自动化学报
ACTA AUTOMATICA SINICA

面向自动驾驶的跨视图定位技术综述

段星宇 吴涛 李峻翔 王耀南 刘流

A Review of Cross-view Localization Techniques for Autonomous Driving

DUAN Xing-Yu, WU Tao, LI Jun-Xiang, WANG Yao-Nan, LIU Liu

在线阅读 View online:
http://www.aas.net.cn/article/current/bbff30cdcefa26ea9512178597f5c2ac3ddb87944ae5f0a1bb4f448fe17fcc44

您可能感兴趣的其他文章

基于混合生成对抗网络的多视角图像生成算法
Cross-view Image Generation via Mixture Generative Adversarial Network
自动化学报. 2021, 47(11): 2623-2636 https://doi.org/10.16383/j.aas.c190743

面向可信自动驾驶策略优化: 一种对抗鲁棒强化学习方法
Toward Trustworthy Policy Optimization for Autonomous Driving: An Adversarial Robust Reinforcement Learning Approach
自动化学报. 2025, 51(11): 2473-2485 https://doi.org/10.16383/j.aas.c250193

基于深度学习的单幅图像超分辨率重建算法综述
A Review of Single Image Super-resolution Reconstruction Algorithms Based on Deep Learning
自动化学报. 2021, 47(10): 2341-2363 https://doi.org/10.16383/j.aas.c190859

图像异常检测研究现状综述
Review of Image Anomaly Detection
自动化学报. 2022, 48(6): 1402-1428 https://doi.org/10.16383/j.aas.c200956

面向行人重识别的局部特征研究进展、挑战与展望
Research Progress, Challenge and Prospect of Local Features for Person Re-Identification
自动化学报. 2021, 47(12): 2742-2760 https://doi.org/10.16383/j.aas.c190821

提示学习在计算机视觉中的分类、应用及展望
The Classification, Applications, and Prospects of Prompt Learning in Computer Vision
自动化学报. 2025, 51(5): 1021-1040 https://doi.org/10.16383/j.aas.c240177

面向自动驾驶的跨视图定位技术综述

段星宇 [1] 吴 涛 [1] 李峻翔 [1] 王耀南 [2,3] 刘 流 [1]

摘 要 跨视图定位技术正经历从通用框架向面向自动驾驶的专有化架构深刻转变. 为系统梳理这一领域, 首先回顾其从粗略检索到精准 3-DoF 姿态估计的发展脉络. 进而, 提出一个从"输入数据、模型架构、评价指标、损失函数"四个核心维度出发的分析框架, 以此对主流方法进行系统性解构与归因, 揭示不同技术路径的设计权衡. 此外, 详细评述自动驾驶跨视图定位数据集, 并指出当前评估在泛化能力考量上的不足; 为此, 构建新的包含结构化道路与越野场景的 CSC-Shanmao 数据集, 作为评估模型环境适应性的新基准, 并在其上验证了多种代表性方法, 结果表明面向自动驾驶设计的方法在精度与鲁棒性上优势显著, 但同时也暴露出精度、效率与泛化性之间的突出矛盾. 最后, 总结该技术在轻量化部署、多模态融合可靠性及评估体系等方面面临的严峻挑战, 并展望轻量化设计、可信评估基准及主动感知等未来关键研究方向.

关键词 跨视图定位; 自动驾驶; 深度学习; 计算机视觉

引用格式 段星宇, 吴涛, 李峻翔, 王耀南, 刘流. 面向自动驾驶的跨视图定位技术综述. 自动化学报, xxxx, **xx**(x): x−xx

DOI 10.16383/j.aas.c250243 **CSTR** 32138.14.j.aas.c250243

A Review of Cross-view Localization Techniques for Autonomous Driving

DUAN Xing-Yu[1] WU Tao[1] LI Jun-Xiang[1] WANG Yao-Nan[2,3] LIU Liu[1]

Abstract Cross-view localization technique is undergoing a profound shift from general-purpose frameworks to specialized architectures tailored for autonomous driving. To systematically review this field, we first trace its evolution from coarse retrieval to precise 3-DoF pose estimation. We then propose an analytical framework centered on four core dimensions (input data, model architecture, evaluation metrics, and loss functions) to deconstruct and attribute mainstream methods systematically, revealing the design trade-offs among different technical approaches. Furthermore, we provide a detailed review of autonomous driving cross-view localization datasets and identify the current evaluation shortcomings regarding generalization capability. To address this, a new dataset named CSC-Shanmao, containing both structured road and off-road scenarios, is constructed as a novel benchmark for assessing model environmental adaptability. Various representative methods are validated on this dataset, demonstrating that approaches designed specifically for autonomous driving exhibit significant advantages in accuracy and robustness, while also revealing prominent trade-off among accuracy, efficiency, and generalization. Finally, we summarize the critical challenges faced by this technique in lightweight deployment, reliability of multi-modal fusion, and evaluation systems, and discuss key future research directions focusing on lightweight design, trustworthy evaluation benchmarks, and active perception.

Keywords cross-view localization; autonomous driving; deep learning; computer vision

Citation Duan Xing-Yu, Wu Tao, Li Jun-Xiang, Wang Yao-Nan, Liu Liu. A review of cross-view localization techniques for autonomous driving. *Acta Automatica Sinica*, xxxx, **xx**(x): x−xx

高精度定位是自动驾驶系统在复杂环境中实现可靠导航的前提. 然而, 在城市峡谷、隧道或受干扰环境中, 全球导航卫星系统 (GNSS) 的信号常会衰减或失效, 导致其提供的定位精度无法满足自动驾驶厘米级控制的需求. 针对这一缺陷, 跨视图定位技术通过将车载摄像头获取的实时图像与带有地理参考的卫星视图或地图数据库进行特征匹配, 能够在 GNSS 失效时构建视觉与地理空间的几何对应关系, 从而提供连续、高精度的绝对位置估计, 为自动驾驶导航提供可靠的保障方案. 该技术最早起源于机器人导航 (robot navigation) 的研究, 目的是解决多传感器数据 (如视觉−激光雷达 (LiDAR)) 的空间对齐问题, 其本质是根据图像的信息得出图像所在的地理坐标.

收稿日期 2025-06-03 录用日期 2026-04-12
Manuscript received June 3, 2025; accepted April 12, 2026
湖南省自然科学基金 (2026JJ40065) 资助
Supported by Natural Science Foundation of Hunan Province (2026JJ40065)
本文责任编委 朱鹏飞
Recommended by Associate Editor ZHU Peng-Fei
1. 国防科技大学智能科学学院 长沙 410073 2. 湖南大学人工智能与机器人学院 长沙 410082 3. 湖南大学机器人视觉感知与控制国家工程研究中心 长沙 410082
1. College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073 2. School of Artificial Intelligence and Robotics, Hunan University, Changsha 410082 3. National Engineering Research Center of Robot Visual Perception and Control Technology, Hunan University, Changsha 410082

现有面向通用地理定位任务的跨视图定位技术 (后文统一简称为通用方法) 大多采用深度学习框架来搭建算法, 这些方法 [1−48] 通过匹配不同视角的图像 (如地面视角与卫星地图视角) 来预测拍摄者大致所处的地理位置, 通过得到这些与图像相关联的地理位置信息, 就能推断出许多有价值的信息, 比如拍摄者的活动轨迹甚至身份信息, 在公共安全、网络安全治理方面具有广泛的应用. 除此之外, 这种定位方式还在其他一些通用地理定位任务 (如景区导览、土地利用变化的宏观分析等) 中得到应用, 能够实现十几米至几十米精度的定位功能, 让操作者了解目标地大致所在区域.

跨视图定位技术的演进经历了从传统特征匹配到深度学习方法的范式跃迁. 早期方法依赖于手工设计特征 (如 SIFT、ORB) 及几何模型, 在视角差异有限的场景中有效, 但在应对地面与空中视图间巨大的视角、尺度与外观差异时, 其泛化能力与鲁棒性面临根本性挑战, 难以实现高精度匹配[49−50]. 自动驾驶场景对定位技术提出了极高要求: 在 GNSS 受限时, 仍需实现厘米级位置精度与度数级朝向精度的实时、可靠估计, 这一严苛需求放大了传统方法的瓶颈——对环境变化敏感、缺乏高层语义理解能力、无法端到端优化, 也驱动研究重心全面转向数据驱动框架.

与传统方法相比, 面向自动驾驶的深度学习范式在性能与适用场景上实现了根本性跨越. 传统方法的定位精度通常为数十米级, 适用于宏观地理信息检索; 而深度学习方法通过端到端学习跨视角的稠密语义特征, 直接输出车辆的精确 3-DoF 姿态, 将定位精度提升到厘米至分米级. 在能力范畴上, 深度学习方法 (比如文献 [51]) 实现了从“区域检索”到“精准实时姿态估计”的质变. 文献 [2, 33, 52−91] 借助 KITTI、Ford Multi-AV 等自动驾驶数据集, 加入车辆位姿这一先验信息, 验证了这类方法在真实行车环境中的实时性与可靠性, 证明了其相较于通用方法的显著优势.

尽管跨视图定位的通用方法已取得长足发展, 但在综述研究方面, 相关工作主要聚焦于通用跨视图地理定位任务. 例如, 周博文等[49] 和刘棟等[50] 对基于深度学习的通用跨视角图像匹配与检索方法进行系统梳理; Durgam 等[92] 则对包括传统方法与深度学习方法在内的技术演进进行全面调研. 然而, 上述综述的视角与自动驾驶场景下的高精度定位需求之间存在着显著差异. 通用方法的目标多在于城市级地理信息检索, 其评价指标 (如 Top-K 召回率) 和精度目标 (十米级) 均与自动驾驶厘米级定位、度数级朝向估计的严苛要求相去甚远.

与上述综述不同, 本文首次系统性地聚焦于“面向自动驾驶”这一特定高要求的应用场景, 旨在为这一新兴交叉领域提供专门的梳理与分析. 本文的核心贡献如图 1 所示, 可概括为以下四个方面:

一是场景驱动的新分类体系: 摒弃通用的技术流分类, 提出“输入数据、模型架构、评价指标、损失函数”四个自动驾驶任务驱动的核心维度, 对方法进行解构与分析, 揭示其设计逻辑与性能边界;

二是任务对齐的评估视角: 重点梳理在自动驾驶数据集 (如 KITTI-CVL、FordAV-CVL) 上评估的方法, 强调其 3-DoF 姿态估计能力, 明确区分“区域检索”与“精准定位”两种范式;

三是挑战与需求的深度剖析: 深入阐述自动驾驶场景下独有的挑战 (如实时性、轻量化、多模态融合的工程难题) 与未被满足的需求 (如对极端天气、场景泛化的鲁棒性);

四是基准数据集的构建与验证: 引入包含结构化与越野场景的 CSC-Shanmao 数据集, 为评估模型的泛化能力提供新基准, 并验证代表性方法的性能.

本文的结构安排如下: 第 1 节对本领域的发展历程与关键技术流程进行总结与梳理; 第 2 节从四个维度对主流方法进行分类剖析; 第 3 节梳理自动驾驶中的跨视图定位数据集与评价指标; 第 4 节介绍面向自动驾驶的跨视图定位方法在不同数据集上的实验对比结果; 第 5 节讨论跨视图定位技术面临的挑战并对未来研究进行展望; 第 6 节总结本文的主要内容和贡献.

1 发展历程与关键技术流程

面向自动驾驶的跨视图定位技术并非一蹴而就, 其演进深深植根于计算机视觉、机器人学与自动驾驶技术的交叉融合. 本节将简要回顾其从萌芽到专有化的发展脉络, 并勾勒出该任务的核心技术流程, 为后续对具体方法的深入剖析提供背景与框架.

1.1 技术发展历程

该领域的发展大致经历了三个阶段, 如图 2 所示.

第一阶段: 通用方法的初步应用 (2020 年以前). 此阶段的研究直接沿用为通用地理定位任务设计的跨视图匹配方法. 核心思想是将地面车载摄像头拍摄的图像视为“街头查询图像”, 通过检索与之最相似的卫星图像块来实现粗略定位. 方法多基于

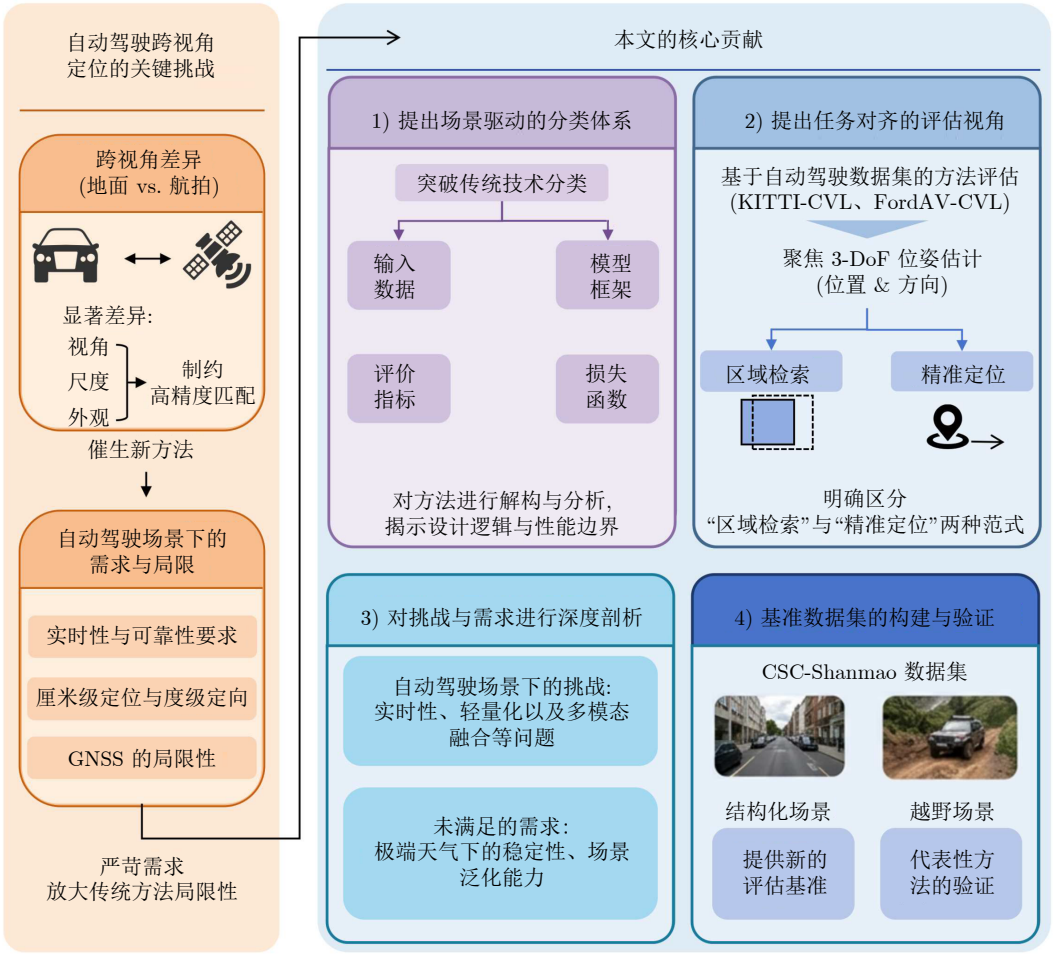


图 1 本文贡献总览

Fig. 1 Overview of contributions in this paper

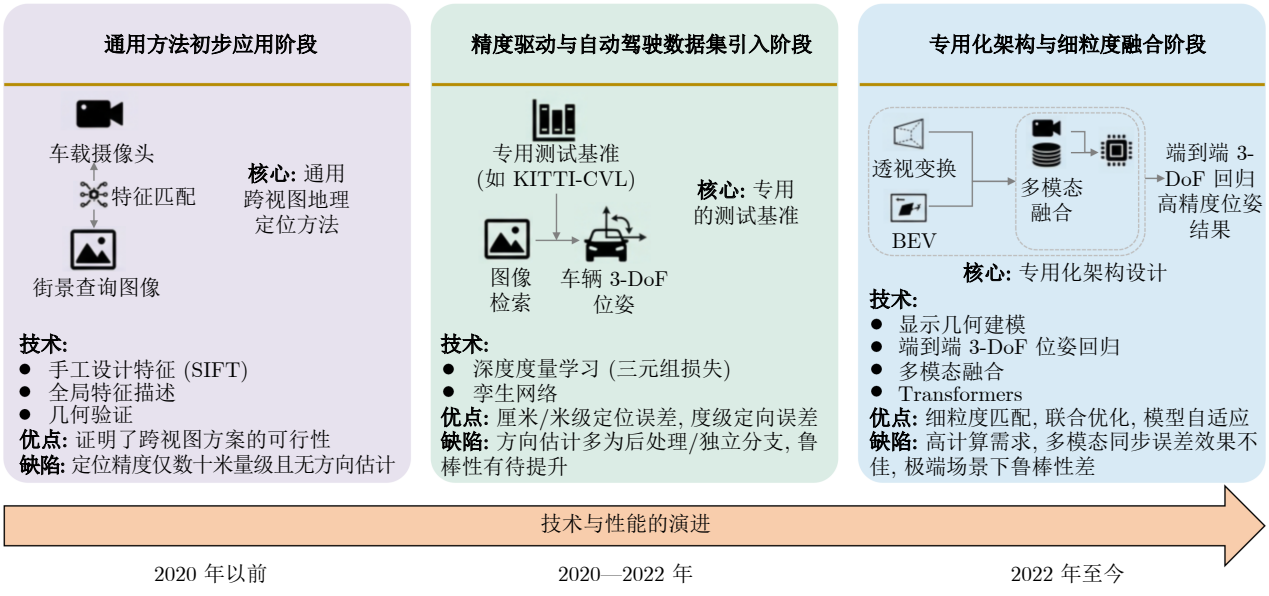


图 2 面向自动驾驶的跨视图定位方法的发展历程

Fig. 2 Evolution of cross-view localization methods for autonomous driving

全局图像描述子匹配或手工特征 (如 SIFT) 的几何验证, 定位精度在数十米量级, 且完全不估计车辆朝向. 这类工作证明了跨视图匹配用于车辆定位的可行性, 但精度与功能远未达到自动驾驶导航的需求.

第二阶段: 精度驱动与自动驾驶数据集的引入 (约 2020—2022 年). 随着自动驾驶公开数据集 (如 KITTI、FordAV) 的成熟, 研究者开始为其匹配卫星图像, 构建出 KITTI-CVL 等专用评估基准. 这一转变具有里程碑意义: 1) 评估指标从“Top-K 召回率”转向“厘米/米级位置误差”与“度数级方向误差”; 2) 任务目标从“图像检索”明确为“车辆 3-DoF 姿态估计”. 方法上, 基于深度度量学习 (如三元组损失) 的孪生网络成为主流, 通过拉近匹配的“地面–卫星”图像对特征、推远不匹配对的特征, 学习视角不变的表示. 然而, 此阶段方法仍普遍将方向估计作为后处理或独立分支, 精度与鲁棒性有待提升.

第三阶段: 专用化架构与细粒度融合 (约 2022 年至今). 为应对自动驾驶的极致要求, 研究进入专用化架构设计阶段. 其显著特征包括: 1) 显式几何建模: 引入可微的透视投影、单应性变换或 BEV 生成模块, 直接将地面视图特征映射至与卫星视图对齐的俯视空间, 实现像素级或区域级的细粒度匹配; 2) 姿态联合优化: 设计端到端网络, 直接回归车辆的 3-DoF 位姿, 或将定位与方向估计在统一框架下进行联合优化; 3) 多模态融合探索: 开始融合 LiDAR 点云等额外模态, 利用其精确的几何信息来增强纯视觉方法在纹理缺失、光照变化下的鲁棒性; 同时, Transformer 等新架构因其强大的全局上下文建模能力而被引入, 以解决大视角差异下的匹配难题.

上述发展脉络表明, 技术演进的核心驱动力是自动驾驶对高精度、强鲁棒姿态估计的迫切需求.

1.2 关键技术流程

尽管具体方法多样, 但一个典型的面向自动驾驶的跨视图定位系统通常遵循由输入数据、数据预处理, 到特征提取与跨视角匹配对齐, 最后输入位姿解算模块得到估计位姿的流程, 其核心在于解决“如何从异构的跨视图数据中精准、鲁棒地估计出车辆的自车位姿”这一问题.

下文将依据此流程中涉及的关键设计决策, 从输入数据、模型架构、评价指标及损失函数四个维度, 对现有方法进行系统性分类与深入分析.

2 方法分类与分析

由于绝大多数面向自动驾驶场景的跨视图定位方法都是基于深度神经网络构建的, 基于第 1 节对发展动因的梳理, 本节旨在构建一个系统性分析框架, 以厘清各类方法的设计逻辑. 本文将从决定方法能力边界的“输入数据”、实现特征转换的“模型架构”、定义优化目标的“评价指标”以及指导模型训练的“损失函数”四个核心维度展开交叉分析. 这一分类框架旨在刻画一个完整的技术设计链条: “输入数据”定义了模型所依赖的信息模态与形式, 是解决跨视图匹配问题的物理基础; “模型架构”决定了如何从输入中提取与融合特征, 是实现跨视图转换的核心计算引擎; “评价指标”明确了算法优化所追求的外在性能目标, 直接反映了自动驾驶任务对精度与鲁棒性的需求; “损失函数”则指导了模型在训练过程中内部参数的优化方向, 是将外部目标转化为可学习信号的关键桥梁.

这四个维度分别对应算法设计中不同且关键的决策层面, 在概念上职责清晰、各有侧重. 尽管在实际设计中它们互有联系, 但本文的多维度交叉审视能够避免单一归类的局限性, 更全面地揭示各类方法的设计哲学、优势权衡与技术演进的内在逻辑. 下文将以此框架对主流方法展开深入分析.

2.1 模型输入数据的差异

对于通用方法, 模型的输入数据通常为地面图像与带有地理坐标的卫星图像. 在面向自动驾驶场景的跨视图定位方法中, 具体又可以细分为地面图像+鸟瞰/卫星图像、地面图像+鸟瞰/卫星图像+点云数据以及鸟瞰/卫星图像+点云数据这 3 类. 下面将从这 3 种不同的输入数据类别展开介绍这些方法.

2.1.1 地面图像+鸟瞰/卫星图像

这是研究最广泛、系统复杂度最低的输入模式, 其核心是建立地面视觉观测与卫星地理上下文之间的关联, 不依赖额外的测距传感器 (如 LiDAR). 这类方法使用的地面图像有两种: 一种是地面前视图, 另一种是地面全景图. 模型输入为地面全景图与鸟瞰/卫星图像的具体方法见文献 [53–55, 57–59, 62, 64–66, 70, 78, 86–89], 模型输入为地面前视图与鸟瞰/卫星图像的方法见文献 [56, 60–61, 63, 68, 72, 75–76, 91]. 另外, 方法 CCVPE[67] 与 C2F-CCPE[77] 在实际测试过程中两种类型的输入都有使用.

这一划分结果并不是指其方法只能用地面前视图或全景图作为输入, 而是与其论文中所使用的数据集有关: 使用地面全景图作为输入的模型大多使用 CVUSA、CVACT 以及 VIGOR 等跨视图定位通用基准数据集作为实验数据集, 而使用地面前视

图作为输入的模型都是使用 KITTI、Ford Multi-AV 以及 Oxford RobotCar 等自动驾驶数据集作为实验数据集.

基于上述不同的数据基础, 方法的研究侧重点也有所不同.

在通用数据集上, 研究侧重于提升跨视图匹配的鲁棒性与精度. 例如, Liu 等[51] 率先在 CVUSA 数据集中引入方向先验, 通过方向对齐提升匹配精度, 启发了对几何线索的利用. HADGEO[78] 则采用困难样本挖掘 (HSM) 策略, 增强了模型对难例的判别能力, 从而在 CVUSA 等数据集上达到了领先的检索精度. 然而, 这些方法因训练数据缺乏自动驾驶场景的复杂性, 其是否能直接应用于动态、复杂的行车环境仍有待验证.

在自动驾驶数据集上, 研究则直接聚焦于车辆 3-DoF 姿态的精确、鲁棒估计. 例如, LM[60]、CCVPE[67] 等方法直接利用数据集中提供的精确位姿真值进行端到端训练. 另一些工作则探索与其他定位模块的融合, 如 SLAM-G2S-Fusion[57] 结合视觉 SLAM 的局部精度与跨视图匹配的全局一致性, 以减少累积误差. BEVRender[70] 专注于非结构化场景, 通过生成局部鸟瞰图 (BEV) 来与卫星图进行匹配, 但其对季节性变化的泛化能力仍需提升.

纯视觉输入模式 (地面+卫星) 无需依赖 LiDAR 等高成本硬件, 主要数据源 (卫星图) 易于获取, 因此具有系统成本低、可扩展性好的优势; 但是地面视图图像容易受天气、光照以及季节变化等因素的影响, 导致特征提取效果不佳. 尽管如此, 仅使用地面视图图像与卫星图像的方法仍具有较大的研究价值和应用前景, 尤其是在一些对成本敏感且环境相对稳定的自动驾驶场景中得到广泛应用, 这也是目前大多数方法所采用的数据输入模式.

2.1.2 地面图像+鸟瞰/卫星图像+点云数据

考虑到纯视觉输入的方法可能受到不同天气和场景条件影响而导致定位能力不稳定, 特别是在军用无人车所处环境中, 由于卫星信号受限情况多发, 单一的 ground-air 模态可能无法满足高精度的定位需求, 这时通常就需要与其他传感器数据 (如惯性测量单元 (IMU)、LiDAR 等) 融合, 以提高定位系统的鲁棒性和适应性. 使用此类数据输入的方法见文献 [52, 71, 90–91]. SIBCL[52] 能在给定一个空间一致的卫星图像的情况下, 使用单个地面图像与对应的 3D LiDAR 点云匹配估算出准确的车辆 3-DoF 姿态, 实现空间误差小于 1 m、方向误差小于 1° 的效果. Geo-tracking[71] 方法利用与地面数据同步捕获的 LiDAR 点云, 将地面图像中提取的特征投影到空中图像上, 实现了米级的连续地理追踪. 3DGS[90] 通过将街景图像转换为伪鸟瞰图像, 有效减少了地面视角与卫星视角之间的特征差异, 实验结果表明, 该方法在 KITTI 数据集上显著提升了跨视角匹配性能, 尤其在大视角差异和复杂场景下表现优异, 并在处理遮挡、几何一致性优化和计算效率方面展现了强大鲁棒性, 为自动驾驶及其他智能系统提供了技术支持. 为应对视觉与 LiDAR 传感系统中的时序误差, CPC-CVPE[91] 提出一种基于时间戳近似的多模态传感器同步策略, 该方法通过将 LiDAR 点云数据作为时间基准, 降低了硬件同步复杂性, 同时实现了 98.2% 的有效对齐精度.

上述方法面向的任务有所区别, 比如 Geo-tracking[71] 是为了解决长期轨迹跟踪中的误差累积问题, 而 SIBCL[52]、3DGS[90]、CPC-CVPE[91] 更注重于对车辆的精确姿态估计. 虽然 SIBCL 在方位误差方面表现更好, 但其在位置误差上仍有很大的改进空间, 并且该方法依赖于 LiDAR 深度信息, 使得其应用场景受到限制. 这类方法融合了地面图像的局部细节、卫星图像的大范围地理信息以及点云数据的高精度三维结构, 能够更全面准确地构建环境模型, 从而显著提升定位的精度与系统的鲁棒性, 尤其是在复杂环境和军事场景中能够更好地应对各种挑战.

然而, 这种多模态融合方案在工程落地层面仍面临严峻挑战. 首先是极为严苛的时空校准要求: 视觉与 LiDAR 特征有效融合的前提是像素级或体素级的精确对齐. 在实际系统中, 相机与 LiDAR 之间的外参标定极易受车辆振动、温度变化影响而发生微小漂移, 导致特征错位; 同时, 不同传感器采样频率不一致 (如相机 30 Hz 与 LiDAR 10 Hz) 带来的时间同步误差, 在车辆高速运动时会转化为显著的位置偏差, 严重破坏融合的一致性. 其次是计算开销与系统鲁棒性矛盾: 引入 LiDAR 点云处理分支 (通常基于 Voxel 或 Point-based 网络) 会成倍增加显存占用与推理延迟, 对边缘计算设备的带宽与算力构成巨大压力. 此外, 过度依赖多模态融合可能导致系统在单一传感器失效 (如摄像头被遮挡或雷达受干扰) 时出现“灾难性遗忘”或性能断崖式下跌, 这对融合机制的容错性与模态退化处理能力提出了极高要求.

2.1.3 鸟瞰/卫星图像+点云数据

仅使用鸟瞰/卫星图像+点云数据的输入模式相比于三模态输入模式而言, 减少了对地面视角图像的采集和处理需求, 在一定程度上降低了系统的复杂性和数据处理负担 (如文献 [33, 73–74]). 虽然

该模式省去了相机的标定与同步环节, 在一定程度上降低了硬件集成难度, 但其核心挑战转移到了稀疏点云与稠密卫星图的异构特征匹配上. 由于缺乏地面视觉的纹理与颜色信息辅助, 仅依靠几何结构进行匹配时, 在几何特征单一或空旷的场景 (如长直隧道、开阔平原) 中极易产生定位歧义.

Hu 等[33] 提出的方法利用卫星图像和地面点云抽象的道路结构作为特征匹配的输入, AGL-Net[74] 利用本地地面 LiDAR 和空中视图作为输入, 这样的输入使得方法能避免一些由于天气变化或场景因素的影响而导致的定位偏差, 在一些特殊场景下能起到很好的效果. 但是这种输入模式由于缺少地面图像的纹理和语义信息, 可能导致在复杂场景下定位的准确性和鲁棒性不足.

综合前面的介绍与分析可知: 输入模态的选择本质上是信息完备性、获取成本与系统复杂性之间的权衡. 纯视觉 (地面+卫星) 方案成本最低、适用最广, 是当前的主流方案, 但其性能受环境外观变化制约. 引入 LiDAR 点云能提供几何不变性, 显著提升模型在恶劣天气与复杂场景下的鲁棒性, 但代价是高昂的传感器成本、严苛的标定要求与融合计算开销. 仅使用鸟瞰图与点云的方案则放弃了地面视角的丰富纹理, 虽简化了系统, 却在纹理缺失场景下面临匹配歧义. 因此, 在后续的研究中, 自适应模态选择与更高效的轻量级融合架构将是突破此权衡的关键.

2.1.4 不同模态的融合机理与互补性分析

上述三类输入模式的选择, 本质上是对“几何精度”、“语义丰富度”与“获取成本”三者间的不同权衡, 其融合旨在实现模态间的优势互补. 深层融合机理主要体现在两个层面:

一是特征层融合. 这是最主流的方式, 旨在学习一个跨模态的共享特征空间. 例如, 在“地面图像+鸟瞰/卫星图像+点云数据”模式中, 点云提供的精确 3D 几何结构被用作“空间锚点”. 通常先将 LiDAR 点云投影到鸟瞰图 (BEV) 或地平面, 生成一个与卫星图像空间对齐的几何特征图. 地面图像的丰富纹理特征和卫星图像的全局上下文特征则通过卷积或 Transformer 网络提取后, 在 BEV 空间中通过注意力机制与该几何特征图进行对齐与融合. 点云的几何不变性有助于校正由视角和光照引起的视觉特征畸变, 而视觉信息则为冰冷的几何结构赋予语义, 从而在复杂场景 (如植被区) 中实现更鲁棒的匹配.

二是决策层融合. 这种方式保留了各模态的独立性, 在后期进行结果集成. 例如, 可以分别用“地面+卫星”和“卫星+点云”两个子系统进行定位, 然后基于各自的置信度分数或几何一致性对位姿估计进行加权或筛选. 这种方式更具灵活性, 当某一模态 (如相机被遮挡) 失效时, 系统可降级依赖另一模态, 但其难点在于如何设计精准的跨模态置信度评估机制.

因此, 选择何种输入模式及融合策略, 取决于任务对精度、鲁棒性、实时性及成本的综合要求. 精细的特征层融合能最大化性能, 但对数据标定与计算资源要求高; 决策层融合更易于实现容错, 但可能损失潜在的性能上限.

2.2 模型网络架构的差异

2.2.1 卷积神经网络架构

卷积神经网络 (convolutional neural network, CNN) 是最常见的深度学习模型架构, 主要用于处理具有网格结构的数据, 如图像和视频. CNN 的核心思想是利用卷积操作自动提取数据中的空间特征, 常见的 CNN 架构有 VGGNet、ResNet 和 EfficienNet 等.

VGGNet 包含很多级别的网络, 深度从 11 层到 19 层不等, 其采用的是一种预训练的方式, 即先训练浅层的简单网络 VGG11, 再复用 VGG11 的权重初始化 VGG13, 如此反复训练并初始化 VGG19, 使得训练时的收敛速度更快. 得益于其结构简单且特征提取能力强的特点, VGGNet 在跨视图定位的许多方法 [52, 54–55, 62, 65, 89] 中都有应用. 需要注意的是, VGG16 网络的计算量较大、计算速度慢并且模型体积较大, 因此容易影响模型的性能和泛化能力, 对计算资源提出了较高的要求. 比如 CVL-GSA[89] 采用 VGG16 作为骨干网络, 通过选择性地关注几何可靠的区域、优化相机姿态估计, 显著提高了跨视角地理定位的准确性和稳定性.

EfficienNet 也属于 CNN 架构, 它的核心思想是在网络设计中平衡网络的深度、宽度和分辨率以达到最佳的性能和效率. EfficienNet 包括多个变体, 如 EfficienNet-B0 到 EfficienNet-B7. 方法 MapLocNet[53] 和 HC-Net[59] 使用 EfficienNet-B0 作为主干网络, 这样的主干网络使得方法的参数量大大减少, 在相同的计算资源下能实现更高的性能, 适合硬件资源受限的场景, 但是其在移动 GPU 和 EdgeTPU 等加速器上的算子可能不够完善, 需要进行模型优化才能更好地在边缘设备上运行.

AGL-Net[74] 是一种自定义的 CNN 架构, 通过单独的编码器处理 LiDAR 点云和卫星图, 分别使用预训练的 ResNet-50 和 VGG19 作为特征提取

器, 以生成全局地图特征. AGL-Net 有效匹配了 LiDAR 数据和卫星图像, 并解决了这两种模态之间的尺度差异. Geo-tracking[71] 分别使用 ConvNeXt-T 和 UperNet 作为编码器与解码器处理航空图像和地面图像的特征, 充分利用了这两种网络在处理视觉任务上的潜力, 更适于高分辨率图像的任务.

CNN 架构 (如 VGG、ResNet) 通过局部卷积核提取层次化特征, 擅长捕捉纹理、边缘等局部视觉模式, 这对于从地面和卫星图像中识别共有的局部地标 (如窗户、斑马线) 至关重要. 然而, 其在处理巨大的跨视图尺度与视角变化时存在固有局限: 标准的池化操作会损失精确的空间位置信息, 而感受野有限导致其难以建立长距离的、非局部的特征对应 (如地面一条道路的曲线与卫星图中对应的大尺度道路形状). 因此, 纯 CNN 方法常依赖额外的几何建模 (如极坐标变换) 来缓解视角差异.

2.2.2 孪生神经网络架构

孪生神经网络 (Siamese neural networks) 是一种特殊的神经网络架构, 主要用于处理两个输入之间的相似性度量问题, 这一应用与跨视图定位方法有很强的相关性, 能有效提取不同视角之间的特征并学习相似性, 对于视角差异比较鲁棒, 因此也有学者使用孪生神经网络作为主干网络, 比如文献 [58, 67, 78, 87].

其中, CVML[58] 使用的是类孪生神经网络, 即网络结构中不共享权重. 另外两个方法 CCVPE[67] 与 HADGEO[78] 都根据各自的设计进行了共享权重这一步骤. 比如 CCVPE[67] 虽然在特征提取器的分支之间没有共享权重, 但在空地特征投影模块使用了共享全连接层, 让模型在跨视图定位任务上更快收敛. 使用孪生神经网络也存在一定的局限性, 其训练对数据的需求较大, 并且调参较为困难, 容易受视角干扰, 从而影响相似性度量的置信度. Geo-SSK[87] 通过设计基于交叉注意力的跨视图融合交互模块 (CIM), 学习地面图像与航拍图像之间的局部语义相似性, 从而增强模型的全局特征表示能力. 同时, 采用语义相似性知识蒸馏 (SSKD) 技术, 将教师模型中的知识传递给学生模型, 并引入错误知识过滤器 (IKF) 来提升蒸馏效果. 但是教师模型的引入会额外增加计算资源需求, 并且极坐标转化预处理方式存在限制, 可能不适用于某些视角极端偏离的场景.

标准的孪生网络通常进行全局特征匹配, 这种方式难以建模两个视图间复杂的、非线性的空间对应关系. 当面临巨大视角差异 (如地面 45° 斜视与卫星垂直俯瞰) 时, 简单的全局相似性度量容易失效. 此外, 其性能高度依赖训练数据的质量和数量, 需要大量精准对齐的跨视图图像对来学习有效的度量空间, 且对训练样本中未出现的极端视角或尺度泛化能力有限. 因此, 先进的孪生网络方法常需引入注意力机制、几何变换层 (如空间变换网络) 或局部特征匹配模块, 以增强其对细粒度几何关系的建模能力.

2.2.3 Transformer 架构

Transformer 架构的自注意力机制使得模型能有效捕捉到全局特征, 使用 Transformer 模型架构的有文献 [66, 70, 86, 88, 90]. 其中 Song 等[66] 所提出的模型基于 RAFT (recycle attention feature Transformer) 架构构建, 其主要差异在于地面特征投影模块和 BEV 特征细化模块, 此外他们还提出一种可微的最小二乘姿态回归模块, 用于筛选匹配的内点并用密集流场计算相机姿态. 该模型虽在精度上表现突出, 但其自注意力机制带来的二次方计算复杂度, 使其在部署于边缘设备时面临显著的实时性挑战.

BEVRender[70] 主要用于在无 GNSS 信号的野外环境中生成局部 BEV 图像以实现地面车辆的定位. 其采用基于 ViT (vision Transformer) 的注意力学习网络, 通过输入的地面图像生成对应的 BEV 图像以实现地面到空中两种视角图像的匹配, 相比于其他方法更强调逐帧匹配并减少了对全局轨迹优化的依赖性. 但是该方法的训练与测试没有充分考虑季节变化导致的环境条件波动问题, 因此特征泛化的能力还有提升的空间.

针对 HC-Net[59] 等方法过度依赖静态几何投影、难以应对复杂几何畸变的问题, Xia 等[86] 提出一种基于 Transformer 查询机制的细粒度定位网络——FG^2. 该方法不再使用固定的单应性变换, 而是初始化一组可学习的 BEV 查询向量, 通过 6 层级联的几何感知注意力模块, 利用相机内参将 3D 空间中的查询点投影至 2D 地面图像平面进行特征采样, 实现了从地面视图到 BEV 空间的动态特征聚合. 3DGS[90] 利用 COLMAP 生成初始相机位姿和稀疏点云, 并结合 3DGS 实现高效的伪鸟瞰图像合成. 实验结果表明, 该方法在 KITTI 数据集上表现出色, 尤其在大视角差异和复杂场景下展现了强大的鲁棒性和实用性.

Transformer 架构在全局表示方面具有明显优势, 其核心是自注意力机制, 能建立图像内任意两个像素间的全局依赖关系, 这使其天生擅长解决跨视图匹配中的“全局上下文对齐”问题. 它可以将地面图像中的一个局部区块与卫星图像中所有可能的

区域进行相似度计算, 从而在存在大量视觉歧义 (如重复纹理) 时, 依靠全局上下文做出更可靠的匹配. 然而, 这种强大的能力以巨大的计算复杂度为代价, 且其对精确的几何关系建模能力较弱, 通常需要与 CNN 或显式的几何模块 (如位置编码、可变形注意力) 结合, 以同时保证匹配的语义正确性与几何合理性.

2.2.4 U-Net

U-Net 具有独特的对称编码器–解码器结构, 其利用跳跃连接将编码器部分的特征图直接与解码器部分对应层次的特征图连接, 能够有效解决 CNN 在下采样过程中细节信息丢失的问题, 从而提取到更精细的空间信息和语义信息. 具体使用 U-Net 作为主干网络的有文献 [57, 60–61, 63–64, 68, 77].

但是由于 U-Net 含有大量的卷积层和跳跃连接, 导致其计算过程相对复杂, 对硬件的计算能力提出了较高的要求, 使用 U-Net 作为主干网络的跨视图定位方法需要高性能的硬件设备支持以应对自动驾驶场景下的实时性要求.

由前述介绍与分析可以发现: 网络架构的演进反映了对跨视图匹配问题本质理解的深化. CNN 善于提取局部特征, 但建模全局上下文的能力有限; 孪生网络天然适合度量学习, 但对细粒度几何关系建模不足; Transformer 凭借自注意力机制擅长全局语义对齐, 但计算复杂度高且对精确几何不敏感; U-Net 类结构则在保留空间细节上具有优势. 未来的研究可以利用混合架构, 例如用 CNN 提取局部特征, 再用 Transformer 进行全局关联, 或将几何变换层嵌入网络中, 以同时实现语义匹配的鲁棒性与几何对齐的精确性.

图 3 从“输入模态”与“网络架构”两个基础维度, 对主流方法进行了系统性归类与可视化, 为后续结合“评价指标” (第 2.3 节) 与“损失函数” (第 2.4 节) 的交叉分析提供了清晰的基准框架.

2.3 模型评价指标的变化

在跨视图定位问题中, 一些用 CVUSA、CVACT 和 VIGOR 等传统跨视图定位基准数据集进行评估实验的方法, 采取的是 Top-K 召回率指标来衡量定位系统的性能. 这里的 Top-K 召回率是指在给定一组查询数据和对应的候选匹配数据中, 目标位置出现在前 K 个推荐结果中的比例.

另外, 大多数面向自动驾驶的方法通常用位置与方向上的误差召回率这两个一级指标来评估模型的优劣, 其中位置误差包括横向位置误差和纵向位置误差, 方向误差是指估计方向与真实方向之间的差值. 为更准确地衡量定位系统的准确性, 研究者通常用中位或平均误差作为性能评估指标, 其中中位误差可以被用来定义一个误差圆, 帮助人们更直观地理解定位精度; 平均误差是指在一组定位数据中, 每个定位点与真实位置之间误差的平均值.

面向自动驾驶的跨视图定位方法在评估自身算法性能时, 用上面提到的模型评价指标进行评估测试, 通过各种对比实验体现出自身的优越性. 下面将依据所使用模型评价指标的区别对这些方法进行更精细的分类, 注意这里只是对方法进行分类, 更为详细的对于数据集以及各方法的指标对比将于第 3 节重点阐述.

2.3.1 Top-K 召回率

使用 Top-K 召回率作为模型评价指标的有文献 [8, 51, 55, 61–62, 78, 87–88]. 表 1、表 2 中的 r@1 表示查询图像的第 1 个检索结果就是正确匹配卫星图像的比例, r@5 和 r@10 同理, r@1(%) 表示查询图像的前 1% 候选卫星图像中包含正确匹配的比例. 从表 1、表 2 可以看出, GeoSSK[87] 在传统跨视图定位数据集 CVUSA 的测试中性能最优, 而 DSTG[88] 在 CVACTval 的测试中属于最优模型.

2.3.2 定位与方向估计的误差及召回率

使用定位误差召回率+平均/中位误差作为模型评价指标的有文献 [33, 52–54, 56–60, 63–68, 70–71, 74, 77, 86, 89, 91], 这些文献大多在 KITTI、Ford Multi-AV 这些自动驾驶基准数据集上通过增加卫星图像形成新的面向自动驾驶的跨视图定位数据集, 如 KITTI-CVL 和 FordAV-CVL. 在评估模型性能时, 不同的模型即使在使用同一数据集进行评估时, 也会由于引入噪声的不一致而导致结果不一, 因此在下文进行汇总时会考虑到测试时的噪声差异问题.

如表 3 所示, 这些模型的性能对比结果分别是在 KITTI-CVL 上引入不同范围的方向噪声和横/纵向位置噪声得到的.

依据表 3, 本文使用图 4、图 5 分别展示不同模型在同区域/跨区域测试中方向扰动较小情况下 ($\pm$ 10°) 的性能对比情况. 从这些图表可以得出, 在基于 KITTI 数据集进行测试的模型中, T2GA[68] 在小范围或大范围的初始先验误差下都能达到较优的性能, 在几乎所有模型都由于跨区域测试而出现性能下降的情况时, T2GA[68] 在应对跨区域测试时的性能表现较好, 没有出现明显的精度下降情况; C2F-CCPE[77] 在方向估计这一指标上最为稳定, 并且一直处于最优水平. 具体分析如下:

在同区域测试中, T2GA[68] 与 C2F-CCPE[77] 在

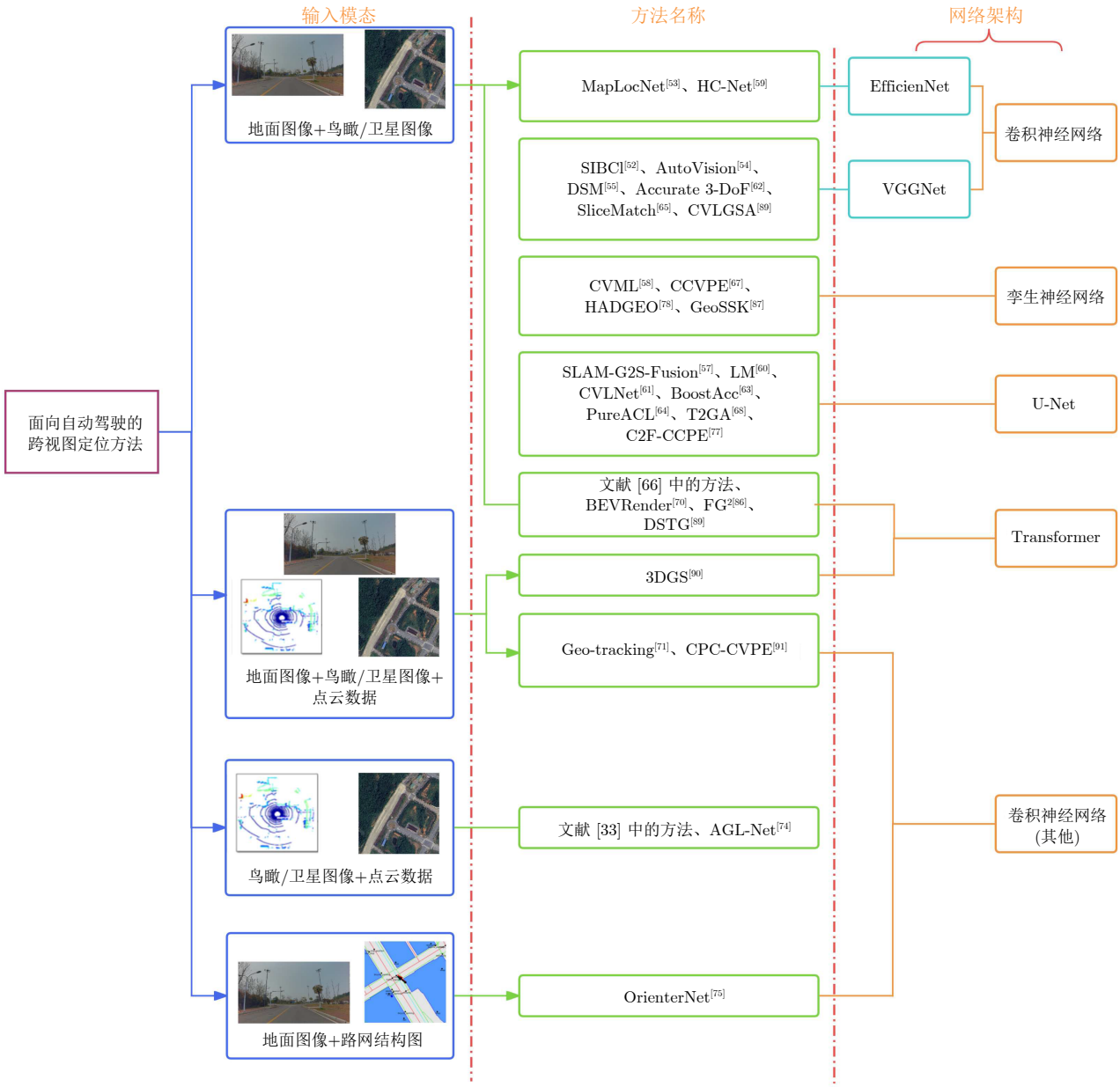


图 3　面向自动驾驶的跨视图定位方法一览

Fig. 3　Overview of cross-view localization methods for autonomous driving

表 1　CVUSA 数据集上的跨视图定位方法性能对比 (%)

Table 1　Performance comparison of cross-view localization methods on the CVUSA dataset (%)

方法名称	r@1	r@5	r@10	r@1(%)
CVM-Net[8]	22.47	49.98	63.18	93.62
文献 [51] 中的方法	40.79	66.82	76.36	96.12
DSM[55]	91.96	97.50	98.54	99.67
Accurate 3-DoF[62]	92.69	97.78	98.60	99.61
HADGEO[78]	95.01	98.45	99.10	—
GeoSSK[87]	94.13	**98.75**	**99.27**	**99.85**
DSTG[88]	**95.22**	98.66	99.13	99.79

表 2　CVACTval 数据集上的跨视图定位方法性能对比 (%)

Table 2　Performance comparison of cross-view localization methods on the CVACTval dataset (%)

方法名称	r@1	r@5	r@10	r@1(%)
CVM-Net[8]	20.15	45.00	56.87	87.57
文献 [51] 中的方法	46.96	68.28	75.48	92.01
DSM[55]	82.49	92.44	93.99	97.32
Accurate 3-DoF[62]	82.70	92.50	94.24	97.65
HADGEO[78]	**86.48**	94.21	95.50	—
GeoSSK[87]	85.56	94.93	96.17	98.54
DSTG[88]	86.29	**95.43**	**96.85**	**98.81**

表 3 KITTI-CVL 数据集上不同实验条件下的跨视图定位方法性能对比

Table 3 Performance comparison of cross-view localization methods under different experimental settings on the KITTI-CVL dataset

方法	定位 ↓ (m)		横向 ↑ (%)					纵向 ↑ (%)					方向 ↑ (%)					方向 ↓ (°)	
	平均	中位	0.25 m	0.5 m	1 m	3 m	中位 ↓	0.25 m	0.5 m	1 m	3 m	中位 ↓	1°	2°	3°	4°	5°	平均	中位
方向先验噪声为 ±15°, 横/纵位置先验噪声为 ±5 m																			
SIBCL[52]	—	—	25.59	46.26	72.63	—	0.54	21.91	41.22	64.47	—	0.64	**56.05**	**79.70**	—	**90.89**	—	—	**0.85**
DSM[55]	—	—	8.51	13.97	23.44	—	3.48	7.67	11.02	20.03	—	4.97	3.13	8.29	—	17.44	—	—	12.73
CVML[58]	—	—	6.11	12.24	23.78	—	2.40	5.97	11.68	23.73	—	2.46	—	—	—	—	—	—	—
LM[60]	—	—	16.51	32.05	57.65	—	0.83	7.14	14.11	27.41	—	2.01	29.83	53.41	—	76.51	—	—	3.23
PureACL[64]	—	—	**84.58**	**99.54**	**99.98**	—	**0.12**	**98.55**	**100.00**	**100.00**	—	**0.09**	31.18	54.13	—	76.00	—	**3.57**	1.78
方向先验噪声为 ±10°, 横/纵位置先验噪声为 ±20 m (同区域测试)																			
DSM[55]	—	—	—	—	10.12	30.67	—	—	—	4.08	12.01	—	3.58	—	13.81	—	24.44	—	—
HC-Net[59]	0.80	0.50	—	—	99.01	—	—	—	—	92.20	—	—	91.35	—	—	—	99.84	0.45	0.33
LM[60]	12.08	11.42	—	—	35.54	70.77	—	—	—	5.22	15.88	—	19.64	—	51.76	—	71.72	3.72	2.83
BoostAcc[63]	10.01	5.19	—	—	76.44	96.34	—	—	—	23.54	50.57	—	99.10	—	**100.00**	—	**100.00**	0.55	0.42
PureACL[64]	2.42	0.42	—	—	91.95	93.40	—	—	—	91.86	92.12	—	32.17	—	53.89	—	71.66	3.97	1.77
SliceMatch[65]	7.96	4.39	—	—	49.09	—	—	—	—	15.19	—	—	13.41	—	—	—	64.17	4.12	3.65
文献 [66] 中的方法	1.48	0.47	—	—	95.47	—	—	—	—	87.89	—	—	89.40	—	—	—	99.31	0.49	0.30
CCVPE[67]	1.22	0.62	—	—	97.35	98.65	—	—	—	77.13	96.08	—	77.39	—	99.47	—	99.95	0.67	0.54
T2GA[68]	**0.20**	**0.17**	—	—	**99.97**	**99.97**	—	—	—	**99.97**	**99.97**	—	50.95	—	83.97	—	95.45	1.53	0.93
C2F-CCPE[77]	0.92	0.44	—	—	98.52	99.02	—	—	—	88.97	96.77	—	**99.97**	—	99.99	—	**100.00**	**0.01**	**0.01**
FG^2[86]	0.75	0.52	—	—	99.73	—	—	—	—	86.99	—	—	61.17	—	—	—	95.65	1.28	0.74
CVLGSA[89]	1.15	0.35	—	—	98.22	99.04	—	—	—	92.47	96.13	—	99.86	—	**100.00**	—	**100.00**	0.28	0.23
CPC-CVPE[91]	1.03	0.55	—	—	97.87	—	—	—	—	86.21	—	—	86.21	—	—	—	97.98	1.01	0.65
方向先验噪声为 ±10°, 横/纵位置先验噪声为 ±20 m (跨区域测试)																			
DSM[55]	—	—	—	—	10.77	31.37	—	—	—	3.87	11.73	—	3.53	—	14.09	—	23.95	—	—
HC-Net[59]	8.47	4.57	—	—	75.00	—	—	—	—	58.93	—	—	33.58	—	—	—	83.78	3.22	1.63
LM[60]	12.58	12.11	—	—	27.82	59.79	—	—	—	5.75	16.36	—	18.42	—	49.72	—	71.00	3.95	3.03
BoostAcc[63]	13.01	9.06	—	—	57.72	86.77	—	—	—	14.15	34.59	—	**98.98**	—	**100.00**	—	**100.00**	0.56	0.43
PureACL[64]	6.20	0.61	—	—	67.24	87.26	—	—	—	64.81	73.63	—	23.45	—	48.69	—	59.24	4.26	2.48
SliceMatch[65]	13.50	9.77	—	—	32.43	—	—	—	—	8.30	—	—	46.82	—	—	—	46.82	4.20	6.61
文献 [66] 中的方法	7.97	3.52	—	—	54.19	—	—	—	—	23.10	—	—	43.44	—	—	—	89.31	2.17	1.21
CCVPE[67]	9.16	3.33	—	—	44.06	81.72	—	—	—	23.08	52.85	—	57.72	—	92.34	—	96.19	1.55	0.84
T2GA[68]	**0.21**	**0.18**	—	—	**99.92**	**99.96**	—	—	—	**99.91**	**99.91**	—	38.31	—	80.12	—	92.48	2.04	1.38
C2F-CCPE[77]	8.16	3.14	—	—	49.55	87.39	—	—	—	20.96	55.20	—	98.85	—	98.98	—	99.02	**0.26**	**0.01**
FG^2[86]	7.45	4.03	—	—	89.46	—	—	—	—	12.42	—	—	30.34	—	—	—	81.17	3.33	1.88
CVLGSA[89]	9.50	4.34	—	—	60.00	91.24	—	—	—	21.36	44.68	—	99.53	—	**100.00**	—	**100.00**	0.31	0.24
方向先验噪声为 ±45°, 横/纵位置先验噪声为 ±20 m (同区域测试)																			
SIBCL[52]	—	—	17.01	33.34	61.88	—	0.70	20.68	31.40	55.58	—	0.86	35.50	55.04	—	71.06	—	8.90	1.69
BoostAcc[63]	—	—	21.92	41.51	67.72	—	0.63	5.17	9.38	16.54	—	5.42	18.45	35.25	—	62.20	—	3.88	2.98
PureACL[64]	—	—	58.75	66.34	70.60	—	0.23	45.65	68.23	74.87	—	0.37	37.99	52.57	—	64.13	—	8.13	2.80
CCVPE[67]	—	—	11.05	22.05	42.09	—	1.22	7.45	13.36	24.97	—	3.35	34.66	60.72	—	84.03	—	3.43	1.53
T2GA[68]	—	—	**76.20**	**97.73**	**99.97**	—	**0.14**	**98.18**	**99.89**	**99.97**	—	**0.07**	**48.97**	**66.29**	—	**90.54**	—	**2.19**	**1.07**
方向先验噪声为 ±45°, 横/纵位置先验噪声为 ±20 m (跨区域测试)																			
SIBCL[52]	—	—	16.90	32.62	58.24	—	0.71	19.58	30.74	49.25	—	1.05	25.02	50.18	—	60.16	—	9.67	1.96
BoostAcc[63]	—	—	14.88	29.20	52.64	—	0.93	3.38	5.56	11.73	—	8.92	12.46	24.87	—	47.26	—	5.61	4.26
PureACL[64]	—	—	50.25	62.63	64.56	—	0.25	7.41	62.04	64.23	—	0.44	19.95	36.25	—	56.49	—	10.96	3.16
CCVPE[67]	—	—	4.96	10.05	38.12	—	2.79	3.13	6.14	12.45	—	8.16	3.38	10.24	—	17.41	—	19.46	15.39
T2GA[68]	—	—	**75.96**	**97.47**	**99.97**	—	**0.15**	**97.49**	**99.87**	**99.97**	—	**0.07**	**36.10**	**63.08**	—	**86.18**	—	**2.88**	**1.44**

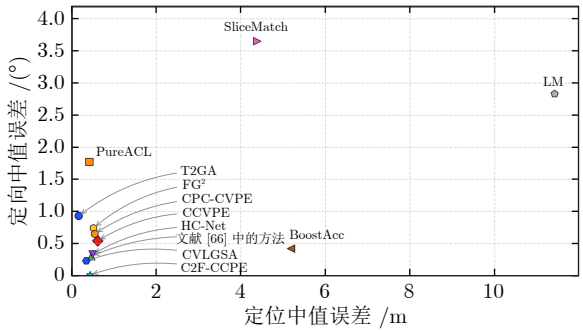


图 4　基于 KITTI-CVL 数据集的同区域测试

Fig. 4　Same-region testing based on the KITTI-CVL dataset

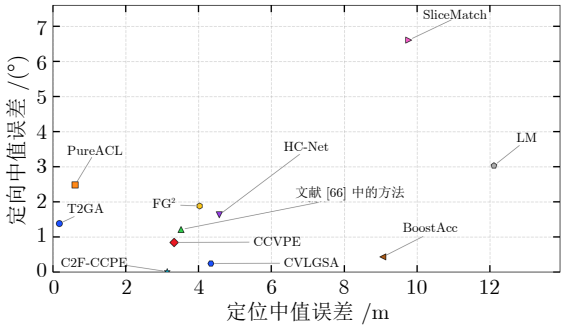


图 5　基于 KITTI-CVL 数据集的跨区域测试

Fig. 5　Cross-region testing based on the KITTI-CVL dataset

定位与方向上均展现出最顶尖且稳定的性能 (定位中值误差小于 0.5 m, 方向误差小于 1°), 其卓越表现可归因于协同优化的架构设计. 另外, 各方法的性能在跨区域测试或不同先验噪声下出现显著分化, 这揭示了其不同的鲁棒性来源. 比如 CCVPE[67] 在小先验误差、同区域测试中定位精度极高 (中值误差为 0.62 m), 但其性能在跨区域或大噪声下剧烈下滑. 这表明其高度依赖数据驱动的密集特征匹配, 当测试场景分布偏离训练数据或存在较大几何不确定性时, 匹配质量迅速恶化, 泛化能力是其弱点所在. 而 LM[60] 作为经典方法, 其绝对精度虽非最高, 但在所有测试条件下 (包括 CSC-Shanmao 的越野子集) 都表现出最佳的均衡性与稳定性. 这得益于其显式的几何投影与 Levenberg-Marquardt 优化框架, 这种基于模型的优化对局部特征噪声相对不敏感, 提供了更强的跨场景可靠性. 相比之下, 绝大多数纯视觉方法在非结构化场景下性能普遍衰减. 这根本上是由于纹理缺失、特征重复、强光照干扰破坏了跨视图匹配所依赖的视觉外观一致性, 凸显了纯视觉模态的固有瓶颈.

根据上面的模型评价指标分类可以看到, 使用传统跨视图定位数据集进行测试的模型基本上都是使用 Top-K 召回率作为优化指标, 这是由于 CVUSA 和 CVACTval 数据集本身的限制导致的. 这类数据集仅包含地面视角图像和卫星图像, 不包含车辆/相机的 IMU 数据, 因此定位精度受卫星图分辨率影响较大, 且仅能评估 2-DoF 的估计精度, 缺少对方向/姿态指标的评估. 相比之下, 使用 KITTI-CVL 数据集进行评估的模型大多针对自动驾驶领域中的跨视图定位任务, 这类模型对定位精度要求更高, 并且能够实现 3-DoF 的评估, 包含了对方向误差的评估. 这样的模型评价指标更有利于后续方法明确优化方向, 并且能够规范未来研究者的评估基准. 表 4 中简要列出了面向自动驾驶的跨视图定位方法, 以模型的评价指标作为分类标准, 包括方法提出的时间、发表期刊、输入以及创新点.

从表 4 中可以看出, 依赖路网/矢量地图先验的方法 (如部分输入为“地面+卫星”的方法通过隐式学习或专门设计模块来利用路网) 通常在城市结构化道路中表现稳健, 因为先验信息提供了强大的空间约束, 能有效消除视觉歧义. 然而, 这类方法在缺乏清晰道路结构的野外、非结构化环境中可能失效. 另外, 纯视觉方法 (地面+卫星) 成本低、适用广, 但在极端天气或纹理缺失场景下面临挑战; 融合 LiDAR 点云的方法几何精度高、对光照鲁棒, 尤其适合高精度定位、军事无人车及恶劣环境, 但系统复杂、成本高昂.

因此, 方法的选择并非孤立地追求最高精度, 而应首先对齐任务场景的核心约束: 在城市 L4 级自动驾驶中, 可优先考虑融合轻量化先验的 3-DoF 纯视觉方法以平衡成本与性能; 在野外勘探或军事领域, 则倾向于采用多模态融合方案以保证在 GNSS 拒止环境下的鲁棒性与精度.

评价指标直接定义了研究的优化目标与应用导向. 根据前面的分析可以发现, 使用 Top-K 召回率的方法源于通用图像检索任务, 其目标是“找到大致区域”, 适用于对自动驾驶车辆进行粗略初始化; 而采用定位与方向估计误差的方法, 则直面自动驾驶对厘米级位置与度数级朝向的精确控制需求, 标志着任务范式从“区域检索”到“姿态估计”的根本转变. 随着技术走向实际部署, 未来研究中的效率指标 (如帧率、功耗) 以及在挑战性子集 (如跨季节、跨天气) 上的泛化误差必须被纳入核心评价体系.

2.4　模型损失函数的设计

作为模型训练的核心部分, 损失函数对于定位精度、特征学习以及模型的泛化能力都有着至关重要的作用, 它不仅能为模型优化提供明确的目标和方向, 还在很大程度上影响着模型对复杂场景的适

表 4 面向自动驾驶的跨视图定位方法分类 (依据评价指标)

Table 4 Taxonomy of cross-view localization methods for autonomous driving based on evaluation metrics

模型评价指标	方法	时间及论文来源	输入	创新点
Top-K 召回率	DSM[55]	2020 CVPR	地面+鸟瞰/卫星	极坐标变换对齐航拍图像与地面图像, 两流卷积网络学习深度特征
	CVLNet[61]	2022 ACCV	地面+鸟瞰/卫星	通过几何驱动模块和场景先验匹配机制确定车载相机相对于卫星图中心的位置
	Accurate 3-DoF[62]	2022 TPAMI	地面+鸟瞰/卫星	设计双孪生网络结构, 提出两阶段定位机制, 使网络聚焦特定区域
	HADGEO[78]	2024 ICASSP	地面+鸟瞰/卫星	设计新的损失函数, 结合全可学习卷积网络 (ALFCN) 提取特征
	GeoSSK[87]	2025 MMM	地面+鸟瞰/卫星	设计基于交叉注意力的跨视图融合交互模块 (CIM) 和语义相似性知识蒸馏 (SSKD)
	DSTG[88]	2025 TGRS	地面+鸟瞰/卫星	提出一种在不同分辨率级别之间进行教师模型与学生模型特征对齐的方法
定位与方向估计的误差及召回率	SIBCL[52]	2023 ICRA	地面+点云+卫星	用地面图像和点云估计车辆的 3-DoF 姿态
	MapLocNet[53]	2024 IROS	地面+鸟瞰/卫星	使用从粗到精的特征注册策略进行视觉重定位
	AutoVision[54]	2019 ICRA	地面+鸟瞰/卫星	用粒子滤波方法结合 CVM-Net 进行跨视图匹配
	SLAM-G2S-Fusion[57]	2024 ICRA	地面+鸟瞰/卫星	将 vSLAM 与 G2S 方法结合形成一种从粗到精的选择方法
	CVML[58]	2022 ECCV	地面+鸟瞰/卫星	利用密集的卫星描述符和相似性匹配来生成密集概率分布
	HC-Net[59]	2023 NeurIPS	地面+鸟瞰/卫星	提出单应性估计模块消除重复采样并忽略不可观察内容
	LM[60]	2022 CVPR	地面+鸟瞰/卫星	通过几何投影模块, 将卫星图像的深度特征投影到地面视角下, 再利用 LM 模块细化相机的姿态估计
	BoostAcc[63]	2023 ICCV	地面+鸟瞰/卫星	提出几何引导交叉视图变换器, 解决了对象模糊和相机倾斜的问题
	PureACL[64]	2023 ICCV	地面+鸟瞰/卫星	检测一致性关键点以及深度特征, 减少纯视觉匹配的模糊性
	SliceMatch[65]	2023 CVPR	地面+鸟瞰/卫星	将水平视场分割成多个切片, 学习与定位和方向估计相关的特征
	文献 [66] 中的方法	2023 NeurIPS	地面+鸟瞰/卫星	通过密集像素级流场来估计车载相机的 3-DoF 姿态
	CCVPE[67]	2023 TU Delft	地面+鸟瞰/卫星	构建方向感知的图像描述符, 实现位置和方向的联合估计
	T2GA[68]	2022 CVPR	地面+鸟瞰/卫星	T2GA 模块整合离地特征, CycDA 损失确保特征一致性, 等距投影损失 (ERP) 平衡关键点影响
	BEVRender[70]	2024 IROS	地面+鸟瞰/卫星	生成局部鸟瞰图图像并将其与航拍视图对齐
	Geo-tracking[71]	2022 IROS	地面+点云+卫星	将车载摄像头和激光雷达数据与地理配准的正射影像对齐
	文献 [33] 中的方法	2024 JSTARS	点云+卫星	引入道路结构作为不同模态之间的桥梁
	AGL-Net[74]	2024 IROS	点云+卫星	利用尺度分类器和特征差值方法处理尺度差异
	C2F-CCPE[77]	2024 ICME	地面+鸟瞰/卫星	通过多尺度特征融合模块 (LOFFM) 同时进行方向和位置的精确预测
	FG2[86]	2025 CVPR	地面+鸟瞰/卫星	利用几何感知注意力与高度选择机制实现细粒度特征匹配
	CVLGSA[89]	2025 IROS	地面+鸟瞰/卫星	通过视角驱动的注意力融合模块和投影稳定补丁引导优化器
	CPC-CVPE[91]	2025 Scientific Data	地面+点云+卫星	通过多传感器融合实现精确的 GNSS 独立定位
其他	OrienterNet[75]	2023 CVPR	地面+卫星	用卷积神经网络提取图像语义特征并将其提升为鸟瞰图表示, 同时将开放街图编码为神经地图

应性和鲁棒性. 因此, 选择和设计合适的损失函数能够引导模型学习到更具有判别力的特征表示, 从而提高跨视图特征之间的匹配准确性, 在多变的自动驾驶环境中实现更稳定可靠的定位. 下文将深入探讨面向自动驾驶的跨视图定位方法中各式各样的损失函数的特点及多样化应用.

2.4.1 L1 损失

L1 损失也被称为平均绝对值误差 (MAE), 是指模型预测值 $f(x)$ 与真实值 y 之间绝对差值的平均值, 主要用于回归问题和一些简单的模型, 定义为:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f(x_i) - y_i| \tag{1}$$

式中, $f(x_i)$ 和 y_i 分别表示第 i 个样本 x_i 的预测值及相应的真实值; n 为样本的个数. 由于 L1 损失是对误差取绝对值, 因此异常值对于整体损失的影响相对较小, 不会过分影响模型的训练方向, 这能让 L1 损失对异常值保持鲁棒性, 适用于对异常值敏感的场景. 另外, L1 损失的稀疏性会使一些权重系数变为 0, 从而实现特征选择, 适用于需要对特征进行筛选的场景; 但是 L1 损失梯度固定的特点也会导

致误差较小时收敛速度过慢的问题.

2.4.2 L2 损失

L2 损失也被称为欧氏距离, 通常用于度量数据点之间的相似度, 又被称为均方误差 (MSE), 定义为:

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)^2 \tag{2}$$

LM[60] 损失函数通过最小化当前姿态 ξ 和真实姿态 ξ^* 之间的 L2 范数来进行优化, 定义为:

$$L=\|\xi-\xi^*\|_2^2 \tag{3}$$

在训练过程中, 当通过 LM 优化算法得到的相机姿态偏离真实值时, 误差会被反向传播到特征提取网络以更新其参数. L2 损失由于平方操作会放大异常值的影响, 可能会导致模型过于关注这些异常点, 影响对正常数据的拟合效果. 此外, L2 损失不具有稀疏性, 会保留所有特征的信息, 因此适用于对预测精度要求较高并且数据质量较好的场景.

2.4.3 对比损失

对比损失最早由 Hadsell 等[93] 提出, 主要用于度量一对样本之间的相似性或差异性, 表示为:

$$\begin{aligned}\mathrm{Contrast}(\boldsymbol{x},\ \boldsymbol{y},\ l)=(1-l)\,\frac{1}{2}lD+\\ \frac{1}{2}l\left(\max(0,\ m-D)\right)^2\end{aligned} \tag{4}$$

式中, $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 表示输入与输出图像; l 是标签, $l\in\{0,\ 1\}$ 表示输入与输出是否匹配, 如果匹配, 则 $l=1$, 否则 $l=0$; $D=\|f(\boldsymbol{x})-f(\boldsymbol{y})\|_2$ 表示输入与输出在特征空间中的欧氏距离; $m>0$ 是不匹配图像之间的函数间隔值. 对比损失多用于相似性度量学习任务, 能有效拉近相似样本的距离并分散不同的类别, 但是它需要正负样本对并且难以直接用于多分类任务.

2.4.4 Smooth L1 损失

Smooth L1 损失由 Girshick[94] 提出, 该损失结合了 L1 与 L2 损失的优点, 通过引入一个平滑参数, 使得其在误差较小时表现为 L2 损失, 在误差较大时表现为 L1 损失, 公式如下:

$$\mathrm{smooth}_{L_1}(x,\ y)=\begin{cases}0.5(x-y)^2, & |x-y|<1\\ |x-y|-0.5, & \text{其他}\end{cases} \tag{5}$$

式中, x 是预测值; y 是目标值. 当 $|x-y|<1$ 时, 它的倒数部分会变得平滑, 反之当 $|x-y|\geq 1$ 时, 其线性部分更加重要. 这样的特性使得 Smooth L1 损失对异常值较为鲁棒, 但是也导致它相对于 L1 损失以及 L2 损失的实现更为复杂.

2.4.5 BCE 损失

BCE (binary cross entropy) 损失主要适用于二分类任务, 由于多标签分类任务可以简化为多个二分类任务叠加, 所以 BCE 经过修改也可以用于多标签分类任务, 公式如下:

$$\begin{aligned}Loss=-\frac{1}{N}\sum_{i=1}^{N}y_i\log(p(y_i))+\\ (1-y_i)\log(1-p(y_i))\end{aligned} \tag{6}$$

式中, y 是二元标签 0 或 1; $p(y)$ 是输出属于标签的概率; N 表示模型预测对象的组数. 对于标签 y 为 1 的情况, 如果预测值 $p(y)$ 趋近于 1, 那么 $Loss$ 的值应趋近于 0; 反之, 如果此时预测值 $p(y)$ 趋近于 0, 则 $Loss$ 的值应该很大. BCE 损失具有可解释性强的特点, 其损失值可以解释为模型预测错误的程度. 另外, BCE 损失函数形式简单, 易于实现和理解, 适合在实际应用中快速开发. 但是当正负样本数量差异较大时, BCE 损失可能会使模型过于关注多数类, 而忽略少数类, 导致模型在少数类上的预测性能较差.

2.4.6 InfoNCE 损失

InfoNCE 损失是一种用于自监督的损失函数, 它基于信息论的思想, 通过对比正样本和负样本的相似性来学习模型参数, 表示为:

$$\mathrm{InfoNCE}=-\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(\frac{q_i k_{i+}}{\tau})}{\sum\limits_{j=1}^{N}\exp(\frac{q_i k_{j-}}{\tau})}\right) \tag{7}$$

式中, N 是样本的数量; q_i 是查询样本 i 的编码向量; k_{i+} 是与查询样本 i 相对应的正样本的编码向量; k_{j-} 是与查询样本 i 不对应的正样本的编码向量; τ 是温度系数, 用于调节相似度得分的分布. InfoNCE 损失具有强大的特征表达能力、良好的对比学习效果以及较高的灵活性, 能够有效地学习到数据的深层特征表示, 但是它的计算复杂度高、对温度超参数敏感并且可能面临信息丢失的问题, 这使得 InfoNCE 在实际应用中需要进行仔细调整和优化.

SliceMatch[65] 的损失函数是对原始的 InfoNCE 损失进行修改后设计得到的, 该损失函数使用 K 个训练姿态, 定义为:

$$\mathrm{SMLoss}=-\log\left(\frac{\exp(\frac{c^{GT}}{\tau})}{\frac{\alpha}{K}\sum\limits_{k=1}^{K}\exp(\frac{c^{k}}{\tau})+\exp(\frac{c^{GT}}{\tau})}\right) \tag{8}$$

式中, α 是超参数, 用于衡量 K 个姿态对于学习的

贡献; 变量 c^{GT} 是地面全局描述符与空中全局描述符在候选真值位姿处的余弦相似度; c^k 是地面全局描述符与空中全局描述符在预测姿态处的余弦相似度. 原始的 InfoNCE 损失可以通过 $\alpha = K$ 获得, 通过这一损失函数, SliceMatch 对比了真实姿态与其他 K 个不同位置和方向的姿态, 以此来训练模型提取用于定位和方向估计的特征. 此外, FG^2[86] 也是一种基于 InfoNCE 损失来驱动的端到端架构, 这种架构在处理具有显著视角差异的场景时优势较为明显.

2.4.7 三元组损失

三元组损失 (triplet loss) 的优势在于细节区分, 即当两个输入相似时, 三元组损失能够更好地对细节进行建模, 学习到输入的更好表示. 在跨视图定位任务中三元组损失是被使用最多的损失函数, 它随机选择一张图像作为锚点, 把相同标签的图像拉近, 把不同标签的图像推远. 这种损失函数能让相同标签但不同视角的特征向量距离变小, 恰好符合跨视图定位任务中的数据集组成结构, 公式如下:

$$\text{TripletLoss} = \max(d(a,\ n) - d(a,\ p) + m,\ 0) \quad (9)$$

式中, a 表示锚点样本; p 表示与 a 标签相同的样本, 即正样本; n 表示与 a 标签不同的样本, 即负样本; d 表示锚点样本 a 和正样本 p 之间的距离度量; m 是一个大于 0 的超参数, 较大的 m 能够增强模型对不同类样本的区分度, 但是过大会增加模型训练的难度, 出现网络不收敛的情况, 因此在模型训练初期可以使用一个较小的 m 进行初始化训练, 之后再根据测试的结果对 m 进行适当的增大或缩小, 这样就可以在保证网络收敛的同时让模型也能有较好的性能. 三元组损失的优点在于其能有效学习样本之间的相对距离, 从而提升模型的度量学习能力, 它适用于多种相似性度量任务且无需类别标签, 但是它的训练复杂度相对于其他损失函数而言更高, 并且对超参数敏感, 收敛速度较慢, 这些因素也都可能会影响模型的训练效率和最终的性能. 尽管如此, 基于三元组损失训练的模型仍然是当前数量最多的.

T2GA[68] 在三元组损失的基础上增加了自定义的循环领域适应损失和等距重投影损失, 构造了一个新的混合损失函数, 其中增加的循环领域适应损失明确强制同一物体的视图不变表示, 增强了跨视图定位方法的鲁棒性, 等距重投影损失允许模型利用距离车辆更远的关键点, 与此同时也更注重修正车辆姿态估计, 三种损失共同作用, 有效提高了模型在不同视角下的特征提取能力和定位精度.

其他也使用了三元损失作为模型损失函数的有文献 [52, 54–55, 61–64, 71, 78], 这里不作详细介绍, 具体可参阅原文献.

2.4.8 自适应损失函数

正如上文所提到的, 不同的损失函数各有优缺点, 并且也都有各自适用的场景, 因此不同的模型在使用损失函数的过程中, 往往不是单一地调用或者仅使用某一种损失函数, 研究人员会根据任务的特性对损失函数进行改进和组合, 以适应自己所面向的任务场景. 下面将介绍不同模型面对自身任务时构建的自适应损失函数.

MapLocNet[53] 的损失函数设计包含 Smooth L1 损失与 BCE 损失, 总体的损失函数由二者用权重因子加权结合而成, 如下式:

$$\begin{aligned}\text{MaplocLoss} = \lambda_c L_c + \lambda_f L_f + \\ \lambda_{bev} L_{bev} + \lambda_{map} L_{map}\end{aligned} \quad (10)$$

式中, λ_c、λ_f、λ_{bev}、λ_{map} 为权重因子, 在训练过程中均设置为 1; L_c 为粗定位位姿损失, 用于监督第一阶段 (粗特征配准) 预测的位姿偏移量; L_f 为精定位位姿损失, 用于监督第二阶段 (精特征配准) 预测的位姿偏移量; L_{bev} 为 BEV 语义分割辅助损失, 用于对 BEV 模块 (图像→BEV 特征) 进行语义分割监督; L_{map} 为地图语义分割辅助损失, 用于对 U-Net 进行语义分割监督. 通过这些损失函数的组合, MapLocNet 能够进行有效的姿态估计和特征对齐, 从而加快模型的收敛速度并提高模型的定位性能.

HC-Net[59] 提出一种新的混合损失函数, 该函数定义为:

$$\text{HCLoss} = \alpha_1 L_{dis} + \alpha_2 L_{ori} + \alpha_3 L_{info} \quad (11)$$

式中, α_1、α_2、α_3 为不同的权重; $L_{dis} = (u_s - u_s^*)^2 + (v_s - v_s^*)^2$ 为位置损失, 通过计算预测结果与真实值之间的 L2 损失得出, 用于反映预测中心点与地面真实位置之间的差距; $(u_s^*,\ v_s^*)$ 是对应的真值 GPS 在卫星图像上的像素坐标; $(u_s,\ v_s)$ 是预测得到的像素坐标; $L_{ori} = |\theta - \theta^*|$ 为方向损失, 用 L1 损失衡量预测的相机朝向 θ 与真实朝向 θ^* 之间的差异; L_{info} 是信息损失, 使用 InfoNCE 损失强化 BEV 点与卫星图像之间的相关性.

Song 等[66] 提出的方法中的总损失函数表示为:

$$\text{SLoss} = \beta L_p + \sum_{i=1}^{l} (L_m^l + \alpha L_c^l) \quad (12)$$

该损失也由 3 部分组成, 分别是匹配损失 L_m^l、置信度损失 L_c^l 和位置损失 L_p, 三部分损失共同作

用, 通过最小化这些损失来优化相机的位置参数, 从而提升 BEV 图像与卫星图像之间的对齐精度; α、β 为权重系数.

AGL-Net[74] 整合了一个新颖的损失函数 AGL-Loss, 包括姿态损失、尺度损失和骨架损失三部分, 定义为:

$$\text{AGLLoss} = L_{uv\theta} + L_S + L_{skeleton} \tag{13}$$

式中, $L_{uv\theta} = -\log P(\eta^*)$ 为姿态损失, 它衡量了估计的相机姿态与真实姿态之间的差异, 这里用 L2 范数度量位置误差, L1 范数度量角度误差, $P(\eta^*)$ 表示网络预测的位姿 η^* 在离散概率分布中的概率值; $L_S = \sum_{i=1}^{B}(S_i - S_{gt,\,i})/B$ 为尺度损失 (B 为批量大小), 该损失专注于约束从 LiDAR 扫描中提取的特征的尺度, 它通过均方误差来计算预测的尺度因子 S 与真实地面尺度 S_{gt} 之间的差异; $L_{skeleton}$ 为骨架损失, 该损失函数旨在捕捉特征的骨架信息, 确保提取的特征在空间结构上的准确性, 作者在构造 $L_{skeleton}$ 时参考了 BCE 损失, 定义为:

$$\begin{aligned} L_{skeleton} = & -\frac{1}{X \times Y} \sum_{p \in X \times Y} (y_{gt}(p)\log(f^s_{map}) + \\ & (1 - y_{gt}(p))\log(1 - f^s_{map})) \end{aligned} \tag{14}$$

式中, $X \times Y$ 表示特征图的空间尺寸; f^s_{map} 表示网络预测的地图骨架掩码; p 表示 f^s_{map} 中的单个 2D 点; $y_{gt}(p)$ 是地面地图的真实骨架掩膜. 通过整合上面三个部分的损失函数, AGL-Net[74] 的模型性能得到显著提升.

CCVPE[67] 的损失函数由对比损失 L_M、定位损失 L_D 以及方向损失 L_Y 三部分组成, 表示为:

$$\text{CCVPELoss} = L_D + \alpha' L_Y + \beta' L_M \tag{15}$$

式中, α'、β' 为权重系数; 对比损失 L_M 是 InfoNCE 损失的加权求和; 定位损失 L_D 是交叉熵损失. 由于该方法将方向预测视为回归问题, 并且使用了平滑的真实地面 D_{gt}, 因此其从平滑的真实位置贡献中求和, 并定义方向损失为:

$$\begin{aligned} L_Y = & \sum_{i=1}^{W}\sum_{j=1}^{W} D_{gt}^{i,\,j}((\cos(O_{gt}) - Y_1^{i,\,j})^2 + \\ & (\sin(O_{gt}) - Y_2^{i,\,j})^2) \end{aligned} \tag{16}$$

其中, W 表示航拍图像的边长; O_{gt} 表示真实姿态; $D_{gt}^{i,\,j}$ 是在真实位置 $(i_{gt},\, j_{gt})$ 处放置的 2D 高斯平滑分布; Y_1 和 Y_2 分别表示预测姿态向量场 Y 的两个通道.

与 CCVPE 类似, C2F-CCPE[77] 的损失函数也是由定位交叉熵损失、方向损失和对比损失整合而成, 表示为:

$$\text{C2FLoss} = \alpha_1' L_{dis} + \beta_1' L_{ori} + \gamma L_{const} \tag{17}$$

式中, L_{dis} 为定位交叉熵损失; L_{ori} 为方向损失; L_{const} 为对比损失; α_1'、β_1'、γ 为权重系数.

另外, 还有一些方法也自定义了专有化的自适应损失函数, 比如 Hu 等[33] 的损失函数是一个结合了数据保真度与正则化的复合函数, CVML[58] 的总损失是输出损失与瓶颈损失的结合.

损失函数是连接任务目标与模型学习的桥梁. 从上述介绍与分析可以看出, 其设计已从推动特征空间对齐的单一度量学习 (如三元组损失), 发展为融合了几何一致性约束与多任务 (位置、朝向) 协同优化的复合范式. 这种演进表明, 领域共识已从“学习看起来像”转变为“学习在几何上对齐”. 在未来的研究中, 不确定性感知的损失权重调节以及针对极端困难样本的自适应加权机制, 将是进一步提升模型鲁棒性与校准能力的重要方向.

2.4.9 面向跨视图定位的损失函数设计范式

前述多样化的损失函数及其组合, 反映了面向自动驾驶的跨视图定位任务对损失函数设计的多层次、结构化需求. 一个有效的损失函数不仅需要驱动模型收敛, 更需要显式地编码任务先验, 引导模型学习跨越视角鸿沟的本质特征. 通过梳理, 可以总结出以下关键设计范式与原则.

一是度量学习范式: 构建视角不变的嵌入空间. 这是解决跨视图外观差异的核心. 以三元组损失、对比损失为代表, 其设计哲学是拉近正样本对 (匹配的跨视图图像)、推离负样本对, 从而学习一个视角不变的公共特征子空间. 其有效性高度依赖于难样本挖掘策略, 以应对跨视图中大量“简单负样本”和“困难正样本”的挑战.

二是几何一致性范式: 引入跨视图几何约束. 这是应对几何歧义的关键. 单纯的度量学习可能只保证语义相似, 而忽略精确的空间对齐. 因此, 许多方法 (如 Song 等[66] 提出的密集光流损失) 在损失函数中直接融入几何约束, 例如计算预测的像素级对应关系与真实位姿变换的一致性. 这种范式将几何推理从后处理阶段前置到训练目标中, 使特征学习与几何对齐协同优化.

三是任务驱动多目标权衡范式. 自动驾驶定位任务要求同时估计位置与朝向 (3-DoF), 且对不同误差的容忍度不同. 因此, 当前主流方法 (如 CCVPE[67]、C2F-CCPE[77]) 普遍采用复合损失函数, 通过可学习的权重将定位损失与方向损失相结合. 更先进的设计还会引入不确定性估计, 让网络自适应地调整

各任务损失的权重.

综上所述, 损失函数的设计已从早期简单的度量学习, 演进为融合几何约束与多任务权衡的复杂范式. 这一演进路径清晰地反映了领域认知的深化: 研究者逐渐意识到, 解决跨视图定位问题不能仅靠学习外观相似性, 而必须将物理世界的几何一致性作为核心约束嵌入优化目标. 这种从“数据驱动相似”到“物理规律驱动对齐”的思维转变, 是推动定位精度突破的关键所在.

2.5 技术贡献的归因与可解释性分析

上文依据输入数据、网络架构、评价指标与损失函数对现有方法进行了系统性分类. 在此基础上, 一个随之而来的关键问题是: 如何准确评估与归因各类技术组件对最终性能提升的具体贡献. 这将直接关系到跨视图定位技术发展的清晰度与可积累性.

尽管当前大多数研究都通过消融实验来验证其核心模块的有效性, 但这些实验的设计深度与归因粒度仍存在提升空间. 许多工作的消融实验侧重于验证“有无”某个宏观模块 (如整个特征提取分支、匹配网络或特定的损失函数) 对整体精度的影响. 然而, 对于日益复杂的模型架构 (如包含多个注意力头的 Transformer、多路径融合网络、由多项加权损失组成的复合目标函数), 这种粗粒度的验证方式难以精确量化每个子模块 (如特定的注意力机制、损失函数中的某一项约束、特征投影中的归一化层) 的独立贡献, 也无法充分揭示不同模块之间的协同或拮抗作用.

例如, 一个采用了 Transformer 与 CNN 混合架构的方法, 其性能提升可能同时源于自注意力机制的全局建模能力和 CNN 的局部特征提取能力, 但二者各自的贡献比例以及是否存在“$1+1>2$”的协同效应, 往往未被细致量化. 同样, 一个融合了多种约束的自适应损失函数, 其中每一项对于解决视角差异、光照变化或几何失配等具体问题的效用, 也值得更精细的剖析.

技术贡献归因的模糊性可能带来两个潜在问题: 一是造成创新点评估失真, 性能的提升可能被笼统地归功于一个宏观的“新架构”或“新损失”, 而其中真正起决定性作用的细微设计容易被忽视, 不利于后续研究抓住关键进行突破; 二是导致技术复现与迁移困难, 当其他研究者试图复用或改进某个方法时, 由于缺乏对组件敏感性的深入理解, 在调整模型或适应新场景时可能面临“黑箱”调试的困境.

因此, 推动更细致、更可解释的消融实验与分析, 是未来研究值得强化的方向. 这包括但不限于: 设计控制变量更严格的消融实验以分离子模块效应; 利用可视化工具 (如特征图、注意力热图) 直观展示关键模块的工作机制; 在不同挑战性场景子集上测试各技术组件的鲁棒性. 通过增强技术创新的透明度和可归因性, 将有助于本领域研究建立起更坚实、更清晰的知识积累路径, 从“经验驱动”的设计范式逐步转向“原理驱动”的深度创新.

通过四个维度的交叉分析, 本文揭示了不同技术路径的内在权衡. 但是这些设计差异在实际性能上如何体现? 其泛化能力究竟如何? 这需要基于统一的数据基准进行实证检验. 因此, 第 3 节将首先系统介绍领域内的关键数据集与评估体系, 进而呈现综合性的实验结果.

3 数据集与评价指标

传统的跨视图定位数据集无法准确估计车辆的 3-DoF 姿态, 对于自动驾驶导航系统而言, 准确的 3-DoF 姿态信息具有重要意义: 通过结合 3-DoF 姿态信息和高精度地图, 车辆可以更准确地识别当前位置和行驶方向, 并且可以帮助车辆在行驶过程中保持平衡, 避免侧翻和失控. 可以说, 传统的跨视图定位数据集仅提供位置信息, 而没有车辆本身的位姿信息, 并且定位精度受限于卫星图像的划分精度, 不能满足自动驾驶场景下的定位需求. 在早期的跨视图定位技术研究中, 主要针对地面和卫星视角展开, 并且多用于城市环境的跨视图定位, 因此研究者们构造出 CVUSA[2]、San Francisco[2]、CVA-CT[51] 和 VIGOR[14] 等通用标准数据集. 空中−卫星跨视图定位任务随着无人机领域的迅速发展开始被重视, University-1652 数据集在 2020 年被 Zheng 等[10] 提出, SUES-200 数据集在 2023 年被 Zhu 等[81] 提出. 本文主要介绍适用自动驾驶场景的跨视图定位数据集, 因此对传统方法数据集不做详细阐述, 仅列出作为参考, 见表 5.

近几年提出的针对自动驾驶车辆的跨视图定位的专用数据集, 主要基于自动驾驶经典数据集 (如 KITTI[82]、Ford Multi-AV Seasonal[83]、Oxford RobotCar[84−85] 等) 构建得到, 研究者们根据数据集中的 GNSS 信息, 通过使用谷歌地图收集卫星图像, 构建了如 KITTI-CVL[52]、FordAV-CVL[52]、Oxford RobotCar(+)[58] 等这样的更适合自动驾驶场景下跨视图定位方法评估的数据集.

然而, 当前的主流数据集存在固有的时空与分布偏差, 且评估范式严重忽视了模型的跨域泛化能力, 这为自动驾驶系统在真实复杂环境中的可靠部

表 5 传统跨视图定位数据集

Table 5 Traditional cross-view localization datasets

数据集	发布时间	图像分辨率 (像素)	图像类型	数据量	地点
CVUSA[2]	2015	1232 × 224 (地面), 750 × 750 (卫星)	地面图像和卫星图像	44416 对地面和卫星图像	美国
San Francisco[2]	2015	256 × 256	地面图像和卫星图像	278561 幅地面图像和 174217 幅卫星图像	旧金山弗朗西斯科
CVACT[51]	2019	多尺寸	地面图像和卫星图像	44416 对地面和卫星图像以及 92802 幅额外测试图像	澳大利亚堪培拉
University-1652[10]	2020	512 × 512	地面图像、无人机图像和卫星图像	1652 组地面、无人机和卫星图像	世界各地 72 所大学
VIGOR[14]	2021	2048 × 1024 (地面), 640 × 640 (卫星)	地面图像和卫星图像	238696 幅地面图像和 90618 幅卫星图像	纽约市、旧金山弗朗西斯科、芝加哥和西雅图
SUES-200[81]	2023	1080 × 1080 (无人机), 512 × 512 (卫星)	无人机图像和卫星图像	24120 对无人机和卫星图像	上海工程科技大学周围

署带来了潜在风险. 一方面, 数据集在地理覆盖、季节变化、天气条件与道路类型上高度同质化; 另一方面, 绝大多数研究仅进行“数据集内”评估, 极少考察模型在分布外场景下的性能. 这种偏差导致模型的“高精度”可能仅是对特定数据分布的过拟合, 其真实鲁棒性无法被有效度量.

为系统揭示这一问题并提供一个面向泛化能力评估的新基准, 本节首先介绍现有数据集并分析其局限性, 随后介绍本文构建的 CSC-Shanmao 数据集及其为解决上述问题所做的针对性设计. 面向自动驾驶的数据集信息如表 6 所示, 下面将详细介绍这些数据集.

3.1 由 KITTI 衍生的数据集

Shi 等[60] 在 2022 年设计了一个双分支 CNN 模型 LM, 该方法分别从地面和卫星图像中提取特征. 他们在两个标准的自动驾驶数据集中评估所提出的研究方法, 分别是 KITTI 和 Ford Multi-AV 数据集. 基于这两个数据集中提供的 GNSS 信息, 他们从谷歌地图中收集卫星图像, 并以 18 倍变焦下载. 其中 KITTI 数据集对应的卫星图栅格分辨率为 0.2 m 每像素, 由于其包含车辆在不同时间的不同轨迹, 因此他们将整个数据集划分为三个子集, 即一个训练集和两个测试集, 分别用 training、Test1 和 Test2 表示, 其中 training 和 Test1 来自同一区域, Test2 位于不同区域, 这样的划分可以让 Test2 作为评价算法泛化能力的指标. Wang 等[52] 在 2023 年整理并规范了 Shi 等[60] 的 KITTI 增强数据集, 进而正式提出 KITTI-CVL 数据集. 该数据集结合标准的自动驾驶数据集 KITTI 以及从谷歌地图中获取的空间一致的卫星图像, 并根据提供的 GNSS 标签进行卫星图像的选择和校正. 数据集中的图像类型包括地面视角图像和卫星图像, 其中地面视角图像来自 KITTI 数据集中的前左相机视角, 卫星图像来自谷歌地图 API, 分辨率为 1280 × 1280 像素, 区域大小为 256 m × 256 m, 卫星图栅格分辨率约为 0.2 m 每像素. 这两个基于 KITTI 的数据集来源相同, 不同之处仅在于其数据集的组成格式不太一样, 后续的研究基本上沿用了基于 KITTI 构建的数据集 KITTI-CVL. 为便于分析对比, 本文将在这两个数据集上进行性能测试的模型视为在同一基准上的测试.

该数据集已成为领域内最常用的基准之一. 然而 KITTI-CVL 存在明显的时空局限性: 其地面数据源于单次、短时间跨度的采集, 本质上缺乏对季节更替、天气变化及长期场景演变的覆盖; 同时, 其场景完全集中于德国卡尔斯鲁厄市的城市结构化道路, 地理与道路类型多样性严重不足. 因此, 在该数据集上表现优异的模型, 其对动态环境和未知区域的泛化能力仍然是一个未被充分验证的问题.

3.2 由 Ford Multi-AV Seasonal 衍生的数据集

Ford Multi-AV 跨视图定位数据集由 Shi 等[60] 在 2022 年正式提出, 该数据集基于 Ford Multi-AV Seasonal[83] 构建, 同样包含了与从谷歌地图获取的卫星图像相对应的地面视图, 他们在 6 条 Ford Multi-AV Seasonal 轨迹中选择了“log1”轨迹进行方法评估, 数据集中的地面视角图像来自 Ford Multi-AV Seasonal 中 V2 车辆的前左视角, 卫星图像来自谷歌地图 API, 分辨率为 1280 × 1280 像素, 区域大小为 281.6 m × 281.6 m, 卫星图栅格分辨率约为 0.22 m 每像素. Wang 等[52] 在 2023 年提出一种 LiDAR-视觉混合的细粒度跨视图定位方法 SIBCL, 他们也是在两个标准数据集 KITTI 和 Ford Multi-AV Seasonal 中评估所提出的方法. 根据数据集中提供的 GNSS 标签, 从谷歌地图中收集对应的卫星图像, 正式提出 FordAV-CVL 数据集. 具体而言, 他们先根据 GNSS 信息找到覆盖车辆轨迹的区域, 然后将该区域进行均匀划分得到重叠的卫星图像块, 这些卫星图像块的尺寸为 1280 × 1280

表 6 面向自动驾驶的跨视图定位数据集

Table 6 Cross-view localization datasets for autonomous driving

数据集	发布时间	图像分辨率 (像素)	图像类型	数据量	地点
KITTI-CVL[52]	2023	1242 × 375 (地面), 1280 × 1280 (卫星)	地面图像和卫星图像	30973 幅地面视角图像和 2750 幅卫星图像	德国卡尔斯鲁厄市
FordAV-CVL[52]	2023	1656 × 860 (地面), 1280 × 1280 (卫星)	地面图像和卫星图像	30209 对地面视角和卫星图像	美国密歇根州
Oxford RobotCar(+)[58]	2022	1280 × 960 (地面), 600 × 600 (卫星)	地面图像和卫星图像	24000 对地面视角和卫星图像	英国牛津
KITTI(+)[60]	2022	1242 × 375 (地面), 1280 × 1280 (卫星)	地面图像和卫星图像	30973 幅地面视角图像和 2750 幅卫星图像	德国卡尔斯鲁厄市
Ford Multi-AV[60]	2022	1656 × 860 (地面), 1280 × 1280 (卫星)	地面图像和卫星图像	30209 对地面视角和卫星图像	美国密歇根州
CSC-Shanmao (本文数据集)	2025	960 × 540 (地面), 1280 × 1280 (卫星)	地面图像和卫星图像	24115 对地面视角和卫星图像	中国湖南省长沙市以及内蒙古包头市

像素. 与前述工作一样, KITTI-CVL 数据集的卫星图栅格分辨率为 0.2 m 每像素, 而 FordAV-CVL 数据集的卫星图栅格分辨率为 0.22 m 每像素. 他们在测试中发现, 从谷歌地图中获取的卫星图像存在轻微的偏移, 因此在使用 FordAV-CVL 数据集中的地面真值时会出现卫星图像与地面图像不匹配的情况. 通过检查 Ford Multi-AV Seasonal 数据集的轨迹, 他们最终选择具有最佳卫星配准的“log4”轨迹进行方法评估. 同年, Wang 等[64] 还提出一种名为 PureACL 的跨视图定位方法, 用于准确的 3-DoF 姿态估计, 它也是首个基于纯视觉的稀疏视觉定位方法. 该方法所采用的数据集也是 KITTI-CVL 和 FordAV-CVL.

与 KITTI-CVL 相比, FordAV-CVL 的主要优势在于其地面数据包含了不同季节 (夏、冬) 的采集结果, 为模型提供了初步的跨季节泛化测试基础. 但其局限性同样突出: 首先, 其季节和天气变化仍然有限, 缺乏对暴雨、浓雾等极端条件的系统性覆盖; 其次, 其地理范围仍局限于美国密歇根州的特定区域; 最后, 从谷歌地图获取的卫星图像存在难以消除的配准偏差, 为算法评估引入了额外的噪声.

3.3 由 Oxford RobotCar 衍生的数据集

Oxford RobotCar 数据集是一个包含多传感器测量的丰富数据集, 其涵盖了自动驾驶车辆在牛津一条路线上的多次行驶数据记录.

Xia 等[58] 在 2022 年的工作中使用了 Oxford RobotCar 作为基础数据集, 并在其基础上使用已有的卫星图像补丁, 同时他们收集额外的卫星数据, 创建了覆盖目标区域的连续卫星地图, 总共获得来自 15 条路径的 23854 对匹配数据, 得到跨视图定位领域专用的 Oxford RobotCar(+) 数据集. 在 Oxford RobotCar(+) 数据集中, 训练、验证和测试集分别包含 17067、1698 和 5089 张地面前视图像, 每张卫星图像的分辨率为 600 × 600 像素, 区域大小为 55.44 m × 55.44 m, 卫星图栅格分辨率约为 0.01 m 每像素. 该数据集提供了长达一年的数据采集跨度, 包含更为丰富的光照和天气变化. 然而, 其适用性仍然受限: 场景单一 (仅牛津一条固定路线) 和相对有限的卫星图像分辨率与覆盖范围, 制约了其作为评估模型在多样地理与拓扑结构中泛化能力的通用基准.

3.4 CSC-Shanmao 数据集

鉴于上述主流数据集在时间跨度、场景多样性与地理分布上的固有局限性, 以及领域内普遍缺乏对模型跨域泛化能力的系统性评估, 本文构建一个全新的自动驾驶跨视图定位数据集——CSC-Shanmao. 该数据集由国防科技大学自研的“山猫”全地形无人作战平台采集得到, 旨在提供更具时效性和多样性的测试基准. 数据集的采集时间跨度覆盖 2024 年 11 月至 2025 年 3 月, 包含了从深秋到初春的季节变化以及晴天、阴天等多种天气条件, 以增加环境多样性.

为获取高精度真值, 数据采集平台搭载了 NovAtel PwrPak7-E2 双频 GNSS/INS 紧组合系统, 在开阔的天空条件下可实现厘米级的实时定位精度与 0.05° 的姿态精度. 对于卫星图像, 我们根据车辆的高精度 GNSS/IMU 坐标, 以 19 级缩放级别匹配了分辨率为 1280 × 1280 像素的对应卫星正射影像. 为进一步消除系统性偏差, 我们采用基于特征点的图像配准方法, 对约 5% 的样本进行了人工抽查与微调, 确保地面视点与卫星图像的中心对齐误差平均小于 3 个像素 (约合 0.6 m).

为全面评估模型在不同环境下的适应能力, CSC-Shanmao 数据集被设计为包含两个具有显著环境差异的子集, 共计包含 24115 组数据对.

一是结构化道路子集 (常规场景): 包含 13 条

轨迹, 共 15 070 组车辆前视视角图像及其对应的 3-DoF 位姿真值. 该子集主要覆盖具备清晰车道线和规则路面特征的结构化道路场景. 我们按照 9:3:1 的比例将其划分为训练集 (9 条轨迹, 10 430 张图像)、测试集 (3 条轨迹, 3 480 张图像) 及验证集 (1 条轨迹, 1 160 张图像).

二是非结构化越野子集 (泛化性挑战场景): 为响应对模型鲁棒性测试的需求, 我们新增了采集于包头户外试验场的独立测试数据, 共包含 9 045 组前视图像及位姿数据. 该场景具有显著的北方冬季荒漠草原地貌特征: 植被呈现枯黄状态, 路侧树木落叶, 背景色调单一, 能有效模拟季节变化带来的视觉特征迁移; 另外, 采集时段光照强烈, 非铺装土路的高反光特性导致路面纹理弱化, 且路面颜色与周围枯草背景极度接近, 造成严重的视觉混叠, 缺乏清晰的道路边界与车道线特征. 该子集仅作为独立的测试集使用, 不参与模型训练, 专门用于评估模型在未知复杂环境下的零样本泛化能力.

对于所有数据, 我们根据车辆的高精度 GPS/IMU 坐标, 以 19 级缩放级别匹配了分辨率为 1 280 × 1 280 像素的对应卫星正射影像, 确保了视图对齐的精确性.

3.5 现有数据生态的综合分析

综合来看, 当前研究依赖于几个高度同质化且存在固有偏差的数据基准. 这导致了两个核心挑战:

一是数据偏差固化模型能力边界. 模型在 KITTI-CVL、FordAV-CVL 等数据集上的“高精度”, 本质是在特定地理、季节和道路类型的狭窄分布内优化的结果, 其泛化能力未被充分检验.

二是评估范式缺失关键一环. 主流研究范式几乎完全依赖“数据集内”评估, 严重缺乏跨数据集的、标准化的泛化性能测试. 一个模型在 KITTI-CVL 上表现最佳, 但在 FordAV-CVL 或真实场景中可能完全失效.

因此, 构建如 CSC-Shanmao 中“越野子集”这类针对性挑战集, 并推动将其纳入模型评估的必选项目, 是打破数据偏差闭环、驱动研发真正鲁棒算法的关键一步. 未来的基准建设应更加注重地理、季节、天气和道路类型的均衡覆盖, 并强制要求报告模型在分布外挑战集上的性能.

3.6 评价指标

本文沿用文献 [66] 中的评价体系, 使用所有测试样本的平均误差和中位误差作为定位与方向预测的主要评估指标. 对于定位, 误差是指预测位置与地面真实位置之间的距离, 单位为米; 对于方向, 误差是预测位置处预测相机的方向与地面真实相机的方向之间的角度差, 单位为度. 此外, 当误差低于一定阈值时, 如定位中的 1 m、3 m 和 5 m, 方向中的 1°、3° 和 5°, 这些测试样本的百分比也将被统计. 在定位指标中, 记录了沿车辆行驶方向的纵向和横向误差, ±10° 表示给定的方位先验范围.

3.7 实验细节

为验证 CSC-Shanmao 数据集的有效性及评估基准算法的泛化能力, 本文基于该数据集进行了相关测试工作. 所有实验均在统一的硬件平台 NVIDIA GeForce RTX3090 和软件环境 PyTorch1.12.0、CUDA11.6 下进行. 对于 CCVPE[67] 与 LM[60], 我们严格遵循其原始论文的超参数设置; 对于其他测试方法 (详见第 4 节), 我们同样尽力复现其官方开源代码中的默认配置. 需要说明的是, 尽管我们尽力扩大验证范围, 但部分表现优异的方法 (如 T2GA[68]、C2F-CCPE[77]) 因未公开代码而无法在本数据集上验证; BoostAcc[63] 等方法则因模型权重文件缺失导致复现失败. 这一现象本身也凸显了当前领域在代码开源与可复现性方面面临的普遍挑战.

本节系统梳理了支撑自动驾驶跨视图定位研究的关键数据集及其评价体系. 基于上述数据集, 尤其是 KITTI-CVL、FordAV-CVL 以及本文提出的 CSC-Shanmao, 第 4 节将集中呈现与分析各主流方法的定量性能, 并深入探讨不同方法在不同数据集上所展现出的性能特点与泛化能力.

4 实验结果展示

基于第 3 节介绍的数据集与评估指标, 本节对现有方法的性能进行系统性对比与分析. 依据表 3 以及图 4、图 5 的信息, 本文总结出当前面向自动驾驶的跨视图定位方法的性能对比图, 如图 6 所示. 这些模型都是基于 KITTI 数据集进行实验测试的, 不同的颜色表示方法使用了不同的初始先验误差进行实验, 具体的图例已在图中标注. 对于有中值误差结果的模型, 我们选取方向或定位的中值误差作为核心评价指标; 对于没有中值误差结果的模型, 本文通过综合对比横向、纵向误差来实现模型性能的比较.

为验证算法及自制数据集的有效性, 我们在 CSC-Shanmao 数据集上对多种方法进行测试. 本文的选择标准兼顾了性能代表性与可复现性: LM[60] 方法作为被广泛对比的基准模型; CCVPE[67] 作为公开的高精度模型代表. 为进一步增强验证的覆盖度与结论的稳健性, 我们还成功复现并测试了同样提供完整开源代码的文献 [66] 中的方法和 FG^2[86]

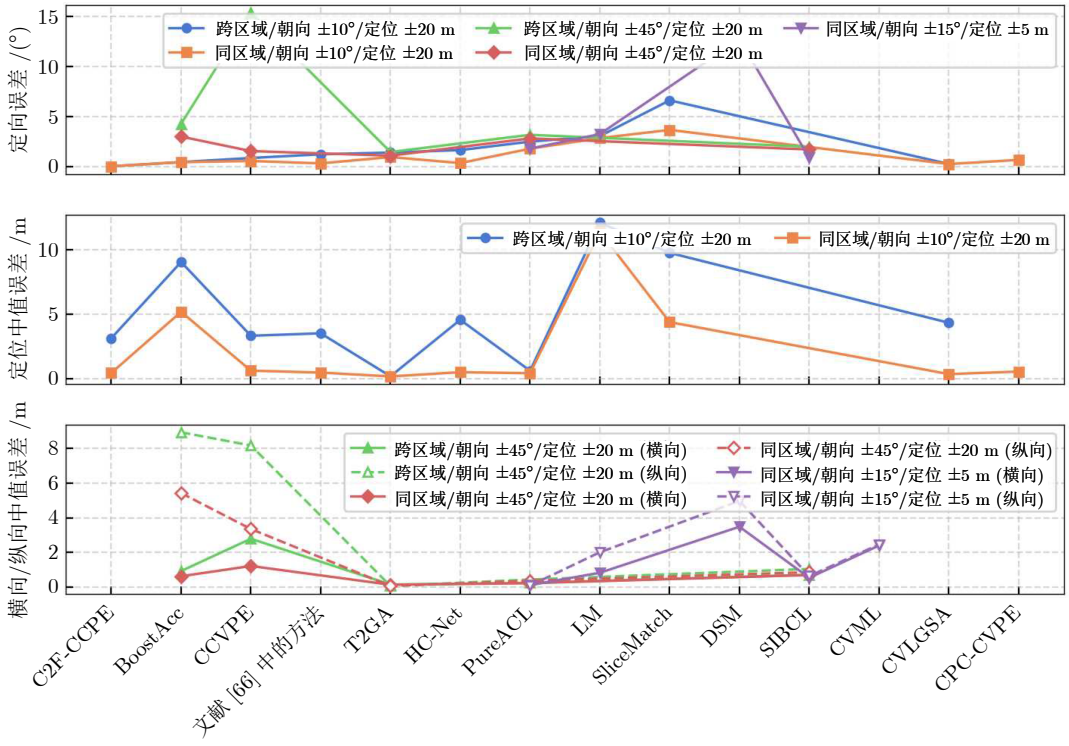


图 6 面向自动驾驶的跨视图定位方法性能对比

Fig. 6 Performance comparison of cross-view localization methods for autonomous driving

方法. 所有实验均在统一的硬件与软件环境下进行, 并严格遵循各方法官方开源代码的默认配置与训练策略 (详见第 3.7 节).

综合测试结果 (如表 7 所示) 表明, 不同方法在 CSC-Shanmao 数据集上展现出显著差异的性能特性与条件适应性: CCVPE 在第一测试条件 (结构化数据) 下展现出卓越的定位精度, 其横向误差 1 m 内的召回率高达 73.02%, 定位中值误差仅为 1.58 m, 但其性能在第二测试条件 (非结构化数据) 下急剧下降, 定位误差大幅增加, 凸显了其对测试条件极为敏感、稳定性不足的特点. 同时, 其方向估计能力在两个条件下均较弱. LM 方法表现出最佳的鲁棒性与均衡性, 其方向估计在两个条件下均保持领先 (第一测试条件下中值误差低至 1.25°), 定位性能在第二测试条件下最优. 这表明经典的基准方法在泛化能力和可靠性上仍具有重要价值. 文献 [66] 中的方法在两个测试条件下, 定位与方向性能均不突出, 说明该方法可能对该数据集或特定任务场景的适应性有限. FG^2 方法在定位性能上表现稳定且适中, 在两个条件下未出现剧烈波动, 但其方向估计精度有待提升, 这体现了该方法有一定的泛化能力, 但在绝对精度上存在优化空间. 具体结果如表 7 所示.

进一步深入分析可以发现, 这些性能差异与各方法的核心技术设计密切相关.

CCVPE[67] 在第一测试条件下的卓越表现得益于其基于对比学习的特征匹配机制与密集特征投影设计, 能够在纹理丰富、结构清晰的道路场景中实现高精度像素级对齐. 然而, 在第二测试条件下, 其性能急剧下降 (定位中值误差增至 37.69 m), 凸显了其对纹理与结构先验的强依赖——当面对光照强烈、纹理退化、道路边缘模糊的越野场景时, 纯视觉匹配机制易受环境干扰, 难以维持稳定性能; 同时, 其方向估计能力在两个条件下均较弱 (误差大于 5°), 这可能与其将方向回归作为辅助任务的损失函数设计有关, 未显式建模几何约束.

LM[60] 表现出最佳的鲁棒性与均衡性. 这源于其经典的几何投影模块与 Levenberg-Marquardt 优化框架, 通过显式建模地面–卫星视图间的投影关系, 增强了对几何结构的理解, 因此在场景外观剧烈变化时仍能保持较好的泛化能力. 这表明在复杂多变场景中, 显式几何建模仍具有不可替代的鲁棒性优势.

文献 [66] 中的方法在两个测试条件下定位与方向性能均不突出, 可能与其采用的 RAFT 架构有关: 该架构依赖稠密光流估计, 对图像质量与纹理连续性要求较高, 在低纹理、高反光的越野场景中易产生匹配歧义.

表 7　现有先进方法在 CSC-Shanmao 数据集上的性能评测

Table 7　Performance evaluation of state-of-the-art methods on the CSC-Shanmao dataset

测试条件与方法		定位 ↓ (m)		横向 ↑ (%)			纵向 ↑ (%)			方向 ↓ (°)		方向 ↑ (%)		
		均值	中值	1 m	3 m	5 m	1 m	3 m	5 m	均值	中值	1°	3°	5°
±10°, 结构化数据	LM[60]	8.69	4.96	24.74	57.93	65.95	31.21	62.07	76.55	**1.68**	**1.25**	**42.16**	**83.45**	**95.09**
	文献 [66] 中的方法	16.06	15.61	0.66	4.45	16.12	4.59	14.59	25.43	3.98	3.31	15.40	46.26	69.74
	CCVPE[67]	**5.99**	**1.58**	**73.02**	**88.62**	**90.43**	**47.16**	**71.63**	**78.53**	5.99	5.64	6.98	22.93	43.19
	FG^2[86]	8.09	7.58	16.49	50.46	76.12	8.62	26.67	42.27	7.80	7.00	5.66	18.28	33.45
±10°, 非结构化数据	LM[60]	**4.77**	**3.86**	**31.47**	**56.70**	**81.23**	**41.48**	**66.08**	**77.10**	**2.46**	**1.56**	**41.64**	**66.40**	**83.44**
	文献 [66] 中的方法	20.23	19.20	2.44	8.20	16.51	3.48	11.28	18.50	4.84	4.20	11.75	36.14	58.04
	CCVPE[67]	35.59	37.69	0.03	0.07	0.09	0.02	0.07	0.11	7.96	8.48	0.08	0.21	0.31
	FG^2[86]	8.82	7.31	13.88	41.18	63.38	9.35	31.52	50.36	10.32	9.19	4.20	12.35	23.50

FG^2[86] 在定位性能上表现稳定且适中, 在两个条件下未出现剧烈波动, 体现了其基于 Transformer 的全局注意力机制在一定程度上缓解了局部特征退化问题. 但其方向估计精度仍有提升空间, 可能源于其查询机制对旋转变化的建模不够敏感.

这些结果表明, 面向自动驾驶的跨视图定位方法在性能与鲁棒性之间存在显著的设计权衡: 基于稠密匹配的方法在理想条件下精度突出, 但对环境变化敏感; 引入显式几何建模的方法虽在绝对精度上可能不占优, 却展现出更好的跨场景稳定性. CSC-Shanmao 数据集通过结构化与越野子集的对照, 有效揭示了这一关键权衡, 为算法选型与改进提供了重要参考.

同时, 本次验证也揭示了一个与第 5.2.3 节论述相呼应的关键局限: 部分在 KITTI-CVL 上报告了顶尖性能的方法 (T2GA[68]、C2F-CCPE[77]), 由于其代码或模型权重未开源, 我们无法在 CSC-Shanmao 数据集上对其进行评估与比较. 这一事实恰恰凸显了领域内在代码开源与实验复现方面的普遍挑战. 因此, CSC-Shanmao 数据集作为一个公开可用的基准, 其另一项重要价值在于推动构建更公平、透明、可复现的算法比较环境, 以促进领域内更扎实的知识积累.

从前面的论述以及图表可以看到, 跨视图定位技术作为自动驾驶高精度定位的关键支撑, 在自动驾驶中的应用已然取得了一定进展, 正历经从单模态朝着多模态、从静态场景迈向动态适应性、从粗粒度转向细粒度感知的技术转变. 借助卫星图像与地面图像、LiDAR 等多源数据的异构匹配, 这些方法打破了传统定位方法对高精度地图的依赖, 还达成了亚米级位置估计与姿态解算的平衡.

5　挑战与未来展望

基于前文对技术发展、方法设计与实验性能的系统性回顾, 可以清晰地凝练出该领域在迈向实际落地过程中所面临的“结构性矛盾”与“共性挑战”. 本节将不再罗列现象, 而是致力于将这些挑战归因于更深层的技术或评估范式局限, 并据此指明有突破潜力的未来方向. 图 7 列出了面向自动驾驶的跨视图定位技术目前存在的关键问题以及未来潜在的研究方向.

5.1　算法模型层面的挑战

5.1.1　模型轻量化部署与计算效率瓶颈

随着模型复杂程度不断提升 (例如 Transformer 架构的广泛应用), 对计算资源的需求急剧增加. 比如 TransGeo 模型在处理大规模数据集时, 其训练时间以及推理时间都有了明显增长. 当这些模型部署于无人机、自动驾驶汽车等计算资源严格受限的边缘平台时, 其庞大的计算开销与设备有限的算力之间的矛盾日益突出, 构成了模型落地的核心瓶颈.

当前, 大多数研究在追求更高定位精度的同时, 未对模型的计算效率进行系统性优化. 部分复杂模型 (如基于 RAFT 架构的方法) 在典型车载嵌入式硬件上的推理帧率较低, 难以直接满足高频率定位输出的需求. 然而, 对“实时性”的理解应超越单一的帧率阈值. 在实际的自动驾驶系统中, 视觉定位模块的输出通常与高频 IMU 进行紧耦合传感器融合. 这种系统级设计可以在一定程度上补偿单一模块的瞬时帧率不足, 通过 IMU 插值维持位姿输出的连续性与高频刷新.

因此, 纯粹的推理帧率只是问题的一个侧面, 更本质的挑战在于: 如何在给定苛刻的车载计算资源预算内, 实现定位算法精度、鲁棒性与计算开销三者的最优平衡, 并为感知、规划等其他关键任务预留必要的算力. 这要求未来的研究从“单纯优化精度”转向“系统级的效率协同优化”.

具体而言, 当前为追求高定位精度而设计的复

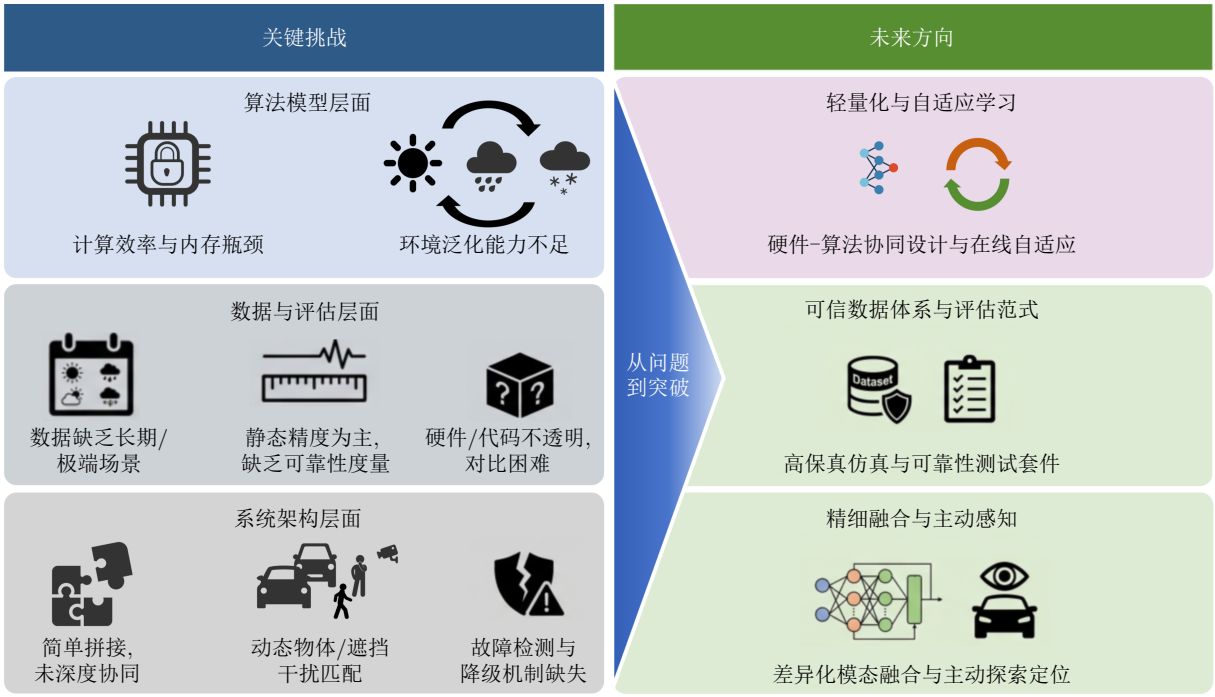


图 7 跨视图定位方法未来研究方向

Fig. 7 Future research directions of cross-view localization methods

杂模型结构, 在部署于边缘设备时普遍面临三大瓶颈. 一是内存瓶颈, 许多先进模型依赖庞大的参数量与深层网络. 例如, CNN 的连续卷积层累积产生大量中间特征图, 而基于 Transformer 的架构则需存储多头自注意力的巨大权重矩阵和键值缓存, 这些内存需求极易超出边缘设备有限的片上内存容量 (通常在数吉字节 (GB) 至数十吉字节 (GB) 范围), 导致昂贵的离片数据交换, 严重拖慢推理速度. 二是功耗瓶颈, 多分支或多模态融合网络因并行处理多条数据流而显著增加计算单元的活动因子; 自注意力机制的二次方复杂度在处理高分辨率输入时可能导致计算量爆炸, 而用于处理点云等数据的 3D 卷积或图神经网络同样计算密集. 这些操作都可能突破移动平台严格的功耗预算 (通常要求从数瓦到数十瓦不等). 三是推理速度瓶颈, 上述计算复杂度和内存瓶颈共同转化为过高的推理延迟. 对于需要高频定位的自动驾驶系统, 端到端推理延迟成为关键瓶颈, 难以满足实时闭环控制的刚性要求. 这些瓶颈并非孤立存在, 而是常见于追求高性能的模型设计中. 例如, 基于 Transformer 或大型 CNN 的编码器主要用来提取鲁棒的跨视图特征, 但会带来高计算与内存开销; 为提升定位精度而引入密集匹配或光流估计网络时, 其本身的计算成本就极高; 多模态融合模块又会由于额外的特征对齐与融合分支增加并行计算和内存占用的复杂度.

因此, 如何在内存、功耗、推理延迟这三重严格约束下维持或权衡定位精度, 是模型能否实际部署的核心挑战. 尽管模型剪枝、量化等轻量化技术为缓解这一挑战提供了可能, 但当前研究对其缺乏系统性应用与硬件层面的实证评估, 使得“轻量化部署”本身仍是一个悬而未决的问题.

5.1.2 纯视觉模型的精度上限与环境泛化局限

尽管现有的跨视角定位方法已取得显著进展, 但在追求高精度、全天候定位目标时, 仍面临着严峻挑战, 尤其是模型在复杂多变环境下的泛化能力问题尤为突出.

纯视觉模型对训练数据的分布具有高度依赖性. 以 SOTA 模型 C2F-CCPE 为例, 在 KITTI 数据集上, 当给定 $\pm 10^\circ$ 先验方向时, 其在同区域测试中的定位中值误差可低至 0.44 m; 然而在跨区域测试中, 该误差急剧上升至 3.14 m. 这表明当测试场景特征与训练场景特征存在差异时, 模型性能会显著衰退.

更为关键的是, 面对从“城市结构化道路”到“野外非结构化复杂环境”的域偏移时, 现有模型的鲁棒性表现出明显的局限性. 我们在 CSC-Shanmao 数据集上的实验进一步证实了这一点: 当将在城市道路上训练的模型 (如 LM 和 CCVPE) 直接应用于新增的包头越野测试集时, 各项指标出现了明显的下降.

这种性能衰减主要归因于环境特征的剧烈变化: 强光照下的高曝光导致路面纹理信息丢失, 而冬季枯黄的植被背景使得道路边缘与周围环境在色彩与纹理上高度相似, 极易混淆. 模型无法依赖城市环境中常见的车道线、路沿石等强语义特征进行定位. 这充分说明, 在视觉特征纹理匮乏或存在强光干扰的情况下, 仅依靠单一视觉模态难以维持高精度定位.

5.2 数据与评估体系层面的挑战

5.2.1 数据基础的时空局限性与真实性缺失

当前模型的泛化能力在面对光照、季节、天气等环境变化时表现不足, 其深层原因可追溯至支撑模型训练与评估的数据基础存在固有缺陷.

首先, 现有数据集存在显著的时空局限性. 主流数据集如 KITTI-CVL 本质上是单次、短时间跨度的采集, 完全缺乏对同一区域四季更替与长期场景演变 (如植被荣枯、道路维修) 的覆盖; FordAV-CVL 数据集虽然包含了不同季节的采集, 为跨季节泛化提供了初步基础, 但仍缺少对连续性时序变化 (如从黎明到黄昏的全天候光照循环) 的系统记录. 同时, 这些数据集在“场景多样性”上覆盖不足, 普遍缺少对暴雨、浓雾等极端天气的充分呈现. 这种时空覆盖的狭隘性使得模型的训练与测试过程处于一个被简化的、“时间纯净”的环境假设中, 无法充分验证其对真实世界动态复杂性的适应能力, 构成了模型泛化瓶颈的关键数据根源.

其次, 更为严峻的是, 现有数据与评估几乎完全忽略了传感器故障这一现实风险. 在真实驾驶环境中, 传感器部分失效 (如相机镜头被遮挡、LiDAR 点云因污损而稀疏) 是常见的意外情况. 此外, 即使传感器硬件工作正常, 多模态融合系统还面临更深层的“数据真实性”挑战: 异源传感器之间的时间同步误差, 以及因车辆振动、热胀冷缩导致的外参标定 (相机、LiDAR、IMU 的相对位姿关系) 漂移. 这些时空层面的细微误差在高速运动或长距离推算中会被显著放大, 直接损害特征对齐与融合的一致性, 是工程部署中影响定位精度的关键瓶颈. 然而, 当前研究普遍缺乏对这些故障场景的系统性模拟与测试, 导致无法评估多模态定位系统在面临单模态或跨模态数据缺失/退化时的定位容错能力与性能衰减边界.

这一“真实性”的缺失带来了本质性的安全性验证缺口. 一个在理想多模态数据上表现优异的系统, 可能在某个关键传感器失效时性能发生不可预测的骤降, 而系统自身并无相应的故障检测与自适应重构机制. 因此, 传感器故障模拟的缺失, 不仅是数据层面的不足, 更是系统级安全性评估的盲区, 它使得当前许多研究的“鲁棒性”结论在实际部署中面临严峻挑战.

由此可以分析得到, 数据集的时空局限性与传感器故障场景的缺失, 共同构成了当前跨视图定位技术从“实验室验证”迈向“道路应用”的关键数据瓶颈. 因而构建涵盖长期时序、多极端条件并包含系统化故障注入的新一代基准数据集, 是评估与提升系统实际泛化能力与功能安全性的迫切前提.

5.2.2 评估指标难以衡量系统的实际可靠性

与数据局限并行的是评估体系的片面性. 当前工作严重依赖定位误差与召回率等静态精度指标, 缺乏对系统在实际运行中关键特性的衡量: 一是定性平滑性指标, 如定位结果的空间一致性与帧间稳定性, 这对于自动驾驶控制的平顺性至关重要; 二是模型输出的置信度评估, 未能通过精度–置信度曲线分析系统“何时值得信赖”; 三是边界场景下的专项评估, 这种“唯精度论”的评估导向, 难以驱动高可靠、可解释的定位系统发展.

值得注意的是, 上述“定性指标”的引入并非仅意味着评估维度的简单扩充, 更深层的解决方案在于: 将空间一致性、帧间稳定性等系统级要求转化为模型训练或后处理中的优化约束. 例如, 可以在损失函数中引入基于时序的平滑性正则项, 或利用卡尔曼滤波等运动模型对定位轨迹进行后优化, 从而在提升定量精度指标的同时, 自然满足实际应用对输出稳定性的内在要求. 同样地, 发展能够提供校准化不确定性估计的模型, 使系统能够依据置信度动态权衡精度与可靠性, 是比静态阈值分析更加本质的进步. 因此, 未来的评估体系应鼓励从“静态精度对比”向“在系统约束下寻求最优平衡”的范式演进.

5.2.3 实验规范与对比公平性挑战

当前研究的可重复性与对比公平性还面临一个基础性挑战, 即实验环境报告的严重不透明. 在本文所综述的大量工作中, 相当一部分未明确报告训练与推理所依赖的硬件配置 (如 GPU 型号、内存大小), 而在已报告的工作中, 硬件平台也差异巨大 (从消费级 GPU 到数据中心级加速器).

这种透明度的缺失对领域发展构成三重危害: 首先, 它直接损害了效率对比的公平性与意义. 不同硬件平台对算法的推理速度 (FPS) 和功耗有决定性影响, 缺乏统一基准将难以评估任何效率指标. 其次, 它为高精度指标的细微对比引入了不可控的噪声. 虽然硬件不直接决定算法上限, 但不同的计算平台可能通过影响训练时的随机性、并行计算顺

序及可支持的批处理大小, 导致复现结果出现波动, 这在顶尖方法进行“厘秒之争”时尤为关键. 最后, 它完全阻断了模型的部署可行性分析, 使研究者无法评估算法在特定算力约束下的实际应用成本. 因此, 推动建立强制公开硬件配置、完整代码与核心超参数的标准化实验规范, 并鼓励在共享基准平台或提供完整 Docker 环境上进行效率评估, 是确保研究可复现、对比公平并最终能导向实际部署的关键步骤.

需要说明的是, 本文作为综述, 对各类方法性能的比较与分析均是基于其原始文献在相同或相近实验条件设置下 (如初始先验误差范围) 所报告的公开结果. 这一分析框架旨在梳理技术发展的脉络与相对性能趋势. 然而, 正如前文所述, 由于各文献在硬件环境、代码实现细节上缺乏统一标准, 这种跨研究的对比, 尤其是在推理效率以及毫厘之间的精度差异方面, 其结论可能受到底层不可控因素的潜在干扰. 这进一步凸显了建立统一、透明实验规范的必要性——它不仅关乎单个研究的可复现性, 也决定了整个领域能否在坚实、可比的基础上进行有效的知识积累与性能评估.

5.3 系统架构层面的挑战

5.3.1 多模态融合策略的粗放性

当前采用多模态输入 (如地面图像+LiDAR 点云+卫星图像) 的方法为提升鲁棒性提供了可能, 但其融合策略往往较为粗放. 多数研究采用“先编码后拼接”或简单的特征图叠加方式, 未能深度挖掘并协同利用不同模态的固有优势. 例如, LiDAR 数据提供精确的 3D 几何与距离不变性, 但对纹理语义不敏感; 卫星图像蕴含丰富的全局地理上下文与拓扑信息, 但分辨率有限且存在视角畸变; 地面图像则包含细致的纹理与动态物体信息, 但视角狭窄. 现有的“简单拼接+统一编码”模式很容易导致模态间信息冗余或互补性信息丢失, 例如, 用相同的卷积网络处理本质不同的点云与图像数据时, 无法最优地提取各自的独特特征.

未来研究需要探索更精细的差异化特征提取与动态融合机制, 比如设计模态专用的编码器、引入基于注意力的跨模态特征交互或在几何约束下进行特征对齐与选择, 以实现“$1+1+1>3$”的真正协同.

5.3.2 动态场景与严重遮挡带来的匹配挑战

当前绝大多数跨视图定位方法都建立在“静态、无遮挡环境”的隐含假设之上. 然而, 真实城市道路充满动态物体 (如车辆、行人) 和各类遮挡 (如树木、桥梁、临时路障), 它们会严重干扰地面视图与卫星视图之间的特征提取与匹配, 构成实际部署中的严峻挑战.

动态物体会在地面图像中引入瞬时、局部的错误特征 (即“特征污染”). 例如, 一辆临时停靠的卡车可能完全遮挡住用于匹配的关键道路标志或建筑立面; 大量移动的车辆会改变道路的纹理模式. 这些特征在静态的卫星图像中不存在, 导致基于全局外观匹配的方法 (尤其是早期方法) 极易产生误匹配. 尽管以 Transformer 为代表的新架构, 其自注意力机制在一定程度上具备从全局上下文中滤除局部噪声的能力, 但这种能力是隐式、被动的, 且尚未经过针对动态场景的专门设计和系统验证.

因此, 将“动态与遮挡鲁棒性”作为跨视图定位的一个核心研究议题被明确提出, 并发展相应的评估基准、数据增强方法和专用网络模块, 是推动该技术从实验室走向复杂现实世界的必经之路.

5.3.3 系统级的安全可靠性与工程化挑战

自动驾驶定位系统作为安全关键系统的核心模块, 其价值不仅在于平均精度, 更在于最差情况下的表现 (如定位失败、性能断崖式下降) 以及从异常中恢复的能力. 当前研究多聚焦于提升理想条件下的精度, 对以下工程化安全挑战的系统性探讨严重不足.

一是对边缘情况与异常输入的脆弱性研究不足. 现有方法极少测试在极端恶劣天气 (如暴雨)、剧烈运动模糊、传感器瞬时故障 (如单帧丢图) 或非预期输入 (如驶入在建工地) 下的表现. 模型在这些“分布外”场景下可能给出一个高置信度但完全错误的位置, 这比直接报告失败更为危险.

二是缺乏故障检测与功能降级机制. 一个健壮的系统应具备自监控能力, 包括: 输入可信度评估, 即实时检测传感器数据质量 (如图像模糊度、点云密度), 并在质量过低时触发警告; 输出置信度校准, 即模型的置信度分数必须与真实误差概率严格校准, 当前模型输出的置信度往往过于乐观 (过度自信), 无法用于可信的决策, 置信度校准技术 (如温度缩放) 在该领域尚未得到重视; 故障隔离与功能降级, 即当检测到某模态 (如视觉) 失效时, 系统应能无缝切换至降级模式 (如依赖更稳定的 IMU/轮速计进行航位推算, 或提示驾驶员接管), 而非完全崩溃.

三是系统冗余设计的缺失. 工程上的高可靠性往往通过冗余实现, 包括传感器冗余 (如多摄像头互为备份) 和算法冗余 (如多个独立定位算法投票). 当前研究多为单个“最优”模型设计, 缺乏对异构冗余架构如何设计、如何低成本实现以及如何管理其

一致性的探索.

5.4 未来展望

针对上述挑战, 面向自动驾驶的跨视图定位技术正朝着更高效、更可靠、更智能的方向演进. 以自动驾驶数据集为基础的深度学习方法将跨视图定位技术的研究推向新的高度, 也使得跨视图定位技术具有更广阔、更现实的应用场景. 未来的研究将不仅聚焦于提升模型精度, 更需关注其在真实复杂环境下的系统性能. 本节将前述挑战转化为具体的技术攻关方向, 旨在为后续研究提供明确的突破切入点.

5.4.1 轻量化与自适应学习: 实现精度–效率–泛化性的协同进化

为应对模型计算效率与环境泛化能力不足的挑战, 未来的算法研究需从模型本质与信息输入维度双向出发进行优化.

一方面, 针对实时性需求, 需超越简单的模型压缩, 研究硬件–算法协同设计. 具体路径包括: 开发面向车载 SoC (如 NVIDIA Orin、地平线 J5) 的自动神经网络架构搜索框架, 在给定的内存、功耗预算下直接搜索最优网络; 探索动态稀疏化与混合精度推理, 使模型能根据输入场景复杂度自适应调整计算图.

另一方面, 针对泛化性瓶颈, 应发展在线自适应与持续学习框架, 并超越简单的多模态特征堆叠, 深入挖掘语义地图与传感器数据中的“环境不变性”作为强先验, 指导视觉模型在不同域之间实现特征对齐, 使模型能够利用少量新数据在部署后持续适应新的环境、季节甚至传感器特性, 打破静态数据集带来的泛化壁垒.

此外, 真实的商业部署对“快速适配”提出了更高要求. 当前模型在面对数据采集不足的陌生环境时, 因缺乏小样本学习能力, 往往需要代价高昂的重新训练.

5.4.2 可信数据体系与评估范式: 构建模型能力的“压力测试场”

为系统解决数据偏差与评估缺陷, 未来的研究必须重建基准体系.

一方面, 构建开源、可编程的仿真与数据生成平台, 鼓励开发基于游戏引擎 (如 Unreal Engine) 或物理渲染的高保真开源仿真器, 允许研究者参数化控制季节、天气、光照、传感器噪声甚至故障模式, 系统性生成覆盖长尾场景的无限量合成数据. 同时, 可以利用扩散模型等生成式 AI, 从真实数据中学习并生成难以采集的极端条件场景.

另一方面, 建立分级、强制的可靠性评估标准, 倡议业界建立“自动驾驶定位可靠性测试套件”. 该套件应包含基础精度集 (如现有数据集)、泛化挑战集 (如 CSC-Shanmao 的越野子集)、故障注入集 (如模拟传感器遮挡、数据丢失). 此外还要求论文报告在泛化集和故障集上的性能, 并将轨迹平滑度、不确定性校准误差列为与定位误差同等重要的核心指标.

5.4.3 精细融合与主动感知: 从被动匹配到智能定位

为系统性应对前述动态场景干扰与安全可靠性的严峻挑战, 未来的研究必须超越对静态、理想数据的被动匹配, 从系统架构层面探索更具适应性与智能性的新范式.

一方面, 研究基于注意力的差异化模态融合方法. 这不仅是提升精度的手段, 更是实现鲁棒感知与故障容错的关键. 可以设计非对称的跨模态 Transformer, 其中不同模态使用专用的编码器, 并通过可学习的交叉注意力机制进行动态特征交互与选择. 例如, 让 LiDAR 的几何特征作为“鲁棒查询”去检索卫星图像中的全局结构, 或让地面图像的纹理作为“语义查询”去增强点云的上下文理解. 这种细粒度、可解释的融合机制, 有助于模型在面临部分传感器数据污染或缺失时, 自适应地依赖更可靠的模态, 从而提升系统整体的容错能力与安全边界.

另一方面, 探索“具身智能”驱动的主动感知与定位. 这为解决定位模糊、置信度校准问题以及在异常场景下主动获取信息提供了新思路. 可以研究初步的交互式定位框架, 使车辆在定位置信度低或遭遇严重遮挡时, 能主动采取信息收集行动. 例如, 模型可联合输出定位结果以及一个同时编码不确定性与信息增益的地图, 决策模块据此规划一段短距离探索轨迹 (如稍微改变航向以获取新视角), 通过主动观察来消除歧义. 这不仅有望提高定位精度, 更能将系统从被动的“匹配者”升级为主动的“探索者”, 为实现“失效可感知、降级可操作”的高可用性定位系统奠定基础.

6 总结

本文对面向自动驾驶的跨视图定位技术进行了系统性综述. 首先, 我们梳理了该技术从通用地理检索迈向自动驾驶专用高精度姿态估计的三个发展阶段, 明确了其范式演进的驱动力. 进一步, 为深入解构各类方法, 本文创新性地提出一个从“输入数据、模型架构、评价指标与损失函数”四个维度出发

的交叉分析框架, 系统揭示了不同技术路线的设计哲学、优势及其性能边界.

为评估算法在复杂真实环境中的泛化能力, 本文构建并开源了 CSC-Shanmao 数据集, 其涵盖结构化城市道路与非结构化越野场景, 填补了现有基准在挑战性跨域测试方面的空白. 基于此数据集及主流公开基准的实验表明, 面向自动驾驶专有化设计的方法在精度与鲁棒性上优势显著, 但同时也深刻揭示了当前技术面临“精度−效率−泛化性”难以兼得的根本矛盾, 以及多模态融合系统在工程落地时面临的同步、标定与可靠性挑战.

综上所述, 未来研究需在以下方向寻求突破: 发展硬件感知的轻量化架构以平衡模型性能与车载算力约束; 构建涵盖极端条件与传感器故障模拟的标准化评估体系以驱动可信算法的研发; 探索融合主动感知与在线自适应能力的下一代定位范式, 使车辆能在未知或动态环境中主动减少定位歧义. 本文期望通过上述梳理与分析, 为领域研究者提供一份清晰的技术图谱与扎实的参考依据.

参考文献

1 Lin T Y, Cui Y, Belongie S, Hays J. Learning deep representations for ground-to-aerial geolocalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 5007−5015

2 Workman S, Jacobs N. On the location dependence of convolutional neural network features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 70−78

3 Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S A, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015, **115**(3): 211−252

4 Zhou B L, Lapedriza A, Xiao J X, Torralba A, Oliva A. Learning deep features for scene recognition using places database. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014. 487−495

5 Bromley J, Guyon I, LeCun Y, Säckinger E, Shah P. Signature verification using a “Siamese” time delay neural network. In: Proceedings of the 7th International Conference on Neural Information Processing Systems. Denver, USA: Morgan Kaufmann Publishers Inc., 1993. 737−744

6 Tian Y C, Chen C, Shah M. Cross-view image matching for geo-localization in urban environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 1998−2006

7 Vo N N, Hays J. Localizing and orienting street views using overhead imagery. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 494−509

8 Hu S X, Feng M D, Nguyen R M H, Lee G H. CVM-Net: Cross-view matching network for image-based ground-to-aerial geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 7258−7267

9 Arandjelović R, Gronat P, Torii A, Pajdla T, Sivic J. NetVLAD: CNN architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(6): 1437−1451

10 Zheng Z D, Wei Y C, Yang Y. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In: Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM, 2020. 1395−1403

11 Zheng Z D, Zheng L, Garrett M, Yang Y, Xu M L, Shen Y D. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2020, **16**(2): Article No. 51

12 Sun B, Chen C, Zhu Y Y, Jiang J M. GEOCAPSNET: Ground to aerial view image geo-localization using capsule network. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). Shanghai, China: IEEE, 2019. 742−747

13 Zhu Y Y, Sun B, Lu X F, Jia S. Geographic semantic network for cross-view image geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, **60**: Article No. 4704315

14 Zhu S J, Yang T J N, Chen C. VIGOR: Cross-view image geo-localization beyond one-to-one retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 5316−5325

15 Guo Y L, Choi M, Li K H, Boussaid F, Bennamoun M. Soft exemplar highlighting for cross-view image-based geo-localization. *IEEE Transactions on Image Processing*, 2022, **31**: 2094−2105

16 Gu B, Ling H F, Shi Y X, Li Z Y, Zhao C, Li P, et al. Mutual relative position learning Transformer for cross-view geo-localization. In: Proceedings of the IEEE International Conference on Image Processing (ICIP). Kuala Lumpur, Malaysia: IEEE, 2023. 286−290

17 Xia Z. Ground-to-aerial Image Matching for Vehicle Localization [Ph.D. dissertation], Delft University of Technology, The Netherlands, 2024.

18 Park J, Sung C, Lee S, Kang D, Myung H. Cross-view geo-localization via effective negative sampling. In: Proceedings of the 24th International Conference on Control, Automation and Systems (ICCAS). Jeju, South Korea: IEEE, 2024. 1078−1083

19 Zhai M H, Bessinger Z, Workman S, Jacobs N. Predicting ground-level scene layout from aerial imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 4132−4140

20 Cai S D, Guo Y L, Khan S, Hu J W, Wen G J. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2019. 8390−8399

21 Shi Y J, Liu L, Yu X, Li H D. Spatial-aware feature aggregation for cross-view image based geo-localization. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2019. Article No. 905

22 Wang J J, Li X. Cross-attention network for cross-view image geo-localization. In: Proceedings of the 6th International Symposium on Autonomous Systems (ISAS). Nanjing, China: IEEE, 2023. 1−6

23 Xia P W, Wan Y, Zheng Z, Zhang Y J, Deng J W. Enhancing cross-view geo-localization with domain alignment and scene consistency. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, **34**(12): 13271−13281

24 Zhang X H, Li X Y, Sultani W, Zhou Y, Wshah S. Cross-view geo-localization via learning disentangled geometric layout correspondence. In: Proceedings of the 37th AAAI conference on artificial intelligence. Washington, USA: AAAI Press, 2023. 3480−3488

25 Zhang X H, Li X Y, Sultani W, Chen C, Wshah S. GeoDTR+: Toward generic cross-view geolocalization via geometric disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(12): 10419−10433

26 Wang T Y, Zheng Z D, Yan C G, Zhang J Y, Sun Y Q, Zheng B L, et al. Each part matters: Local patterns facilitate cross-

view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, **32**(2): 867−879

27 Yagi R, Iwasaki A. Lightweight CNN for cross-view geo-localization using aerial image. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. Pasadena, USA: IEEE, 2023. 6266−6269

28 Ahn W J, Park S Y, Pae D S, Choi H D, Lim M T. Bridging viewpoints in cross-view geo-localization with Siamese vision Transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, **62**: Article No. 4707312

29 Ge F W, Zhang Y Z, Wang L, Liu Y X, Si P J, Zhang J J, et al. Joint representation learning based on feature center region diffusion and edge radiation for cross-view geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 2025, **63**: Article No. 5601513

30 Lin J L, Zheng Z D, Zhong Z, Luo Z M, Li S Z, Yang Y, et al. Joint representation learning and keypoint detection for cross-view geo-localization. *IEEE Transactions on Image Processing*, 2022, **31**: 3780−3792

31 Rodrigues R, Tani M. SemGeo: Semantic keywords for cross-view image geo-localization. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023. 1−5

32 Wang T Y, Zheng Z D, Zhu Z J, Sun Y Q, Yan C G, Yang Y. Learning cross-view geo-localization embeddings via dynamic weighted decorrelation regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, **62**: Article No. 5647112

33 Hu D, Yuan X, Xi H Y, Li J, Song Z B, Xiong F C, et al. Road structure inspired UGV-satellite cross-view geo-localization. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024, **17**: 16767−16786

34 Rodrigues R, Tani M. Are these from the same place? Seeing the unseen in cross-view image geo-localization. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2021. 3752−3760

35 Yan Y M, Wang M Y, Su N, Hou W, Zhao C H, Wang W X. IML-Net: A framework for cross-view geo-localization with multi-domain remote sensing data. *Remote Sensing*, 2024, **16**(7): Article No. 1249

36 Regmi K, Shah M. Bridging the domain gap for ground-to-aerial image matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2019. 470−479

37 Isola P, Zhu J Y, Zhou T, Efros A A. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 5967−5976

38 Zhang H, Xu T, Li H S, Zhang S T, Wang X G, Huang X L, et al. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 5908−5916

39 Toker A, Zhou Q J, Maximov M, Leal-Taixè L. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 6484−6493

40 Tian X Y, Shao J, Ouyang D Q, Zhu A J, Chen F Y. SMDT: Cross-view geo-localization with image alignment and Transformer. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). Taipei, China: IEEE, 2022. 1−6

41 Yang H J, Lu X F, Zhu Y Y. Cross-view geo-localization with layer-to-layer Transformer. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2021. Article No. 2222

42 Dai M, Hu J H, Zhuang J D, Zheng E H. A Transformer-based feature segmentation and region alignment method for UAV-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, **32**(7): 4376−4389

43 Zhu S J, Shah M, Chen C. TransGeo: Transformer is all you need for cross-view image geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 1152−1161

44 Wang Y F, Xia Y M, Lu T B, Zhang X Y, Yao W B. An efficient method based on multi-view semantic alignment for cross-view geo-localization. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN). Gold Coast, Australia: IEEE, 2023. 1−8

45 Zhang W H, Li J, Chen H, Wu J J. MT-GEO: A multi-scale feature extraction network for cross-view geo-localization between street-view and remote sensing imagery. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. Athens, Greece: IEEE, 2024. 6964−6968

46 Zhang W H, Zhong Z N, Chen H, Jing N. GeoViewMatch: A multi-scale feature-matching network for cross-view geo-localization using swin-Transformer and contrastive learning. *Remote Sensing*, 2024, **16**(4): Article No. 678

47 Li G P, Qian M, Xia G S. Unleashing unlabeled data: A paradigm for cross-view geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 16719−16729

48 Shore T, Mendez O, Hadfield S. PEnG: Pose-enhanced geo-localisation. *IEEE Robotics and Automation Letters*, 2025, **10**(4): 3835−3842

49 Zhou Bo-Wen, Li Yang, Ma Xin-Ji, Miao Zhuang, Zhang Rui. A survey of cross-view geo-localization methods based on deep learning. *Journal of Image and Graphics*, 2024, **29**(12): 3543−3563
(周博文, 李阳, 马鑫骥, 苗壮, 张睿. 深度学习的跨视角地理定位方法综述. 中国图象图形学报, 2024, **29**(12): 3543−3563)

50 Liu Lian, Li Jiang-Shan, Xu Jing-Qian, Yang Chun-Fang, Qin Yu-Lin, Liu Yan. Cross-view image geo-localization review. *Artificial Intelligence Security*, 2023, **2**(1): 53−61
(刘棟, 李江杉, 许经乾, 杨春芳, 秦雨林, 刘琰. 跨视角图像地理定位技术研究综述. 智能安全, 2023, **2**(1): 53−61)

51 Liu L, Li H D. Lending orientation to neural networks for cross-view geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 5617−5626

52 Wang S, Zhang Y H, Vora A, Perincherry A, Li H D. Satellite image based cross-view localization for autonomous vehicle. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). London, UK: IEEE, 2023. 3592−3599

53 Wu H, Zhang Z H, Lin S Y, Mu X R, Zhao Q, Yang M, et al. MapLocNet: Coarse-to-fine feature registration for visual re-localization in navigation maps. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, United Arab Emirates: IEEE, 2024. 13198−13205

54 Heng L, Choi B, Cui Z P, Geppert M, Hu S X, Kuan B, et al. Project AutoVision: Localization and 3D scene perception for an autonomous vehicle with a multi-camera system. In: Proceedings of the International Conference on Robotics and Automation (ICRA). Montreal, Canada: IEEE, 2019. 4695−4702

55 Shi Y J, Yu X, Campbell D, Li H D. Where am I looking at? Joint location and orientation estimation by cross-view matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 4063−4071

56 Dixit D, Verma S, Tokekar P. Evaluation of cross-view matching to improve ground vehicle localization with aerial perception. arXiv preprint arXiv: 2003.06515, 2020.

57 Zhang Y H, Shi Y J, Wang S, Vora A, Perincherry A, Chen Y B, et al. Increasing SLAM pose accuracy by ground-to-satellite image registration. In: Proceedings of the IEEE International

Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2024. 8522−8528

58 Xia Z M, Booij O, Manfredi M, Kooij J F P. Visual cross-view metric localization with dense uncertainty estimates. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 90−106

59 Wang X L, Xu R S, Cui Z F, Wan Z Y, Zhang Y. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 234

60 Shi Y J, Li H D. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 16989−16999

61 Shi Y J, Yu X, Wang S, Li H D. CVLNet: Cross-view semantic correspondence learning for video-based camera localization. In: Proceedings of the 16th Asian Conference on Computer Vision. Macao, China: Springer, 2022. 123−141

62 Shi Y J, Yu X, Liu L, Campbell D, Koniusz P, Li H D. Accurate 3-DoF camera geo-localization via ground-to-satellite image matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(3): 2682−2697

63 Shi Y J, Wu F, Perincherry A, Vora A, Li H D. Boosting 3-DoF ground-to-satellite camera localization accuracy via geometry-guided cross-view Transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 21459−21469

64 Wang S, Zhang Y H, Perincherry A, Vora A, Li H D. View consistent purification for accurate cross-view localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 8163−8172

65 Lentsch T, Xia Z M, Caesar H, Kooij J F P. SliceMatch: Geometry-guided aggregation for cross-view pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 17225−17234

66 Song Z B, Ze X H, Lu J F, Shi Y J. Learning dense flow field for highly-accurate cross-view camera localization. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. 70612−70625

67 Xia Z M, Booij O, Kooij J F P. Convolutional cross-view pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(5): 3813−3831

68 Wang S, Nguyen C, Liu J W, Zhang Y H, Muthu S, Maken F A, et al. View from above: Orthogonal-view aware cross-view localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 14843−14852

69 Zhang Z H, Xu M, Zhou W Q, Peng T, Li L, Poslad S. BEV-locator: An end-to-end visual semantic localization network using multi-view images. *Science China Information Sciences*, 2025, **68**(2): Article No. 122106

70 Jin L H, Dong W, Wang W S, Kaess M. BEVRender: Vision-based cross-view vehicle registration in off-road GNSS-denied environment. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, United Arab Emirates: IEEE, 2024. 11032−11039

71 Fervers F, Bullinger S, Bodensteiner C, Arens M, Stiefelhagen R. Continuous self-localization on aerial images using visual and LiDAR sensors. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Kyoto, Japan: IEEE, 2022. 7028−7035

72 Fervers F, Bullinger S, Bodensteiner C, Arens M, Stiefelhagen R. Uncertainty-aware vision-based metric cross-view geolocalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 21621−21631

73 Li L J, Ma Y K, Tang K, Zhao X R, Chen C, Huang J X, et al. Geo-localization with Transformer-based 2D-3D match network. *IEEE Robotics and Automation Letters*, 2023, **8**(8): 4855−4862

74 Guan T R, Xian R Q, Wang X J, Wu X Y, Elnoor M, Song D, et al. AGL-Net: Aerial-ground cross-modal global localization with varying scales. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, United Arab Emirates: IEEE, 2024. 8161−8161

75 Sarlin P E, DeTone D, Yang T Y, Avetisyan A, Straub J, Malisiewicz T, et al. OrienterNet: Visual localization in 2D public maps with neural matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 21632−21642

76 Guo X Y, Peng H C, Hu J J, Bao H J, Zhang G F. From satellite to ground: Satellite assisted visual localization with cross-view semantic matching. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2024. 3977−3983

77 Tang Y, Huang Q, Zhu Y Y. C2F-CCPE: Coarse-to-fine cross-view camera pose estimation. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). Niagara Falls, Canada: IEEE, 2024. 1−6

78 Li C R, Yan C, Xiang X J, Lai J, Zhou H, Tang D Q. HADGEO: Image based 3-DoF cross-view geo-localization with hard sample mining. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, South Korea: IEEE, 2024. 3520−3524

79 Zhu J L, Yan S, Cheng X Y, Wu R W, Liu Y X, Zhang M J. CLaSP: Cross-view 6-DoF localisation assisted by synthetic panorama. *IET Computer Vision*, 2024, **18**(7): 859−874

80 Miao J Y, Wen T P, Luo Z A, Qian K A, Fu Z, Wang Y L, et al. Efficient end-to-end visual localization for autonomous driving with decoupled BEV neural matching. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hangzhou, China: IEEE, 2025. 10719−10726

81 Zhu R Z, Yin L, Yang M Z, Wu F, Yang Y C, Hu W B. SUES-200: A multi-height multi-scene cross-view image benchmark across drone and satellite. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, **33**(9): 4825−4839

82 Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 2013, **32**(11): 1231−1237

83 Agarwal S, Vora A, Pandey G, Williams W, Kourous H, Mcbride J. Ford Multi-AV seasonal dataset. *The International Journal of Robotics Research*, 2020, **39**(12): 1367−1376

84 Maddern W, Pascoe G, Linegar C, Newman P. 1 year, 1000 km: The Oxford RobotCar dataset. *The International Journal of Robotics Research*, 2017, **36**(1): 3−15

85 Maddern W, Pascoe G, Gadd M, Barnes D, Yeomans B, Newman P. Real-time kinematic ground truth for the Oxford RobotCar dataset. arXiv preprint arXiv: 2002.10152, 2020.

86 Xia Z M, Alahi A. FG^2: Fine-grained cross-view localization by fine-grained feature matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 6362−6372

87 Chen G L, Huang G H, Yuan X C, Chen X H, Zhong G, Pun C M. Cross-view geo-localization via learning correspondence semantic similarity knowledge. In: Proceedings of the 31st International Conference on MultiMedia Modeling. Nara, Japan: Springer, 2025. 220−233

88 Liang J X, Bao M W, Dong H T, Xie L, Liu R W, Chen N C. DSTG: Distillation swin Transformer for cross-view geolocalization. *IEEE Transactions on Geoscience and Remote Sensing*, 2025, **63**: Article No. 5632917

89 Qin R Y, Liu Z Y, Wang K Y, Yuan X. Focusing on projection-

stable patch: Cross-view localization with geometric-semantic alignment. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hangzhou, China: IEEE, 2025. 15438−15444

90 Ding X K, Zhang X Y, Song S Z, Li B, Hui L, Dai Y C. Cross-view geo-localization via 3D Gaussian splatting-based novel view synthesis. *Remote Sensing*, 2025, **17**(22): Article No. 3673

91 Yuan X, Wang K Y, Qin R Y, Xu J C. A multi-modality ground-to-air cross-view pose estimation dataset for field robots. *Scientific Data*, 2025, **12**(1): Article No. 754

92 Durgam A, Paheding S, Dhiman V, Devabhaktuni V. Cross-view geo-localization: A survey. *IEEE Access*, 2024, **12**: 192028−192050

93 Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR′06). New York, USA: IEEE, 2006. 1735−1742

94 Girshick R. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1440−1448



段星宇 国防科技大学智能科学学院博士研究生. 2023 年获得国防科技大学学士学位. 主要研究方向为智能驾驶技术.
E-mail: duanxingyu19@nudt.edu.cn
(**DUAN Xing-Yu** Ph.D. candidate at the College of Intelligence Science and Technology, National University of Defense Technology. He received his bachelor degree from National University of Defense Technology in 2023. His main research interest is intelligent driving technology.)



吴 涛 国防科技大学智能科学学院研究员. 2002 年获得国防科技大学博士学位. 主要研究方向为自动驾驶, 感知规划. 本文通信作者.
E-mail: wutao@nudt.edu.cn
(**WU Tao** Researcher at the College of Intelligence Science and Technology, National University of Defense Technology. He received his Ph.D. degree from National University of Defense Technology in 2002. His research interests include autonomous driving and perception and planning. Corresponding author of this paper.)



李峻翔 国防科技大学智能科学学院副研究员. 2019 年获得国防科技大学博士学位. 主要研究方向为无人车规划与感知.
E-mail: lijunxiang15@nudt.edu.cn
(**LI Jun-Xiang** Associate researcher at the College of Intelligence Science and Technology, National University of Defense Technology. He received his Ph.D. degree from National University of Defense Technology in 2019. His research interests include unmanned ground vehicle planning and perception.)



王耀南 中国工程院院士, 湖南大学人工智能与机器人学院教授. 1995 年获得湖南大学博士学位. 主要研究方向为机器人学, 智能控制和图像处理.
E-mail: yaonan@hnu.edu.cn
(**WANG Yao-Nan** Academician at Chinese Academy of Engineering, professor at the School of Artificial Intelligence and Robotics, Hunan University. He received his Ph.D. degree from Hunan University in 1995. His research interests include robotics, intelligent control, and image processing.)



刘 流 国防科技大学智能科学学院副工程师. 2013 年获得北京航空航天大学硕士学位. 主要研究方向为无人车决策控制与测试.
E-mail: liuliu124252@163.com
(**LIU Liu** Associate engineer at the College of Intelligence Science and Technology, National University of Defense Technology. He received his master degree from Beihang University in 2013. His research interests include unmanned ground vehicle decision control and testing.)