

跨视角图像地理定位数据集综述

张硝, 高艺, 夏宇翔, 赵春雪

61175 部队, 南京 210049

摘要: 跨视角图像地理定位 (Cross-view Image Geo-localization) 旨在通过不同视角图像间的匹配, 检索参考图库中相似度最大的图像, 进而利用其 GPS 标签实现定位功能。传统的单一视角图像地理定位受限于数据集质量、规模以及定位精度等因素, 因而近年来众多研究人员和机构发布了一系列跨视角地理定位数据集, 为提升地理定位精度打下数据基础。然而目前尚缺乏对跨视角图像地理定位数据集的系统性分析。本文首先梳理了跨视角图像地理定位发展以来的 32 个经典数据集, 从视角信息、构建类型、真实程度、时相信息 4 个维度构建分类体系并对数据集基本信息进行归纳总结; 其次从元数据、影响力、关键词、获取来源以及应用领域 5 个方面对跨视角图像地理定位数据集进行深入分析, 整理概括了跨视角图像地理定位目前的主流算法; 最后从数据集多模态趋势、大模型方法、图像干扰物处理以及模型优化 4 个角度探讨了跨视角定位数据集未来的发展方向, 可以为相关领域研究人员提供参考。

关键词: 跨视角, 图像地理定位, 数据集, 深度学习, 无人机, 图像检索, 图像匹配, 计算机视觉

中图分类号: P2

引用格式: 张硝, 高艺, 夏宇翔, 赵春雪. 2025. 跨视角图像地理定位数据集综述. 遥感学报, 29(8): 2511-2530

Zhang X, Gao Y, Xia Y X and Zhao C X. 2025. Review of cross-view image geo-localization datasets. National Remote Sensing Bulletin, 29(8): 2511-2530 [DOI: 10.11834/jrs.20254348]

1 引言

图像地理定位 VPR (Visual Place Recognition) 是估计图像所在的地理位置, 即给定一张未知地点的查询图像, 通过图像检索技术从数据库中检索出与查询图像最具相关性的参考图像, 利用参考图像的 GPS 标签估计查询图像的地理位置 (刘兴, 2023)。该技术在网络空间测绘、情报获取、自动驾驶以及战场态势分析等领域都发挥着重要作用, 尤其在全球导航卫星系统 (GNSS) 受限或不可用的情况下, 可以作为一种有效的地理定位手段, 在民用和军用领域受到了越来越广泛地关注。

跨视角图像地理定位最早起源于机器人导航研究。2008 年, Leung 等 (2008) 首次设计了一套以单目摄像机作为传感器、基于地面视角图像和航拍图像匹配算法的机器人户外定位系统。2013 年, Lin 等 (2013) 第一次在计算机视觉国际

顶级学术会议 CVPR 中提出“跨视角图像地理定位” (Cross-view Image Geo-localization) 的概念, 通常被视为图像地理定位技术在特定场景下的应用与发展。其中, 跨视角指从不同视角 (地面、无人机和卫星) 图像之间检索相似度最高的图像。

一直以来, 跨视角图像地理定位都是一项具有挑战性的任务, 面临的主要问题是学习大视点变化的鲁棒特征。早期的跨视角图像地理定位方法主要依赖于手工特征, 如 SIFT (Lowe, 2004)、SURF (Bay 等, 2006) 和 ORB (Rublee 等, 2011) 等。虽然在一些应用场景下手工特征可以较好地解决图像匹配问题, 但手工特征对光照、视角以及噪声变化鲁棒性较差, 难以提取图像中复杂且具有辨识力的视点不变特征。同时, 手工特征还存在泛化能力差、强依赖于设计者经验等问题, 在实际应用上受到了一定程度的限制。

近年来, 深度学习的快速发展为跨视角图像地理定位提供了新的解决思路, 而依赖于数据驱

收稿日期: 2024-08-27; 预印本: 2025-01-10

第一作者简介: 张硝, 研究方向为摄影测量与遥感, 人工智能。E-mail: 709645587@qq.com

通信作者简介: 高艺, 研究方向为摄影测量与遥感, 人工智能。E-mail: iamninigao@163.com

动的深度学习模型离不开高质量、大规模的数据集。

如表 1 所示，不同视角图像地采集手段和获取途径呈现多平台特点。与此同时，跨视角图像数据集不仅在数量上呈现连年增多的趋势，而且在质量和多样性上也不断优化。如主流数据集 CVUSA（Cross-View USA，美国跨视角数据集）（Workman 等，2015）、CVACT（Cross-view Australian Capital Territory，澳大利亚首府地区跨视角数据集）（Liu 和 Li，2019）等提供了包含卫星和地面两个视角的图像对，通常来自于卫星和机载摄像头或手持摄像头；University-1652 数据集（Zheng 等，2020）提供了包含来自卫星、无人机和地面 3 个视角的数据，来自于卫星、谷歌地球模拟的无人机和街景数据 3 个平台。除此以外，近年来电子游戏中的虚拟世界也成为跨视角数据集的一个重要数据源。

表 1 数据采集平台及获取途径

Table 1 Data acquisition platform and access

视角	采集平台	获取途径
卫星视角	开源平台	OSM
		谷歌地图
		谷歌地球
		Bing 地图
		高德地图
	真实卫星	ArcGIS Maps
		NWPU-RESISCS-45
		高分二号
		哨兵-2
		合成数据
无人机视角	无人机平台	虚拟世界
	直升机平台	无人机载摄像头
	直升机平台	直升机载摄像机
	合成数据	虚拟世界
地面视角	车载摄像机	谷歌街景
		Bing 街景
		自动驾驶数据集
	手持相机	图片分享网站
		个人手机获取
	合成数据	虚拟世界

本文在前人研究的基础上，从广度和深度上进一步拓展，收集整理了 2011 年—2024 年间 32 个跨视角图像地理定位数据集，并从多个角度对数据集进行了分析。首先对该领域数据集进行梳理

和归纳，按照视角信息、构建类型、真实程度以及及时相信息 4 个维度对跨视角定位数据集进行系统性分类，并对每个数据集进行基本信息介绍；然后从元数据、应用领域以及关键词等多个维度对跨视角定位数据集进行分析，同时，对跨视角图像地理定位相关算法进行总结；最后对跨视角地理定位数据集未来发展趋势进行讨论。本文的研究为理解、设计和选择适合特定应用需求的数据集时提供了新的视角，也为后续相关领域的研究提供有价值的参考。

2 跨视角图像地理定位数据集概述

跨视角图像地理定位技术发展至今，涌现了大批相关数据集，面对数量众多的数据集，当前的一些研究多侧重于介绍部分主流数据集，重点关注基于数据集的相关算法和模型。如 Wilson 等（2023）对视觉地理定位进行了全面综述，并概括了流行的算法和数据集；刘棟等（2023）详细介绍并比较了跨视角图像地理定位领域的常见方法；黄高爽等（2023）对图像地理定位研究进展进行了梳理，归纳总结了图像地理定位方法、常见数据集及其评价方法；Durgam 等（2024）全面综述跨视角地理定位领域的前沿方法、技术及挑战，并重点介绍了基于特征和深度学习的策略。目前，尚缺乏对该领域发展至今的数据集系统性、综合性的梳理与总结。

2.1 跨视角图像地理定位数据集分类

本节根据目前主流跨视角图像地理定位数据集的属性特征，从视角信息、构建类型、真实程度以及及时相信息 4 个维度构建分类体系，如图 1 所示。

按图像拍摄的不同视角分为：地面—无人机视角数据集、无人机—卫星视角数据集、地面—无人机视角数据集、地面—无人机—卫星视角数据集，反映了数据集图像所属的视角信息。

按构建类型分为：自建数据集和融合数据集，主要反映了数据集建立的独立程度。自建数据集指从不同平台收集原始数据并处理为符合自身研究需求的新的跨视角数据集，例如 Workman 等（2015）收集了两种视角的图像，其中地面视角的图像来自于 Flickr 平台和谷歌街景平台，一共约 150 万图像，对应的卫星视角图像来自 Bing Map 平台 14、16 和 18 层数据，经去重处理后共约 80 万张

构成自建数据集。融合数据集指的是将多个已有数据集的部分或全部数据拼凑在一起组成新的数据集。如饶子昱等（2023）将谷歌地球 19—20 级郑州地区卫星影像、部分 CVUSA 城市地区卫星影像、部分 Vo & Hays 城区卫星影像、University1652 部分影像及摄取的无人机影像作为空中视角影像，将手持设备采集的目标图像和街景图像作为地面视角图像构建融合数据集。

按真实程度分为：真实数据集、虚拟合成数据集以及真实—虚拟混合数据集。主要反映了数据的真实性，例如 SVA 数据集（Palazzi 等，2017）和 Minecraft 数据集（Yagi 等，2023）均来自游戏中的虚拟世界；University-1652（Zheng 等，2020）数据集中的地面和卫星视角来自真实世界，但无人机视角来自于谷歌地球引擎模拟，属于真实—虚拟混合数据集。

按照时相信息分为：单时相数据集和多时相数据集。主要反映了数据集中图像的时间信息。多时相数据集是图像可能有不同季节或一天中的不同时段，例如 ZHcity750 数据集（Yan 等，2024）中每座目标建筑有 4 张不同时相的卫星图像，SUES-200 数据集（Zhu 等，2023）有雾、雨、雪 3 种不同天气的无人机图像，多时相能有效加强模型的鲁棒性和泛化性。

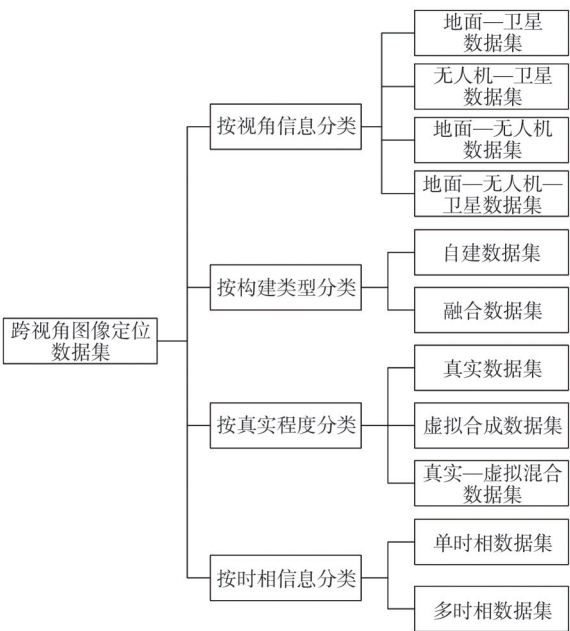


图 1 跨视角图像地理定位数据集分类

Fig. 1 Classification of cross-view image geo-localization datasets

2.2 跨视角图像数据集相关信息

本文对收集的 32 个跨视角图像地理定位数据集，按照元数据信息，包括数据集的名称、基本信息、数据规模、包含场景、图像特征等，从多方面完整地表达跨视角定位数据集的基本属性特征，相关信息汇总如表 2 所示。

表 2 跨视角图像定位数据集
Table 2 Cross-view image localization dataset

名称	基本信息	数据规模	包含场景	图像特征	网址
Ottawa (Bansal 等, 2011)	包含空中视角和倾斜鸟瞰视角,使用微软 Bing 网络服务下载,图像尺寸固定	—	加拿大城市 (建筑)	单时相	
BEV&SV (Bansal 等, 2012)	包含鸟瞰视角和地面视角,鸟瞰视角使用微软 Bing 服务下载,街景视角来自图片分享平台和街景平台	—	美国城市(建筑)	单时相	
Charleston (Lin 等, 2013)	由街景、卫星以及土地覆盖属性图三元组构成,街景视角包含 6756 张图像,卫星视角包含 182988 张图像	~24.8 k	美国城市,农业区,森林,沙滩等	单时相	
San Francisco (Workman 和 Jacobs, 2015)	包含 174217 张地面图像及其对应的航拍图像,覆盖美国旧金山附近 200 km × 200 km 区域	~45.2 k	美国城市	单时相	
CVUSA (Workman 等, 2015)	包含地面视角和卫星视角,其中地面视角由 1036804 张谷歌街景图像和 551851 张 Flickr 图像构成,卫星图像 879318 张	~2.48 m	美国陆地国土随机取样	单时相	http://cs.uky.edu/~jacobs/datasets/cvusa/ [2024-08-27]
DeepGeo (Lin 等, 2015)	包含 78 K 对齐的跨视角图像对,地点覆盖全世界 7 个城市	~78 k	美国、日本、意大利、法国城市	单时相	
Vo&Hays (Vo 和 Hays, 2016)	从谷歌地图中随机收集了美国 11 个城市超过 100 万个街景图像与俯视图像对,每对匹配的图像都包含一张地面图像和一张卫星影像	~1 m	美国城市	单时相	https://github.com/lugiavn/gt-crossview [2024-08-27]

续表

名称	基本信息	数据规模	包含场景	图像特征	网址
UrbanGeo (Tian 等, 2017)	使用 DualMaps 分别在匹兹堡、奥兰多和曼哈顿的 1586、1324 和 5941 个位置采集了街景和鸟瞰视图	~18 k	美国城市	单时相	
CVACT (Liu 和 Li, 2019)	包含 92802 个地面图像与卫星图像对, 其中地面和卫星图像分辨率分别为 1664×832 和 1200×1200	~128 k	澳大利亚城市	单时相	https://github.com/Liumouliu/OriCNN [2024-08-27]
OP (Regmi 和 Shah, 2019)	包含 1910 对训练和 722 对测试用的航拍和地面全景图像。航拍图像的分辨率为 640×640, 全景图像的分辨率为 416×832	~2.6 k	美国城市	单时相	https://github.com/kregmi/cross-view-image-matching [2024-08-27]
CrossViewRet (Khurshid 等, 2019)	包括 6 种类别图像, 每个类别包含 700 张高分辨率街景和卫星图像, 其中卫星图像分辨率为 256×256	~35.4 k	美国公路、山脉、宫殿、河流、船舶和体育场	多时相	https://cvlab.lums.edu.pk/category/projects/imageretrieval [2024-08-27]
VIGOR (Zhu 等, 2021)	包含 105214 张街景图像和 90618 张航拍图像, 分辨率分别为 2048×1024 和 640×640	~48 k	美国城市	单时相	https://github.com/Jeff-Zilence/VIGOR [2024-08-27]
AiRound (Machado 等, 2021)	包含地面视角图像, 高分辨率 RGB 航空影像以及哨兵 2 号卫星拍摄的多光谱影像, 分辨率均为 500×500, 由 11753 张图像组成	—	美国城市(建筑)	单时相	
CV-BrCT (Machado 等, 2021)	包含 24 k 对图像, 由卫星和街景视图组成的图像对构成, 分辨率为 500×500	~48 k	巴西城市	单时相	
Satellite-augmented AD (Shi 和 Li, 2022)	使用卫星数据对两个自动驾驶数据集 KITTI(Geiger 等, 2013)和 Ford multi-AV(Agarwal 等, 2020)进行补充构建新的跨视角数据集。卫星图像的分辨率为 512×512, 地面图像分辨率为 256×1024	—	德国、美国城市(住宅区, 道路等)	多时相	https://github.com/shiyujiao/HighlyAccurate.git [2024-08-27] http://www.cvlibs.net/datasets/kitti [2024-08-27] https://github.com/Ford/AVData [2024-08-27]
KITTI-CVL (Shi 等, 2023c)	包含车载地面视角和卫星视角, 分辨率分别为 256×1024 和 1280×1280。分为训练集、验证集和两个测试集, 数量分别为 23905、2362、2473 和 2473	~31 k	德国城市(住宅区, 道路等)	单时相	
CV-KITTI (Han 等, 2023)	包含 4217 张街景视角图像, 分辨率为 1242×375, 以及 1735 张对应卫星视角图, 分辨率为 466×475	~6 k	德国城市(住宅区, 道路等)	单时相	
Oxford RobotCar-CV (Xia 等, 2021)	地面视角图像为自动驾驶数据集 Oxford RobotCar(Maddern 等, 2020), 分辨率为 1280×960, 卫星视角分辨率为 600×600, 共包含 23854 个图像对	~24 k	英国城市	多时相	https://github.com/tudelft-iv/Visual-Localization-with-Spatial-Prior [2024-08-27] https://robotcar-dataset.robots.ox.ac.uk/ [2024-08-27]
大尺度跨视角图像序列定位数据集 (Zhang 等, 2023c)	包含 38863 张卫星视角图像和 118549 张地面视角图像, 分辨率分别为 640×640 以及 1920×1080, 相机视域约为 120°	~157 k	美国城市、郊区和双向高速公路	多时相	https://gitlab.com/vail-uvms/seqgeo [2024-08-27]
SVA (Palazzi 等, 2017)	数据来自虚拟游戏世界 GTA V, 包含地面和航拍视角, 分辨率均为 256×256, 共 68284 个跨视角图像对	~68 k	虚拟城市(道路)	多时相	http://imagelab.ing.unimore.it/scene-awareness [2024-08-27]
Minecraft (Yagi 等, 2023b)	数据来自虚拟游戏世界 Minecraft, 分为训练集、测试集一、测试集二, 图像数量约 89.3 k。无人机视角分辨率为 128×156, 航拍视角为 256×256, 图像在不同天气和时间条件下获取	~89.3 k	虚拟城市	多时相	
Shetty (Shetty 和 Gao, 2019)	包含无人机视角和卫星视角, 分为训练集和测试集, 分别包含 108 k 和 3 k 个图像对	~111 k	美国城市	单时相	
University-1652 (Zheng 等, 2020)	包含无人机、卫星和地面 3 种视角, 分为训练集、查询集和候选集, 图像数量约 146 k, 分辨率 256×256	~146 k	全世界城市(建筑)	单时相	https://github.com/layumi/University1652-Baseline [2024-08-27]

续表

名称	基本信息	数据规模	包含场景	图像特征	网址
AirLoc (Yan 等, 2023)	包括地面和无人机视角图像, 分别为 1970 张参考图像和 1432 张查询图像, 其中查询图像采集方式包括智能手机和无人机, 图像时相包括白天和黑夜	~3.3 k	中国城市(建筑、街道、植被)	多时相	
CVOGL (Sun 等, 2023)	包含地面、无人机和卫星 3 种视角, 分辨率分别为 256×512、256×256 和 1024×1024, 数量总共约 16 k	~16 k	美国城市, 郊区	单时相	
SUES-200 (Zhu 等, 2023b)	包含不同高度无人机视角和卫星视角图像, 分辨率分别为 1080×1080 和 512×512, 图像数量为 40000 张无人机图像和 200 张卫星图像	~13 k	中国城市(大学, 湖泊, 公园, 学校)	多时相	https://github.com/Reza-Zhu/SUES-200-Benchmark [2024-08-27]
ALTO (Wang 等, 2024b)	包含训练、验证集和测试集, 图像数量约 19 k, 由通用地点识别竞赛提供, 图像分辨率均为 256×256	~19 k	美国城市, 郊区, 乡村, 丛林	单时相	http://gprcompetition.com [2024-08-27]
UL14 (Dai 等, 2024a)	包含不同高度的无人机视角图像和卫星视角图像, 分辨率分别为 512×512 和 1280×1280, 图像数量约 54 k	~54 k	中国城市(大学)	单时相	
NEWYORKFLY (Wang 等, 2024b)	包含无人机视角和卫星视角, 包括训练集、验证集和测试集, 图像数量约 12 k, 图像分辨率均为 256×256	~12 k	美国城市	单时相	
ZHcity-750 (Yan 等, 2024)	由两个子数据集组成, 分别是 Zurich550 和 Harbin200, 包含无人机视角和卫星视角图像, 分辨率均为 256×256, 图像数量约 14.8 k。其中无人机视角图像包含对目标建筑的东西南北以及正上方多个视角	~13.8 k	瑞士 中国城市(建筑)	多时相	
DenseUAV (Dai 等, 2024b)	包含不同高度的无人机视角和卫星视角图像, 图像时相为早晨—傍晚, 分为训练集、测试集和候选集, 图像数量约 54 k	~54 k	中国城市(大学)	多时相	
UAV-VisLoc (Xu 等, 2024)	包含 6742 张不同高度无人机图像和 10 张卫星图像	~6 k	中国城市, 城镇, 农场, 河流	多时相	

注: “~”符号表示约等于, k、m 分别表示图像量级为千级和百万级。

3 相关数据集分析

3.1 元数据分析

元数据描述了数据集的发布时间、数据集规模、数据集覆盖地域、数据集图像特征等信息, 可以从多方面完整地表达跨视角定位数据集的基本属性特征。本节分别从这 4 个方面进行简要分析。

(1) 数据集发布时间分析。根据表 2 中统计的数据集发布年份, 2011 年—2015 年共发布了 6 个, 2015 年—2019 年共发布了 8 个, 2020 年—2024 年发布了 19 个。由图 2 可以看出, 数据集的发布整体呈上升趋势, 2020 年之后增长明显, 这一变化与深度学习以及计算机视觉的发展紧密相关, 也表明跨视角图像定位领域的研究热度在不断攀升。

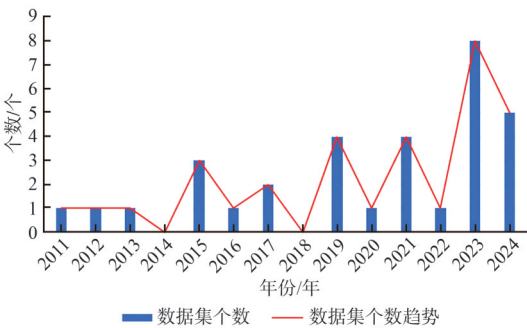


图 2 跨视角图像地理定位数据集年发布量与趋势

Fig. 2 Annual publication volume and publication trend of cross-view image geo-localization dataset

(2) 数据集规模分析。统计表 2 各数据集大小可以得到图 3 所示数据集规模占比。可以看出百万级别的数据集占比为 7%, 共有两个分别为 CVUSA 和 Vo&Hays。万级占比最多为 65%, 千级和十万级均为 14%。结合数据集发布时间分析该现象的主要原因是跨视角图像地理定位发展早期由于技

术不成熟、数据质量各异因而数据集体量不大。随着街景、遥感影像以及无人机数据获取成本的降低、深度学习和计算机视觉技术的发展，很多学者开始自建数据集。根据不同的研究目的，所需的数据集体量也不尽相同。随着图像质量提高，万级数据集因其规模较小且便于共享，因而可以满足基本研究需求，成为大部分学者构建数据集的首选。

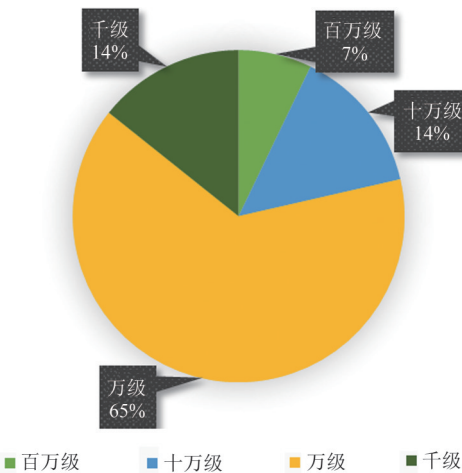


图3 数据集规模占比
Fig. 3 Dataset scale proportions

(3) 数据集覆盖地域分析。从表2中可以看出，32个数据集中覆盖美国地区的数据集有17个，占数据集总数的53%；覆盖欧洲地区的数据集有7个，占据22%，包括英、法、德、意和瑞士；第3为覆盖中国地区的数据集，共有6个，占总数据集的18%。主要原因是美国在跨视角数据集构建与应用领域起步较早，技术手段和各类平台相对成熟，覆盖中国的数据集从2023年以后开始逐渐出现。

(4) 数据集图像特征分析。高度和角度信息描述了数据集图像特征的丰富程度，其中多高度是无人机图像的特有属性，表示无人机在不同高度获取的图像。多角度指的是一个目标地点存在多个角度的图像。早期发布的数据集特征较为单一，后来学者们为了提高模型的泛化性和鲁棒性开始从高度、时相和角度等特征对数据集进行扩展，提高了数据集的丰富程度。例如SUES-200数据集（Zhu等，2023b）中无人机视角图像包括150 m、200 m、250 m、300 m这4种高度，且包含了不同天气条件，如图4所示；ZHcity-750数据集（Yan等，2024）中无人机视角图像包括东、西、南、北和正上方5种视角。



图4 不同天气条件的无人机视角图像
Fig. 4 Drone's view images under different weather conditions

3.2 影响力分析

由于数据集的图像质量、属性丰富程度以及体量大小存在差异，所以数据集的影响力也各不相同。本节以数据集被Google学术平台记录的文献引用次数为依据，对各数据集进行影响力分析，如图5所示。图5中气泡面积大小反映了影响力大小。

可以看出DeepGeo、UrbanGeo、CVUSA、Charleston、Vo&Hays、CVACT以及University-1652数据集影响力较大。Charleston和DeepGeo数据集的作者最早提出利用跨视角匹配的方法进行地理定位，因此，这两个数据集在跨视角地理定位领域具有里程碑意义，其文献通常被奉为经典而被

广泛引用。但由数据集本身由于发布较早且并未公开，因此后期并未受到广泛使用。CVUSA、Vo&Hays 和 CVACT 是目前较为流行的数据集，也是被学者们最广泛使用的数据集。University-1652 是首个地面—无人机—空卫星 3 种视角相结合的跨视角数据集，具有多源性、多视角、密集性等特点，一经发布后也受到很多学者的青睐，其部分图像对示例如图 6 所示。

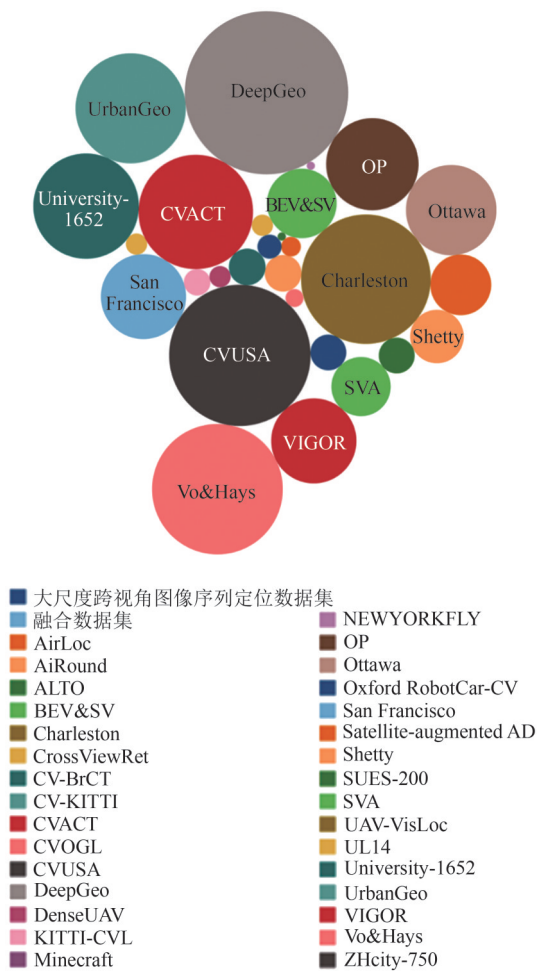


图 5 跨视角图像地理定位数据集影响力分析
Fig. 5 Analysis of the influence of cross-view image geo-localization datasets

3.3 关键词分析

通过在 Web of Science 核心收录数据库检索“跨视角定位”、“无人机定位与导航”、“数据集”等关键词获得相关文献 677 篇。利用 VOSviewer 软件筛选了出现频率大于 3 的共 128 个符合条件的关键词，并进行可视化如图 7 (a) 所示。由于 2019 年之前的铺垫，2019 年—2023 年跨视角图像

地理定位领域得以迅速发展，可以看出图中共有 8 个大聚类族群，分别为地理定位、特征提取、跨视角定位、无人机、位置感知、任务分析、深度学习以及 transformer，这些都是近几年跨视角定位领域发展的重要方向。较小的聚类族群有孪生网络、胶囊网络、卷积神经网络、表示学习、跨视角匹配以及图像合成等，预示着旧技术发展势头减弱以及新技术开始涌现。



图 6 University-1652 数据集图像对示例
Fig. 6 Example of image pairs in University-1652 dataset

在中国知网检索“跨视角定位”获得相关文献 38 篇，关键词 89 个，由 VOSviewer 生成关键词共现图谱如图 7 (b) 所示。可以看出，国内跨视角图像地理定位的发展集中在 2020 年—2023 年，研究热点主要包括图像地理定位、视角转换、无人机等。

3.4 数据获取来源分析

图 8 以柱状图形式总结了上述 32 个跨视角数据集不同视角数据的获取平台。可以看出，航拍视角数据多来自于 Google Map 和 Bing Map 两大平台，因其发展较早、用户多、数据应用技术较为成熟，形成了较完整的数据生产使用链。但随着国内地图平台数据应用和开发技术的发展，高德地图等平台也开始走进研究人员的视野。此外，画面更加精细逼真的虚拟世界作为数据获取平台也慢慢崭露头角。

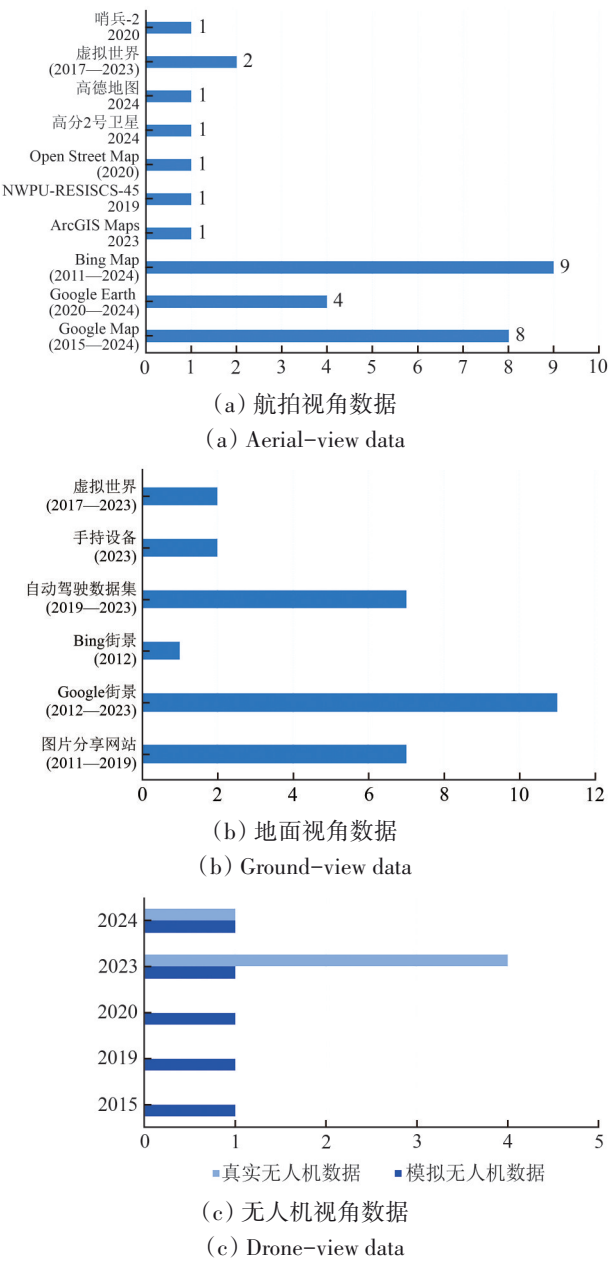


图8 数据获取平台统计图
Fig. 8 Statistical chart of data acquisition platform

无人机视角数据在2023年之前基本由Google Earth Studio地图平台模拟，该平台是谷歌推出的一款无人机飞行模拟引擎，可以从多个角度观察城市和景观（Wang等，2024b）。由于其低成本、模拟效果好，因而被广泛使用。但随着民用无人机技术的发展，使用成本大大降低，且成像多样、精度高，近年来，越来越多学者开始使用真实无人机构建跨视角数据集。

3.5 应用领域分析

跨视角图像地理定位技术发展至今，在民用乃至军用等诸多领域被广泛使用。Durgam等

（2024）将跨视角定位应用领域分为汽车应用、机器人应用、增强现实以及无人机导航和定位，Wilson等（2021）相机位置估计、自动驾驶车辆以及室外增强现实导航。盛怡宁等（2024）认为跨视角图像地理定位可以为机器人导航、自动驾驶以及三维重建提供技术支持。

本文经过梳理跨视角图像地理定位相关文献，整合近年来该技术在各个领域的应用，总结9个应用领域并介绍了跨视角地理定位在各个领域的用法以及相关文献，具体如表3所示。

表 3 跨视角图像地理定位应用领域		
Table 3 Cross-view image geo-localization application		
应用领域	应用方法	相关文献
自动驾驶	通过跨视角定位技术、相机姿态和朝向估计实现车辆精确定位进而辅助自动驾驶	Shi等(Shi and Li, 2022; Shi等, 2023a) Zhang等(2023) Wang等(2023a) Xia等(2024)
	通过将机器人拍摄图像与卫星图像匹配,可实现机器人定位、环境监测、地图绘制以及灾害管理等应用	Xia等(2021) Downes等(2023)
增强现实应用	通过确定相机位于3D场景的精确位置和方向,AR应用可以高精度地将数字内容无缝叠加到物理世界中	Emmaneel等(2023) Wang等(2023b)
无人机导航	给定卫星图像,在无人机图像候选库中匹配同一位置的图像	Zheng等(2020) Liang等(2021) Dai等(2022) Bui等(2023a)
	给定无人机图像,在卫星图像候选库匹配同一位置图像	Luo等(2022) Zhu等(2023a) Shen等(2024)
无人机视觉定位	将无人机摄像头垂直向下获取图像并与卫星图像进行匹配以确定无人机位置	Dai等(2022, 2024a)
无人机姿态估计	通过场景定位网络和摄像头定位网络,借助带有地理参考的卫星图像获得无人机的加权全局姿态	Shetty和Gao(2019)
激光雷达测程与制图	利用激光雷达SLAM方法和空地跨视角姿态校正神经网络实现其定位和制图	Zhu等(2020)
水下潜航器定位	通过跨视角和跨域的图像匹配技术将声学数据和带有地理标签的航空影像进行匹配进而实现水下地理定位	Dos Santos等(2020, 2021, 2022)

4 跨视角图像地理定位算法

上述的公开数据集被广泛应用于基于深度学习的跨视角图像地理定位研究领域，并为相关算法的发展提供了坚实的数据基础。本节依据黄高爽等（2023）的分类，从网络结构优化、损失函数优化、注意力机制、ViT、视角转换以及其他方面对深度学习跨视角图像地理定位算法进行总结和分析。

4.1 网络结构优化

网络结构是深度学习模型的核心，直接影响着算法的定位精度，许多学者围绕网络结构进行了一系列的改进。Workman和Jacobs（2015）等开创性地将卷积神经网络引入跨视角图像检索领域，使用预训练完成的AlexNet进行图像深度特征的提取。Lin等（2015）受到DeepFace（Taigman等，2014）的启发提出了Where-CNN，该网络是首个网使用孪生网络解决跨视角图像地理定位的CNN模型。实验结果表明Where-CNN在不同数据集上具有较好的泛化性。Vo和Hays（2016）探讨了分类、混合、孪生和三生网络这4种深度卷积神经网络架构，研究了视角以及方向信息对跨视角图像匹配带来的挑战，实验结果表明明确的方向监督可以提升位置预测精度。Hu等（2018）使用NetVLAD模块结合两个不共享权重的神经网络实现图像特征提取，用于跨视角图像匹配。薛朝辉等（2021）提出一种融合NetVLAD和全连接层的三元神经网络定位方法（Tri-NetVLAD），该方法通过加强特征之间的关联性提升了模型的定位精度。Li等（2024b）创新性地提出一种四分支框架来减少图像匹配中天空干扰问题。该框架通过裁剪天空和地一空视角对齐进行跨视角地理定位，在CVUSA和CVACT数据集上的实验证明了四分支网络的R@K均优于三分支网络。

4.2 损失函数优化

通过最小化损失函数可以使模型更好地拟合数据，以提高跨视角定位精度。在损失函数优化方面，Vo和Hays（2016）提出了一种基于距离的软边界损失函数和一个辅助网络分支进行方向预测，实验证明该模型在随机方向上具有鲁棒性。Hu等（2018）首次将人脸识别领域三元组损失函

数（Triplet Loss）引入跨视角图像匹配领域，有效提高了网络训练的收敛速度。Cai等（2019）提出了一种新的批内重加权三元组损失函数（in-Batch Reweighting Triplet Loss）从而可以在端到端训练时加强困难样本的积极作用。Li（2024b）设计了一种多约束损失函数MCL（Multiple Constraint Loss）来优化其提出的四分支深度学习框架，减少特征融合过程中的信息丢失导致的不匹配。Liu等（2024）提出了一种新的交叉视角匹配损失函数可以缓解无人机视角图像在同一位置不同高度、不同视角图像中存在的偏差。Li（2024a）通过整合硬样本挖掘策略定义了一种新的加权软边界三元组损失，该损失函数增加了训练难度，迫使网络对方向感知特征更具辨别能力。

4.3 注意力机制

注意力机制作为一种特征映射重校准技术，广泛应用于计算机视觉任务中（林缪鹏等，2024）。注意力机制可以提高卷积神经网络的关注度，减少背景等不相关特征对结果的影响（Woo等，2018），对提高检索精度具有重要作用。Shi等（2019）使用极坐标变换处理航拍图像，然后加入一种空间感知的注意力模块来根据特征位置重新加权特征。以上两步实现了更具辨别性的深度表达，相较于CVM-Net（Hu等，2018），其R@1在CVUSA数据集上提高到89.8%，在CVACT数据集上提高到81.0%。Cai等（2019）提出了一种轻量级的注意力模块FCAM（Feature Context-based Attention Module），并将其集成到基础残差网络中形成孪生网络，其性能优于普通模型（不含FCAM）。何思瑾（2021）设计了一种端到端的多任务联合训练的网络模型，其中图像检索子网络使用空间注意力特征聚合模块进行不同视角之间的特征转化，在CVUSA数据集上的R@1达到了91.59%。Zhang等（2022）设计了一种空间尺度注意力SSA（Spatial-Scale Attention）模块来突出不同尺度下的显著对应布局特征，该模块可以学习到更具判别性的深度特征表达，其在GRAL数据集上将R@1从27.6%提高至40.5%。Li等（2023）提出的多尺度注意力机制能够抑制图像中的变形区域，使得使网络能够理解高层次的语义信息，通过与逆极坐标转换和全局挖掘策略结合，其r@1%在CVUSA和CVACT数据集上分别可达到99.53%

和97.08%。Wang和Li(2023)提出了一种基于交叉注意的交换编码器模块SEM(Swap Encoder Module),该模块可以进行有效的特征对齐,实验证明在CVUSA数据集上R@1和R@10分别从61.4%和90.49%提高到71.45%和92.30%。蔡美玉等(2024)提出一种全局注意力和多粒度特征融合GAMF(Global Attention and Multi-granularity feature Fusion)深度神经网络,该网络嵌入了全局关系注意力模块和通道全局注意力模块以更好地捕捉全局信息,该模型在University-1652数据集的实验结果为:无人机定位任务的平均精度达到87.41%,无人机导航任务的召回率(R@1)达到90.30%。此外,Bui等(2022)提出一种基于通道的注意力机制与基于部分的表征学习方法相结合的新框架,增强特征提取过程。该作者(Bui等,2023b)还将注意力机制应用到特征学习过程中,以提高从输入图像中学习基本特征的能力。

4.4 ViT

自从Transformer被Vaswani等(2017)在自然语言处理领域提出以来,其在深度学习领域一直保持着较高的热度。最近,计算机视觉领域基于Transformer的研究也开始涌现。在图像识别领域,Dosovitskiy等(2021)提出了一个纯Transformer网络,即视觉Transformer(Vision Transformer, ViT)网络,该网络对图像块进行排序并在图像分类任务中表现良好。ViT可以通过自注意力机制模型和位置编码建立各个分析单元之间的空间关系,为空间关系的建模提供了解决途径(鲁伟鹏等,2024)。Yang等(2021)设计了一种地理定位网络EgoTR,它利用Transformer中的自注意力机制来模拟全局依赖关系,从而显著降低跨视角地理定位中的视觉差异性,同时还利用Transformer的位置编码来帮助EgoTR理解与对应地面和俯视图像之间的几何关系。Dai等(2022)发现已有的方法主要为了挖掘更全面的细粒度信息,但忽略了提取鲁棒性特征与特征匹配的重要性,因此提出了一种特征分割与区域匹配结构FSRA(Feature Segmentation and Region Alignment),以增强模型了解上下文的能力。FSRA能根据Transformer特征图的热量分布对区域进行划分,然后将不同视图中的多个特定区域对齐。Zhuang等(2022)提出了一种基于Transformer的网络用来匹配无人机图

像与遥感图像。该网络通过语义引导模块SGM(Semantic Guidance Module)匹配特征,SGM通过基于像素的注意力对图像中的每个像素进行分类,以匹配两幅图像中的相同语义部分。将ViT结构整合到跨视角检索模型中可以进一步提高检索精度,但由于ViT本身的原因(Dosovitskiy等,2021),需要大量的数据进行训练才能获得不错的效果。

4.5 视角转换

视角转换一般分为两种方式,一是通过几何变换将跨视角图像中的一种视角转换为与另一种视角相似的图像再进行特征匹配,其二是利用生成对抗网络GAN(Generative Adversarial Networks)生成不同视角图像,再进行特征匹配。

几何转换方面,Shi等(2019)通过将极坐标变换和空间注意力机制结合实现了更深层次的图像特征表达,从而使跨视角地理定位更加准确,R@1在CVUSA上提高到89.8%,在CVACT上提高到81.0%。Li等(2023)使用逆极坐标变换结合多尺度注意力编码器来提高图像匹配精度。类似的,刘旭东和余平(2023)也使用逆极坐标转换结合深度学习方法进行跨视角地理定位。

生成对抗网络方面,Zhai等(2017)使用PixelNet技术将卫星图像转换为地面视角的语义图,再利用GAN生成地面全景图。Deng等(2018)首次使用条件生成网络cGAN(Conditional Generative Adversarial Networks)将卫星视角图像转换为地面视角。Regmi和Borji(2018)提出了X-Fork与X-Seq架构实现了地面和空中视角图像的双向生成。Tang等(2019)提出了一种多通道注意力选择对抗生成网络(Multi-Channel Attention Selection GAN),可以根据场景图像和语义图像,在任意视角下生成自然场景的图像。Toker等(2021)利用极坐标转换和GAN生成地面图像和图像检索相结合的方法。王先兰等(2023)提出多任务联合模型进行跨视角地理定位,其中前置模型使用逆透视映射结合跨视角生成对抗网络进行图像合成任务。Zhu等(2023c)提出了一种渐进式并行生成对抗网络PPGAN(Progressive Parallel GAN),PPGAN采用并行生成范式,使生成器能够并行产生多分辨率图像,另外通过与特征融合模块以及通道空间注意选择模块的结合,只需一张输入图像即可生成质量较好的地面图像。

除此之外，学者们还在方向信息（Liu 和 Li，2019）、数据集视角样本平衡（Ding 等，2021）、最优特征传输（Shi 等，2020b）、局部特征对齐（Hu 和 Chang，2020；Wang 等，2022）、哈希策略（Ren 等，2023）以及模型轻量化（Han 等，2023）等方面优化跨视角图像定位算法。表 4 和表 5 分别展示了发布于 2016 年—2024 年的跨视角定位方法在 3 个经典数据集 CVUSA（卫星—街景图像匹配任务）、CVACT（卫星—街景图像匹配任务）以及 University-1652（无人机定位和导航任务）上的结

果。表 4 可以通过对比 LPN 和 L2LTR 算法看出极坐标转换方法可以一定程度提高算法的精度，因为极坐标转换可以减小地面和航拍图像的视角差异，并且可以将图像旋转变换转换为平移变换，有效利用了卷积神经网络的平移不变形（Shi 等，2020a）。表 5 可以对比相同算法不同图片分辨率对精度产生的影响。另外，学者们通过消融实验证明损失函数（Cai 等，2019）、池化方式（Bui 等，2022）、图像旋转角度（Tian 等，2022a）以及无人机图像的拍摄高度（Ge 等，2024）也会对精度产生影响。

表 4 跨视角定位方法在 CVUSA 和 CVACT 数据集上的结果比较

Table 4 Comparison of results of cross-view localization methods on CVUSA and CVACT Datasets

		/%							
方法	骨干网络	CVUSA				CVACT			
		R@1	R@5	R@10	R@TOP1%	R@1	R@5	R@10	R@TOP1%
没有使用 极坐标 转换方法	Vo 和 Hays(2016)	AlexNet	—	—	63.7	—	—	—	—
	Regmi 和 Shah(2019)	X-Fork	48.75	—	81.27	95.98	—	—	—
	Rknet(Lin 等,2022)	USAM	52.50	—	—	96.52	40.53	—	89.12
	CVM-Net (Hu 等,2018)	VGG16	18.80	44.42	57.47	91.54	20.15	45.00	56.87
	Orientation(Liu 和 Li,2019)	VGG16	27.15	54.66	67.54	93.91	46.96	68.28	75.48
	Zheng等(2020)	VGG16	43.91	66.38	74.58	91.78	31.20	53.64	63.00
	CVFT(Shi 等,2020b)	VGG16	61.43	84.69	90.94	99.02	61.05	81.33	86.52
	SIRNet(Lu 等,2022)	VGG16	81.82	93.39	96.24	99.49	75.37	88.76	91.90
	Siam-FCANet(Cai 等,2019)	ResNet-34	—	—	—	98.30	—	—	—
	Yagi 和 Iwasaki (2023a)	ResNet-34	89.99	97.17	98.18	99.66	86.37	94.77	96.22
	LPN(Wang 等,2022)	ResNet-50	85.79	95.38	96.80	99.41	79.99	90.63	92.56
	SMDT(Tian 等,2022b)	ViT	95.02	98.97	99.25	99.87	85.52	94.97	96.28
	GeoNet-II(Zhu 等,2022)	ResNetX	—	—	—	98.70	58.90	81.80	88.30
	L2LTR(Yang 等,2024)	ViT	91.99	97.68	98.65	99.75	83.14	93.84	95.51
	Hashing(Ren 等,2023)	ViT+Hash	93.98	97.82	99.27	99.88	—	—	—
	TransGeo(Zhu 等,2022)	ViT	94.08	98.36	99.04	99.77	—	—	—
	TransGeo+4SCIG(Li 等,2024)	ViT	94.10	98.74	99.22	99.81	83.73	94.41	95.79
	GeoViewMatch(Zhang 等,2024b)	ViT	<u>95.52</u>	<u>99.22</u>	<u>99.67</u>	<u>99.83</u>	<u>89.71</u>	<u>95.93</u>	<u>96.92</u>
	Gu 等(2023)	ViT	95.05	98.56	99.09	99.82	87.16	95.27	96.35
	Sample4Geo(Deuser 等,2023)	ConvNeXt	98.68	99.68	99.78	99.87	90.81	96.74	97.48
使用 极坐标 转换方法	Toker 等(2021)	VGG16	92.56	97.55	98.33	99.57	83.28	93.57	95.42
	SAFA(Shi 等,2019)	VGG16	89.84	96.93	98.14	99.64	81.03	92.80	94.84
	DSM(Shi 等,2020a)	VGG16	91.96	97.50	98.54	99.67	82.49	92.44	93.99
	Shi 等(2023b)	VGG16	92.69	97.78	98.60	99.61	82.70	92.50	94.42
	SIRNet(Lu 等,2022)	VGG16	93.74	98.02	98.85	99.76	86.02	94.45	96.02
	LPN(Wang 等,2022)	ResNet-50	93.78	98.50	99.03	99.72	82.87	92.26	94.09
	LPN+USAM(Lin 等,2022)	ResNet-50	91.22	—	—	99.67	82.02	—	—
	DeHi(Wang 等,2024)	VGG16+ViT	94.34	98.63	99.22	99.82	84.96	94.48	95.98
	L2LTR(Yang 等,2024)	ViT	94.05	98.27	98.99	99.67	84.89	94.59	95.96

注：加粗字体为最优方法，下划线则表示次优方法。

表 5 跨视角定位方法在 University-1652 数据集上的结果比较

Table 5 Comparison of results of cross-view localization methods on University-1652 datasets

方法	骨干网络	损失函数	图像尺寸	无人机→卫星		卫星→无人机	
				R@1	AP	R@1	AP
Zheng 等 (2020)	ResNet-50	Instance loss	256×256	58.49	63.13	71.18	58.74
LPN (Wang 等, 2022)	ResNet-101	cross-entropy losses	256×256	76.13	79.29	85.45	75.45
LCM (Ding 等, 2021)	ResNet-50	Cross entropy	512×512	66.65	70.82	79.89	65.38
MSBA (Zhuang 等, 2021)	ResNet-50	Cross-Entropy	256×256	82.33	84.78	90.58	81.61
		Loss+KL loss	384×384	86.61	88.55	92.15	84.45
PCL (Tian 等, 2022a)	ResNet-50	CGAN loss+L1 loss	512×512	83.27	87.32	91.78	82.18
SGM (Zhuang 等, 2022)	Swin-Tiny	Cross-Entropy loss	256×256	82.14	84.72	88.16	81.81
FSRA (Dai 等, 2022b)	Vit-S	Triplet Loss	256×256	82.25	84.82	87.87	81.53
			512×512	85.50	87.53	89.73	84.94
PAAN (Bui 等, 2022)	SE-ResNet50	KL loss	256×256	84.51	86.78	91.01	82.28
Wang 等 (2023c)	Vit-S	Triplet Loss	256×256	85.34	86.73	88.23	85.52
DPN (Dong 等, 2023)	RegNet	—	256×256	90.99	92.33	93.01	90.11
Song 等 (2023)	OSNet	categorized cluster loss	384×384	84.79	87.07	93.15	84.93
MBF (Zhu 等, 2023a)	ResNet+ViT	Cross-Entropy Loss+Triplet loss	384×384	89.05	90.61	93.15	88.17
MCCG (Shen 等, 2024)	ConvNeXt-Tiny	Cross-Entropy Loss+ Triplet loss	256×256	89.28	91.01	94.29	89.29
			384×384	89.64	91.32	94.30	89.39
Sample4Geo (Deuser 等, 2023)	ConvNext-B	InfoNCE Loss	384×384	<u>92.65</u>	<u>93.81</u>	<u>95.14</u>	<u>91.39</u>
TirSA (Sun 等, 2024)	ResNet-50-ibn	KL Loss+MES Loss	256×256	93.37	94.53	96.15	94.41
TransFG (Zhao 等, 2024)	ViT	TC-Loss	256×256	84.01	86.31	90.16	84.61
GeoFormer (Li 等, 2024)	E-Swin-L	Cross-Entropy Loss	224×224	89.08	90.83	92.30	88.54
SeGCN (Liu 等, 2024)	SwinV2-T	Triplet Loss	256×256	89.18	90.89	94.29	89.65

注：加粗字体为最优方法，下划线则表示次优方法。

5 讨 论

基于数据驱动的深度学习技术在跨视角图像地理定位研究领域，其精度和泛化能力依赖于样本数据集的规模和质量。当前，还存在高质量数据集不多、数据来源单一、模型泛化能力不高以及难以做到精简高效的问题。对此，本文从数据集多模态趋势、大模型方法、干扰物处理以及模型优化等角度，对相关研究方法就目前存在的难点和挑战进行讨论。

(1) 匹配精度问题。由于获取平台和传感器的不同，图像之间存在的“五差异”（成像特性差异、几何差异、尺度差异、视角差异以及维度差异）和“三不同”（不同环境、天气以及气候）使得精确匹配存在较大困难（Jiang 等，2021）。如何让模型有效融合图像不同方面的特征，增加模型

对图像中细节的感知和表达鲁棒特征的能力，是跨视角图像地理定位模型亟需解决的关键问题（Li 等，2024b）。随着 C2Seg（Hong 等，2023），GeoText-1652 数据集（Chu 等，2025）等多模态遥感数据集的不断涌现，不同类型的数据可以提供互补的信息，如文本或音视频等，包含更丰富的上下文信息，有助于更好地理解图像内容。与此同时，不同模态间的组合也可以创造出更多的虚拟样本，利用多模态数据集来提升跨视角图像地理定位精度是值得进一步研究的方向。

(2) 泛化能力弱问题。跨视角地理定位由于空中和地面图像之间空间对应关系的脆弱性使得数据增广非常有限，且容易被轻微的干扰破坏（Zhang 等，2023b），容易出现过拟合，泛化能力较弱。遥感大模型的发展为跨视角地理定位提供新的思路和方法。大语言模型和视觉大模型在语

言理解和视觉任务中表现出来的强大泛化性引起了遥感领域学者的广泛关注。基于不同任务的大模型被陆续开发,如为处理光谱遥感数据而设计,基于场景分类、语义分割以及变化监测任务的基础模型SpectralGPT (Hong等, 2024), 基于图像理解任务的EarthGPT (Zhang等, 2024a) 和RSGPT (Hu等, 2025) 以及基于街景图像定位任务的AddressCLIP (Xu等, 2025) 等。大模型可作为数据源为跨视角地理定位任务提供新的数据增广方法,生成更多视角的图像。同时,大模型可以学习到更加泛化的特征表示,有助于提高模型在不同数据集上的迁移能力。

(3) 图像干扰物问题。地面影像、低高度无人机图像由于噪声、其他目标物体、光源变化以及遮挡物等因素会导致图像中关键特征信息的丢失,从而影响网络学习特征的效率。而跨视角图像地理定位本身由于视角的剧烈变化导致不同视角图像间巨大视觉外观变化,从而对图像特征提取效率和精度提出了更高的要求。学者们试图通过各种方式提升精度,主要有注意力机制 (Woo等, 2018)、最优特征传输 (Shi等, 2020b) 以及去除干扰物 (Li等, 2024b) 等。然而各类方法目前存在参数较多以及计算量较大的问题,如何以更高效的方式减少图像的噪声以及干扰物带来的影像等是未来重要的研究方向之一。

(4) 模型优化问题。一直以来跨视角图像地理定位技术都面临着效率和模型复杂度、计算量以及体积不可兼得的情况,模型优化技术可以在模型效率和计算成本之间获得一定程度的平衡。模型优化技术种类繁多,其中模型轻量化技术可以在不降低精度的前提下有效减少参数和计算量,对于较为复杂网络,轻量化设计可以加快训练和推理速度,提升模型的效率,并且可以更好地适应边缘计算和移动以及嵌入式设备 (自动驾驶汽车和增强现实) 的需求 (Li等, 2024c)。目前很多学者已经开始研究跨视角定位领域的轻量级网络设计 (Han等, 2023; Yagi和Iwasaki, 2023)。模型压缩技术通过减少模型的参数数量、降低模型的计算复杂度等方式来降低模型的计算和存储开销,比如参数剪枝和知识蒸馏等。此外,迁移学习、并行和分布式训练、超参数调优以及新的硬件加速技术和也是未来重要的发展方向。

6 结 语

本文首先归纳梳理了近年来跨视角图像地理定位领域的32个数据集,立足于视角信息、构建类型、真实程度以及时相信息4种特征,实现不同特征视角下对数据集分类体系的构建,并以表格形式对数据集进行详细介绍。其次,从元数据、影响力、关键词、数据获取来源以及应用领域对跨视角地理定位数据集进行全面的对比分析,总结了跨视角图像地理定位发展以来的典型的、具有代表性的算法和技术,并以表格形式将2016至今的跨视角定位算法在3个经典数据集上进行展现。此外,针对跨视角图像地理定位数据集存在的不足,讨论了数据集的多模态趋势、大模型方法、图像干扰物处理以及模型优化技术4种发展方向的可能性,为相关领域研究人员提供数据支撑和算法参考。

参考文献 (References)

- Agarwal S, Vora A, Pandey G, Williams W, Kourous H and McBride J. 2020. Ford multi-AV seasonal dataset. The International Journal of Robotics Research, 39(12): 1367-1376 [DOI: 10.1177/0278364920961451]
- Bansal M, Daniilidis K and Sawhney H. 2012. Ultra-wide baseline facade matching for geo-localization//Proceedings of the European Conference on Computer Vision. Florence: Springer: 175-186 [DOI: 10.1007/978-3-642-33863-2_18]
- Bansal M, Sawhney H S, Cheng H and Daniilidis K. 2011. Geo-localization of street views with aerial image databases//Proceedings of the 19th ACM international conference on Multimedia. Scottsdale: ACM: 1125-1128 [DOI: 10.1145/2072298.2071954]
- Bay H, Tuytelaars T and Van Gool L. 2006. SURF: speeded up robust features//Proceedings of the 9th European Conference on Computer Vision Computer Vision. Graz: Springer [DOI: 10.1007/11744023_32]
- Bui D V, Kubo M and Sato H. 2022. A part-aware attention neural network for cross-view geo-localization between UAV and satellite. Journal of Robotics, Networking and Artificial Life, 9(3): 275-284 [DOI: 10.57417/jrnal.9.3_275]
- Bui D V, Kubo M and Sato H. 2023a. Cross-view geo-localization for autonomous UAV using locally-aware transformer-based network. IEEE Access, 11: 104200-104210 [DOI: 10.1109/ACCESS.2023.3317950]
- Bui D V, Kubo M and Sato H. 2023b. Attention-based neural network with Generalized Mean Pooling for cross-view geo-localization

- between UAV and satellite. *Artificial Life and Robotics*, 28(3): 560-570 [DOI: 10.1007/s10015-023-00867-x]
- Cai M Y, Zhu R Z, Wu F, Zhang K Y and Li J L. 2024. Cross-view matching model based on attention mechanism and multi-granularity feature fusion. *Journal of Computer Applications*, 44(3): 901-908 (蔡美玉, 朱润哲, 吴飞, 张开昱, 李家乐. 2024. 基于注意力机制和多粒度特征融合的跨视角匹配模型. *计算机应用*, 44(3): 901-908) [DOI: 10.11772/j.issn.1001-9081.2023040412]
- Cai S D, Guo Y L, Khan S, Hu J W and Wen G J. 2019. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*. Seoul: IEEE [DOI: 10.1109/ICCV.2019.00848]
- Chu M, Zheng Z, Ji W, Wang T and Chua T S. 2025. Towards natural language-guided drones: GeoText-1652 benchmark with spatial relation matching//*Proceedings of the 18th European Conference on Computer Vision*. Milan: Springer [DOI: 10.1007/978-3-031-73247-8_13]
- Dai M, Zheng E H, Chen J H, Lei Q, Feng Z H, Wan K Y. 2024a. Drone Referring Localization: An Efficient Heterogeneous Spatial Feature Interaction Method For UAV Self-Localization. *arXiv*: 2208.06561
- Dai M, Hu J H, Zhuang J D and Zheng E H. 2022. A transformer-based feature segmentation and region alignment method for UAV-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7): 4376-4389 [DOI: 10.1109/TCSVT.2021.3135013]
- Dai M, Zheng E H, Feng Z H, Qi L, Zhuang J D and Yang W K. 2024b. Vision-based UAV self-positioning in low-altitude urban environments. *IEEE Transactions on Image Processing*, 33: 493-508 [DOI: 10.1109/TIP.2023.3346279]
- Deng X Q, Zhu Y and Newsam S. 2018. What is it like down there?: generating dense ground-level views and image features from overhead imagery using conditional generative adversarial networks//*Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. Seattle: ACM: 43-52 [DOI: 10.1145/3274895.3274969]
- Deuser F, Habel K and Oswald N. 2023. Sample4Geo: hard negative sampling for cross-view geo-localisation//*Proceedings of 2023 IEEE/CVF International Conference on Computer Vision*. Paris: IEEE [DOI: 10.1109/ICCV51070.2023.01545]
- Ding L R, Zhou J, Meng L X and Long Z Y. 2021. A practical cross-view image matching method between UAV and satellite for UAV-based geo-localization. *Remote Sensing*, 13(1): 47 [DOI: 10.3390/rs13010047]
- Dong L Y, Wang Y H, Huang J S, Qian X M, Fan M Y and Lai S Q. 2023. Dual path network for cross-view geo-localization//*Proceedings of the 2023 Workshop on UAVs in Multimedia: Capturing the World from a New Perspective*. Ottawa: ACM: 45-49 [DOI: 10.1145/3607834.3616573]
- Dos Santos M M, De Giacomo G G, Drews Jr P L J and Botelho S S C. 2022. Cross-view and cross-domain underwater localization based on optical aerial and acoustic underwater images. *IEEE Robotics and Automation Letters*, 7(2): 4969-4974 [DOI: 10.1109/LRA.2022.3154482]
- Dos Santos M M, De Giacomo G G, Drews Jr P L J, Botelho S S C and Mello C D. 2021. A framework for underwater vehicle localization based on cross-view and cross-domain acoustic and aerial images//*Proceedings of 2021 Latin American Robotics Symposium (LARS), 2021 Brazilian Symposium on Robotics (SBR), and 2021 Workshop on Robotics in Education (WRE)*. Natal: IEEE [DOI: 10.1109/LARS/SBR/WRE54079.2021.9605368]
- Dos Santos M M, De Giacomo G G, Drews P L J and Botelho S S C. 2020. Matching color aerial images and underwater sonar images using deep learning for underwater localization. *IEEE Robotics and Automation Letters*, 5(4): 6365-6370 [DOI: 10.1109/LRA.2020.3013852]
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houshy N. 2021. An image is worth 16x16 words: transformers for image recognition at scale//*Proceedings of the 9th International Conference on Learning Representations*. [s.l.]: OpenReview.net
- Downes L M, Steiner T J, Russell R L and How J P. 2023. Wide-area geolocalization with a limited field of view camera//*Proceedings of 2023 IEEE International Conference on Robotics and Automation*. London: IEEE [DOI: 10.1109/ICRA48891.2023.10160607]
- Durgam A, Paheding S, Dhiman V and Devabhaktuni V. 2024. Cross-view geo-localization: a survey. *IEEE Access*, 12: 192028-192050 [DOI: 10.1109/ACCESS.2024.3507280]
- Emmaneel R, Oswald M R, de Haan S and Datcu D. 2023. Cross-view outdoor localization in augmented reality by fusing map and satellite data. *Applied Sciences*, 13(20): 11215 [DOI: 10.3390/app132011215]
- Ge F W, Zhang Y Z, Liu Y X, Wang G Y, Coleman S, Kerr D and Wang L. 2024. Multibranch joint representation learning based on information fusion strategy for cross-view geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 5909516 [DOI: 10.1109/TGRS.2024.3378453]
- Geiger A, Lenz P, Stiller C and Urtasun R. 2013. Vision meets robotics: the KITTI dataset. *The International Journal of Robotics Research*, 32(11): 1231-1237 [DOI: 10.1177/0278364913491297]
- Gu B, Ling H F, Shi Y X, Li Z Y, Zhao C, Li P and Cao Q. 2023. Mutual relative position learning transformer for cross-view geo-localization//*Proceedings of 2023 IEEE International Conference on Image Processing*. Kuala Lumpur: IEEE [DOI: 10.1109/ICIP49359.2023.10222254]
- Han Z H, Li Y C and Xue J R. 2023. LightningSpike: a lightweight hy-

- brid intelligent system for cross-view geo-localization//Proceedings of 2023 China Automation Congress. Chongqing: IEEE [DOI: 10.1109/CAC59555.2023.10451619]
- He S J. 2021. Research on Cross-View Image Geo-Localization Technology Based on Deep Learning. Wuhan: Huazhong University of Science and Technology (何思瑾. 2021. 基于深度学习的跨视角图像地理定位技术研究. 武汉: 华中科技大学) [DOI: 10.27157/d.cnki.ghzku.2021.002046]
- Hong D F, Zhang B, Li H, Li Y X, Yao J, Li C Y, Werner M, Chanussot J, Zipf A and Zhu X X. 2023. Cross-city matters: a multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. *Remote Sensing of Environment*, 299: 113856 [DOI: 10.1016/j.rse.2023.113856]
- Hong D F, Zhang B, Li X Y, Li Y X, Li C Y, Yao J, Yokoya N, Li H, Ghamisi P, Jia X P, Plaza A, Gamba P, Benediktsson J A and Chanussot J. 2024. SpectralGPT: spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5227-5244 [DOI: 10.1109/TPAMI.2024.3362475]
- Hu S X, Feng M D, Nguyen R M H and Lee G H. 2018. CVM-Net: cross-view matching network for image-based ground-to-aerial geo-localization//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE [DOI: 10.1109/CVPR.2018.00758]
- Hu S Y and Chang X J. 2020. Multi-view drone-based geo-localization via style and spatial alignment. *arXiv*: 2006.13681
- Hu Y, Yuan J L, Wen C C, Lu X N, Liu Y and Li X. 2025. RSGPT: a remote sensing vision language model and benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*. 224: 272-286. [DOI: <https://doi.org/10.1016/j.isprsjprs.2025.03.028>]
- Huang G S, Zhou Y, Hu X F, Zhao L Y and Zhang C L. 2023. A survey of the research progress in image geo-localization. *Journal of Geo-Information Science*, 25(7): 1336-1362 (黄高爽, 周杨, 胡校飞, 赵璐颖, 张呈龙. 2023. 图像地理定位研究进展. 地球信息科学学报, 25(7): 1336-1362) [DOI: 10.12082/dqxxkx.2023.230073]
- Jiang X Y, Ma J Y, Xiao G B, Shao Z F and Guo X J. 2021. A review of multimodal image matching: methods and applications. *Information Fusion*, 73: 22-71 [DOI: 10.1016/j.inffus.2021.02.012]
- Khurshid N, Hanif T, Tharani M and Taj M. 2019. Cross-view image retrieval - ground to aerial image retrieval through deep learning//Proceedings of the 26th International Conference on Neural Information Processing. Sydney: Springer [DOI: 10.1007/978-3-030-36711-4_19]
- Leung K Y K, Clark C M and Huissoon J P. 2008. Localization in urban environments by matching ground level video images with an aerial image//Proceedings of 2008 IEEE International Conference on Robotics and Automation. Pasadena: IEEE [DOI: 10.1109/ROBOT.2008.4543264]
- Li C R, Yan C, Xiang X J, Lai J, Zhou H and Tang D Q. 2024a. HAD-GEO: image based 3-DoF cross-view geo-localization with hard sample mining//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul: IEEE [DOI: 10.1109/ICASSP48485.2024.10445839]
- Li J S, Yang C F, Qi B J, Zhu M and Wu N. 2024b. 4SCIG: a four-branch framework to reduce the interference of sky area in cross-view image geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 4703818 [DOI: 10.1109/TGRS.2024.3379376]
- Li Q G, Yang X G, Fan J W, Lu R T, Tang B, Wang S Y and Su S. 2024c. GeoFormer: an effective transformer-based siamese network for UAV geolocalization. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17: 9470-9491 [DOI: 10.1109/JSTARS.2024.3392812]
- Li S L, Tu Z G, Chen Y J and Yu T. 2023. Multi-scale attention encoder for street-to-aerial image geo-localization. *CAAI Transactions on Intelligence Technology*, 8(1): 166-176 [DOI: 10.1049/cit2.12077]
- Liang Z B, Liu X, Chen B, Yuan Y F, Song Y and Jiang H F. 2021. UAV-based cross-view geo-localization fusion spatial attention mechanism and Netvlad//Proceedings of the 7th International Conference on Systems and Informatics. Seoul: IEEE [DOI: 10.1109/ICSAI53574.2021.9664015]
- Lin J L, Zheng Z D, Zhong Z, Luo Z M, Li S Z, Yang Y and Sebe N. 2022. Joint representation learning and keypoint detection for cross-view geo-localization. *IEEE Transactions on Image Processing*, 31: 3780-3792 [DOI: 10.1109/TIP.2022.3175601]
- Lin L P, Li J and Shen H F. 2024. PolSAR image deep learning super-resolution model based on multiscale attention mechanism. *National Remote Sensing Bulletin*, 28(9): 2362-2371 (林缪鹏, 李杰, 沈焕锋. 2024. 基于多尺度注意力机制的PolSAR深度学习超分辨率模型. 遥感学报, 28(9): 2362-2371) [DOI: 10.11834/jrs.20233002]
- Lin T Y, Belongie S and Hays J. 2013. Cross-view image geolocalization//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE [DOI: 10.1109/CVPR.2013.120]
- Lin T Y, Yin C, Belongie S and Hays J. 2015. Learning deep representations for ground-to-aerial geolocalization//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE [DOI: 10.1109/CVPR.2015.7299135]
- Liu C W, Li S S, Du C and Qiu H J. 2024. Adaptive global embedding learning: a two-stage framework for UAV-view geo-localization. *IEEE Signal Processing Letters*, 31: 1239-1243 [DOI: 10.1109/LSP.2024.3392676]
- Liu L and Li H D. 2019. Lending orientation to neural networks for cross-view geo-localization//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE [DOI: 10.1109/CVPR.2019.00577]
- Liu L, Li J S, Xu J Q, Yang C F, Qin Y L and Liu Y. 2023. Cross-view

- image geo-localization review. *Artificial Intelligence Security*, 2(1): 53-61 (刘棟, 李江杉, 许乾乾, 杨春芳, 秦雨林, 刘琰. 2023. 跨视角图像地理定位技术研究综述. *智能安全*, 2(1): 53-61) [DOI: 10.12407/j.issn.2097-2075.2023.01.053]
- Liu X. 2023. Research on Cross-View Geo-Localization Method Based on Drone and Satellite Image. Chongqing: Chongqing Jiaotong University (刘兴. 2023. 基于无人机和卫星图像的跨视角地理定位方法研究. 重庆: 重庆交通大学) [DOI: 10.27671/d.cnki.gcjtc.2023.000234]
- Liu X, Wang Z, Wu Y, Miao Q. 2024. SeGCN: A Semantic-Aware Graph Convolutional Network for UAV Geo-Localization. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17: 6055-6066. [DOI: 10.1109/JSTARS.2024.3370612]
- Liu X D and Yu P. 2023. Cross-view geo-visual localization. *Computer Science*, 50(11A): 221100066 (刘旭东, 余平. 2023. 跨视角地理视觉定位. *计算机科学*, 50(11A): 221100066) [DOI: 10.11896/jcs-jkx.221100066]
- Lowe D G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2): 91-110 [DOI: 10.1023/B:VISI.0000029664.99615.94]
- Lu W P, He Q K, Li J L, Li S Y and Tao C. 2024. Object units and Transformer networks combined with urban functional zone classification method. *National Remote Sensing Bulletin*, 28(8): 1927-1939 (鲁伟鹏, 贺清康, 李佳铃, 李诗逸, 陶超. 2024. 结合对象单元和Transformer网络的城市功能区分类. *遥感学报*, 28(8): 1927-1939) [DOI: 10.11834/jrs.20233036]
- Lu X F, Luo S Q and Zhu Y Y. 2022. It's okay to be wrong: cross-view geo-localization with step-adaptive iterative refinement. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 4709313 [DOI: 10.1109/TGRS.2022.3210195]
- Luo X B, Tian Y L, Wan X, Xu J Z and Ke T. 2022. Deep learning based cross-view image matching for UAV geo-localization//*Proceedings of 2022 International Conference on Service Robotics*. Chengdu: IEEE [DOI: 10.1109/ICoSR57188.2022.00028]
- Machado G, Ferreira E, Nogueira K, Oliveira H, Brito M, Gama P H T and Santos J A D. 2021. AiRound and CV-BrCT: novel multiview datasets for scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 488-503 [DOI: 10.1109/JSTARS.2020.3033424]
- Maddern W, Pascoe G, Gadd M, Barnes D, Yeomans B and Newman P. 2020. Real-time kinematic ground truth for the oxford robotcar dataset. *arXiv*: 2002.10152
- Palazzi A, Borghi G, Abati D, Calderara S and Cucchiara R. 2017. Learning to map vehicles into bird's eye view//*Proceedings of the 19th International Conference on Image Analysis and Processing*. Catania: Springer [DOI: 10.1007/978-3-319-68560-1_21]
- Rao Z Y, Lu J, Guo H T, Yu D H and Hou Q F. 2023. A cross-view image matching method with viewpoint conversion. *Journal of Geo-Information Science*, 25(2): 368-379 (饶子昱, 卢俊, 郭海涛, 余东行, 侯青峰. 2023. 利用视角转换的跨视角影像匹配方法. *地球信息科学学报*, 25(2): 368-379) [DOI: 10.12082/dqxxkx.2023.220312]
- Regmi K and Borji A. 2018. Cross-view image synthesis using conditional GANs//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE [DOI: 10.1109/CVPR.2018.00369]
- Regmi K and Shah M. 2019. Bridging the domain gap for ground-to-aerial image matching//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*. Seoul: IEEE [DOI: 10.1109/ICCV.2019.00056]
- Ren P, Tao Y M, Han J P and Li P. 2023. Hashing for geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 5623213 [DOI: 10.1109/TGRS.2023.3325884]
- Rublee E, Rabaud V, Konolige K and Bradski G. 2011. ORB: an efficient alternative to SIFT or SURF//*Proceedings of 2011 International Conference on Computer Vision*. Barcelona: IEEE [DOI: 10.1109/ICCV.2011.6126544]
- Shen T R, Wei Y M, Kang L, Wan S S and Yang Y H. 2024. MCCG: a ConvNeXt-based multiple-classifier method for cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3): 1456-1468 [DOI: 10.1109/TCSVT.2023.3296074]
- Sheng Y N, Zhao L J, Zhang Z, Cui S L, Rao M B, Tang P. 2024. Review of cross-view image geolocalization methods. *Journal of Image and Graphics*, 29(09): 2716-2736 (盛怡宁, 赵理君, 张正, 崔绍龙, 饶梦彬, 唐婷. 2024. 跨视角图像地理定位方法综述. *中国图象图形学报*, 29(09): 2716-2736) [DOI: 10.11834/jig.230585]
- Shetty A and Gao G X. 2019. UAV pose estimation using cross-view geolocalization with satellite imagery//*Proceedings of 2019 International Conference on Robotics and Automation*. Montreal: IEEE [DOI: 10.1109/ICRA.2019.8794228]
- Shi Y J and Li H D. 2022. Beyond cross-view image retrieval: highly accurate vehicle localization using satellite image//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans: IEEE [DOI: 10.1109/CVPR52688.2022.01650]
- Shi Y J, Liu L, Yu X and Li H D. 2019. Spatial-aware feature aggregation for cross-view image based geo-localization//*Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Vancouver: Curran Associates Inc
- Shi Y J, Wu F, Perincherry A, Vora A and Li H D. 2023a. Boosting 3-DoF ground-to-satellite camera localization accuracy via geometry-guided cross-view transformer//*Proceedings of 2023 IEEE/CVF International Conference on Computer Vision*. Paris: IEEE [DOI: 10.1109/ICCV51070.2023.01967]
- Shi Y J, Yu X, Campbell D and Li H D. 2020a. Where am I looking at? Joint location and orientation estimation by cross-view matching//*Proceedings of 2020 IEEE/CVF Conference on Computer Vi-*

- sion and Pattern Recognition. Seattle: IEEE [DOI: 10.1109/CVPR42600.2020.00412]
- Shi Y J, Yu X, Liu L, Campbell D, Koniusz P and Li H D. 2023b. Accurate 3-DoF camera geo-localization via ground-to-satellite image matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 2682-2697 [DOI: 10.1109/TPAMI.2022.3189702]
- Shi Y J, Yu X, Liu L, Zhang T and Li H D. 2020b. Optimal feature transport for cross-view image geo-localization//*Proceedings of the 34th AAAI Conference on Artificial Intelligence*. New York: AAAI Press [DOI: 10.1609/aaai.v34i07.6875]
- Shi Y J, Yu X, Wang S and Li H D. 2023c. CVLNet: cross-view semantic correspondence learning for video-based camera localization//*Proceedings of the 16th Asian Conference on Computer Vision*. Macao: Springer [DOI: 10.1007/978-3-031-26319-4_8]
- Song H S, Wang Z, Lei Y, Shi D X, Tong X C, Lei Y X and Qiu C P. 2023. Learning visual representation clusters for cross-view geo-location. *IEEE Geoscience and Remote Sensing Letters*, 20: 6011805 [DOI: 10.1109/LGRS.2023.3326005]
- Sun J, Sun H, Lei L, Ji K F and Kuang G Y. 2024. TirSA: a three stage approach for UAV-satellite cross-view geo-localization based on self-supervised feature enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9): 7882-7895 [DOI: 10.1109/TCSVT.2024.3382717]
- Sun Y X, Ye Y M, Kang J, Fernandez-Beltran R, Feng S S, Li X T, Luo C Y, Zhang P Z and Plaza A. 2023. Cross-view object geo-localization in a local region with satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 4704716 [DOI: 10.1109/TGRS.2023.3307508]
- Taigman Y, Yang M, Ranzato M and Wolf L. 2014. DeepFace: closing the gap to human-level performance in face verification//*Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus: IEEE [DOI: 10.1109/CVPR.2014.220]
- Tang H, Xu D, Sebe N, Wang Y Z, Corso J J and Yan Y. 2019. Multi-channel attention selection GAN with cascaded semantic guidance for cross-view image translation//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE [DOI: 10.1109/CVPR.2019.00252]
- Tian X Y, Shao J, Ouyang D Q and Shen H T. 2022a. UAV-satellite view synthesis for cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7): 4804-4815 [DOI: 10.1109/TCSVT.2021.3121987]
- Tian X Y, Shao J, Ouyang D Q, Zhu A J and Chen F Y. 2022b. SMDT: cross-view geo-localization with image alignment and transformer//*Proceedings of 2022 IEEE International Conference on Multimedia and Expo*. Taipei, China: IEEE [DOI: 10.1109/ICME52920.2022.9859992]
- Tian Y C, Chen C and Shah M. 2017. Cross-view image matching for geo-localization in urban environments//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE: 1998-2006 [DOI: 10.1109/cvpr.2017.216]
- Toker A, Zhou Q J, Maximov M and Leal-Taixé L. 2021. Coming down to earth: satellite-to-street view synthesis for geo-localization//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville: IEEE [DOI: 10.1109/CVPR46437.2021.00642]
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//*Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach: Curran Associates Inc: 6000-6010
- Vo N N and Hays J. 2016. Localizing and orienting street views using overhead imagery//*Proceedings of the 14th European Conference on Computer Vision*. Amsterdam: Springer [DOI: 10.1007/978-3-319-46448-0_30]
- Wang J J and Li X. 2023. Cross-attention network for cross-view image geo-localization//*Proceedings of the 6th International Symposium on Autonomous Systems*. Nanjing: IEEE: 1-6 [DOI: 10.1109/isas59543.2023.10164457]
- Wang S, Zhang Y H, Perincherry A, Vora A and Li H D. 2023a. View consistent purification for accurate cross-view localization//*Proceedings of 2023 IEEE/CVF International Conference on Computer Vision*. Paris: IEEE [DOI: 10.1109/ICCV51070.2023.00753]
- Wang S, Zhang Y H, Vora A, Perincherry A and Li H D. 2023b. Satellite image based cross-view localization for autonomous vehicle//*Proceedings of 2023 IEEE International Conference on Robotics and Automation*. London: IEEE [DOI: 10.1109/ICRA48891.2023.10161527]
- Wang T, Li J W and Sun C Y. 2024a. DeHi: a decoupled hierarchical architecture for unaligned ground-to-aerial geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3): 1927-1940 [DOI: 10.1109/TCSVT.2023.3293514]
- Wang T Y, Zheng Z D, Yan C G, Zhang J Y, Sun Y Q, Zheng B L and Yang Y. 2022. Each part matters: local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2): 867-879 [DOI: 10.1109/TCSVT.2021.3061265]
- Wang X L, Zhou J K, Mu N and Wang C. 2023. Cross-view geo-localization method based on multi-task joint learning. *Journal of Computer Applications*, 43(5): 1625-1635 (王先兰, 周金坤, 穆楠, 王晨. 2023. 基于多任务联合学习的跨视角地理定位方法. *计算机应用*, 43(5): 1625-1635) [DOI: 10.11772/j.issn.1001-9081.2022040541]
- Wang Y F, Xia Y M, Lu T B, Zhang X Y and Yao W B. 2023c. An efficient method based on multi-view semantic alignment for cross-view geo-localization//*Proceedings of 2023 International Joint Conference on Neural Networks*. Gold Coast: IEEE [DOI: 10.1109/IJCNN54540.2023.10191537]

- Wang Z, Shi D X, Qiu C P, Jin S C, Li T Y, Shi Y Y, Liu Z and Qiao Z T. 2024b. Sequence matching for image-based UAV-to-satellite geolocalization. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 5607815 [DOI: 10.1109/TGRS.2024.3359605]
- Wilson D, Zhang X H, Sultani W and Wshah S. 2023. Visual and object geo-localization: a comprehensive survey. *arXiv*: 2112.15202
- Woo S, Park J, Lee J Y and Kweon I S. 2018. CBAM: convolutional block attention module//*Proceedings of the 15th European Conference on Computer Vision*. Munich: Springer: 3-19 [DOI: 10.1007/978-3-030-01234-2_1]
- Workman S and Jacobs N. 2015. On the location dependence of convolutional neural network features//*Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Boston: IEEE [DOI: 10.1109/CVPRW.2015.7301385]
- Workman S, Souvenir R and Jacobs N. 2015. Wide-area image geolocalization with aerial reference imagery//*Proceedings of 2015 IEEE International Conference on Computer Vision*. Santiago: IEEE: 3961-3969 [DOI: 10.1109/ICCV.2015.451]
- Xia Z M, Booi O and Kooij J F P. 2024. Convolutional cross-view pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5): 3813-3831 [DOI: 10.1109/TPAMI.2023.3346924]
- Xia Z M, Booi O, Manfredi M and Kooij J F P. 2021. Cross-view matching for vehicle localization by learning geographically local representations. *IEEE Robotics and Automation Letters*, 6(3): 5921-5928 [DOI: 10.1109/LRA.2021.3088076]
- Xu S X, Zhang C H, Fan L B, Meng G F, Xiang S M and Ye J P. 2025. AddressCLIP: empowering vision-language models for city-wide image address localization//*Proceedings of the 18th European Conference on Computer Vision*. Milan: Springer [DOI: 10.1007/978-3-031-73390-1_5]
- Xu W J, Yao Y X, Cao J Q, Wei Z W, Liu C B, Wang J N and Peng M G. 2024. UAV-VisLoc: a large-scale dataset for UAV visual localization. *arXiv*: 2405.11936
- Xue Z H, Zhou Y Y, Qiang Y G, Liu G F and Lin H. 2021. Cross-view scene image localization with Triplet Network integrating NetVLAD and Fully Connected Layers. *National Remote Sensing Bulletin*, 25(5): 1095-1107 (薛朝辉, 周逸飏, 强永刚, 刘弋锋, 林晖. 2021. 融合NetVLAD和全连接层的三元神经网络交叉视角场景图像定位. *遥感学报*, 25(5): 1095-1107) [DOI: 10.11834/jrs.20210188]
- Yagi R and Iwasaki A. 2023. Lightweight CNN for cross-view geo-localization using aerial image//*Proceedings of 2023 IEEE International Geoscience and Remote Sensing Symposium*. Pasadena: IEEE [DOI: 10.1109/IGARSS52108.2023.10281631]
- Yagi R, Yairi T and Iwasaki A. 2023. Navigating the metaverse: UAV-based cross-view geo-localization in virtual worlds//*Proceedings of the 2023 Workshop on UAVs in Multimedia: Capturing the World from a New Perspective*. Ottawa: ACM: 13-17 [DOI: 10.1145/3607834.3616568]
- Yan S, Cheng X Y, Liu Y X, Zhu J L, Wu R W, Liu Y and Zhang M J. 2023. Render-and-compare: cross-view 6-DoF localization from noisy prior//*Proceedings of 2023 IEEE International Conference on Multimedia and Expo*. Brisbane: IEEE [DOI: 10.1109/ICME55011.2023.00371]
- Yan Y M, Wang M Y, Su N, Hou W, Zhao C H and Wang W X. 2024. IML-Net: a framework for cross-view geo-localization with multi-domain remote sensing data. *Remote Sensing*, 16(7): 1249 [DOI: 10.3390/rs16071249]
- Yang H J, Lu X F and Zhu Y Y. 2021. Cross-view geo-localization with evolving transformer. *arXiv*: 2107.00842
- Yang H J, Lu X F and Zhu Y Y. 2024. Cross-view geo-localization with layer-to-layer transformer//*Proceedings of the 35th International Conference on Neural Information Processing Systems*. [s.l.]: Curran Associates Inc., 2222
- Zhai M H, Bessinger Z, Workman S and Jacobs N. 2017. Predicting ground-level scene layout from aerial imagery//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE [DOI: 10.1109/CVPR.2017.440]
- Zhang H H, Xie C, Toriya H, Shishido H and Kitahara I. 2023. Vehicle localization in a completed city-scale 3D scene using aerial images and an on-board stereo camera. *Remote Sensing*, 15(15): 3871 [DOI: 10.3390/rs15153871]
- Zhang W, Cai M X, Zhang T, Zhuang Y and Mao X R. 2024a. EarthGPT: a universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 5917820 [DOI: 10.1109/TGRS.2024.3409624]
- Zhang W H, Zhong Z N, Chen H and Jing N. 2024b. GeoViewMatch: a multi-scale feature-matching network for cross-view geo-localization using swin-transformer and contrastive learning. *Remote Sensing*, 16(4): 678 [DOI: 10.3390/rs16040678]
- Zhang X H, Li X Y, Sultani W, Zhou Y and Wshah S. 2023b. Cross-view geo-localization via learning disentangled geometric layout correspondence//*Proceedings of the 37th AAAI Conference on Artificial Intelligence*. Washington: AAAI: 3480 -3488 [DOI: 10.1609/aaai.v37i3.25457]
- Zhang X H, Sultani W and Wshah S. 2023c. Cross-view image sequence geo-localization//*Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision*. Waikoloa: IEEE [DOI: 10.1109/WACV56688.2023.00293]
- Zhang X W, Meng X C, Yin H L, Wang Y X, Yue Y Z, Xing Y H and Zhang Y N. 2022. SSA-net: spatial scale attention network for image-based geo-localization. *IEEE Geoscience and Remote Sensing Letters*, 19: 8022905 [DOI: 10.1109/LGRS.2021.3120658]
- Zhao H, Ren K Y, Yue T Y, Zhang C and Yuan S. 2024. TransFG: a cross-view geo-localization of satellite and UAVs imagery pipeline using transformer-based feature aggregation and gradient

- guidance. IEEE Transactions on Geoscience and Remote Sensing, 62: 4700912 [DOI: 10.1109/TGRS.2024.3352418]
- Zheng Z D, Wei Y C and Yang Y. 2020. University-1652: a multi-view multi-source benchmark for drone-based geo-localization//Proceedings of the 28th ACM International Conference on Multimedia. Seattle: ACM: 1395-1403 [DOI: 10.1145/3394171.3413896]
- Zhu M Z, Yang Y, Song W J, Wang M L and Fu M Y. 2020. AGCV-LOAM: air-ground cross-view based LiDAR odometry and mapping//Proceedings of 2020 Chinese Control And Decision Conference. Hefei: IEEE [DOI: 10.1109/CCDC49329.2020.9164147]
- Zhu R Z, Yang M Z, Yin L, Wu F and Yang Y C. 2023a. UAV's status is worth considering: a fusion representations matching method for geo-localization. Sensors, 23(2): 720 [DOI: 10.3390/s23020720]
- Zhu R Z, Yin L, Yang M Z, Wu F, Yang Y C and Hu W B. 2023b. SUES-200: a multi-height multi-scene cross-view image benchmark across drone and satellite. IEEE Transactions on Circuits and Systems for Video Technology, 33(9): 4825-4839 [DOI: 10.1109/TCSVT.2023.3249204]
- Zhu S J, Shah M and Chen C. 2022. TransGeo: transformer is all you need for cross-view image geo-localization//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE [DOI: 10.1109/CVPR52688.2022.00123]
- Zhu S J, Yang T J N and Chen C. 2021. VIGOR: cross-view image geo-localization beyond one-to-one retrieval//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE [DOI: 10.1109/CVPR46437.2021.00364]
- Zhu Y Y, Chen S H, Lu X F and Chen J Y. 2023c. Cross-view image synthesis from a single image with progressive parallel GAN. IEEE Transactions on Geoscience and Remote Sensing, 61: 4701513 [DOI: 10.1109/TGRS.2023.3257861]
- Zhu Y Y, Sun B, Lu X F and Jia S. 2022. Geographic semantic network for cross-view image geo-localization. IEEE Transactions on Geoscience and Remote Sensing, 60: 4704315 [DOI: 10.1109/TGRS.2021.3121337]
- Zhuang J D, Chen X R Y, Dai M, Lan W B, Cai Y H and Zheng E H. 2022. A semantic guidance and transformer-based matching method for UAVs and satellite images for UAV geo-localization. IEEE Access, 10: 34277-34287 [DOI: 10.1109/ACCESS.2022.3162693]
- Zhuang J D, Dai M, Chen X Y and Zheng E H. 2021. A faster and more effective cross-view matching method of UAV and satellite images for UAV geolocalization. Remote Sensing, 13(19): 3979 [DOI: 10.3390/rs13193979]

Review of cross-view image geo-localization datasets

ZHANG Xiao, GAO Yi, XIA Yuxiang, ZHAO Chunxue

Troops 61175, Nanjing 210049, China

Abstract: Cross-view image geo-localization aims to retrieve the most similar image from a reference database through matching images captured from different viewpoints, which subsequently leverages the GPS tag of images to fulfill localization tasks. Traditional single-view image geo-localization is limited by factors such as dataset quality, scale, and positioning accuracy. Therefore, in recent years, numerous researchers and institutions have released multiple cross-view geo-localization datasets, which lay the data foundation for improving geo-localization accuracy. However, systematic analysis of these cross-view image geo-localization datasets is still lacking. Therefore, this study aimed to provide a comprehensive review of published cross-view image geo-localization datasets. Based on the literature review, we collect and organize 32 cross-view image geo-localization datasets spanning from 2011 to 2024. We review 32 classic datasets that have emerged since the development of cross-view image geo-localization, which constructs a classification system from four dimensions: viewpoint information, construction type, authenticity, and temporal information. We summarize the basic information of these datasets in a tabular form, including the name, image resolution, data scale, and encompassed scenes of the dataset. Fully expressing the fundamental attributive characteristics of cross-view geolocation datasets from multiple perspectives. Then, we delve into these cross-view image geo-localization datasets from five aspects: metadata, influence, keywords, acquisition sources, and application fields. In addition, we collate and summarize the mainstream algorithms for cross-view image geo-localization (e.g., network structure optimization, loss function optimization, and attention mechanism). Finally, we discuss the future development directions of cross-view localization datasets from four perspectives: the trend of multimodal datasets, the approach of large language models, image distraction handling, and model optimization. In summary, we offer a comprehensive review of cross-view image geo-localization datasets from various perspectives. To the best of our knowledge, this study is the first to review such datasets in the field, which can provide a reference for researchers in related fields. However, the current datasets still involve issues such as low data quality, single source of data, and weak generalization ability. Thus, further research is needed.

Key words: cross-view, image geo-localization, datasets, deep learning, unmanned aerial vehicle, image retrieval, image matching, computer vision