文章编号:1005-0523(2026)02-0001-15

无人机具身智能综述

王　森,孙一铭,朱鹏飞

(东南大学自动化学院,江苏 南京 210096)

摘要:近年来,随着无人机技术与具身智能的深度融合,低空无人机凭借其高机动性、多模态感知能力与自主决策能力,在智慧物流、城市巡检、灾害救援等领域展现出巨大潜力。然而,复杂动态的低空环境(如密集障碍物、气象扰动)对无人机的感知精度、实时决策与系统鲁棒性提出了严峻挑战。尽管已有大量研究围绕无人机具身智能展开,但大多聚焦于单一技术模块,缺乏系统性技术链梳理与跨领域协同分析。为此,文章从"感知-决策-控制"一体化系统出发,系统综述了低空无人机具身智能的关键技术体系与最新进展,梳理其在多模态感知、任务推理、物理交互与仿真训练等方面的发展路径,并展望未来的研究趋势。

关键词:无人机;具身智能;多模态感知;自主决策;交互执行

中图分类号:TP 391.4　　**文献标志码:**A

本文引用格式:王森,孙一铭,朱鹏飞. 无人机具身智能综述[J]. 华东交通大学学报,2026,43(2):1-15.

A Survey on Embodied Intelligence for Unmanned Aerial Vehicle

Wang Sen, Sun Yiming, Zhu Pengfei

(School of Automation, Southeast University, Nanjing 210096, China)

Abstract: In recent years, with the deep integration of unmanned aerial vehicle (UAV) technology and embodied intelligence, low-altitude UAV has demonstrated great potential in fields such as smart logistics, urban inspection, and disaster response, owing to their flexibility, multimodal perception, and autonomous decision-making capabilities. However, the complex and dynamic low-altitude environments, characterized by dense obstacles and meteorological disturbances, pose significant challenges to UAV in terms of perception accuracy, real-time decision-making, and system robustness. Although extensive research has been conducted on UAV embodied intelligence, most efforts focus on isolated technical modules, lacking a systematic overview and cross-domain integration. To address this gap, this paper provides a comprehensive review of key technologies and recent advances in low-altitude UAV embodied intelligence from the perspective of the perception-decision-control loop. It summarizes the development of multimodal perception, task reasoning, physical interaction, and simulation-based training, and further discusses promising future research directions.

Key words: unmanned aerial vehicle (UAV); embodied intelligence; multimodal perception; autonomous decision-making; interactive execution

---

收稿日期:2026-01-09

基金项目:国家自然科学基金重点项目(62436002);国家自然科学基金青年科学基金项目C类(62506073);国家重点研发计划雄安新区科技创新专项(2025XAGG0039);中国博士后科学基金国家资助博士后研究人员计划B档(GZB20250395);江苏省卓越博士后计划B档(2025ZB294)

Citation format: WANG S, SUN Y M, ZHU P F. A survey on embodied intelligence for unmanned aerial vehicles[J]. Journal of East China Jiaotong University, 2026, 43(2): 1−15.

近年来,具身智能作为人工智能(AI)发展的重要方向,强调智能体在真实环境中通过感知、认知与动作的闭环过程实现自主学习与适应。具身智能自主无人系统通过物理形态与智能算法的深度耦合,能够在非结构化环境中实现更鲁棒的主动感知与交互。相较于传统虚拟智能体,具身智能体能够利用物理存在,与动态环境进行复杂交互,从而展现出更强的环境适应性与任务泛化能力[1-2]。在这一背景下,无人机(unmanned aerial vehicle,UAV)作为一种特殊的具身智能体,凭借其三维空间高机动性、远距离感知与低成本部署等优势,成为空中具身智能研究的重要载体。这些能力使无人机在复杂动态环境中具备良好的自主操作潜力,并在多种领域展现出应用优势。

具身智能无人机与多机协同调度、智能任务分配系统结合后,已实现复杂场景下的多目标动态响应能力,形成规模化自动作业体系。针对单机感知能力的局限性,於志文等[3]系统综述了无人机集群协同感知计算架构,指出通过跨平台信息融合可显著降低环境不确定性,为复杂任务决策提供更完备的状态空间。Dong等[4]提出的热成像与YOLOv3融合的幸存者检测系统,为无人机执行灾后搜救提供了技术验证。Yan等[5]针对城市复杂环境下的目标逃逸与遮挡问题,提出了一种基于意图推理与深度强化学习的长期跟踪框架。

1 技术全景与框架创新

1.1 无人机具身智能技术链

在已有研究中,无人机具身智能的关键技术体系大致可以分为以下4个核心模块。

1) 多模态感知建模。无人机通过搭载RGB相机、激光雷达、惯性测量单元(IMU)、热成像与多光谱传感器等设备,构建对复杂环境的实时认知能力。近年来,随着深度学习的引入,目标检测、语义分割与三维重建等感知任务的精度和速度均显著提升。

2) 语言引导的认知推理能力。以视觉语言导航与视觉问答为代表的研究,推动了无人机在执行高层语义任务方面的能力。Liu等[6]针对长时序任务中的时空上下文缺失问题,提出了LongFly框架,通过历史感知时空建模策略将碎片化的历史观测转化为结构化的紧凑表示,显著提升了无人机在复杂动态环境下的指令跟随精度与导航稳定性。

3) 高维决策与动作生成。无人机需在连续动作空间中完成路径规划、轨迹生成与避障控制。强化学习、模仿学习与基于大模型的策略生成方法正成为主流途径。此外,模型预测控制与鲁棒控制技术在处理风扰与多目标协同中的应用也日益广泛。

4) 仿真训练与部署迁移。高保真仿真平台(如AirSim)提供了多模态感知模拟与动态交互环境,使无人机能在安全环境下进行大规模训练,并通过仿真到现实(Sim2Real)策略逐步迁移至真实场景[7]。

1.2 系统框架演进与研究模式革新

从系统层面来看,无人机具身智能系统正从早期的模块化设计范式,逐步向端到端统一建模与多模态交互增强框架演进。以AlMahamid等[8]提出的VizNav框架为代表,研究者通过将感知模块与导航策略分离,实现了在三维动态环境中的高效自主飞行。

然而,在复杂动态环境中,这种分离式设计往往存在信息丢失、响应滞后与任务泛化能力不足等问题。近年来,端到端架构成为新趋势,通过统一模型将图像、语言、状态等多源输入直接映射为动作决策,减少人工设定与模块耦合。与此同时,多模态与多智能体协同成为研究热点。具身智能体不再被视为孤立个体,而是与人类(自然语言交互)、环境(语义图谱理解)、群体(多无人机编队)等形成动态耦合关系。这些研究表明,在高层任务规划中引入大语言模型,有助于统一任务目标与环境感知之间的语义桥接。王文晟等[9]进一步将基于大模型的具身智能系统划分为观察、想象、规划等核心模块,并指出大模型赋予了无人系统在开放世界中极强的零样本泛化能力,推动了从“指令跟随”向

“自主推理”的范式转变。

2　无人机具身感知

具身感知作为具身智能的第一环节，是无人机自主理解环境、制定任务策略与执行决策的前提。在动态、非结构化的低空场景中，无人机需通过多模态传感器主动采集环境信息，并在计算资源有限的条件下高效处理视觉、几何、语义等多维数据，从而形成稳定、连续的状态感知能力。相较于传统的被动感知机制，具身智能更强调“感知即交互”，即感知过程与任务目标密切耦合，体现出对环境理解与操作意图的同步推进。

2.1　无人机主动视觉感知

2.1.1　定义

主动视觉感知是具身智能体系中最为核心的一部分，强调感知行为和动作目标之间的闭环耦合。与传统的被动视觉采集和处理不同，主动视觉感知采用基于任务驱动的视角采样和注意力调度，完成感知资源的合理配置，提高信息的获取效率以及交互的适应性[10]。

2.1.2　方法介绍

1）目标检测。在无人机具身智能系统中，目标检测技术必须同时满足实时性、精确性与复杂环境适应能力的要求。为应对资源受限平台的挑战，研究者提出了多种轻量化模型设计策略。例如，Cheng等[11]通过引入轻量级特征提取模块、注意力融合机制和改进的损失函数，在保持高精度的同时显著降低了模型复杂度，并成功部署于Jetson TX2等嵌入式设备中，具备应用潜力。Han等[12]提出的LUFFD-YOLO专为森林火灾等空中监测任务设计，结合GhostNetV2和改进的注意力机制，在提升检测精度的同时大幅压缩参数量，较好地兼顾了速度与精度，适用于嵌入式平台上的高效部署。

为提升在复杂环境下的目标检测能力，部分方法侧重于环境鲁棒性增强。Qing等[13]在主干网络模型中引入结构重参数化机制与角度分类分支，显著提升了模型在航拍图像中对任意角度目标的检测性能，使其适用于旋转目标密集区域的监控任务。Fang等[14]进一步整合了图像去雾与目标检测的多任务学习框架，在保留语义特征的同时有效去除视觉噪声。

近年来，大模型与多模态技术的兴起推动了语言引导检测范式的形成。Liao等[15]结合了Segment Anything模型（SAM）的提示分割机制与YOLO检测框架，实现了语义引导下的交互式检测，特别适用于任务初始化阶段的目标指定与人机协同交互。Guo等[16]基于Transformer检测架构，引入多尺度注意力聚合模块与场景感知机制，实现了对潜在威胁活动的精准识别与快速响应，在城市治理、边境监控与人群行为分析等场景中展现出良好性能。

2）语义分割。无人机具身智能场景下的语义分割必须在有限算力、复杂光照与持续飞行等多重约束下，稳定提供像素级环境语义信息，以支撑避障、路径规划和操作决策。近年来的主流研究大体沿着三条技术脉络演进：一是聚焦推理速度的轻量化网络；二是借助Transformer捕获全局上下文；三是利用多模态信息克服逆境感知难题。

由于低空无人机计算资源与电源受限，语义分割模型需在严格的延迟预算内完成前向推理。Zhang等[17]提出的KDP-Net结合了可分组动态卷积、边缘感知机制与语义增强分支，增强了对不规则地形与植被的适应能力，尤其适用于灾后环境中无人机快速识别可降落区域。Yu等[18]通过细节分支与语义分支解耦建模，并融合引导聚合模块，在保持图像细节的同时兼顾语境理解，特别适用于道路标线与小目标（如行人、路灯）共存的城市航拍场景。Huang等[19]提出的Mamba-UAV-SegNet将状态空间模型Mamba嵌入多尺度特征融合通道，具备长距离建模能力，能够更清晰地区分农田、灌溉沟渠等结构相似区域，在农业无人机巡检中展现出良好性能。

在算力稍宽裕、需要全局上下文的任务中，层级视觉Transformer显示出强大的表示能力。Yi等[20]提出的UAVformer通过将飞行状态信息（如高度、航向、俯仰）编码为空间先验嵌入视觉token中，实现了对多飞行角度图像的统一理解，有效提升了高低空混合航线下的识别一致性。

为解决单RGB相机在云雾、逆光条件下信息缺失的问题，学界正将深度、LiDAR、热红外等互补信息引入分割框架。Yin等[21]利用边缘引导机制对LiDAR点云与可见光影像进行对齐优化，提升了模型在山地林区等高差较大区域的地形识别精度，适合滑坡预警与林地监管。Qu等[22]针对夜航任务提出低照度增强与语义对齐机制，结合特定夜间数据集的设计，显著提升了低光环境下的分割稳定性，适

合夜间巡逻与应急场景的快速部署。

综上所述，面向无人机具身智能的语义分割研究正持续向轻量化部署、全局建模、多模态融合与时序一致性协同发展，不断提升系统在复杂环境下的感知稳健性与任务泛化能力。

3）深度估计。与检测、分割相比，深度估计的技术路径更多取决于成像方式与监督策略。目前常见的4条路径可归纳为：① 单目轻量估计；② 立体/多视角重构；③ 多模态补偿退化；④ 视频自监督与尺度一致性。

Florea等[23]提出的TanDepth通过将全球数字高程模型（DEM）投影到相机视域实现“推理阶段尺度恢复”，使同一网络可迁移到不同飞行高度且保持米级精度。Cheng等[24]提出的TinyDepth采用一种层级式Transformer编码器，结合卷积残差分支，实现了自监督单目深度估计的轻量化，特别适合续航受限的微型无人机实时导航。

Dhafani等[25]提出的FIReStereo-Net针对双红外成像设计了温度归一化层，使其在夜航林区场景中仍可恢复树干与地面层级纹理。Madhuanand等[26]提出的Self-Sup Oblique-UAV利用连续帧重投影与对比损失，在UAVid 4K视频上实现了纯视觉自监督深度估计，解决了倾斜摄影缺乏真值的问题。Yu等[27]提出的Scene-Aware Refine-Net通过光照一致性和动态物体掩码来细化无监督误差，提升了复杂街区航拍的几何连贯性。

无人机深度估计的最新研究正沿着单目轻量化、多视立体重构、多模态补偿、自监督4条路径并进，旨在将多模态感知和自监督策略统一到端到端架构中，在降低标注与算力成本的同时，提升全场景、全天候的三维感知稳健性。

2.1.3 方法总结与局限性分析

如表1所示，现有研究主要在“精度-速度”权衡与“环境适应性”两个维度上寻求突破。轻量化网络（如YOLOv5s-ngn）解决了资源受限下的实时性问题，但在夜间或强遮挡场景下泛化能力不足；而引入大模型（如SAM系列）虽然提升了语义理解能力，却难以在无人机机载端实现实时推理。当前的局限性在于缺乏一种能够同时兼顾低功耗部署与开放世界语义理解的统一架构。未来的突破口在于发展“云-边协同感知”架构，或探索基于脉冲神经网络（SNN）的神经形态视觉感知，以极低功耗实现高动态范围的视觉处理。

2.2 无人机三维视觉感知与定位

2.2.1 定义

无人机三维视觉感知与定位在本综述中指同时定位与建图（simultaneous localization and mapping，SLAM）在低空自主飞行平台上的应用：利用机载视觉、惯性、激光等多模态传感器数据流，在未知或仅具粗略先验的环境中实时估计自身六自由度位姿，并同步生成尺度一致的三维地图，该过程构建了“度量-语义”一体的时空坐标系。

现有UAV-SLAM研究大体可归纳为三条技术谱系。几何驱动SLAM以传感器几何约束为核心，通过特征匹配或点云配准完成高精度“位姿-地图”估计，代表工作包括面向RGB-D的ORB-SLAM2及其“视觉-惯性”扩展等。语义增强SLAM在几何框架中显式识别和剔除动态目标，提升城市、室内等高动态场景下的鲁棒性。随着大规模“视觉-语言”预训练模型的兴起，“视觉-语言”模型（vision-language model，VLM）辅助定位将跨模态对齐能力引入SLAM闭环，通过对齐图像与自然语言指令中的多粒度语义，实现“用语言找位置”的高层语义重定位，为人机协作与任务指令导航提供了新范式。

受空域六自由度高速机动、俯视/斜视视角下尺

表1 典型无人机具身视觉感知方法对比

Tab.1 Comparison of typical embodied visual perception methods for UAV

方法类型	核心机制	优势	局限性	典型应用场景
轻量化检测[11-12]	注意力机制剪枝、GhostNet轻量化	推理速度快，适合部署于边缘设备（如Jetson TX2）	在极端光照条件或目标尺度极小时，检测精度易下降	实时避障、林火监测
鲁棒性增强[13,17]	边缘高斯驱动、去雾与多任务学习	对夜间、雾霾等退化环境具有较强鲁棒性	计算开销较大，难以维持高帧率运行	灾害救援、全天候巡逻
语义交互引导[15]	提示（Prompt）驱动的分割、语言对齐	支持零样本交互，具备开放世界理解能力	依赖大算力，端侧部署时延较高	人机交互、复杂指令执行

度剧变、风扰振动及机载算力约束等因素影响，UAV-SLAM仍面临实时性与精度的双重挑战。几何、语义与语言三类技术的融合正在成为趋势：在轻量化图优化框架内，同时利用点云、深度语义与语言提示进行约束补充，有望在复杂低空空域持续输出稳健、可解释的“地图-位姿”流，进一步夯实无人机具身智能闭环的感知基础。

2.2.2　方法介绍

1）几何驱动SLAM。几何驱动SLAM完全依赖传感器间的几何一致性（像素视差、点云配准或惯性积分）来推断位姿与地图，不借助外部语义先验，因此在机载算力有限、通信受限的低空无人机平台中仍是主流方案。“视觉-惯性”耦合与“激光-视觉”互补已成为提升空域SLAM稳健性的两条核心路线，研究重点正从纯视觉里程计向多传感器闭环优化扩展。

在视觉分支中，特征基方法以ORB-SLAM系列为代表：Mur-Artal等[28]提出的ORB-SLAM2首次统一单目、双目与RGB-D输入并在多场景中达到厘米级精度。Qin等[29]提出的VINS-Mono将IMU预积分嵌入滑动窗口光束法平差框架，使单目系统在剧烈俯仰与快速侧移时依然保持尺度可观测性，成为多旋翼自主飞行的标配定位前端。

激光/多传感器分支以LOAM的实时点云“边-面”特征配准为开端，验证了轻量化激光里程计在无人机六自由度运动中的厘米级稳定性。为适应高速飞行，Xu等[30]的FAST-LIO采用紧耦合的迭代扩展卡尔曼滤波器（IEKF）替代传统的批量优化方法，实现了高频率、低延迟的LiDAR-IMU状态估计，适用于实时性要求高的应用场景。

整体来看，几何驱动SLAM正沿着“特征-半直接-光度稀疏”多元视觉里程计与“边-面-强度-多模态”激光前端两条路线并行演进，并通过因子图将IMU、视觉、LiDAR等异质观测统一到同一优化框架中；在低空无人机平台，这些方法已能够在高速机动、尺度剧变与光照/纹理退化等极端条件下稳定输出厘米级轨迹与稠密地图，为后续语义增强与语言提示定位提供了坚实的几何基础。

2）语义增强SLAM。针对低空自主飞行中全球导航卫星系统（GNSS）难以长期可用、动态目标密集且纹理退化明显的现实约束，UAV领域正将传统几何SLAM升级为可同时输出“度量-语义”一体化地图的框架。近年来，多项工作直接在无人机平台上验证了语义信息带来的定位增益：Liu等[31]的RT-Semantic SLAM系统在针叶林下方实现了1 km级自主巡航，利用LiDAR中树干/地面的语义约束抑制累积漂移，并维持厘米级建图精度；Fanta-Jende等[32]提出的Semantic RT-Mapping-UAV则将轻量级深度卷积神经网络（CNN）分割结果实时投射到稠密体素，在无人机板载Jetson平台上实现了10 Hz全帧语义重建，面向灾害监测场景验证了其遮挡补全能力。

除了针对特定场景的系统化方案，研究者亦在“着陆、安全巡检与长航测绘”等关键任务上探索语义先验的引入：Yang等[33]的Semantic SLAM-Landing在未知室内依靠CNN提取“平整-无遮挡”着陆区语义，叠加ORB-SLAM2构建的稀疏地图，实现了仅依赖机载RGB-D的安全降落。综合来看，UAV-SLAM正沿着语义滤动、实例补全、多模态互补三条主线推进：在动态、弱纹理乃至夜间场景下，语义先验已被证明能够成倍削弱漂移并显著提升地图可解释性，为后续语言引导的高层具身决策奠定坚实基础。

3）VLM辅助定位。基于大规模VLM的辅助定位，将“几何-语义”SLAM进一步升级为“几何-语义-语言”三重对齐范式：系统首先在传统SLAM框架内完成稠密或稀疏三维重建，然后将CLIP、X-VLM等预训练模型生成的多粒度“文本-图像”嵌入回投到网格/点云上，使每一个体素或网格面都带有可检索的语言特征。VRLM结合数字孪生构造的大尺度VDUAV数据集，在城市、丘陵等多场景空域实现了20 m以内的定位精度，并通过多尺度特征融合保证板载Jetson实时运行[34]。CCA-UAV则利用跨视图一致注意力为“无人机-卫星”配准引入Transformer特征，对抗大俯仰角差异，刷新了University-1652数据集的Top-1记录[35]。

整体来看，VLM辅助定位专注于“用语言找位置”而非直接产生控制命令：几何SLAM确保尺度一致，语义分割缓解动态遮挡，VLM嵌入则赋予地图开放词汇检索能力。三者耦合后，无人机在GNSS受限、弱纹理或任务多变的空域可实现零样本重定位与地标标注，为后续“视觉-语言”导航提供可靠、清晰且互补的先验定位支撑。

2.3 无人机触觉感知

2.3.1 定义与任务分类

无人机触觉感知是指在飞行平台上集成“力觉、振动觉或视觉”传感器，使无人机能够在空中或贴壁操作时直接量化接触力分布、材质特征与外界扰动力，并将这些物理信号反馈至飞控或末端执行器。与传统非接触的“视觉-激光”感知相比，触觉信息提供了高带宽、抗光照的局部交互感知，可在复杂环境中弥补视觉遮挡与纹理缺失问题。近期研究已展示了将软质光学触觉皮肤安装于四旋翼或全向推力平台，从而在动态壁面接触中实时估算三维力向量并感知表面纹理的可行性[34]。

面向任务，本综述将无人机触觉感知划分为两大子方向：表面识别（surface recognition）与外力估计（exogenous force estimation）。

表面识别关注通过“触觉-振动”模式区分墙体材质、结构粗糙度及摩擦系数，以指导抓取或贴壁巡检策略；外力估计则侧重于感知风力脉动、碰撞冲击或承载力变化等环境扰动，为姿态稳定控制与安全回路提供实时观测。基于扰动观测器（DOB）的风场估计算法已在轻型多旋翼上实现3 m/s级阵风识别，并可在无额外风速计的情况下有效降低横滚向漂移；更早的机体系统辨识工作亦表明，通过机载“惯性-气压”数据融合可重建局部风梯度，为路径规划器提供风场先验。这两类任务共同构成无人机触觉感知的核心应用：前者直接服务于抓取、挂载与贴壁检测等“主动交互”场景，后者则支撑飞行稳定性与能耗优化等“被动适应”需求。

2.3.2 方法介绍

1）表面识别。近期研究聚焦于在机体表面安装“轻量化-高分辨率”的视觉型触觉皮肤，将触觉图像映射为材质类别或局部形貌特征。Aucone等[36]设计的大曲面光学触觉传感器将环形LED与微型高速相机封装于柔性硅胶壳内，利用多任务网络同时输出像素级深度形变与三维接触力，材质分类Top-1准确率达93%，并在阵风5 m/s情况下维持法向位置均方根偏差<2 cm。此外，Vigara-Puche等[37]的障碍物接触传感器将六轴力力矩传感器与弹性防撞外壳结合，通过神经网络将柔性位移映射为接触方向与粗糙度标签，实现飞行中“碰-即-分”式材质甄别和避障触发。

2）外力估计。外力估计方法集中于从机载惯性、姿态与气动数据中在线重建扰动力，以提升姿态稳定性和能耗效率。Hattenberger等[38]利用机体动力学模型与标准导航传感器提出风估测观测器，能在无风速计的情况下将平均风速估计误差控制在0.5 m/s以内，并为路径规划提供风场先验。Asignacion等[39]通过Takagi-Sugeno模糊非线性干扰观测器（NDOB）在100 Hz速率下分离阵风脉动；实机试验在8 m/s突风场景中将姿态误差降低约40%。这些方法共同展示了从低成本机载数据中推断外部扰动的可行路径，为高风速、强扰动环境下的无人机稳定控制与触觉交互奠定了动力学基础。

2.4 “视觉-语言”导航

2.4.1 定义

无人机“视觉-语言”导航（vision-language navigation，VLN）指以旋翼或固定翼无人机为具身智能体，接收自然语言指令（如“沿河向西飞到红色砖塔上方30 m处”），仅依赖机载视觉与惯性等原生传感器，在未知或仅部分可见的三维空域中自主规划安全航迹并精确抵达目标。该任务源自地面机器人场景中VLN原型R2R（Room-to-Room）数据集与Matterport3D模拟器，但在扩展至空域后呈现出全新的技术难点。首先，俯视/斜视视角导致目标呈现强透视与尺度变形，视觉特征与语言描述的对齐需跨高度、跨视点完成；其次，无人机需遵循六自由度动力学和能耗限制，并实时应对风扰、禁飞区等安全约束；最后，室外场景往往跨越“城市-郊野”尺度，导航系统须在稀疏语言线索与稠密视觉观测之间动态维护多尺度语义拓扑图。

现有研究将UAV-VLN细分为室内低空（仓储、厂房）与室外城市/郊野两类任务。评价体系沿用地面VLN的成功率（success rate，SR）、路径效率（success weighted by path length，SPL）与归一化DTW（normalized dynamic time warping，nDTW），并新增飞行能耗效率与安全违规率，以度量三维航迹的经济性与风险。2023年AerialVLN首次发布包含25座近真实城市场景与8×10^3条人类演示“指令-轨迹”的任务框架，奠定了空域VLN的方法与指标基线[6]；2024年的CityNav将规模扩展到32×10^3条指令并引入真实地理信息系统（GIS）地标，系统化揭示了跨高度、跨城市导航的难度[40]。与此同时，方法范式正从传统跨模态注意力模型迈向大型多模

态模型（LMM）驱动，2025年的OpenFly-Agent通过关键帧筛选与VLN对齐显著提升了长程规划的效率与安全性[41]。

2.4.2　方法介绍

无人机VLN方法大体沿两条技术谱系演进：一条继承室内VLN的“显式语义地图+跨模态对齐”范式，另一条借助大规模预训练模型在语言推理与视觉理解上的零样本能力，形成“多粒度表示+LMM规划”新框架。

AerialVLN率先将跨模态注意力（CMA）迁移至25座城市场景，引入高度编码解决俯视透视畸变，成为后续工作的默认基线[6]。

LMM驱动范式借助大型多模态模型的开放词汇推理与上下文记忆能力，实现了指令理解、动态规划与安全控制的端到端一体化。OpenFly-Agent通过“关键帧感知+视觉token合并”压缩序列冗余，在10×10^4条轨迹基准上将SPL提升至34%，并在Jetson NX实机上达到15 Hz推理频率[41]。CityNavAgent冻结CLIP多尺度特征，仅微调LoRA适配器，就在“跨城市-跨高度”场景中的零样本SR提高到38%[40]。

总体来看，传统路线通过改进视图离散化、拓扑度量融合和交互图注意力，不断提高跨视角语义对齐与长程规划性能；LMM路线则依托大模型的语言理解、世界知识与序列推理优势，以多粒度token表示和安全控制层相结合，展现出强大的零样本泛化与人机交互潜力。国内与国际团队交替领跑，使无人机VLN成为具身智能领域最活跃、方法最丰富的研究前沿之一。

2.4.3　现有基准的局限与趋势

纵观现有的无人机VLN导航数据集（表2），虽然场景规模已从单一街区扩展至城市级区域，但Sim2Real的差距依然巨大。现有数据集多依赖AirSim等仿真器生成，缺乏真实飞行中存在的气动扰动、信号丢失及动态光照变化等数据。未来的研究亟需构建基于实测数据的真实世界空中导航基准，并引入LMM作为评估器，以解决长程导航任务中因奖励稀疏导致的训练效率低下问题。

3　无人机具身决策

3.1　动态路径规划与避障

3.1.1　定义

动态路径规划与避障是指无人机在执行任务过程中，依托在线环境感知（机载LiDAR、视觉重建的障碍物几何、实时风速与风向估计、禁飞区边界等）与机载/边缘算力，通过滚动重规划或强化学习策略等方式持续更新飞行轨迹，从而在安全约束（最小安全距离、可行姿态包络、能耗上限）与任务效率指标（航时、路径长度、目标时限）之间实现最优权衡。该问题通常可拆分为三个耦合子任务：① 局部地图构建，将传感器数据流抽象为可实时查询的欧氏距离场/占据网格；② 可行轨迹采样与评估，依据动力学模型及“风-载荷”扰动力，对候选轨迹进行碰撞检测、能耗与时延评分；③ 自适应重规划，当发现新障碍或风场发生剧变时，通过强化学习策略或改进采样算法在10~100 Hz频率范围内快速生成替代航迹。动态路径规划与避障是具身决策层连接低层控制与高层任务调度的关键环节，其核心研究难点在于时变环境的不确定性量化以及高维非线性动力学下的实时最优解搜索。

3.1.2　方法介绍

针对无人机在动态环境中的路径规划与避障问题，研究者提出了多种方法体系，大体可分为基于规划算法的经典方法与基于深度学习的智能方法两大类，同时在多机协同场景下还衍生出分布式路径规划技术。经典方法包括图搜索、采样基和优化等方法，这些算法最初多借鉴于地面机器人路径规划，但随着无人机三维飞行和动态环境需求的增加而不断演进改进，以增强对复杂动态环境的适应

表2　主流VLN导航数据集与基准对比

Tab.2　Comparison of mainstream VLN datasets and benchmarks

数据集	场景类型	规模/指令数	核心特点	关键挑战
AerialVLN[6]	仿真城市	25个场景/8×10^3条指令	首个空中视觉与语言导航基准，引入高度维度	俯视视角的透视畸变与跨模态对齐
CityNav[40]	真实GIS重建	32×10^3条指令	整合GIS数据，支持大尺度城市导航	跨区域长程规划与语言地标匹配
OpenFly[41]	混合场景	大规模基准	包含长时序任务与安全约束	“视觉-语言-行动”的长距离时序关联

性。近年来,强化学习等学习方法开始应用于无人机自主导航,能够通过交互体验来学习避障策略。此外,对于多智能体无人机编队,需要考虑多机间的避撞协调和通信约束,发展出专门的协同规划与避障方法。下面分别介绍各类方法及其演进脉络。

1)几何优化范式。几何优化范式依旧是高动态飞行的安全基线:为解决未知环境中的高速安全飞行问题,Tordesillas等[42]设计的FASTER规划器将"主轨迹+安全备份轨迹"双通道思想融入混合整数二次规划,在实机测试中于未知林地实现了7.8 m/s的无碰撞飞行,成为近年高动态轨迹规划的经典基线。Ye等[43]提出的TGK-Planner通过拓扑图引导采样和免欧氏距离场(ESDF-free)梯度下降,将局部规划时延压缩至毫秒级且无需重建稠密地图。这些工作在分层搜索、风险显式建模与走廊约束方面持续逼近几何算法的极限,为后续学习方法提供了可解释、安全的参照路径。

2)深度学习范式。深度学习范式则凭借端到端策略,显著提升了对未知环境和强扰动的适应性。Song等[44]结合强化学习(reinforcement learning,RL)与模仿学习,将12 m/s林区穿越成功率提升了25个百分点,并在Agilicious开源平台公开了硬件/软件栈,支持5 kg载荷、70 km/h的视觉自主飞行。神经形态路线将推理延迟降至毫秒级,Paredes-Vallés等[45]发布的全神经形态视觉控制系统,依靠事件相机与Loihi芯片实现了2.1 ms的闭环控制,并在40 km/h穿门实验中保持零失误。深度学习的"感知-决策"一体化和硬件友好推理正将无人机自主导航推向"高鲁棒-低功耗"的新阶段。

高速安全飞行仍依赖几何优化方法的安全可行性分析,深度学习带来对未知扰动与感知噪声的强鲁棒策略,而分布式协同则解决多机规模化部署与通信约束难题;三条技术路线相辅相成,共同推动无人机具身决策迈向"实时-安全-可解释-可迁移"的下一代自主飞行框架。

3.1.3 总结与突破方向

传统的几何优化方法在已知环境中具备极高的安全性与可解释性,但在面对非结构化未知环境时,往往受限于地图构建的延迟;相比之下,基于学习的方法(如RL、深度强化学习DRL)展现了卓越的动态适应性,但缺乏明确的安全保障机制。当前研究的痛点在于UAV敏捷性与安全性的二元对立。未来的突破口在于"神经-符号"混合架构(Neuro-Symbolic AI),即利用深度学习处理感知与策略生成,同时利用控制理论工具(如控制障碍函数,CBF)作为安全过滤器,实现既敏捷又安全的具身决策。

3.2 多机协同任务分配

3.2.1 定义

多机协同任务分配(multi-UAV task allocation,MUTA)是指在共享空域与有限通信带宽下,针对任务集合T和无人机集合U,综合考虑飞行器的动力学可行性、载荷/能量余量以及航迹安全约束,对"谁、何时、以何顺序"执行子任务做出联合决策的组合优化问题。其核心目标是在保证时空冲突消解与资源平衡的同时,最小化全局完工时间或能耗,或最大化任务收益与系统鲁棒性,因此常被表述为多目标非确定性多项式困难(NP-hard)优化问题,并普遍采用分层规划、市场博弈或强化学习等近似求解策略。该问题区别于单机航迹规划:它不仅要求对单架飞行器规划可行路径,还需在任务与资源层面对跨平台协同、通信异步和动态任务或UAV失效等不确定性作出实时调度。

3.2.2 方法介绍

多无人机系统在任务分配与资源调度方面已取得一定研究成果。当前的研究主要集中在三大方向:基于传统优化算法的改进、基于分布式的协作框架、基于AI的泛化方式。

表3 无人机动态路径规划主流范式特性对比

Tab.3 Comparison of main characteristics of mainstream approaches for UAV dynamic path planning

技术范式	实时性	安全性/可解释性	未知环境适应性	典型局限
几何优化法[42-43]	高(毫秒级延迟)	高(具备理论安全边界)	中(依赖先验地图质量)	在高动态障碍物密集区域易陷入局部最优
深度强化学习[44]	极高(端到端)	低(黑盒模型)	高(具备探索策略)	训练收敛困难,Sim2Real迁移存在安全性风险
神经形态计算[45]	极高(微秒级延迟)	中	中	需专用硬件(如Loihi),算法生态系统尚不成熟

1）基于传统优化算法的改进。传统的优化方法基于建模进行任务分配，采用启发式或进化算法进行求解，以提高任务完成率及系统资源利用率等。Alqefari等[46]提出的Hybrid Auction采用分层拍卖机制与优化后的粒子群策略来支持任务分配过程中的快速调整，并能适应异构无人机在能力和状态上的差异，使整个系统具有良好的鲁棒性和灵活性。Han等[47]提出的FESGA引入了模糊逻辑和精英策略，在多目标环境中根据解的优势以及群体的多样性度量进行权衡，增强了全局搜索能力和收敛速度。Chen等[48]使用蚁群优化机制，模拟信息素更新过程来引导任务选择路径，不仅能够考虑异构平台及多任务限制，还能对任务执行过程中环境发生的变化做出响应。

2）基于分布式的协作框架。通过应用去中心化的分配方式以加强系统的鲁棒性和适应性，是分布式协作框架的一个核心思想。Xu等[49]提出的RL-DTA结合Q-learning与局部搜索策略，使得无人机能够在实时环境中根据历史反馈动态调整其分配策略，有效提升了系统应对突发任务与局部故障的能力。Wang等[50]提出的TS-DTA采用任务先分配后优化的两阶段机制，能够在考虑不同无人机性能差异以及时间、空间约束条件的同时，降低对中心调度环节的依赖，从而提升全局资源利用率。

3）基于AI的泛化方式。深度强化学习和大型语言模型为多无人机系统的任务分配提供了新的解决方案，增强了系统的泛化与自适应能力。Yin等[51]提出的Deep Transfer RL先在模拟器中进行预训练，再迁移到真实世界中进行微调以得到最终策略，以此解决少样本问题，从而提高多场景下任务分配的通用性。Fan等[52]提出的SA-NNO-DRL将多无人机侦察任务规划问题分解为目标分配与路径规划两部分，采用模拟退火与基于最近邻最优的深度强化学习相结合的策略，提高了规划效率和泛化性。

3.3　“视觉-语言-行动”模型

3.3.1　定义

“视觉-语言-行动”（vision-language-action，VLA）模型是指将自然语言指令与环境视觉信息联合建模，形成从自然语言到无人机行动序列/控制命令端到端映射的多模态智能系统。与以往的分阶段架构（先进行语言理解，再进行路径规划，最后执行动作）不同，VLA模型采用单一网络结构，利用语言语义引导视觉感知，并同步生成时序动作计划。这类模型广泛应用于无人机的自然语言任务执行，如灾害搜救、航线巡查与目标交互等场景，能够根据语言指令自主完成复杂的操控行为。典型模型如VIMA和RT-2已在地面平台上取得了较好的效果，最近的研究也开始将此范式迁移至无人机平台，使空中智能体获得对高层语义任务的感知与响应一体化能力。

3.3.2　方法介绍

近年来，随着多模态学习与具身智能的融合发展，VLA模型在机器人及无人机领域均取得了重要进展。在机器人研究中，Google DeepMind提出的RT-2模型通过在大规模“视觉-语言”数据上进行多模态预训练，将从互联网学习到的视觉与语言知识直接迁移至机器人控制任务中，实现了从语言指令到动作决策的端到端映射，并展现出良好的跨平台泛化能力[53]。

在此基础上，VLA模型理念被逐步引入具身智能无人机系统，以增强其复杂任务执行能力与多模态理解水平。UAV-VLA系统融合卫星图像、VLM与大型语言模型（如GPT），实现了从自然语言描述到飞行路径与动作计划的自动生成，在任务规划效率上显著优于人类操作员。为增强推理能力与实时响应，CognitiveDrone构建了包含人类识别与符号任务的模拟飞行数据集，基于第一人称视觉与语言输入生成四维动作命令，其改进版本Cognitive-Drone-R1在认知任务中的成功率达到77.2%[54]。VLA模型正逐步形成统一的VLA融合框架，并在机器人及无人机平台上实现从感知理解到任务执行的闭环优化，为具身智能体系统在多任务环境中的自主决策提供了有效支撑。

3.4　“感知-决策-控制”一体化

3.4.1　定义

“感知-决策-控制”（perception-decision-control，PDC）一体化系统是具身智能体系中最为核心的闭环结构之一，强调无人机在复杂动态环境中以多模态感知驱动策略决策，并将其高效映射为物理控制行为，从而实现自主、稳定、连续的任务执行过程。

与传统的分层式任务管线不同，PDC一体化系统更加注重“信息流-决策链-动作反馈”三者之间的耦合与协同，力求在任务响应速度、轨迹精度与系统稳健性之间实现最优平衡。

在典型PDC一体化系统中，无人机首先通过搭载的多模态传感器（如RGB相机、IMU、事件相机、LiDAR等）构建对环境的结构性认知，包括自由空间构型、可通行区域、目标状态及动态障碍等关键要素；随后，系统在此基础上进行任务状态解析与高维轨迹或控制策略生成，如时间最优路径、避障重规划或动作原语序列等；最终，生成的控制指令将通过飞控系统或低层控制器精确执行，从而完成飞行、避障、操作等具体动作任务，并借助执行反馈更新下一轮感知与策略，形成闭环。与语言引导的VLA模型相比，PDC一体化系统虽然不包含语言理解模块，但其在底层架构、任务响应与飞控适配方面具备极高的通用性，常作为构建高层语义系统（如VLA、具身问答系统等）的基础框架或部署骨架。

3.4.2 方法介绍

近年来，多个研究团队围绕PDC一体化系统展开深入探索，致力于提升无人机在动态、复杂环境中的自主操作能力。浙江大学FAST-Lab提出了面向特技飞行的闭环控制框架，融合视觉、IMU与事件流等多模态输入，借助强化学习策略网络实现了高动态动作的自主生成与高精度控制，飞行成功率达100%[55]。为提升系统的部署效率，该团队进一步构建了RAPTO区域膨胀系统，在障碍空间构建方面展现出极高的实时性与资源友好性[56]。

同时，其他研究者也在不同方向上提出了一系列关键进展。香港科技大学沈劭劼团队提出D^2 SLAM架构，聚焦分布式“视觉-惯性-语义”地图构建任务，通过联邦建图机制实现多无人机间状态同步与任务分配优化，显著增强了城市环境下的协同导航能力[57]。此外，该团队还提出了RACER系统，结合探索边界构图与协同控制策略，实现了多无人机在未知环境中的快速部署与分区导航[58]。

上述研究表明，PDC一体化系统已从单机任务逐步扩展至多智能体协作、语义地图构建与极端环境下的高鲁棒控制等方向，形成了“多模态感知-任务推理-低层控制”全链路优化的新范式，为具身智能系统的集成奠定了坚实基础。

4 物理操作

4.1 定义与任务分类

具身物理操作是具身智能研究的一大核心问题，指智能体利用其身体部件（如机械臂、末端执行器、推进装置等）与外部世界发生物理接触，在“感知-决策-反馈”的闭环机制下主动干预外部环境，以完成目标任务。这就要求智能体具备对环境状态的感知能力、对物理规律的认知以及对物理交互影响的动态响应能力，体现智能体“行动即认知”的特性。相比于传统的基于视觉的非接触感知任务，具身操作则是让动作本身进入认知系统，使得智能体能够通过实际的物理交互获取信息，在交互的过程中实现对复杂环境的适应与干预。

具身操作任务的形式多种多样，涵盖从大尺度、低精度操作（如推动、搬运）到小尺度、高精度操作（如钻孔、螺丝固定），可以归纳为以下几类：一是推动类操作，需要实现目标物沿指定方向运动，在实现大行程直线运动的基础上，还需要具备连续调整接触力的能力，常见应用有搬运、路径清障等；二是旋转类操作，如工业现场中对阀门、旋钮等零件的扭动，需要控制系统输出足够稳定且能提供角度反馈；三是精密装配类任务，包括钻孔、拧螺丝等工艺操作，在电子制造与维修领域普遍存在，需要良好的姿态保持能力和轨迹跟踪精度。部分具身操作任务强调协同搬运或协作交互能力，如多臂协作托举、人机协作等，这些任务进一步提出了对时序协调与人机共享控制的要求。

4.2 推动操作

4.2.1 定义

通过机械臂控制推动物体至目标位置。该类任务常用于工业生产线、路径清障、设备触发等场景，要求无人系统能够施加稳定连续的接触力，并兼顾飞行平台的姿态控制与力反馈调节。

4.2.2 方法介绍

Jimenez-Cano等[59]研发了一种基于柔性腱驱动的空中机械臂平台，在发生碰撞时可主动吸收冲击能量。该系统结合轻量化结构设计与柔顺控制策略，具备良好的环境适应能力，适合在复杂地形或易碎物体表面执行稳定的推动操作。Lee等[60]提出了一种基于DOB的鲁棒控制方法，用于移动式结构表面的物体推动任务，通过补偿机制使无人机

在模型不确定性大、外部干扰强的复杂环境下完成控制任务。目前，针对无人机推动类具身交互任务已有诸多相关设计尝试，从机架结构到控制策略均有所涉及，并逐步提升了平台的操作稳定性、环境适应性及任务执行精度。

4.3　旋转操控

4.3.1　定义

旋转操控任务是典型的具身物理交互操作，指智能体通过末端执行器对具有转动结构的对象（如阀门、门把手）施加扭矩，实现其在闭环控制系统中的功能启动或空间结构的物理开启。该类操作常出现在工业巡检、设备维护以及能源管控等任务中。在此类任务环境下，无人机通常需要在受限空间内完成精确对位、柔顺接触和受控施力等多阶段动作序列。

4.3.2　方法介绍

在旋转操控任务方面，研究者围绕阀门转动与门体开启提出了多种具备高自由度控制与精细操作能力的空中系统。Martinez等[61]通过引入附加推力矢量单元，显著提升了无人机平台在地面接触状态下施加扭矩的能力，突破了传统UAV的力矩输出极限。与此同时，具备三臂并联结构的空中机器人平台通过冗余自由度与多点支撑机制，使其能够在倾斜或动态表面上稳定完成旋钮类操作[62]。综合来看，以上几种方法均在结构设计、控制策略与环境适应性方面提出了不同方案以实现空中旋转任务。

4.4　钻孔与拧螺丝

4.4.1　定义

钻孔与拧螺丝任务是具身交互中典型的高精度操作行为，主要涉及对目标材料或结构进行局部加工作业，如在预定位置实施孔洞开凿、紧固件旋入或结构拼接等。该类任务广泛应用于工业装配、基础设施维修、航空制造与空间舱维护等场景，通常要求执行器具备较高的姿态稳定性、末端力矩控制能力与路径跟踪精度。在空中平台场景下，需额外解决平台自身姿态扰动对操作过程的影响，确保操作过程中的力/位控制稳定性与工具路径可控性。因此，钻孔与拧螺丝任务是空中具身智能系统最具挑战性的操作类型之一。

4.4.2　方法介绍

近年来，围绕无人机在空中执行钻孔与拧螺丝等高精度作业的研究持续取得进展，多个研究团队提出了创新性的结构设计与控制方法，为复杂环境下的具身操作提供了技术支撑。Ding等[63]提出了一种新型全向推进飞行平台，结合自适应鲁棒控制与选择性阻抗控制策略，实现了对末端操作器的稳定接触力调节，首次在飞行状态下完成多角度钻孔与拧螺丝任务。Park等[64]的ODAR（全向Delta式空中机器人）平台通过对双向桨布置与最小保证推力优化，实现了在任意姿态下输出全向扭矩，为多角度钻孔提供了充足的力矩余量。在毫米级定位、连续150 N推力、全姿态大扭矩输出等方面，空中钻孔与螺钉紧固研究取得了突破，为灾后巡检、高空维护及空间装配等应用奠定了坚实技术基础。

5　结束语

本文从PDC一体化系统出发，系统回顾并梳理了近年来低空无人机具身智能领域的关键技术路径与前沿研究进展。首先，在具身感知方面，无人机通过主动视觉、三维重建、触觉感知与视觉语言导航等手段，实现了对复杂环境的高效认知与结构建模，支撑了精细化任务执行与交互操作。其次，在具身决策层面，研究者围绕动态路径规划、多机任务分配与多模态指令理解等任务，构建了从语义理解到运动规划的智能化决策链条。特别是在VLA系统方面，多项新兴模型融合自然语言、视觉感知与动作策略，在搜索救援、巡检作业等实际应用中展现出跨模态自主操作能力。最后，在交互与物理执行环节，研究围绕物理操控展开了深入探索，推动无人机逐步从信息采集工具向具身交互主体演进。

总体来看，无人机具身智能正逐步从感知驱动走向任务驱动，从模块分离走向系统协同。其技术体系日益呈现出多模态融合、跨任务联动、端到端优化的演进趋势，推动具身智能从单一领域的性能优化转向多场景、多任务下的系统泛化能力构建。本文旨在为该领域的持续发展提供系统性技术梳理与研究启发，构建无人机具身智能技术图谱的基础认知框架。

参考文献：

[1] 孙长银, 袁心, 王远大, 等. 具身智能自主无人系统技术[J]. 自动化学报, 2025, 51(4): 762-777.
SUN C Y, YUAN X, WANG Y D, et al. Embodied intelligence autonomous unmanned systems technology[J]. Acta Automatica Sinica, 2025, 51(4): 762-777.

[2] ANDERSON P, WU Q, TENEY D, et al. Vision-and-language navigation: interpreting visually-grounded navigation instructions in real environments[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 3674-3683.

[3] 於志文, 孙卓, 程岳, 等. 智能无人机集群协同感知计算研究综述[J]. 航空学报, 2024, 45(20): 1-16.
YU Z W, SUN Z, CHENG Y, et al. A review of intelligent UAV swarm collaborative perception and computation[J]. Acta Aeronautica et Astronautica Sinica, 2024, 45(20): 1-16.

[4] DONG J, OTA K, DONG M X. UAV-based real-time survivor detection system in post-disaster search and rescue operations[J]. IEEE Journal on Miniaturization for Air and Space Systems, 2021, 2(4): 209-219.

[5] YAN P, GUO J F, SU X J, et al. Long-term tracking of evasive urban target based on intention inference and deep reinforcement learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(11): 16886-16900.

[6] LIU S B, ZHANG H S, QI Y K, et al. AerialVLN: vision-and-language navigation for UAVs[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV), October 1-6, 2023, Paris, France. New York: IEEE, 2024: 15338-15348.

[7] SHAH S, DEY D, LOVETT C, et al. AirSim: high-fidelity visual and physical simulation for autonomous vehicles [C]//HUTTER M, SIEGWART R, eds. Field and Service Robotics. Cham: Springer, 2018: 621-635.

[8] ALMAHAMID F, GROLINGER K. VizNav: a modular off-policy deep reinforcement learning framework for vision-based autonomous UAV navigation in 3D dynamic environments[J]. Drones, 2024, 8(5): 173.

[9] 王文晟, 谭宁, 黄凯, 等. 基于大模型的具身智能系统综述[J]. 自动化学报, 2025, 51(1): 1-19.
WANG W S, TAN N, HUANG K, et al. Embodied intelligence systems based on large models: a survey[J]. Acta Automatica Sinica, 2025, 51(1): 1-19.

[10] GIRISHA S, VERMA U, MANOHARA PAI M M, et al. UVid-net: enhanced semantic segmentation of UAV aerial videos by embedding temporal information[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 4115-4127.

[11] CHENG Q, WANG Y Z, HE W J, et al. Lightweight air-to-air unmanned aerial vehicle target detection model[J]. Scientific Reports, 2024, 14: 2609.

[12] HAN Y H, DUAN B C, GUAN R X, et al. LUFFD-YOLO: a lightweight model for UAV remote sensing forest fire detection based on attention mechanism and multi-level feature fusion[J]. Remote Sensing, 2024, 16(12): 2177.

[13] QING Y H, LIU W Y, FENG L Y, et al. Improved YOLO network for free-angle remote sensing target detection[J]. Remote Sensing, 2021, 13(11): 2171.

[14] FANG W X, ZHANG G Q, ZHENG Y H, et al. Multi-task learning for UAV aerial object detection in foggy weather condition[J]. Remote Sensing, 2023, 15(18): 4617.

[15] LIAO J W, JIANG S Y, CHEN M H, et al. SAM-YOLO: an improved small object detection model for vehicle detection[J]. The European Journal on Artificial Intelligence, 2025, 38(3): 279-295.

[16] GUO J T, CAO S, WANG T, et al. Transformer-based InspecNet for improved UAV surveillance of electrical infrastructure[J]. International Journal of Applied Earth Observation and Geoinformation, 2025, 137: 104424.

[17] ZHANG Z Q, ZHANG Y F, XIANG S, et al. KDP-net: an efficient semantic segmentation network for emergency landing of unmanned aerial vehicles[J]. Drones, 2024, 8 (2): 46.

[18] YU C Q, GAO C X, WANG J B, et al. BiSeNet V2: bilateral network with guided aggregation for real-time semantic segmentation[J]. International Journal of Computer Vi-

sion, 2021, 129(11): 3051-3068.

[19] HUANG L Y, TAN J T, CHEN Z H. Mamba-UAV-SegNet: a multi-scale adaptive feature fusion network for real-time semantic segmentation of UAV aerial imagery[J]. Drones, 2024, 8(11): 671.

[20] YI S, LIU X, LI J J, et al. UAVformer: a composite transformer network for urban scene segmentation of UAV images[J]. Pattern Recognition, 2023, 133: 109019.

[21] YIN X Q, LI X, NI P Z, et al. A novel real-time edge-guided LiDAR semantic segmentation network for unstructured environments[J]. Remote Sensing, 2023, 15(4): 1093.

[22] QU R K, TAN J T, LIU Y L, et al. NoctuDroneNet: real-time semantic segmentation of nighttime UAV imagery in complex environments[J]. Drones, 2025, 9(2): 97.

[23] FLOREA H, NEDEVSCHI S. TanDepth: leveraging global DEMs for metric monocular depth estimation in UAVs [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2025, 18: 5445-5459.

[24] CHENG Z Y, ZHANG Y, YU Y, et al. TinyDepth: lightweight self-supervised monocular depth estimation based on transformer[J]. Engineering Applications of Artificial Intelligence, 2024, 138: 109313.

[25] DHAFANI D, LIU Y F, JONG A, et al. FIReStereo: forest InfraRed stereo dataset for UAS depth perception in visually degraded environments[J]. IEEE Robotics and Automation Letters, 2025, 10(4): 3302-3309.

[26] MADHUANAND L, NEX F, YANG M Y. Self-supervised monocular depth estimation from oblique UAV videos[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 176: 1-14.

[27] YU K L, LI H, XING L J, et al. Scene-aware refinement network for unsupervised monocular depth estimation in ultra-low altitude oblique photography of UAV[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 205: 284-300.

[28] MUR-ARTAL R, TARDÓS J D. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras[J]. IEEE Transactions on Robotics, 2017, 33(5): 1255-1262.

[29] QIN T, LI P L, SHEN S J. VINS-mono: a robust and versatile monocular visual-inertial state estimator[J]. IEEE Transactions on Robotics, 2018, 34(4): 1004-1020.

[30] XU W, ZHANG F. FAST-LIO: a fast, robust LiDAR-inertial odometry package by tightly-coupled iterated Kalman filter[J]. IEEE Robotics and Automation Letters, 2021, 6 (2): 3317-3324.

[31] LIU X, NARDARI G V, CLADERA F, et al. Large-scale autonomous flight with real-time semantic SLAM under dense forest canopy[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 5512-5519.

[32] FANTA-JENDE P, STEININGER D, KERN A, et al. Semantic real-time mapping with UAVs[J]. PFG-Journal of Photogrammetry, Remote Sensing and Geoinformation Science, 2023, 91(3): 157-170.

[33] YANG L J, YE J, ZHANG Y, et al. A semantic SLAM-based method for navigation and landing of UAVs in indoor environments[J]. Knowledge-Based Systems, 2024, 293: 111693.

[34] YAO Y W, SUN C, WANG T, et al. UAV geo-localization dataset and method based on cross-view matching[J]. Sensors, 2024, 24(21): 6905.

[35] CUI Z F, ZHOU P W, WANG X L, et al. A novel geo-localization method for UAV and satellite images using cross-view consistent attention[J]. Remote Sensing, 2023, 15(19): 4667.

[36] AUCONE E, SFERRAZZA C, GREGOR M, et al. Optical tactile sensing for aerial multicontact interaction: design, integration, and evaluation[J]. IEEE Transactions on Robotics, 2025, 41: 364-377.

[37] VIGARA-PUCHE V, FERNANDEZ-GONZALEZ M J, FUMAGALLI M. Design and validation of an obstacle contact sensor for aerial robots[J]. Sensors, 2024, 24(23): 7814.

[38] HATTENBERGER G, BRONZ M, CONDOMINES J P. Estimating wind using a quadrotor[J]. International Journal of Micro Air Vehicles, 2022, 14: 17568293211070824.

[39] ASIGNACION A, SUZUKI S, NODA R, et al. Frequen-

cy-based wind gust estimation for quadrotors using a nonlinear disturbance observer[J]. IEEE Robotics and Automation Letters, 2022, 7(4): 9224-9231.

[40] LEE J, MIYANISHI T, KURITA S, et al. CityNav: a large-scale dataset for real-world aerial navigation[EB/OL]. (2025-08-02)[2025-11-01]. https://arxiv.org/abs/2406.14240.

[41] GAO Y P, LI C H, YOU Z R, et al. OpenFly: a comprehensive platform for aerial vision-language navigation [EB/OL]. (2025-07-31)[2025-11-01]. https://arxiv.org/abs/2502.18041.

[42] TORDESILLAS J, LOPEZ B T, EVERETT M, et al. FASTER: fast and safe trajectory planner for navigation in unknown environments[J]. IEEE Transactions on Robotics, 2022, 38(2): 922-938.

[43] YE H K, ZHOU X, WANG Z P, et al. TGK-planner: an efficient topology guided kinodynamic planner for autonomous quadrotors[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 494-501.

[44] SONG Y L, SHI K X, PENICKA R, et al. Learning perception-aware agile flight in cluttered environments[C]// 2023 IEEE International Conference on Robotics and Automation (ICRA), May 29-June 2, 2023, London, United Kingdom. New York: IEEE, 2023: 1989-1995.

[45] PAREDES-VALLÉS F, HAGENAARS J J, DUPEYROUX J, et al. Fully neuromorphic vision and control for autonomous drone flight[J]. Science Robotics, 2024, 9(90): eadi0591.

[46] ALQEFARI S, EL BACHIR MENAI M. A hybrid method to solve the multi-UAV dynamic task assignment problem [J]. Sensors, 2025, 25(8): 2502.

[47] HAN S, FAN C C, LI X B, et al. A modified genetic algorithm for task assignment of heterogeneous unmanned aerial vehicle system[J]. Measurement and Control, 2021, 54(5/6): 994-1014.

[48] CHEN L Z, LIU W L, ZHONG J H. An efficient multi-objective ant colony optimization for task allocation of heterogeneous unmanned aerial vehicles[J]. Journal of Computational Science, 2022, 58: 101545.

[49] XU Y, LI X B, MENG X P. A Q-learning based iterated local search algorithm for human-UAV cooperation in restoring transmission network[J]. Expert Systems with Applications, 2024, 252: 124200.

[50] WANG G, LV X, YAN X H. A two-stage distributed task assignment algorithm based on contract net protocol for multi-UAV cooperative reconnaissance task reassignment in dynamic environments[J]. Sensors, 2023, 23(18): 7980.

[51] YIN Y F, GUO Y, SU Q R, et al. Task allocation of multiple unmanned aerial vehicles based on deep transfer reinforcement learning[J]. Drones, 2022, 6(8): 215.

[52] FAN M F, LIU H, WU G H, et al. Multi-UAV reconnaissance mission planning via deep reinforcement learning with simulated annealing[J]. Swarm and Evolutionary Computation, 2025, 93: 101858.

[53] BROHAN A, BROWN N, CARBAJAL J, et al. RT-2: vision-language-action models transfer web knowledge to robotic control[EB/OL]. (2023-07-28)[2025-11-01]. https://arxiv.org/abs/2307.15818.

[54] SAUTENKOV O, YAQOOT Y, LYKOV A, et al. UAV-VLA: vision-language-action system for large scale aerial mission generation[C]//2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI), March 4-6, 2025, Melbourne, Australia. New York: IEEE, 2025: 1588-1592.

[55] WANG M Y, WANG Q H, WANG Z, et al. Unlocking aerobatic potential of quadcopters: autonomous freestyle flight generation and execution[J]. Science Robotics, 2025, 10(101): eadp9905.

[56] WANG Q H, WANG Z P, WANG M Y, et al. Fast iterative region inflation for computing large 2-D/3-D convex regions of obstacle-free space[J]. IEEE Transactions on Robotics, 2025, 41: 3223-3243.

[57] XU H, LIU P Z, CHEN X Y, et al. D^2 SLAM: decentralized and distributed collaborative visual-inertial SLAM system for aerial swarm[J]. IEEE Transactions on Robotics, 2024, 40: 3445-3464.

[58] ZHOU B Y, XU H, SHEN S J. RACER: rapid collaborative exploration with a decentralized multi-UAV system [J]. IEEE Transactions on Robotics, 2023, 39(3): 1816-

1835.

[59] JIMENEZ-CANO A E, SANCHEZ-CUEVAS P J, GRAU P, et al. Contact-based bridge inspection multirotors: design, modeling, and control considering the ceiling effect [J]. IEEE Robotics and Automation Letters, 2019, 4(4): 3561-3568.

[60] LEE D, SEO H, JANG I, et al. Aerial manipulator pushing a movable structure using a DOB-based robust controller[J]. IEEE Robotics and Automation Letters, 2021, 6 (2): 723-730.

[61] MARTINEZ R R, PAUL H, SHIMONOMURA K. Aerial torsional work utilizing a multirotor UAV with add-on thrust vectoring device[J]. Drones, 2023, 7(9): 551.

[62] PAUL H, MIYAZAKI R, KOMINAMI T, et al. A versatile aerial manipulator design and realization of UAV take-off from a rocking unstable surface[J]. Applied Sciences, 2021, 11(19): 9157.

[63] DING C W, LU L, WANG C, et al. Design, sensing, and control of a novel UAV platform for aerial drilling and screwing[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 3176-3183.

[64] PARK S, LEE J, AHN J, et al. ODAR: aerial manipulation platform enabling omnidirectional wrench generation [J]. IEEE/ASME Transactions on Mechatronics, 2018, 23 (4): 1907-1918.

第一作者：王森（2000—），男，博士研究生，研究方向为无人机具身智能。E-mail：230259115@seu.edu.cn。

通信作者：朱鹏飞（1986—），男，教授，博士，博士生导师，国家优秀青年科学基金项目和天津市杰出青年科学基金项目获得者。研究方向为低空智能感知、低空具身智能。E-mail：zhupengfei@seu.edu.cn。

（责任编辑：姜红贵）