）。此外，表 2 和表 3 分别将所提出的采用超网络架构的基线绝对位姿回归模型的中位位置误差和方向误差，与近期最先进的定位算法在剑桥地标和 7Scenes 数据集上的结果进行了比较。对于室内和室外数据集，除领先的绝对位姿回归方法的结果外，我们还报告了基于结构、基于序列、IR 和相对位姿回归方法的中位位置误差和方向误差。尽管 DSAC* [6] 达到了最先进的准确度，但与所有其他绝对位姿回归方法相比，HyperPose 在剑桥室外数据集上取得了最准确的平均 $1.03m$、 $2.67°$，在 7Scenes 室内数据集上取得了最准确的平均$0.19m$、 $7.48°$ 。这些定位误差表明，其性能可与领先的相对位姿回归技术相媲美；增强超网络的基线模型在位置精度和旋转精度方面分别排名第三和第二/第四（7Scenes/剑桥）。接下来，我们将比较所提出的 MS-HyperPose 架构与其他多场景绝对位姿回归方法。表 2和 3 分别展示了在剑桥地标和 7 Scenes 数据集上的中位数位置误差和方向误差。在室内和室外基准测试中，我们的 MS-HyperPose 方法取得了最准确的平均 1.27米， $2.41°$ 在剑桥数据集上；以及 0.18米， $7.12°$ 在 7Scenes 上，优于其他单场景和多场景绝对位姿回归方法。此外，在剑桥数据集的 8 个参数中有 6 个参数，以及 7 Scenes 的 14 个参数中有 6 个参数上，MS-HyperPose 在位置误差和旋转误差方面均最为准确。MS-HyperPose 的运行时间为 30.93毫秒，与参考架构 [46]相近。其模型大小为 571MB，明显大于原始模型的 71.1MB。我们将在补充材料的局限性与未来工作 部分讨论存储方面的考量。

表4 报告了 AtLoc [61]与我们使用 Oxford RobotCar 数据集的超网络增强变体相比的位置和方向误差的均值和中位数 [30]。我们还将增强模型与 [9], 该数据集使用图像序列时的最先进的绝对位姿回归进行了比较。定位结果的可视化内容可在补充材料中找到。

我们还评估了扩展剑桥数据集（ECL）上基线绝对位姿回归模型的两种配置，即使用和不使用超网络的配置。在所有场景和变化下，超网络增强模型的性能均优于基线，展现出更强的适应能力

表 3。**与单场景最先进方法的比较 - 7Scenes：**我们报告位置/方向误差中位数，单位分别为米/度。最有效的单场景和多场景绝对位姿回归方法以**粗体**标示。

		国际象棋	Fire	头部	办公室	南瓜	厨房	楼梯	Avg.
SCR	DSAC* [6]	0.02,1.11°	0.02,1.24°	0.01,1.82°	0.03,1.18°	0.04,1.41°	0.03,1.70°	0.04,1.42°	0.03,1.41°
	Marepo[15]	0.02,1.24°	0.02,1.39°	0.02,2.03°	0.03,1.26°	0.04,1.48°	0.04,1.71°	0.06,1.67°	0.03,1.54°
Seq.	LsG [66]	0.09,3.3°	0.26,10.9°	0.17,12.7°	0.18,5.5°	0.20,3.7°	0.23,4.9°	0.23,11.3°	0.19,7.74°
	MapNet[9]	0.08,3.3°	0.27,11.7°	0.18,13.3°	0.17,5.2°	0.22,4.0°	0.23,4.9°	0.30,12.1°	0.21,7.78°
	GL-Net[67]	0.08,2.8°	0.26,8.9°	0.17,11.4°	0.18,13.3°	0.15,2.8°	0.25,4.5°	0.23,8.8°	0.19,7.50°
IR	VLAD [56]	0.21,12.5°	0.33,13.8°	0.15,14.9°	0.28,11.2°	0.31,11.2°	0.30,11.3°	0.25,12.3°	0.26,12.46°
	VLAD+Inter[42]	0.18,10.0°	0.33,12.4°	0.14,14.3°	0.25,10.1°	0.26,9.4°	0.27,11.1°	0.24,14.7°	0.24,11.71°
RPR	NN-Net [26]	0.13,6.5°	0.26,12.7°	0.14,12.3°	0.21,7.4°	0.24,6.4°	0.24,8.0°	0.27,11.8°	0.21,9.30°
	RelocNet[2]	0.12,4.1°	0.26,10.4°	0.14,10.5°	0.18,5.3°	0.26,4.2°	0.23,5.1°	0.28,7.5°	0.21,6.73°
	EssNet [68]	0.13,5.1°	0.27,10.1°	0.15,9.9°	0.21,6.9°	0.22,6.1°	0.23,6.9°	0.32,11.2°	0.22,8.03°
	NC-EssNet [68]	0.12,5.6°	0.26,9.6°	0.14,10.7°	0.20,6.7°	0.22,5.7°	0.22,6.3°	0.31,7.9°	0.21,7.50°
	RelocGNN[21]	0.08,2.7°	0.21,7.5°	0.13,8.70°	0.15,4.1°	0.15,3.5°	0.19,3.7°	0.22,6.5°	0.16,5.24°
	CamNet [17]	0.04,1.7°	0.03,1.7°	0.05,2.00°	0.04,1.60°	0.04,1.6°	0.04,1.6°	0.04,1.5°	0.04,1.70°
APR	PoseNet [25]	0.32,8.12°	0.47,14.4°	0.29,12.0°	0.48,7.68°	0.47,8.42°	0.59,8.64°	0.47,13.8°	0.44,10.4°
	BayesianPN [23]	0.37,7.24°	0.43,13.7°	0.31,12.0°	0.48,8.04°	0.61,7.08°	0.58,7.54°	0.48,13.1°	0.47,9.81°
	LSTM-PN [60]	0.24,5.77°	0.34,11.9°	0.21,13.7°	0.30,8.08°	0.33,7.00°	0.37,8.83°	0.40,13.7°	0.31,9.85°
	GPoseNet [11]	0.20,7.11°	0.38,12.3°	0.21,13.8°	0.28,8.83°	0.37,6.94°	0.35,8.15°	0.37,12.5°	0.31,9.95°
	PoseNetLearn[24]	0.14,4.50°	0.27,11.8°	0.18,12.1°	0.20,5.77°	0.25,4.82°	0.24,5.52°	0.37,10.6°	0.24,7.87°
	GeoPoseNet[24]	0.13,4.48°	0.27,11.3°	0.17,13.0°	0.19,5.55°	0.26,4.75°	0.23,**5.35**°	0.35,12.4°	0.23,8.12°
	IRPNet[44]	0.13,5.64°	**0.25**，9。67°	**0.15**，13。1°	0.24,6.33°	0.22,5.78°	0.30,7.29°	0.34,11.6°	0.23,8.49°
	AtLoc[61]	0.10,**4.07**°	**0.25**,11.4°	0.16,11.8°	0.17,5.34°	0.21,**4.37**°	0.23,5.42°	**0.26**，10。5°	0.20,7.56°
	TransBoNet[49]	0.11,4.48°	**0.25**,12.5°	0.18,14.0°	0.20,**5.08**°	0.19,4.77°	0.17,**5.35**°	0.30,13.0°	0.20,8.45°
	基线绝对位姿回归								
	使用超网络	**0.09**,4.47°	0.28,**9.44**°	**0.15**,**11.4**°	**0.16**,6.29°	**0.19**,4.62°	**0.18**,6.95°	0.28,**9.15**°	**0.19**,**7.48**°
MS-APR	MSPN[3]	**0.09**，4。76°	0.29,10.5°	0.16,13.1°	**0.16**，6。80°	0.19,5.50°	0.21,6.61°	0.31,11.6°	0.20,7.56°
	MS-Trans[46]	0.11,4.66°	0.24,9.60°	0.14，12.2°	0.17，**5.66**°	0.18，**4.44**°	**0.17**，**5.94**°	**0.26**，8.45°	0.18，7.28°
	MS-HyperPose	0.11,**4.34**°	**0.23**,9.79°	**0.13**,**10.7**°	0.17,6.05°	**0.16**，5。24°	**0.17**，6。86°	0.27，**6。00**°	**0.18**，**7。00**°

表 4。**针对 Oxford RobotCar 数据集，对采用与未采用所提出超网络的不同 APR架构进行比较分析：**我们报告以米/度为单位的位置/方向误差的均值和中位数。

	MapNet[9]		**AtLoc**[61]		**AtLoc（含超网络）**	
序列	Mean	中位数	Mean	中位数	Mean	中位数
LOOP1	8.76, 3.46	5.79, 1.54	8.61, 4.58	5.68, 2.23	**8.42, 3.30**	**4.91, 1.33**
LOOP2	9.84, 3.96	4.91, 1.67	8.61, 4.58	5.68, 2.23	**8.10, 3.04**	**4.73, 1.27**
FULL1	41.4, 12.5	17.94, 6.68	29.6, 12.4	11.10, 5.28	**29.58, 6.77**	**7.86, 1.38**
FULL2	59.3, 14.8	20.04, 6.39	48.2, 11.1	12.2, 4.63	**45.96, 9.60**	**8.38, 1.45**

能够适应场景中固有的季节变化和多样的光照条件。关于 HyperPose 对场景变化的鲁棒性和韧性的详细分析见补充材料。

6. 结论

在这项工作中，我们提出了一种基于超网络的方法，以改进 APR 模型。实验结果表明，超网络能够提高单场景 APR 模型的精度。我们还提出了 MS-HyperPose，这是一种新颖的多场景相机姿态回归模型，通过超网络增强，在室内和室外定位数据集上实现了新的最先进精度。我们推出了扩展剑桥地标（ECL）数据集，以促进对鲁棒定位方法的研究。

参考文献

[1] Eduardo Arnold、Jamie Wynn、Sara Vicente、Guillermo Garcia-Hernando、Aron Monszpart、Victor Prisacariu、Dani-yar Turmukhambetov 和 EricBrachmann。无地图视觉

重定位：相对于单幅图像的度量位姿。载于《欧洲计算机视觉会议论文集》（ECCV），第690–708页。施普林格，2022年。3[2] Vassileios Balntas、Shuda Li 和 Victor Prisacariu。Reloc-net：使用神经网络进行连续度量学习重定位。载于《欧洲计算机视觉会议论文集》（ECCV），2018年。1, 8[3] Hunter Blanton、Connor Greenwell、Scott Workman 和 Nathan Jacobs。将绝对位姿回归扩展至多个场景。载于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），第38–39页，2020年。2, 3, 7, 8[4] E. Brachmann 和 C. Rother。少即是多——通过3D表面回归实现6D相机定位。载于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），第4654–4662页，2018年。1, 3[5] Eric Brachmann 和 Carsten Rother。将专家样本共识应用于相机重新定位。载于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），第7525–7534页，2019。[6]E. Brachmann 和 C. Rother。使用RGB和RGB-D图像进行视觉相机重新定位：DSAC。IEEE《模式分析与机器智能汇刊》，44(09):5847–5865，2022年。3, 7, 8[7]Eric Brachmann、Frank Michel、Alexander Krull、Michael Ying Yang、Stefan Gumhold 和 Carsten Rother。从单幅RGB图像对物体和场景进行不确定性驱动的6D位姿估计。载于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），第3364–3372页，2016年。[8] E. Brachmann、A. Krull、S. Nowozin、J. Shotton、F. Michel、S. Gumhold 和 C. Rother。DSAC——用于相机定位的可微RANSAC。载于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），第2492–2500页，美国加利福尼亚州洛斯阿拉米托斯，2017年。1, 3[9] Samarth Brahmbhatt、Jinwei Gu、Kihwan Kim、James Hays 和 Jan Kautz。面向相机定位的地图几何感知学习。载于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），2018年。3, 6, 7, 8[10] T. Brooks、A. Holynski 和 A. A. Efros。InstructPix2Pix：学习遵循图像编辑指令。载于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），第18392–18402页，美国加利福尼亚州洛斯阿拉米托斯，2023年。IEEE计算机学会。6[11] Ming Cai、Chunhua Shen 和 Ian Reid。相机重定位的混合概率模型。载于《英国机器视觉会议论文集》，第238页，2018年。2, 3, 7, 8[12] Nicolas Carion、Francisco Massa、Gabriel Synnaeve、Nicolas Usunier、Alexander Kirillov 和 Sergey Zagoruyko。使用Transformer进行端到端目标检测。载于《欧洲计算机视觉会议论文集》（ECCV），第213–229页，尚，2020年。3, 5[13] Yoav Chai、Raja Giryes 和 Lior Wolf。参数化颜色增强的监督与无监督学习。载于

IEEE/CVF冬季计算机视觉应用会议（WACV）论文集，第992–1000页，2020年。3[14] Kefan Chen、Noah Snavely和Ameesh Makadia。利用方向学习进行宽基线相对相机位姿估计。2021年IEEE。在CVF计算机视觉与模式识别会议（CVPR）中，第3257–3267页，2021年。1[15] Shuai Chen、Tommaso Cavallari、Victor Adrian Prisacariu和Eric Brachmann。相对于地图的姿态回归用于视觉重新定位。在IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集中，第20665–20674页，2024年。1、2、3、8[16] Yinbo Chen和Xiaolong Wang。将Transformer作为隐式神经表示的元学习器。在欧洲计算机视觉会议（ECCV）论文集中，第170–187页。施普林格，2022年。2、3[17] 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。在IEEE国际计算机视觉会议（ICCV）论文集中，2019年。1、8[18] Siyan Dong、Qingnan Fan、He Wang、Ji Shi、Li Yi、Thomas Funkhouser、Baoquan Chen和Leonidas J Guibas。通过空间划分实现鲁棒神经路由，用于动态室内环境中的相机重定位。在IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集中，第8544–8554页，2021年。2[19] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszkoreit和Neil Houlsby。一幅图像胜过16×16个词：用于大规模图像识别的Transformer。在国际学习表征会议（ICLR）论文集中，2021年。3[20] Sovann En、Alexis Lechervy和Fr´ed´eric Jurie。Rpnet：用于相对相机位姿估计的端到端网络。在欧洲计算机视觉会议（ECCV）研讨会论文集中，第0–0页，2018年。1[21] B. Glocker、S. Izadi、J. Shotton和A. Criminisi。实时RGB-D相机重定位。在2013年IEEE国际混合与增强现实研讨会（ISMAR）中，第173–179页，2013年。6、7、8[22] David Ha、Andrew Dai和Quoc V Le。超网络。arXiv预印本arXiv:1609.09106，2016年。2、3[23] Alex Kendall和Roberto Cipolla。深度学习中相机重定位的不确定性建模。在IEEE机器人与自动化国际会议（ICRA）论文集中，第4762–4769页，2016年。2、3、7、8[24] A. Kendall和R. Cipolla。利用深度学习进行相机姿态回归的几何损失函数。在IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集中，第6555–6564页，2017年。2、3、5、7、8[25] A. Kendall、M. Grimes和R. Cipolla。PoseNet：用于实时六自由度相机重新定位的卷积网络。在IEEE国际计算机视觉会议（ICCV）论文集中，第2938–2946页，2015年。2、3、6、7、8[26] Z. Laskar、I. Melekhov、S. Kalia和J. Kannala。通过卷积神经网络计算成对相对姿态进行相机重定位。在IEEE国际计算机视觉会议研讨会（ICCVW）论文集中，第920–929页，2017年。1、8

积神经网络。在IEEE国际计算机视觉会议研讨会（ICCVW）论文集中，第920–929页，2017年。1、8[27] Xiaotian Li、Shuzhe Wang、Yi Zhao、Jakob Verbeek和Juho Kannala。用于视觉定位的层次化场景坐标分类与回归。在IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集中，第11983–11992页，2020年。1、2[28] Gidi Littwin和Lior Wolf。用于形状表示的深度元函数。在IEEE国际计算机视觉会议（ICCV）论文集中，第1824–1833页，2019年。3 – 。[29] Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin和Baining Guo。S win Transformer：使用移位窗口的层次化视觉Transformer。在IEEE国际计算机视觉会议（ICCV）论文集中，第10012–10022页，2021年。3[30] Will Maddern、Geoffrey Pascoe、Chris Linegar和Paul Newman。一年，1000公里：牛津RobotCar数据集。《国际机器人研究杂志》，36（15），2017年，1、3、6、7。[31] Luke Melaskyriazi。EfficientNet-PyTorch，2019年。5[32] Iaroslav Melekhov、Juha Ylioinas、Juho Kannala和Esa Rahtu。使用沙漏网络的基于图像的定位。在IEEE国际计算机视觉会议研讨会（ICCVW）论文集中，第870–877页，2017年。2 – 。[33] Tayyab Naseer和W. Burgard。户外环境中基于单目相机的六自由度全局定位深度回归。2017年IEEE/RSJ智能机器人与系统国际会议（IROS），第1525–1530页，2017年。2、3、7 – 。[34] Yuval Nirkin、Lior Wolf和Tal Hassner。HyperSeg：用于实时语义分割的逐块超网络。在IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集中，第4061–4070页，2021年。2、3 – 。[35]N. Radwan、A. Valada和W. Burgard。Vlocnet++：用于语义视觉定位和里程计的深度多任务学习。IEEE机器人与自动化快报，3(4)：4407–4414，2018年。2[36] Prajit Ramachandran、Barret Zoph和Quoc V Le。搜索激活函数。arXiv预印本 arXiv:1710.05941，2017年。3、4、5[37] P. Sarlin、C. Cadena、R. Siegwart和M. Dymczyk。从粗到细：大规模鲁棒层次化定位。在IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集中，第12708–12717页，2019年。1、2[38] Paul-Edouard Sarlin、Daniel DeTone、Tomasz Malisiewicz和Andrew Rabinovich。Super Glue：利用图神经网络学习特征匹配。在IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集中，第4938–4947页，2020年。[39] Paul-Edouard Sarlin、Ajaykumar Unagar、Mans Larsson、Hugo Germain、Carl Toft、Viktor Larsson、Marc Pollefeys、Vincent Lepetit、Lars Hammarstrand、Fredrik Kahl等。

回到特征：从像素到位姿学习鲁棒的相机定位。载于《IEEE/CVF计算机视觉与模式识别会议》（CVPR）论文集，第3247–3257页，2021年。[40] Torsten Sattler、Michal Havlena、Filip Radenovic、Konrad Schindler和Marc Pollefeys。用于大规模地点识别的超点与精细词汇表。载于《IEEE国际计算机视觉会议》（ICCV）论文集，第2102–2110页，2015年。[41] T. Sattler、B. Leibe和L. Kobbelt。面向大规模基于图像定位的高效有效优先匹配。IEEE模式分析与机器智能汇刊，39(9)：1744–1756，2017年。1，2[42] T. Sattler、Q. Zhou、M. Pollefeys和L. Leal-Taix´e。理解基于CNN的绝对相机位姿回归的局限性。载于《IEEE/CVF计算机视觉与模式识别会议》（CVPR）论文集，第3297–3307页，2019年。7，8[43]Yoli Shavit和Ron Ferens。深度学习相机位姿估计导论。载于arXiv预印本 arXiv:1907.05272，2019年。2，3，5[44] Yoli Shavit和Ron Ferens。我们真的需要特定于场景的位姿编码器吗？载于《国际模式识别会议》（ICPR）论文集，第3186–3192页。IEEE，2021年。2，3，5，7，8[45] Yoli Shavit、Ron Ferens和Yosi Keller。使用Transformer学习多场景绝对位姿回归。载于《IEEE国际计算机视觉会议》（ICCV）论文集，第2713–2722页，2021年。2，4，5[46] Yoli Shavit、Ron Ferens和Yosi Keller。使用Transformer学习多场景绝对位姿回归。载于《IEEE国际计算机视觉会议》（ICCV）论文集，2021年。2，3，7，8[47] Jamie Shotton、Ben Glocker、Christopher Zach、Shahram Izadi、Antonio Criminisi和Andrew Fitzgibbon。用于RGB-D图像相机重定位的场景坐标回归森林。载于《IEEE/CVF计算机视觉与模式识别会议》（CVPR）论文集。IEEE，2013年。1[48] Vincent Sitzmann、Julien Martel、Alexander Bergman、David Lindell和Gordon Wetzstein。具有周期激活函数的隐式神经表示。《神经信息处理系统进展》（NIPS），33：7473，2020年。3[49] Xiaogang Song、Hongjuan Li、Li Liang、Weiwei Shi、Guo Xie、Xiaofeng Lu和Xinhong Hei。Transbonet：利用Transformer瓶颈和注意力学习相机定位。《模式识别》，146：109975，2024年。2，3，8[50] Przemysław Spurek、Artur Kasymov、Marcin Mazur、Diana Janik、Slawomir K Tadeja、J Tabor、T Trzci´nski等。Hyperpocket：生成式点云补全。载于2022年IEEE/RSJ智能机器人与系统国际会议（IROS），第6848–6853页。IEEE，2022年。3[51] Hajime Taira、Masatoshi Okutomi、Torsten Sattler、Mircea Cimpoi、Marc Pollefeys、Josef Sivic、Tomas Pajdla和Akihiko Torii。Inloc：通过密集匹配和视图合成实现室内视觉定位。IEEE模式分析与机器智能汇刊，

43(4)：1293–1307，2021年。1，2[52] Mingxing Tan和Quoc Le。EfficientNet：重新思考卷积神经网络的模型缩放。第6105–6114页，美国加利福尼亚州长滩，2019年。5[53] Mingxing Tan和Quoc Le。Efficientnet：重新思考卷积神经网络的模型缩放。载于《国际机器学习会议》（ICML）论文集，第6105–6114页。PMLR，2019年。2[54] Shitao Tang、Sicong Tang、Andrea Tagliasacchi、Ping Tan和Yasutaka Furukawa。Neumap：通过自动转译码器实现神经坐标映射，用于相机定位。载于《IEEE/CVF计算机视觉与模式识别会议》（CVPR）论文集，第929–939页，2023年。1，2[55] Yi Tay、Zhe Zhao、Dara Bahri、Don Metzler和Da-Cheng Juan。HyperGrid Transformer：迈向用于多任务的单一模型。载于《国际学习表征会议》（ICLR）论文集，2021年。2[56] Akihiko Torii、Relja Arandjelov´c、Josef Sivic、Masatoshi Okutomi和Tomas Pajdla。通过视图合成实现全天候地点识别。载于《IEEE/CVF计算机视觉与模式识别会议》（CVPR）论文集，第1808–1817页，2015年。7，8[57] Mehmet Ozgur Turkoglu、Eric Brachmann、Konrad Schindler、Gabriel J Brostow和Aron Monszpart。使用图神经网络和相对位姿监督实现视觉相机重定位。载于2021年三维视觉国际会议（3DV），第145–155页。IEEE，2021年。1[58] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszko-reit、Llion Jones、Aidan N Gomez、Ł ukasz Kaiser和Illia Polosukhin。Attention is All You Need。载于《神经信息处理系统进展》（NIPS），第5998–6008页，2017年。2，5[59] Johannes Von Oswald、Christian Henning、Jo˜ao Sacramento和Benjamin F Grewe。利用超网络实现持续学习。arXiv预印本arXiv:1906.00695，2019年。3[60] F. Walch、C. Hazirbas、L. Leal-Taix´e、T. Sattler、S. Hilsen-beck和D. Cremers。使用长短期记忆网络进行结构化特征相关性的基于图像的定位。载于《IEEE国际计算机视觉会议》（ICCV）论文集，第627–637页，2017年。2，7，8[61] Bing Wang、Changhao Chen、Chris Xiaoxuan Lu、Peijun Zhao、Niki Trigoni和Andrew Markham。Atloc：注意力引导的相机定位。载于AAAI人工智能会议论文集，第10393–10401页，2020年。2，3，4，5，6，7，8[62] Dominik Winkelbauer、Maximilian Denninger和Rudolph Triebel。从合成训练数据中学习在新环境中进行定位。载于2021年IEEE机器人与自动化国际会议（ICRA），第5840–5846页。IEEE，2021年。1[63] J. Wu、L. Ma和X. Hu。深入研究用于相机重定位的卷积神经网络。载于《IEEE机器人与自动化国际会议》（ICRA）论文集，第5644–5651页，2017年。2[64] Tete Xiao、Mannat Singh、Eric Mintun、Trevor Darrell、Piotr Doll´ar和Ross Girshick。早期卷积有助于Transformer……

Transformer表现更佳。 神经信息处理系统进展（*NIPS*），34:30392–30400，2021年。4[65] 徐萌、王有辰、徐彬、张军、任健、Stefan Poslad和徐鹏飞。图像相机位姿估计技术的批判性分析。 *arXiv*预印本 *arXiv:2201.05816*，2022年。2[66] 薛飞、王鑫、闫子科、王秋原、王俊秋和查宏斌。局部支持全局：通过序列增强实现深度相机重定位。载于 *IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2019。8[67] 薛飞、吴鑫、蔡少军和王俊秋。利用图神经网络学习多视图相机重定位。载于 *IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2020年。7，8[68] 周群杰、Torsten Sattler、Marc Pollefeys和Laura Leal-Taixe。学习还是不学习：基于本质矩阵的视觉定位。载于 *IEEE*机器人与自动化国际会议（*ICRA*）论文集，第3319–3326页。IEEE，2020年。3，7，8