©中国图象图形学报版权所有

中图法分类号:TP391;TP18　　**文献标识码**:A　　**文章编号**:1006-8961(2026)09-3251-15

论文引用格式:Lang J W, Li Y X, Liu S, Kang X D and Cheng J Q. 2026. Progress and perspectives on large models for aerial remote sensing intelligent interpretation. Journal of Image and Graphics, 31(9):3251-3265(郎晋伟, 李娅星, 刘帅, 康晓东, 程俊强. 2026. 大模型赋能航空遥感智能解译的进展与思考. 中国图象图形学报, 31(9):3251-3265)[DOI:10.11834/jig.260217]

大模型赋能航空遥感智能解译的进展与思考

郎晋伟*,李娅星,刘帅,康晓东,程俊强

中国航空工业集团西安航空计算技术研究所,西安　710065

摘　要:随着航空遥感多源传感器的发展及多模态数据融合与智能解译技术的推进,航空遥感技术正经历从单一模态到多模态的深刻变革。这一进展为精准农业、城市环境监测、生态保护及自然灾害评估等领域的智能化监测与决策提供了重要的应用前景。传统的遥感影像解译方法主要依赖单源传感器数据和单任务学习模型,难以应对地表地物目标尺度多变、语义层次丰富、时空异质性强的复杂场景。近年来,多模态大模型的快速发展推动了遥感技术的显著进步,尤其在视觉—语言融合、跨模态推理,以及基于任务指令的跨模态分析与生成等方面取得了重要成果。然而,如何有效处理多源异构数据、提升模型的可解释性,同时保证实时推理能力,仍是当前面临的核心挑战。针对上述问题,对近年来航空遥感领域多模态数据融合及大模型技术的研究进展进行了系统评述,并深入分析了关键技术瓶颈与未来发展方向。研究重点包括跨传感器、多源及多尺度数据融合、空间语义推理、模型的可信推理与可解释性评估,以及在边缘计算环境下的高效部署与数据隐私保护策略。这些问题不仅是推动航空遥感智能解译技术从理论研究向实际应用转化的关键所在,也是实现技术落地并确保可持续发展的现实挑战。

关键词:航空遥感;多模态融合;视觉—语言模型(VLM);大模型;语义理解

Progress and perspectives on large models for aerial remote sensing intelligent interpretation

Lang Jinwei*, Li Yaxin, Liu Shuai, Kang Xiaodong, Cheng Junqiang

Xi'an Aeronautics Computing Technique Research Institute, AVIC, Xi'an　710065, China

Abstract: With the proliferation of multisource sensors and the rapid progress of multimodal data fusion and intelligent interpretation techniques, aerial remote sensing continues to evolve from traditional single-modal perceptions toward multimodal perception and understanding. This progress demonstrates the significant potential for intelligent monitoring and decision-making in precision agriculture, urban environmental monitoring, ecological protection, and natural disaster assessment. Due to ongoing advancements in Earth observation capabilities, multisource remote sensing data, including optical imagery, synthetic aperture radar (SAR), and hyperspectral imagery are being increasingly characterized by high spatial resolution, multitemporal coverage, and multidimensional richness. These developments, in turn, offering a solid foundation for fine-grained land-cover identification and dynamic change analysis. Nevertheless, traditional remote sensing interpretation methods, which rely on single-source data and single-task learning models, struggle to cope with the complexities of real-world scenarios in which object scales vary dramatically, semantic layers are rich and entangled, and spa-

收稿日期:2026-04-16;**修回日期**:2026-06-02;**预印本日期**:2026-06-09

*通信作者:郎晋伟　langjw@mail.ustc.edu.cn

基金项目:航空科学基金项目(2024Z015031002)

Supported by:Aerospace Science Foundation(2024Z015031002)

tiotemporal heterogeneity is strong. Because such methods cannot adequately balance holistic scene semantics with local fine-grained details, this limits their capacity for high-level semantic understanding and comprehensive decision-making. Compared with satellite remote sensing, aerial platforms that typically employ unmanned aerial vehicles (UAVs) and low-altitude aircraft offer high spatial resolution and great observational flexibility, facilitating the completion of several tasks, such as fine urban modeling, small-object recognition, and emergency monitoring. However, these advantages also introduce severe scale variations, highly complex backgrounds, and inconsistent imaging conditions, making it extremely difficult for conventional models to capture both global scene semantics and subtle local textures simultaneously. Recent advances in deep learning and artificial intelligence (AI) have transformed remote sensing interpretation from manual inspection toward automation and intelligence, achieving notable results in scene classification, object detection, and semantic segmentation. Yet single-task and single-modality paradigms remain inadequate in terms of fully exploiting complementary information in heterogeneous multisource data. As such, these paradigms fall short of supporting the high-level semantic comprehension and integrated decision-making required for complex applications. The emergence of multimodal large models provides a transformative pathway for intelligent aerial remote sensing interpretation. These models can jointly perform image understanding and linguistic reasoning by integrating visual encoders with large language models (LLMs) through instruction tuning and cross-modal alignment, thus achieving breakthroughs in visual-language fusion, cross-modal reasoning, and task-instruction-guided analysis. At the same time, within the remote sensing community, researchers have begun to construct large-scale image-text datasets to train vision-language models and employ linguistic guidance to enhance the comprehension of complex land-cover semantics, thus marking a critical shift from purely visual modeling toward semantically augmented modeling and laying the groundwork for the subsequent development of multimodal large models. A systematic review reveals that the evolution of remote sensing large models has followed a clear paradigm progression: from early approaches that relied solely on single visual modalities, through vision-language models that introduced textual semantics but are still confined to specific tasks, to the current stage in which unified multimodal large models enable cross-task collaboration and complex reasoning. Despite these promising advances, the application of multimodal large models to aerial remote sensing still faces several formidable challenges. The substantial disparities in spatiotemporal references and radiometric properties among heterogeneous data sources, such as optical, SAR, and hyperspectral imagery, render cross-modal alignment extremely difficult, often giving rise to semantic misalignment and information loss. While fine-grained spatial structures and multiscale objects inherent in aerial imagery impose stringent demands on spatial cognition and multiscale reasoning, general-purpose large models often lack sufficient domain-specific spatial priors to accurately capture geometric and structural relationships. High-resolution scenes that impose massive computational and storage burdens also make real-time inference on edge devices highly challenging, resulting in the failure to meet the stringent timeliness and generalization requirements of time-sensitive tasks, such as disaster response and UAV-based inspection. Moreover, the black-box nature of model decision-making leads to insufficient interpretability and trustworthiness that hinders deployment in safety-critical scenarios. Meanwhile, data privacy concerns have become increasingly prominent within distributed collaborative learning frameworks. In response to the abovementioned challenges, this paper systematically reviews recent advances in multimodal data fusion and large model technologies for aerial remote sensing by tracing the shift from single-source to multimodal integration. Driven by large models, remote sensing interpretation has gradually evolved from low-level perception to cross-modal reasoning and semantic understanding. Nevertheless, substantial challenges remain in multimodal alignment, spatial cognition, task reliability, and practical deployment in real-world scenarios. To help resolve these problems, we analyze cross-sensor fusion and emphasize how disparities in geometry, radiometry, and acquisition timing among optical, SAR, and LiDAR data severely impede cross-modal semantic alignment. Furthermore, despite advances in vision-language pretraining, high resolution and complex backgrounds often degrade alignment accuracy, thus requiring fine-grained multimodal representations through geometric correction, radiometric normalization, and spatiotemporal consistency modeling. For multiscale spatial reasoning, we emphasize that complex tasks require holistic understanding of structural relations and spatial distributions. We argue that while existing models partially strengthen spatial reasoning via region-based interaction and spatial question answering, a unified framework for global spatial relation modeling is needed to elevate task awareness and multilevel reasoning. Generative AI for few-shot interpreta-

tion has the potential to alleviate annotation scarcity and support time-critical missions, such as disaster response. Regarding trustworthiness and interpretability, peculiar data distributions and insufficient instruction data in remote sensing exacerbate hallucinations in large vision-language models. Thus, domain knowledge, task constraints, and dedicated evaluation systems are required to enforce reasoning consistency and reliable high-stakes decisions. For efficient edge deployment and privacy, high-resolution large-format data impose heavy computational and storage burdens. Furthermore, although lightweight networks continue to emerge, trade-offs among real-time performance, stability, and complex scene adaptability remain, which means that model compression, inference optimization, and hardware-software co-design become critical. Related to this, federated learning offers a privacy-preserving mechanism for multisource data integration. In summary, addressing these challenges is important not only for advancing the theoretical understanding of cross-modal semantic reasoning and multiscale spatial cognition but also for facilitating the practical deployment of intelligent aerial remote sensing systems in high-stakes, real-world applications.

Key words: aerial remote sensing; multimodal fusion; visual-language model(VLM); large model; semantic comprehension

0 引　言

航空遥感技术作为获取地球表面信息的重要手段，在资源调查、生态环境监测、灾害评估及国防安全等领域发挥着关键作用(Sun 等，2024)。随着对地观测能力的不断提升，多源遥感数据(如光学影像、合成孔径雷达(synthetic aperture radar, SAR)及高光谱数据)呈现出高分辨率、多时相与多维度的发展趋势，为精细化地物识别与动态变化分析提供了丰富的数据基础(许文嘉 等，2026；Zhou 等，2025；Zhu 等，2017)。近年来，随着深度学习与人工智能(artificial intelligence，AI)技术的发展，遥感智能解译逐渐从传统的人工判读向自动化与智能化方向演进，尤其是在场景分类、目标检测与语义分割等任务中取得了显著进展(Kazanskiy 等，2025；Li 等，2024，2025c；Ma 等，2019)。然而，面对持续增长的多源异构遥感数据，仅依赖传统单任务、单模态方法，仍难以有效支撑复杂应用中的高层语义理解与综合决策。

相较于航天遥感，航空遥感在数据获取方式与应用场景上具有显著差异(Bazrafkan 等，2025；Xing 等，2026)。航空遥感通常依托无人机或低空飞行平台，具备更高的空间分辨率与更灵活的观测能力，其在城市精细建模、小目标识别及应急监测等场景中具有独特优势(Alvarez-Vanhard 等，2021；Chang 等，2025；Zhang 等，2021)。然而，这种高分辨率与灵活性也带来了新的挑战，例如，目标尺度变化剧烈、背景复杂以及成像条件多变等问题，使得传统遥感模型难以兼顾全局语义与局部细节(Ji 等，2025；Jiang 等，2020；Liu 等，2025)。此外，航空遥感应用往往具有较强的任务驱动特性，例如灾害响应与无人机巡检等，对模型的实时性、泛化能力及交互能力提出了更高要求(Cao 等，2023；Xu 等，2025a；Zhang 和 Zhu，2023)。因此，面向航空场景的遥感智能解译方法亟需在模型结构与语义表达能力上进行进一步提升。

近年来，大模型特别是多模态大模型(multimodal large language model, MLLM)的快速发展，为航空遥感智能解译提供了新的技术路径。航空遥感成像平台与多模态大模型之间的相互承接关系如图 1 所示。MLLM 通过融合视觉编码器与大语言模型(large language model, LLM)，并结合指令微调与跨模态对齐机制，能够同时处理图像理解与语言推理任务(Radford 等，2021；Wu 等，2025a；Zhan 等，2025)。在遥感领域，相关研究开始探索将视觉—语言模型(vision-language model, VLM)应用于遥感场景，例如通过构建大规模图文数据集实现跨模态语义对齐，或通过引入语言指导提升模型对复杂地物语义的理解能力(Qiu 等，2025；Tao 等，2025；Zhang 等，2025b)。这些研究标志着航空遥感从传统视觉建模向语义增强建模的重要转变，为后续多模态大模型的发展奠定了基础。

然而，现有研究表明，遥感大模型的发展并不仅限于参数量的增加，而是经历了从单模态到多模态的范式演进过程，各类大模型对航空遥感的协同贡献如图 2 所示。早期方法主要依赖单一视觉信息进行建模，而视觉—语言模型通过引入文本信息实现语义增强，但仍局限于特定任务(Bazi 等，2024；Liu 等，2024a)。近年来，多模态大模型在统一建模框架下实现了跨任务协同与复杂推理能力的提升(Wang

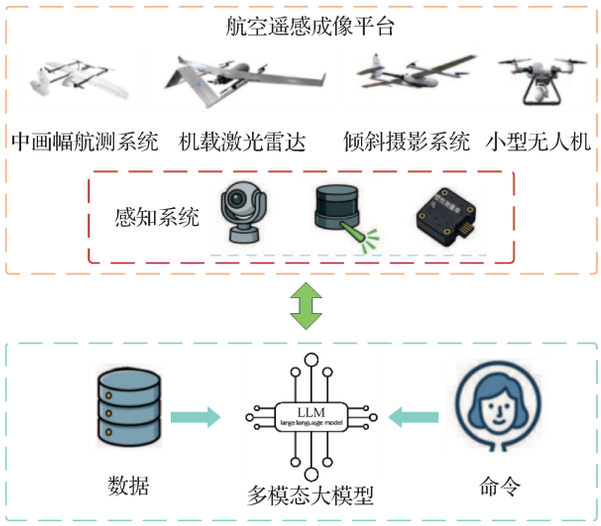


图1　航空遥感成像平台与多模态大模型

Fig. 1　Aerial remote sensing imaging platform and multimodal large model

等，2024b；Zhang 等，2024b）。与此同时，航空遥感场景的特殊性，如多源异构数据、复杂空间结构及高分辨率特征，使得多模态大模型在应用过程中面临新的挑战，包括跨模态对齐困难、空间认知能力不足以及高计算成本等问题（Huang 等，2025；Yang 等，2025a）。因此，有必要系统梳理遥感大模型从单模态到多模态的发展路径，以明确其技术演进逻辑与关键问题。

在此背景下，本文以“从单源到多模态”为主线，系统梳理了航空遥感智能解译的演进轨迹及关键挑战。本文从方法体系演进的视角出发，深入剖析了单源视觉建模、视觉—语言融合以及多模态大模型3个阶段的发展脉络，重点分析了多模态大模型在通用表征能力、典型遥感任务适配及关键支撑技术等方面的研究进展与发展趋势。结合航空遥感实际应用需求，进一步提炼当前面临的主要挑战，包括高分辨率场景下的建模压力、多源异构与时空不一致带来的融合难题、通用大模型与领域知识之间的鸿沟，以及面向真实任务部署中的效率、可信与可解释性问题。

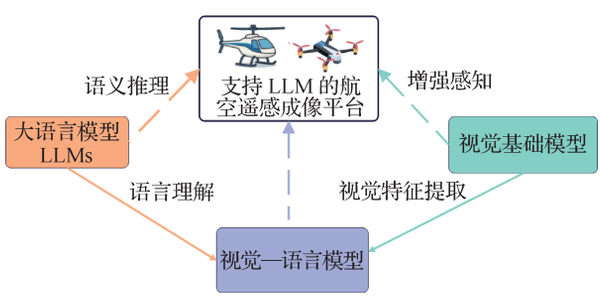


图2　各类大模型对航空遥感的协同贡献

Fig. 2　The collaborative contributions of various large-scale models to aerial remote sensing

1　大模型赋能航空遥感智能解译的演进与应用

1.1　单源视觉建模

航空遥感影像具有超高空间分辨率、复杂地物结构以及目标尺度差异显著等典型特征（Zhang 和 Zhu，2023）。相比卫星遥感（Ji 等，2025；Zhu 和 Zhang，2025），航空遥感能够提供更为精细的纹理与结构信息，但同时也带来了数据冗余高、类别区分困难以及背景干扰复杂等问题。围绕这一特性，遥感智能解译最初主要依赖单源视觉建模方法，即仅利用图像数据本身进行自动分析，其核心目标是完成对地物的感知级理解。早期的遥感影像解译主要依赖人工构建特征，而随着深度学习技术的兴起，单一视觉信息的建模逐渐由传统方法转向以数据驱动为核心的方法。以卷积神经网络为代表的方法能够通过多层非线性结构自动学习空间表征，在场景分类与土地覆盖制图等任务中显著优于传统方法（Wang 等，2022b）。随着研究的深入，模型结构不断演进以适应遥感影像中尺度变化大与背景复杂的特点，多尺度特征融合方法（如特征金字塔）被引入以提升不同尺度目标的表征能力，同时注意力机制通过强化关键区域响应、抑制冗余信息，提高了复杂场景下的特征判别能力。在此过程中，面向具体任务的模型结构不断优化。例如，在目标检测任务中，以 YOLO（you only look once）为代表的单阶段检测框架，以及结合状态空间模型（如 Mamba）的改进方法，通过增强多尺度特征表征能力与长距离空间依赖建模能力，有效提升了复杂背景及多尺度目标场景下的检测精度（Guo 等，2026）。与此同时，Transformer 架构逐渐应用于遥感视觉任务，尤其在变化检测领域，通过构建跨时相特征之间的全局关联关系与时序交互机制，增强了变化区域特征的判别能力，并在一定程度上抑制了伪变化带来的干扰（Wang 等，2022a）。这一阶段的模型以任务性能为导向，尽管在分类、检测与分割等任务中取得显著进展，但整体仍依赖实际任务来实现定制设计，缺乏统一表征能力。

随着数据规模与算力水平的提升，单模态视觉建模逐步向大规模预训练范式演进。基于海量遥感数据构建的视觉基础模型相继出现，例如RingMo模型通过自监督学习具有较强迁移能力的通用视觉表征，在多类下游任务中表现出良好的适应性与性能优势（Wang 等，2025）；RemoteCLIP（remote contrastive language-image pretraining）则通过统一视觉表征空间的构建，提升了模型在遥感场景分类、跨模态检索及少样本学习等任务中的泛化能力（Liu 等，2024b）。这类模型不再围绕单一任务进行优化，而是通过“预训练—微调”实现统一表征学习，并能够同时支撑分类、目标检测、语义分割与变化分析等多类任务（陈诗琪 等，2026）。相比前一阶段以任务为中心的模型设计，单模态大模型在表达能力与迁移性能上进一步提升，同时也标志着遥感视觉建模从任务驱动向表征驱动的转变，为后续多模态融合与统一建模奠定了基础。

从整体来看，单源视觉建模仍然主要停留在视觉感知层面，在实际应用中逐渐显示出一定的局限性。这类方法高度依赖大规模标注数据，在跨区域迁移或成像条件变化时，性能往往显著下降（Sun 等，2026）。同时，复杂背景干扰与小目标检测困难的问题也始终未能得到根本解决（高广帅 等，2025；Xu 和 Ma，2026）。更值得关注的是，单模态模型的输出形式通常局限于类别标签或空间位置，缺乏对语义关系与场景含义的深层理解和表达能力。这意味着单模态模型虽然能够完成地物识别，但在高层语义解释与目标间关系建模方面能力有限，难以支撑复杂场景下的语义推理需求（李海峰 等，2025）。基于上述特征，单源视觉建模阶段本质上可以归结为以感知任务为核心的自动解译阶段。随着应用需求从识别地物逐步向理解场景及其语义关系转变，研究者开始尝试引入自然语言信息，以构建更具表达能力的模型。由此，遥感智能解译开始从单模态视觉建模，逐步迈向视觉—语言融合的新阶段（Tuia 等，2026）。

1.2 视觉—语言融合模型

在单源视觉建模阶段取得显著进展的基础上，遥感图像解译逐步从视觉感知迈向语义理解。其核心转变在于：不再仅依赖图像特征进行地物识别，而是通过引入自然语言，实现视觉信息与语义表达之间的对齐与统一建模。近年来，视觉—语言模型在自然图像领域取得突破性进展，其中代表性工作CLIP（contrastive language-image pretraining）通过大规模图文对比学习，构建了统一的图像—文本嵌入空间，为跨模态理解提供了基础框架（Radford 等，2021）。大量研究表明，该模型具备良好的零样本泛化能力，也为遥感领域的跨模态研究与应用提供了重要启发。表1列举了近3年航空遥感领域的代表性大模型，按照视觉基础模型、视觉—语言大模型和任务特定模型进行分类。分类依据主要为模型的输入模态和任务目标，其中任务特定模型包括变化检测、区域描述和开放词表目标检测等。表1对各模型的训练方式、方法策略、核心优势及典型应用进行了简要整理。

在此基础上，研究者开始将视觉—语言模型引入遥感场景，推动解译任务由“分类与检测”向“语义匹配与跨模态检索”拓展（Liu 等，2024b；Xu 等，2025b）。例如，RSICD（remote sensing image caption dataset）为遥感图像描述任务提供了标准数据支撑（Lu 等，2018）；进一步地，针对遥感领域特征构建的模型，如RemoteCLIP，通过遥感图文数据进行再训练，有效提升了模型在遥感语义检索与分类任务中的表现（Liu 等，2024b）。有研究进一步指出，在遥感基础模型构建中，视觉—语言联合预训练通过建立视觉表征与语义信息之间的跨模态对齐关系，不仅缓解了遥感领域标注数据不足的问题，同时增强了模型在多任务场景下的迁移能力与语义理解能力（Xiao 等，2025）。这些研究表明，视觉—语言融合正在成为遥感智能解译的重要发展方向。

然而，与自然图像相比，遥感影像在视觉—语言对齐过程中面临更为复杂的挑战。航空遥感影像具有俯视成像、尺度变化显著及语义表达抽象等特征，导致视觉内容与自然语言之间难以建立稳定的对应关系。同时，遥感领域缺乏大规模高质量图文配对数据，进一步制约了视觉—语言模型的训练效果与泛化能力（Cheng 等，2026；Lin 等，2025；Yang 等，2024b；Zhang 等，2025a）。表2汇总了近年来航空遥感视觉—语言建模相关的典型数据集，包括数据规模、图像数量、任务类型及主要应用场景，反映出该领域在数据资源建设与任务体系拓展方面的发展趋势。在此背景下，构建面向遥感场景的视觉—语言预训练框架，并提升跨模态语义对齐与语义表征能力，成为该阶段研究的重要方向。总体而言，视觉—

表1　典型航空遥感大模型汇总与对比

Table 1　Summary and comparison of recent large models for aerial remote sensing

模型类型	模型名称	训练方式	方法策略	核心优势	典型应用
视觉基础模型	Scale-MAE(Reed等, 2023)	自监督预训练	多尺度感知编码	多尺度泛化强	场景分类
	RingMo-Aerial(Diao等, 2024)	对比学习预训练	仿射增强+多尺度注意力	小目标表现佳	分类/检测
	AnySat (Astruc等, 2025)	自监督预训练	多分辨率/多尺度/多模态建模	跨模态能力强	地表覆盖分类、变化检测、环境监测
视觉—语言大模型	GeoChat(Kuckreja等, 2024)	指令微调	区域输入+空间定位	区域描述+对话+定位一体	区域解释、交互系统
	RemoteCLIP(Liu等, 2024b)	图文对比学习	图文对齐	零样本分类能力强	开放类别分类、检索
	RSGPT(Hu等, 2025)	高质量图文微调	视觉—语言对齐+指令微调	推动遥感标注标准化	图像描述、视觉问答
	RS-LLaVA(Bazi等, 2024)	LoRA(low-rank adaptation)微调	视觉—语言融合+多任务指令微调	训练成本低、部署简单	多模态系统
	RS-MoE(Lin等, 2025)	专家混合训练	多专家路由机制	高效+高性能平衡	高质量描述生成
任务特定模型	RAT(Zhao和Xiong, 2024)	监督训练	区域特征+关系建模	局部目标描述精细	对象级描述
	RSVG(Zhan等, 2023)	监督训练	多层跨模态对齐	跨模态定位, 尺度与背景鲁棒	文本定位目标
	MTP(Wang等, 2024a)	多任务预训练	分割+检测联合训练	与遥感任务一致性强	通用检测
	MSVG(Ding等, 2025)	监督训练	多维跨模态交互	精细对齐能力强	小目标定位
	LAE-DINO(Pan等, 2024)	大规模预训练	文本驱动检测	支持未知类别	开放世界检测

注：表中选取的航空遥感的代表性大模型起始时间为2023年，且同时包含部分预印本文献。

语言融合推动遥感智能解译由视觉识别逐步向语义理解演进，并为后续多模态大模型的发展奠定了基础。

1.3　多模态大模型

1.3.1　通用能力与推理机制

多模态大模型的核心突破在于其从单一视觉识别能力向统一任务理解与推理能力的跃迁。通过大规模视觉—语言联合建模，模型不仅能够执行传统的目标检测与分类任务，还逐步实现了基于自然语言指令的跨任务推理与决策能力。这种能力转变本质上体现为从感知驱动向认知驱动的升级。已有研究表明，随着视觉与语言模态在语义空间的深度对齐与联合训练，多模态模型在指令理解、复杂语义推理及多任务泛化方面的能力均显著提升(Zhu等, 2025)。

在遥感场景中，这种能力尤为关键。传统方法通常针对单一任务构建专用模型，而多模态大模型通过引入指令驱动机制，可以实现任务即输入的统一建模。例如，RingMoGPT通过构建多任务指令调优数据，使模型能够在同一框架下完成目标检测、变化描述和视觉问答等多种任务(Wang等, 2025)。类似地，Change-Agent进一步结合大语言模型的推理能力，实现了变化检测、变化描述及变化原因分析的统一推理过程(Liu等, 2024a)。同时，这种统一建模的思路也在多项研究中得到进一步拓展，EarthGPT通过统一指令调优，将多传感器、多任务遥感解译纳入同一模型接口(Zhang等, 2024b)；EarthMarker则进一步将视觉提示与语言指令联合建模，实现了统一指令输入从纯文本扩展到视觉—语言协同交互(Zhang等, 2025d)；SPEX的研究表明，该方法已能够推广至光谱遥感的像素级地物提取任务(Si等, 2026)。Xue等人(2026)的研究指出，多模态大模型

表 2　支持视觉—语言建模的航空遥感数据集

Table 2　A survey of aerial remote sensing datasets for vision-language modeling

数据集	数据规模	主要任务类型	典型用途
RSICD（Lu 等，2018）	10 921 幅图像，每幅图像 5 条描述	图像描述、图文检索	遥感图像描述、图文匹配、跨模态检索
RSITMD（Yuan 等，2022）	4 743 幅图像，23 715 条文本描述	图文检索、细粒度图文匹配	遥感跨模态检索、细粒度语义匹配
NWPU-Captions（Cheng 等，2022）	31 500 幅图像，157 500 条描述	图像描述、图文检索	大规模遥感图像描述、图文检索
RSVQA（Lobry 等，2020）	100 659 幅图像，100 660 316 条问答记录	视觉问答	遥感视觉问答、目标计数、属性识别、空间关系推理
FloodNet-VQA（Rahnemoonfar 等，2021）	3 200 幅图像，包含视觉问答数据	灾害场景视觉问答	灾害场景问答、灾后损毁识别、应急监测
DIOR-RSVG（Zhan 等，2023）	17 402 幅图像，38 320 个图像—文本查询对	遥感数据视觉定位任务	文本引导目标定位、视觉锚定、小目标定位
LEVIR-CC（Liu 等，2022）	10 077 对双时相图像，50 385 条变化描述	变化描述、双时相图像理解	变化描述、土地利用变化解释、灾害前后对比分析
RSIEval（Hu 等，2025）	100 条人工描述，936 个开放式问答对	图像描述、视觉问答	遥感视觉语言模型综合评测、开放式问答、描述生成

正由单一变化任务的统一建模，演进为在统一表征下协同处理变化描述、定位与计数等多任务的交互式推理框架。

此外，生成式 AI 模型（类 GPT 模型）在跨模态推理中表现出显著优势。它们能够结合视觉信息与外部知识进行多步推理，从而在视觉问答、复杂场景解析等任务中实现更深层次的语义理解。例如，RS_DeepReason 通过构建多阶段推理链，将复杂问题拆解为多个子任务并逐步求解，有效提升了复杂场景下的推理准确性与可解释性（Yang 等，2026）。

1.3.2　多模态对齐与模型适配机制

多模态大模型实现复杂任务统一建模的关键，在于其底层的对齐机制与适配机制。首先，在多模态对齐方面，模型需要解决不同模态之间的语义一致性问题，同时兼顾遥感数据特有的几何、尺度与辐射差异。例如，CLV-Net 通过语义一致性与关系一致性联合约束，实现了视觉区域与文本描述之间的精细对齐（Zhang 等，2026）；BITA 方法则通过两阶段对比学习构建生成模型，提高了遥感图像与文本之间的跨模态一致性（Yang 等，2024b）。

其次，针对遥感领域样本规模有限及数据分布差异显著等问题，研究者提出了多种参数高效适配方法，以提升大模型在多任务场景下的迁移与泛化能力。其中，基于专家混合机制（mixture of experts，MoE）的方法通过动态分配任务相关子模型，实现模型容量与计算效率之间的平衡。例如，RS-MoE（mixture of experts for remote sensing）引入指令路由机制，使不同专家模块分别面向特定任务进行学习，从而提升模型整体性能（Lin 等，2025）。此外，轻量化微调策略也广泛应用于遥感大模型适配过程。CLIP-MoA 通过多适配器融合机制实现跨任务知识共享与快速迁移（Fu 等，2025）；LVM-StARS（large vision model soft adaption for RS）则采用软适配策略，在降低参数更新开销的同时，提高了模型的适配效率与训练稳定性（Yang 等，2024a）。

此外，近年来无参数或少参数适配方法（如知识注入与检索增强）也逐渐兴起。相关研究表明，在无需额外参数训练的条件下，通过构建面向遥感任务的知识库，并在推理阶段动态引入与当前任务相关的先验信息，可以有效提升多模态大模型在遥感图像理解中的适应能力与推理性能（Li 等，2025b）。多模态对齐与高效适配分别面向表征一致性与应用效率，是支撑多模态大模型在遥感复杂任务中有效运行的重要基础。

1.3.3　目标级理解与精细感知

多模态大模型在航空遥感领域的应用正由单任

务优化向多任务协同与复杂场景应用演变。在变化检测、目标识别与语义分割等基础任务中，已有研究验证了多模态建模方法在特征表达与语义理解方面的优势（潘子哲 等，2026；陶超 等，2025）。例如，MGCR-Net（multimodal graph-conditioned vision-language reconstruction network）模型通过视觉—语言联合建模与图结构约束，实现了变化检测中的细粒度语义交互与特征重构，从而增强了复杂场景下的变化表征能力（Wang等，2026）。针对多源遥感船舶检测任务，Popeye模型能够统一处理光学与SAR等异构模态数据，并同时支持水平框、旋转框及像素级分割等多种标注形式，表明多模态模型正逐步突破传统单任务检测框架的限制，向跨模态、跨任务的精细感知方向发展（Zhang等，2024a）。此外，在开放场景遥感解译中，SPEX（spectral instruction extraction）将视觉—语言模型扩展至光谱遥感地物提取任务（Si等，2026），通过构建面向指令驱动的数据组织与交互框架，实现了统一指令条件下的像素级地物提取，体现出多模态大模型由图像级语义理解向精细化空间解译能力的进一步拓展。

在此基础上，多模态大模型逐渐展现出跨任务协同能力。EarthGPT通过统一多源遥感数据与语言指令，实现了场景分类、图像描述、区域描述、视觉问答、视觉定位和目标检测等任务的统一建模，体现出“一个模型支持多类任务”的基础能力（Zhang等，2024b，2025c）。与之相似，RingMoGPT通过构建大规模多任务指令调优数据，将目标检测、视觉问答、图像描述和变化描述等任务纳入同一框架，进一步验证了遥感大模型在多任务统一学习中的可行性（Wang等，2025）。此外，EarthMarker通过融合点、框等视觉提示与文本指令，实现了图像级、区域级及点级任务的统一解译，提升了多任务场景下统一交互框架的适应能力（Zhang等，2025d）。在模型结构层面，RS-MoE通过引入多专家协同结构与指令路由机制（Lin等，2025），将图像描述与视觉问答任务的建模过程分解为多个子任务，由不同专家分别处理，在两项任务上均取得了良好的泛化效果。这表明多任务协同正在由传统基于参数共享的统一模型，逐步演进为基于功能分工的结构化推理模式。

除此之外，这类模型已开始面向更复杂的应用场景，如灾害响应、环境监测与智能决策等。Remote Sensing ChatGPT模型通过引入任务规划机制，将复杂遥感需求分解为多个子任务并逐步执行，构建了“任务理解—任务分解—结果整合”的推理流程，反映出遥感智能解译正由单一模型调用向流程化推理与自主执行方向演进（Guo等，2024）。在变化理解领域，Liu等人（2024a）提出的Change-Agent进一步结合变化检测、变化描述、变化计数和变化原因分析等功能，实现了交互式综合变化解译，表明多模态大模型已具备处理复杂时序遥感任务的协同推理能力。相关研究通过构建交互式多任务指令数据体系，将变化描述、分类、计数与定位等任务纳入统一建模框架，促进了变化分析由单一结果输出向多任务协同理解的转变（Xue等，2026）。总体而言，多模态大模型正由面向单一功能的工具型模型，逐步发展为具备任务编排、交互推理与结果整合能力的智能代理框架。

2 面向航空遥感的关键挑战与思考

2.1 高分辨率场景下的建模压力

对于多模态大模型而言，这一问题并未因模型规模扩大而缓解，反而在计算与建模复杂度层面进一步放大。现有多模态模型通常依赖视觉编码器将图像划分为词元（token）进行建模，而航空遥感影像的高分辨率特性会导致token数量急剧增长，从而在显存占用、推理效率以及信息保真之间形成更加突出的矛盾。相关研究指出，遥感数据具有的高冗余性与多尺度结构特征，难以直接沿用自然图像的建模方法，需要通过分层表示、多尺度特征融合与区域选择机制进行适配（Chen等，2024）。因此，高分辨率虽然有数据层面的优势，但同时也成为制约模型设计与推理效率的重要因素。

此外，高分辨率建模的挑战不仅体现在计算数据量的负担上，还来源于信息表达过程中可能产生的失真问题。对于小目标而言，在下采样或图像切块过程中，其关键结构信息容易丢失，从而影响目标检测与识别精度；但如果模型过度依赖局部细节，有可能会削弱对全局语义信息的理解，使模型难以把握场景的整体结构和目标之间的关联关系。如何实现局部信息与全局上下文语义的平衡，已成为遥感视觉建模中的核心问题之一，并且该问题在小目标检测等任务中尤为突出（Al-Hababi等，2024）。因此，未来模型设计需重点关注多尺度特征建模、局部

与全局信息协同以及关键区域选择机制，以期在有限计算资源下获得高效且可靠的遥感解译能力。

2.2 多源异构与时空不一致带来的融合难题

航空遥感的多模态融合难题，其核心并不是模态数量的增加，而在于不同来源数据难以在同一语义空间中稳定对齐。正如高分辨率问题主要考验模型的视觉感知能力，多源异构问题更侧重于模型的跨模态对齐能力。实际任务中，航空遥感通常依赖可见光、近红外以及多光谱等机载传感数据，并常结合同一区域不同时段获取的观测结果开展联合分析。然而，由于不同传感器及不同时段数据之间存在辐射响应、成像几何和观测条件上的差异，往往会对后续目标识别与场景解译结果产生影响。已有多模态遥感图像匹配的相关研究指出，高精度融合的关键在于几何、辐射与结构特征之间的一致性表达，而非对多源数据的简单叠加（Liao 等，2024；Yang 等，2025b）。

与卫星平台相比，无人机（unmanned aerial vehicle，UAV）/低空平台观测更灵活，但也伴随更明显的视角扰动、覆盖范围变化和获取条件波动；而灾害监测、结构巡检以及作物长势分析等任务通常要求模型能够联合解析不同时间、不同平台和不同模态的数据（Bazrafkan 等，2025）。相关综述表明，UAV—卫星协同观测和时空融合在实际应用中展现出较大的潜力，但其实际效果高度依赖跨尺度、跨平台数据能否实现可靠匹配与统一解释（Ebrahimy 等，2025）；同时，时空融合方法虽然能够在一定程度上弥补单次观测信息不足的问题，但其仍然普遍受到时空错配、模型泛化能力不足以及复杂场景下稳定性的制约（Lian 等，2025；Xiao 等，2023）。

多模态大模型为这一问题提供了新的解决框架，即通过统一表征将影像、文本与任务指令关联起来，使多源融合从早期的特征拼接转向表征和任务协同。RemoteCLIP 模型指出遥感视觉—语言基础模型能够借助大规模图文对齐获得更强的语义迁移能力（Liu 等，2024b），EarthGPT-X 则进一步展示了其在面向高分辨率、多传感器场景的区域理解、视觉问答与多任务统一建模能力（Zhang 等，2025c）。然而，这些研究也从侧面反映出，当前高质量遥感多模态训练资源仍然不足，尽管模型已经具备一定的跨模态表达能力，却尚未真正充分掌握不同传感器之间的互补关系及时空一致性规律。在航空遥感场景中，未来研究应更多关注多源融合结果的可靠性与应用有效性，并加强跨模态配准、时空约束与任务导向数据基准等方面的研究。

2.3 通用大模型与航空领域知识之间的鸿沟

当前多模态大模型之所以备受关注，很大程度上在于其在统一接口下具备开放任务处理能力，即同一模型可支持图像描述、区域问答、目标定位以及一定程度的语义推理等任务。以 EarthGPT-X 为代表的大模型表明，通过多任务指令调优，模型能够在高分辨率遥感场景中同时完成图像理解、区域分析与多任务协同处理，并在少样本条件下表现出较强的泛化能力（Zhang 等，2025c）。与此同时，关于大视觉语言模型的研究普遍认为，其核心优势正逐渐由传统的单一识别性能，转向跨模态对齐、指令跟随以及复杂推理能力等更高层次的综合能力。

然而，在航空遥感领域，通用能力并不等同于领域理解能力。航空影像中的关键问题往往不仅是识别有什么，更在于判断空间关系、时空异常和任务相关风险，这类决策通常依赖测绘规则、空间先验、业务流程和领域知识。近期关于地理空间基础模型的研究已明确指出，地理空间人工智能面临的核心挑战并不仅是性能提升，更在于构建稳健、可解释且可靠的知识应用方式。相应地，已有研究也开始强调：现有遥感大模型在精确空间理解、多粒度视觉证据约束以及复杂场景下的可靠回答方面仍存在明显不足。

在航空遥感场景中，通用多模态大模型与实际领域需求之间的差距主要体现在多个层面。首先，现有模型在精细空间关系建模方面仍存在不足，对于邻近、相交或沿线分布等具有明确空间约束的关系，往往难以形成稳定且准确的表达；其次，针对具体业务任务的知识支撑仍比较有限，在基础设施巡检、灾后评估以及耕地监测等应用中，模型输出的结果虽然在语言上具备一定合理性，但缺乏与实际判读规则相一致的依据；此外，在复杂场景或分布外数据下，构建结果与真实观测之间的一致性仍难以保证。已有针对大规模视觉语言模型的综述将此类“幻觉”现象视为影响模型可靠性的关键因素之一（Liu 等，2024c）。因此，在航空遥感应用中，仅依赖通用多模态能力难以满足实际需求，更需要在模型设计与应用过程中引入空间约束机制、任务规则及专家知识，以提升模型在复杂场景中的判读一致性

与结果可信度。

2.4 面向真实任务部署的效率、可信与可解释性

从研究原型到实际业务应用，航空遥感大模型首先面临的重要挑战是部署效率的限制。在灾害响应、应急巡检及无人机自主作业等场景中，模型除需具备较高精度外，还需满足实时处理、算力受限环境适配及通信资源约束等要求。针对机载视觉任务的相关综述指出，真实系统中的瓶颈往往不仅来自检测算法本身，还涉及传感器配置、数据传输、边缘计算、系统集成与时延控制等完整工作流（Habash等，2025），这意味着离线实验中的性能优势，并不能直接转化为实际任务中的应用优势。对于多模态大模型而言，这一矛盾尤为突出，因为其通常具有更庞大的参数规模、更长的推理链路和更复杂的输入形式，因此在机载平台、边缘端或低带宽环境中的部署成本明显高于传统的单任务网络（Wu等，2025b）。

除部署效率外，可信性是航空遥感大模型能够融入实际决策流程的另一关键前提。航空遥感服务通常涉及基础设施监控、安全保障和灾害应对等领域，一旦模型输出应用于辅助决策，其评价标准不能仅停留在结果输出的流畅性层面。在可靠的地理空间基础模型研究框架下，模型的可靠性、责任界定以及可持续应用能力正逐步成为核心关注点。同时，大规模视觉语言模型中普遍存在的“幻觉”现象，也被视为影响其在实际应用中可信度的潜在风险因素（Wu等，2025b）。对于航空遥感而言，受领域专业性与应用决策敏感性的影响，模型可靠性问题更为突出。若模型缺乏不确定性表征、异常输入识别及分布外场景适应能力，其实际应用与推广将受到明显制约。

可解释性已不再是可有可无的附加属性，而是推动模型实际应用的关键前提。多模态大模型将视觉表征、跨模态对齐与语言生成过程紧密耦合，导致错误可能源于感知、检索、推理或生成等任一阶段。相比之下，航空遥感领域的实际应用更加注重协同系统的构建，不仅要求模型能够标注关键证据区域，还需要能够呈现出推理过程并表达结果的不确定性。EarthMarker模型通过融合视觉提示与语言指令的输入方式，增强了区域级证据的表达能力（Zhang等，2025d），这体现出遥感大模型在输出形式上正逐步由单一结果生成向可支持验证与分析的表达方式发展。总体来看，面向实际应用场景的航空遥感大模型仍需在轻量化部署、可信推理与可解释交互等方面进一步完善。一方面，需要通过轻量化模型设计与系统级优化降低边缘端部署成本；另一方面，应增强模型对异常输入与不确定性结果的识别能力，并结合人机协同机制提升输出可靠性。此外，还需构建具备可解释性的交互框架，以支持结果验证、过程追踪与误差修正。

3 讨论与总结

在大模型技术推动下，遥感解译正由传统视觉感知逐步向跨模态推理、语义理解与任务导向分析等更高层次演进。围绕这一发展过程，本文系统梳理了从单源视觉模型到多模态大模型的演变，重点分析了多模态融合、跨模态推理及大模型应用的关键进展。本文进一步指出，遥感智能解译能力的提升并非仅仅依赖于模型规模的扩展，多模态对齐、空间认知能力、任务可靠性及可部署性等关键问题，仍是制约技术突破与实际应用的核心瓶颈。

多模态对齐依然是限制遥感大模型性能提升的重要因素。航空遥感数据来源多样（如光学影像、SAR与激光雷达（light detection and ranging，LiDAR）数据），在几何结构、辐射特性以及获取时序方面存在显著差异，使跨模态语义对齐面临较大挑战（Geng等，2025；Zhang等，2024c）。尽管基于图文对齐的模型（如CLIP）在通用视觉任务中取得了一定进展，但在遥感场景中，仍受到高分辨率、复杂背景和噪声干扰，导致对齐精度下降（Lu等，2025）。未来的研究应重点提升跨传感器对齐的鲁棒性，从几何校正、辐射归一化到时空一致性建模等方面，发展更精细的多模态表征方法。

空间认知能力是实现高层次遥感解译与推理的重要基础。相较于传统基于像素或局部特征的建模方式，复杂遥感任务更依赖对地物结构关系、空间分布及动态变化的整体理解。为增强模型的空间推理能力，已有研究引入区域交互与空间问答等空间引导机制。GeoChat等模型进一步表明，遥感视觉语言模型能够在统一框架下支持空间约束对话、区域问答及目标指代定位等多类任务（Kuckreja等，2024）。然而，现有模型在大规模数据与复杂地理环境中的泛化能力仍有限。如何构建具备全局空间关系建模能力的统一框架，是提升任务感知与推理水平的重

要方向。

随着应用场景的拓展，任务可靠性成为多模态大模型落地的关键约束。在灾害评估与应急响应等高风险任务中，模型不仅需具备高精度，还需保证结果的可解释性与可信性。现有研究表明，大视觉语言模型普遍存在“幻觉”问题，而在遥感场景下，由于数据分布特殊、指令数据不足以及区域级语义复杂等因素，该问题可能更加突出（Kuckreja 等，2024；Li 等，2025a；Liu 等，2024c）。为此，需要通过引入领域知识、任务约束及针对性微调，增强模型推理过程的一致性，并建立面向遥感应用的专业评估体系，以提升整体稳定性与可靠性。

此外，模型的可部署性直接决定了其在实际应用中的价值。遥感数据的高分辨率和大幅面特性使得检测与解译模型在计算和存储上面临较大的压力，与此同时，无人机和边缘平台通常受到算力、能耗和时延的限制，这使得轻量化设计和高效推理成为当前研究的重点（Wu 等，2024）。近年来，面向 UAV/边缘设备的轻量化检测网络不断涌现，相关研究表明，在压缩模型参数和计算量的同时保持可接受的精度是可行的，但实时性、稳定性与复杂场景适应能力之间仍然存在明显权衡（Habash 等，2025）。因此，模型压缩、推理优化以及软硬件协同设计仍是推动多模态遥感模型实用化部署的关键研究方向。

4 展　望

未来航空遥感智能解译的发展仍将依赖多模态大模型能力的持续提升，但其关键已不仅在于模型规模的扩展，更在于复杂应用场景下的有效落地与协同优化。跨传感器异构数据的高精度对齐仍是多模态遥感建模的基础问题，需要进一步加强几何一致性约束与辐射差异建模，以提升多源数据融合质量。针对复杂地理场景，多尺度空间关系表征与跨尺度推理能力仍有待增强，以适应目标尺度变化显著和空间结构复杂的遥感任务需求。与此同时，生成式大模型的发展为小样本遥感解译提供了新的技术路径，有望缓解遥感领域标注数据稀缺的问题，并提升灾害应急等高时效场景下的任务适应能力。在实际应用过程中，模型的可信推理能力与可解释性至关重要，其直接关系到遥感辅助决策结果的可靠性与稳定性。工程应用层面，轻量化建模与边缘部署能力将成为制约大模型在无人机等平台应用的重要因素，而面向隐私保护的协同学习机制（如联邦学习）也将为多源数据共享与联合建模提供支撑。总体来看，航空遥感大模型仍处于由单源建模向多模态统一建模演进的关键阶段。未来，模型性能、系统效率与实际应用约束之间的协同优化能力，将在很大程度上决定其在航空遥感领域的应用深度与推广价值。

参考文献（References）

Al-Hababi M A M, Habib A, Thabit F and Liu Y. 2024. A novel preprocessing approach and benchmarking analysis for faster, robust, and improved small object detection methods. Remote Sensing, 16(20): #3753 [DOI: 10.3390/rs16203753]

Alvarez-Vanhard E, Corpetti T and Houet T. 2021. UAV and satellite synergies for optical remote sensing applications: a literature review. Science of Remote Sensing, 3: #100019 [DOI: 10.1016/j.srs.2021.100019]

Astruc G, Gonthier N, Mallet C and Landrieu L. 2025. AnySat: one earth observation model for many resolutions, scales, and modalities//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 19530-19540 [DOI: 10.1109/CVPR52734.2025.01819]

Bazi Y, Bashmal L, Al Rahhal M M, Ricci R and Melgani F. 2024. RS-LLaVA: a large vision-language model for joint captioning and question answering in remote sensing imagery. Remote Sensing, 16(9): #1477 [DOI: 10.3390/rs16091477]

Bazrafkan A, Igathinathane C, Bandillo N and Flores P. 2025. Optimizing integration techniques for UAS and satellite image data in precision agriculture — a review. Frontiers in Remote Sensing, 6: #1622884 [DOI: 10.3389/frsen.2025.1622884]

Cao Z, Kooistra L, Wang W S, Guo L F and Valente J. 2023. Real-time object detection based on UAV remote sensing: a systematic literature review. Drones, 7(10): #620 [DOI: 10.3390/drones7100620]

Chang B H, Li F, Hu Y C, Yin H, Feng Z H and Zhao L. 2025. Application of UAV remote sensing for vegetation identification: a review and meta-analysis. Frontiers in Plant Science, 16: #1452053 [DOI: 10.3389/fpls.2025.1452053]

Chen S Q, Yang X, Zhu R Q, Liao N and Zhao W W. 2026. Parameter-efficient fine-tuning for remote sensing image interpretation: a survey. Journal of Image and Graphics, 31(1): 212-242 (陈诗琪, 杨学, 朱荣强, 廖宁, 赵卫伟. 2026. 面向遥感图像解译的参数高效微调研究综述. 中国图象图形学报, 31(1): 212-242) [DOI: 10.11834/jig.250105]

Chen Z Y, Wang H Y, Wu X Y, Wang J, Lin X R, Wang C, et al. 2024. Object detection in aerial images using DOTA dataset: a survey. International Journal of Applied Earth Observation and Geoin-

formation, 134: #104208 [DOI: 10.1016/j.jag.2024.104208]

Cheng K D, Liu J L, Mao R, Wu Z L and Cambria E. 2026. CMCap: cross-modal meta-captioning for few-shot remote sensing image captioning. IEEE Geoscience and Remote Sensing Letters, 23: #6006705 [DOI: 10.1109/LGRS.2026.3674723]

Cheng Q M, Huang H Y, Xu Y, Zhou Y Z, Li H Y and Wang Z Y. 2022. NWPU-Captions dataset and MLCA-Net for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing, 60: #5629419 [DOI: 10.1109/TGRS.2022.3201474]

Diao W H, Yu H C, Kang K Y, Ling T, Liu D, Feng Y C, et al. 2024. RingMo-Aerial: an aerial remote sensing foundation model with affine transformation contrastive learning [EB/OL]. [2026-04-16]. https://arxiv.org/pdf/2409.13366.pdf

Ding Y L, Wang D, Li K, Zhao X H and Wang Y F. 2025. Visual grounding of remote sensing images with multi-dimensional semantic-guidance. Pattern Recognition Letters, 189: 85-91 [DOI: 10.1016/j.patrec.2025.01.013]

Ebrahimy H, Yu T and Zhang Z. 2025. Developing a spatiotemporal fusion framework for generating daily UAV images in agricultural areas using publicly available satellite data. ISPRS Journal of Photogrammetry and Remote Sensing, 220: 413-427 [DOI: 10.1016/j.isprsjprs.2024.12.024]

Fu Z Z, Yan H P and Ding K. 2025. CLIP-MoA: visual-language models with mixture of adapters for multitask remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 63: #4703817 [DOI: 10.1109/TGRS.2025.3565552]

Gao G S, Shang Y Q and Dong Y. 2025. Review of deep learning algorithms for small object detection in optical remote sensing images. Journal of Image and Graphics, 30(11): 3479-3505 (高广帅, 尚云琦, 董燕. 2025. 光学遥感图像小目标检测深度学习算法综述. 中国图象图形学报, 30(11): 3479-3505) [DOI: 10.11834/jig.240740]

Geng Z L, Liu H B, Duan P H, Wei X H and Li S T. 2025. Feature-based multimodal remote sensing image matching: benchmark and state-of-the-art. ISPRS Journal of Photogrammetry and Remote Sensing, 229: 285-302 [DOI: 10.1016/j.isprsjprs.2025.08.028]

Guo B H, Liu D H, Shen Z and Wang T B. 2026. FF-Mamba-YOLO: an SSM-Based benchmark for forest fire detection in UAV remote sensing images. Journal of Imaging, 12(1): #43 [DOI: 10.3390/jimaging12010043]

Guo H N, Su X, Wu C, Du B, Zhang L P and Li D R. 2024. Remote sensing ChatGPT: solving remote sensing tasks with ChatGPT and visual models//Proceedings of 2024 IGARSS IEEE International Geoscience and Remote Sensing Symposium. Athens, Greece: IEEE: 11474-11478 [DOI: 10.1109/IGARSS53475.2024.10640736]

Habash N, Alqumsan A A and Zhou T. 2025. Recent real-time aerial object detection approaches, performance, optimization, and efficient design trends for onboard performance: a survey. Sensors, 25(24): #7563 [DOI: 10.3390/s25247563]

Hu Y, Yuan J L, Wen C C, Lu X N, Liu Y and Li X. 2025. RSGPT: a remote sensing vision language model and benchmark. ISPRS Journal of Photogrammetry and Remote Sensing, 224: 272-286 [DOI: 10.1016/j.isprsjprs.2025.03.028]

Huang Z Y, Yan H X, Zhan Q Q, Yang S, Zhang M M, Zhang C K, et al. 2025. A survey on remote sensing foundation models: from vision to multimodality [EB/OL]. [2026-04-16]. https://arxiv.org/pdf/2503.22081.pdf

Ji J Y, Zhao Y F, Li A H, Ma X L, Wang C L and Lin Z L. 2025. Dense small object detection algorithm for unmanned aerial vehicle remote sensing images in complex backgrounds. Digital Signal Processing, 158: #104938 [DOI: 10.1016/j.dsp.2024.104938]

Jiang S, Jiang C and Jiang W S. 2020. Efficient structure from motion for large-scale UAV images: a review and a comparison of SfM tools. ISPRS Journal of Photogrammetry and Remote Sensing, 167: 230-251 [DOI: 10.1016/j.isprsjprs.2020.04.016]

Kazanskiy N, Khabibullin R, Nikonorov A and Khonina S. 2025. A comprehensive review of remote sensing and artificial intelligence integration: advances, applications, and challenges. Sensors, 25(19): #5965 [DOI: 10.3390/s25195965]

Kuckreja K, Danish M S, Naseer M, Das A, Khan S and Khan F S. 2024. GeoChat: grounded large vision-language model for remote sensing//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 27831-27840. [DOI: 10.1109/CVPR52733.2024.02629]

Li H D, Zhang X F and Qu H C. 2025a. DDFAV: remote sensing large vision language models dataset and evaluation benchmark. Remote Sensing, 17(4): #719 [DOI: 10.3390/rs17040719]

Li H F, Guo W, Wu M W, Peng C L, Zhu Q, Liu Y, et al. 2025. Visual-language joint representation and intelligent interpretation of remote sensing geo-objects: principles, challenges and opportunities. Acta Geodaetica et Cartographica Sinica, 54(5): 853-872 (李海峰, 郭旺, 吴梦伟, 彭程里, 朱庆, 刘瑜, 等. 2025. 视觉—语言联合的遥感地物概念表达与智能解译: 原理、挑战与机遇. 测绘学报, 54(5): 853-872) [DOI: 10.11947/j.AGCS.2025.20240244]

Li H F, Li Q J, Yang C, Guo W, Li M Y, Yuan H Y, et al. 2025b. Task knowledge injection: training-free adaptation of multimodal large language models for remote sensing image understanding. IEEE Geoscience and Remote Sensing Letters, 22: #6010605 [DOI: 10.1109/LGRS.2025.3581558]

Li J Y, Cai Y X, Li Q, Kou M Y and Zhang T X. 2024. A review of remote sensing image segmentation by deep learning methods. International Journal of Digital Earth, 17(1): #2328827 [DOI: 10.1080/17538947.2024.2328827]

Li L Z, Han L, Ye Y X, Xiang Y M and Zhang T Y. 2025c. Deep learning in remote sensing image matching: a survey. ISPRS Journal of Photogrammetry and Remote Sensing, 225: 88-112 [DOI: 10.1016/j.isprsjprs.2025.04.001]

Lian Z L, Zhan Y L, Zhang W H, Wang Z J, Liu W B and Huang X H.

2025. Recent advances in deep learning-based spatiotemporal fusion methods for remote sensing images. Sensors，25(4)：#1093［DOI：10.3390/s25041093］

Liao Y F，Xi K，Fu H J，Wei L，Li S，Xiong Q，et al. 2024. Refining multi-modal remote sensing image matching with repetitive feature optimization. International Journal of Applied Earth Observation and Geoinformation，134：#104186［DOI：10.1016/j. jag. 2024. 104186］

Lin H，Hong D F，Ge S H，Luo C Y，Jiang K，Jin H，et al. 2025. RS-MoE：a vision-language model with mixture of experts for remote sensing image captioning and visual question answering. IEEE Transactions on Geoscience and Remote Sensing，63：#5614918［DOI：10.1109/TGRS.2025.3547988］

Liu C Y，Chen K Y，Zhang H T，Qi Z P，Zou Z X and Shi Z W. 2024a. Change-Agent：toward interactive comprehensive remote sensing change interpretation and analysis. IEEE Transactions on Geoscience and Remote Sensing，62：#5635616［DOI：10.1109/TGRS. 2024.3425815］

Liu C Y，Zhao R，Chen H，Zou Z X and Shi Z W. 2022. Remote sensing image change captioning with dual-branch transformers：a new method and a large scale dataset. IEEE Transactions on Geoscience and Remote Sensing，60：#5633520［DOI：10.1109/TGRS. 2022. 3218921］

Liu F，Chen D L，Guan Z Q Y，Zhou X C，Zhu J L，Ye Q L，et al. 2024b. RemoteCLIP：a vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing，62：#5622216［DOI：10.1109/TGRS.2024.3390838］

Liu H C，Xue W Y，Chen Y F，Chen D P，Zhao X T，Wang K，et al. 2024c. A survey on hallucination in large vision-language models［EB/OL］.［2026-04-16］. https://arxiv.org/pdf/2402.00253.pdf

Liu H Q，Long Q，Yi B and Jiang W. 2025. A survey of sensors based autonomous unmanned aerial vehicle（UAV）localization techniques. Complex and Intelligent Systems，11(8)：#371［DOI：10. 1007/s40747-025-01961-2］

Lobry S，Marcos D，Murray J and Tuia D. 2020. RSVQA：visual question answering for remote sensing data. IEEE Transactions on Geoscience and Remote Sensing，58(12)：8555-8566［DOI：10.1109/TGRS.2020.2988782］

Lu Q Q，Xie Y X，Zhang J，Guo Y M，Wei Y M，Jiang J，et al. 2025. CLIP-Driven with dynamic feature selection and alignment network for referring remote sensing image segmentation. Remote Sensing，17(22)：#3675［DOI：10.3390/rs17223675］

Lu X Q，Wang B Q，Zheng X T and Li X L. 2018. Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing，56(4)：2183-2195［DOI：10.1109/tgrs.2017.2776321］

Ma L，Liu Y，Zhang X L，Ye Y X，Yin G F and Johnson B A. 2019. Deep learning in remote sensing applications：a meta-analysis and review. ISPRS Journal of Photogrammetry and Remote Sensing，152：166-177［DOI：10.1016/j.isprsjprs.2019.04.015］

Pan J C，Liu Y X，Fu Y Q，Ma M Y，Li J H，Paudel D P，et al. 2024. Locate anything on earth：advancing open-vocabulary object detection for remote sensing community［EB/OL］.［2026-04-16］. https://arxiv.org/pdf/2408.09110.pdf

Pan Z Z，Gao F，Gong C Z，Gan Y H and Dong J Y. 2026. Remote sensing image semantic segmentation with selective attention and directional feature enhancement. Journal of Image and Graphics，31(4)：1272-1284（潘子哲，高峰，宫传政，甘言海，董军宇. 2026. 选择注意力与方向特征增强的遥感图像语义分割. 中国图象图形学报，31(4)：1272-1284）［DOI：10.11834/jig.250317］

Qiu J T，Chang W，Ren W，Hou S S and Yang R. 2025. MMFNet：a mamba-based multimodal fusion network for remote sensing image semantic segmentation. Sensors，25(19)：#6225［DOI：10.3390/s25196225］

Radford A，Kim J W，Hallacy C，Ramesh A，Goh G，Agarwal S，et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning.［s.l.］：PMLR：8748-8763

Rahnemoonfar M，Chowdhury T，Sarkar A，Varshney D，Yari M and Murphy R R. 2021. FloodNet：a high resolution aerial imagery dataset for post flood scene understanding. IEEE Access，9：89644-89654［DOI：10.1109/ACCESS.2021.3090981］

Reed C J，Gupta R，Li S F，Brockman S，Funk C，Clipp B，et al. 2023. Scale-MAE：a scale-aware masked autoencoder for multiscale geospatial representation learning//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris，France：IEEE［DOI：10.1109/ICCV51070.2023.00378］

Si D C，Wang D，Gao E Z，Qin X L，Zhao L，Zhang J，et al. 2026. SPEX：a vision-language model for land cover extraction on spectral remote sensing images. IEEE Transactions on Geoscience and Remote Sensing，64：#5403416［DOI：10.1109/TGRS. 2026. 3670308］

Sun Y，Cheng Q X，Xie W J，Huang H X and Gu C C. 2026. Low-data cross-modal adaptation for remote sensing with proxy-enhanced multi-granularity feature caching. Scientific Reports，16（1）：#10895［DOI：10.1038/s41598-026-39823-7］

Sun Y，Wang D H，Li L，Ning R S，Yu S L and Gao N Y. 2024. Application of remote sensing technology in water quality monitoring：from traditional approaches to artificial intelligence. Water Research，267：#122546［DOI：10.1016/j.watres.2024.122546］

Tao C，Guo X，Hu K Y，Shen Y X and Wang H. 2025. Language-guided cross-spatiotemporal domain adaptation for remote sensing image semantic segmentation. Journal of Image and Graphics，30(9)：3153-3170（陶超，郭鑫，胡柯彦，沈羽翔，王昊. 2025. 以语言为媒介的遥感图像跨时空领域自适应语义分割. 中国图象图形学报，30(9)：3153-3170）［DOI：10.11834/jig.240640］

Tao L J，Zhang H K，Jing H Z，Liu Y，Yan D W，Wei G T，et al. 2025. Advancements in vision-language models for remote sensing：datasets，capabilities，and enhancement techniques. Remote Sensing，17(1)：#162［DOI：10.3390/rs17010162］

Tuia D, Bazi Y, Demir B, Jacobs N and Lobry S. 2026. Editorial to the special issue in vision language models for remote sensing analysis and interpretation. ISPRS Journal of Photogrammetry and Remote Sensing, 231: 307-309 [DOI: 10.1016/j.isprsjprs.2025.10.001]

Wang C M, Fan G D, Li J J, Gan M and Chen C L P. 2026. MGCR-Net: multimodal graph-conditioned vision-language reconstruction network for remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing, 64: #4701515 [DOI: 10.1109/TGRS.2026.3654629]

Wang D, Zhang J, Xu M Q, Liu L, Wang D S, Gao E Z, et al. 2024a. MTP: advancing remote sensing foundation model via multitask pretraining. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17: 11632-11654 [DOI: 10.1109/jstars.2024.3408154]

Wang G H, Li B, Zhang T and Zhang S B. 2022a. A network combining a transformer and a convolutional neural network for remote sensing image change detection. Remote Sensing, 14(9): #2228 [DOI: 10.3390/rs14092228]

Wang P J, Hu H Y, Tong B Y, Zhang Z Q, Yao F L, Feng Y C, et al. 2025. RingMoGPT: a unified remote sensing foundation model for vision, language, and grounded tasks. IEEE Transactions on Geoscience and Remote Sensing, 63: #5611320 [DOI: 10.1109/TGRS.2024.3510833]

Wang W Q, Chen Y S and Ghamisi P. 2022b. Transferring CNN with adaptive learning for remote sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing, 60: #5533918 [DOI: 10.1109/tgrs.2022.3190934]

Wang Y Q, Chen W T, Han X T, Lin X D, Zhao H T, Liu Y F, et al. 2024b. Exploring the reasoning abilities of multimodal large language models (MLLMs): a comprehensive survey on emerging trends in multimodal reasoning [EB/OL]. [2026-04-16]. https://arxiv.org/pdf/2401.06805.pdf

Wu K, Zhang Y Y, Ru L X, Dang B, Lao J W, Yu L, et al. 2025a. A semantic-enhanced multi-modal remote sensing foundation model for Earth observation. Nature Machine Intelligence, 7(8): 1235-1249 [DOI: 10.1038/s42256-025-01078-8]

Wu W N, Liu A, Hu J W, Mo Y, Xiang S, Duan P H, et al. 2024. EUAVDet: an efficient and lightweight object detector for UAV aerial images with an edge-based computing platform. Drones, 8(6): #261 [DOI: 10.3390/drones8060261]

Wu Y M, Mu X F, Shi H and Hou M X. 2025b. An object detection model AAPW-YOLO for UAV remote sensing images based on adaptive convolution and reconstructed feature fusion. Scientific Reports, 15(1): #16214 [DOI: 10.1038/s41598-025-00239-4]

Xiao A R, Xuan W H, Wang J J, Huang J X, Tao D C, Lu S J, et al. 2025. Foundation models for remote sensing and earth observation: a survey. IEEE Geoscience and Remote Sensing Magazine, 13(4): 297-324 [DOI: 10.1109/mgrs.2025.3576766]

Xiao J, Aggarwal A K, Duc N H, Arya A, Rage U K and Avtar R. 2023. A review of remote sensing image spatiotemporal fusion: challenges, applications and recent trends. Remote Sensing Applications: Society and Environment, 32: #101005 [DOI: 10.1016/j.rsase.2023.101005]

Xing Y, Liu X N and Wang X K. 2026. Integrating UAVs, satellite remote sensing, and machine learning in precision agriculture: pathways to sustainable food production, resource efficiency, and scalable innovation. Frontiers in Agronomy, 7: #1670380 [DOI: 10.3389/fagro.2025.1670380]

Xu H, Lu C, Zhang H, Shao Z F, Liu G C and Ma J Y. 2025a. Artificial intelligence-assisted remote sensing observation, understanding, and decision. The Innovation, 6(12): #101016 [DOI: 10.1016/j.xinn.2025.101016]

Xu L J and Ma H. 2026. Dual-path adaptive feature elevation system for detecting small targets in remote sensing imagery. Engineering Applications of Artificial Intelligence, 169: #114132 [DOI: 10.1016/j.engappai.2026.114132]

Xu L X, Wang L Y, Zhang J Z, Ha D and Zhang H S. 2025b. A review of cross-modal image-text retrieval in remote sensing. Remote Sensing, 17(24): #3995 [DOI: 10.3390/rs17243995]

Xu W J, Yu R Q, Xue M H, Wang X Y, Zhang Y B, Wei Z W, et al. 2026. A survey on earth observation multimodal large language models: framework, core technologies, and future perspectives. Journal of Radars, 15(1): 361-386 (许文嘉, 于睿卿, 薛铭浩, 汪雪怡, 张源奔, 魏智威, 等. 2026. 多模态对地观测大模型: 架构、关键技术和未来展望. 雷达学报(中英文), 15(1): 361-386) [DOI: 10.12000/JR25088]

Xue J X, Deng Q, Wu X C, Yao K L, Yin X Y, Yu F, et al. 2026. Toward comprehensive interactive change understanding in remote sensing: a large-scale dataset and dual-granularity enhanced VLM. IEEE Transactions on Geoscience and Remote Sensing, 64: #4401516 [DOI: 10.1109/TGRS.2025.3650151]

Yang B H, Chen Y S and Ghamisi P. 2024a. LVM-StARS: large vision model soft adaption for remote sensing scene classification. IEEE Geoscience and Remote Sensing Letters, 21: #6013905 [DOI: 10.1109/LGRS.2024.3432069]

Yang C, Li Z C and Zhang L F. 2024b. Bootstrapping interactive image-text alignment for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing, 62: #5607512 [DOI: 10.1109/TGRS.2024.3359316]

Yang C, Zhang J, Li Q J, Guo W and Li H F. 2026. RS_DeepReason: LLM-Driven deep reasoning for multigranularity remote sensing scene interpretation. IEEE Geoscience and Remote Sensing Letters, 23: #6005305 [DOI: 10.1109/LGRS.2026.3663856]

Yang L L, Chen N, Yue J, Liu Y D, Ma J Y, Ghamisi P, et al. 2025a. Survey of multimodal geospatial foundation models: techniques, applications, and challenges [EB/OL]. [2026-04-16]. https://arxiv.org/pdf/2510.22964.pdf

Yang Y, Liu S, Zhang H T, Li D C and Ma L. 2025b. Multi-modal remote sensing image registration method combining scale-invariant feature transform with co-occurrence filter and histogram of ori-

ented gradients features. Remote Sensing, 17(13): #2246 [DOI: 10.3390/rs17132246]

Yuan Z Q, Zhang W K, Fu K, Li X, Deng C B, Wang H Q, et al. 2022. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. IEEE Transactions on Geoscience and Remote Sensing, 60: #4404119 [DOI: 10.1109/TGRS.2021.3078451]

Zhan Y, Xiong Z T and Yuan Y. 2023. RSVG: exploring data and models for visual grounding on remote sensing data. IEEE Transactions on Geoscience and Remote Sensing, 61: #5604513 [DOI: 10.1109/TGRS.2023.3250471]

Zhan Y, Xiong Z T and Yuan Y. 2025. SkyEyeGPT: unifying remote sensing vision-language tasks via instruction tuning with large language model. ISPRS Journal of Photogrammetry and Remote Sensing, 221: 64-77 [DOI: 10.1016/j.isprsjprs.2025.01.020]

Zhang C, Ren Z L, Hou B, Xu C H, Meng J H, Li W B, et al. 2025a. Adaptive scale-aware semantic memory network for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing, 63: #5653418 [DOI: 10.1109/TGRS.2025.3636596]

Zhang H D, Wang L Q, Tian T and Yin J H. 2021. A review of unmanned aerial vehicle low-altitude remote sensing (UAV-LARS) use in agricultural monitoring in China. Remote Sensing, 13(6): #1221 [DOI: 10.3390/rs13061221]

Zhang S Y, Shan L L and Qiu R H. 2025b. Multimodal interpretation of remote sensing images: dynamic resolution input strategy and multi-scale vision-language alignment mechanism [EB/OL]. [2026-04-16]. https://arxiv.org/pdf/2512.23243.pdf

Zhang W, Cai M X, Ning Y Q, Zhang T, Zhuang Y, Lu S J, et al. 2025c. EarthGPT-X: a spatial MLLM for multilevel multisource remote sensing imagery understanding with visual prompting. IEEE Transactions on Geoscience and Remote Sensing, 63: #4709221 [DOI: 10.1109/TGRS.2025.3626941]

Zhang W, Cai M X, Zhang T, Lei G Q, Zhuang Y and Mao X R. 2024a. Popeye: a unified visual-language model for multisource ship detection from remote sensing imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17: 20050-20063 [DOI: 10.1109/JSTARS.2024.3488034]

Zhang W, Cai M X, Zhang T, Zhuang Y, Li J and Mao X R. 2025d. EarthMarker: a visual prompting multimodal large language model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 63: #5604219 [DOI: 10.1109/TGRS.2024.3523505]

Zhang W, Cai M X, Zhang T, Zhuang Y and Mao X R. 2024b. EarthGPT: a universal multimodal large language model for multisensor image comprehension in remote sensing domain. IEEE Transactions on Geoscience and Remote Sensing, 62: #5917820 [DOI: 10.1109/TGRS.2024.3409624]

Zhang X, Fang J B, Ding Z M, Yuan J, Liu X, Zhang Q J, et al. 2026. Cross-modal context-aware learning for visual-prompt-guided multimodal image understanding in remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 64: #5613614 [DOI: 10.1109/TGRS.2026.3673739]

Zhang Y J, Zhang W F, Yao Y X, Zheng Z, Wan Y and Xiong M T. 2024c. Robust registration of multi-modal remote sensing images based on multi-dimensional oriented self-similarity features. International Journal of Applied Earth Observation and Geoinformation, 127: #103639 [DOI: 10.1016/j.jag.2023.103639]

Zhang Z X and Zhu L X. 2023. A review on unmanned aerial vehicle remote sensing: platforms, sensors, data processing methods, and applications. Drones, 7(6): #398 [DOI: 10.3390/drones7060398]

Zhao K and Xiong W. 2024. Exploring region features in remote sensing image captioning. International Journal of Applied Earth Observation and Geoinformation, 127: #103672 [DOI: 10.1016/j.jag.2024.103672]

Zhou G Q, Qian L H and Gamba P. 2025. Advances on multimodal remote sensing foundation models for earth observation downstream tasks: a survey. Remote Sensing, 17(21): #3532 [DOI: 10.3390/rs17213532]

Zhu J, Li J Y, Li S Y, Yang Y F, Xing J C, Shen Y, et al. 2025. PMNET: fine-grained extraction of tidal flat farmland based on sandglass module codec for remote sensing fundamental model. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18: 28771-28781 [DOI: 10.1109/JSTARS.2025.3625585]

Zhu X G and Zhang Z Z. 2025. Efficient vision transformers with edge enhancement for robust small target detection in drone-based remote sensing. Frontiers in Remote Sensing, 6: #1599099 [DOI: 10.3389/frsen.2025.1599099]

Zhu X X, Tuia D, Mou L C, Xia G S, Zhang L P, Xu F, et al. 2017. Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine, 5(4): 8-36 [DOI: 10.1109/MGRS.2017.2762307]

作者简介

郎晋伟，男，博士，工程师，主要研究方向为航空计算机视觉、飞行视觉引导。E-mail: langjw@mail.ustc.edu.cn

李娅星，女，工程师，主要研究方向为飞控计算机设计和飞行视觉引导。E-mail: liyx174@avic.com

刘帅，男，研究员，主要研究方向为嵌入式高可靠智能计算和飞行器管理计算平台。E-mail: liushuaijiayou@126.com

康晓东，男，硕士，高级工程师，主要研究方向为高可靠多余度容错技术。E-mail: kangxd@avic.com

程俊强，男，研究员，主要研究方向为容错计算技术。E-mail: junstrong@hotmail.com