文章编号：1005-6734(2025)10-0963-09 doi.10.13695/j.cnki.12-1222/o3.2025.10.002

一种用于无人机景象匹配定位的异源图像快速检索方法

张小国，李天宇，史志豪，况余进
（东南大学 仪器科学与工程学院，南京 210096）

摘要：无人机景象匹配定位时，由于无人机图像和卫星基准图像之间域、观察角度等因素不同，容易出现误匹配甚至检索失败。针对上述问题，提出了一种基于显著位置特征的异源图像快速检索方法。首先，针对无人机图像与基准图像因获取场景和时间差异导致匹配失败的问题，设计了显著位置特征提取模块，在降低计算复杂度的同时能够提取更有效的上下文信息。其次，引入标签平滑损失函数，提升了模型的泛化能力。最后，提出分块微调策略以缓解大模型视觉 Transformer 在有限训练数据条件下的过拟合问题。实验结果表明，所提方法在 DenseUAV 数据集上 R@1 和 R@5 分别达到了 86.01%和 96.52%，mAP 达到了 76.04%，较现有主流方法 ViT-S 分别提升 5.83%、3.53%和 9.49%，单张图像检索时间为 9.55 ms，表明所提方法在无人机异源景象匹配中的有效性。

关 键 词：GNSS 拒止；无人机视觉定位；遥感影像；异源图像检索
中图分类号：V249.3 **文献标志码：**A

A fast heterogeneous image retrieval method for UAV scene matching and positioning

ZHANG Xiaoguo, LI Tianyu, SHI Zhihao, KUANG Yujin
(School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China)

Abstract: The scene matching and positioning of Unmanned aerial vehicles (UAVs) are prone to mismatching or even retrieval failure due to the differences in domain, observation angle and other factors between UAV images and satellite reference images. To address this issue, a rapid cross-source image retrieval method based on salient location features is proposed. Firstly, to solve the matching failure caused by scene and time differences between UAV images and reference images, a salient position feature extraction module is designed, which can extract more effective context information while reducing the computational complexity. Secondly, a label smoothing loss function is introduced to enhance the generalization ability of the model. Finally, a block-wise fine-tuning strategy is proposed to alleviate the overfitting problem of large models like vision transformer (ViT) under limited training data conditions. The experimental results show that the proposed method achieves 86.01% and 96.52% respectively in R@1 and R@5 on the DenseUAV dataset, and 76.04% in mAP, which is improved by 5.83%, 3.53% and 9.49% respectively compared with ViT-S. The retrieval time for a single image is 9.55 ms on the DenseUAV dataset, indicating the effectiveness of the proposed method in UAV cross-source scene matching.

Key words: GNSS-denied; UAV visual positioning; remote sensing imagery; heterogeneous image retrieval

近年来，无人机（Unmanned Aerial Vehicle, UAV）在精准农业、地面侦察和民用航空摄影等多个领域中发挥着愈加重要的作用。高精度的定位与导航能力是无人机完成各项任务的关键保障。通常无人机可采用全球卫星导航系统（Global Navigation Satellite System, GNSS）定位。然而，实际工作场景下 GNSS 信号容易因遮挡或者干扰丧失服务能力[1,2]，从而严重影响无人机的自主性和任务执行能力。

收稿日期：2024-10-15；**修回日期：**2025-07-30
基金项目：国家自然科学基金（62073078）
作者简介：张小国（1973—），男，教授，从事视觉导航定位。

为了应对这一问题，近年来基于景象匹配的无人机视觉绝对定位技术逐渐受到关注。该技术通过构建无人机图像特征，在预先构建的无人机或者卫星遥感影像数据库中实现精确检索和匹配，从而利用基准影像中的地理信息完成无人机的绝对位置估计。由于该技术不依赖于外部系统，并且能够在 GNSS 拒止的情况下提供高精度的定位信息，因此成为研究的热点。然而，由于无人机图像与卫星遥感影像来源不同，在分辨率、光照、视角等方面存在显著差异，传统同源图像检索方法在异源图像检索中难以取得理想效果。例如，传统的人工设计特征在处理异源图像时表现出明显的局限性，如尺度不变特征变换（Scale-Invariant Feature Transform, SIFT）[3]等，这类方法通常基于局部特征的匹配，难以全面捕捉图像的全局上下文信息，导致在特征匹配时受到限制。而卷积神经网络（Convolutional Neural Networks, CNN）在特征提取方面的强大能力使其成为图像检索研究的热点，通过孪生网络（Siamese Network）对地面和俯视图像进行训练，能够有效提取特征并实现检索。Tian 等人[4]采用 Faster R-CNN 检测建筑物，并结合孪生网络和多近邻匹配，进一步提升了匹配精度。王等人[5]通过引入可训练软分配深度学习框架 NetVLAD，结合内容检索技术，提出了一种聚合深度学习特征的无人机影像检索方法，可提取更稳定的特征，然而该方法平均检索一张影像耗时 3.7 s，实时性仍有待提升。

近期研究显示，将视觉 Tranformer（Vision Transformer, ViT）[6]应用于异源图像检索领域，已经取得了显著成效。Dai 等人[7]提出了一种创新的特征分割与区域匹配结构（Feature Segmentation and Region Alignment, FSRA），通过分析 Transformer 特征图的热量分布来划分区域，并巧妙地将不同视图中的特定区域进行对齐，以提升匹配精度。Yang 等人提出了一种地理定位网络（Evolving geo-localization Transformer, EgoTR）[8]，利用 Transformer 的自注意力机制捕捉异源图像间的全局依赖关系，有效减小了无人机图像与卫星影像之间的视觉差异。通过位置编码功能，EgoTR 进一步增强了对地面图像与俯视图像几何关系的理解，为异源图像检索中的跨视角匹配问题提供了有效解决方案。Dai 等人[9]提出了一个基于 TransFormer 的异源图像检索模型 DenseUAV，利用孪生网络学习无人机影像和卫星影像两个不同模态的表示。然而该模型在应对复杂异源图像检索任务时，表现出对上下文信息的利用不足的问题，影响了整体检索和匹配性能。

综上所述，跨视角、跨域及时间等因素对无人机异源图像检索的效率和成功率产生了严重影响。尽管目前基于 Transformer 和注意力机制的深度学习网络在提升检索和匹配性能方面取得了一定进展，但对上下文信息的提取仍存在不足，且对尺度、旋转和光照变化的适应性较弱，导致误匹配或匹配失败的情况时有发生。为此，本文提出了一种基于显著位置特征的无人机图像快速检索算法，针对 DenseUAV 基准模型进行改进，在骨干网络后引入显著位置特征提取模块（Salient Positions based K-NN Attention, SPKA），提升了上下文信息的提取能力并降低计算复杂度。同时采用标签平滑损失（Label Smoothing Cross Entropy, LSCE）函数替代交叉熵损失函数，增强模型的泛化能力及对尺度、旋转和光照变化的适应性，并通过分块微调策略缓解模型过拟合问题。

1 算法框架

本文算法基于 DenseUAV 基准模型改进，算法框架如图 1 所示。首先基于本文提出的 SPNet 模型，对查询图像和基准数据库图像进行特征提取，得到查询特征向量和基准特征向量库，然后将查询特征向量与基准特征向量数据库进行余弦相似度计算，按照相似度分数从高到低排序，实现对无人机图像快速准确的检索。

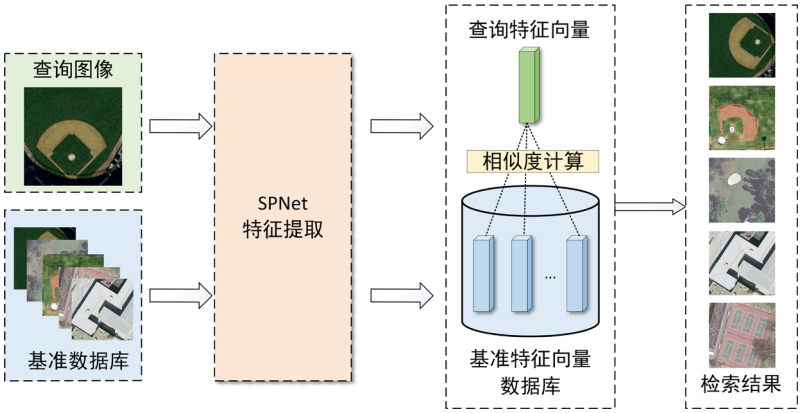

图 1 异源图像快速检索算法

Fig.1 Fast algorithm for heterogeneous image retrival

2　模型与改进方法

2.1　基准模型

本文提出的结合 SPKA、LSCE 损失函数和分块微调策略的改进模型 SPNet，用于无人机异源图像的高效检索，该模型的结构如图 2 所示。

模型采用孪生网络架构，网络中两个分支共享权重，以提高模型的参数效率和特征对齐能力。Zhang 等人[10]指出该架构在图像匹配和检索任务中已被广泛应用。模型首先接收无人机和卫星图像作为输入，并通过数据增强模块进行预处理，以增强模型在异源图像匹配中的鲁棒性。接着，主干网络提取图像特征，特征被送入 Head 模块进行集成并映射到特定的特征空间。其中主干网络为 ViT-S（Vision Transformer-Small）[6]，与传统的卷积神经网络相比，ViT 在特征提取过程中能够更好地平衡细粒度信息和计算效率，尤其在大规模数据集和高分辨率图像的推理速度方面表现出色，可以在性能和推理速度之间取得平衡。Head 模块通过 Global Pooling 将主干网络输出的 768 维的特征转换为用于分类的 512 维特征向量。

在训练阶段，基准模型通过全连接层和 softmax 函数计算类别概率，并结合三种监督学习方法优化损失函数，分别为表示学习（Representation Learning）、度量学习（Metric Learning）和互学习（Mutual Learning）。表示学习通过交叉熵损失（Cross Entropy, CE）对模型进行优化，具体函数如式(2)。

$$p_i = \begin{cases} 1, \text{if}(i = y) \\ 0, \text{if}(i \neq y) \end{cases} \tag{1}$$

$$L_{cls} = CELoss = -\sum_{i=1}^{K} p_i \log q_i \tag{2}$$

其中 q_i 表示的是模型输出的 logits 后经过 softmax 的结果，p_i 表示的是对应的 label。

度量学习则通过软加权三元组损失（Soft-Weighted Triplet Loss），在特征空间中缩小相似样本的距离，拉远不相似样本的距离，有效应对无人机与卫星图像之间的模态差异问题，提升模型的判别能力。传统三元组损失函数和软加权三元组损失函数定义为：

$$TriLoss(a,p,n) = \max(0, D(a,p) - D(a,n) + m) \tag{3}$$

$$L_{tri} = SWTriLoss(a,p,n) = \log(1 + e^{\alpha \times (D(a,p) - D(a,n))}) \tag{4}$$

其中 a 为锚点样本的特征向量，p 为锚点样本的正样本特征向量，n 为负样本特征向量，m 是控制正负样本之间距离期望差异的间隔，$D(a,b)$表示样本 a 与 b 之间的余弦相似度。

互学习引入了 KLLoss，通过知识蒸馏的方式，确保无人机图像和卫星图像类别分布的对齐，促进不同模态之间的并行学习。KLLoss 表达式为：

$$\text{KLDiv}(O_p \,\|\, O_q) = \sum_{i=1}^{N} O_p(i) \times \log\left(\frac{O_p(i)}{O_q(i)}\right) \tag{5}$$

$$L_{kl} = \text{KLLoss} = \text{KLDiv}\left(O_d \,\|\, O_s\right) + \text{KLDiv}\left(O_s \,\|\, O_d\right) \tag{6}$$

其中 O_p 和 O_q 分别表示教师和学生类别向量通过 softmax 的概率分布，O_d 表示无人机图像的类向量输出，O_s 表示卫星图像的类向量输出。

综合上述三部分，相加得到基准模型整体损失函数如式(7)所示。

$$L = L_{cls} + L_{tri} + L_{kl} \tag{7}$$

在推理阶段，模型利用余弦相似度进行图像检索和排序，从而实现高效的图像匹配。

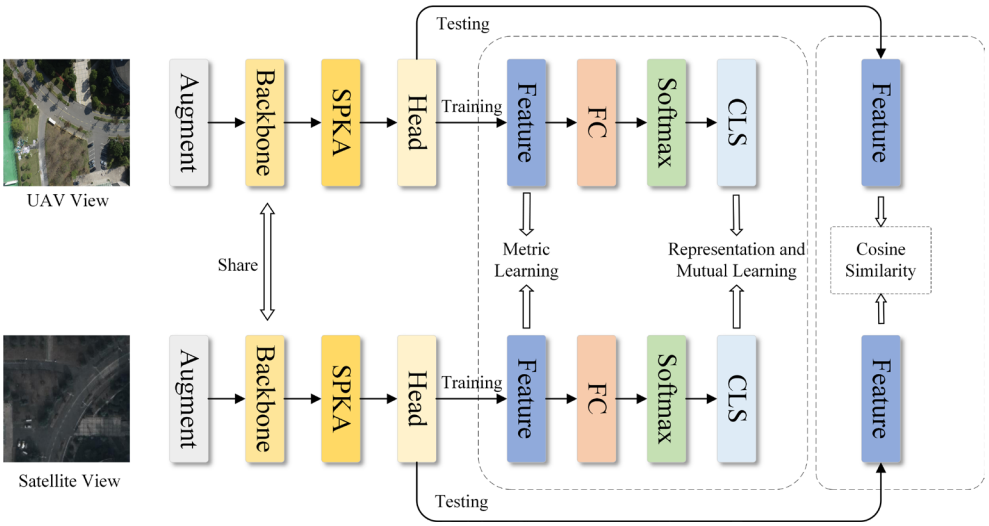

图 2　SPNet 模型结构图

Fig.2　SPNet model structure diagram

2.2 基于注意力机制的显著位置特征提取模块设计

虽然基准模型在 DenseUAV 数据集上效果显著，但由于无人机影像和卫星影像不同模态之间的域差异、时间差异和视角差异等，检索结果仍存在误匹配甚至是匹配失败的问题。此外，在深度学习领域，尤其是在处理序列数据和图像识别任务中，注意力机制已经成为提升模型性能的关键技术。Yuan 等人[11]的研究指出，传统的全连接自注意力机制由于其高昂的计算成本且噪声敏感，限制了其在大规模数据集或高分辨率图像处理上的应用。

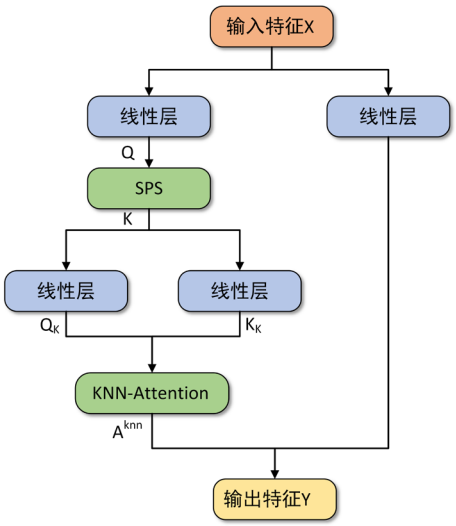

图 3 SPKA 模块图

Fig.3 SPKA module diagram

针对上述问题，设计了显著位置特征提取模块 SPKA，旨在提高网络训练和特征的质量，同时可以有效提取上下文信息和降低计算复杂度。SPKA 模块采取两个关键步骤来优化注意力计算：

1）通过显著位置选择算法（Salient Positions Selection, SPS），预先筛选出图像中的关键特征点。SPS 算法通过计算特征矩阵沿通道维度的平方和，选择前 k 个最显著位置作为关注点。该方法显著减少了 K-NN（K-Nearest Neighbors）注意力机制[12]需要处理的位置数量，从而降低了计算复杂度，同时避免了非相关特征的干扰。Fang 等人[13]的研究表明，在低网络层中，SPS 算法能通过选择显著特征位置蒸馏输入特征中的正确信息，有效减少背景噪声的影响。通过稀疏化注意力矩阵的计算，SPS 不仅节省了内存资源，还显著加快了注意力机制的计算速度。Gao 等人[14]进一步指出，SPS 算法能够显著优化模型在处理高分辨率图像时的表现。

2）K-NN 注意力机制通过关注 k 个与查询最相似的键，过滤掉噪声或不相关特征。该机制通过仅处理筛选出的显著位置矩阵 $\boldsymbol{K}$，进一步减少了计算负担，同时提高了特征对齐的质量。相比于传统的全局注意力机制，这种稀疏化方法不仅降低了计算复杂度，还能更好地捕捉局部特征中的关键上下文信息。因此，SPKA 模块在保持性能的同时显著减少了模型的内存消耗，并加快了训练速度。此外，SPKA 模块通过结合 SPS 算法和 K-NN 注意力机制，在处理长距离依赖时，能够更精确地捕捉当前任务最重要的特征并建模，从而提升了模型对全局信息的泛化能力。SPKA 模块的具体结构如图 3 所示。

本文提出的 SPKA 的注意力机制可用如下方程解释：

$$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}_{\boldsymbol{Q}},\boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}_{\boldsymbol{V}} \tag{8}$$

$$\boldsymbol{K}=M_{SPS}(\boldsymbol{Q}) \tag{9}$$

$$\boldsymbol{Q}_k=\boldsymbol{K}\boldsymbol{W}_{\boldsymbol{Q}_k},\boldsymbol{K}_k=\boldsymbol{K}\boldsymbol{W}_{\boldsymbol{V}_k} \tag{10}$$

$$\boldsymbol{A}^{knn}=\mathrm{softmax}(T_k(\frac{\boldsymbol{Q}_k\boldsymbol{K}_k^{\mathrm{T}}}{c})) \tag{11}$$

$$[T_k(A)]_{ij}=\begin{cases}A_{ij},A_{ij}\in top-k(rowj)\\-\infty,\ \text{otherwise}\end{cases} \tag{12}$$

$$\boldsymbol{Y}=\boldsymbol{A}^{knn}\cdot\boldsymbol{V} \tag{13}$$

其中 $\boldsymbol{Q}\in\boldsymbol{R}^{n\times c}$ 为查询矩阵，$\boldsymbol{K}\in\boldsymbol{R}^{k\times c}$ 为显著位置矩阵，$\boldsymbol{X},\boldsymbol{Y}\in\boldsymbol{R}^{n\times c}$ 分别为输入输出特征，$\boldsymbol{W}$ 为线性投影矩阵，T_k 为 K-NN 算法，M_{SPS} 为显著位置选择算法。

在 SPKA 中，SPS 算法起着重要作用，显著位置的选择将降低计算复杂度，并且在建模全局依赖时提取更有效的上下文信息，SPS 算法可以由算法 1 表示：

算法 1：SPS 算法

Input: - 大小为[c, h*w]的矩阵 $\boldsymbol{Q}$

- 超参数 k

Output: - 大小为[c, k]的矩阵 $\boldsymbol{K}$

1.计算 $\boldsymbol{Q}^{\mathrm{T}}$ 在通道维度上的平方

2.按照通道维度对 $\boldsymbol{Q}^{\mathrm{T}}$ 求和，得到 $\boldsymbol{Q}pow$

3.选择 $\boldsymbol{Q}pow$ 中最大的 k 个位置，记为 $indexk$

4.返回矩阵 $\boldsymbol{K}$= $\boldsymbol{Q}$(c, $indexk$)

SPKA 模块的具体步骤为：

1）显著位置的选择：对于输入特征 X 通过线性投影生成对应的查询矩阵 $\boldsymbol{Q}$ 和值矩阵 $\boldsymbol{V}$，对 $\boldsymbol{Q}$ 应用 SPS 算法选择显著位置，得到显著位置矩阵 $\boldsymbol{K}$。

2）对显著位置矩阵 $\boldsymbol{K}$ 应 0.

用 K-NN 注意力机制：对于显著位置选择算法选出的显著位置矩阵 $\boldsymbol{K}$，通过线性投影生成对应的查询矩阵 $\boldsymbol{Q}_k$ 和键矩阵 $\boldsymbol{K}_k$。对于每个查询向量，计算它与

所有键向量之间的相似度，选择与每个查询最相似的 k 个键，然后使用这些选定的键来构建一个稀疏的注意力矩阵 $\boldsymbol{A}^{knn}$ 。

3）输出特征：将通过 K-NN 注意力机制加权的输出特征重塑回原始特征图的维度。

2.3　损失函数引入

DenseUAV 基准模型中采用了交叉熵损失函数。然而交叉熵损失对错误的预测非常敏感，尤其是当数据集中包含不同的类别时，模型将会偏向于多数类，且交叉熵损失倾向于促使模型输出过于确定性的预测值（概率值接近 0 或 1），这增加了过拟合的风险。针对上述问题，本文引入了标签平滑交叉熵（Label Smoothing Cross-Entropy, LSCE）[15]损失函数。如式(14)和式(15)所示：

$$p_i=\begin{cases}(1-\varepsilon), \text{if}(i=y)\\ \dfrac{\varepsilon}{K-1}, \text{if}(i\neq y)\end{cases} \tag{14}$$

$$L_{cls\_LS}=CELoss_i=\begin{cases}(1-\varepsilon)\cdot CELoss, \text{if}(i=y)\\ \varepsilon\cdot CELoss, \text{if}(i\neq y)\end{cases} \tag{15}$$

其中 ε 表示对标签进行平滑的数值，一般设为 0.1。

最后，模型整体损失函数定义为：

$$L=L_{cls\_LS}+L_{tri}+L_{kl} \tag{16}$$

LSCE 的引入具有以下几个优势：

1）LSCE 鼓励模型不对任何单个类别过于自信，特别是在数据量有限的情况下有助于减少过拟合；

2）LSCE 通过平滑标签，使得模型不会对训练数据中的特定样本过于敏感，从而提高了模型对新图像的适应能力，能够泛化到新图像；

3）无人机影像和卫星影像之间存在尺度、旋转和光照等变化，LSCE 可以通过减少对特定特征的过度依赖，帮助模型学习到更加适应性的特征表示来应对这些变化。

2.4　分块微调算法

本文模型的主干网络通过基于 timm 迁移学习库提供的 ViT 预训练模型来实现。Touvron 等人[16]的研究表明，从零开始训练一个 ViT 通常需要 1400 万到 3 亿张图像。然而，由于无人机和卫星图像数据集的收集成本较高，通常规模较小，直接在这些数据集上训练 ViT 模型可能会导致模型过拟合。因此，针对在无人机和卫星图像数据有限的情况下训练大型 ViT 模型的问题，本文提出了一种分块微调策略（BlockWise Fine-Tuning, BW-FT），以充分利用预训练模型的优势，提高模型在小规模数据集上的泛化能力。分块微调策略的算法步骤可以由算法 2 表示：

算法 2：分块微调

1. 冻结所有的 Transformer 块 B
2. 初始化参数：
   $t=2, b=12, lr=3e-4, lr_{decay}=0.85$
3. **while** $0 <= i < epochs$ **do**
   if $i\%t == 0$ *and* $b > 0$ **then**
   unfreeze B[b]
   $b \leftarrow b-1$
   $lr \leftarrow lr * lr_{decay}$

分块微调策略具体步骤如下：首先冻结除瓶颈层外的所有 Transformer 块，并在每隔 t 个 epoch 后解冻一个新的 Transformer 块。同时，学习率会随着解冻过程根据衰减因子逐步降低。此策略有效缓解了预训练模型训练过程中出现的灾难性遗忘问题[16]，即新知识覆盖旧知识而导致信息丢失。通过对模型不同层次采用差异化的微调策略，分块微调降低了早期层次知识丢失的风险，并且结合学习率衰减机制，有效控制了后续层的梯度更新，避免权重大幅波动。这一策略不仅加快了模型收敛速度，还提升了整体性能，相较于传统微调方法，可以显著改善训练效果。

3　实验与分析

3.1　实验平台与数据集

本文实验环境具体配置如表 1 所示。

表 1　实验环境配置

Tab.1　Experimental environment configuration

实验环境	配置
CPU	Intel(R) Xeon(R) Gold 6330CPU @ 2.00GHz
GPU	NVIDIA GeForce RTX 3090
深度学习框架	Pytorch 1.10.0
编程语言	Python 3.8
操作系统	Ubuntu 18.04

图 4　数据集无人机视角

Fig.4　UAV view of the dataset

图 5　数据集卫星视角

Fig.5　Satellite view of the dataset

本文采用 Dai 等人[9]于 2023 年发布的 DenseUAV

数据集来验证所提算法，DenseUAV 是首个专为无人机景象匹配定位任务设计的公开数据集，图 4 和图 5 分别展示了部分无人机视角和卫星视角影像。其数据采集完全基于真实世界场景，避免了合成数据与实际环境之间的偏差，这使得该数据集在无人机图像匹配定位任务中，尤其是异源图像检索方面，具备更高的应用价值和实用性。此外，该数据集覆盖了不同类型的环境，包括植被较多的区域以及建筑物密集的区域，这些场景的图像特征与其他乡村或城市区域存在一定的相似性。通过这些多样化的场景，能够全面评估所提方法在不同环境下的鲁棒性与泛化能力，从而验证算法在实际应用中的表现。无人机图像部分，该数据集在三个不同高度（80 m、90 m、100 m）采集无人机图像，以控制尺度变化，且使用 RTK 技术将采样点的误差控制在 1 米以内。为减少天气和光照的影响，采用随机天气（晴天、阴天）和随机时间（6:00-18:00）采样。卫星图像部分，该数据集使用 20 级谷歌地图图像，包含 2020 年和 2022 年两个不同年份的卫星图像，以及三种不同的缩放比例，这有助于增强模型在空间尺度和时间变化下的鲁棒性。因此，该数据集通过提供高质量的真实世界数据，以及高密度图像采样和多场景、多时间条件下的数据采集，使其在无人机异源图像检索任务中具有良好的代表性和适用性。

表 2 DenseUAV 数据集构成

Tab.2 Dense UAV dataset composition

Subset	UAV	Satellite	Classes	Universities
Training	6768	13536	2256	10
Query	2331	4662	777	4
Gallery	9099	18198	3033	14

DenseUAV 的数据集具体构成如表 2 所示。训练集由 10 所大学的 2256 个采样点组成，包括 6768 张无人机图像和 13536 张卫星图像。测试集包含 4 所大学的 777 个采样点，共 2331 张无人机图像和 4662 张卫星图像。Gallery 集合涵盖所有 14 所大学，共 3033 个采样点，包含 27297 张图像。

3.2 评价指标

本文实验采用 Recall@K（R@K）和平均检索精度（mean Average Precision, mAP）作为图像检索的精度评价指标。R@K 是图像检索领域最常用的评价指标，其代表检索系统在给定查询图像的前 K 个结果中，正确匹配目标的概率。该指标能够反映模型在给定检索结果数量下，检索到相关图像的能力，以 R@1 为例，一个样本是否正确匹配可以表示为：

$$I(l_q,l_i)=\begin{cases}1,\text{if}\left(l_q=l_i\right)\\0,\text{if}\left(l_q\neq l_i\right)\end{cases} \tag{17}$$

其中 l_q 对应于查询的类别，l_i 对应于按计算的欧氏距离升序排序的第 i 个图像的类别。如果属于同一类别，结果值为 1，否则为 0。对于所有样本，R@1 定义为：

$$\text{R@1}=\frac{1}{\|S\|}\sum_{q\in S}I(l_q,l_1) \tag{18}$$

其中 S 是所有查询图像的集合，$\|S\|$表示 S 中的图像数量。只有查询的类别与图库中距离最近的图片的类别相同时，R@1 指标的数值才会增加。本文中选择 R@1 和 R@5 作为主要评价指标，R@1 反映了模型在检索任务中能否将正确匹配目标作为首选结果，这对于无人机景象匹配定位任务至关重要，因为首个检索结果通常直接用于后续处理；而 R@5 则评估了模型在前五个候选结果中找到正确匹配目标的能力，这在实际应用中可以为后续处理提供备选方案，从而提升系统的鲁棒性。

在此基础上，本文还引入 *mAP* 作为评估检索系统性能的重要指标，*mAP* 综合考虑了检索系统在所有查询图像上的平均精度，通过计算每个查询图像的精度并求均值来得到。对于每个查询图像，首先按检索结果的排序位置计算精度，定义为每个位置的平均精度（Average Precision, AP）。具体计算公式为：

$$AP=\frac{1}{n}\sum_{i=1}^{n}\frac{i}{k_i} \tag{19}$$

其中 n 是相关样本总数，k_i 是排序列表中第 i 个相关样本的索引位置，对于整个查询集合 S，*mAP* 的定义为：

$$mAP=\frac{1}{\|S\|}\sum_{q\in S}AP_q \tag{20}$$

其中$\|S\|$表示查询集合 S 中的图像数量，AP_q 表示查询图像 q 对应的 *AP* 值。

3.3 实验结果与分析

为验证本文提出的基于基准模型改进的 SPNet 算法的有效性，设计了一系列消融实验，旨在评估显著位置特征提取模块的引入、损失函数的修改以及微调策略对模型性能的影响。骨干网络的预训练采用 timm 框架，并移除了额外的分类层。无人机和卫星图像均被调整为 224 × 224 的输入分辨率。训练过程中，采用随机梯度下降优化器，初始学习率设为 0.003，批处理大小为 8。骨干网络的学习率被设定为其他网络层学习率的 0.3 倍。模型总共训练了 120 个 epoch，以充分评估不同改进策略对模型性能的提升效果。

表 3　不同改进方法对模型性能的影响

Tab.3　The impact of different improvement methods on model performance

Method	R@1	R@5
Baseline	80.18%	93.99%
Baseline+LSCE	81.77%	94.55%
Baseline+SPKA	84.98%	95.71%
Baseline+BW-FT	80.49%	94.15%
Baseline+LSCE+SPKA+BW-FT	**86.01%**	**96.52%**

如表 3 所示，引入 LSCE 损失函数后模型的 R@1 和 R@5 分别提升了 1.59 和 0.56 个百分点，加入 SPKA 模块后，模型的 R@1 和 R@5 分别提升了 4.80 和 1.72 个百分点。引入分块微调策略进一步提升了 0.31 和 0.16 个百分点。当这些策略结合使用时，R@1 和 R@5 的性能分别提升了 5.83 和 2.53 个百分点。综上所述，本文提出的方法具有显著的有效性和可行性，能够显著提升网络的整体性能。

表 4　DenseUAV 数据集上不同方法性能对比

Tab.4　Performance comparison of different methods on DenseUAV dataset

Method	Params	InferTime	R@1	R@5	mAP
ResNet50	27.8 M	10.20 ms	16.52%	39.30%	23.14%
EfficientNet-B3	14.1 M	23.40 ms	42.81%	64.52%	39.7%
EfficientNet-B5	32.3 M	33.85 ms	44.96%	67.78%	47.25%
ConvNext-T	30.1 M	8.45 ms	60.23%	81.94%	46.27%
DeiT-S	23.7 M	9.60 ms	71.77%	89.70%	59.35%
PvTv2-B2	26.8 M	20.45 ms	77.99%	92.79%	67.76%
Swinv2-T	29.9 M	19.25 ms	77.99%	92.49%	69.05%
ViT-S(Baseline)[9]	23.3 M	9.45 ms	80.18%	93.99%	69.45%
FSRA	26.0 M	10.55 ms	82.58%	94.94%	69.80%
LPN	26.0 M	10.60 ms	83.05%	94.89%	73.12%
SPNet(ours)	23.6 M	9.55 ms	**86.01%**	**96.52%**	**76.04%**

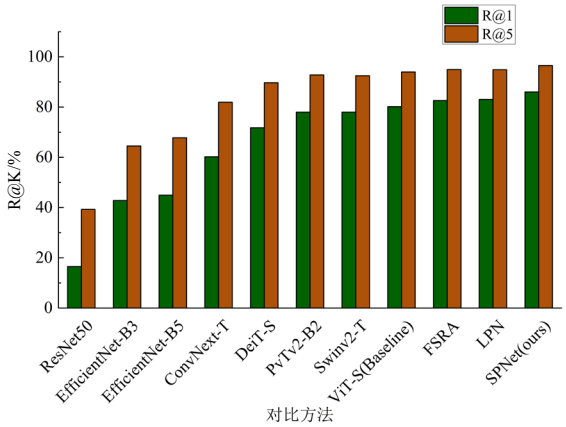

图 6　不同方法 R@K 指标对比

Fig.6　Comparison of R@K indicators of different methods

在 DenseUAV 数据集上，采用 R@K、mAP 和单张图片推理时间为性能评价指标，将本文模型与不同主干网络及 head 模块的方法进行对比，实验结果如表 4 所示，图 6 清晰地展示了各方法在 R@1 和 R@5 上的差异。本文提出的 SPNet 在在 R@1、R@5 以及 mAP 上均取得了最优表现，其中 R@1 和 R@5 分别达到 86.01%和 96.52%，相较于基准模型（R@1 为 80.18%，R@5 为 93.99%）分别提升了 5.83 和 2.53 个百分点，mAP 达到 76.04%，相较于基准模型提升了 6.59 个百分点。同时，SPNet 单张图片推理时间约为 9.55 ms，与基准模型相当，表明所提方法在性能与效率间实现了良好平衡。

此外，为了进一步验证所提方法的优势，选取了在异源图像检索任务中具有代表性或在相近任务表现优异的多种 SOTA 方法进行对比，如表 5 所示。

表 5　不同 SOTA 方法性能对比

Tab.5　Comparison of performance with different SOTA methods

Method	BackBone	R@1	R@5
MSBA	ResNet50	46.13%	64.22%
LPN	ResNet50	32.43%	56.80%
SDPL	ResNet50	7.08%	14.07%
LPN	ViT-S	83.05%	94.89%
Baseline	ViT-S	80.18%	93.99%
SPNet(Ours)	ViT-S	**86.01%**	**96.52%**

MSBA[18]和 LPN[19]在使用 ResNet50 主干网络时在 DenseUAV 数据集上的 R@1 仅分别达到 46.13%和 32.43%，SDPL[20]的 R@1 和 R@5 更是仅有 7.08%和 14.07%，尽管该方法在 University-1652 数据集中表现良好，但在 DenseUAV 这样跨视角变化、成像模式复杂的异源检索场景中适应性不足。相比之下，LPN 在采用 ViT-S 主干网络时的 R@1 与 R@5 则提升至 83.05% 和 94.89%，这也进一步说明了 vision Transformer 在处理跨视角、跨分辨率影像时的潜力。最终，SPNet 以 86.01%的 R@1 和 96.52%的 R@5 取

得最优效果。实验结果表明，SPNet 能够有效提取上下文信息和显著位置特征，显著提升了无人机景象匹配定位任务中的图像检索性能。

为了更好展示本文所提算法的检索效果，图 7 和图 8 分别展示了测试集中的检索结果。蓝线左侧为无人机视图图像，即查询图像，右侧为卫星视图中最接近查询图像的 5 幅图像。正确匹配的图像由绿色框框出，错误匹配的图像由红色框框出表示。

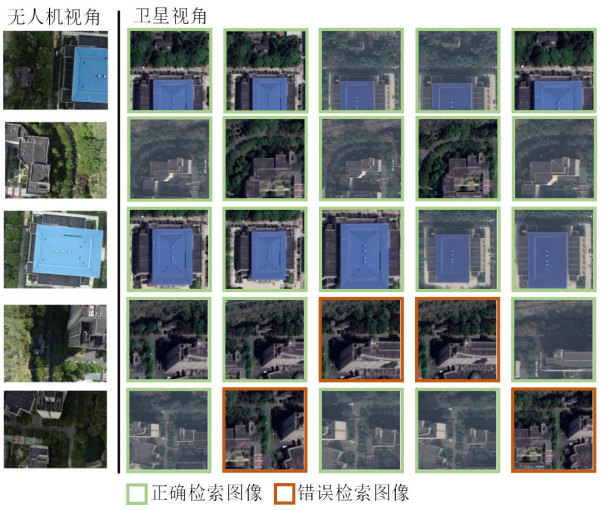

图 7 SPNet 检索结果

Fig.7 SPNet Retrieval results

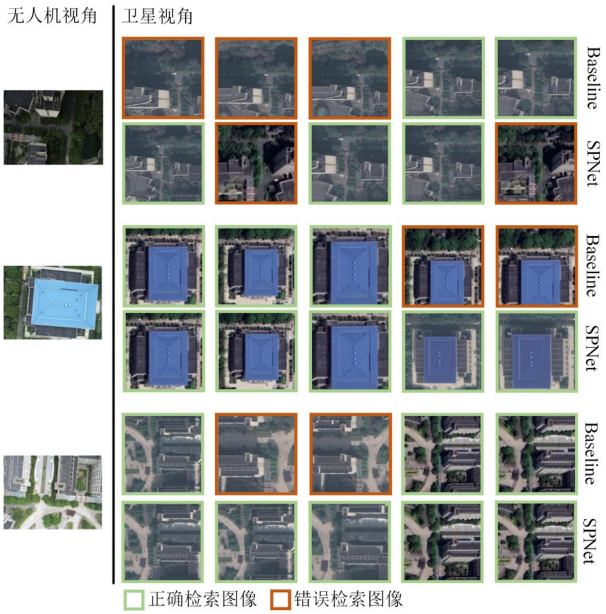

图 8 SPNet 与基准模型检索结果对比

Fig.8 Comparison of retrival results between SPNet and baseline

如图 7 所示，排名第一的均为正确的检索结果，表明 SPNet 所学习到的显著位置特征能够在一些极为相似的场景中实现有效检索。在图 8 中，通过对比 SPNet 与基准模型的检索结果可以发现：基准模型在一些难以区分、极为相似的场景中，容易发生误匹配，而改进后的 SPNet 能够在相同的查询图像下有效降低误匹配率，排名前 5 的图像中误匹配情况显著减少。该结果直观地证明了 SPNet 相较于基准模型在检索任务中表现出更高的准确性和鲁棒性，进一步验证了其优越性。

4 结 论

针对 GNSS 拒止环境下的无人机景象匹配定位任务中，由于无人机图像与卫星基准影像之间域、观察角度、时间等因素的不同，在检索过程中容易出现误匹配甚至检索失败的情况，本文提出了一种异源图像快速检索方法。通过设计显著位置特征提取模块、引入标签平滑损失函数以及分块微调策略，有效提升了模型在异源图像检索任务中的泛化能力和抗过拟合能力。

实验结果表明，本文方法在 DenseUAV 数据集上的表现优于现有主流方法，在 R@1 和 R@5 指标上分别从 80.18%和 93.99%提升至 86.01%和 96.52%，在 mAP 指标上提升了 9.49%，表明了其在复杂异源图像检索场景下的有效性。同时，本文方法还能够显著减少检索时间，达到了单张图像 9.55 ms 的检索速度，证明了其在实际应用中的高效性。此外，由于本文所使用数据集包括植被较多的区域和建筑物密集的区域，涵盖了类似于乡村和城市区域的图像特征。因此所提方法具备一定的跨环境适应性。

尽管所提方法在 DenseUAV 数据集上取得了优异表现，尤其是在 R@1 和 mAP 指标上表现出色，并展现了一定的跨环境适应性，但仍有进一步优化的空间。未来工作将聚焦于以下两方面：一是扩展至更加多样化的环境和场景，包括不同地理条件（如山区、高空）和复杂天气条件（如雨、雪、雾）下的数据，以全面评估模型在多变环境下的表现；通过在更广泛的数据集上进行验证，可进一步增强模型的跨场景适应性和实际应用效果；二是优化算法设计，以进一步提升模型的效率和精度，更好地应对实际应用中的复杂挑战。

参考文献（References）：

[1] 尚克军，赵亮，张伟建，等. 基于深度特征正射匹配的无人机视觉定位方法[J]. 中国惯性技术学报，2024, 32(01): 052-057.
Shang K, Zhao L, Zhang W, et al. Unmanned aerial vehicle visual localization method based on deep feature orthorectification matching[J]. Journal of Chinese Inertial Technology, 2024, 32(01): 052-057.

[2] 韩勇强，于潇颖，纪泽源，等. 面向城市复杂环境的GNSS/INS高精度图优化算法[J]. 中国惯性技术学报，2022, 30(05): 582-588.
Han Y, Yu X, Ji Z, et al. The high-precision factor graph optimization algorithm of GNSS/INS for urban complex environment[J]. Journal of Chinese Inertial Technology, 2022, 30(05): 582-588.

[3] Lowe D G. Distinctive image features from

scale-invariant keypoints[J]. International Journal of Computer Vision, 2004,60(2):91-110.
[4] Tian Y, Chen C, Shah M. Cross-view image matching for geo-localization in urban environments[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 1998-2006.
[5] 王小攀，李建胜，王安成，等. 面向无人机绝对定位的遥感影像快速检索方法[J]. 中国惯性技术学报，2024, 32(04): 363-370+378.
Wang X, Li J, Wang A, et al. Fast retrieval method of remote sensing image for UAV absolute location[J]. Journal of Chinese Inertial Technology, 2024, 32(04): 363-370+378.
[6] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale[J]. arxiv preprint arxiv:2010.11929, 2020.
[7] Dai M, Hu J H, Zhuang J D, et al. A transformer-based feature segmentation and region alignment method for UAV-view geo-localization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(7): 4376-4389.
[8] Yang H, Lu X, Zhu Y. Cross-view geo-localization with evolving transformer[J]. arxiv preprint arxiv:2107.00842, 2021.
[9] Dai M, Zheng E, Feng Z, et al. Vision-based UAV self-positioning in low-altitude urban environments[J]. IEEE Transactions on Image Processing, 2023, 33: 493-508.
[10] Zhang K, Qi S, Cai J, et al. Content-based image retrieval with a convolutional siamese neural network: Distinguishing lung cancer and tuberculosis in CT images[J]. Computers in biology and medicine, 2022, 140: 105096.
[11] Yuan Z, Zhang H, Lu P, et al. Ditfastattn: Attention compression for diffusion transformer models[J]. arXiv preprint arXiv:2406.08552, 2024.
[12] Wang P, Wang X, Wang F, et al. Kvt: k-nn attention for boosting vision transformers[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 285-302.
[13] Fang S, Li K, Li Z. Salient positions based attention network for image classification[J]. arxiv preprint arxiv:2106.04996, 2021.
[14] Gao T, Li Z, Wen Y, et al. Attention-free global multiscale fusion network for remote sensing object detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 62: 5603214.
[15] Huo J. A study of spatial attention and squeeze excitation block fusion improved resnet for identifying bank notes[J]. Security and Communication Networks, 2021: 1-8.
[16] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[C]//International conference on machine learning. PMLR, 2021: 10347-10357.
[17] Howard J, Ruder S. Fine-tuned language models for text classification[J]. arxiv preprint arxiv:1801.06146, 2018.
[18] Zhuang J, Dai M, Chen X, et al. A faster and more effective cross-view matching method of UAV and satellite images for UAV geolocalization[J]. Remote Sensing, 2021, 13(19): 3979.
[19] Wang T, Zheng Z, Yan C, et al. Each part matters: Local patterns facilitate cross-view geo-localization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(2): 867-879.
[20] Chen Q, Wang T, Yang Z, et al. SDPL: Shifting-dense partition learning for UAV-view geo-localization[J]. arXiv preprint arXiv: 2403.04172, 2024.

（上接第 962 页）

[6] 杨秀建，皇甫尚昆，颜绍祥. 基于改进 UKF 的 UWB/IMU/里程计融合定位方法[J]. 中国惯性技术学报，2023, 31(5): 462-471.
Yang X, Huangfu S, Yan S. Fusion positioning method with UWB/IMU/odometer based on the improved UKF[J].
Journal of Chinese Inertial Technology, 2023, 31(5): 462-471.
[7] 韩勇强，于潇颖，纪泽源，等. 面向城市复杂环境的 GNSS/INS 高精度图优化算法[J]. 中国惯性技术学报，2022, 30(5): 582-588.
Han Y, Yu X, Ji Z, et al. The high-precision factor graph optimization algorithm of GNSS/INS for urban complex environment[J]. Journal of Chinese Inertial Technology, 2022,30(5): 582-588.
[8] Mi J, Wang Q, Liu P, et al. A performance enhancement method for redundant IMU based on neural network and geometric constraint[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 1-11.
[9] Yan H, Shan Q, Furukawa Y. RIDI: Robust IMU double integration[C]//15th European Conference on Computer Vision (ECCV). 2018: 621-636.
[10] Herath S, Yan H, Furukawa Y. RONIN: Robust neural inertial navigation in the wild: Benchmark, evaluations & new methods[C]//2020 IEEE International Conference on Robotics and Automation (ICRA). France, 2020: 3146-3152.
[11] Chen B, Zhang R, Wang S, et al. Deep-learning-based inertial odometry for pedestrian tracking using attention mechanism and Res2NET module[J]. IEEE Sensors Letters, 2022, 6(11): 1-4.
[12] Zeinali B, Zanddizari H, Chang M J. IMUNet: Efficient regression architecture for inertial IMU navigation and positioning[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 1-13.
[13] Wang Y, Cheng H, Meng MQH. Inertial odometry using hybrid neural network with temporal attention for pedestrian localization[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-10.
[14] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1251-1258.
[15] Ashish V, Noam S, Niki P, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017: 5999-6009.
[16] Li J, Yuan G, Duan H. Adaptive Kalman filter for SINS/GPS integration system with measurement noise uncertainty[J]. IEEE Transactions on Industrial Electronics, 2022, 69(12): 13925-13935.
[17] Sun S, Melamed D, Kitani K. IDOL: Inertial deep orientation-estimation and localization[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(7): 6128-6137.
[18] Zhu Q, Zhuang H, Zhao M, et al. A study on expression recognition based on improved mobilenetV2 network[J]. Scientific Reports, 2024, 14(1): 8121.
[19] Li Y, Yu A, Meng T, et al. Deepfusion: LiDAR-camera deep fusion for multi-modal 3D object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 17182-17191.
[20] Arora L, Singh S K, Kumar S, et al. Ensemble deep learning and EfficientNet for accurate diagnosis of diabetic retinopathy[J]. Scientific Reports, 2024, 14(1): 30554.