Software Engineering and Applications 软件工程与应用, 2025, 14(3), 703-713
Published Online June 2025 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2025.143062

Hans汉斯

基于图像检索的无人机相机重定位方法

任昭扬

南京邮电大学物联网学院，江苏 南京

收稿日期：2025年4月1日；录用日期：2025年6月20日；发布日期：2025年6月30日

摘 要

无人机视觉重定位技术是其在GPS拒止环境下实现自主导航的核心支撑，广泛应用于城市巡检、灾害救援等场景。然而，复杂动态环境导致传统方法面临特征匹配歧义性高、位姿解算累积误差大等挑战。针对此，本文提出一种基于图像检索的无人机相机重定位方法，采用改进的MobileNetV2骨干网络，结合深度可分离卷积与反向残差模块的通道扩展－压缩策略，在减少参数量的同时保留关键几何信息；全局分支引入可微分NetVLAD层动态聚合局部特征，生成紧凑的4096维描述符；局部特征提取分支设计双解码器架构，利用子像素卷积与双三次插值实现高精度关键点检测与连续性描述子生成，并结合图注意力网络动态筛选几何一致性强的匹配对，通过Sinkhorn算法迭代优化软匹配矩阵，以自适应阈值剔除低置信度噪声。实验表明，该算法在弱纹理、动态遮挡等复杂条件下显著降低位姿误差。

关键词

无人机相机重定位，图像检索，卷积神经网络，注意力机制

UAV Camera Relocation Method Based on Image Retrieval

Zhaoyang Ren

School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing Jiangsu

Received: Apr. 1st, 2025; accepted: Jun. 20th, 2025; published: Jun. 30th, 2025

Abstract

UAV visual repositioning technology is the core support for its autonomous navigation in GPS denial environments, which is widely used in urban inspection, disaster rescue and other scenarios. However, the complex dynamic environment causes the traditional method to face challenges such as high ambiguity in feature matching and large cumulative error in position solving. To address this, this paper proposes a UAV camera relocation method based on image retrieval, which adopts an

文章引用: 任昭扬. 基于图像检索的无人机相机重定位方法[J]. 软件工程与应用, 2025, 14(3): 703-713.
DOI: 10.12677/sea.2025.143062

improved MobileNetV2 backbone network, combining a channel expansion-compression strategy with depth-separable convolution and inverse residual module, to reduce the number of parameters while retaining the key geometrical information; a global branch introduces a differentiable NetVLAD layer to dynamically aggregate local features, and generates compact 4096-dimensional descriptors; a local feature extraction branch introduces a local feature extraction branch to dynamically aggregate local features, and generates compact 4096-dimensional descriptors. The global branch introduces a differentiable NetVLAD layer to dynamically aggregate local features and generate compact 4096-dimensional descriptors; the local feature extraction branch designs a dual-decoder architecture, uses sub-pixel convolution and dual cubic interpolation to achieve high-precision keypoint detection and continuity descriptor generation, and combines with the graph-attention network to dynamically screen matching pairs with strong geometric consistency, and optimises the soft-matching matrix iteratively through the Sinkhorn algorithm to reject low-confidence noise by an adaptive thresholding. Experiments show that the algorithm significantly reduces the positional error under complex conditions such as weak texture and dynamic occlusion.

Keywords

UAV Camera Repositioning, Image Search, Convolution Neural Network, Attention Mechanism

1. 引言

随着无人机在智慧城市巡检、灾害救援、农业监测等领域的广泛应用，实时、精准的自主定位能力成为其核心需求。无人机在复杂动态环境中(如高楼林立的城市峡谷、光照多变的农田)常面临 GPS 信号缺失或漂移问题，亟需基于视觉的鲁棒重定位技术实现位姿恢复。

相机重定位技术[1]主要分为传统方法与深度学习方法。传统方法通过特征匹配(如 SIFT、ORB)结合几何约束(如 PnP [2]算法)求解位姿，虽计算高效但对动态干扰与光照变化敏感。深度学习方法中，混合位姿学习依赖图像检索(如 NetVLAD [3]、CamNet [4])或场景坐标回归间接推算位姿，而端到端方法(如 PoseNet、AtLoc)直接回归 6DoF 位姿。然而，无人机在未知区域飞行时面临两大挑战：其一，无规则运动导致先验场景信息(如 3D 地图)难以获取，场景坐标回归方法[5]因依赖离线建模且计算资源消耗指数级增长而失效；其二，动态环境与大规模地理范围下，现有算法因无法有效利用轻量级先验知识导致定位鲁棒性不足。为此，本文提出一种基于图像检索的无人机重定位算法，通过构建动态感知的层级检索框架与轻量化特征匹配模型，在不依赖 3D 重建的前提下，利用图像数据库实现高效位姿估计，突破传统方法与深度学习在未知场景中的局限性。

2. 基础知识

2.1. 相机位姿估计理论

2.1.1. 三维几何变换

三维几何变换通过刚体运动描述物体在空间中的旋转与平移，数学上由旋转矩阵 $R \in SO(3)$ 和平移向量 $t \in \mathbb{R}^3$ 组成，表示为齐次坐标下的变换矩阵 $T = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \in SE(3)$。其中 $SO(3)$ 是三维旋转群，满足

$R^{\mathrm{T}}R=I$，且 $\det(R)=1$，而 $SE(3)$ 是包含旋转和平移的特殊欧氏群。通过齐次坐标可将三维点 $P_{\text{world}}=(X,Y,Z,1)^{\mathrm{T}}$ 转换为相机坐标系下的坐标：$P_{\text{camera}}=T\cdot P_{\text{world}}=\begin{bmatrix}R & t\\ 0 & 1\end{bmatrix}\begin{bmatrix}X\\ Y\\ Z\\ 1\end{bmatrix}=\begin{bmatrix}R\cdot(X,Y,Z)^{\mathrm{T}}+t\\ 1\end{bmatrix}$。这一变换是视觉定位的基石，例如无人机通过自身位姿 T 将环境点从全局坐标系映射到相机视角。

2.1.2. 投影几何

投影几何基于针孔相机模型，将三维点 P_{world} 映射到二维像素平面。首先通过外参矩阵 $[R\,|\,t]$ 将世界坐标转换为相机坐标 $P_{\text{camera}}=R\cdot P_{\text{world}}+t$，再经内参矩阵 $K=\begin{bmatrix}f_x & 0 & c_x\\ 0 & f_y & c_y\\ 0 & 0 & 1\end{bmatrix}$ 投影到归一化平面 $(x,y,1)=\left(\frac{X_c}{Z_c},\frac{Y_c}{Z_c},1\right)$，最终得到像素坐标 (u,v)：$u=f_x\cdot x+c_x, v=f_y\cdot y+c_y$。实际应用中还需校正镜头畸变，例如径向畸变模型 $x_{\text{corrected}}=x\left(1+k_1r^2+k_2r^4\right)\left(r=\sqrt{x^2+y^2}\right)$，确保投影精度。无人机在动态环境中需实时求解逆投影问题(即从像素反推位姿)，依赖 PnP 算法或深度学习模型克服遮挡与噪声干扰。

2.2. 卷积神经网络

卷积神经网络(Convolutional Neural Network, CNN)是一种专为处理图像等网格化数据设计的深度学习模型，其核心通过卷积层的局部感知和参数共享机制高效提取空间特征：卷积核在输入数据上滑动计算局部区域特征(如边缘、纹理)，生成特征图并通过 ReLU 激活函数引入非线性，池化层(如最大池化)逐步降维并增强平移不变性，最终由全连接层完成分类或回归任务。这种结构以少量参数保留平移不变性，在图像分类、目标检测等领域表现卓越，例如 PoseNet 直接通过 CNN 回归相机 6DoF 位姿，而轻量化设计(如 MobileNet)可适配无人机等资源受限平台，实现实时视觉定位。

3. 基于图像检索的无人机相机重定位算法

本文提出一种三阶段图像检索无人机相机重定位模型，其原理图如图 1 所示，针对无人机动态未知场景实现高效相机位姿估计：全局检索模块通过共享骨干网络提取查询图像全局特征，经全连接层压缩后与数据库图像计算相似度，筛选 Top-K 候选图像；局部匹配模块复用骨干网络提取候选图像局部特征，结合 SuperPoint [6]关键点解码器与描述子解码器生成像素级关键点及描述子，通过匹配网络回归查询–候选图像对的几何一致性匹配矩阵；位姿回归模块融合候选图像的匹配矩阵权重及其真实位姿，基于加权优化策略恢复查询图像的绝对 6DoF 相机位姿。该方法通过全局粗筛、局部精配、位姿回归的三阶段设计，在避免 3D 重建资源消耗的同时，利用共享骨干网络降低计算冗余，结合 SuperPoint 的动态关键点检测增强对遮挡与光照变化的鲁棒性，为无人机无先验地图环境提供实时精准的视觉定位解决方案。

3.1. 全局检索

本文提出的全局检索模块基于改进的 MobileNetV2 [7]轻量化网络构建，如图 2 所示，其骨干网络包含三部分：标准卷积层用于提取初始图像特征；反向残差模块通过扩展层(逐点卷积扩展通道数)、深度可分离卷积(逐通道卷积提取空间特征+逐点卷积融合跨通道信息)及投影层(线性压缩通道维度)的级联结构，平衡特征多样性与计算效率；全局平均池化层将多维特征图压缩为一维向量。传统全连接层被移除，替换为 NetVLAD 层以增强全局特征表达能力，最终输出固定维度的紧凑特征向量。

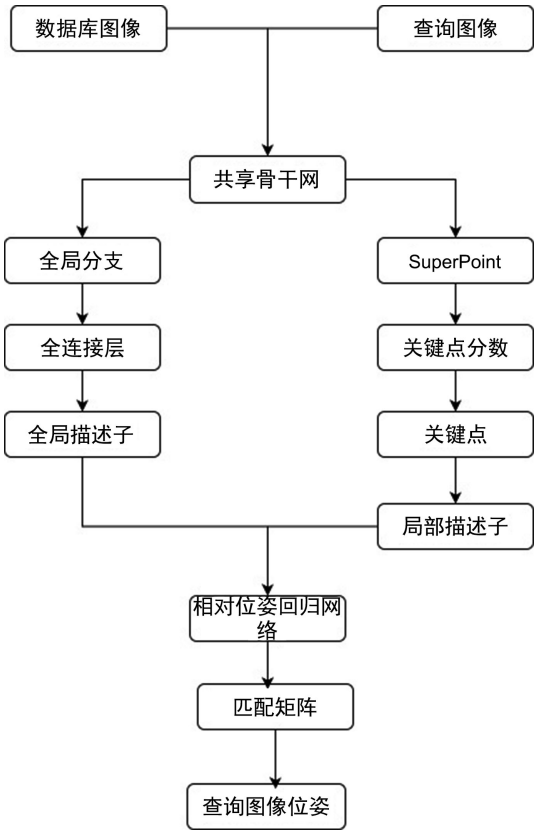

Figure 1. Model structure diagram
图 1. 模型结构图

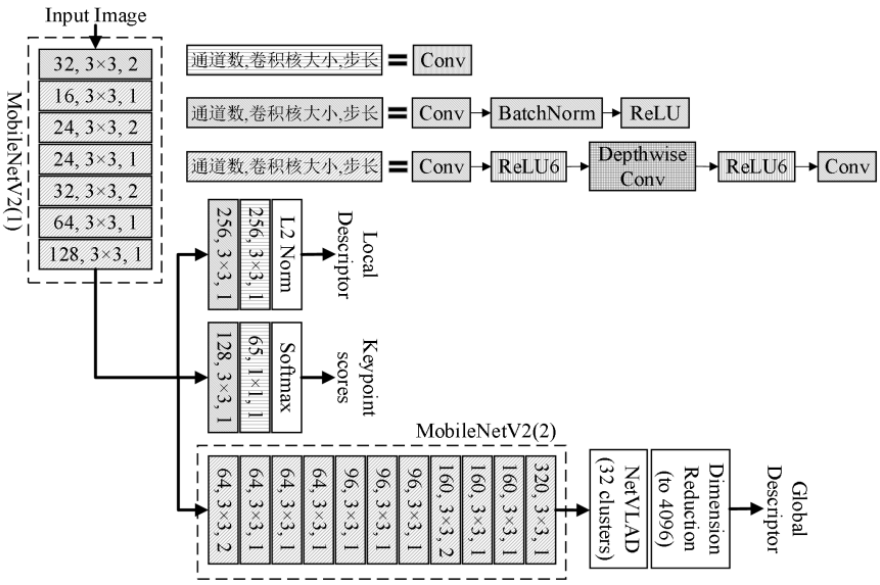

Figure 2. Global search module
图 2. 全局检索模块图

为降低计算量与参数量，全局分支采用深度可分离卷积替代标准卷积。其核心步骤为：1) 逐通道卷积独立提取各通道空间特征；2) 逐点卷积跨通道融合信息。对于输入特征图(尺寸 $H \times W \times C$，卷积核

$N\times N$)，标准卷积计算量为 $H\times W\times C^2\times N^2$，而深度可分离卷积计算量仅为 $H\times W\times C\times\left(N^2+1\right)$，参数量减少至标准卷积的 $1/\left(C\times N^2\right)$。尽管其可能因通道压缩损失部分特征表达能力，但通过反向残差模块中扩展层的通道升维与投影层的线性降维，可在保留关键信息的同时抑制噪声与过拟合。

全局特征提取器采用改进的 NetVLAD 层，基于可微分视觉词典对局部特征动态聚类：预设 K 个视觉词(如 64 个)，计算各词簇内特征残差的加权和，拼接形成高维全局描述符。相比传统 VLAD [8]，NetVLAD 通过端到端训练自适应优化视觉词典权重，使其适配无人机航拍图像的尺度与视角变化特性。为进一步压缩存储开销，NetVLAD 输出后接入全连接层，将特征维度从 16,384 降至 4096，并结合弱监督策略，以自监督方式学习场景专属的视觉词典与特征分布，避免人工标注成本。此设计在保证检索精度的前提下，使特征向量内存占用降低，支持汉明距离或余弦相似度的快速匹配，满足无人机端实时重定位需求。

3.2. 局部匹配

局部匹配分支旨在提取图像关键点及其局部描述子，实现跨图像的精准特征匹配。该分支由共享骨干网、关键点解码器、描述子解码器及匹配网络构成[6]，如图 3 所示。

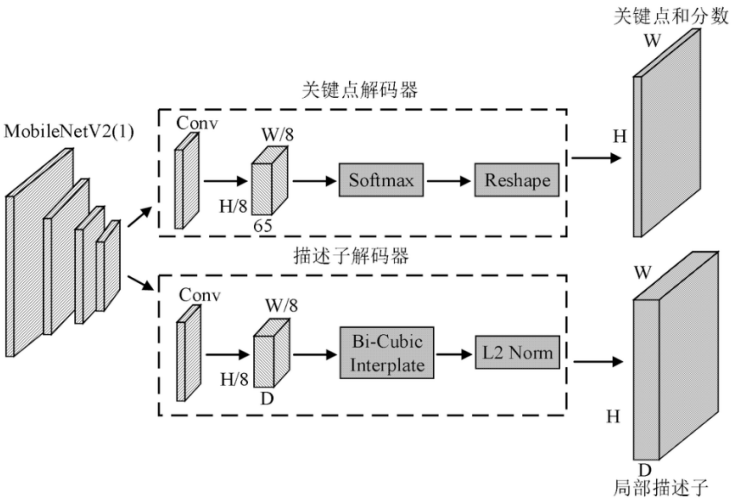

Figure 3. Local matching module
图 3. 局部匹配模块图

其中，共享骨干网复用全局分支改进的 MobileNetV2 前 7 层作为编码器，输入图像经其处理后输出 $H_c\times W_c\times 256$ 特征张量($H_c=H/8$，$W_c=W/8$)，在保留高分辨率空间信息的同时降低后续计算负荷。

关键点解码器通过卷积解码结构对编码器输出进行上采样与分类，生成 $H_c\times W_c\times 65$ 概率张量。其中 64 个通道对应输入图像 8 × 8 网格划分后的局部区域关键点分布，另设垃圾箱(Dustbin)通道过滤无关键点区域，减少误检。为实现原始分辨率重建，采用子像素卷积将每个通道值映射至 8 × 8 像素，得到 $H\times W\times 64$ 概率图，并通过非极大值抑制筛选置信度最高的关键点。此设计在提升空间分辨率的同时，确保关键点定位精度与稀疏性。

描述子解码器通过双三次插值对编码特征进行上采样，恢复至原始图像尺寸($H\times W\times 256$)，再经 L2 归一化生成单位长度描述子向量。双三次插值基于 16 邻域像素加权计算插值点，虽增加计算开销，但相比双线性或最近邻插值，可避免锯齿伪影并保持描述子空间连续性，显著提升特征匹配鲁棒性。描述子解码器与关键点解码器共享编码器特征，形成端到端联合训练框架，使关键点检测与描述子学习相互促进，增强局部特征对视角、光照变化的适应性。

3.3. 相对位姿回归网络

相对位姿回归模块基于图神经网络(GNN) [9]与最优匹配层构建，旨在通过稀疏特征匹配实现跨图像的几何一致性建模，其结构如图 4 所示。

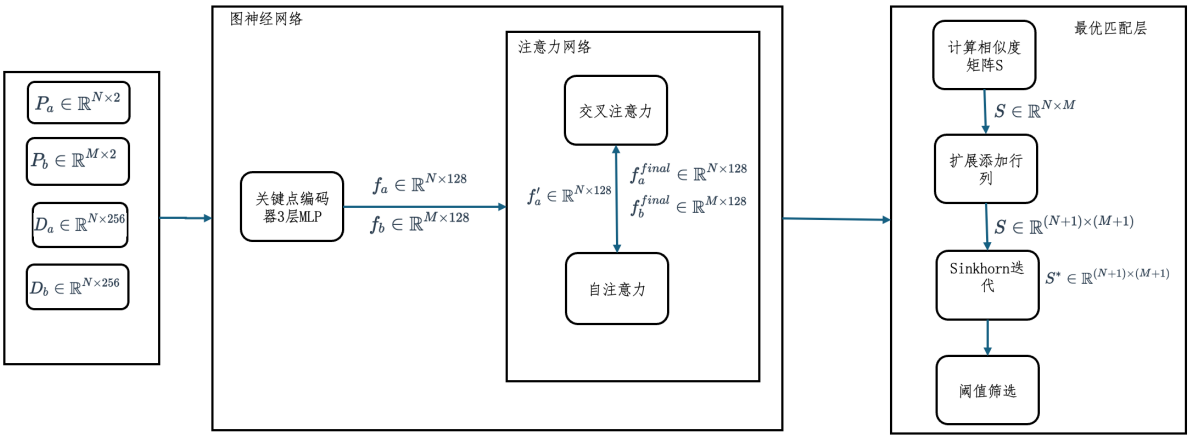

Figure 4. Relative positional regression network
图 4. 相对位姿回归网络

其核心流程分为三阶段：1) 关键点编码：输入两组图像的关键点坐标与描述子，通过多层感知机(MLP)编码为高维特征向量，MLP 采用 ReLU 激活函数逐层传递，融合空间坐标 (x,y) 与描述子(256 维)的几何与语义信息；2) 注意力图神经网络：将编码后的特征向量构建为图结构(节点为关键点，边为潜在匹配关系)，利用双向消息传递机制聚合上下文信息——通过注意力权重动态计算节点间相似度(内积 + Softmax 归一化)，更新节点特征以增强匹配判别性；3) 最优匹配层：基于 Sinkhorn 算法[10]对两组节点特征计算软匹配矩阵(相似度内积→Softmax 归一化→迭代行列归一化)，再通过自适应阈值 α 二值化生成稀疏匹配矩阵，剔除低置信度匹配对，输出最终特征对应关系用于后续位姿解算。

相对位姿回归网络通过注意力机制实现自适应的特征交互：通过三层 MLP 编码器将几何坐标与语义描述融合，生成 128 维联合特征 $f_a \in \mathbb{R}^{N\times 128}$ 和 $f_b \in \mathbb{R}^{M\times 128}$。在注意力网络阶段，首先通过自注意力机制(输入/输出维度均为 $N\times 128$)对同源特征进行上下文增强，采用 $Q \in \mathbb{R}^{N\times d_k}, K \in \mathbb{R}^{N\times d_k}, V \in \mathbb{R}^{N\times d_v}$ 的投影计算(本文设 $d_k = d_v = 128$)，通过 $\text{softmax}\left(QK^{\mathrm{T}}/\sqrt{d_k}\right)V$ 实现特征重校准，保持输出维度 $N\times 128$ 不变。

继而通过交叉注意力机制建立跨图像特征交互通道：以 $f_a \in \mathbb{R}^{N\times 128}$ 作为 Query 源，通过可学习权重矩阵 $W_q \in \mathbb{R}^{128\times 128}$ 生成 $Q_a \in \mathbb{R}^{N\times 128}$；以 $f_b \in \mathbb{R}^{M\times 128}$ 作为 Key/Value 源，分别通过 $W_k \in \mathbb{R}^{128\times 128}$ 和 $W_v \in \mathbb{R}^{128\times 128}$ 生成 $K_b \in \mathbb{R}^{M\times 128}$ 和 $V_b \in \mathbb{R}^{M\times 128}$。计算跨注意力权重矩阵 $A_{cross} \in \mathbb{R}^{N\times M}$：

$$A_{cross} = \text{softmax}\left(\frac{Q_a K_b^{\mathrm{T}}}{\sqrt{128}}\right) \tag{1}$$

最终输出增强特征 $f_a^{final} \in \mathbb{R}^{N\times 128}$ 和 $f_b^{final} \in \mathbb{R}^{M\times 128}$，计算式为：

$$f_a^{final} = A_{cross} V_b,\ f_b^{final} = A_{cross}^{\mathrm{T}} V_a \tag{2}$$

最优匹配层采用可微分的 Sinkhorn 算法求解，将相似度矩阵迭代优化为双随机矩阵(每行/列和为 1)，生成连续值的软匹配矩阵。为平衡匹配数量与精度，设定阈值 $\alpha = 0.2$ 进行二值化，仅保留置信度高于阈值的关键点对。相比传统最近邻匹配，此方法通过全局优化避免局部次优解，同时稀疏化策略减少误匹配对位姿估计的干扰。

3.4. 损失函数设计

本文采用两阶段训练策略优化无人机相机重定位模型。第一阶段聚焦全局特征与局部关键点学习，第二阶段优化关键点匹配网络。

3.4.1. 一阶段损失函数设计

在一阶段中全局分支引入改进的 HardTriplet Loss [11]，其定义为：

$$L_{\text{HardTriplet}} = \max\left(d\left(S_{\text{anchor}}, S_p\right) - d\left(S_{\text{anchor}}, S_n\right) + m_{\text{preset}}, 0\right) \tag{3}$$

S_{anchor} 为锚点图像特征，S_p 和 S_n 分别表示正样本(高重叠度且相对姿态角小)和负样本(低重叠度且姿态角大)特征，$d(\cdot)$ 为欧氏距离，$m_{\text{preset}} = 1$ 控制正负样本间距。

局部分支采用知识蒸馏策略，以预训练 SuperPoint 为教师模型生成关键点真值，学生模型通过 MSE 损失和交叉熵损失对齐教师输出。MSE 损失约束局部描述子相似性：

$$L_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left(R_p^i - R_g^i\right)^2 \tag{4}$$

其中 R_p 为学生模型预测描述子，R_g 为教师输出。交叉熵损失优化关键点概率分布：

$$L_{\text{CrossEntropy}} = -\frac{1}{N}\sum_{i=1}^{N} R_g^i \log\left(\text{softmax}\left(R_p^i\right)\right) \tag{5}$$

保留 Dustbin 通道过滤背景噪声，总损失为：

$$L_{\text{Total}} = L_{\text{MSE}} + L_{\text{CrossEntropy}} + L_{\text{HardTriplet}} \tag{6}$$

3.4.2. 二阶段损失函数设计

第二阶段优化关键点匹配网络，定义软分配损失[12]：

$$L_{\text{Soft}} = -\sum_{(i,j)\in S_{\text{match}}} \log s_{i,j} - \sum_{i\in S_{\text{unmatch}}^{\text{query}}} \log\left(1 - s_{i,j}\right) - \sum_{j\in S_{\text{unmatch}}^{\text{retrieval}}} \log\left(1 - s_{i,j}\right) \tag{7}$$

其中 S_{match} 为由相机绝对位姿几何验证的真实匹配集合，$S_{\text{unmatch}}^{\text{query}}$ 和 $S_{\text{unmatch}}^{\text{retrieval}}$ 分别为未匹配的查询与检索关键点集，$s_{i,j}$ 为 Sinkhorn 算法生成的软匹配概率。该损失最大化正确匹配置信度，抑制误匹配噪声，适配无人机动态场景下的鲁棒位姿回归需求。

4. 实验

4.1. 数据集

为了训练和评估我们基于图像检索的无人机相机重定位算法，我们选择了两个具有代表性的数据集：University-1652 和 7-Scenes。University-1652 数据集由全球 72 所大学的 1652 座建筑物图像组成，涵盖三个视角平台：无人机视角、卫星视角和地面视角[13]。其中，无人机视角图像是基于谷歌地球上的 3D 模型模拟生成的，每个目标点包含 54 张无人机图像和一张卫星图像。该数据集被划分为训练集和测试集，训练集包含 33 所大学的 701 座建筑物，共计 50,218 张图像；测试集包含 39 所大学的 951 座建筑物，包括查询和库图像。这种多视角、多源的数据集为训练无人机相机重定位算法提供了丰富的资源，有助于模型学习视角不变的特征，并在复杂的现实场景中提升泛化能力。

在评估阶段，我们采用了微软研究院发布的 7-Scenes 数据集[14]。该数据集包含七个不同的室内场景：Chess (棋盘)、Fire (火焰)、Heads (头部)、Office (办公室)、Pumpkin (南瓜)、Red Kitchen (红色厨房)和 Stairs (楼梯)。所有场景均由手持式 Kinect RGB-D 相机以 640 × 480 分辨率录制，利用 KinectFusion 系统获取了精确的相机轨迹和密集的 3D 模型。每个场景包含多个由不同用户录制的序列，每个序列由 500

至 1000 帧组成，数据采用 PNG 格式存储，每帧提供 RGB 彩色图像(24 位深度)、深度图(16 位，以毫米为单位)以及摄像头相对于世界坐标系的位姿(4 × 4 齐次变换矩阵)。该数据集的高精度位姿信息使其成为评估相机重定位算法性能的可靠基准。

通过在 University-1652 数据集上训练模型，使其学习多视角下的特征表示，然后在 7-Scenes 数据集上进行评估，我们能够全面验证算法在不同场景和视角下的性能表现。

4.2. 实验细节

4.2.1. 实验设备

实验使用的计算机的 GPU 是 Tesla K80 (11 GB)*12。CPU 是 40 vCPU Intel(R) Xeon(R) Silver 4210 CPU @ 2.20GHz。PyTorch 版本为 1.8.1，Python 版本为 3.8.0，CUDA 版本为 10.1。

4.2.2. 训练策略

本文提出两阶段联合训练框架，基于 University-1652 无人机航拍数据集训练模型，并在 7-Scenes 室内数据集测试性能，验证模型对异构环境的泛化能力。训练仅依赖图像与相机位姿真值(无需 3D 点云或深度)，通过端到端梯度回传机制将关键点检测后处理整合至张量运算中，确保参数可学习性。核心训练策略包括：1) 全局–局部联合优化：全局分支采用改进的 HardTriplet Loss，按重叠度(SuperGlue 特征匹配率)与相对姿态角划分样本为正样本(重叠度 ＞40%、姿态角 ＜20˚)、中等样本(重叠度 ＞10%、姿态角＜30˚)、负样本(重叠度 ＜5%)，通过随机采样平衡三类样本比例为 1:1:1；2) 局部特征蒸馏：以 Aachen 预训练 SuperPoint 为教师模型，通过 MSE 损失约束描述子相似性，交叉熵损失对齐关键点分布，结合 Dustbin 通道抑制背景噪声。输入图像经随机缩放(0.5~1.5 倍)及随机仿射变换增强，模拟无人机高度变化引起的尺度与透视形变，推理时统一缩至 512 × 342 像素(保持 3:2 长宽比)。

初始学习率 1e−4，Batch Size 16，训练 100 轮。相对位姿回归网络通过 Sinkhorn 算法生成软匹配矩阵。为避免梯度爆炸，训练初期启用全精度浮点(FP32)，后期切换为混合精度(FP16)。真实匹配标签由本质矩阵几何验证生成(重投影误差 ＜3 像素)，冻结全局分支参数以专注局部匹配优化。

4.3. 实验结果

4.3.1. 跨场景定位精度与动态鲁棒性验证

本文提出的基于图像检索的无人机相机重定位算法在 7-Scenes 数据集上展现出显著的跨场景重定位能力。如表 1 所示。

Table 1. Localization performance of different methods on seven scenarios of the 7-Scenes dataset
表 1. 不同方法在 7-Scenes 数据集 7 个场景上的定位性能

场景	CamNet	PoseNet	NN-Net	Ours
Chess	0.16 m, 7.61˚	0.32 m, 8.12˚	0.13 m, 6.46˚	0.12 m, 5.41˚
Fire	0.11 m, 6.32˚	0.47 m, 14.4˚	0.26 m, 12.72˚	0.10 m, 9.88˚
Heads	0.25 m, 8.71˚	0.29 m, 12.0˚	0.14 m, 12.34˚	0.12 m, 11.74˚
Office	0.19 m, 11.95˚	0.48 m, 7.68˚	0.21 m, 7.35˚	0.16 m, 8.42˚
Pumpkin	0.25 m, 9.35˚	0.47 m, 8.42˚	0.24 m, 6.35˚	0.23 m, 6.21˚
RedKitchen	0.19 m, 7.45˚	0.59 m, 8.64˚	0.24 m, 8.03˚	0.20 m, 7.13˚
Stairs	0.23 m, 8.56˚	0.47 m, 13.8˚	0.27 m, 11.82˚	0.21 m, 7.22˚

如图 5 所示，在长达 12.6 米的厨房场景遍历中，本文方法(红色轨迹)与真值轨迹(绿色)保持高度吻

合。其他传统方法出现均轨迹漂移现象，而本方法维持了定位稳定性。

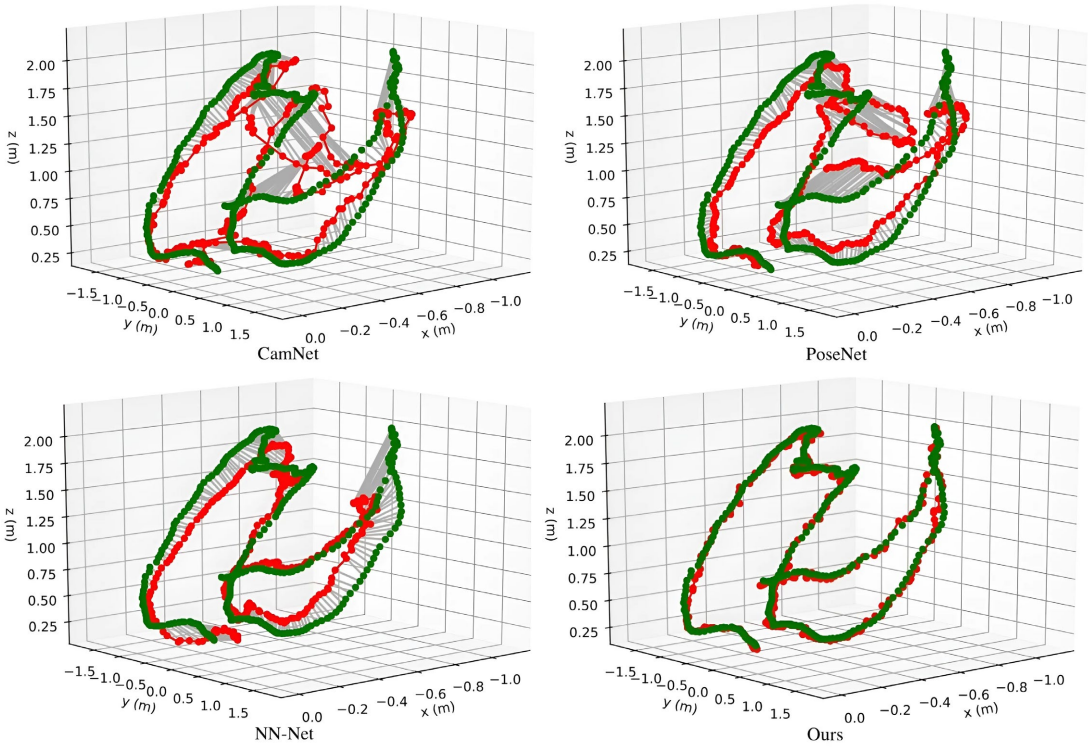

Figure 5. Visualization of the positioning accuracy of the Redkitchen-seq-12 scene (the green path represents the actual camera motion trajectory, and the red path presents the algorithm's output of the 6-DOF position estimation results)
图 5. Redkitchen-seq-12 场景的定位精度可视化(绿色路径代表实际相机运动轨迹，红色路径呈现算法输出的 6-DOF 位姿估计结果)

在小尺度结构化场景中，我们的方法以 0.12 m/5.41˚ (Chess)和 0.12 m/11.74˚ (Heads)的位姿误差显著优于其他对比算法，较 CamNet 误差降低 25%~52%，其核心优势源于层次化特征融合策略：通过 MobileNetV2 提取的全局描述符引导 SuperPoint 局部关键点检测，有效解决狭窄空间中纹理重复导致的误匹配问题。在高动态干扰场景中，本文提出的方法表现尤为突出，Fire 场景误差仅为 0.10 m/9.88˚ (对比 NN-Net 的 0.26 m/12.72˚)，验证了其对烟雾、运动模糊的鲁棒性，这得益于动态关键点筛选模块对噪声匹配的抑制能力。

4.3.2. 轻量化模型性能评估

为验证无人机场景下的实用性，本实验在 7-Scenes 上对比了 CamNet、PoseNet 及 NN-Net 算法。所有模型统一输入分辨率(224 × 224)且在相同训练集 University-1652 上训练，实测结果如表 2 所示。

Table 2. Comparison of the number of model parameters and inference speed of different methods
表 2. 不同方法的模型参数量与推理速度对比

方法	参数量(M)	推理速度(ms)
CamNet	7.21	48
PoseNet	64.28	103
NN-Net	33.29	62
Ours	7.01	35

本方法以 7.01 M 参数量与 35 ms 推理速度实现精度–效率的协同优化，其性能优势源于算法架构的针对性设计。在全局检索模块中，改进的 MobileNetV2 通过深度可分离卷积(Depthwise Separable Convolution)替代标准卷积，使特征提取阶段的参数量减少，同时保留反向残差结构的通道扩展能力，避免轻量化导致的特征表达退化。NetVLAD 层的自监督视觉词典优化进一步强化了全局特征判别性。

相对位姿回归网络的注意力 GNN 是关键的速度–精度平衡器：通过动态稀疏连接与轻量化 MLP (3 层 128 维)，在保持几何一致性建模能力的同时，降低匹配耗时。

此实验表明在无人机相机重定位场景中，本文轻量化设计通过深度可分离卷积与模块参数共享显著降低计算负载，使算法可直接部署于机载嵌入式平台，在保障实时性的同时，适应动态飞行环境中的资源约束与复杂干扰。

5. 结论

本文提出的一种基于图像检索的无人机相机重定位算法通过全局特征引导的层次化检索与动态局部匹配机制，在跨场景相机重定位任务中实现了鲁棒性与泛化能力的显著提升，在 7-Scenes 数据集上以 7.01 M 参数量与 35 ms 端到端推理速度实现了精度–效率的平衡。其深度融合全局上下文感知与局部几何约束的策略，有效解决了复杂环境下因纹理缺失、动态干扰及尺度变化导致的定位退化问题。

实验表明，该算法在弱纹理区域(如 Stairs 场景楼梯墙面)、动态物体遮挡(如 Fire 场景烟雾干扰)及多尺度视角切换(如 Pumpkin 场景长距离运动)等复杂条件下，仍能维持稳定定位精度，较传统方法显著降低因环境异质性引发的性能衰减。其通过全局检索与局部几何约束的协同优化，有效克服了传统算法在低纹理、动态噪声及运动尺度突变场景下的局限性。

参考文献

[1] Bianchi, M. and Barfoot, T.D. (2021) UAV Localization Using Autoencoded Satellite Images. *IEEE Robotics and Automation Letters*, **6**, 1761-1768. https://doi.org/10.1109/lra.2021.3060397

[2] Xu, D., Li, Y.F. and Tan, M. (2008) A General Recursive Linear Method and Unique Solution Pattern Design for the Perspective-N-Point Problem. *Image and Vision Computing*, **26**, 740-750. https://doi.org/10.1016/j.imavis.2007.08.008

[3] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T. and Sivic, J. (2018) Netvlad: CNN Architecture for Weakly Supervised Place Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 1437-1451. https://doi.org/10.1109/tpami.2017.2711011

[4] Ding, M., Wang, Z., Sun, J., *et al*. (2019) CamNet: Coarse-to-Fine Retrieval for Camera Re-Localization. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October 2019-2 November 2019, 2871-2880. https://doi.org/10.1109/ICCV.2019.00296

[5] 王静, 胡少毅, 郭苹, 等. 改进场景坐标回归网络的室内相机重定位方法[J]. 计算机工程与应用, 2023, 59(15): 160-168.

[6] DeTone, D., Malisiewicz, T. and Rabinovich, A. (2018) SuperPoint: Self-Supervised Interest Point Detection and Description. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Salt Lake City, 18-22 June 2018, 337-33712. https://doi.org/10.1109/cvprw.2018.00060

[7] Michele, A., Colin, V. and Santika, D.D. (2019) MobileNet Convolutional Neural Networks and Support Vector Machines for Palmprint Recognition. *Procedia Computer Science*, **157**, 110-117. https://doi.org/10.1016/j.procs.2019.08.147

[8] Delhumeau, J., Gosselin, P., Jégou, H. and Pérez, P. (2013) Revisiting the VLAD Image Representation. *Proceedings of the* 21*st ACM international conference on Multimedia*, *Barcelona*, 21-25 October 2013, 653-656. https://doi.org/10.1145/2502081.2502171

[9] Li, Y., Huang, Y., Liu, Z., *et al*. (2024) A Distributed Scheme for the Taxi Cruising Route Recommendation Problem Using a Graph Neural Network. *Electronics*, **13**, Article 574. https://doi.org/10.3390/electronics13030574

[10] Luise, G., Rudi, A., Pontil, M., *et al*. (2018) Differential Properties of Sinkhorn Approximation for Learning with Wasserstein Distance. arXiv.1805.11897.

[11] Hao, L.Y., Min, T. and Yun-Bo, Z. (2020) Cross-Modality Person Re-Identification Framework Based on Improved

Hard Triplet Loss.

[12] Ghasemi, S. and Moshtagh, J. (2014) A Novel Codification and Modified Heuristic Approaches for Optimal Reconfiguration of Distribution Networks Considering Losses Cost and Cost Benefit from Voltage Profile Improvement. *Applied Soft Computing*, **25**, 360-368. https://doi.org/10.1016/j.asoc.2014.08.068

[13] Zheng, Z., Wei, Y. and Yang, Y. (2020) University-1652: A Multi-View Multi-Source Benchmark for Drone-Based Geo-Localization. *Proceedings of the* 28*th ACM International Conference on Multimedia*, Seattle WA, 12-16 October 2020, 1395-1403. https://doi.org/10.1145/3394171.3413896

[14] Bilasco, I.M., Gensel, J., Villanova-Oliver, M. and Martin, H. (2005) On Indexing of 3D Scenes Using MPEG-7. *Proceedings of the* 13*th annual ACM international conference on Multimedia*, Hilton, 6-11 November 2005, 471-474. https://doi.org/10.1145/1101149.1101254