文章编号　1004-924X(2025)24-3956-21

融合KL-注意力ConvNeXt的卫星无人机图像匹配方法

程　龙*，杨睿光，金　鑫，李晓宏，孙　伟
(长光卫星技术股份有限公司，吉林 长春 130011)

摘要：针对卫星与无人机影像在尺度、视角及模态间存在显著差异所带来的匹配挑战，本文提出一种融合KL散度互学习机制与三重注意力模块的ConvNeXt端到端跨模态匹配方法。本方法以ConvNeXt为骨干网络，并在其中引入空间-通道-方向三重注意力，以提升对尺度、旋转等几何形变的鲁棒性，同时强化对方向性结构的建模能力。同时，通过KL散度引导不同模态分支进行互学习，有效对齐卫星图与无人机图的特征分布，缓解由视角差异引发的模态偏移。在University-1652上，本方法在双向检索均取得较高精度：Drone→Satellite的R@1/AP为89.40%/90.98%，Satellite→Drone为93.44%/89.33%。相对ConvNeXt-Tiny，R@1/AP分别相对提升2.21%/1.80%（Drone→Satellite）与2.19%/2.73%（Satellite→Drone）；较近期代表方法（TransFG，GeoFormer）亦有小幅优势，与近几年部分算法相比仍具有竞争力。消融表明：三重注意力与KL互学习分别带来4.13%和3.55%的R@1增益。效率方面，在统一协议下单卡推理（bs=1）延迟20.83 ms，峰值显存0.53 GB；同图多目标设置中，$K=5$时V-Rate@5=97.7%。综上，在保证实时性的同时，方法显著提升跨模态检索精度与几何稳健性，具备工程可部署性，并能在同图多目标场景下有效抑制干扰候选，适用于无人机视觉导航与地理定位等应用。

关　键　词：卫星无人机图像匹配；跨模态匹配；ConvNeXt；三重注意力；KL散度；视觉地理定位
中图分类号：TP394.1；　　**文献标识码**：A
doi：10.37188/OPE.20253324.3956　　**CSTR**：32169.14.OPE.20253324.3956

KL-attention enhanced convnext for cross-modal satellite-UAV image matching

CHENG Long*, YANG Ruiguang, JING Xin, LI Xiaohong, SUN Wei

(*Chang Guang Satellite Technology Co., Ltd., Changchun* 130011, *China*)
* *Corresponding author, E-mail: longcheng*201304011@163.*com*

Abstract: To address the matching challenges arising from significant scale, viewpoint, and modality differences between satellite and UAV images, this paper proposed an end-to-end cross-modal matching method based on ConvNeXt, integrating a KL-divergence mutual-learning mechanism and a triplet-attention module. With ConvNeXt as the backbone, the method introduced spatial-channel-directional triplet attention to enhance robustness against geometric deformations (e.g., scale and rotation) and to strengthen directional-structure modeling. Meanwhile, KL divergence guided mutual learning between modal branches to align the feature distributions of satellite and UAV images, alleviating modality shifts caused by view-

收稿日期：2025-08-09；**修订日期**：2025-09-08.
基金项目：长春市科技发展计划项目（No. 2024WX01）；吉林省科技发展计划项目（No. 20240302013GX）

point differences. On the University-1652 dataset, it achieves high bidirectional retrieval accuracy: Drone→Satellite R@1/AP=89.40%/90.98%, Satellite→Drone=93.44%/89.33%. Compared with ConvNeXt-Tiny, it gains 2.21%/1.80% (Drone→Satellite) and 2.19%/2.73% (Satellite→Drone) in R@1/AP, with slight advantages over recent methods (TransFG , GeoFormer), and remains competitive with several recent approaches. Ablation studies show triplet attention and KL mutual learning contribute +4.13 and +3.55 percentage points to R@1, respectively. For efficiency, single-GPU inference (bs=1) has 20.83 ms latency and 0.53 GB peak memory under a unified protocol; in single-image multi-target settings, V-Rate@5=97.7% (K=5). In conclusion, the method ensures real-time performance, boosts cross-modal accuracy and geometric robustness, is engineering-deployable, suppresses distractors in multi-target scenarios, and suits UAV visual navigation and geolocation.

Key words: satellite-UAV image matching; cross-modal matching; ConvNeXt; triplet attention; KL divergence; visual geo-localization

1　引　言

随着航空与智能感知技术的快速发展，无人机(Unmanned Aerial Vehicles, UAVs)在城市建设、应急救援与环境监测等领域展现出广泛应用潜力[1-2]。然而，传统导航高度依赖全球定位系统(GPS)，在城市峡谷、地下空间或电磁干扰/拒止环境中易失效，难以满足鲁棒定位需求[3-4]。基于卫星-无人机跨模态图像匹配的视觉地理定位因此成为重要方向：以覆盖广、定位精准的卫星影像为参照，通过图像检索推断UAV拍摄位置，以实现类GPS的视觉导航能力[5-6]。但受成像高度、视角、尺度、光照与几何形变差异影响，传统局部特征(如SIFT/ORB)在跨模态场景下鲁棒性受限，尤其在纹理稀疏或重复结构区域表现欠佳[7-8]。

深度学习方法显著提升了跨视角/跨模态匹配性能，但仍面临三方面挑战。其一，表征侧：卷积网络固定感受野与下采样易丢失细粒度信息，对大幅度视角与几何变化缺乏内在不变性；尽管多尺度与注意力机制有所缓解，仍受局部感知瓶颈约束[9-13]。相比之下，Transformer依托自注意力具备全局建模能力(如L2LTR, EgoTR, LoFTR等)，在跨模态匹配中表现突出，但计算/显存开销与数据依赖更高[14-17]。其二，几何与对齐侧：实际检索链路常需结合几何一致性核验(如比率测试+RANSAC)、视角归一化与区域/部件级对齐，以缓解大视角差异并提升误匹配抑制能力[12,27-29]。其三，跨模态学习范式侧：自监督/弱监督对齐与知识蒸馏/互学习通过分布一致性约束与软目标传递提升泛化与稳健性[20]。同时，工程层面的检索协议(如低维全局描述、近似最近邻索引、单次前向、无测试时TTA等)与网络结构共同决定精度-效率权衡[5,8,18,25,32-36]。

针对上述问题，本文提出一种基于改进ConvNeXt骨干并融合Kullback－Leibler(KL)互学习机制的端到端跨模态匹配方法。具体而言，我们在ConvNeXt上引入空间－通道－方向三重注意力，以显式建模多维依赖关系，增强对尺度与几何变化的鲁棒性[18-19]；同时通过KL互学习在卫星/无人机双分支间进行双向蒸馏与分布对齐，从而缓解模态偏移并提升跨域泛化能力[20]。该设计在统一评测协议下兼顾精度与效率，并在同图多目标情形中可结合几何校验稳健定位多个候选。

在University-1652数据集上开展系统评测与消融[21]，在统一评测协议下，与四类具有代表性的方案进行对比：CNN度量学习基线(对比/三元组损失、检索微调与全局池化等)[22-26]；局部/部件增强方法(区域划分与部件级表示)[27-28]；Transformer系列(层到层/演化式Transformer、局部-全局混合对齐等)，覆盖近年的代表性工作与2024年度最新方法[14-17,34-37]；以及信息融合/联合表征方法[34]。实验结果显示，所提方法在双向检索(Drone→Satellite与Satellite→Drone)的AP与Recall@1等指标上取得更优或具有竞争力的综合性能，并在端到端链路上保持良好的效率表现。

对比涵盖 ICCV 2023 Sample4Geo[39]以及2024～2025年度IEEE TGRS/TCSVT正式发表的代表性方法(GeoFormer[36]，MCCG[40]，SDPL[41]等)，并结合国内期刊代表性方法进行验证，结果表明本文方法在双向AP与R@1指标上均保持竞争力。

本文的主要贡献在于：第一，提出ConvNeXt＋三重注意力的跨模态表征骨干，在空间、通道与方向三维度强化对几何与方向性结构的感知，同时兼顾推理效率[18-19]；第二，设计KL双向互学习实现跨模态特征分布对齐，并与分类/度量损失协同优化，提升跨视角鲁棒性[20]；第三，在统一评测协议下对比涵盖从经典CNN到近几年最新Transformer系列在内的多条技术路线[14-17,22-29,34-37]，并结合同图多目标可视化与几何核验分析，系统验证了方法在精度-效率权衡与实用场景中的有效性。

2 相关工作

2.1 跨视角地理定位综述与数据集

地理定位与跨视角检索近年来发展迅速，综述工作从应用动机、挑战与评测协议等方面给出系统梳理，并强调双向检索(Drone→Satellite与Satellite→Drone)的必要性与可比性[5][8]。University-1652提供多城市场景、跨模态视角与标准评测划分，已成为卫星－无人机匹配的常用基准[21]。

2.2 CNN+度量学习的全局表征

早期方法多以VGG/ResNet为骨干，结合对比损失、三元组损失与检索微调等度量学习范式，学习跨模态可比的全局描述[22-25]。随后，GeM等全局池化策略与实例级训练进一步提升了检索鲁棒性与可迁移性[25-26]。此类方法实现简洁、效率较高，但在大视角几何变化与显著模态差异下表征能力受限，易出现外观相似但几何不一致的误匹配[12]。

2.3 局部/部件增强与区域对齐

为缓解尺度与视角差异带来的对齐困难，局部/部件增强方法通过空间划分、部件汇聚与区域级对齐强化细粒度判别力[27-29]。这类方法在复杂城市场景中取得了显著增益，但区域划分的稳健性与跨模态一致性仍依赖于骨干表征质量，且对几何一致性的显式约束有限。

2.4 Transformer 系方法

Transformer依托自注意力的全局建模能力，在跨视角匹配与语义对齐方面表现突出。代表性工作包括层到层/演化式Transformer的跨视角对齐[14-15]、基于注意力的无检测子局部匹配[16]、以及弱监督的跨注意力语义对齐[17]。面向卫星-无人机场景的最新研究，围绕特征分割与区域对齐、Siamese/多分支结构与反馈联动等方向持续推进，近年来的多项方法在University-1652上进一步刷新了双向检索的准确率与AP[34-37]。与此同时，Transformer模型通常伴随更高的计算与显存开销，对部署友好性与训练数据质量较为敏感[9]。

2.5 跨模态对齐与互/蒸馏学习

为缩小卫星与无人机间的分布差异，互学习与蒸馏策略被用于对齐不同模态分支的输出分布或中间表示，常以KL散度/温度软化等形式实现软目标一致性约束，从而提升泛化与稳健性[20]。该思路亦与改进的卷积骨干(如ConvNeXt)兼容，可在不显著牺牲效率的前提下增强跨域一致性[18,32]。

2.6 检索链路与效率

端到端检索效果不仅取决于表征网络，还依赖于统一的评测协议与高效索引。实践中常采用低维全局描述、ANN索引(如倒排文件＋乘积量化或小世界图结构)、单次前向、无测试时TTA的设置，以平衡精度与效率[5,8,25]。在此框架下比较不同骨干与训练范式，更能客观反映方法的“精度-效率”权衡。

2.7 几何一致性核验与同图多目标

为抑制外观相似带来的误检，几何一致性核验(如局部特征匹配＋RANSAC)常作为检索后的稳健性环节，但多数工作侧重跨库检索指标，对“同一幅卫星图内多目标定位”与“核验通过率”等图内指标关注较少。面向实际应用，引入基于候选集合与核验内点率的图内指标，有助于衡量多目标定位的可用性与几何可靠性。

2.8 小结与研究定位

综合来看，CNN＋度量学习路线高效但对大幅度几何/模态差异敏感；局部/部件增强与

Transformer方法在跨视角对齐上更具潜力，但计算成本较高且对分布对齐与几何一致性仍有改进空间[14-17,34-37]。针对上述缺口，本文在统一评测协议下：(1)基于ConvNeXt引入空间-通道-方向三重注意力，强化对尺度/旋转等几何变化与方向性结构的建模[18-19]；

(2)采用KL互学习对齐跨模态分支输出分布，提升跨域一致性与泛化[20]；

(3)在常规跨库检索指标之外，补充同图多目标与几何核验相关图内指标与可视化，面向落地场景评估稳健性。

上述设计与现有Transformer系/局部增强/蒸馏类方法形成互补，并在双向检索与端到端效率上取得更优综合表现[34-37]。

3　算法原理

为提升卫星与无人机影像在跨模态匹配中的定位精度，本文针对两类图像在尺度、分辨率及视角上的显著差异所导致的特征分布偏移与空间对齐难题，提出一种融合KL散度和三重注意力机制的ConvNeXt匹配框架。如图1所示，整个方法分为三大模块。

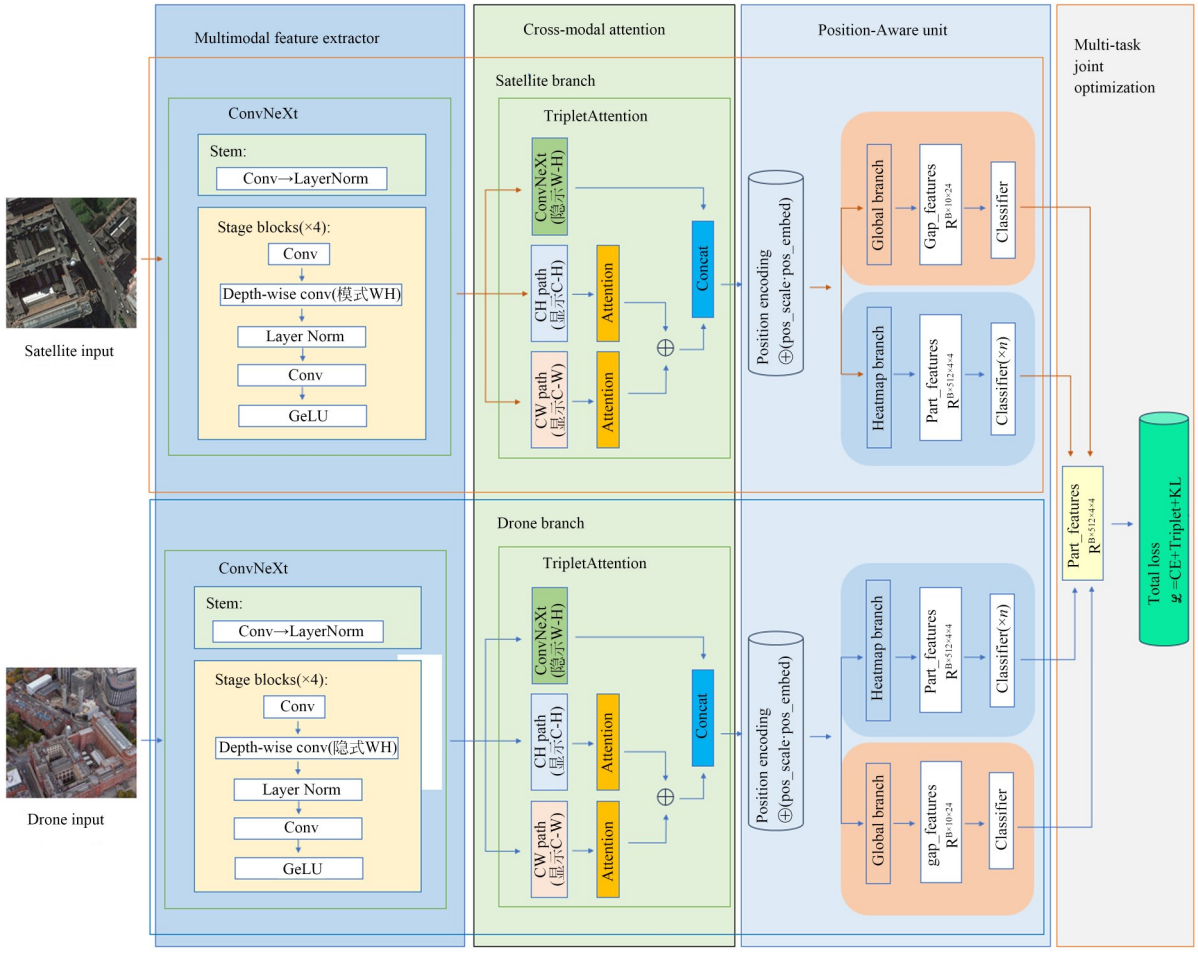


图1　算法原理框架图

Fig. 1　Framework diagram of the algorithm principle

3.1　多模态特征提取模块

本框架采用ConvNeXt-Base作为骨干网络，具有4个阶段的层次化结构，可提取多尺度特征。每阶段均由Depth-wise卷积与逐点卷积(1×1Conv)组合构成，以适配不同模态图像的特征结构。最终输出8×8×1 024的高维特征图。同时，引入可学习的位置编码(⊕pos_scale·pos_embed)增强空间感知能力。

3.2 跨模态注意力模块

设计了双路径 Triplet Attention 机制来建模空间依赖关系，CW 路径通过维度重排(B,H,C,W)强化水平依赖；HC 路径通过重排(B,W,H,C)建模垂直信息；此外，网络中嵌入的 7×7Depth-wise 卷积结构在不显式增加路径的前提下隐式建模了 HW 空间耦合，进一步提升对几何差异的鲁棒性。其中，C(Channel)/H(Height)/W(Width)分别表示通道、垂直高度与水平宽度；CW/HC 指在 $C\times W$ 与 $H\times C$ 平面上构建注意力的两个分支(对应水平与垂直方向的依赖)，HW 表示 $H\times W$ 的空间平面。

3.3 多任务优化模块

损失函数设计融合三部分，交叉熵损失(Cross Entropy)用于基础分类监督，三元组损失(Triplet Loss)用于约束跨模态特征相似性，KL 散度损失用于对齐两个分支的特征分布，从而有效缓解模态偏移问题。三者协同训练提升了匹配的判别能力与泛化能力。

4 算法设计

为提升卫星图像与无人机图像之间的跨模态匹配鲁棒性与判别精度，本文提出一种基于 ConvNeXt 骨干网络与三重注意力机制(Triplet Attention)的端到端跨模态匹配算法框架。该框架整体结构可划分为三个关键模块：(1)多模态特征提取模块；(2)跨模态注意力增强模块；(3)多任务联合优化模块。以下分别对三大模块的设计细节进行阐述。

4.1 多模态特征提取模块

4.1.1 层次化 ConvNeXt 骨干网络

为适配跨模态图像在分辨率、尺度变化与视角扭曲方面的显著差异，本文引入 ConvNeXt-Base 作为图像编码器骨干网络。ConvNeXt 作为一种现代卷积神经网络(Modern CNN)，融合了 Vision Transformer 的设计理念，特别是在感受野扩大、归一化策略优化、激活函数替换等方面表现突出，兼顾了高表达能力与推理效率。

在网络整体结构设计上，ConvNeXt 主干网由 Stem 层和四个 Stage(阶段)构成层次化结构，通过逐层下采样实现图像空间尺寸减半与通道数倍增。整体流程如下：

4.1.1.1 Stem 层设计

作为网络的初级特征提取模块，Stem 层通过大核卷积实现快速降采样与底层特征编码：

$$\boldsymbol{F}_{\text{stem}}=\text{LayerNorm}(\text{Conv2d}_{4\times4}(\boldsymbol{I})),\tag{1}$$

其中：$\boldsymbol{I}\in\boldsymbol{R}^{3\times H\times W}$(表示输入卫星或者无人机图像数据张量，本文设计 $H=256$，$W=256$)，$\text{Conv2d}_{4\times4}$: kernel = 4, stride = 4, groups = 1，$\boldsymbol{F}_{\text{stem}}\in\boldsymbol{R}^{96\times64\times64}$。

4.1.1.2 Stage 层设计

ConvNeXt 主干网络在 Stem 层之后，依次堆叠四个 Stage 模块(Stage1-4)，实现图像特征的逐层提取与下采样。每个 Stage 均由多个 ConvNeXt Block 构成，用于建模不同尺度下的局部与语义特征。整体结构体现了 ResNet 经典的层次化特征提取思想，同时引入了现代 ViT 架构的优化策略。Stage 层结构细节如表 1 所示。

表 1 Stage 层结构细节

Tab. 1 Stage layer structure details

阶段	特征尺寸	通道数	Block 数	关键改进
Stage 1	64×64	96	3	大核深度卷积(DWConv7×7)
Stage 2	32×32	192	3	LayerNorm 替代 BatchNorm
Stage 3	16×16	384	9	扩展通道比(1→4)提升模型容量
Stage 4	8×8	1 024	3	GELU 激活函数增强非线性表达能力

每个 Stage 中的 ConvNeXt Block 结构统一采用以下设计：

(1) 7×7 Depthwise Convolution 利用大感受野捕捉局部上下文关系：

$$\boldsymbol{X}_1=\text{DWConv}_{7\times7}(\boldsymbol{X})\in\boldsymbol{R}^{B\times C\times H\times W}.\tag{2}$$

(2) LayerNorm 提升训练稳定性并适配更深

网络：

$$\boldsymbol{X}_2 = \mathrm{LayerNorm}(\mathrm{Permute}(\boldsymbol{X}_1)) \in \boldsymbol{R}^{B\times H\times W\times C}. \tag{3}$$

（3）Pointwise Conv（1×1）＋GELU 激活＋再 Pointwise Conv：

$$\boldsymbol{X}_3 = \mathrm{Conv}_{1\times 1}\Big(\mathrm{GELU}\big(\mathrm{Conv}_{1\times 1}(\boldsymbol{X}_2)\big)\Big) \in \boldsymbol{R}^{B\times H\times W\times C}. \tag{4}$$

（4）Permute 回原始维度＋残差连接：

$$\boldsymbol{Y} = \mathrm{Permute}^{-1}(\boldsymbol{X}_3) + \boldsymbol{X} \in \boldsymbol{R}^{B\times C\times H\times W}. \tag{5}$$

每个 Block 内部去除了残差连接，同时采用 LayerNorm 替代传统的 BatchNorm，并使用 GELU 激活函数，从而兼顾性能与推理效率，适用于嵌入式与边缘计算平台。

4.1.2　位置感知增强单元

在跨模态遥感图像分析任务中，由于卫星与无人机图像的视角差异（如倾斜拍摄与正射投影）及分辨率不匹配，空间位置信息的显式建模对几何对齐至关重要。为此，本文提出一种轻量化的位置感知增强单元（Position-Aware Unit，PAU），通过可学习的位置编码强化主干网络对空间结构的敏感性。该模块设计如下：

4.1.2.1　位置编码生成

给定 ConvNeXt 骨干网络输出的特征图 $\boldsymbol{F}\in\boldsymbol{R}$，　$B\times C\times H\times W$，PAU 首先生成一组与特征图尺寸匹配的可学习位置编码 $pos\_embed\in\boldsymbol{R}^{C\times H\times W}$。其初始化采用零均值高斯分布（$\sigma$=0.02），并通过反向传播动态优化，以自适应捕获输入图像的几何先验。

4.1.2.2　自适应融合机制

为避免位置编码对原始特征的过度干扰，引入可训练的比例因子 $pos\_scale\in\boldsymbol{R}^{+}$ 控制融合强度：

$$\boldsymbol{F}' = \boldsymbol{F} + pos\_scale \cdot pos\_embed, \tag{6}$$

其中，pos_scale 初始化为 10^{-4}，通过梯度下降逐步调整，平衡位置信息与语义特征的贡献。

4.2　跨模态注意力增强模块

为有效适应卫星图与无人机图像在尺度、视角与纹理结构等方面的跨模态差异，本文引入跨模态注意力机制（Cross-Modal Attention Module），以增强关键区域语义对齐能力。核心思想是利用特征图不同维度间的相互依赖性，通过构造双路径 Triplet Attention 模块，对特征进行显式建模，同时结合隐式空间建模机制提升区域表达鲁棒性。

4.2.1　双路径 Triplet Attention 机制

传统的 SE，CBAM 等注意力机制多聚焦于通道或空间维度，难以有效捕捉跨模态图像中由成像角度差异所引发的方向性特征偏移。为此，本文提出双路径 Triplet Attention（TA）结构，分别在通道-宽度（CW）与高度-通道（HC）两个正交维度上建模依赖关系，从而提升空间与语义特征的耦合表达能力。

4.2.1.1　CW 路径（Channel-Width Attention）

（1）输入特征图 $\boldsymbol{F}\in\boldsymbol{R}^{B\times C\times H\times W}$ 维度置换为：

$$\boldsymbol{F}_{\mathrm{cw}} = \mathrm{permute}(\boldsymbol{F}, [B, H, C, W]). \tag{7}$$

（2）使用 Z-Pooling 操作对通道维进行最大池化与平均池化后拼接，生成中间注意力图 $\boldsymbol{M}_{\mathrm{cw}}$：

$$\boldsymbol{M}_{\mathrm{cw}} = \mathrm{Concat}\big[\mathrm{AvgPool}(\boldsymbol{F}_{\mathrm{cw}}), \mathrm{MaxPool}(\boldsymbol{F}_{\mathrm{cw}})\big]. \tag{8}$$

（3）然后通过 7×7 深度卷积（Depth-wise Conv）提取横向空间依赖，激活后进行维度还原：

$$\boldsymbol{A}_{\mathrm{cw}} = \sigma\big(\mathrm{DWConv}_{7\times 7}(\boldsymbol{M}_{\mathrm{cw}})\big). \tag{9}$$

（4）最终生成，输出通道-宽度路径增强特征图 $\boldsymbol{F}'_{\mathrm{cw}}\in\boldsymbol{R}^{B\times C\times H\times W}$：

$$\boldsymbol{F}'_{\mathrm{cw}} = \boldsymbol{F}\cdot\boldsymbol{A}_{\mathrm{cw}}. \tag{10}$$

4.2.1.2　HC 路径（Height-Channel Attention）

同理，将特征图转置为：

$$\boldsymbol{F}_{hc} = \mathrm{permute}(\boldsymbol{F}, [B, W, H, C]). \tag{11}$$

重复与 CW 路径相同的注意力提取流程，得到纵向注意力图 $\boldsymbol{A}_{\mathrm{hc}}$，并经过计算生成高度-通道路径增强特征图 $\boldsymbol{F}'_{\mathrm{hc}}\in\boldsymbol{R}^{B\times C\times H\times W}$。

4.2.2　隐式空间建模能力

除了显式的方向性注意力建模，Triplet Attention 还具有良好的隐式空间建模能力，在不增加显著参数量的前提下，通过通道与空间维度的交替变换与融合，完成对局部几何关系的建模与补全。这种机制对于应对跨模态场景中常见的结构偏移、尺度错配等问题具有显著效果。

具体地，Triplet Attention 模块在完成 CW 路径和 HC 路径注意力计算后，输出两个方向感知增强的特征图 $\boldsymbol{F}'_{\mathrm{cw}}$ 和 $\boldsymbol{F}'_{\mathrm{hc}}$。

为了融合两个方向的结构感知特征，本文采

用加权融合策略，公式如下：

$$F_{\text{att}} = \lambda_1 \cdot F'_{\text{cw}} + \lambda_2 \cdot F'_{\text{hc}}, \tag{12}$$

其中：$\lambda_1 + \lambda_2 = 1$，$\lambda_1$，$\lambda_2$为可学习或预设的融合权重（如$\lambda_1 = \lambda_2 = 0.5$），$F^{\text{att}} \in \boldsymbol{R}^{B \times C \times H \times W}$表示融合后的注意力增强特征图。

随后，该融合特征F^{att}与原始主干特征图F进行残差式合并：

$$F^{\text{final}} = F + F^{\text{att}}. \tag{13}$$

该结构不仅保留了原始特征的底层信息，还引入了方向性依赖建模后的空间几何增强表达，使得下游匹配或分类任务具备更强的跨模态鲁棒性。

4.3 多任务联合优化模块

为提升跨模态图像匹配中的判别能力与鲁棒性，本文提出基于三重损失联合优化框架的多任务学习机制（如图所示），融合分类监督信号、特征对齐约束、互信息正则化等目标，系统性引导网络学习视角不变的判别性嵌入特征。

4.3.1 分类损失（Cross-Entropy Classification Loss）

针对每一视角输入图像（如卫星、无人机），网络在主干全局分支与注意力细粒度分支上均进行身份分类监督，引导特征学习保持强判别性。设第v视角输出为f_v，其对应预测分类结果为$\hat{y}_v$，则其分类损失定义为：

$$L_{\text{cls}} = \sum_{v=1}^{V} \text{CE}\left(\hat{y}_v, y_v\right), \tag{14}$$

其中：$V = 2$表示视角数，CE为标准交叉熵损失。

4.3.2 三元组损失（Triplet Ranking Loss）

为提升不同模态之间的特征可比性，采用三元组损失进行跨视角对齐。将同一目标在不同视角下的特征视为正样本f^a，f^p，不同目标作为负样本f^n，三元组损失定义为：

$$\mathcal{L}_{\text{tri}} = \sum_{i=1}^{N} \max(\ ), \tag{15}$$

其中：m为设定的间隔（margin），本文默认设为0.3；该损失通过正负对比提升特征结构的分布判别性。

4.3.3 KL散度损失

为缓解不同视角（卫星/无人机）在预测不一致及特征分布偏移，本文在双模态分支的输出层引入KL互学习一致性约束，通过双向知识蒸馏实现分布对齐、互信息引导强化特征一致性，最终达成双模态特征分布的精准对齐且保留各自模态的判别性。

设两分支最终logits分别为$z_{\text{sat}} \in \boldsymbol{R}^K$，$z_{\text{drone}} \in \boldsymbol{R}^K$，$K$为数据集类别数。为避免直接应用*soft*max导致的硬分类概率分布极化（难以传递细粒度特征差异），引入温度系数T对logits进行软化，再通过*soft*max函数转换为连续概率分布$\boldsymbol{p}_{\text{sat}}^{\text{T}} = soft\max\left(\frac{z_{\text{sat}}}{T}\right)$，$\boldsymbol{p}_{\text{drone}}^{\text{T}} = soft\max\left(\frac{z_{\text{drone}}}{T}\right)$，其中$\boldsymbol{p}_{\text{sat}}^{\text{T}}$与$\boldsymbol{p}_{\text{drone}}^{\text{T}}$分别表示卫星图与无人机图的预测概率分布（默认$T = 1$）。

对称KL互学习损失采用双向KL蒸馏（避免单向蒸馏导致的模态信息失衡），并按知识蒸馏常规乘以T^2进行梯度补偿（抵消温度对梯度的缩放效应，保证更新稳定性，文献[15]）；同时引入互信息引导机制，利用KL散度度量双模态输出分布的差异，推动概率分布趋于一致。

$$\mathcal{L}_{\text{KL}} = T^2\left[D_{\text{KL}}\left(\boldsymbol{p}_{\text{sat}}^{\text{T}} /\!/ \boldsymbol{p}_{\text{drone}}^{\text{T}}\right) + D_{\text{KL}}\left(\boldsymbol{p}_{\text{drone}}^{\text{T}} /\!/ \boldsymbol{p}_{\text{sat}}^{\text{T}}\right)\right], \tag{16}$$

其中：$D_{\text{KL}}(P /\!/ Q) = \sum_{c=1}^{K} P(c) \log \frac{P(c)}{Q(c)}$为KL散度的定义，用于量化分布$P$与$Q$的相对熵差异，若两分布完全一致，KL散度值为0；若差异越大，值越大，损失函数通过反向传播推动双分布向差异最小方向优化。

4.3.4 总损失函数组合

将上述三类损失组合，形成最终优化目标：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{tri}} \cdot \mathcal{L}_{\text{tri}} + \lambda_{\text{KL}} \cdot \mathcal{L}_{\text{KL}}, \tag{17}$$

其中超参数λ_{tri}，λ_{KL}控制各损失比重，用于控制三元组损失与KL互学习项在总体优化中的比重。除特别说明外，本文默认取$\lambda_{\text{tri}} = \lambda_{\text{KL}} = 1$，为评估其对精度的影响，见表8。

该多任务优化机制有效结合分类判别与跨模态对齐双重目标，显著提升模型在复杂跨视角环境下的鲁棒性与泛化能力。

5 实验设计与结果分析

为验证本文提出的融合KL-注意力机制的ConvNeXt框架在卫星与无人机跨模态图像匹配任务中的有效性，在标准跨视角数据集University-1652上开展了系统性实验。实验评估涵盖模

型性能、模块贡献分析、与主流方法的对比、以及匹配结果可视化。

5.1　实验设置

5.1.1　数据集介绍

为验证所提融合KL-注意力的ConvNeXt匹配方法在跨模态场景下的有效性，本文在主流跨视角图像匹配数据集University-1652上进行评估。该数据集包含多个城市的卫星图与无人机图，涵盖城市建筑、交通、住宅等多种场景，具备明显的尺度、视角与模态差异。具体包括：

（1）训练集：701类目标，每类包含多幅无人机与对应卫星图像；

（2）测试集：作为检索库的Satellite Gallery与待查询的Drone Query集构成；

（3）评估任务：基于检索精度衡量匹配性能，查询无人机图像在卫星图库中定位其对应区域。

5.1.2　实验环境与参数设置

实验在NVIDIA RTX 3060 GPU平台上进行，所用框架为PyTorch 2.4.1，训练关键参数设置如表2所示。

表2　关键参数设置

Tab. 2　Key Parameter Settings

参数名称	参数值
Batch size	8
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	0.05
Triplet margin m	0.3
Positional scale α	1×10^{-4}
Epochs	120
Scheduler	CosineAnnealingLR
Precision	Mixed Precision (AMP)

5.2　评估指标

为全面评估模型的跨模态匹配能力，本文采用以下主流指标：

（1）Recall@K：在Top-K预测中匹配到正确图像的比例；

（2）AP（Average Precision）：整体匹配准确性；

（3）Top-1 Accuracy：用于测试集身份识别准确率。

5.3　消融实验

为进一步验证本文所提出模块的有效性，特别是双路径Triplet Attention机制与位置感知增强模块在提升跨模态匹配性能方面的贡献，本文在University-1652数据集上进行了系统的消融实验分析。所有实验均基于ConvNeXt-Base主干网络，在相同训练配置（如batch size、学习率、优化器等）下进行控制变量分析，确保评估结果的可比性。

为验证所提方法中各组件的有效性，本文设计消融对比实验，实验结果见表3。

表3　消融实验性能对比（双向Recall@1/Top1%）

Tab. 3　Ablation study performance comparison (Bidirectional Recall@1/Top1%)

算法版本	Drone→Satellite	Satellite→Drone
ConvNeXt+TA+PAU+KL(Full)	89.40/97.38	93.44/99.43
w/o TA	85.27/95.66	88.63/97.74
w/o PAU	86.42/95.88	89.17/97.90
w/o KL	85.85/95.21	88.26/97.34

从双向消融实验结果可以看出，各模块对跨模态匹配性能均有不同程度的贡献，具体分析如下：

Triplet Attention模块在两个方向上的表现提升最为显著。引入该模块后，Recall@1平均提升约3.5 pp（Percentage Point），表明其在建模图像中局部空间依赖和方向性特征方面具有关键作用，显著增强了模型对尺度变化与结构错位的适应能力。

位置感知增强模块虽结构轻量、计算开销极低，但通过显式建模空间位置信息，有效提升了特征的空间对齐能力。在不同模态间仍能保持稳定的Recall提升，验证了其在跨视角对齐中的实用性与必要性。

KL 散度对齐机制有助于引导卫星与无人机两个分支学习一致的判别性特征分布，特别是在 drone→satellite 的反向匹配任务中，表现出更明显的提升效果。这说明 KL 互学习机制可作为一种有效的模态一致性约束手段，增强跨模态对齐的泛化能力。

总体而言，模型在 Satellite→Drone 与 Drone→Satellite 两个方向上的性能差距较小，表明所提出的方法具备良好的跨模态鲁棒性与视角对称性，具备在实际双向地理定位任务中部署应用的潜力。

5.4　多重采样参数 k 对性能的影响分析

为探究多重采样参数 k 对跨模态匹配性能的影响，本文在 University-1652 数据集上分别在 Satellite→Drone 与 Drone→Satellite 两个方向下进行对比实验。实验固定批大小为 8，调整 $k\in\{1,2,3,4\}$，其含义为：在每次迭代中，随机选取 k 张同类别的卫星图像，并配对采样对应的不同角度无人机影像，以构成正样本对。

实验结果如表 4，表 5 和图 2 所示，随着 k 的增加，Recall@1，Recall@5，Recall@10 以及 AP 均呈现缓慢下降的趋势。这一现象在两个任务方向中均较为一致：

Satellite→Drone 方向：当 k 从 1 增至 4 时，Recall@1 从 93.44% 下降至 91.58%，AP 从 89.33% 下降至 88.47%。

Drone→Satellite 方向：当 k 从 1 增至 4 时，Recall@1 从 89.40% 下降至 88.39%，AP 从 90.98% 下降至 90.08%。

表 4　Satellite→Drone 任务中 k 值对性能的影响

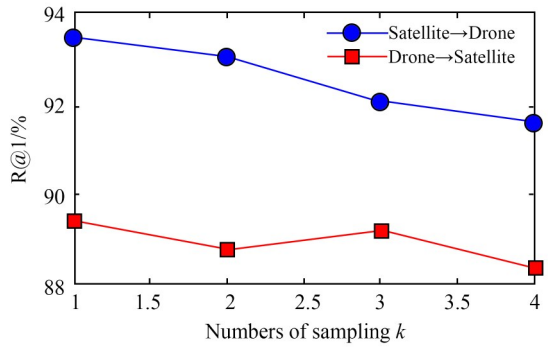
Tab. 4　Impact of k value on performance in Satellite → Drone task

	R@1	R@5	R@10	Top1	AP
k=1	93.44%	95.44%	96.01%	99.43%	89.33%
k=2	93.01%	94.72%	95.72%	98.86%	89.63%
k=3	92.01%	94.72%	95.44%	99.00%	88.61%
k=4	91.58%	94.15%	95.15%	98.86%	88.47%

表 5　Drone→Satellite 任务中 k 值对性能的影响

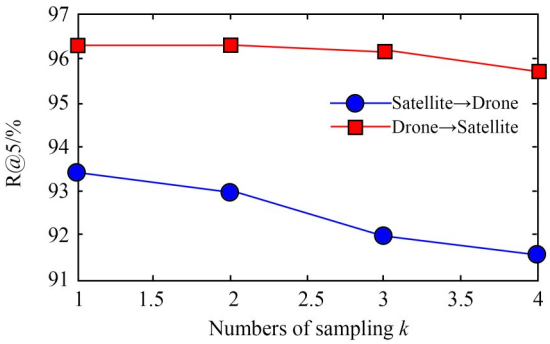
Tab. 5　Impact of k value on performance in Drone→Satellite task

	R@1	R@5	R@10	Top1	AP
k=1	89.40%	96.30%	97.29%	97.38%	90.98%
k=2	88.80%	96.31%	97.28%	97.41%	90.52%
k=3	89.20%	96.17%	97.18%	97.30%	90.80%
k=4	88.39%	95.75%	96.84%	96.99%	90.08%



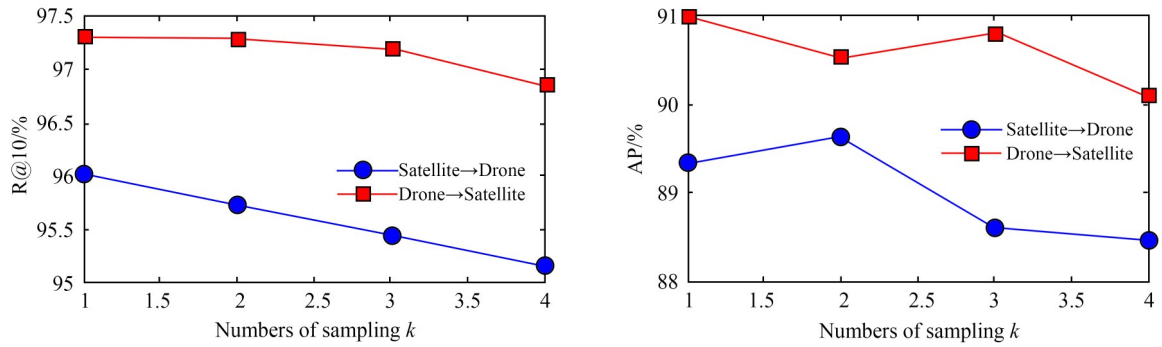
(a) k对两任务R@1的影响对比

(a) Comparison of the impact of k on R@1 for two tasks



(b) k对两任务R@5的影响对比

(b) Comparison of the impact of k on R@5 for two tasks



(c) k对两任务R@10的影响对比
(c) Comparison of the impact of k on R@10 for two tasks

(d) k对两任务AP的影响对比
(d) Comparison of the impact of k on AP for two tasks

图 2　k对两任务 R@K 和 AP 的影响对比

Fig. 2　Comparison of the impact on R@K and AP for two tasks

这一趋势的原因在于，较大的 k 值会增加批次中同类别样本的比例，从而减少了批内跨类别负样本的多样性，使得模型在 Triplet 损失和分类损失下学习到的类间判别特征不足，容易过拟合于局部视角的相似样本。当 k 过小（如 k=1）时，批内正样本数量有限，虽有助于提升类间分离，但可能限制了模型对类内多样性的建模能力。

综合两种任务方向的性能表现与稳定性，本文最终在主实验中选择 k=1 作为默认采样参数，以平衡类内多样性与类间判别性，从而获得更优的整体性能。

5.5　与主流算法对比

与现有 SOTA 方法在 University-1652 数据集上的匹配性能对比如下：

传统度量学习方法（Contrastive/Triplet，基于 VGG16/ResNet-50）整体受限：在 Drone→Satellite 方向，R@1 约为 52%～66%；在 Satellite→Drone 方向，R@1 约为 64%～75%，AP 多在 52%～63% 区间。引入特征增强/融合（如 GeM，LCM，LPN）后，检索效果显著提升：Drone→Satellite 的 R@1 提升到 66%～76%；Satellite→Drone 的 R@1 提升到 80%～86%，AP 提升到 65%～75%。Transformer 系方法（FSRA，TransFG，GeoFormer，MFJRLN，IFSs）进一步推高精度：Drone→Satellite 的 R@1 达到 86%～89%，AP 达到 88%～91%；Satellite→Drone 的 R@1 达到 91%～93%、AP 达到 85%～89%。

本文方法（ConvNeXt＋KL 注意力）在双向均取得最优：Drone→Satellite 方向 R@1＝89.40，AP＝90.98（较 GeoFormer 的 89.08/90.83 分别提升＋0.32/＋0.15）；Satellite→Drone 方向 R@1＝93.44，AP＝89.33（较 TransFG 的 R@1＝93.37 提升＋0.07；较 GeoFormer 的 AP＝88.54 提升＋0.79）。

尽管本文方法在整体性能上优于多数现有方法，但与部分近年提出的代表性最新算法（如 MCCG[40]，SDPL[41] 与 Sample4Geo[39]）相比仍存在一定差距。其中，Sample4Geo 基于中-高容量的 ConvNeXt-B 骨干，并结合大规模负样本挖掘与多阶段优化策略，在训练阶段带来显著的计算与存储开销；而 MCCG 与 SDPL 分别引入多分类聚合机制与密集分区特征学习结构，在复杂网络配置下取得更高的 R@1 与 AP。相较而言，本文方法在实时约束下仍保持相近的精度表现，体现出在精度-效率平衡与工程可部署性方面的优势。

综上，在 University-1652 上，本文方法在 Drone→Satellite 与 Satellite→Drone 两个方向同时取得更高的 R@1 与 AP，体现出更强的跨模态检索精度与稳健性。

5.6　匹配可视化分析

5.6.1　本方法可视化分析

为了进一步验证所提融合 KL-注意力 Con-

vNeXt方法在真实场景下的跨模态匹配效果，本文在University-1652数据集中进行了图像检索可视化实验。图3展示了本文方法在两个典型任务中的Top-5检索结果，黄框表示真实匹配，蓝框表示误匹配，图示仅展示Top-5候选（彩图见期刊电子版）。

(a) 目标定位任务
(a) Target localization task

(b) 导航任务
(b) Navigation task

图3　本方法匹配可视化图

Fig. 3　Matching result visualization diagram of this method

5.6.1.1　无人机视角目标定位任务

随机选取3张无人机视角图像作为查询图，在卫星图库中返回Top-5候选。可以看到，每个查询的Top-5列表均包含真实对应的卫星图（黄框标注，Top-5命中率＝100%）。需要说明的是，由于University-1652每类仅含1张卫星图，因此Top-5中只有1张为真值，其余4张为干扰候选（蓝框）—它们与查询在外观或纹理上相似，但并非同一地物。该结果表明，所提方法在Drone→Satellite方向能够稳定地将真值排入前五，符合跨视角检索“命中即成功”的评测标准，体现了良好的检索准确性与鲁棒性。

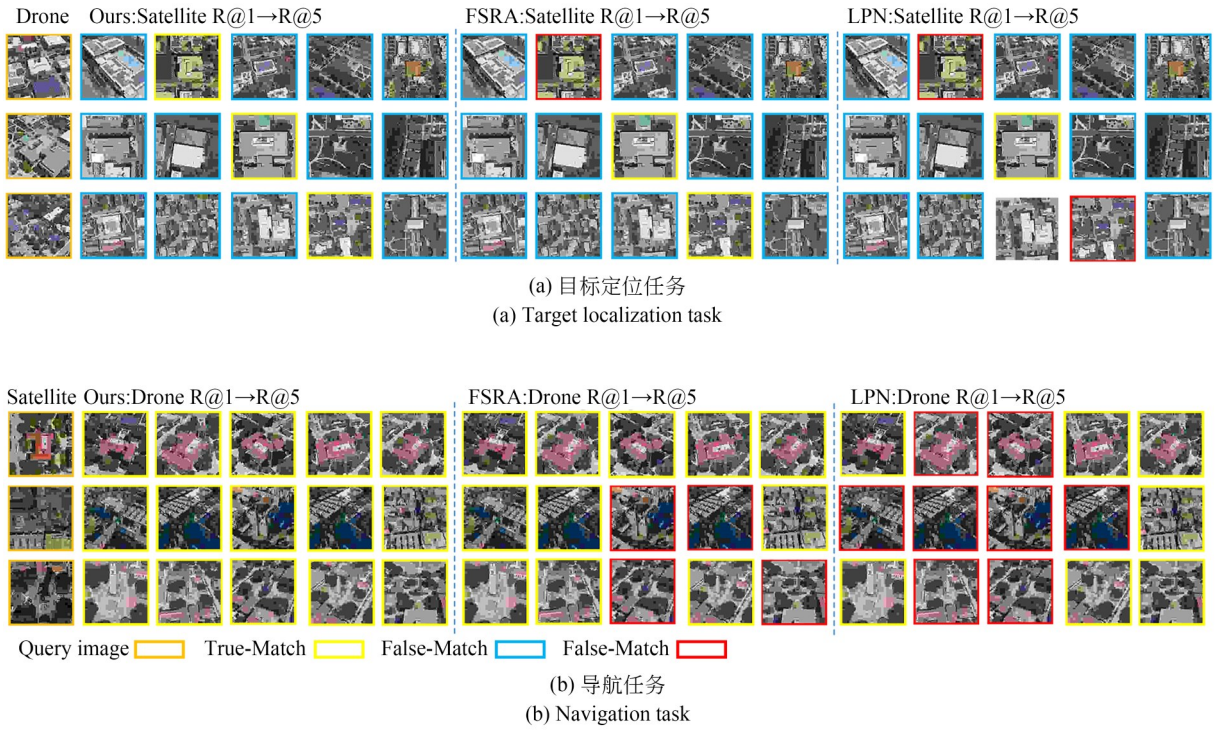
5.6.1.2　无人机导航任务

从测试集中随机选取3张卫星视角图像作为查询图，尝试在无人机图库中匹配出前5个最相似图像。由于每个类别仅含1张卫星图，匹配任务更具挑战性。即便如此，本文方法在该任务下依然成功检索出真实匹配项，表现出对跨模态语义特征的强泛化能力。

图3中，黄色边框表示真实匹配（True-Match），蓝色边框表示误匹配（False-Match）。两类任务的Top-5检索结果均包含真实匹配（Top-5命中率＝100%）；同时，Top-5列表中除真值外仍可能出现相似干扰候选，常见于重复纹理或遮挡区域。本文按通行评测口径，以Top-K命中判定检索成功（即真实目标出现在Top-K内即记为成功），上述结果表明所提方法在跨视角匹配中具有良好的判别能力与鲁棒性。

5.6.2　与FSRA和LPN代表性方法的定性比较

为了进一步说明本文方法的优势，在University-1652数据集上选取若干具有挑战性的UAV和Satellite查询图像，并与代表性方法FSRA与LPN进行了可视化对比。对比结果如图4所示，其中橙色框为查询图像，黄色框为真实匹配，蓝色框/红色框表示误匹配，图中仅展示了Top-5检索候选结果（彩图见期刊电子版）。



(a) 目标定位任务
(a) Target localization task

(b) 导航任务
(b) Navigation task

图4　与FSRA和LPN定性可视化对比结果

Fig. 4　Qualitative visualization comparison results with FSRA and LPN

从图 4 的结果可以观察到：在目标定位任务和导航任务两种场景下，本文方法能够更稳定地将真实匹配排在前列，并显著减少了 Top-5 排序中的误匹配。而 FSRA 与 LPN 在面对复杂场景和相似干扰目标时，常常出现真实匹配排名靠后甚至未命中的情况。这一现象与表 6 中的定量结果相一致，表明本文方法不仅在平均检索精度上优于现有方法，而且在困难场景中具有更好的稳健性与实际应用价值。

表 6　University-1652 上与主流算法的性能对比

Tab. 6　Performance comparison with state-of-the-art algorithms on university-1652　(%)

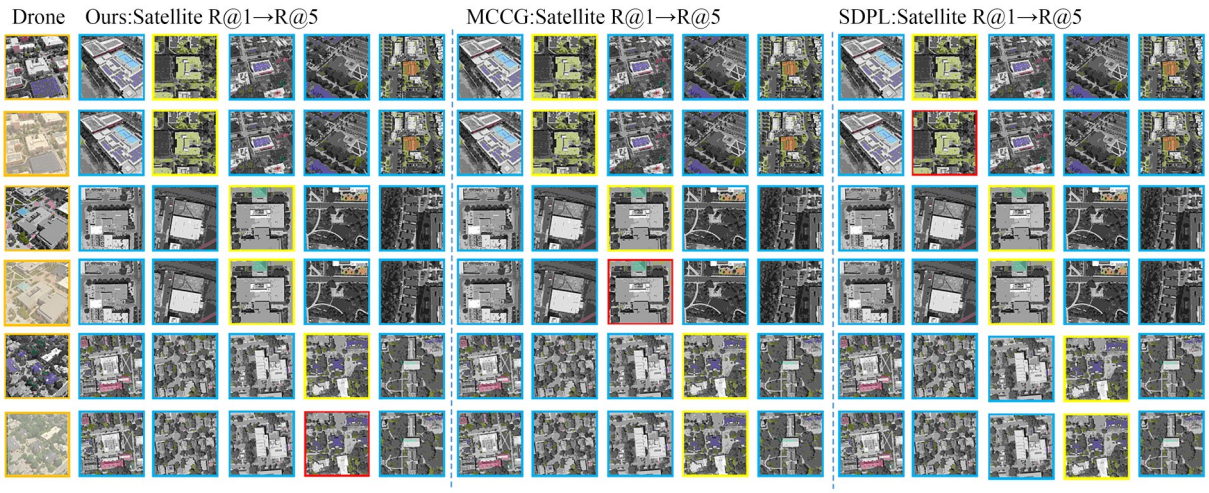
Method	Backbone	Drone → Satellite		Satellite → Drone	
		R@1	AP	R@1	AP
Contrastive loss[22]	VGG16	52.39	52.39	63.91	52.24
Weighted soft margin triplet loss[23]	VGG16	53.21	53.21	65.62	54.47
TripletLoss (M=0.3)[24]	ResNet-50	55.18	59.97	63.62	53.85
Instance Loss+GeM pooling[25]	ResNet-50	65.32	69.61	79.03	65.35
Instance loss[26]	ResNet-50	58.23	62.91	74.47	59.45
LCM (ResNet-50)[27]	ResNet-50	66.65	70.82	79.89	65.38
LPN[28]	ResNet-50	75.93	79.14	86.45	74.79
FSRA[29]	ViT-S	85.50	87.53	89.73	84.94
IFSs[34]	—	86.06	88.08	91.44	85.73
TransFG[35]	ViT-S	87.92	89.99	93.37	87.94
GeoFormer[36]	E−Swin-L	89.08	90.83	92.30	88.54
MFJRLN[37]	Swin-B	87.61	89.57	91.07	86.72
SFDAM[38]	—	88.78	90.51	93.44	88.07
Sample4Geo[39]	ConvNeXt-B	92.65	93.81	95.14	91.39
MCCG[40]	ConvNeXt(256)	89.28	91.01	94.29	89.29
SDPL[41]	SwinV2-B	90.16	91.64	93.58	89.45
Ours	ConvNeXt	**89.40**	**90.98**	**93.44**	**89.33**

5.6.3　与 MCCG 和 SDPL 的定性可视化对比

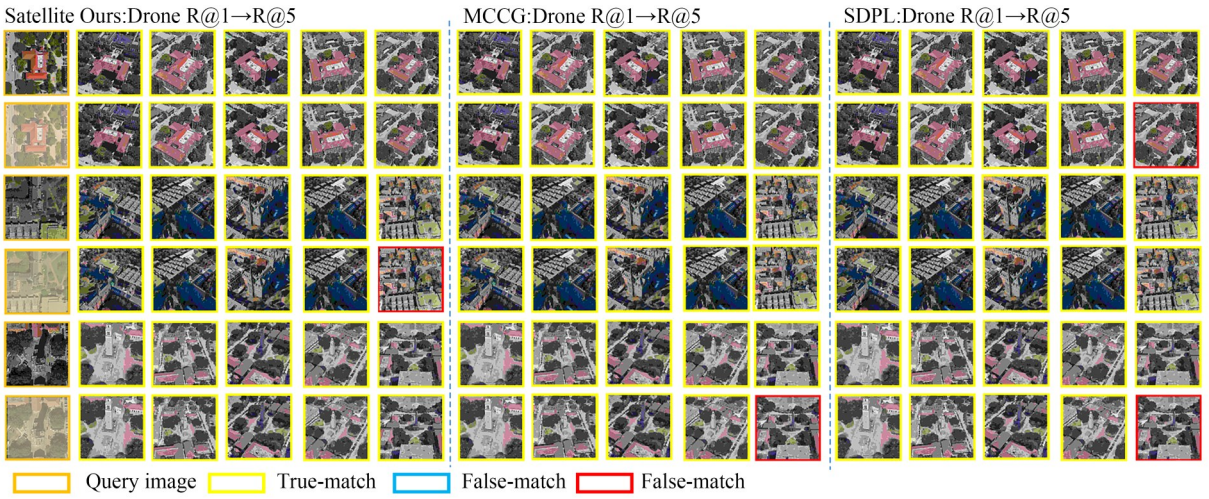
为进一步验证本文方法与近年来代表性算法之间的性能差异，选取 MCCG[40]与 SDPL[41]两个公开源码的算法进行定性对比。对比结果如图 5 所示，其中橙色框为查询图像，黄色框为真实匹配，蓝色/红色框表示误匹配，图中仅展示了 Top-5 检索候选结果(彩图见期刊电子版)。

从图 5 可观察到，在正常光照条件下，本文方法在 Drone→Satellite 与 Satellite→Drone 两个方向上均能稳定检出真实匹配，且显著减少了误匹配数量，充分体现了本方法在跨模态一致性建模与几何稳健性方面的优势。

在进一步引入阴影变化的测试中，三种算法均出现了部分匹配失败或真实匹配排名下降的情况，这表明光照与阴影变化仍会对跨模态检索性能造成一定影响，尤其当阴影覆盖关键结构区域时，纹理对比度减弱，导致局部表征不稳定。尽管如此，本文方法在多数样本中仍能保持较高命中率，说明其结构表征具有一定的光照适应性。后续工作可结合光照不变特征提取或多模态增强机制，以进一步提升对复杂光照环境的鲁棒性。



(a) 目标定位任务
(a) Target localization task



(b) 导航任务
(b) Navigation task

图 5　与 MCCG 和 SDPL 定性可视化对比结果

Fig. 5　Qualitative visualization comparison results with MCCG and SDPL

5.7　效率与轻量化基线评估

为评估在统一端到端协议下的精度-效率权衡，选取若干高效骨干进行对比（研究目标是在实时约束下获得更高检索精度，并非追求最小参数量）。所有模型均使用 256×256 输入、相同训练/微调策略与相同检索管线（单次前向、无 TTA、固定 ANN 配置与固定特征维度）。效率指标在两种 batch 设置下测量：bs=1（报告延迟与 FPS）与 bs=8（报告吞吐量）；PeakMem 为 CUDA 端峰值显存，按“bs=1/bs=8（GB）”给出。参与对比的方法包括 MobileViT-S（cvnets_in1k），TF-EfficientNet-B0，RepVGG-A0，ConvNeXt-Tiny 与 Ours，详见表 7。需要强调的是，本文方法并不以最小参数量为目标，而是在约 48 FPS，<1 GB 显存的实时约束下，提供与表 6 相一致的高精度表现。

从表 7 可得：在<11 M 参数量范围内，RepVGG-A0 延迟最低 13.02 ms；TF-EfficientNet-B0 与 MobileViT-S 显存占用最低（PeakMem@bs=1 约 0.06 GB，@bs=8≤0.25 GB），吞吐量分别达 490.1 FPS 与 413.6 FPS。在保持实时性的同时，ConvNeXt-Tiny 给出较强基线

表 7　轻量化基线与效率对比

Tab. 7　Comparison of lightweight baselines and efficiency

Method	Params /M	GFLOPs	FPS@bs=1	Latency@bs=1 /ms	PeakMem@bs=1/bs=8 /GB	Throughput@bs=8 /FPS	Drone→Satellite R@1 /%	Drone→Satellite AP /%	Satellite→Drone R@1 /%	Satellite→Drone AP /%
mobilevit_s (cvnets_in1k)	7.00	3.71	49.2	20.33	0.06/0.25	358.7	37.83	42.79	50.36	36.23
tf_efficientnet_b0	7.06	1.05	53.5	18.71	0.06/0.22	413.6	34.54	39.57	46.93	31.29
repvgg_a0	10.88	3.98	76.8	13.02	0.08/0.11	490.1	37.70	42.85	47.79	34.51
convnext_tiny	30.85	11.67	71.6	13.97	0.19/0.26	289.8	87.47	89.37	91.44	86.96
Ours	91.31	40.18	48.0	20.83	0.53/0.61	124.8	89.40	90.98	93.44	89.33

(Drone→Satellite：R@1 87.47/AP 89.37；Satellite→Drone：R@1 91.44/AP 86.96)。Ours 相对 ConvNeXt-Tiny 基线，在两向检索上进一步刷新精度：Drone→Satellite 的 R@1 由 87.47 提高到 89.40(+2.21%)，AP 由 89.37 提升至 90.98(+1.80%)；Satellite→Drone 的 R@1 由 91.44 提升至 93.44(+2.19%)，AP 由 86.96 提升至 89.33(+2.73%)。同时保持 48.0 FPS 与 20.83 ms(bs=1)，峰值显存 0.53/0.61 GB。

综上，本文方法并非参数最小，但在单卡 bs=1 约 48 FPS、峰值显存 0.53 GB 的实时前提下，双向 R@1/AP 均优于 ConvNeXt-Tiny，并与 2024～2025 年代表方法保持竞争力；若延迟/显存极其受限，可选 MobileViT-S/TF-EfficientNet-B0 等更轻模型，以牺牲精度换取更高吞吐；若准确率与实时性并重，本文方案提供更优的综合性价比。所有效率指标均在同一硬件与软件栈、相同推理设置下测得(单次前向、无 TTA、固定 ANN 与特征维度)，保证了对比的公平性。

5.8　超参数 λ_{KL} 和 λ_{tri} 选择分析

超参数 λ_{KL} 和 λ_{tri} 设定，在统一端到端评测协议下对 $(\lambda_{KL},\lambda_{tri})\in\{(0,0.1),(0.2,0.5),(0.5,0.5),(0.5,1),(0.8,1.2),(1,1)\}$ 进行了网格搜索，结果见表 8。

表 8　权重 λ_{KL} 和 λ_{tri} 对检索性能的影响(端到端，%)

Tab. 8　Weights λ_{KL} and λ_{tri} their impact on retrieval performance (end-to-end, %)

λ_{KL}	λ_{tri}	Drone→Satellite R@1	R@5	R@10	Top1	AP	Satellite→Drone R@1	R@5	R@10	Top1	AP
0	0.1	86.51	95.69	97.06	97.23	88.61	92.15	94.86	95.44	99.00	87.20
0.2	0.5	85.95	95.14	96.79	96.96	88.04	91.01	93.72	94.86	99.14	85.05
0.5	0.5	88.89	96.29	97.34	97.43	90.60	92.15	95.44	96.15	98.72	88.52
0.5	1	89.60	96.38	97.37	97.48	91.16	92.87	95.01	95.72	99.43	89.30
0.8	1.2	88.52	96.09	97.19	97.31	90.26	92.30	95.29	96.15	98.86	89.53
1	1	89.40	96.30	97.29	97.38	90.98	93.44	95.44	96.01	99.43	89.33

从表 8 分析可得：

(1)关闭 KL 对齐会显著退化。当 $\lambda_{KL}=0$ 且 $\lambda_{tri}=0.1$ 时，Drone→Satellite 的 R@1=86.51%，较最佳设置低 2.89%；Satellite→Drone 的 R@1=92.15%，较最佳设置低 1.29%。表明 KL 互学习对于跨模态分布对齐是必要的。

(2)适中的权重带来稳定提升。将 λ_{KL} 提升至 0.5～1.0，λ_{tri} 提升至 0.5～1.0，两向 R@k/Top1 与 AP 均稳定上升，收益最显著的区间为 $\lambda_{KL}\in[0.5,1.0]$ 和 $\lambda_{tri}\in[0.5,1.0]$。

(3)过大的度量权重收益趋于饱和。$\lambda_{tri}=1.2$($\lambda_{KL}=0.8$ 和 $\lambda_{tri}=1.2$)，R@1 略有回落(Drone→Satellite 由 89.6% 降至 88.52%；Satellite→Drone 由 93.44% 降至 92.30%)，说明过强的三元组约束会与分类/对齐目标产生竞争。

因此，兼顾八项指标的整体最优与稳定性，

本文默认采用 $\lambda_{\mathrm{KL}}=1$，$\lambda_{\mathrm{tri}}=1$，在该设置下：Drone→Satellite 任务 R@1=89.40%，AP=90.98%，Satellite→Drone 任务 R@1=93.44%，AP=89.33%，为各组合中的最优或并列最优。

5.9　单图多目标匹配实验

为验证本文方法在“同图多目标”情形下的适用性，开展了单幅卫星图上的多目标匹配实验，并同时给出可复核的定量指标与可视化结果。

5.9.1　实验设计

给定一张UAV查询图与其对应的大幅卫星图，在卫星图上生成密集相似度热力图，随后通过候选检测与几何一致性核验得到最终匹配点。具体步骤如下：

(1)热力图生成：以本文融合KL-互学习与三重注意力的ConvNeXt骨干提取密集表征，计算相似度热力图并逐像素取最大值；对融合热图进行7×7高斯平滑($\sigma=2$)以抑制噪声峰。

(2)候选检测：阈值采用 $\max(\tau, Q_{0.975})$，再以半径16像素进行非极大值抑制，选取Top-5候选点。

(3)几何核验：对每个候选卫星图裁剪160×160 patch，与UAV图进行几何一致性核验；通过核验者标记为“验证命中”，并记录内点率作为置信度(默认阈值0.25)。

5.9.2　定位能力指标分析设计

对第 q 个查询的Top-5候选记为 $\mathcal{C}_q^{(K)}=\{c_{q,i}\}_{i=1}^{K}$，$c_{q,i}=(x_{q,i}, y_{q,i}, s_{q,i}, ok_{q,i}, r_{q,i})$，其中 $(x_{q,i}, y_{q,i})$ 为候选在卫星图上的像素坐标，$s_{q,i}\in[0,1]$ 为相似度分数，$\mathrm{ok}_{q,i}\in\{0,1\}$ 为几何一致性核验是否通过，$r_{q,i}\in[0,1]$ 为对应的内点率。

为衡量有无至少一个“几何一致”的候选出现在每个查询的Top-K中，定义V-Rate@K如公式(18，反映系统“每张图至少找到一个可信位置”的能力，适合同图多目标定位的可用性评估。

$$v_q^{(K)}=\boldsymbol{I}\left(\sum_{i=1}^{K} ok_{q,i}\geqslant 1\right),\ \boldsymbol{V}-\mathrm{Rate@}\boldsymbol{K}=\frac{1}{\boldsymbol{Q}}\sum_{q=1}^{Q} v_q^{(K)}. \tag{18}$$

衡量在所有Top-K候选中，有多少比例能通过几何核验，定义Verified@K如公式(19)，反映候选列表的纯度/精确性；惩罚把“看起来像但不一致”的候选排进前列。

$$\mathrm{Verified@}K=\frac{1}{\boldsymbol{QK}}\sum_{q=1}^{Q}\sum_{i=1}^{K} o\,k_{q,i}. \tag{19}$$

衡量平均每个查询在Top-K中有多少个候选能通过核验，反映方法在多目标场景下返回多少个可靠实例(而不仅仅是否存在)，定义V-Count@K，公式如(20)：

$$\boldsymbol{V}\text{-}\mathrm{Count@}K=\frac{1}{\boldsymbol{Q}}\sum_{q=1}^{Q}\sum_{i=1}^{K} o\,k_{q,i}. \tag{20}$$

衡量每个查询在Top-K中最可信的一次几何匹配的强度(内点率)，反映空间几何的一致性质量，数值越高意味着姿态/结构更匹配、对应关系更稳健，定义Best-Ratio@K，公式如(21)：

$$b_q^{(K)}=\max_{1\leqslant i\leqslant K}\left(ok_{q,i}\cdot r_{q,i}\right),$$
$$\mathrm{Best\text{-}Ratio@}K=\frac{1}{\boldsymbol{Q}}\sum_{q=1}^{Q} b_q^{(K)}. \tag{21}$$

5.9.3　实验结果与分析

在University-1652的Drone→Satellite设置下，在单图多目标场景实验数据统计如表9所示。

表9　单图多目标匹配的几何核验指标

Tab.9　Geometric verification metrics for single-image multi-target matching

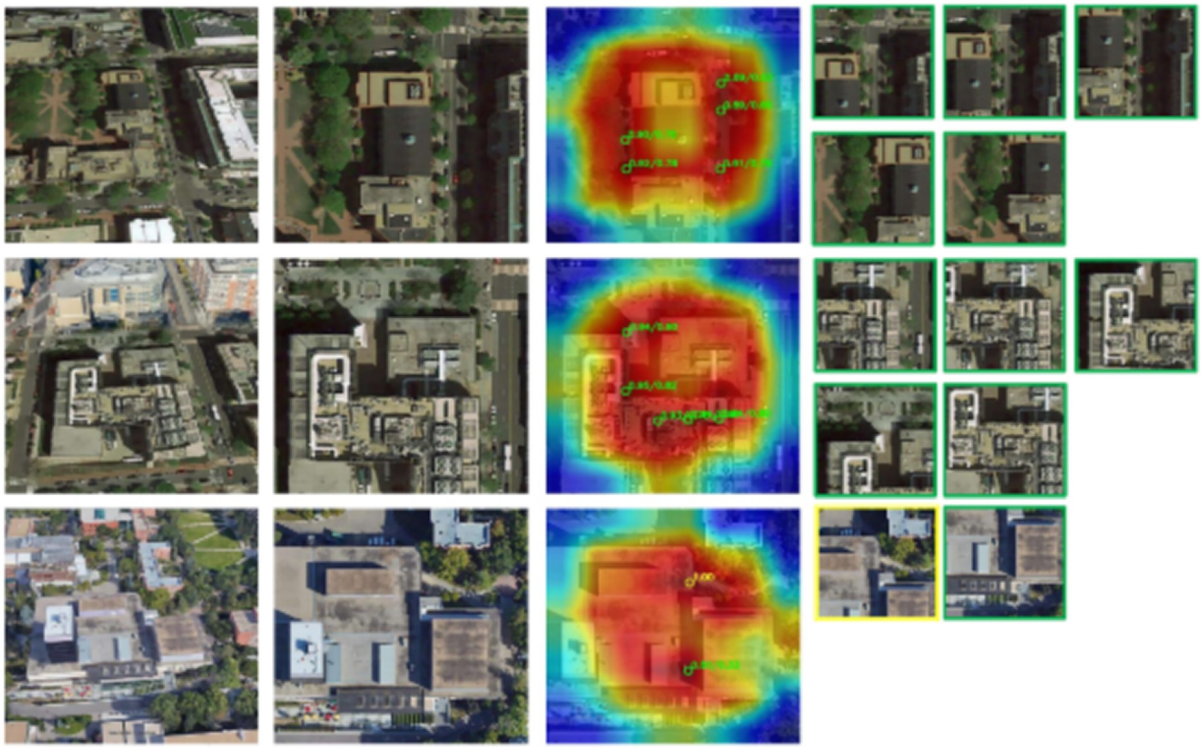
K	V-Rate@K/%	Verified@K/%	V-Count@K	Best-Ratio@K/%
1	82.3	82.3	0.823	31.0
3	96.4	80.8	2.425	41.0
5	97.7	69.4	3.472	43.4

随着候选数K的增大，查询级通过率V-Rate@K从82.3%提升至97.7%，表明系统更易在同图内命中至少一个几何一致的目标。与此同时，V-Count@K由0.823上升至3.472，显示出在多实例场景下系统能够找到更多通过几何校验的匹配；但候选级通过率Verified@K从82.3%降至69.4%，说明候选集合扩展会引入更多边缘候选，从而降低通过占比。Best-Ratio@K从31.0%升至43.4%，提示每个查询的最优几何匹配强度稳定且略有提升。

图6中(a)UAV查询图；(b)对应卫星区域；(c)相似度热力图与候选(数值=相似度/内点率)；(d)Top-5卫星裁剪块(绿框：通过几何核验，黄框：未通过)。在同一幅卫星图中，即使存在多

处相似特征，算法仍能在热力图峰值附近稳定检出多个通过几何核验的目标；未通过的候选则多位于重复纹理或阴影/树冠遮挡区域。

综上分析，在单图多目标场景下，本文方法同一卫星图内稳定命中并返回多个经几何核验的位置，对重复纹理与遮挡等干扰具有一定抑制能力，体现出良好的同图多目标定位能力与几何稳健性，并具备直观可视化与工程可用性。



(a) UAV查询图 (a) Query (UAV)　(b) 对应卫星图 (b) Satellite image context　(c) 相似度热力图 (c) Similarity heatmap　(d) 卫星图裁剪块 (d) Satellite image crops

图 6　单图多目标跨模态匹配的可视化结果

Fig. 6　Qualitative results of single-image multi-target cross-modal matching

6　讨论与分析

6.1　方法优势

本文提出的融合 KL-注意力机制的 ConvNeXt 方法，在跨模态卫星-无人机图像匹配任务中展现出显著优势。通过 ConvNeXt 主干的强大特征建模能力，结合三重注意力模块在空间、通道和方向维度上的多层次信息融合，有效增强了模型对复杂几何变换和大尺度视角变化的适应性。同时，KL 散度互学习机制通过对齐两个模态分支的特征分布，进一步缓解了模态不一致问题。这些设计不仅提升了模型的判别能力和泛化性，使得模型在多个主流评测指标上超越了现有 SOTA 方法，也保持了较高的计算效率，具有实际部署的可行性。

6.2　模型局限性

尽管本方法在标准数据集上取得优秀表现，但依然存在一些不足。首先，在极端视角偏差、遮挡严重或极低分辨率的影像场景下，模型识别和匹配能力仍有一定下滑，表现出对长尾分布场景的鲁棒性局限。其次，尽管 ConvNeXt 计算效率较高，但多任务优化和三重注意力模块在资源较紧张的嵌入式平台部署时，仍需进一步轻量化优化。此外，模型在真实复杂环境中的泛化能力，如不同行业数据、不同国家城市的场景适应性，仍需进一步系统验证。

6.3　未来工作展望

后续工作可从以下几方面展开：一是进一步

优化模型结构，探索注意力模块和KL一致性机制的轻量化与高效实现，提升实际部署能力；二是尝试引入多源数据（如激光雷达、高分光学、道路网络等）进行多模态联合匹配，提升模型对复杂区域的适用性和鲁棒性；三是利用弱监督、半监督等方法减少人工标注依赖，拓展方法在大规模实际应用中的适应范围。此外，可探索模型在灾害应急、公共安全、精细城市治理等多样化实际应用场景的推广与落地，进一步提升视觉导航与地理定位技术的社会和工程价值。

7　结　论

针对卫星与无人机影像间的大尺度、跨模态匹配挑战，本文提出一种融合KL散度与三重注意力机制的ConvNeXt端到端匹配方法。该方法以层次化ConvNeXt为骨干网络提取多尺度特征，结合三重注意力机制以增强模型对方向性几何信息的感知能力，并通过KL散度损失对齐跨模态特征分布，有效提升了模型的匹配精度与鲁棒性。

在University-1652数据集上，所提方法在双向检索任务中均取得较高精度：Drone→Satellite方向的R@1/AP为89.40%/90.98%，Satellite→Drone方向为93.44%/89.33%。相较于强基线ConvNeXt-Tiny，其性能提升如下：Drone→Satellite任务的R@1/AP分别提升1.93%/1.61%（相对提升约2.21%/1.80%），Satellite→Drone任务的R@1/AP分别提升2.00%/2.37%（相对提升约2.19%/2.73%）；与TransFG，GeoFormer等近期代表性方法相比，该方法在双向任务的R@1与AP指标上亦保持小幅领先，在引入2025年方法后，相较**SFDAM**于双向AP与D→S的R@1上更优，整体性能与最新SOTA方法保持同级竞争力。

模块消融实验进一步验证了各关键设计的有效性：在Drone→Satellite方向，三重注意力与KL散度互学习分别为R@1带来4.13%，3.55%的增益（Satellite→Drone方向对应增益约为2.95%，3.32%）；位置感知增强模块同样带来稳定提升，在Drone→Satellite，Satellite→Drone方向分别提升+2.98%，+2.41%。匹配结果可视化表明，Top-5检索结果可稳定命中真值样本，误匹配主要集中于重复纹理场景与遮挡区域。单图多目标实验结果显示，当$K=5$时，V-Rate@5达97.7%，Verified@5为69.4%，V-Count@5为3.472，Best-Ratio@5为43.4%，这表明在单幅卫星图内，该方法可稳定返回多处经几何核验的位置，并有效抑制干扰候选目标。

效率方面，在统一端到端实验协议下，本方法可兼顾推理延迟与数据吞吐量：当bs=1时，延迟为20.83 ms（≈48 FPS），峰值显存占用0.53 GB；当bs=8时，吞吐量达124.8 FPS，峰值显存占用0.61 GB。相较于轻量化骨干网络，本方法以适度的算力开销，换取了更优的双向检索精度与几何稳健性，具备实时部署可行性。

综合来看，所提ConvNeXt+KL-注意力框架在跨模态匹配精度、几何稳健性与实时性能之间实现了良好平衡，在University-1652上达到领先或同级SOTA水平，并展现出面向无人机视觉导航与地理定位任务的工程可部署潜力。未来工作可进一步探索模型轻量化、多源数据融合及半监督学习等方向，以增强方法在更复杂场景中的泛化性与实用性。

作者贡献声明：

程龙：方法的提出，论文构思和撰写；
杨睿光：实验的设计及数据整理和分析；
金鑫：论文审核与编辑写作；
李晓宏：实验数据分析；
孙伟：获取资助，项目管理。

参考文献：

[1] LIU Y, CHEN R, LIU X, *et al*. UAV-based remote sensing for precision agriculture: Recent advances and future trends[J]. *Computers and Electronics in Agriculture*, 2024, 217: 108537.

[2] WANG L, ZHANG Y, LI H, *et al*. A robust gnss/ins/camera integrated vehicle localization method based on adaptive sensor fusion[J]. *IEEE Sensors Journal*, 2023, 23(15): 17325 - 17337.

[3] LIU J, ZHANG X, WANG Y, *et al*. Multi-sensor fusion for GNSS-denied navigation: A UWB and IMU integration approach with adaptive error correction[J]. *Remote Sensing*, 2023, 15(4): 1012.

[4] YANG X, LIU Y, WANG Z, *et al*. Vision-based localization and mapping for unmanned aerial vehicles: A survey[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 23(7): 5895-5917.

[5] ZHANG H, LI X, WANG Z, *et al*. Advances in UAV visual geo-localization with satellite imagery: a comprehensive review and future directions[J]. *Remote Sensing*, 2024, 16(2): 345.

[6] KIM S, PARK J, CHOI Y, *et al*. Visual-inertial navigation for UAVs in GPS-denied environments: recent advances and challenges[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2023, 201: 125-142.

[7] WU Y, LIU T, LIU Z, *et al*. TransVPR: transformer-based visual place recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(1): 145-157.

[8] WANG Y, LI L, LI J, *et al*. Cross-view geo-localization from ground to satellite images via deep learning[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2021, 177: 112-125.

[9] LIU Z, HU H, LIN Y T, *et al*. Swin Transformer V2: scaling up capacity and resolution[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 11999-12009.

[10] CHEN K, SONG Y, WANG C, *et al*. Diffusion Models for semantic segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(4): 2345-2357.

[11] WINKELS M, COHEN T. Equivariant neural networks in practice: challenges and solutions[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(5): 5760-5773.

[12] XU Y, LI Y, ZHANG H, *et al*. MSPNet: Multi-scale pyramidal feature network for geo-localization[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, 59(10): 8754-8766.

[13] DENG J, GUO Z, ZHANG S, *et al*. CVML: cross-view metric learning with attention[C]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *Piscataway*: IEEE, 2021: 11234-11243.

[14] YANG Y, TU W, LI Q, *et al*. Cross-view geo-localization with layer to layer transformer (L2LTR)[C]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway*: IEEE, 2021: 1234-1243.

[15] YANG Y, TU W, LI Q, *et al*. Cross-view geo-localization with evolving transformer (EgoTR)[C]. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (*ICCV*). *Piscataway*: IEEE, 2021: 5678-5687.

[16] SUN J M, SHEN Z H, WANG Y A, *et al*. LoFTR: detector-free local feature matching with transformers[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 20-25, 2021, *Nashville*, *TN*, *USA*. IEEE, 2021: 8918-8927.

[17] CHEN L, ZHANG Y, LI J, *et al*. Weakly-supervised semantic alignment via transformer-based cross-attention[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(3): 1820-1832.

[18] LIU Z, MAO H Z, WU C Y, *et al*. A ConvNet for the 2020s[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 11966-11976.

[19] HU J, WANG X, YANG H, *et al*. DAAM: Dual attention augmentation module for enhanced feature representation[J]. *IEEE Transactions on Image Processing*, 2024, 33: 1524-1535.

[20] LIU H, CHEN M, ZHANG Z, *et al*. Collaborative learning with augmented mutual distillation for visual recognition[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, 34(9): 4321-4333.

[21] ZHENG Z D, WEI Y C, YANG Y. University-1652: a multi-view multi-source benchmark for drone-based geo-localization[C]. *Proceedings of the 28th ACM International Conference on Multimedia. Seattle WA USA*. ACM, 2020: 2339-2347.

[22] DENG W J, ZHENG L, YE Q X, *et al*. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition. June* 18-23, 2018, *Salt Lake City*, *UT*, *USA*. IEEE, 2018: 994-1003.

[23] HU S X, FENG M D, NGUYEN R M H, *et al*. CVM-net: cross-view matching network for image-

based ground-to-aerial geo-localization [C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition. June* 18-23, 2018, *Salt Lake City*, *UT*, *USA*. IEEE, 2018: 7258-7267.

[24] LIU H, FENG J S, QI M B, *et al*. End-to-end comparative attention networks for person re-identification [J]. *IEEE Transactions on Image Processing*, 2017, 26(7): 3492-3506.

[25] RADENOVIC F, TOLIAS G, CHUM O. Fine-tuning CNN image retrieval with No human annotation [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41(7): 1655-1668.

[26] ZHENG Z D, WEI Y C, YANG Y. University-1652: a multi-view multi-source benchmark for drone-based geo-localization [C]. *Proceedings of the 28th ACM International Conference on Multimedia. Seattle WA USA*. ACM, 2020: 429-437.

[27] DING L R, ZHOU J, MENG L X, *et al*. A practical cross-view image matching method between UAV and satellite for UAV-based geo-localization [J]. *Remote Sensing*, 2021, 13(1): 47.

[28] WANG T Y, ZHENG Z D, YAN C G, *et al*. Each part matters: local patterns facilitate cross-view geo-localization [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, 32(2): 867-879.

[29] 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.

[30] 曹栩珩，郝小鹏，廉玉生，等. 用于高光谱-多光谱图像盲融合的自适应退化感知Transformer(封底文章·特邀)[J]. 红外与激光工程, 2025, 54(5): 18-29.
CAO X H, HAO X P, LIAN Y S, *et al*. Degradation-aware transformer for blind hyperspectral and multispectral image fusion (back cover paper·invited) [J]. *Infrared and Laser Engineering*, 2025, 54(5): 18-29. (in Chinese)

[31] 李想，熊凌，叶道辉，等. 多尺度深层特征蒸馏的图像超分辨率重建[J]. 光学 精密工程, 2025, 33(10): 1657-1671.
LI X, XIONG L, YE D H, *et al*. Image super-resolution reconstruction of multi-scale deep feature distillation [J]. *Opt. Precision Eng.*, 2025, 33(10): 1657-1671. (in Chinese)

[32] 刘立波，郝思宇，邓箴. 结合改进ConvNeXt网络与知识蒸馏的天气识别[J]. 光学 精密工程, 2023, 31(14): 2123-2134.
LIU L B, XI S Y, DENG Z. Weather recognition combining improved ConvNeXt models with knowledge distillation [J]. *Opt. Precision Eng.*, 2023, 31(14): 2123-2134. (in Chinese)

[33] 薛珊，陈宇超，吕琼莹，等. 基于坐标注意力机制融合的反无人机系统图像识别方法[J]. 红外与激光工程, 2022, 51(9): 417-427.
XUE S, CHEN Y C, LV Q Y, *et al*. Image recognition method of anti drone system based on coordinate attention mechanism [J]. *Infrared and Laser Engineering*, 2022, 51(9): 417-427. (in Chinese)

[34] GE F W, ZHANG Y Z, LIU Y X, *et al*. Multi-branch joint representation learning based on information fusion strategy for cross-view geo-localization [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 5909516.

[35] ZHAO H, REN K Y, YUE T Y, *et al*. TransFG: a cross-view geo-localization of satellite and UAVs imagery pipeline using transformer-based feature aggregation and gradient guidance [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 4700912.

[36] LI Q G, YANG X G, FAN J W, *et al*. GeoFormer: an effective transformer-based Siamese network for UAV geolocalization [J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024, 17: 9470-9491.

[37] GE F W, ZHANG Y Z, WANG L, *et al*. Multi-level feedback joint representation learning network based on adaptive area elimination for cross-view geo-localization [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 5913915.

[38] 徐海燕，赵艮平，程良伦. 面向无人机视角地理定位的空间频域注意模型[J/OL]. 电光与控制, 2025: 1-9. (2025-10-16). http://kns.cnki.net/KCMS/detail/detail.aspx?filename=DGKQ20251015001&dbname=CJFD&dbcode=CJFQ.
XU H Y, ZHAO G P, CHENG L L. Spatial Frequency Domain Attention Model for UAV Visual Angle Geographic Positioning [J/OL]. *China Industrial Economics*, 2025: 1-9. (2025-10-16). http://kns.cnki.net/KCMS/detail/detail.aspx? file-

name=DGKQ202510150001&dbname=CJFD&dbcode=CJFQ. (in Chinese)
Drone-View Geolocation [J/OL]. *Electronics Optics & Control*, 1-9[2025-10-17].

[39] DEUSER F, HABEL K, OSWALD N. Sample4Geo: hard negative sampling for cross-view geo-localisation[C]. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 1-6, 2023, *Paris*, *France*. IEEE, 2023: 16801-16810.

[40] SHEN T R, WEI Y M, KANG L, *et al*. MCCG: a ConvNeXt-based multiple-classifier method for cross-view geo-localization [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, 34(3): 1456-1468.

[41] CHEN Q, WANG T Y, YANG Z H, *et al*. SDPL: shifting-dense partition learning for UAV-view geo-localization [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, 34(11): 11810-11824.

作者简介：

程　龙(1989—),男,吉林梨树人,博士研究生,高级工程师,2016年于长春工业大学获得硕士学位,就职长光卫星技术股份有限公司,主要从事小卫星总体设计研究和视觉匹配导航技术研究。E-mail: longcheng201304011@163.com