DOI: 10.19650/j.cnki.cjsi.J2413414

基于混合频域 Transformer 的相机位姿估计方法*

杨傲雷[1,2],甘少英[1],杨帮华[1],苗中华[1],徐昱琳[1]
(1. 上海大学机电工程与自动化学院 上海 200444; 2. 上海市电站自动化技术重点实验室 上海 200444)

摘 要:针对移动机器人相机位姿估计问题,提出一种基于混合频域 Transformer 的相机位姿估计方法,旨在从 RGB 图像中预测相机的位置与方向。首先,构建了室内场景数据集 RotIndoor,每个样本包含场景 RGB 图像和通过 VICON 系统获取的相机位姿真值;其次,提出位姿回归网络模型 CamPose,该模型融合空间域和频域的信息,提升了图像特征表达能力,进而实现高精度的相机位姿估计。具体而言,CamPose 引入基于差分卷积网络的特征增强模块,捕获图像细粒度特征;设计了频域编码层,通过傅里叶变换提取频率特征,并整合频域注意力模块,使模型感知不同频率成分的重要性。最后,在公开数据集 7Scenes 和 RotIndoor 上进行了实验验证表明,该方法在 7Scenes 数据集上的位姿估计误差为 0.17 m/7.85°,在 RotIndoor 上定位精度提高了 23%。

关键词: 相机位姿估计;深度学习;特征增强;频域编码;Transformer
中图分类号: TP391 TH86 **文献标识码**: A **国家标准学科分类代码**: 510.4050

Camera pose estimation based on hybrid frequency domain and Transformer

Yang Aolei[1,2], Gan Shaoying[1], Yang Banghua[1], Miao Zhonghua[1], Xu Yulin[1]
(*1. School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China;*
2. Shanghai Key Laboratory of Power Station Automation Technology, Shanghai 200444, China)

Abstract: To address the challenges of camera pose estimation and mobile robot localization, a camera pose estimation method is proposed based on a hybrid frequency domain Transformer to predict the position and orientation of a camera from RGB images. Firstly, a camera pose estimation dataset, RotIndoor, is constructed based on indoor scenes, with each sample containing an RGB image of the scene and the ground truth camera poses obtained from a VICON system. Secondly, a pose regression network model, CamPose, is introduced, which effectively integrates spatial and frequency domain information to enhance the representation capability of image features, ultimately achieving higher accuracy in camera pose estimation. Specifically, CamPose incorporates a feature enhancement module based on differential convolution networks to capture fine-grained features within the images. Additionally, a frequency domain encoding layer is designed that applies Fourier transformation to extract frequency characteristics while integrating a frequency domain attention module, enabling the model to sensitively perceive the importance of different frequency components. Finally, experiments are implemented on the public datasets 7Scenes and RotIndoor. The experimental results show that the pose estimation error on the 7Scenes dataset is reduced to 0.17 m/7.85°, and the positioning accuracy on RotIndoor is improved by 23% compared to other methods.

Keywords: camera pose estimation; deep learning; feature enhancement; frequency domain encoding; Transformer

0 引 言

相机位姿估计作为计算机视觉领域中的关键任务之一,在自动驾驶、移动机器人[1-3]、增强现实等多种应用中发挥着基础性作用。相机位姿估计也被称为相机定位,其核心目标是通过分析图像数据,准确估计相机在三维空间中的位置和方向,从而实现对场景的有效感知与理解。相机定位不仅是技术上的需求,更是实现智能系统自主导航与交互的基石。

收稿日期:2024-10-26 Received Date: 2024-10-26
*基金项目:国家重点研发计划(2023YFF1203503)、上海市自然科学基金(22ZR1424200)项目资助

传统的相机位姿估计方法主要包括几何结构法和图像检索法。几何结构法通常依赖于局部特征描述子如尺度不变特征转换(scale-invariant feature transform, SIFT)、加速鲁棒特征(speeded up robust features, SURF)等来提取和匹配关键点,通过解决 n 点透视定位问题(perspective-n-point, PnP)问题来推断相机的姿态。这类方法的优点在于其较强的可解释性和相对较低的计算复杂度,然而,它们严重依赖于准确的 2D~3D 匹配,易受到噪声和光照变化的影响,并且在处理纹理较弱或存在遮挡的场景时,往往表现不佳,限制了其在复杂环境中的应用[4-5]。相比之下,图像检索法通过匹配全局描述符,从大型图像数据库中寻找与查询图像相似的参考图像[6]。尽管这种方法在理论上可以通过已知的参考图像的三维模型提高定位精度,但它在实际应用中受到图像分辨率和计算资源消耗等因素的制约。此外,图像检索往往难以找到高相似度的参考图像,从而进一步影响位姿解算的准确性。

近年来,深度学习的迅猛发展为相机位姿估计带来了新的机遇。基于卷积神经网络(convolutional neural network, CNN)的方法能够直接从图像中学习特征表示,并通过回归模型实现相机位姿的预测。其中,Kendall 等[7]提出的 PoseNet 是首个完全基于深度学习的端到端相机位姿估计方法,能够从单幅 RGB 图像中直接回归相机的绝对位姿。然而,这些方法在泛化能力和鲁棒性方面仍存在不足,尤其在复杂真实场景中表现不够稳定。为了解决这些问题,研究者们提出了更多基于 RGB 图像的绝对位姿估计方法,这些方法通常通过更深层的网络架构[8-10]或设计不同的损失函数[11-12]来提升性能。例如,过度拟合可以通过对多个随机丢弃激活模型的预测结果进行平均来缓解[8],或者通过引入长短期记忆(long short-term memory, LSTM)层以降低全局图像编码的维度[9]。关于损失函数,Kendall 等[12]建议优化参数以平衡不同损失,从而提高预测精度,避免依赖人工微调,该公式已被众多位姿回归器所采用。尽管已取得了一定的进展,但在具有相似视觉特征和光照变化的真实场景中,相机定位的精度仍然较低。

在此背景下,Transformer 模型凭借其独特的多头自注意力机制,能够有效捕捉图像中的全局信息[13],并在许多视觉任务中展现出良好的性能。Shavit 等[14]首次将 Transformer 应用于相机绝对位姿估计,成功解决了多场景下的位姿回归问题[15],显示出其在处理复杂场景时的潜力。然而,尽管 Transformer 具有较强的特征学习能力,它仍然对噪声较为敏感,可能导致异常值的出现。为了解决这一问题,Song 等[16]提出引入洗牌注意力机制(shuffle attention, SA),通过集成空间和通道维度的特征信息来增强模型对噪声和光照变化的鲁棒性,同时减少动态物体的影响,从而提升相机定位精度。

尽管多头自注意力机制在处理图像序列时展现了强大的并行处理能力和对空间域信息敏锐的捕捉能力。但图像作为一种复杂信号,包含丰富的频率成分,这些频率信息不仅反映了局部细节(如边缘和纹理等高频信息),还揭示了全局结构(如形状和布局等低频信息)。然而,目前的研究主要集中于空域信息,往往忽略了频域信息的重要性。

基于以上讨论,针对相机定位问题,提出了一种混合频域 Transformer 方法。该方法通过快速傅里叶变换将空间域转换为频域信息,并结合频域注意力模块和 Transformer 多头注意力机制,设计了一种分层结构的编码器,以有效捕捉图像中的高频和低频信息。此外,为了分别回归相机的位置和方向,采用了两个独立的编解码器,以便更有效地处理特征并提升定位精度。为验证所提方法的广泛适用性与实际应用价值,构建了一个专为室内移动机器人定位设计的相机位姿估计数据集。该数据集包含由不同速度的移动机器人搭载的相机所采集的图像,这些图像具有较多相似的视觉特征和弱纹理,并附有从 VICON 系统获取的相机位姿真值。数据集旨在提供丰富的训练和测试样本,以支持算法性能的全面评估,从而确保其在实际应用中的有效性和可靠性。

1 问题描述及方法架构

1.1 相机位姿估计问题

相机成像过程包括世界坐标系、相机坐标系、图像坐标系之间的转换,可将三维世界中点的映射到二维图像坐标系。其中世界坐标系 $\{\boldsymbol{W}\}$ 是现实世界中的绝对坐标系,用于描述相机和场景中物体的位置;相机坐标系 $\{\boldsymbol{C}\}$ 以相机光心为原点,相机光轴为 Z 轴;图像坐标系 $\{\boldsymbol{I}\}$ 通常取主光轴与图像平面的交点为原点。相机坐标系与图像坐标系之间的转换可根据相似三角形可建立等比关系:

$$\frac{X_C}{x_I}=\frac{Y_C}{y_I}=\frac{Z_C}{f} \tag{1}$$

相机坐标系与世界坐标系之间的转换可以看作是相机在世界坐标系中的刚体运动,可以用一个旋转矩阵 $\boldsymbol{R}$ 和平移向量 $\boldsymbol{t}$ 表示为 $P_C=\boldsymbol{R}P_W+\boldsymbol{t}$,上述过程中 $p_I(x_I,y_I)$ 表示图像坐标系中 f 表示焦距,$P_C(X_C,Y_C,Z_C)$ 和 $P_W(X_W,Y_W,Z_W)$ 分别表示相机坐标系中的点与世界坐标系中的点。相机坐标系与世界坐标系之间的关系可表示为 ${}^W_C\boldsymbol{T}=[\boldsymbol{R},\boldsymbol{t}]$,则相机位姿的参数化描述为 $\boldsymbol{T}=\{\boldsymbol{X},\boldsymbol{Q}\}$,$\boldsymbol{X}\in\boldsymbol{R}^3$ 表示相机位置,$\boldsymbol{Q}\in\boldsymbol{R}^4$ 表示相机方向。

CamPose 直接从单幅图像中回归相机位姿,则相机位姿估计问题可表示为 $\hat{\boldsymbol{T}}=\pi(p,\omega)$,$\pi$ 表示相机位姿估

计模型,ω 是模型的可训练参数。用 $\boldsymbol{T}_g$ 表示相机位姿真值,为了训练 ω, 约束函数为:

$$\min_{\omega} L(\pi(p,\omega),\boldsymbol{T}_g) \tag{2}$$

1.2　整体方法与架构

本文提出的相机位姿估计方法以位于移动机器人顶端的相机所采集的室内场景 RGB 图像为基础,通过设计 CamPose 模型,构建场景信息与相机位姿的映射关系,估计出世界坐标系中相机的位置和方向,最后通过坐标系转换,求解移动机器人的位姿,实现移动机器人的室内定位。

整体方法架构如图 1 所示,可划分为以下 3 个阶段:1)数据采集与预处理;2)模型构建与训练;3)模型加载与部署。具体流程如下:首先,将相机采集的场景图像序列与 VICON 系统捕捉的相机刚体位姿序列进行时间戳对齐,构建包含有 RGB 图像与相机位姿真值标签的 RotIndoor 数据集;其次,提出相机位姿估计模型,并使用 RotIndoor 和公开数据集进行训练;最终,通过加载已训练的相机位姿估计模型,实现相机定位和移动机器人定位,并计算分析定位误差。

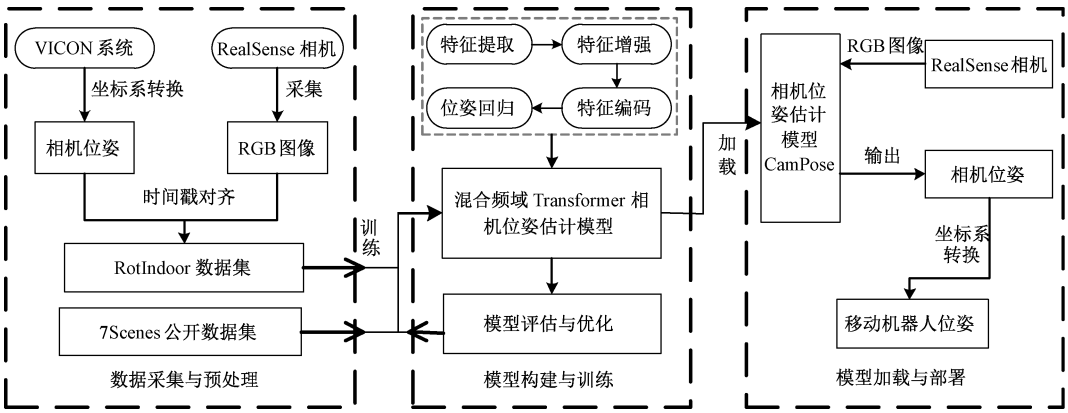

图 1　整体方法架构

Fig. 1　Architecture of the proposed method

2　相机位姿数据集构建

本文构建的 RotIndoor 数据集的室内采集场景如图 2 所示,室内布局包括数据采集区、展示柜、书柜、桌椅等,硬件设备包含已部署在室内上方的 VICON 系统和位于地面上的移动采集设备。

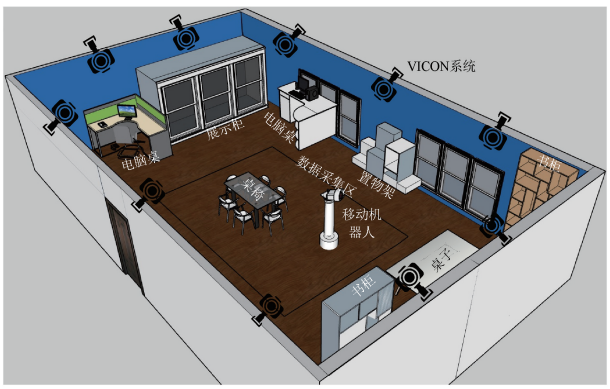

图 2　RotIndoor 数据集采集场景

Fig. 2　Acquisition scene of RotIndoor dataset

具体介绍如下:

1)VICON 系统:一种高性能的光学动作捕捉系统。利用主动式红外光学标记跟踪技术能够实时捕捉粘贴在刚体上的反光球(Marker)的运动轨迹[17],并通过控制软件计算出刚体的空间位置与姿态信息,其定位精度达到了毫米级别。相较于现有的关于相机位姿估计的公开数据集中位姿标签是由三维重建或 SFM 方法生成的"伪位姿标签",该系统为本文构建的数据集提供了更高精度的真实位姿标签。

2)移动采集设备:主体是轮式移动机器人,可通过控制手柄操控其在室内前进后退、旋转,并在顶部安装了相机刚体,由相机支架、反光球、RealSense 相机组成。其中相机支架是一个 3D 打印的多面体,扩大了反光球的可粘贴面积,用于解决相机表面有限,无法满足反光球互不遮挡且尽量异面的布局要求;反光球用于在 VICON 系统中构建可跟踪的刚体,获取相机真实位姿标签;RealSense 相机用于获取室内场景图像,作为模型的输入。

数据采集过程主要包含以下 3 部分:

1)系统配置与校准:

(1)使用控制软件和标定杆校准每个摄像头确保其

正常工作,并初始化 VICON 系统基坐标 $\{\boldsymbol{V}\}$ 为世界坐标系 $\{\boldsymbol{W}\}$;

(2)使用棋盘格标定板对相机内参进行标定;通过手眼标定法求解相机刚体坐标系 $\{\boldsymbol{B}\}$ 与相机坐标系 $\{\boldsymbol{C}\}$ 的转换关系 ${}^{\boldsymbol{B}}_{\boldsymbol{C}}\boldsymbol{T}$;

(3)设置 VICON 系统的数据采集频率为 100 Hz,相机的彩色分辨率设置为 1 080×720,采集帧率设为 15 fps。

2)数据采集:

(1)启动 VICON 数据传输节点、启动相机采集程序,通过 NTP 协议保证各个设备处于同一个时间基准下;

(2)操控手柄控制机器人在数据采集区内移动或原地旋转;通过控制机器人移动速度与移动范围采集多个序列,设置对照组;

(3)将 VICON 系统中传输的相机刚体位姿数据 ${}^{\boldsymbol{W}}_{\boldsymbol{B}}\boldsymbol{T}_t$ 和相机采集的图像数据 $\boldsymbol{I}_t=\{t,p_t\}$ 保存在本地计算机,${}^{\boldsymbol{W}}_{\boldsymbol{B}}\boldsymbol{T}_t$ 可表示为 ${}^{\boldsymbol{W}}_{\boldsymbol{B}}\boldsymbol{T}_t=\{t,\boldsymbol{X}_t,\boldsymbol{Q}_t\}$,$t$ 表示时间戳,$\boldsymbol{X}_t=x_t,y_t,z_t$ 表示 t 时刻相机刚体位姿的位置分量,$\boldsymbol{Q}_t=w_t,q_{1_t},q_{2_t},q_{3_t}$ 表示方向分量,用四元数表示。

3)数据处理:

(1)以采样频率较低的相机传感器中的时间戳数据为同步基准,随后根据最邻近时间戳 $\min\Delta t$ 对齐位姿数据与图像数据;

$$\min\Delta t=\min\{t_I-t_{\boldsymbol{W}_{\boldsymbol{B}}\boldsymbol{T}}\} \tag{3}$$

(2)根据采集过程 1)~(2)中已知的相机与相机刚体之间的转换关系计算相机位姿真值标签 ${}^{\boldsymbol{W}}_{\boldsymbol{C}}\boldsymbol{T}^t={}^{\boldsymbol{W}}_{\boldsymbol{B}}\boldsymbol{T}^t\times{}^{\boldsymbol{B}}_{\boldsymbol{C}}\boldsymbol{T}^t$;

(3)数据对齐后需要进一步筛选数据集以排除异常数值,确保数据集质量满足实验需求。异常数据具有明显的离群效应,因此可使用差分统计方法寻找异常值。相机在机器人上的安装高度不变,故 z 应为定值,但 VICON 存在系统误差,最终获得的位姿数据也存在微小误差,所以可以使用 z 为基准值计算时间序列上相邻的数据差分 Δz_t 并设置它的平均值 $\overline{\Delta z_t}$ 为异常值阈值,即可识别异常值 $\{\Delta z_t>\overline{\Delta z_t}\}$,随后将包含异常值的位姿数据和对齐的图像数据删除。

$$\Delta z_t=z_t-z_{t-1},\overline{\Delta z_t}=\frac{1}{n}\times\sum_{i=1}^{n-1}\Delta z_{t_i} \tag{4}$$

最后得到的数据集中包含已对齐的相机位姿标签 $\{x,y,z,w,q_1,q_2,q_3\}$ 和场景图像 $\{p\}$ 两部分数据。

3 相机位姿模型构建

CamPose 通过输入图像的前向传递来实现位姿预测,网络设计结构如图 3 所示。其核心构成包括骨干网络、特征增强模块、混合频域编解码器模块以及回归器。骨干网络通过多层卷积有效提取图像的高维特征。随后,特征增强模块利用差分卷积网络进一步细化和捕获细节特征。在此基础上,混合频域编码器对位置信息和方向信息进行独立的自适应聚合。解码器模块借鉴视觉 Transformer(vision Transformer, ViT)[13] 中的多层结构,对编码器输出进行解码。回归器则包括一个 Softmax 函数用于场景分类,以及一个多层感知器(multi-layer perceptron, MLP)将编码器输出映射到相机位姿。整体网络体系的构建受 MS-Trans(multi-scene absolute pose regression with transformers)[14] 架构启发,旨在提升相机位姿估计的精度、鲁棒性和泛化能力。值得注意的是,本文还对输入图像进行了预处理,具体随机裁剪参数为 224,随机改变图像亮度、对比度、饱和度和色调的参数为 0.5、0.5、0.5 和 0.2。

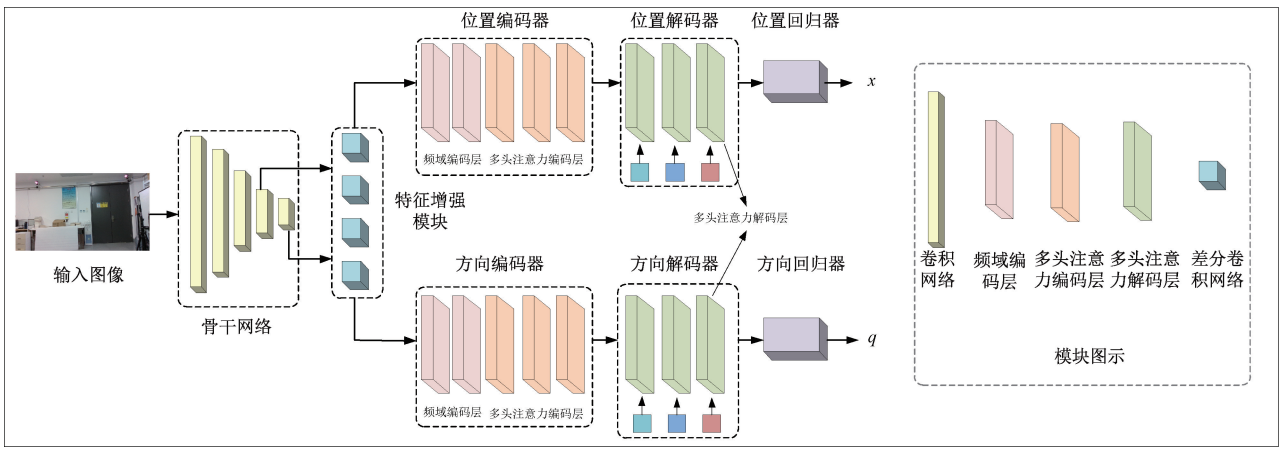

图 3 网络设计结构

Fig. 3 Network architecture design

3.1　骨干网络与特征增强模块

选取具有良好性能的残差网络 ResNet34[18] 作为骨干网络,并使用在大型图像分类数据集 ImageNet 上预训练的权重来初始化 ResNet34 网络。同时为了更好的适应本文研究需求,对 ResNet34 网络进行了调整,移除了全连接层和用于分类的 Softmax 层,选取了不同残差块输出特征图作为后续位置编解码器与方向编解码器的输入。即当输入图像为 $p_i \in \boldsymbol{R}^{3\times H_o\times W_o}$,骨干网络的输出为 $F_{X_i} \in \boldsymbol{R}^{C_x\times H_x\times W_x}$ 和 $F_{Q_i} \in \boldsymbol{R}^{C_q\times H_q\times W_q}$。

位姿的微小变化通常反映在图像的局部细节和边缘信息中,因此本文在骨干网络后设计了特征增强模块,以提高模型对图像细节和边缘的敏感度,从而增强特征表征能力。该模块并行部署了 4 个差分卷积,分别为中心差分卷积(central difference convolution, CDC)、角度差分卷积(angular difference convolution, ADC)、水平差分卷积(horizontal difference convolution, HDC)和垂直差分卷积(vertical difference convolution, VDC)[19],通过调整不同维度卷积核的权重实现细粒度特征提取。同时利用卷积层的可加性实现重参数化技术,将 4 个卷积层简化为单个标准卷积层,以解决 4 个差分卷积并行部署所导致的参数冗余和推理时间延长的问题,简化过程可表达如下:

$$E_{X_i} = \sum_{i=1}^{4} F_{X_i} * K_i = F_{X_i} * \left(\sum_{i=1}^{4} K_i\right) = F_{X_i} * K_{cvt} \tag{5}$$

其中,E_{X_i} 表示特征增强模块的输出,$K_{i=1:4}$ 分别表示 CDC、ADC、HDC 和 VDC 的卷积核,$*$ 代表卷积操作,K_{cvt} 表示转换后的卷积核。

3.2　混合频域编码

鉴于传统多头注意力(multi-head attention, MHA)机制在处理图像时普遍采用统一的全局关注模式,忽略了图像中频率特性的差异性,故提出了一种基于分层结构的混合频域编码器。该编码器融合了频域编码层与传统 Transformer 编码层的优势,既实现了对不同频率的针对性学习,又保留了 Transformer 编码层在全局信息整合与上下文建模方面的优势。

频域编码层的设计如图 4 所示,包括层归一化、快速傅里叶变换(fast Fourier transform, FFT)、频域注意力模块、逆傅里叶变换(inverse fast Fourier transform, IFFT)以及 Dropout 层。

首先,通过 FFT 将图像的空间域信息转换至频域,以获取图像的频率分量。接着,通过频域注意力模块[20]实现对图像频率的加权学习,具体采用两条路径:高频注意力模块通过局部窗口自注意力提取细粒度的高频信息,低频注意力模块通过平均池化获取低频信息。随后,IFFT 将处理后的频域信息逆变换回空间域,以恢复图像特征的空间表达。层归一化与 Dropout 层的引入有效提升了模型的泛化能力,降低了过拟合风险,确保了编码器在复杂图像数据上的稳健表现。频域编码层可表示为:

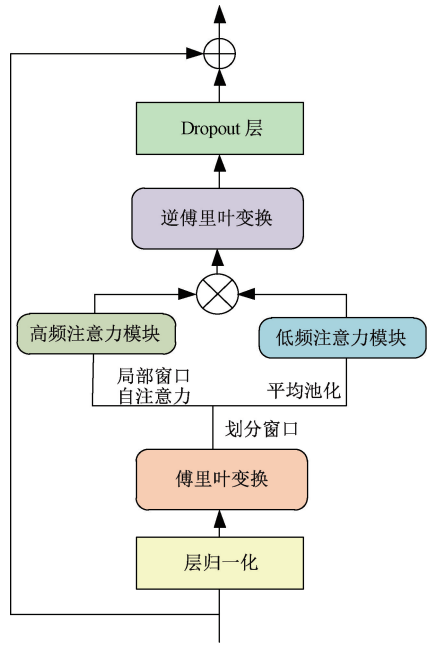

图 4　频域编码层结构

Fig. 4　Structure of frequency domain coding layer

$$S_{X_i} = FFT(LN(E_{X_i})) \tag{6}$$

$$SA_{X_i} = [HiA(S_{X_i}); LoA(S_{X_i})] \tag{7}$$

$$Z_{X_i} = \mathrm{Drop}(IFFT(SA_{X_i})) \tag{8}$$

其中,S_{X_i} 表示经过傅里叶变换后的频域信息,SA_{X_i} 表示频域注意力模块,HiA 表示高频注意力模块,LoA 表示低频注意力模块,Z_{X_i} 表示逆傅里叶变换后的空间域信息,LN 表示层归一化,Drop 表示 Droput 层。混合频域编码器由 α 层频域编码层和 L 层由 MHA 和 MLP 组成的 Transformer 编码层构成。Transformer 编码层可表示为:

$$Z'^{l}_{X_i} = MHA(LN(Z^{l-1}_{X_i})) + Z^{l-1}_{X_i} \tag{9}$$

$$Z^{l}_{X_i} = MLP(LNZ^{l}_{X_i}) + Z^{l}_{X_i} \tag{10}$$

$$Z^{L}_{X_i} = LN(Z^{L}_{X_i}) \tag{11}$$

3.3　损失函数

根据公共数据集和 RotIndoor 的差异性,本文采用了 2 种不同的损失函数来训练模型。其中,针对公共数据集,使用了较为典型的位置损失函数 Loss_X 和方向损失函数 Loss_Q,表达式如下:

$$\mathrm{Loss}_X = \| \boldsymbol{X}_{i=x,y,z} - \widehat{\boldsymbol{X}}_i \|_2 \tag{12}$$

$$\mathrm{Loss}_Q = \left\| \boldsymbol{q} - \frac{\hat{\boldsymbol{q}}}{\| \hat{\boldsymbol{q}} \|} \right\|_2 \tag{13}$$

$$\mathrm{Loss}_p = \mathrm{Loss}_X \exp(-s_X) + s_X + \mathrm{Loss}_Q \exp(-s_q) + s_q \tag{14}$$

其中，$\{\boldsymbol{X},\boldsymbol{Q}\}$，$\boldsymbol{X}\in\boldsymbol{R}^3$，$\boldsymbol{Q}\in\boldsymbol{R}^4$ 表示相机位姿真值，$\boldsymbol{q}$ 表示单位化后的四元数。然后使用 Kendall 等[12]提出的相机位姿损失函数将两个损失函数结合起来并表示为 Loss_p，s_X 和 s_q 表示控制两个损失之间平衡的超参数。回归 N 个场景中的相机位姿，还需对场景图像进行分类，故添加负对数似然函数（NLL），其参数 s 表示模型对场景预测的对数概率，s_0 表示场景真值标签，最终损失函数表示为：

$$\mathrm{Loss}_T=\mathrm{Loss}_p+NLL(s,s_0) \tag{15}$$

本文所采集的数据因设备特性，即相机高度不变且旋转方向单一，避免了万向锁问题，故可使用欧拉角 θ 更直观地表示相机方向。这种表示方式不仅使方向损失函数 Loss_θ 的定义更加明确，还减少了模型的输出维度，提高了计算效率。具体损失函数表示为：

$$\mathrm{Loss}_X=\|\boldsymbol{X}_{i=x,y}-\widehat{\boldsymbol{X}}_i\|_2,\mathrm{Loss}_\theta=\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|_2 \tag{16}$$

$$\mathrm{Loss}_T=\mathrm{Loss}_X\exp(-s_X)+s_X+\mathrm{Loss}_{\boldsymbol{\theta}}\exp(-s_q)+s_q \tag{17}$$

为了提高模型在大场景中的位姿预测精度，假设所采集的数据集中仅包含一个场景，不再细分为多个小场景。因此，场景分类损失函数不再需要，从而进一步减少了模型的计算量。这一设计旨在优化模型的效率，使其更专注于提高位姿估计的准确性

4 模型训练评估与实验

4.1 实验平台与数据集

采用两个室内数据集进行模型训练和测试：

1）7Scenes：该数据集是当前相机位姿估计任务中应用最广泛的公开室内数据集，包含由手持 Kinect RGB-D 相机拍摄的 7 个小规模室内场景，空间范围约为几平方米。每个场景包含 2~7 个序列，每个序列包含 500~1 000 fps 图像，分辨率为 640 pixels×480 pixels。

2）RotIndoor：该数据集由搭载 RealSense D435i 相机的机器人在室内移动过程中采集。包含 6 个序列的样本数据，每个样本包含分辨率为 1 280×720 的 RGB 图像和相机位姿标签，总计 16 200。数据集按照 8∶2的比例随机划分为训练集和测试集。其中，序列 1、2 和 3 分别以机器人移动速度的快、中、慢作为对照组，序列 4 和 5 则考虑了场景中物体位置的变化，而序列 6 则包含大量弱纹理图像。不同序列及对照组的样本数量详见表 1。

实验平台：本文所有实验均在搭载 Intel® Core™ i7-7700 CPU @ 3.60GHzX8，16GRAM 和 8 GB 显存的 NVIDIA GeForce RTX2070 GPU，装载 Ubuntu18.04 LTS 系统的计算机上进行。训练过程中，本文使用 Adam 模型最小化式（18）中的损失函数，其中 $\beta_1=0.9$，$\beta_2=0.999$，$\varepsilon=10^{-10}$。在所有实验过程中，批大小为 8，初始学习率为 $\lambda=10^{-4}$。

表 1 RotIndoor 数据集

Table 1 RotIndoor dataset

序列编号	机器人移动速度	物体位置是否发生变化	弱纹理图像数量	训练集	测试集
序列 1	快	否	少	2 305	576
序列 2	中	否	少	2 500	625
序列 3	慢	否	少	2 378	594
序列 4	慢	是	少	1 691	422
序列 5	慢	否	少	2 444	610
序列 6	慢	否	多	1 644	411
总计	-	-	-	12 962	3 238

4.2 相机位姿估计模型评估与分析

本文采用绝对位姿误差（absolute pose error，APE）评估 CamPose 在六自由度公共数据集上的估计位姿与真实位姿之间的差值，涵盖绝对位置误差 ΔX 和绝对方向误差 $\Delta\theta$ 两部分。

该评估指标衡量了模型在整个运动轨迹上位姿估计的准确性，直观反映其在全局尺度上的精度和一致性。绝对位置误差表示预测位置 $\widehat{\boldsymbol{X}}$ 与真实位置 $\boldsymbol{X}$ 的欧氏距离，单位为 m。

$$\Delta X=\|\boldsymbol{X}-\widehat{\boldsymbol{X}}\|_2 \tag{18}$$

绝对方向误差可以用旋转角度表示，单位为°，当使用四元数表示相机方向时，可表示为：

$$\Delta\theta=2\arccos|\boldsymbol{q}\,\hat{\boldsymbol{q}}|\frac{180}{\pi} \tag{19}$$

当使用欧拉角表示相机方向时，可表示为：

$$\Delta\theta=\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|_2 \tag{20}$$

为全面评估 CamPose 的性能，进行了详尽的对比分析，将其与当前基于单幅图像、序列处理和多场景分析的相机位姿估计算法进行了系统对比。所引结果均来自于被引用文献的原文，具体量化结果见表 2。分析结果显示，本文方法在多数测试序列中表现优异，尤其在处理具有重复性结构（如楼梯场景）及弱纹理区域（如火焰、人群头部等）的复杂环境中，其优势尤为突出。这主要归功于集成的特征增强模块，该模块通过多重差分卷积技术显著提升了对细微特征的捕获与解析能力。此外，CamPose 在平均位置误差这一关键指标上超越了对比方法，除了与特征增强模块有关，还得益于频域编码层中频域变换与注意力模块的协同作用。

表 2　在 7Scenes 数据集上的方法比较

Table 2　Comparison of methods on the 7Scenes dataset　(m/°)

参考文献	方法	国际象棋	火焰	人群头部	办公楼	南瓜	厨房	楼梯	平均
[7]	PoseNet	0.32/8.12	0.47/14.40	0.29/12.00	0.48/7.68	0.47/8.42	0.59/8.64	0.47/13.80	0.45/9.94
[12]	PN-Learnable	0.14/4.50	0.27/11.80	0.18/12.10	0.20/5.77	0.25/4.82	0.24/5.52	0.37/10.60	0.24/7.87
[9]	LSTM-PN	0.24/5.77	0.34/11.90	0.21/13.70	0.30/8.08	0.33/7.00	0.37/8.83	0.40/13.70	0.31/9.85
[11]	BayesianPN	0.37/7.24	0.43/13.70	0.31/12.00	0.48/8.04	0.61/7.08	0.58/7.54	0.48/13.10	0.47/9.81
[10]	GPoseNet	0.20/7.11	0.38/12.30	0.21/13.80	0.28/8.83	0.37/6.94	0.35/8.15	0.37/12.50	0.31/9.95
[21]	MLFBPPose	0.12/5.82	0.26/12.00	0.14/13.54	0.18/824	0.21/7.05	0.22/8.14	0.38/10.26	0.22/9.29
[22]	ViPR	0.22/7.89	0.38/12.74	0.21/16.41	0.35/9.59	0.37/8.45	0.40/9.32	0.31/12.65	0.32/11.01
[8]	IRPNet	0.13/5.64	0.25/9.670	0.15/13.10	0.24/6.33	0.22/5.78	0.30/7.29	0.34/11.60	0.23/8.49
[16]	TransBoNet	0.11/4.48	0.25/12.46	0.18/14.00	0.20/5.08	0.19/4.77	0.17/5.35	0.30/13.04	0.20/8.45
[15]	MSPN	0.09/4.76	0.29/10.50	0.16/13.10	0.16/6.80	0.19/5.50	0.21/6.61	0.31/11.60	0.20/7.56
	本文	0.10/5.78	0.25/10.12	0.12/11.17	0.16/6.37	0.18/5.81	0.16/6.85	0.24/8.90	0.17/7.85

本文还在 RotIndoor 数据集上应用并训练了 2 种开源且具有代表性的算法,即单幅图像定位算法与多场景定位算法。鉴于原算法所依赖的数据集结构与 RotIndoor 数据集存在差异,针对性地调整了这 2 种算法的数据读取模块,以确保其适配 RotIndoor 数据集。表 3 详细汇总了这 2 种调整后的算法与 CamPose 在 RotIndoor 数据集上的定量对比结果,显示出本文方法在各测试序列中的优越性。此外,为了深入探究方向表示策略及损失函数的选取对性能的影响,还对基于四元数与基于欧拉角的损失函数进行了定量比较,结果表明,在构建的数据集上,采用欧拉角方法相较于四元数方法具有更高的预测精度,验证了其在特定数据集上的优越性。图 5 则直观展示了本文方法在多个测试序列中的预测轨迹与真实轨迹的对比(圆圈表示机器人原地旋转),进一步支持了前述量化分析结果的准确性。

表 3　在 RotIndoor 数据集上的方法比较

Table 3　Comparison of methods on the RotIndoor dataset　(m/°)

参考文献	方法	序列 1	序列 2	序列 3	序列 4	序列 5	序列 6	平均
[7]	PoseNet	0.102/5.34	0.121/5.44	0.133/5.63	0.110/5.53	0.103/5.44	0.121/5.41	0.115/5.47
[14]	MS-Trans	0.039/3.14	0.048/3.38	0.050/3.56	0.037/3.19	0.036/3.68	0.042/3.13	0.042/3.35
	四元数	0.029/2.79	0.040/3.18	0.043/3.81	0.033/3.46	0.027/3.12	0.033/3.23	0.034/3.26
	欧拉角	0.029/2.10	0.038/2.20	0.041/2.29	0.028/1.97	0.024/2.14	0.031/1.82	0.032/2.09

4.3　消融实验

本节中设计了一系列消融实验,以验证网络设计的合理性与有效性。首先,针对网络模型结果,设计了多个模型结构对比实验;其次,针对损失函数中的调控因子 s_x 和 s_q, 设计了不同参数组合,并在 RotIndoor 数据集上进行实验;最后,针对编码部分,设计了不同层数组合,通过实验选取最佳组合。

在 RotIndoor 数据集上,比较了 CamPose 与传统编解码结构的网络模型、去特征增强模块的网络模型,以及多场景相机位姿估计领域标杆 MS-Trans 网络模型。实验结果如表 4 所示,通过引入特征增强模块与混合频域编解码器,CamPose 在相机位姿估计精度上有明显提升。特征增强模块对位置预测精度的贡献尤为明显,同时混合频域编解码器不仅进一步增强了位置预测的精度,还显著提高了方向预测的准确性。

表 4　在 RotIndoor 数据集上的比较不同模型结构的表现

Table 4　Comparing the performance of different model architectures on the RotIndoor dataset　(m/°)

模型结构	平均
MS-Trans	0.063/3.74
骨干网络+编解码器+回归模块	0.053/3.58
骨干网络+特征增强+编解码器+回归模块	0.046/3.12
骨干网络+特征增强+混合频域编解码器+回归模块(本文)	0.032/2.09

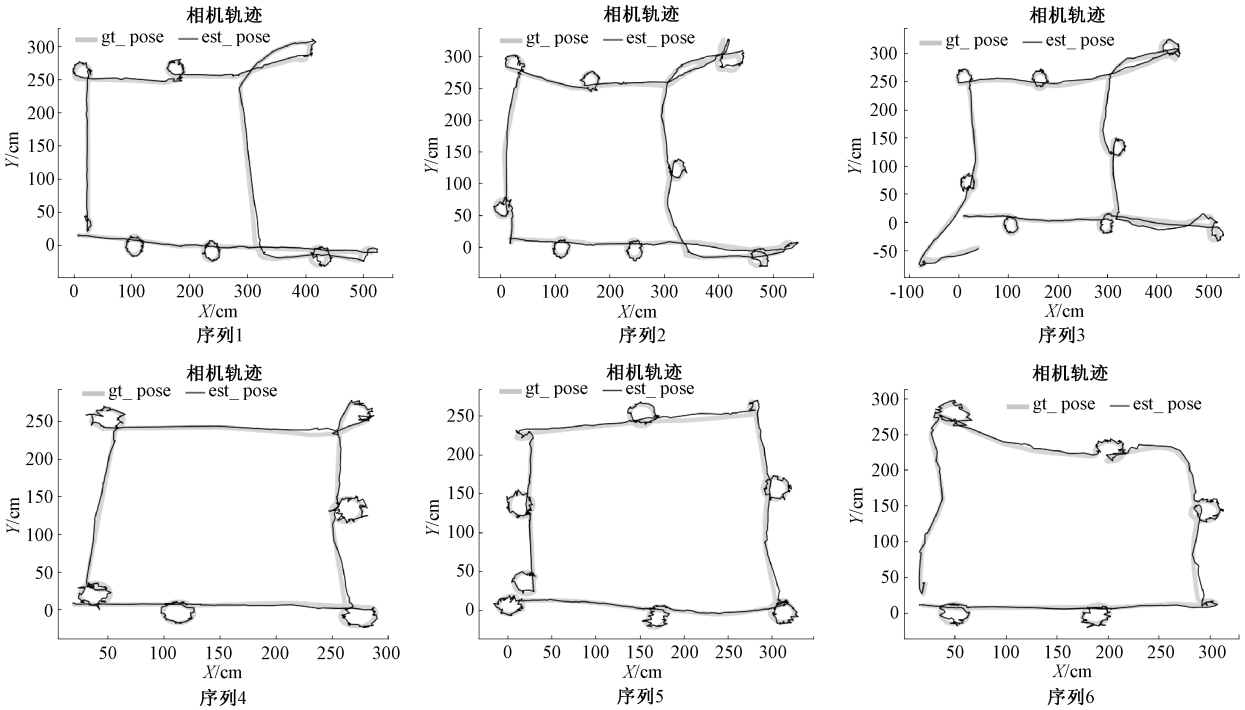

图 5 RotIndoor 数据集序列轨迹

Fig. 5 Sequence trajectory visualizations from the RotIndoor dataset

s_X 与 s_q 作为调控因子,旨在均衡位置误差与姿态误差在模型训练中的影响权重。采用四元数表示相机方向时,已有研究表明 $s_X = 0.0$、$s_q = -3.0$ 时能有效实现两者间的平衡。然而,鉴于本文采用欧拉角表示相机方向,这一平衡状态需重新评估,因此,进行了不同参数设置的灵敏度分析,以探索模型性能的潜在变化。通过在 RotIndoor 数据集上使用不同的 s_X 与 s_q 组合进行训练和验证,汇总位姿估计误差于表 5 中,结果显示当 $s_X = -5.0$、$s_q = 1.0$ 时,位姿预测误差最小,从而验证了该参数组合与模型的适配性。

表 5 在 RotIndoor 数据集上的比较不同 s_X 与 s_q 和的值

Table 5 Comparing the performance of different s_X and s_q on the RotIndoor dataset (m/°)

参数值	序列 1	序列 2	序列 3	序列 4	序列 5	序列 6	平均
$s_X = -2.0, s_q = 2.0$	0.028/2.50	0.038/2.62	0.038/3.07	0.038/3.07	0.025/2.92	0.032/2.81	0.032/2.83
$s_X = -3.0, s_q = 0.0$	0.032/2.73	0.042/3.17	0.042/3.45	0.031/3.11	0.027/3.24	0.035/3.10	0.035/3.13
$s_X = -4.0, s_q = 0.0$	0.031/2.34	0.040/2.70	0.043/2.76	0.031/2.53	0.027/2.76	0.032/2.73	0.034/2.64
$s_X = -5.0, s_q = 1.0$	0.029/2.10	0.038/2.20	0.041/2.29	0.028/1.97	0.024/2.14	0.031/1.82	0.032/2.09
$s_X = -6.0, s_q = 1.0$	0.029/2.54	0.039/2.70	0.041/3.11	0.028/2.56	0.024/2.50	0.031/2.70	0.032/2.69

针对编解码部分,实验探索了 Transformer 编码层数与频域编码层数的配置,以及这些配置对解码器层数的影响,以优化模型性能。系统地对比分析了多种层数组合,重点关注两个关键维度:位置与方向编解码层数的一致性,以及总体编码器与解码器层数的平衡关系。实验结果如表 6 所示,当位置编解码层数(t)与方向编码层数(rot)相等,且 Transformer 编码层数设定为 4 层、频域编码层数为 2 层、解码器层数为 6 层时,模型预测误差达到最小值。这一结果为模型架构的优化提供了坚实的实验依据。

4.4 真实场景实验

为了验证 CamPose 的准确性和有效性,将经过训练的 CamPose 部署到实际的计算平台中,以支持移动机器人系统的应用并进行实时位姿估计。该过程主要包括模型部署和实时定位两个阶段。首先,实验环境与图 2 中的 RotIndoor 数据集采集环境保持一致,确保光照条件和空间布局与日常室内场景相符,未作特殊处理,以反映真实环境中的自然条件。随后,将训练好的模型加载到计算平台并设计实施一个实时处理管道,使计算平台能够持续获取相机拍摄的 RGB 图像,并将其传递至位姿估计

表 6　在 RotIndoor 数据集上的比较不同的编解码器层数组合

Table 6　Comparison of different combinations of encoder-decoder layer counts on the RotIndoor dataset　(m/°)

编码器		解码器	均值
Transformer 编码层	频域编码层		
$t=4, rot=4$	$t=1, rot=1$	$t=4, rot=4$	0.040/2.63
$t=4, rot=6$	$t=1, rot=1$	$t=4, rot=6$	0.039/2.33
$t=6, rot=4$	$t=1, rot=1$	$t=6, rot=4$	0.035/2.42
$t=6, rot=6$	$t=1, rot=1$	$t=6, rot=6$	0.032/2.23
$t=5, rot=3$	$t=1, rot=1$	$t=6, rot=4$	0.031/2.44
$t=5, rot=3$	$t=1, rot=1$	$t=6, rot=6$	0.035/2.40
$t=4, rot=4$	$t=2, rot=2$	$t=4, rot=4$	0.040/2.63
$t=4, rot=6$	$t=2, rot=2$	$t=4, rot=6$	0.039/2.32
$t=6, rot=4$	$t=2, rot=2$	$t=6, rot=4$	0.035/2.42
$t=6, rot=6$	$t=2, rot=2$	$t=6, rot=6$	0.034/2.24
$t=4, rot=2$	$t=2, rot=2$	$t=6, rot=4$	0.034/2.37
$t=4, rot=4$	$t=2, rot=2$	$t=6, rot=6$	0.032/2.09

模型进行处理。实时定位过程中,首先控制机器人在实验环境中以设定速度自由移动,同时启动相机持续捕捉 RGB 图像,实验设备见图 6。随后,对捕获的图像进行实时处理,调用已部署的模型,生成相机在世界坐标系中的位置和方向信息,并保存实验数据。整个实验共进行了 4 次,确保结果的准确性和重复性。

在上述实验中为了更直观的展示 CamPose 模型预测位姿的有效性,依旧使用了 VICON 系统获取位姿真值,

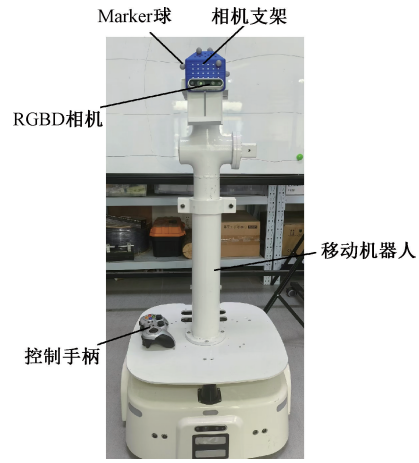

图 6　实验设备

Fig. 6　Experimental equipment

与预测位姿进行对比。在实际应用或在不具备 VICON 系统的室内环境中,可直接由已训练好的模型获取预估相机位姿。实时估计的位姿数据和通过 VICON 系统获取的真值数据的可视化结果如图 7 所示。从图中可以看出,二者的轨迹图重合度较高,实时估计的轨迹与真值轨迹在多个关键点上高度一致,验证了所提出模型的有效性与鲁棒性。此外,预测结果的平滑性与连续性表明模型在室内环境中的表现稳定。模型的平均推理时间为 33 ms,充分满足实时性要求,这一性能得益于采用了 TensorRT 优化技术。整体而言,这些结果证明了 CamPose 模型在实际应用中的优越性能,为后续研究和应用提供了坚实的基础。

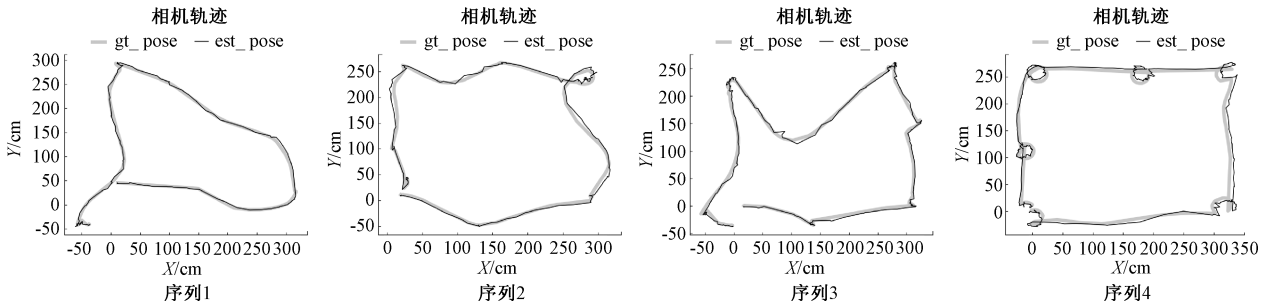

图 7　场景实验轨迹对比

Fig. 7　Comparison graph of real experimental trajectories

5　结　　论

本文提出了基于混合频域 Transformer 的相机位姿估计方法,该方法能够直接从 RGB 图像中恢复相机的位置和姿态信息。通过在广泛认可的 7Scenes 数据集的实验,展示了该模型在多数测试序列中均能实现较低的位姿估计误差,从而验证了其有效性与鲁棒性。本文还进一步将该方法应用于室内移动机器人的定位任务中,实验结果显示模型在复杂室内环境下同样具备一定精度的定位能力,这充分证明了该方法在实际应用中的潜力和价值。尽管本研究取得了令人鼓舞的成果,但仍存在改进空间。具体而言,在 7Scenes 数据集的某些特定序列上,模型的绝对方向误差相对较大,这表明在特定场景或光

照条件下,模型的预测精度有待进一步提升。未来工作可聚焦于优化模型架构,引入更先进的注意力机制或融合多模态信息以进一步提高模型的泛化能力和预测精度。

参考文献

[1] 杨傲雷, 金宏宙, 陈灵, 等. 融合深度学习与粒子滤波的移动机器人重定位方法[J]. 仪器仪表学报, 2021,42(7):226-233.
YANG AO L, JIN H ZH, CHEN L, et al. Mobile robot relocalization method fusing deep learning and particle filtering[J]. Chinese Journal of Scientific Instrument, 2021, 42(7):226-233.

[2] 焦传佳, 江明. 基于 AprilTag 图像识别的移动机器人定位研究[J]. 电子测量与仪器学报, 2021, 35(1): 110-119.
JIAO CH J, JIANG M. Research on positioning of mobile robot based on low complexity AprilTag image recognition[J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(1): 110-119.

[3] 吴亚辉, 刘春阳, 谢赛宝, 等. 基于视觉深度学习的机器人环境感知及自主避障[J]. 电子测量技术, 2021, 44 (20): 99-106.
WU Y H, LIU CH Y, XIE S B, et al. Mobile robotic perception and autonomous avoidance based on visual depth learning[J]. Electronic Measurement Technology, 2021, 44 (20): 99-106.

[4] 余洪山, 郭丰, 郭林峰, 等. 融合改进 SuperPoint 网络的鲁棒单目视觉惯性 SLAM[J]. 仪器仪表学报, 2021, 42 (1): 116-126.
YU H SH, GUO F, GUO L F, et al. Robust monocular visual-inertial SLAM based on the improved SuperPoint network[J]. Chinese Journal of Scientific and Instrument, 2021, 42 (1): 116-126.

[5] 史涛, 校诺政, 丁垚, 等. 动态场景下融合改进 YOLOv7 的视觉 SLAM 算法[J]. 国外电子测量技术, 2024, 43 (7): 90-96.
SHI T, XIAO N ZH, DING Y, et al. Visual SLAM algorithm for fusing improved YOLOv7 in dynamic scenes[J]. Foreign Electronic Measurement Technology, 2024, 43 (7): 90-96.

[6] TAIRA H, OKUTOMI M, SATTLER T, et al. InLoc: Indoor visual localization with dense matching and view synthesis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(4): 1293-1307.

[7] KENDALL A, GRIMES M, CIPOLLA R. PoseNet: A convolutional network for real-time 6-dof camera relocalization[C]. 2015 IEEE International Conference on Computer Vision (ICCV), 2015: 2938-2946.

[8] SHAVIT Y, FERENS R. Do we really need scene-specific pose encoders? [C]. 25th International Conference on Pattern Recognition (ICPR), 2021:3186-3192.

[9] WALCH F, HAZIRBAS C, LEAL-TAIXE L, et al. Image-based localization using LSTMs for structured feature correlation[C]. 2017 IEEE International Conference on Computer Vision, 2017: 627-637.

[10] CAI M P, SHEN CH H, REID I. A hybrid probabilistic model for camera relocalization[C]. Proceedings of the 29th British Machine Vision Conference, 2018: 1-12.

[11] KENDALL A, CIPOLLA R. Modeling uncertainty in deep learning for camera relocalization[C]. 2016 IEEE International Conference on Robotics and Automation (ICRA), 2016: 4762-4769.

[12] KENDALL A, CIPOLLA R. Geometric loss functions for camera pose regression with deep learning[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 5974-5983.

[13] CHENG B W, MISRA I, SCHWING A G, et al. Masked-attention mask transformer for universal image segmentation[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 1280-1289.

[14] SHAVIT Y, FERENS R, KELLER Y. Learning multi-scene absolute pose regression with transformers[C]. 2021 IEEE/CVF International Conference on Computer Vision, 2021: 2733-2742.

[15] BLANTON H, GREENWELL C, WORKMAN S, et al. Extending absolute pose regression to multiple scenes[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 170-178.

[16] SONG X G, LI H J, LIANG L, et al. TransBoNet: Learning camera localization with transformer bottleneck and attention[J]. Pattern Recognition, 2024, 146: 109975.

[17] 陈仁钧, 费敏锐, 杨傲雷. 面向人机交互的手势指向

估计方法[J]. 仪器仪表学报, 2023, 44(3): 200-208.
CHEN R J, FEI M R, YANG AO L. Estimation of gesture pointing for human-robot interaction[J]. Chinese Journal of Scientific Instrument, 2023, 44(3): 200-208.

[18] LI H W, HUANG J X, HUANG H D, et al. Comparison of plain network and resnet classifiers for fruit image classification[C]. 2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT), 2023: 434-437.

[19] CHEN Z X, HE Z W, LU ZH M. DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention[J]. IEEE Transactions on Image Processing, 2024, 33: 1002-1015.

[20] PAN Z ZH, CAI J F, ZHUANG B H. Fast vision transformers with HiLo attention[C]. Advances in Neural Information Processing Systems, 2022, 35: 14541-14554.

[21] WANG X, WANG X, WANG CH, et al. Discriminative features matter: Multi-layer bilinear pooling for camera localization[C]. In 30th British Machine Vision Conference, 2020: 1-12.

[22] OTT F, FEIGL T, LOFFLER C, et al. ViPR: Visual-odometry-aided pose regression for 6dof camera localization[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 187-198.

作者简介

杨傲雷(通信作者),2004 年于湖北工业大学获得学士学位,2009 年于上海大学获得硕士学位,2012 年于英国女王大学获得博士学位,现为上海大学副教授,主要研究方向为机器人与视觉控制、计算机视觉与感知定位等。

E-mail:aolei@ shu. edu. cn

Yang Aolei (Corresponding author) received his B. Sc. degree from Hubei University of Technology in 2004, M. Sc. degree from Shanghai University in 2009 and Ph. D. degree from Queen's University Belfast, UK, in 2012. He is currently an associate professor at Shanghai University. His main research interests include robotics and vision control, computer vision and perception localization, etc.