中图分类号: 学校代码: 10055
UDC: 密级: 公开

南開大學

博 士 学 位 论 文

面向复杂遥感场景的多模态感知基础模型研究

Research on Multimodal Perception Foundation Models for Complex Remote Sensing Scenes

论文作者 李宇轩 指导教师 杨健 教授
申请学位 工学博士 培养单位 计算机学院
学科专业 计算机科学与技术 研究方向 计算机视觉
答辩委员会主席 林宙辰 教授 评 阅 人 匿名评阅

南开大学研究生院

二〇二六年五月

摘要

遥感技术是获取地表空间信息、支撑资源调查、环境监测、灾害响应与国家安全保障的重要手段。随着高分辨率成像载荷、多平台协同观测系统和数据处理基础设施的发展，遥感数据呈现出高分辨率、多时相、多模态并行增长的趋势。与自然图像相比，复杂遥感场景同时具有俯视成像、小目标密集分布、尺度跨度大、背景组织复杂、成像机理异构以及高质量标注稀缺等特点，使得现有视觉骨干、下游任务框架和预训练范式难以直接适配。

围绕“面向复杂遥感场景的多模态感知基础模型”这一核心问题，本文按照“感知基础能力、数据与迁移、统一架构、语言驱动预训练与对齐”的递进路线开展研究。本文从复杂遥感数据特点出发，逐步回答空间先验如何建模、专业数据与跨域迁移如何支撑、异构模态如何统一以及语言语义如何参与基础模型预训练等关键问题。主要贡献如下：

1. 针对遥感目标对广域上下文的依赖及传统骨干感受野固定的问题，提出选择性大核网络 LSKNet。该方法通过大核分解与动态空间选择机制，在可控计算开销下实现自适应上下文建模，为目标检测、场景分类、语义分割和变化检测等遥感任务提供更适配的骨干表示。
2. 针对合成孔径雷达（SAR）目标检测长期存在的数据规模小、类别覆盖有限和跨域迁移困难等问题，构建大规模多类别 SAR 检测基准 SARDet-100K，并提出多阶段滤波增强预训练框架 MSFA。该框架从输入表征、领域过渡和检测器迁移三个层面缩小自然图像到 SAR 图像的迁移差距。
3. 针对 RGB、SAR、红外等异构传感器长期分离建模带来的知识割裂、任务碎片化和部署冗余问题，提出统一多模态检测框架 SM3Det。该框架通过混合专家机制兼顾共享知识与模态特定表示，并通过动态子模块优化缓解联合训练中的收敛失衡。
4. 针对传统“自下而上”预训练难以充分利用高层语义监督、异构多模态后期对齐容易引发优化冲突的问题，提出语言引导的遥感基础模型预训练与跨模态对齐方法。首先，视觉指令预训练范式 ViTP 借助视觉语言模型的高层语义监督反向塑造遥感视觉骨干；进一步地，BabelRS 通过概念共享

指令对齐与层级视觉语义退火，将跨模态语义统一前移到预训练阶段，从而改善异构多模态联合优化的稳定性。

实验结果表明，本文提出的方法在旋转目标检测、SAR 目标检测、异构多模态检测及相关遥感视觉基准上取得了较优性能。总体而言，本文围绕动态空间先验建模、大规模专业基准建设、统一多模态检测以及语言驱动的预训练与对齐机制，系统推进了复杂遥感场景多模态感知基础模型研究，为构建具备表征能力、跨域迁移能力与跨模态统一能力的新一代遥感智能感知系统提供了方法参考。

关键词：复杂遥感场景；多模态感知基础模型；统一多模态检测；语言引导对齐

Abstract

Remote sensing (RS) technology is a fundamental means of acquiring spatial information for resource management, environmental monitoring, disaster response, and national security. With the rapid development of high resolution sensors and multiplatform observation systems, RS data have entered an era of large scale, high resolution, and multimodal coexistence. Compared with natural images, complex RS scenes exhibit overhead imaging geometry, dense small objects, large scale variation, complex spatial layouts, heterogeneous imaging mechanisms, and limited high quality annotations. These characteristics make existing visual backbones, downstream architectures, and pretraining paradigms difficult to transfer directly to RS scenarios.

Focusing on multimodal perception foundation models for complex RS scenes, this dissertation follows a progressive route from perception capability, data and transfer, unified architecture, to language-driven pretraining and alignment. This dissertation starts from the intrinsic characteristics of complex RS data and studies how to model spatial priors, build data and transfer foundations, unify heterogeneous modalities, and use language semantics in foundation-model pretraining. The main contributions are summarized as follows:

1. Spatial-prior-aware backbone design: LSKNet, a Large Selective Kernel Network, is proposed to address the strong dependence of RS targets on wide-range contextual cues. Through large-kernel decomposition and dynamic spatial selection, LSKNet provides adaptive context modeling for RS detection, classification, segmentation, and change detection.
2. SAR benchmark construction and domain transfer: SARDet-100K, a large scale multiclass SAR detection benchmark, is constructed to alleviate the long-standing limitations of SAR data resources. Based on this benchmark, the MSFA pretraining framework is proposed to bridge the gap between natural-image pretraining and SAR detection through input adaptation, domain transition, and detector transfer.

3. Unified heterogeneous multimodal detection: SM3Det is developed to overcome the siloed modeling of RGB, SAR, and infrared data. By introducing a mixture-of-experts design and dynamic sub-module optimization, SM3Det jointly models shared knowledge and modality-specific characteristics within a unified detection framework.
4. Language-driven RS foundation-model pretraining and alignment: ViTP is introduced to use high-level semantic supervision from vision-language models to guide RS visual backbone learning in a top-down manner. Building on this direction, BabelRS moves heterogeneous cross-modal alignment from downstream fine-tuning to the pretraining stage through concept-sharing instruction alignment and hierarchical visual-semantic annealing, improving the stability of multimodal optimization.

Experimental results show that the proposed methods achieve competitive performance on oriented object detection, SAR object detection, heterogeneous multimodal detection, and related RS benchmarks. Overall, this dissertation advances multimodal perception foundation models for complex RS scenes and provides practical methodologies for next-generation intelligent Earth observation systems.

Key Words: Complex Remote Sensing; Multimodal Perception Foundation Models; Unified Multimodal Detection; Language-guided Alignment

目录

第一章 绪论

第一节 研究背景与意义

一、研究背景

遥感技术是国家空间信息基础设施的重要组成部分，也是资源调查、生态环境监测、灾害预警与评估、海洋监管、城市治理和国防安全的重要技术支撑。随着卫星、航空平台、无人机系统和地面接收处理体系的持续发展，对地观测正在由“低频、单源、离线解译”逐步走向“高频、多源、智能处理”的新阶段。光学、合成孔径雷达（Synthetic Aperture Radar, SAR）、红外、多光谱乃至高光谱等多种传感器，可以从反射、散射和热辐射等不同物理维度刻画地表目标，为复杂场景感知提供互补而丰富的信息来源。尤其在海上目标监视、灾后快速评估、城市扩张分析、交通调度和军事侦察等任务中，多平台多模态协同观测已经成为提升感知鲁棒性与时效性的现实需求[1-6]。

作为遥感图像解译中的核心问题之一，目标检测承担着从大范围复杂背景中定位并识别车辆、船舶、飞机、桥梁等目标的任务，并与场景分类、语义分割、变化检测等任务共同构成遥感智能解译的关键能力链条[7-12]。如图 1.1和 1.2所示，相比自然场景图像，遥感影像通常具有俯视成像、小目标密集分布、尺度跨度大、背景组织复杂以及模态间物理差异明显等特点，这使得目标识别往往不仅依赖目标局部外观，更依赖道路网络、港区岸线、机场布局等大范围空间上下文。与此同时，实际业务系统还往往要求模型能够跨区域、跨时相、跨传感器稳定工作，因此单一数据源、单一任务、单一模型的研究范式越来越难以满足开放环境下的应用需求。

从研究范式的演进看，复杂遥感场景感知大致经历了三个相互衔接的发展阶段。第一阶段主要面向特定任务的人工特征阶段，研究者依赖纹理、边缘、和散射机理等显式先验，设计 HOG、Canny、Haar-like、小波特征等描述子，并结合浅层分类器或专家规则完成目标识别与场景分析[13-16]。这一路线在样本有限、任务边界清晰的条件下具有一定有效性，但其表达能力和泛化能力高度依赖人工经验，难以适应复杂场景下尺度变化、背景干扰和跨域迁移等问题。

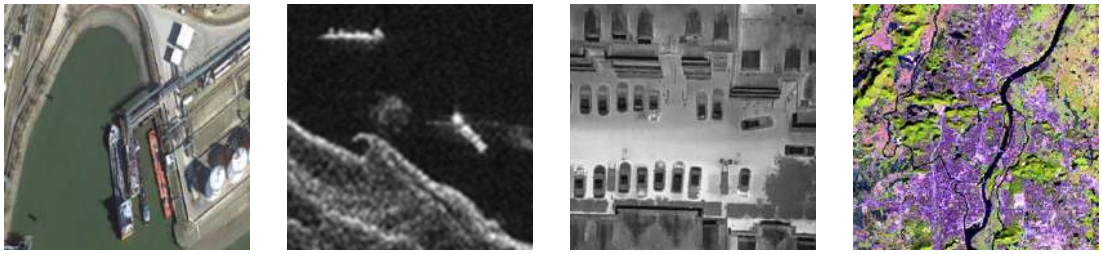


图 1.1 遥感图像包含光学、合成孔径雷达（SAR）、红外、多光谱等多种模态传感器，不同模态间的成像特征差异明显。

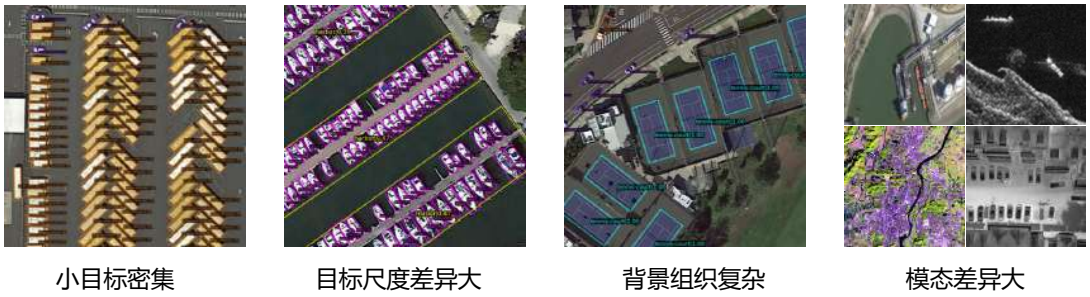


图 1.2 遥感影像通常具有俯视成像、小目标密集分布、目标尺度跨度大、背景组织复杂以及模态间物理差异明显等特点。

第二阶段是以深度学习为代表的端到端表征学习阶段。随着卷积神经网络和大规模数据驱动方法的发展，遥感视觉研究逐步由人工设计特征转向自动表征学习。Faster R-CNN、RetinaNet、FCOS、DETR 等通用检测框架，以及 Oriented R-CNN、R3Det、S^2A Net 等遥感定向检测方法，持续推动了遥感目标检测的发展[7-8,17-21]。与此同时，UNetFormer、ChangeFormer、BIT 等模型则分别在语义分割与变化检测中强化了多尺度交互和长程依赖建模[10-12]。这一阶段的核心贡献，在于将遥感解译由“局部特征匹配”推进到“任务驱动的深层特征学习”，也为跨任务共享骨干和统一优化奠定了基础。

第三阶段则表现为由大规模预训练、跨模态融合与基础模型驱动的统一建模阶段。在自然图像领域，以 ResNet、ViT、ImageNet 和 COCO 为代表的模型和数据体系，以及 MAE、DINOv2、CLIP 等预训练范式，已经形成了较为成熟的“骨干网络 + 任务头 + 预训练”技术框架[22-28]。受此推动，遥感领域也开始探索 RingMo、SatMAE、Scale-MAE、RemoteCLIP、SkySense 等面向专业场景的预训练与跨模态学习方法[29-33]。与此同时，LLaVA、InternVL、ImageBind、LanguageBind 等视觉语言与跨模态对齐方法进一步展示了高层语义监督在统一

异构表征方面的潜力[34-37]。这表明，遥感感知研究正在从“为单一任务设计专门模型”迈向“为一类复杂场景构建基础能力”的新阶段。

然而，遥感场景并不能简单复制自然图像领域已经形成的技术路径。一方面，俯视视角下的空间组织规律、小目标密集分布和多尺度长程依赖，使得自然图像中常见的局部纹理主导型骨干难以充分建模遥感空间先验。另一方面，SAR、光学、红外等模态之间在成像机理、统计分布和标注形态上的明显差异，又使统一表示学习与跨模态迁移远比常规视觉任务更为困难。再加上专业领域高质量标注稀缺、严格空间配对数据有限以及业务部署对稳定性和效率的双重要求，如何构建兼顾空间先验建模、跨模态知识融合和高层语义理解能力的遥感多模态感知基础模型，已成为复杂遥感场景智能解译亟待回答的关键问题。

二、研究意义

面向复杂遥感场景构建兼具空间解析、跨模态协同与高层语义建模能力的感知基础模型，既是遥感智能解译由任务定制化走向能力基础化的内在要求，也是支撑模型高效迁移、统一建模与稳定部署的重要前提。围绕这一目标，本文从单模态骨干设计、大规模 SAR 基准建设、异构多模态统一检测，到语言驱动的预训练与跨模态对齐，开展了系统性的研究。其学术意义主要体现在两个方面。

一方面，本文围绕复杂遥感场景中“空间先验如何建模、异构模态如何统一、语义监督如何反向塑造感知”这些基础问题，提出了一套相互衔接的技术路线。相关研究不仅拓展了遥感视觉骨干、统一检测架构与预训练范式的设计思路，也为理解语言语义如何服务于专业领域感知提供了新的方法视角。

另一方面，本文构建了包含超 11 万张图像的大规模多类别 SAR 目标检测基准及配套工具，为相关算法研究提供了更系统的数据基础。同时，本文提出的轻量化骨干网络、统一多模态检测架构以及语言引导的预训练机制，在提升性能的同时兼顾了模型泛化、迁移与部署效率，为面向复杂遥感场景的新一代智能感知系统提供了理论依据与方法储备。

第二节　研究现状与挑战

近年来，围绕复杂遥感场景感知，国内外研究从目标检测、SAR 解译、多源融合、领域化预训练和视觉语言对齐等方向持续推进，使遥感智能解译逐步由“单任务性能优化”转向“基础能力统一建模”。国内学者也围绕光学遥感目

表 1.1 复杂遥感场景特征、核心挑战与本文方法的对应关系。

复杂场景特征	对应挑战	本文研究方法
小目标密集、背景组织复杂	目标判别依赖广域上下文，固定感受野骨干难以按需建模空间先验	LSKNet 通过大核分解与空间选择机制实现自适应广域上下文建模
尺度跨度大、朝向任意	遥感检测、分割和变化检测均需要兼顾局部细节与大范围空间关系	LSKNet 在多类遥感下游任务中作为统一骨干验证空间先验建模能力
SAR 成像机理特殊、标注稀缺	小规模数据集和自然图像到 SAR 图像的域差异限制模型泛化	SARDet-100K 提供大规模标准化基准，MSFA 通过滤波增强和多阶段迁移缩小跨域差距
RGB、SAR、红外模态异构	多传感器数据难以用单一模型统一处理，分离建模造成知识割裂和部署冗余	SM3Det 以网格级 MoE 和动态子模块优化实现异构多模态统一检测
高质量标注有限、任务需求细粒度	自下而上预训练难以充分利用高层任务语义，视觉骨干与下游感知需求存在错位	ViTP 借助视觉指令目标将语言理解监督反向注入遥感视觉骨干
跨模态对齐缺少严格配对样本，跨模态特征差异大	后期对齐在微调阶段耦合模态对齐和检测优化，容易导致训练不稳定	BabelRS 以语言为语义枢纽，在预训练阶段完成异构模态早期对齐

标检测、检测数据集、旋转框检测、光学-SAR 融合、多源遥感信息融合、SAR 目标检测识别以及遥感基础模型等主题进行了系统总结[38-44]，为本文的问题凝练提供了重要参考。总体来看，现有研究已在局部任务上取得丰富进展，但关键能力仍分散于不同任务、不同模态和不同训练阶段中，尚未形成一条面向复杂遥感场景基础模型的递进技术路线。

从本文的研究主线出发，所要解决的并非若干孤立子任务的局部改进问题，而是在复杂空间组织、异构成像机制和有限监督条件并存的现实约束下，建立一套可迁移、可统一、可扩展的基础感知能力。表 1.1进一步概括了复杂遥感场景特征、核心挑战与本文方法之间的对应关系。本文各章围绕复杂数据特点逐层推进：先补强单模态空间感知和 SAR 数据迁移基础，再走向异构多模态统一架构，最后以语言和高层语义组织基础模型预训练。

挑战一：感知层上，空间先验建模不足与单模态数据基础薄弱。光学遥感目标检测和场景理解长期关注小目标密集、方向任意、尺度变化大和背景复杂等问题。旋转框检测、特征金字塔、多尺度融合、上下文注意力等方法不断提升了具体任务性能[7-8,10,21,45]，国内综述也指出，遥感目标检测的关键难点不仅在检测头设计，还在于如何理解高分辨率场景中的空间结构与目标上下文[38,40]。

然而，许多方法仍沿用自然图像骨干的固定感受野或局部纹理主导设计，难以根据目标类别、尺度和背景关系动态调节上下文范围。另一方面，SAR 目标检测由于成像机制复杂、斑点噪声强、纹理线索弱且标注成本高，长期受限于小规模、类别单一和评测协议不统一的数据资源[43,46-47]。自然图像预训练虽然使用方便，但光学反射图像与 SAR 散射图像之间存在明显分布差异，使直接迁移往往收益有限。归结起来，感知层的核心挑战在于：如何把复杂空间先验建模、有限监督条件下的标准化基准构建以及跨域迁移机制统一到同一表征学习框架中，从而获得兼具广域上下文适应性、数据可扩展性和任务可迁移性的遥感基础表征。

挑战二：架构层上，异构多模态的分离建模与任务碎片化。多源遥感信息融合已经从传统像素级、特征级融合逐步发展到深度学习驱动的语义融合和跨模态表示学习[41-42]。在 RGB-红外、光学-SAR 等场景中，illumination-aware、AR-CNN、TSFADet、C2Former 以及 SkySense 等方法通过成对输入融合互补信息[33,48-51]。这类方法在严格配准条件下有效，但真实遥感系统中的不同传感器往往采集时间不同、分辨率不同、视角与成像几何不同，大规模获得严格空间配对样本并不容易。随着应用系统对统一部署和持续维护的需求提升，研究者开始关注多数据集联合训练和统一模型设计。DA、Universal-RCNN、UniDet 等方法在自然图像或概念相近的数据集之间验证了多数据集学习的潜力[52-54]，GradNorm、不确定性加权等方法则从多任务优化角度缓解目标冲突[55-56]。然而，RGB、SAR 与红外之间的物理机制差异远大于常规数据集域差异，遥感检测还常同时涉及水平框、旋转框等异构标注形式。也就是说，架构层的核心挑战在于：如何在保持模态特异性与任务差异性的同时，实现共享表示、统一训练和高效部署。

挑战三：范式层上，自下而上预训练的局限与跨模态对齐冲突。现代视觉预训练主要包括监督学习、对比学习和掩码图像建模三条主线，代表性工作如 ResNet、ViT、MoCo、BYOL、CLIP、DINOv2 和 MAE 等[22-23,26-28,57-58]。在遥感领域，RingMo、SatMAE、Scale-MAE、RemoteCLIP、CACo、SkySense、GFM 等方法分别从 MIM、图文对齐和图像级对比等角度构建领域基础模型[29-33,59-60]，相关综述也表明，遥感基础模型正从单一视觉表征走向跨模态、跨任务和跨场景泛化[44]。但现有方法多遵循从底层像素逐步构建语义的“自下而上”路径，对高层语言语义如何反向塑造专业视觉骨干关注不足。与此同时，LLaVA、InternVL、

Qwen-VL 等视觉语言模型显示出较强的视觉理解与指令跟随能力[34-35,61-62]，CLIP、ImageBind、LanguageBind 和 UNIALIGN 等方法也展示了共享语义空间对跨模态对齐的潜力[28,36-37,63]。但对遥感异构多模态检测而言，若把模态对齐留到下游微调阶段，模型需要同时处理检测目标与对齐目标，容易引发梯度冲突、训练不稳定和泛化退化[30,52]。归结起来，范式层的核心挑战在于：如何将语言和高层语义从下游辅助信号转化为预训练阶段的主导信号，使其既能反向塑造专业视觉表征，又能在无严格空间配对的条件下实现稳定、语义保真的跨模态早期对齐。

第三节　本文的主要研究内容与贡献

针对上述三层挑战，本文围绕“面向复杂遥感场景的多模态感知基础模型”这一核心主题，从感知层、架构层与范式层三个维度开展研究。主要研究内容与贡献概括如下：

贡献一：提出选择性大核网络 LSKNet，补强复杂空间先验感知能力。针对遥感图像中的广域上下文需求，本文设计了选择性大核网络 LSKNet。该网络通过大核卷积分解与动态空间选择机制，实现大感受野与自适应上下文范围的联合建模，使骨干能够根据输入内容调节有效感受野，并在旋转目标检测、场景分类、语义分割和变化检测等遥感任务上表现出较好的泛化能力。

贡献二：构建 SARDet-100K 并提出 MSFA，补齐 SAR 数据与迁移基础。针对 SAR 目标检测领域数据匮乏、标准不统一和跨域迁移困难的问题，本文构建包含超 11 万张高分辨率 SAR 图像、覆盖 6 类典型目标的大规模检测基准 SARDet-100K。围绕该基准，本文进一步提出多阶段滤波增强预训练框架，通过滤波增强输入、光学遥感桥接预训练和检测器级迁移，缓解自然图像预训练模型向 SAR 领域迁移时面临的差距。

贡献三：提出 SM3Det，实现异构多模态检测统一架构。为实现 RGB、SAR、红外等异构多模态遥感数据的统一感知，本文设计基于网格级稀疏混合专家的多模态检测架构 SM3Det。该架构通过动态路由机制在多个模态专家之间稀疏激活，在保留模态差异的同时实现跨模态知识共享，并通过动态子模块优化缓解多模态、多任务联合训练中的收敛失衡。

贡献四：提出语言引导的遥感基础模型预训练与跨模态对齐方法。针对传统“自下而上”预训练在遥感专业场景中的局限，本文首先提出视觉指令预训

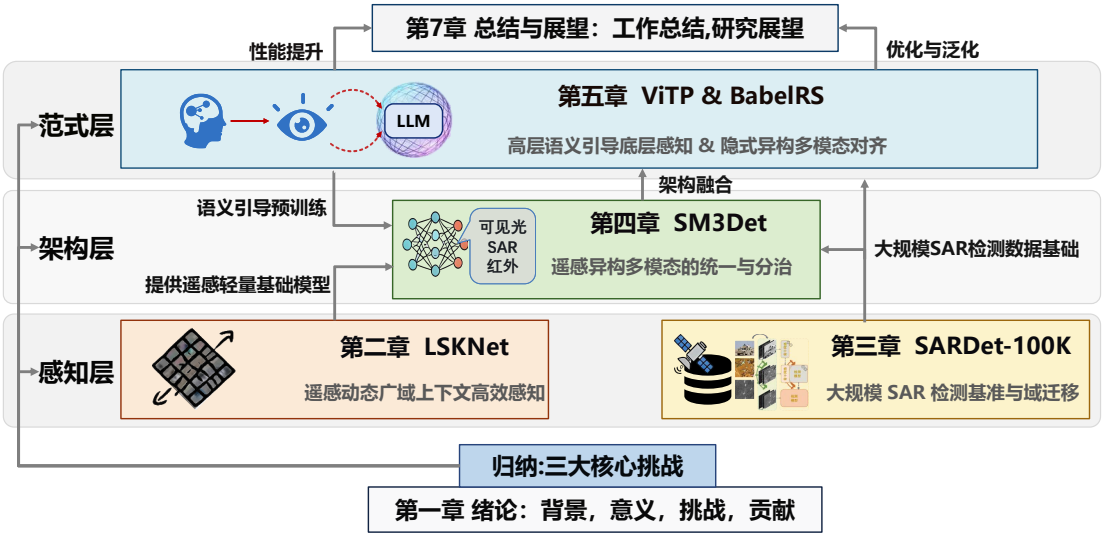


图 1.3 本论文针对“复杂遥感场景的多模态感知基础模型”研究的研究框架和章节关系图。

练框架 ViTP，借助视觉语言模型的理解推理能力，通过端到端持续预训练实现高层语义对底层感知的反向引导，并通过视觉鲁棒学习增强模型对分布偏移和图像退化的适应能力。在此基础上，针对异构多模态后期对齐容易出现的优化冲突，本文进一步提出 BabelRS 预训练框架，以共享语言语义空间为锚点，通过概念共享指令对齐与层级视觉语义退火，将跨模态语义统一前移到预训练阶段。

上述四项贡献分别对应感知基础能力、数据与迁移、统一架构和语言驱动预训练四个层次，共同构成本文围绕“多模态感知基础模型”这一主题的递进式研究路线。

第四节 论文组织结构

本文共分为六章，各章内容组织如下：

第一章为绪论，系统阐述本文的研究背景与意义，综述复杂遥感场景多模态感知基础模型的研究现状，并从感知层、架构层与范式层三个维度凝练关键挑战，进而概述全文的主要研究内容与组织结构。

第二章为空间先验引导的轻量化选择性大核网络，即 LSKNet 的设计与实现。该章围绕遥感场景理解中的广域上下文需求，梳理相关研究进展与不足，介绍大核卷积分解与动态空间选择机制，并通过多个遥感数据集上的实验验证 LSKNet 的适用性。

第三章为大规模 SAR 目标检测基准构建与域适应预训练，即 SARDet-100K 与 MSFA 预训练框架的详细介绍。该章分析 SAR 目标检测在数据基准与迁移链路上的瓶颈，论述 SARDet-100K 数据集的构建过程与特点，并验证 MSFA 连续预训练方法的有效性。

第四章为基于混合专家系统的异构多模态检测统一架构，即 SM3Det 的设计与实现。该章在回顾多数据集检测、多任务学习与多模态融合研究的基础上，定义异构多模态遥感目标检测任务，阐述网格级稀疏 MoE 骨干与动态子模块优化策略，并通过多模态实验验证 SM3Det 的统一建模能力。

第五章为语言引导的遥感基础模型预训练与跨模态对齐，包括 ViTP 与 BabelRS 两个递进部分。该章首先围绕领域化基础模型与视觉指令学习，阐述视觉语言模型引导的端到端预训练机制。随后进一步面向异构遥感多模态条件，介绍概念共享指令对齐与层级视觉语义退火机制，并验证语言引导早期对齐在多模态遥感任务中的作用。

第六章为总结与展望，对本文的研究工作进行系统总结，并展望面向复杂遥感场景的多模态感知基础模型的未来发展方向。

第二章　空间先验引导的轻量化选择性大核网络

遥感影像在俯视视角、尺度跨度、背景组织方式等方面均明显区别于自然图像，这使得分类、目标检测、语义分割和变化检测等任务对空间上下文建模提出了更高要求[7,10,21,45,64-65]。在许多场景中，目标的可靠识别不仅依赖局部纹理，还依赖更大范围的环境线索。与此同时，不同目标所需的上下文范围又存在明显差异。围绕这一特点，本章提出一种轻量级、自适应的大核卷积骨干网络，即选择性大核网络 LSKNet。LSKNet 能够根据输入上下文动态调整有效感受野，更好地刻画遥感场景中不同目标对应的上下文范围。在不引入额外复杂结构的前提下，LSKNet 在遥感分类、目标检测、语义分割和变化检测等标准基准上均取得了较优结果。本章的分析进一步表明，显式建模遥感场景中的空间先验，对于提升基础骨干网络的跨任务适应能力具有重要意义。

第一节　引言

第二章对应绪论中感知层挑战的第一项研究问题，即如何在可控计算成本下实现动态广域上下文感知。对于遥感图像而言，目标识别往往不仅取决于目标自身外观，还取决于其所处场景中的空间关系、背景语义和组织结构。虽然已有研究分别从旋转检测、多尺度特征融合和大规模预训练等方向推动了遥感视觉发展，但现有骨干网络大多仍沿用自然图像中的固定感受野设计，尚不足以支撑“空间先验引导的感知基础模型”这一更底层目标。

航空遥感影像通常以鸟瞰视角捕获。特别是影像中的大多数目标（如车辆、船舶等）可能尺度极小，仅凭自身外观往往难以准确识别。相反，识别这些目标高度依赖于其周围的上下文信息，因为背景环境可以提供有关其类别、方向及空间分布的有价值线索。根据对遥感数据的深入剖析，本章挖掘出两个关键的空间先验，这构成了本章研究的基石：

1. **准确识别通常需要广泛的上下文信息。**如图 2.1所示，在有限的感受野下，船舶和车辆在视觉上可能表现出极高的相似性。区分它们的关键不在于其自身的外观细节，而在于其所处的上下文环境（如海洋之于船舶，道路之于车辆）。

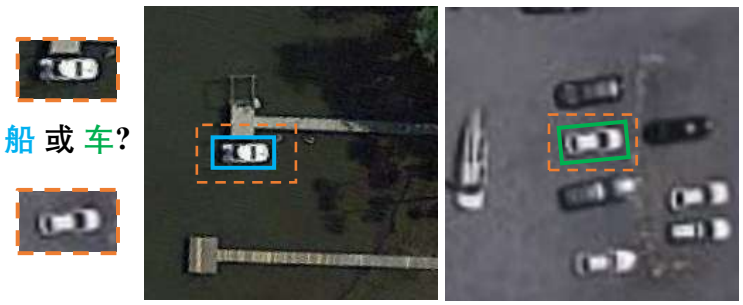


图 2.1 成功检测遥感目标需要利用广泛的上下文信息，而感受野有限的检测器容易误判。

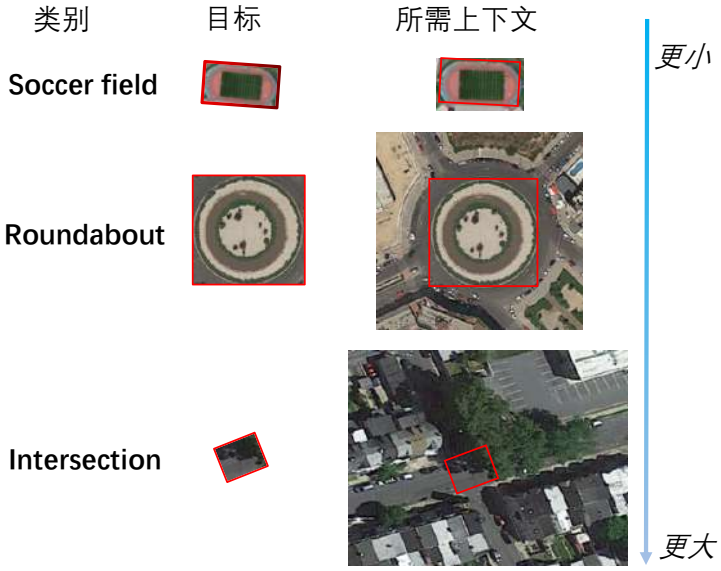


图 2.2 根据人类标准，不同类型目标所需的上下文信息范围差异很大。红色框内的目标为精确的真实标注。

2. **不同目标所需的上下文信息范围差异很大。**如图 2.2所示，由于具备独特且辨识度高的场地边界线，足球场等目标仅需较少的上下文信息即可确认。相比之下，环岛或被树木遮挡的交叉路口则需要跨越数百像素的极广域感受野，以捕捉长程的道路依赖关系。

针对上述问题，本章提出选择性大核网络（Large Selective Kernel Network, LSKNet），将“大感受野”与“上下文自适应选择”统一到同一骨干中。LSKNet 通过对大核卷积进行可分解建模，并在空间维度上引入动态选择机制，使网络能够针对不同目标和场景自动调节有效感受野范围，在保持计算效率的同时更充分地利用遥感场景中的长程上下文先验。

本章的工作不仅给出了一种面向遥感场景的骨干网络设计，也在全文中承担着“打牢单模态感知基础”的角色。后续的 SAR 预训练、多模态统一检测与语言引导预训练，均建立在“感知骨干必须能够适应遥感空间先验”这一前提之上。因此，本章既是独立的方法创新，也是整篇论文技术路线的起点。综上所述，本章的主要贡献包括：

- 深入挖掘并定义了遥感数据中存在的两个重要空间先验，为设计高效感知架构提供了理论指导。
- 提出了 LSKNet 骨干网络，通过大型选择性卷积核动态利用空间先验来提升遥感下游任务性能。
- 尽管结构简单且轻量化，LSKNet 在包含场景分类、目标检测、语义分割和变化检测在内的 14 个公共数据集上均取得了较有竞争力的性能。
- 通过多组定量与定性分析，从实验侧面支持了动态感受野机制在遥感图像分析中的作用，为后续章节构建多模态统一感知模型奠定了单模态特征提取基础。

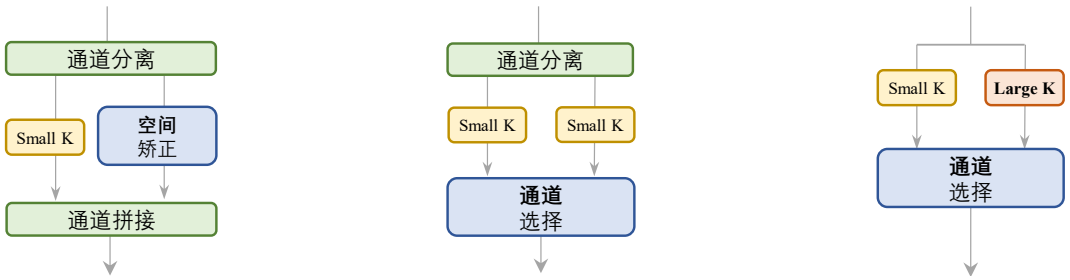
第二节　相关研究与问题分析

一、遥感场景中的广域上下文需求

遥感影像在俯视视角、小目标密集分布、背景组织复杂以及类别间外观相似等方面明显区别于自然图像。许多识别线索并不来自目标自身的局部纹理，而来自道路、岸线、港区、跑道等更大范围的空间关系。围绕这一特点，已有研究分别从旋转检测、多尺度特征融合、语义分割和变化检测等方向提升遥感感知能力[7,10,21,45,64-65]。这些方法在任务层面缓解了尺度变化和上下文不足问题，但大多默认骨干网络的有效感受野相对固定，尚未在骨干层面对“不同目标需要不同上下文范围”这一遥感空间先验进行显式建模。因此，如何以统一的表征机制兼顾广域上下文、计算效率与上下文自适应性，仍是遥感基础网络中的核心问题。

二、大核卷积与大感受野骨干

近年来，扩大感受野已成为视觉骨干设计的重要趋势。基于 Transformer 的模型[69]，如视觉 Transformer（ViT）[70-71]、Swin transformer[72-75]和金字塔 transformer[76-77]，在计算机视觉领域日益流行。研究[78-82]表明，大感受野是它们成



(a) SCNet[66] (b) ResNeSt[67] (c) SKNet[68]

图 2.3 当前主流选择性机制模块的架构对比。图中“K”为卷积核。

功的关键因素之一。更有近期研究显示，设计良好的具有大感受野的卷积网络也能与基于 transformer 的模型相媲美。例如，ConvNeXt[83]在其骨干网络中使用 7×7 深度可分离卷积，提升了下游任务的性能。此外，RepLKNet[84]通过重参数化甚至使用了 31×31 的卷积核，取得了良好的性能。随后的 SLaK[85]工作通过核分解和稀疏分组技术将核大小进一步扩展到 51×51。RF-Next[86]为各种任务自动搜索固定的大核。VAN[87]引入了一种高效的大核分解作为卷积注意力。同样，SegNeXt[88]和 Conv2Former[89]证明了大核卷积在调制具有丰富上下文的卷积特征方面发挥着重要作用。

然而，上述方法大多面向自然图像分类或通用视觉任务，通常采用固定大核或统一上下文范围，难以直接适配遥感场景中“小目标依赖大背景、不同目标依赖范围差异明显”的特点。也就是说，现有研究已经较好回答了“为什么需要大感受野”，但尚未系统回答“遥感骨干应如何按需使用大感受野”。

三、选择性机制与上下文自适应建模

为了使网络能够根据输入上下文动态调整表征，注意力和选择机制被广泛研究。注意力机制[90]是一种简单而有效的方法，可以增强各种任务的神经表示。通道注意力 SE 块[91]使用全局平均信息重新加权特征通道，而空间注意力模块如 GENet[92]、GCNet[93]和 SGE[94]通过空间掩码增强网络建模上下文信息的能力。CBAM[95]和 BAM[96]结合了通道和空间注意力。自注意力机制最初在自然语言处理领域流行[69]，近年来在计算机视觉领域也得到了广泛应用。视觉 Transformer（ViT）[70]利用自注意力捕捉图像中的全局依赖关系和上下文信息。近年来，使用自注意力机制的模型在自然图像分类[97]、检测[98]和分割[99]任务中取得了极具竞争力的性能。然而，在许多遥感图像任务中，如目标检测和分割，

全局上下文信息并非总是必要的。例如，在检测汽车时，数百米外的河流信息并无用处。因此，最近的研究致力于将局部先验信息引入 Transformer 模型，如 Swin[72]、PVT[76,100]、HiViT[101]和 ViTAE[102]。这些模型在遥感场景中相比于原始 ViT 在计算效率和优化方面具有优势[71,103]。

除上述注意力机制外，动态卷积和核选择方法提供了另一条上下文自适应建模路径。CondConv[104]和动态卷积[105]并行核自适应地聚合多个卷积核的特征。SKNet[106]引入了具有不同卷积核的多个分支，并在通道维度上选择性地组合它们。ResNeSt[66]通过将输入特征图划分为多个组，扩展了 SKNet 的理念。类似地，SCNet[67]使用分支注意力捕获更丰富的信息，并使用空间注意力提高定位能力。可变形卷积网络[107-108]为卷积单元引入了灵活的核形状。它们证明了“选择”本身是有效的，但其选择粒度多集中于通道、分支或形变采样位置，较少直接围绕“大核 + 空间位置”这一组合来建模。

四、问题分析与本章定位

综合来看，现有工作已分别证明了大感受野和动态选择机制的重要性，但二者在遥感骨干层面的结合仍不充分，尤其缺少一种既轻量、又能随输入上下文在空间维度上自适应调节上下文范围的方案。对于遥感任务而言，这一缺口尤为关键，因为同一幅图像中不同位置、不同类别目标所需的上下文尺度往往并不一致。基于上述判断，本章将问题聚焦为：在不明显增加复杂度的前提下，如何构建一种能够按需选择大尺度空间上下文的遥感骨干网络。LSKNet 由此通过“分解大核 + 空间选择”的设计，把复杂空间先验转化为可学习、可迁移的骨干表征，并为后续 SAR 预训练、多模态统一检测和语言引导预训练提供更稳固的单模态感知基础。

第三节 方法

一、LSKNet 网络架构

LSKNet 骨干网络的整体架构主要由重复的图 2.4 LSKNet Block 模块构建而成。LSKNet Block 模块的设计灵感来源于 ConvNeXt[109]、MetaFormer[110]、PVT-v2[100]、Conv2Former[89]和 VAN[87]。每个 LSKNet Block 模块由两个残差子模块组成：大核选择（LK Selection）子模块和前馈网络（FFN）子模块。

LK Selection 子模块能够根据需求动态调整网络的感受野。核心的 LSK 模

表 2.1 本章中使用的 LSKNet 变体。C_i：特征通道数；D_i：每个阶段 i 中 LSK 块的数量。

模型	$\{C_1, C_2, C_3, C_4\}$	$\{D_1, D_2, D_3, D_4\}$	参数量
LSKNet-T	{32, 64, 160, 256}	{3, 3, 5, 2}	4.3M
LSKNet-S	{64, 128, 320, 512}	{2, 2, 4, 2}	14.4M

表 2.2 符号、维度及含义诠释。

符号	维度	含义
X	$C \times H \times W$	输入特征
N	1	选择核数量
i	1	分解核索引
$\widetilde{\mathbf{U}}_i$	$C \times H \times W$	富含上下文的特征
$\mathbf{SA}_{max}$	$1 \times H \times W$	通过最大池化得到的空间注意力
$\mathbf{SA}_{avg}$	$1 \times H \times W$	通过平均池化得到的空间注意力
$\widetilde{\mathbf{SA}}_i$	$N \times H \times W$	空间选择注意力
S	$C \times H \times W$	融合后的注意力特征
Y	$C \times H \times W$	输出特征

块（如图 2.5所示）嵌入在 LK Selection 子模块中。该模块由一系列大核卷积和空间尺度的核选择机制组成，具体细节将在后文详细阐述。FFN 子模块用于通道混合和特征细化，由全连接层、深度可分离卷积、GELU[111]激活函数和第二个全连接层依次组成。表 2.1列出了本章所使用的 LSKNet 不同变体的详细配置。此外，表 2.2提供了重要符号的完整列表，包括它们对应的维度和含义。这些符号在图 2.5和后续章节的方程中被广泛引用。

二、大核卷积

第一节中的先验 *2)* 建议对一系列多尺度长程上下文进行建模以实现自适应选择模型感受野。因此，本章提出通过显式分解的方式，将大核卷积构建为一系列具有逐渐增大核尺寸和扩张率的深度可分离卷积。具体而言，对于第 i 个深度可分离卷积，核大小 k、扩张率 d 和感受野 RF 的扩展定义如下：

$$k_{i-1} \leq k_i;\ d_1 = 1,\ d_{i-1} < d_i \leq RF_{i-1}. \tag{2.1}$$

$$RF_1 = k_1,\ RF_i = d_i(k_i - 1) + RF_{i-1}. \tag{2.2}$$

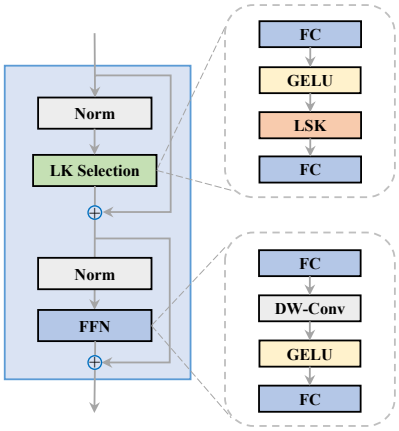


图 2.4 LSKNet Block 模块示意图。

图 2.5 LSK 模块的概念性示意图。

逐渐增大的核大小和扩张率确保了感受野能够快速扩展。本章对扩张率设置了上限，以保证扩张卷积不会在特征图之间引入间隙。例如，如表 2.3所示，可以将大核分解为 2 个或 3 个深度可分离卷积，理论感受野分别为 23 和 29。这种设计具有两个优势：首先，它显式地产生了具有不同大感受野的多个特征，便于后续的核选择。其次，顺序分解比直接应用单个更大的核更加高效。如表 2.3所示，在相同的理论感受野下，本章的分解方法与标准大卷积核相比大大减少了参数数量。

为了从输入 $\mathbf{X}$ 中获取具有不同范围丰富上下文信息的特征，LSKNet 应用了一系列具有不同感受野的被分解的深度可分离卷积：

$$\mathbf{U}_0 = \mathbf{X}, \quad \mathbf{U}_{i+1} = \mathcal{F}_i^{dw}(\mathbf{U}_i), \tag{2.3}$$

其中 $\mathcal{F}_i^{dw}(\cdot)$ 是具有核 k_i 和扩张率 d_i 的深度可分离卷积。假设有 N 个分解核，每个核都通过 1×1 卷积层 $\mathcal{F}^{1\times1}(\cdot)$ 进行进一步处理：

表 2.3 **两个代表性示例的理论效率比较**，本章将单个大型深度可分离卷积核展开为核序列，假设通道数为 64。k：核大小；d：膨胀率。

RF	(k, d) 序列	参数量	FLOPs
23	(23, 1)	40.4K	42.4G
	(5,1) $\longrightarrow$ (7, 3)	11.3K	11.9G
29	(29, 1)	60.4K	63.3G
	(3, 1) $\longrightarrow$ (5, 2) $\longrightarrow$ (7, 3)	11.3K	13.6G

$$\widetilde{\mathbf{U}}_i = \mathcal{F}_i^{1\times1}(\mathbf{U}_i), \quad i \in [1, N], \tag{2.4}$$

这允许对每个空间特征向量进行通道混合。接下来本章提出了一种选择机制，基于获得的多尺度特征动态选择适用于不同目标的核。

三、空间尺度的核选择机制

为了增强网络聚焦于检测目标最相关空间上下文区域的能力，本章采用了空间选择机制，对不同尺度的大卷积核所得到的特征图进行空间选择。首先，本章将通过不同感受野范围的卷积核所获得的特征进行拼接：

$$\widetilde{\mathbf{U}} = [\widetilde{\mathbf{U}}_1; ...; \widetilde{\mathbf{U}}_i], \tag{2.5}$$

然后，通过对 $\widetilde{\mathbf{U}}$ 应用基于通道的平均池化和最大池化（分别表示为 $\mathcal{P}_{avg}(\cdot)$ 和 $\mathcal{P}_{max}(\cdot)$）来高效提取空间关系：

$$\mathbf{SA}_{avg} = \mathcal{P}_{avg}(\widetilde{\mathbf{U}}), \ \mathbf{SA}_{max} = \mathcal{P}_{max}(\widetilde{\mathbf{U}}), \tag{2.6}$$

其中，$\mathbf{SA}_{avg}$ 和 $\mathbf{SA}_{max}$ 分别为平均池化和最大池化后的空间特征描述符。为了实现不同空间描述符之间的信息交互，本章将空间池化特征进行拼接，并使用卷积层 $\mathcal{F}^{2\to N}(\cdot)$ 将池化特征（具有 2 个通道）转换为 N 个空间注意力图：

$$\widehat{\mathbf{SA}} = \mathcal{F}^{2\to N}([\mathbf{SA}_{avg}; \mathbf{SA}_{max}]). \tag{2.7}$$

对于每个空间注意力图 $\widehat{\mathbf{SA}}_i$，本章应用 sigmoid 激活函数以获得每个被分解大卷积核的单独空间选择掩码：

$$\widetilde{\mathbf{SA}}_i = \sigma(\widehat{\mathbf{SA}}_i), \tag{2.8}$$

其中 $\sigma(\cdot)$ 表示 sigmoid 函数。分解大卷积核序列的特征图通过其对应的空间选择掩码进行加权，然后通过卷积层 $\mathcal{F}(\cdot)$ 融合，得到注意力特征 $\mathbf{S}$：

$$\mathbf{S} = \mathcal{F}(\sum_{i=1}^{N}(\widetilde{\mathbf{SA}}_i \cdot \widetilde{\mathbf{U}}_i)). \tag{2.9}$$

LSK 模块的最终输出是输入特征 $\mathbf{X}$ 与 $\mathbf{S}$ 的逐元素乘积，类似于[87-89]中的方法：

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{S}. \tag{2.10}$$

图 2.5展示了 LSK 模块的详细概念图，直观说明了该模块如何依据局部语义特征自适应选择不同范围的感受野。

第四节 实验与分析

在主要结果中，本章采用了在 ImageNet-1K[24]上进行 300 轮的骨干网络预训练策略以追求更高的性能，这与[112-114]的做法类似。然而，对于场景分类任务，本章遵循[115]中概述的预训练设置，在 millionAID 数据集[116]上进行 300 轮预训练。本章直接使用官方或者默认的训练、验证和测试集划分，并遵循每个基准测试的主流设置以确保公平性。在消融研究中，为了实验效率，本章采用了在 ImageNet-1K 上进行 100 轮的骨干网络预训练策略。表格中，最高得分用**粗体**表示，次高得分用下划线 标注。本章实验中的 FLOPs 是通过将 1024×1024 像素的图像输入网络计算得出的。

一、主要实验

（一）场景分类

分类数据集 遥感图像分类的主流方法[115,117-119]通常在三个标准场景识别数据集上进行实验，包括 UC Merced 土地利用（UCM）[120]数据集、航空图像数据集（AID）[121]和西北工业大学收集的图像场景分类（NWPU）[122]数据集。

UCM 是一个相对较小的数据集，仅包含 2,100 张图像和 21 个类别，每个类别有 100 张图像。所有图像的尺寸为 256 × 256。

AID 包含 10,000 张图像，分为 30 个类别，所有图像的尺寸为 600 × 600。

NWPU 是一个相对较大的数据集，包含 31,500 张图像和 45 个类别，每个类别有 700 张图像。所有图像的尺寸为 256 × 256。

遵循遥感分类工作的主流方法[115,117-119]，本章在五个标准基准上进行实验，即 UCM-82、AID-28、AID-55、NWPU-19 和 NWPU-28。

分类结果 表 2.4展示了各种对比方法的分类结果。本章将所提出的 LSKNets 与其他 22 种近期遥感场景分类方法进行了比较。值得注意的是，在不使用任何技巧（如 MBENet[64]和 FENet[65]中的特征集成）的情况下，本章提出的轻量级模型 LSKNet-T 和 LSKNet-S 在多个数据集上都展现出了具有竞争力的性能。这些结果表明，LSKNet 在各种场景下进行准确场景分类方面具有良好的效果，同时也展示了其作为骨干网络进行特征提取的潜力。

表 2.4 不同模型在场景分类上的性能表现。

模型	参数量↓	FLOPs↓	UCM-82	AID-28	AID-55	NWPU-19	NWPU-28
MSANet[123]	>42.3M	>164.3	98.96	93.53	96.01	90.38	93.52
ViT-B[70]	86.0M	118.9G	99.28	93.81	96.08	90.96	93.96
SCCov[124]	13.0M	-	99.05	93.12	96.10	89.30	92.10
MA-FE[125]	>25.6M	>86.3G	99.66	-	95.98	-	93.21
MG-CAP[126]	>42.3M	>164.3G	99.00	93.34	96.12	90.83	92.95
LSENet[127]	25.9M	>86.3G	99.78	94.41	96.36	92.23	93.34
IDCCP[128]	25.6M	86.3G	99.05	94.80	96.95	91.55	93.76
F^2BRBM[118]	25.6M	86.3G	99.58	96.05	96.97	92.74	94.87
EAM[129]	>42.3M	>164.3	98.98	94.26	97.06	91.91	94.29
MBLANet[117]	-	-	99.64	95.60	97.14	92.32	94.66
GRMANet[119]	54.1M	171.4G	99.19	95.43	97.39	93.19	94.72
KFBNet[130]	-	-	99.88	95.50	97.40	93.08	95.11
RSP-R50[115]	25.6M	86.3G	99.48	96.81	97.89	93.93	95.02
RSP-Swin[115]	27.5M	37.7G	99.52	96.83	98.30	94.02	94.51
RSP-ViTAE[115]	19.3M	119.1G	99.90	96.91	98.22	**94.41**	95.60
RVSA[71]	114.4M	301.3G	-	97.01	98.50	93.92	95.66
ConvNext[83]	28.0M	93.7G	99.81	95.43	97.40	94.07	94.76
FSCNet[131]	28.8M	166.1G	**100**	95.56	97.51	93.03	94.76
UPetu[132]	87.7M	>322.2G	99.05	96.29	97.06	92.13	93.79
MBENet[64]	23.9M	108.5G	99.81	96.00	98.54	92.50	95.58
FENet[65]	23.9M	92.0G	99.86	96.45	**98.60**	92.91	95.39
LSKNet-T	**4.3M**	**19.2G**	99.81	96.80	98.14	94.07	95.75
LSKNet-S	14.4M	54.4G	99.81	**97.05**	98.22	94.27	**95.83**

（二）定向目标和合成孔径雷达目标检测

目标检测数据集 为评估所提出模型在遥感检测任务中的适用性，本章在 4 个具有挑战性的数据集上进行了实验。这些数据集包括 3 个广泛使用的定向目标检测数据集：HRSC2016[133]、DOTA-v1.0[134]和 FAIR1M-v1.0[135]，以及一个复杂且具有挑战性的合成孔径雷达（SAR）数据集 SAR-Aircraft[136]。

DOTA-v1.0[134]由 2,806 张遥感图像组成，包含 188,282 个实例，涵盖 15 个类别：飞机（PL）、棒球场（BD）、桥梁（BR）、田径场（GTF）、小型车辆（SV）、大型车辆（LV）、船舶（SH）、网球场（TC）、篮球场（BC）、储罐（ST）、足球场（SBF）、环岛（RA）、港口（HA）、游泳池（SP）和直升机（HC）。

HRSC2016[133]是一个专门用于船舶检测的高分辨率遥感数据集，由 1,061 张图像组成，包含 2,976 个船舶实例。

FAIR1M-v1.0[135]是一个近期发布的遥感数据集，包含 15,266 张高分辨率图像和超过 100 万个实例。该数据集涵盖 5 个主类别和 37 个子类别的目标。

SAR-Aircraft 数据集[136]是一个专为 SAR 模态目标检测收集的最新遥感数据集。与前述 3 个 RGB 模态数据集不同，SAR 数据集中的图像都为灰度图像。该数据集包含 7 个不同类别，分别是 A220、A320/321、A330、ARJ21、Boeing737、Boeing787 和其他。数据集由 3,489 张训练图像和 879 张测试图像组成，总计 16,463 个飞机实例。

检测结果 在定向目标检测实验中，考虑到其出色的性能和效率，本章默认将 LSKNet 构建在 Oriented RCNN[112]框架内。

DOTA-v1.0 数据集结果。本章将 LSKNet 与 20 种近期方法在 DOTA-v1.0 数据集上进行了比较，结果如表 2.5所示。本章提出的 LSKNet-T、LSKNet-S 和 LSKNet-S* 分别取得 **81.37%**、**81.64%** 和 **81.85%** 的 mAP。值得注意的是，LSKNet-S 在单个 RTX3090 GPU 上处理 1024x1024 图像时，推理速度可达 18.1 FPS。

FAIR1M-v1.0 数据集结果。本章将 LSKNet 与其他 6 种模型在 FAIR1M-v1.0 数据集上进行了比较，结果如表 2.6所示。结果表明，本章提出的 LSKNet-T 和 LSKNet-S 分别取得 **46.93%** 和 **47.87%** 的 mAP，在所比较模型中结果更高。

HRSC2016 数据集结果。本章在 HRSC2016 数据集上评估了 LSKNet 与 12 种近期方法的性能。表 2.7中的结果表明，本章提出的 LSKNet-S 在 PASCAL

表 2.5 在 **DOTA-v1.0** 数据集上与先进模型的比较，采用多尺度训练和测试。*：与比较方法类似，使用 EMA 微调[137]。

模型	**Pre.**	**mAP↑**	**参数量↓**	**FLOPs↓**	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC
单阶段																			
R3Det[114]	IN	76.47	41.9M	336G	89.80	83.77	48.11	66.77	78.76	83.27	87.84	90.82	85.38	85.51	65.57	62.68	67.53	78.56	72.62
CFA[138]	IN	76.67	-	-	89.08	83.20	54.37	66.87	81.23	80.96	87.17	90.21	84.32	86.09	52.34	69.94	75.52	80.76	67.96
DAFNe[139]	IN	76.95	-	-	89.40	86.27	53.70	60.51	82.04	81.17	88.66	90.37	83.81	87.27	53.93	69.38	75.61	81.26	70.86
SASM[140]	IN	79.17	-	-	89.54	85.94	57.73	78.41	79.78	84.19	89.25	90.87	58.80	87.27	63.82	67.81	78.67	79.35	69.37
AO2-DETR[141]	IN	79.22	74.3M	304G	89.95	84.52	56.90	74.83	80.86	83.47	88.47	90.87	86.12	**88.55**	63.21	65.09	79.09	**82.88**	73.46
S^2ANet[113]	IN	79.42	-	-	88.89	83.60	57.74	81.95	79.94	83.19	**89.11**	90.78	84.87	87.81	70.30	68.25	78.30	77.01	69.58
R3Det-GWD[142]	IN	80.23	41.9M	336G	89.66	84.99	59.26	82.19	78.97	84.83	87.70	90.21	86.54	86.85	**73.47**	67.77	76.92	79.22	74.92
RTMDet-R[137]	IN	80.54	52.3M	205G	88.36	84.96	57.33	80.46	80.58	84.88	88.08	**90.90**	86.32	87.57	69.29	70.61	78.63	80.97	**79.24**
R3Det-KLD[143]	IN	80.63	41.9M	336G	89.92	85.13	59.19	81.33	78.82	84.38	87.50	89.80	87.33	87.00	72.57	71.35	77.12	79.34	78.68
RTMDet-R[137]	CO	81.33	52.3M	205G	88.01	86.17	58.54	82.44	81.30	84.82	88.71	90.89	**88.77**	87.37	71.96	71.18	81.23	81.40	77.13
两阶段																			
SCRDet[144]	IN	72.61	-	-	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	87.94	86.86	65.02	66.68	66.25	68.24	65.21
ViTDet[145]	IN	74.41	103.2M	502G	88.38	75.86	52.24	74.42	78.52	83.22	88.47	90.86	77.18	86.98	48.95	62.77	76.66	72.97	57.48
Rol Trans.[146]	IN	74.61	55.1M	200G	88.65	82.60	52.53	70.87	77.93	76.67	86.87	90.71	83.83	82.51	53.95	67.61	74.67	68.75	61.03
G.V.[147]	IN	75.02	41.1M	198G	89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32
CenterMap[148]	IN	76.03	41.1M	198G	89.83	84.41	54.60	70.25	77.66	78.32	87.19	90.66	84.89	85.27	56.46	69.23	74.13	71.56	66.06
CSL[149]	IN	76.17	37.4M	236G	**90.25**	85.53	54.64	75.31	70.44	73.51	77.62	90.84	86.15	86.69	69.60	68.04	73.83	71.10	68.93
ReDet[150]	IN	80.10	-	-	88.81	82.48	60.83	80.82	78.34	86.06	88.31	90.87	**88.77**	87.03	68.65	66.90	79.26	79.71	74.67
DODet[151]	IN	80.62	-	-	89.96	85.52	58.01	81.22	78.71	85.46	88.59	90.89	87.12	87.80	70.50	71.54	82.06	77.43	74.47
AOPG[152]	IN	80.66	-	-	89.88	85.57	60.90	81.51	78.70	85.29	88.85	90.89	87.60	87.65	71.66	68.69	82.31	77.32	73.10
O-RCNN[112]	IN	80.87	41.1M	199G	89.84	85.43	61.09	79.82	79.71	85.35	88.82	90.88	86.68	87.73	72.21	70.80	82.42	78.18	74.11
KFloU[153]	IN	80.93	58.8M	206G	89.44	84.41	62.22	82.51	80.10	86.07	88.68	**90.90**	87.32	88.38	72.80	71.95	78.96	74.95	75.27
RVSA[71]	MA	81.24	114.4M	414G	88.97	85.76	61.46	81.27	79.98	85.31	88.30	90.84	85.06	87.50	66.77	**73.11**	**84.75**	81.88	77.58
LSKNet-T	IN	81.37	**21.0M**	**124G**	89.14	84.90	61.78	83.50	81.54	85.87	88.64	90.89	88.02	87.31	71.55	70.74	78.66	79.81	78.16
LSKNet-S	IN	81.64	31.0M	161G	89.57	**86.34**	**63.13**	**83.67**	**82.20**	**86.10**	88.66	90.89	88.41	87.42	71.72	69.58	78.88	81.77	76.52
LSKNet-S*	IN	**81.85**	31.0M	161G	89.69	85.70	61.47	83.23	81.37	86.05	88.64	90.88	88.49	87.40	71.67	71.35	79.19	81.77	**80.86**

表 2.6 在 **FAIR1M-v1.0** 数据集上与近期模型的比较。*：结果引用自 FAIR1M 论文[135]。

模型	G. V.*[147]	RetinaNet*[18]	C-RCNN*[154]	F-RCNN*[17]	RoI Trans.*[146]	O-RCNN[112]	LSKNet-T	LSKNet-S
mAP(%)	29.92	30.67	31.18	32.12	35.29	45.60	46.93	**47.87**

表 2.7 在 **HRSC2016** 数据集上与近期模型的比较。mAP (07/12)：VOC 2007[155]/2012[156] 评价指标。

模型	Pre.	**mAP(07)**	**mAP(12)**	**参数量**	**FLOPs**
DRN[157]	IN	-	92.70	-	-
CenterMap[148]	IN	-	92.80	41.1M	198G
RoI Trans.[146]	IN	86.20	-	55.1M	200G
G. V.[147]	IN	88.20	-	41.1M	198G
R3Det[114]	IN	89.26	96.01	41.9M	336G
DAL[158]	IN	89.77	-	36.4M	216G
GWD[142]	IN	89.85	97.37	47.4M	456G
S^2ANet[113]	IN	90.17	95.01	38.6M	198G
AOPG[152]	IN	90.34	96.22	-	-
ReDet[150]	IN	90.46	97.63	31.6M	-
O-RCNN[112]	IN	90.50	97.60	41.1M	199G
RTMDet[137]	CO	90.60	97.10	52.3M	205G
LSKNet-T	IN	90.54	98.13	**21.0M**	**124G**
LSKNet-S	IN	**90.65**	**98.46**	31.0M	161G

VOC 2007[155]和 VOC 2012[156]指标下分别达到 **90.65%** 和 **98.46%** 的 mAP，在所比较方法中结果更高。

SAR-Aircraft 数据集结果。本章评估了所提出的 LSKNets 与 5 种近期骨干网络在 Cascade Mask RCNN[154]和 RetinaNet[18]检测框架下的性能。结果如表 2.8所示，LSKNets 在 SAR 目标检测任务中带来了稳定收益。

定量分析。在比较的模型中，使用原始 ViT 骨干网络的 ViTDet 具有最大的计算复杂度（相比 LSKNet-T 高 4.0 倍的 FLOPs）和第二大的模型规模（相比 LSKNet-T 多 4.9 倍的参数），但在 DOTA-v1.0 数据集的目标检测任务上表现不佳。另一种基于 ViT 的模型变体 RVSA，以 ViTAE 为基础，融合了多尺度和二维局部性归纳偏置，在建模图像特征方面比原始 ViT 骨干网络更为有效。尽管 RVSA 具有较好效果，但仍存在模型规模庞大（相比 LSKNet-T 多 5.4 倍的参数）和计算复杂度高（相比 LSKNet-T 高 3.3 倍的 FLOPs）的问题。这两种基于 ViT 的模型在该设置下均未超过轻量级的 LSKNet-T。

表 2.8 **SAR-Aircraft** 测试集上的 mAP 结果。

RetinaNet[18] 2x	**参数量**	**mAP$_{50}$**	**mAP$_{75}$**
ResNet-50[159]	25.6M	0.469	0.324
PVT-Tiny[76]	13.2M	0.498	0.335
Res2Net-50[160]	25.7M	0.528	0.339
Swin-T[72]	28.3M	0.586	0.346
ConvNeXt V2-N[161]	15.0M	0.589	0.350
VAN-B1[87]	13.4M	0.603	0.375
LSKNet-T	**4.3M**	0.582	0.354
LSKNet-S	14.4M	**0.624**	**0.387**

Cascade Mask RCNN[154] 2x	**参数量**	**mAP$_{50}$**	**mAP$_{75}$**
ResNet-50[159]	25.6M	0.483	0.339
PVT-Tiny[76]	13.2M	0.502	0.344
Res2Net-50[160]	25.7M	0.544	0.372
ConvNeXt V2-N[161]	15.0M	0.581	0.428
Swin-T[72]	28.3M	0.596	0.416
VAN-B1[87]	13.4M	0.604	0.457
LSKNet-T	**4.3M**	0.586	0.435
LSKNet-S	14.4M	**0.614**	**0.458**

LSKNet 的优势还体现在 DOTA-v1.0 数据集中容易混淆的类别上，如小型车辆（+2.49%）和船舶（+3.59%）（表 2.5），以及 FAIR1M 数据集中需要大量上下文信息的类别上，如交叉路口（+2.08%）、环岛（+6.53%）和桥梁（+6.11%）。这些结果从任务现象层面支持了本章提出的先验 *1* 和先验 *2*，也说明所提出的基础骨干模型能够较好适配遥感场景中的上下文需求。

（三）语义分割

分割数据集 遵循主流分割研究的做法[10,162]，本章通过在五个标准数据集上进行评估来验证所提出模型在遥感分割任务中的有效性：Potsdam[163]、Vaihingen[164]、LoveDA[165]、UAVid[166]和 GID[167]数据集。

Potsdam[163]是一个高分辨率语义分割数据集，包含 38 张高分辨率图像。它由 6 个语义类别组成：不透水表面、建筑物、低矮植被、树木、汽车和一个背景类别（杂波）。

Vaihingen[164]同样是一个高分辨的语义分割数据集，由 33 张高分辨率图像

表 2.9 Potsdam 测试集上的定量比较结果。OA：总体精度

模型	**mF1** ↑	**OA** ↑	**mIOU** ↑
ERFNet[169]	85.8	84.5	76.2
DABNet[170]	88.3	86.7	79.6
BiSeNet[171]	89.8	88.2	81.7
EaNet[172]	90.6	88.7	83.4
MARESU-Net[173]	90.5	89.0	83.9
DANet[174]	88.9	89.1	80.3
SwiftNet[175]	91.0	89.3	83.8
FANet[176]	91.3	89.8	84.2
ShelfNet[177]	91.3	89.9	84.4
ABCNet[178]	92.7	91.3	86.5
Segmenter[179]	89.2	88.7	80.7
BANet[162]	92.5	91.0	86.3
SwinUperNet[72]	92.2	90.9	85.8
UNetFormer[10]	92.8	91.3	<u>86.8</u>
LSKNet-T	<u>92.9</u>	<u>91.7</u>	86.7
LSKNet-S	**93.1**	**92.0**	**87.2**

组成。其语义类别与 Potsdam 相同。

LoveDA[165]是一个多尺度且复杂的遥感语义分割数据集，包含 5,987 张 1024×1024 像素的图像。其中，2522 张用于训练，1,669 张用于验证，1,796 张用于在线测试。该数据集包含 7 个语义类别：建筑物、道路、水体、裸地、森林、农田和背景。

UAVid[166]是一个高分辨率且复杂的无人机（UAV）语义分割数据集。它包含 200 张训练图像、70 张验证图像和 150 张在线测试图像。该数据集由 8 个不同类别组成：建筑物、道路、树木、植被、移动车辆、静止车辆、人类和其他。

GID[167]数据集是一个中等分辨率的土地覆盖分割数据集，地面采样距离（GSD）为 4m，包含 150 张 7,200×6,800 像素的图像。按照[168]的方法，本研究从原始 GID 数据集中选择了 15 张预定义图像，并将所有图像裁剪为 256×256 像素，最终得到 7,830 张训练图像和 3,915 张测试图像。该数据集包含六个语义类别：建成区、农田、森林、草地、水体和其他。

分割结果 本章在上述 5 个数据集上对所提出的 LSKNet-T 和 LSKNet-S 模型与多个近期提出的高水平模型进行了全面比较。对于 Potsdam、Vaihingen、LoveDA

表 2.10 Vaihingen 测试集上的定量比较结果。

模型	**mF1** ↑	**OA** ↑	**mIOU** ↑
PSPNet[180]	79.0	87.7	68.6
ERFNet[169]	78.9	85.8	69.1
DANet[174]	79.6	88.2	69.4
DABNet[170]	79.2	84.3	70.2
Segmenter[179]	84.1	88.1	73.6
BOTNet[181]	84.8	88.0	74.3
FANet[176]	85.4	88.9	75.6
BiSeNet[171]	84.3	87.1	75.8
DeepLabV3+[182]	87.4	89.0	-
ShelfNet[177]	87.5	89.8	78.3
MARESU-Net[173]	87.7	90.1	78.6
EaNet[172]	87.7	89.7	78.7
SwiftNet[175]	88.3	90.2	79.6
ABCNet[178]	89.5	90.7	81.3
BANet[162]	89.6	90.5	81.4
UNetFormer[10]	90.4	91.0	82.7
LSKNet-T	91.7	**93.6**	84.9
LSKNet-S	**91.8**	**93.6**	**85.1**

和 UAVid 数据集，由于 UNetFormer[10]框架较优的性能且开源可用，LSKNet 被集成到该框架中。对于 GID 数据集，本章使用 SegFormer 框架比较了各种骨干网络模型。具体而言，本研究在 Potsdam 数据集上与 14 个模型进行了比较（表 2.9），在 Vaihingen 数据集上与 16 个模型进行了比较（表 2.10），在 LoveDA 数据集上与 13 个模型进行了比较（表 2.11），在 UAVid 数据集上与 16 个模型进行了比较（表 2.12），在 GID 数据集上与 6 个骨干网络模型进行了比较（表 2.13）。值得注意的是，本章提出的 LSKNet-T 和 LSKNet-S 模型在多个数据集的主要指标上取得了较好的比较结果。

（四）变化检测

变化检测数据集 遵循主流变化检测研究的做法[45,195-196]，本章在以下两个标准数据集上进行评估来验证所提出模型在遥感变化检测任务中的有效性：LEVIR-CD[197]和 S2Looking[198]。

LEVIR-CD[197]包含 637 对来自 Google Earth 的双时相图像，每张图像的尺寸为 1024×1024 像素，地面采样距离（GSD）为 0.5 米。该数据集标注了 31,333

表 2.11 LoveDA 测试集上的定量比较结果。

模型	**mIoU** ↑	背景	建筑物	道路	水体	裸地	森林	农田
Segmenter[179]	47.1	38.0	50.7	48.7	77.4	13.3	43.5	58.2
SegFormer[183]	47.4	43.1	52.3	55.0	70.7	10.7	43.2	56.8
DeepLabV3+[182]	47.6	43.0	50.9	52.0	74.4	10.4	44.2	58.5
UNet[184]	47.6	43.1	52.7	52.8	73.0	10.3	43.1	59.9
UNet++[185]	48.2	42.9	52.6	52.8	74.5	11.4	44.4	58.8
SemanticFPN[186]	48.2	42.9	51.5	53.4	74.7	11.2	44.6	58.7
FarSeg[187]	48.2	43.1	51.5	53.9	76.6	9.8	43.3	58.9
PSPNet[180]	48.3	44.4	52.1	53.5	76.5	9.7	44.1	57.9
FactSeg[188]	48.9	42.6	53.6	52.8	76.9	16.2	42.9	57.5
TransUNet[189]	48.9	43.0	56.1	53.7	78.0	9.3	44.9	56.9
BANet[162]	49.6	43.7	51.5	51.1	76.9	16.6	44.9	62.5
HRNet[190]	49.8	44.6	55.3	57.4	78.0	11.0	45.3	60.9
SwinUperNet[72]	50.0	43.3	54.3	54.3	78.7	14.9	45.3	59.6
DC-Swin[191]	50.6	41.3	54.5	56.2	78.1	14.5	**47.2**	62.4
UNetFormer[10]	52.4	44.7	58.8	54.9	79.6	20.1	46.0	62.5
Hi-ResNet[192]	52.5	**46.7**	58.3	55.9	80.1	17.0	46.7	**62.7**
LSKNet-T	53.2	46.4	59.5	57.1	79.9	21.8	46.6	61.4
LSKNet-S	**54.0**	**46.7**	**59.9**	**58.3**	**80.2**	**24.6**	46.4	61.8

表 2.12 UAVid 测试集上的定量比较结果。

模型	**mIoU** ↑	其他	建筑物	道路	树木	植被	移动车辆	静止车辆	人类
MSD[166]	57.0	57.0	79.8	74.0	74.5	55.9	62.9	32.1	19.7
CANet[193]	63.5	66.0	86.6	62.1	79.3	**78.1**	47.8	**68.3**	19.9
DANet[174]	60.6	64.9	85.9	77.9	78.3	61.5	59.6	47.4	9.1
SwiftNet[175]	61.1	64.1	85.3	61.5	78.3	76.4	51.1	62.1	15.7
BiSeNet[171]	61.5	64.7	85.7	61.1	78.3	77.3	48.6	63.4	17.5
MANet[173]	62.6	64.5	85.4	77.8	77.0	60.3	67.2	53.6	14.9
ABCNet[178]	63.8	67.4	86.4	81.2	79.9	63.1	69.8	48.4	13.9
Segmenter[179]	58.7	64.2	84.4	79.8	76.1	57.6	59.2	34.5	14.2
SegFormer[183]	66.0	66.6	86.3	80.1	79.6	62.3	72.5	52.5	28.5
BANet[162]	64.6	66.7	85.4	80.7	78.9	62.1	69.3	52.8	21.0
BOTNet[181]	63.2	64.5	84.9	78.6	77.4	60.5	65.8	51.9	22.4
CoaT[194]	65.8	69.0	**88.5**	80.0	79.3	62.0	70.0	59.1	18.9
UNetFormer[10]	67.8	68.4	87.4	81.5	80.2	63.5	73.6	56.4	31.0
LSKNet-T	69.3	**69.6**	87.9	82.8	80.6	64.8	**77.3**	60.2	31.3
LSKNet-S	**70.0**	**69.6**	84.8	**82.9**	**80.9**	65.5	76.8	64.9	**31.8**

表 2.13 GID 测试集上的定量比较结果。

骨干网络	**mF1** ↑	**OA** ↑	**mIoU** ↑
ConvNext-v2-N[161]	75.1	78.9	62.5
ResNet-50[159]	75.3	80.0	64.1
Swin-T[72]	77.8	80.8	65.6
ResNest-50[66]	79.7	80.3	67.2
VAN-S[87]	80.2	82.1	68.2
MSCAN-S[88]	80.4	81.4	68.4
LSKNet-T	79.4	81.5	67.2
LSKNet-S	**83.2**	**82.3**	**69.6**

个二元变化实例。

S2Looking[198]由全球光学卫星拍摄的 5,000 对双时相图像组成。每张图像的尺寸为 1024×1024 像素，GSD 范围在 0.5 至 0.8 米之间。该数据集标注超过 65,920 个二元变化实例。

变化检测结果 在变化检测实验中，由于 Changer[45]框架较优的性能且开源可用，LSKNet 默认构建于该框架之上。本章在 LEVIR-CD 和 S2Looking 数据集上对所提出的 LSKNet-T 和 LSKNet-S 模型与 17 个近期高性能模型进行了系统比较。表 2.14中的结果表明，LSKNet-T 和 LSKNet-S 在两个数据集的主要指标（F1 和 IoU）上均取得了具有竞争力的结果。

二、分析与对比实验

（一）消融分析

下文报告在 DOTA-v1.0 测试集上进行的消融实验结果。选择 DOTA-v1.0 数据集进行消融研究主要基于两个因素：首先，目标检测是一项实用且具有挑战性的任务，而 DOTA-v1.0 数据集提供了多样化且复杂的目标和场景用于评估。其次，众多可用模型的存在使得全面比较成为可能，从而能够对本章提出方法的有效性进行深入评估。在消融研究中，为了提高实验效率，本章采用了 100 轮的骨干网络预训练计划（表 2.15、 2.16、 2.17、 2.19、 2.18）。

大核分解。确定分解的核数量是 LSK 模块的一个关键选择。本章遵循公式(2.1)来配置分解后的核。表 2.15展示了在理论感受野固定为 29 的情况下，对大核分解数量进行消融研究的结果。结果表明，将大核分解为两个深度可分离

表 2.14 LEVIR-CD 和 S2Looking 数据集上变化检测的定量比较结果。

模型	LEVIR-CD[197]				S2Looking[198]			
	Precision ↑	Recall ↑	F1 ↑	IoU ↑	Precision ↑	Recall ↑	F1 ↑	IoU ↑
FC-EF[199]	86.91	80.17	83.40	71.53	81.36	8.95	7.65	8.77
FC-Siam-Conc[199]	91.99	76.77	83.69	71.96	**83.29**	15.76	13.19	15.28
FC-Siam-Di[199]	89.53	83.31	86.31	75.92	68.27	18.52	13.54	17.05
STANet[197]	83.81	91.00	87.26	77.40	38.75	56.49	45.97	29.84
DTCDSCN[200]	88.53	86.83	87.67	78.05	68.58	49.16	57.27	40.12
HANet[201]	91.21	89.36	90.28	82.27	61.38	55.94	58.54	41.38
CDNet[202]	91.60	86.50	89.00	80.14	67.48	54.93	60.56	43.43
CDMC[203]	93.09	88.07	90.51	82.67	64.88	58.15	61.34	44.23
IFNet[204]	91.17	90.51	90.83	83.22	66.46	61.95	64.13	47.19
SNUNet[205]	92.45	90.17	91.30	83.99	71.94	56.34	63.19	46.19
BiT[12]	91.97	88.62	90.26	82.26	74.80	55.56	63.76	46.80
HCGMNet[206]	92.96	90.61	91.77	84.79	72.51	57.06	63.87	46.91
ChangeFormer[11]	92.59	89.68	91.11	83.67	72.82	56.13	63.39	46.41
C2FNet[207]	93.69	89.47	91.83	84.89	74.84	54.14	62.83	45.80
CGNet[196]	93.15	90.90	92.01	85.21	70.18	59.38	64.33	47.41
DiFormer[195]	**93.75**	90.59	92.15	85.44	72.39	61.19	66.31	49.60
Changer-MiT_b0[45]	93.61	90.56	92.06	85.29	73.01	62.04	67.08	50.47
LSKNet-T	92.56	**91.83**	92.19	85.51	70.44	**64.46**	67.32	50.74
LSKNet-S	93.34	91.23	**92.27**	**85.65**	71.90	63.64	**67.52**	**50.96**

表 2.15 **分解大型卷积核数量对推理 FPS 和 mAP 的影响**，理论感受野为 29。将大型卷积核分解为两个深度可分离卷积核可以在速度和精度方面取得较好平衡。

(k, d) 序列	RF	Num.	FPS	mAP (%)
(29, 1)	29	1	18.6	80.66
(5, 1) ⟶ (7, 4)	29	2	**20.5**	**80.91**
(3, 1) ⟶ (5, 2) ⟶ (7, 3)	29	3	19.2	80.77

大核可以在速度和精度之间取得良好的平衡，在 FPS（每秒帧数）和 mAP（平均精度均值）方面均表现较好。

核感受野大小。基于表 2.15中的评估结果，本章发现将大核分解为两个**串联**的深度可分离卷积核是较优的策略。此外，表 2.16显示，过小或过大的感受野都会影响 LSKNet 的性能，而约为 23 的感受野大小在该设置下较为有效。

（二）空间选择中的池化层

本章进行了实验以分析空间选择中的池化层配置，结果如表 2.17所示。实验表明，在 LSK 模块的空间选择模块中同时使用最大池化和平均池化可以在不牺牲推理速度的情况下获得较好性能。

表 2.16 **LSKNet 关键设计组件的有效性**，大型卷积核被分解为两个深度可分离卷积核序列。CS：通道选择；SS：空间选择（**本章方法**）。当使用具有空间选择的合理大感受野时，LSKNet 取得较好性能。

(k_1, d_1)	(k_2, d_2)	Flow	CS	SS	RF	FPS	mAP	
(3, 1)	(5, 2)	Series	-	-	11	22.1	80.80	
(5, 1)	(7, 3)	Series	-	-	23	21.7	80.94	
(5, 1)	(7, 4)	Series	-	-	29	20.5	80.91	
(7, 1)	(9, 4)	Series	-	-	39	21.3	80.84	
(3, 1)	(5, 1)	Parallel	✓	-	5	23.3	80.19	(SKNet[106])
(5, 1)	(7, 3)	Series	✓	-	23	19.6	80.57	
(5, 1)	(7, 3)	Series	✓	✓	23	18.6	80.82	
(5, 1)	(7, 3)	Series	-	✓	23	20.7	**81.31**	(LSKNet)

表 2.17 关于本章提出的 LSK 模块中**最大和平均池化对空间选择有效性**的消融研究。结果表明，同时使用两种池化方法可获得较好效果。

池化		FPS	mAP (%)
最大	平均		
✓		20.7	81.23
	✓	20.7	81.12
✓	✓	20.7	**81.31**

（三）与其他大核/选择性注意力骨干网络的比较

本章还将 LSKNet 与 9 种流行的大核或选择性注意力骨干网络进行了比较。如表 2.18所示，使用原始 ViT[70]骨干网络的 ViTDet[145]在所比较的模型中具有较大的模型规模和计算复杂度，但在本章评估任务中表现有限。表 2.5中的观察结果显示，它在具有明显细粒度特征的目标（如球场和直升机）上表现尤其差。这表明单纯依赖全局上下文信息建模并不总能适配遥感场景。在相似或更少的模型规模和复杂度预算下，本章提出的 LSKNet 在遥感目标检测（DOTA-v1.0）、分割（Vaihingen）和变化检测（LEVIR-CD）任务上取得了较高结果，体现了其在捕获和处理遥感图像语义特征方面的作用。

LSKNet 与 SKNet 有两个关键区别。首先，本章提出的选择机制依赖于通过核分解实现的一**序列**大核的显式特征流动，这与大多数现有基于注意力方法的做法不同。相比之下，SKNet 采用了**并行**分解技术。其次，LSKNet 在空间维度上自适应地聚合大核信息，而非 SKNet 或 LSKNet-CS 所使用的通道维度。这

表 2.18 LSKNet-S 与其他（大型卷积核或动态/选择性注意力）骨干网络在遥感目标检测（DOTA-v1.0）、分割（Vaihingen）和变化检测（LEVIR-CD）任务上的比较。在相似或更低的复杂度预算下，本章提出的 LSKNet 取得了较高的 mAP。

类别	模型 骨干网络	参数量	Flops	DOTA-v1.0			Vaihingen			LEVIR-CD			
				mAP	@50	@75	F1	OA	mIoU	P.	R.	F1	IoU
基准模型	ResNet-18	11.2M	38.1G	50.54	79.27	55.33	90.15	92.62	82.47	92.97	90.61	91.77	84.80
大核	ViTDet[145]	86.6M	394.9G	45.60	74.41	49.39	81.01	83.74	54.91	80.72	90.59	85.37	74.48
	ConvNeXt v2-N[161]	15.0M	51.2G	52.91	80.81	58.58	89.13	92.15	81.17	93.12	89.73	91.39	84.15
	Swin-T[72]	28.3M	91.1G	51.54	80.81	56.71	90.74	93.01	83.40	93.04	90.25	91.63	84.55
	MSCAN-S[88]	13.1M	45.0G	52.52	81.12	57.92	91.16	93.04	84.10	93.39	91.14	92.25	85.62
	VAN-B1[87]	13.4M	52.7G	52.69	81.15	58.11	91.30	93.12	84.41	93.31	91.20	92.24	85.60
动态/选择性注意力	ResNeSt-14[66]	8.6M	57.9G	49.79	79.51	53.41	90.31	92.84	82.72	92.47	90.38	91.41	84.18
	SCNet-18[67]	14.0M	50.7G	49.91	79.69	53.55	90.50	92.97	83.04	92.03	91.27	91.65	84.58
	DCN-Res50[208]	26.2M	121.2G	49.26	79.74	52.97	90.93	93.07	83.72	92.84	90.67	91.74	84.74
	SKNet-26[106]	14.5M	58.5G	51.53	80.67	56.51	90.83	93.01	83.56	93.09	91.09	92.08	85.32
本章方法	LSKNet-S	14.4M	54.4G	**53.32**	**81.48**	**58.83**	**91.81**	**93.61**	**85.12**	**93.44**	**91.13**	**92.27**	**85.65**

表 2.19 LSKNet-T 与 ResNet-18 作为骨干网络在 DOTA-v1.0 数据集上不同检测框架中的比较。LSKNet-T 在各种框架中均实现了更高的 mAP。

框架	ResNet-18	LSKNet-T
ORCNN[112]	79.27	81.31 **(+2.04)**
RoI Trans.[146]	78.32	80.89 **(+2.57)**
S^2A-Net[113]	76.82	80.15 **(+3.33)**
R3Det[114]	74.16	78.39 **(+4.23)**
参数量 (仅骨干网络)	11.2M	4.3M **(-62%)**
FLOPs (仅骨干网络)	38.1G	19.1G **(-50%)**

种设计对遥感任务而言更为直观和有效，因为通道选择无法捕捉图像空间中不同目标的空间尺度变化。此外，本章还评估了一种同时利用空间和通道选择的 LSKNet 变体。表 2.16中的实验结果表明，在检测任务中，空间信息起着更为关键的作用。然而，同时包含空间和通道选择可能会增加模型优化的难度，导致性能略有下降。

（四）LSKNet 骨干网络在不同检测框架下的性能

为验证所提出的 LSKNet 骨干网络的通用性和有效性，本章在多种遥感检测框架下评估了其性能，包括两阶段框架 O-RCNN[112]和 RoI Transformer[146]，以及单阶段框架 S^2A-Net[113]和 R3Det[114]。表 2.19中的结果显示，与 ResNet-18 相

比，本章提出的 LSKNet-T 骨干网络提高了检测性能，同时仅使用了 38% 的参数量和 50% 的 FLOPs。这些发现说明 LSKNet 骨干网络具有较好的轻量化和通用性特征。

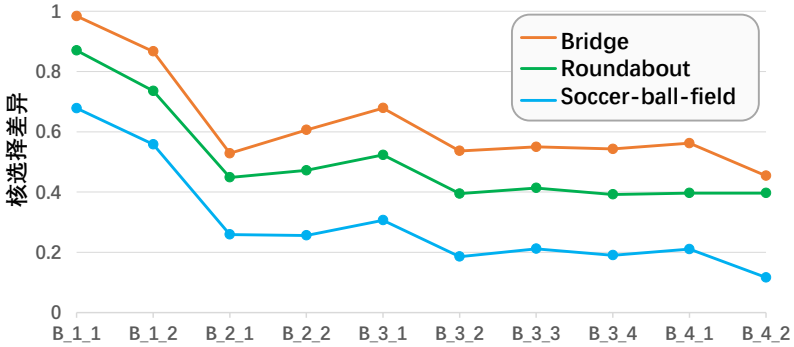


图 2.6 LSKNet-T 块中桥梁、环岛和足球场的归一化核选择差异。B_i_j 表示第 i 阶段的第 j 个 LSK 块。较大的值表示对更广泛上下文的依赖。

（五）核选择机制

本章进一步研究了 LSKNet 中的核选择机制。对于目标类别 c，LSKNet-T 结构块的核选择差异 ΔA_c（即大核选择 − 小核选择）定义如下：

$$\Delta A_c = |\widetilde{\mathbf{SA}}_{larger} - \widetilde{\mathbf{SA}}_{smaller}| \,. \tag{2.11}$$

图 2.6展示了三个典型类别（桥梁、环岛和足球场）在所有图像上的归一化 ΔA_c，以及每个 LSKNet-T 块的 ΔA_c。如预期所示，桥梁类别在所有块中的 ΔA_c 平均比环岛高约 30%，而环岛又比足球场高约 70%。这与常识相符，即足球场确实不需要大量上下文信息，因为其自身的纹理特征已经足够独特和具有辨识度。

本章还意外发现了 LSKNet 在网络深度上的另一种选择模式：LSKNet 通常在浅层使用较大的核，而在高层使用较小的核。第一层块的平均 ΔA_c 为 0.78，第二和第三层块为 0.40，最后一层块仅为 0.33。这表明网络倾向于在低层快速聚焦于捕获大感受野的信息，以便高层语义能包含足够的感受野，从而实现更好的区分。

（六）可视化分析

检测结果可视化。图 2.7展示了检测结果和 Eigen-CAM[209]的可视化。LSKNet 能够捕获与检测目标相关的合理范围的上下文信息，从而在多种困难情况下表

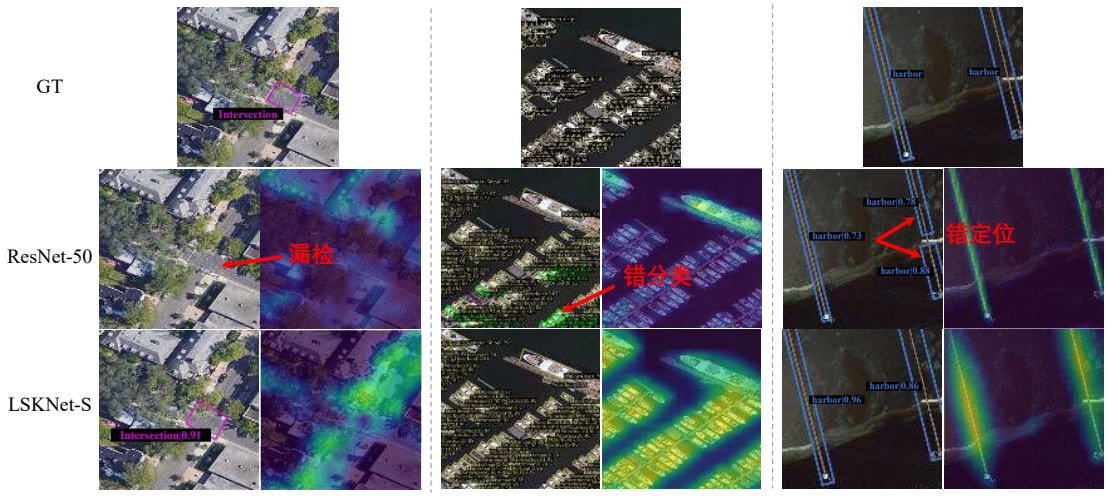


图 2.7 Eigen-CAM 可视化：基于 ResNet-50 和 LSKNet-S 的 Oriented RCNN 检测框架。本章提出的 LSKNet 能够建模合理长程的上下文信息，在各种困难情况下表现更佳。

现更好，这与本章的先验假设 *1)* 相一致。相比之下，ResNet 通常只能捕获有限范围的上下文信息，在目标小而密集时难以建模细粒度细节，在具有挑战性的场景中表现有限。

不同目标的相对上下文范围。为研究每个目标类别的感受野相对范围，本章定义了 R_c 作为类别 c 的预期选择性感受野面积与真实边界框面积之比：

$$R_c = \frac{\sum_{i=1}^{I_c} A_i / B_i}{I_c}, \tag{2.12}$$

$$A_i = \sum_{d=1}^{D}\sum_{n=1}^{N} |\widetilde{\mathbf{SA}}_n^d \cdot RF_n|,\ B_i = \sum_{j=1}^{J_i} Area(\mathrm{GT}_j), \tag{2.13}$$

其中，I_c 是仅包含目标类别 c 的图像数量。A_i 是输入图像 i 在所有 LSK 块中空间选择激活的总和，D 是 LSKNet 中的块数，N 是 LSK 模块中分解大核的数量。B_i 是所有 J_i 个标注的定向目标边界框（真实值）的总像素面积。图 2.8中归一化的 R_c 直观地展示了不同目标类别所需的相对上下文范围。结果表明，桥梁类别相比其他类别需要更多的额外上下文信息，这主要是由于其特征与道路相似，且需要上下文线索来确定其是否被水包围。同样，环岛类别也有相对较高的 R_c 值，为 0.57。相反，球场类别的 R_c 值相对较低，均低于 0.1。由于其独特的纹理属性，特别是场地边界线，它们只需要最少的上下文信息。这与认知相符，进一步支持了本章的先验假设 *2)*，即不同目标类别所需的上下文信息相对范围差异

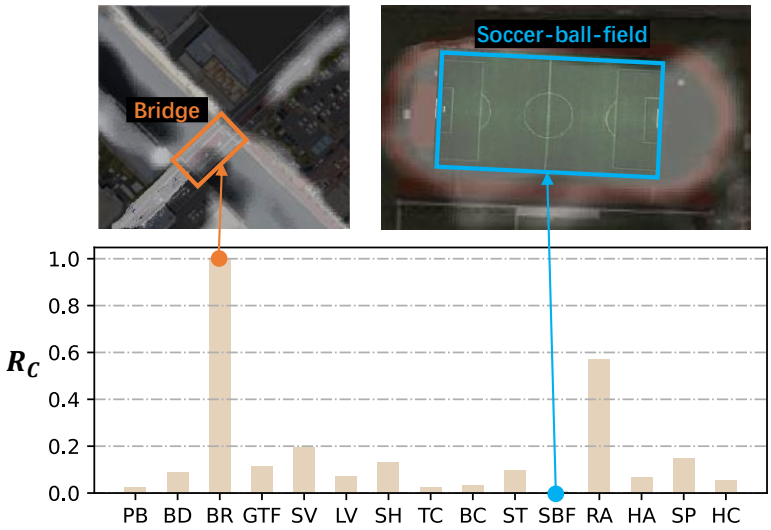


图 2.8 DOTA-v1.0 数据集中各目标类别的预期选择性感受野面积与真实边界框面积的归一化比率 R_c。不同目标类别所需的相对上下文范围差异明显。本章使用公式 (2.8)（即空间激活）可视化训练后的 LSKNet 模型的感受野。

很大。

空间激活图可视化。图 2.9展示了 DOTA-v1.0 数据集中更多目标类别的空间激活图示例，其中激活图是利用训练后的 LSKNet 模型通过公式(2.8)（即空间激活）计算得到。目标类别按照图 2.8所示的预期选择性感受野面积与真实边界框面积之比从左上到右下依次递减排列。空间激活图可视化结果显示，模型行为与本章挖掘出的两个先验假设及上述分析基本一致，为所提出机制提供了定性证据。

第五节 本章小结

本章提出了轻量级自适应大核网络（LSKNet），作为遥感图像分析下游任务（如场景分类、目标检测和语义分割）的骨干网络。LSKNet 面向遥感图像中的广域上下文需求，通过采用大空间感受野和空间选择机制，捕获和建模不同目标类型所呈现的多样化上下文细节。实验结果表明，本章提出的轻量级模型在多个遥感基准上取得了具有竞争力的结果。消融与可视化分析也从不同侧面支持了空间选择大核机制的有效性。

本章工作仍存在一定局限。LSKNet 主要从骨干网络结构角度建模遥感图像中的空间上下文先验，对极端密集小目标、严重遮挡或低质量成像条件下的实

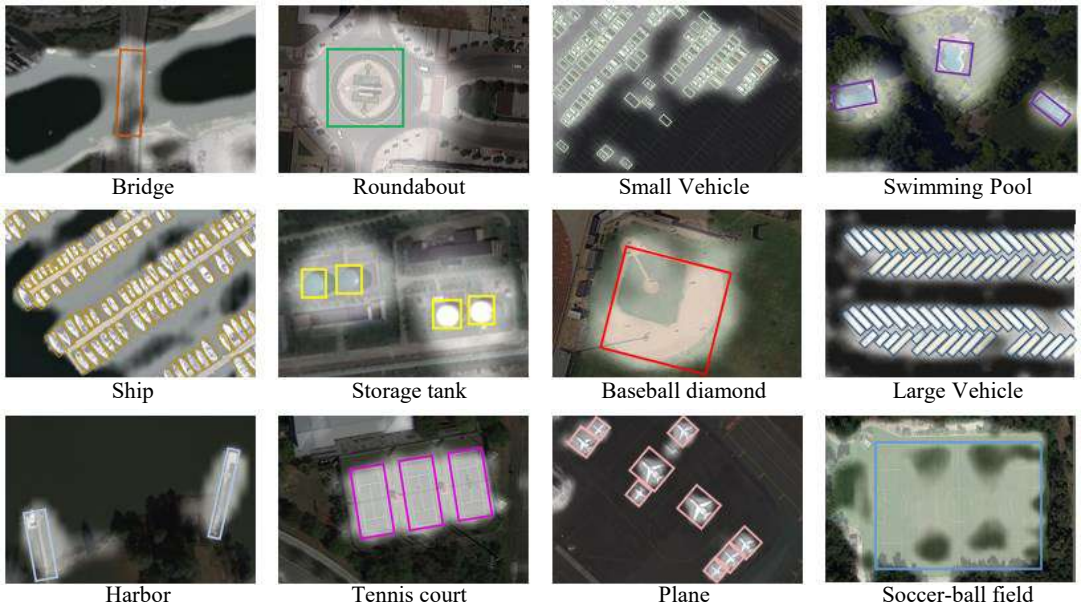


图 2.9 DOTA-v1.0 数据集中更多目标类别的感受野激活图，其中本章使用公式 (2.8)（即空间激活）可视化训练后的 LSKNet 模型的激活图。

例级判别仍依赖检测头、数据增强和训练策略的配合。后续可进一步结合更大规模预训练、更细粒度目标关系建模和可解释性评估，提升空间先验建模的稳定性与适用范围。

第三章 大规模 SAR 目标检测基准构建与域适应预训练

SAR 目标检测因合成孔径雷达在全天时、全天候观测中的独特优势而受到广泛关注。然而，该方向长期受限于公共数据集规模小、类别覆盖有限、评测协议不统一以及开源生态薄弱等问题。为缓解上述瓶颈，本章围绕“大规模基准构建”和“跨域连续预训练”两个方面展开研究。首先，本文系统梳理、收集并标准化 10 个公开 SAR 检测数据集，构建了大规模多类别 SAR 检测基准 SARDet-100K。该数据集提升了 SAR 检测研究的规模与可比性。进一步地，基于 SARDet-100K，本章揭示了自然图像预训练模型迁移到 SAR 检测任务时同时存在领域差距与模型差距，并据此提出多阶段滤波增强（Multi-Stage Filter Augmentation, MSFA）预训练框架，从输入适配、领域过渡与检测器迁移三个层面缩小跨域鸿沟。实验结果表明，MSFA 能够稳定提升 SAR 目标检测性能，并在不同骨干与检测框架上表现出良好的泛化能力。

第一节 引言

第三章对应绪论中感知层挑战的第二项研究问题，即如何围绕 SAR 目标检测构建大规模标准化基准并设计有效的跨域预训练迁移路径。第二章已经从骨干网络层面说明，遥感感知模型必须具备适应复杂空间先验的能力。但对 SAR 场景而言，仅有更适配的骨干并不足以释放性能潜力，数据资源稀缺、基准标准不统一以及自然图像到 SAR 图像的明显域鸿沟，仍然是制约模型发展的关键瓶颈。

合成孔径雷达 (Synthetic Aperture Radar，SAR)[210-211] 是遥感领域的一项关键技术。相较于传统光学传感器，它能够在全天时、全天候条件下稳定成像，不受光照、云层和部分伪装条件的限制，如图 3.1(a) 所示。因此，SAR 已广泛应用于国防安全、人道主义救援、伪装识别和地质勘探等关键场景[1-6]。随着应用需求持续增长，SAR 目标检测研究快速升温，如图 3.1(b) 所示，但研究生态与模型能力之间仍存在明显落差。

基准缺口 高分辨率 SAR 图像高度敏感、采集与标注成本高昂，这直接限制了公共数据集的建设。现有数据集如 SAR-AIRcraft、Air-SARShip、SSDD 和

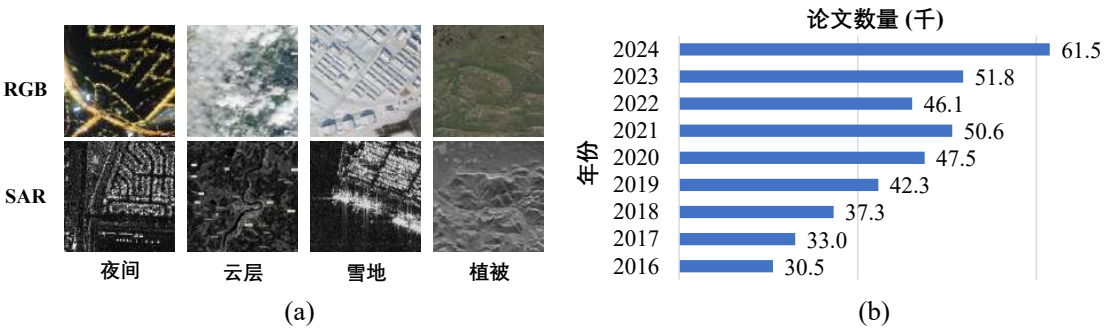


图 3.1 **(a)** SAR 图像的优势：不受天气条件、阳光和地表覆盖影响。**(b)** 从谷歌学术检索，关键词为“SAR detection”的论文数量 (千篇)。

HRSID 等[212-215]，往往规模较小、类别单一，且场景复杂度有限。这种“小而散”的数据生态不仅难以支撑强表征学习，也容易在算法比较中引入偶然性和偏差。与此同时，相关代码和标准化评测流程的缺失，也进一步削弱了该方向的可复现性与社区累积效应。

为解决这一问题，本章首先构建 SARDet-100K。该数据集通过系统梳理、收集并标准化 10 个公开 SAR 检测数据集形成统一基准，最终包含约 11.7 万张图像和 24.6 万个目标实例，覆盖 6 类典型目标。就规模而言，SARDet-100K 将 SAR 目标检测推进到接近 COCO 量级的研究范式。就数据组织而言，它显式解决了不同数据源在标注格式、划分方式和统计口径上的不一致问题，从而为后续模型研究提供更稳定、更公平的实验平台。

迁移缺口 在建立统一基准之后，另一个更隐蔽但更关键的问题随之显现：当研究者将 ImageNet 等自然图像数据上预训练的骨干迁移到 SAR 检测器时，往往会同时遭遇领域差距和模型差距。前者来自 RGB 反射成像与 SAR 散射成像之间的视觉统计差异，后者则来自“仅预训练骨干”与“下游使用完整检测器”之间的结构不一致。这意味着，传统“自然图像预训练 + SAR 微调”的流程并未真正针对 SAR 检测任务做出适配。

为缩小上述双重鸿沟，本章进一步提出多阶段滤波增强（Multi-Stage Filter Augmentation, MSFA）预训练框架，从数据输入、领域过渡和模型迁移三个层面协同设计连续预训练路径。对于数据输入，本章引入手工特征描述符，将原始像素映射到更稳健的特征空间，以减弱 SAR 噪声并缩小 RGB 与 SAR 之间的输入分布差异。对于领域过渡，本章构建了以光学遥感检测数据为桥梁的分层连续预训练过程，使自然图像知识能够经由“与任务更接近”的遥感光学语义平

滑迁移到 SAR 领域。对于模型迁移，本章不再局限于骨干初始化，而是在多阶段预训练中持续使用完整检测器，以减少预训练结构与微调结构之间的偏差。

因此，本章并非简单地“再做一个 SAR 检测模型”，而是试图同时回答两个更基础的问题：领域内究竟需要什么样的基准，模型又应通过怎样的连续预训练路径跨越自然图像与 SAR 之间的鸿沟。MSFA 的意义也正在于此，它把数据、任务和模型迁移三者放在同一框架内统一考虑。

本章对 SAR 目标检测领域的贡献可以总结为以下四点：

- 构建了首个 COCO 量级的大规模多类别 SAR 目标检测基准 SARDet-100K。
- 系统揭示了传统“自然图像预训练 +SAR 微调”范式在 SAR 目标检测中同时存在的领域差距与模型差距。
- 提出了多阶段滤波增强（MSFA）预训练框架，并在多种骨干网络与检测框架上检验了其适用性与稳定收益。
- 通过公开数据集、代码与评测设置，建立了更规范、更可复现的 SAR 目标检测研究基线。

第二节　相关研究与问题分析

一、SAR 成像特点与手工特征

SAR 图像由散射机理形成，与光学图像在成像物理基础上存在本质差异，常伴随乘性斑点噪声、局部伪影和较弱的纹理可分性[210,216]。为了缓解这个问题，早期研究开发或改进了许多传统的手工特征描述符，以便从 SAR 图像中提取更易于辨别的特征。这些描述符包括方向梯度直方图[13] (Histogram of Oriented Gradients，HOG)、Canny 边缘检测器[14]、比率梯度边缘 (Gradient by Ratio Edge，GRE)[217]、Haar-like[15] 特征描述符和小波散射变换[16] (Wavelet Scattering Transform，WST)。早期工作采用了传统算法，例如 HOG 用于 SAR 目标识别[218-219]，Canny[220-221] 用于边缘检测。然而，近年来，SAR 图像分析领域已在很大程度上被深度学习方法所主导。

在从自然图像迁移到 SAR 场景时，若仍直接使用 RGB 统计特性主导的输入假设，模型往往难以稳定继承已有预训练知识。因此，如何把手工先验转化为跨域迁移中的输入桥梁，而不是仅作为独立特征工程，是 SAR 目标检测中的一个关键问题。

二、SAR 目标检测方法

各种流行的基于深度学习的目标检测框架，包括 RetinaNet[18]、FCOS[19]、GFL[222]、RCNN 系列[17,154]、YOLO 系列[223-224] 和 DETR[20]，在通用目标检测领域展现出了较好的通用性。此外，诸如 ConvNext[225]、VAN[226] 和 Swin Transformer[72] 等现代骨干网络旨在高效且有效地建模视觉特征。然而，由于 SAR 图像固有的诸如小目标尺寸、斑点噪声和稀疏信息等因素，SAR 图像目标检测提出了独特的挑战。因此，最近用于 SAR 目标检测的深度学习方法主要集中于网络和模块设计，以应对这些挑战。诸如 MGCAN[227]、MSSDNet[228] 和 SEFEPNet[229] 等方法通过多尺度特征融合来增强目标特征。Quad-FPN[230] 结合了四个不同的特征金字塔网络，用于全面的多尺度特征交互，以减轻噪声干扰和多尺度目标特征错位。PADN[231] 和 EWFAN[232] 采用注意力机制来增强存在 SAR 斑点噪声情况下的目标特征。CenterNet++[233] 是 CenterNet[234] 的扩展，它融入了特征增强、多尺度融合和头部细化模块，以提高检测器针对 SAR 图像的鲁棒性。此外，CRTransSar[235] 构建于高性能 Swin transformer[72] 之上，利用上下文表示学习来增强目标特征。

总体来看，现有 SAR 检测方法主要聚焦于“如何设计更适配 SAR 场景的检测器”，而对“这些检测器应建立在什么样的数据基础和迁移路径之上”关注不足。这导致算法性能常与具体数据集规模、划分策略和初始化方式强耦合，难以形成可稳定复用的研究基线。

三、公开基准与统一评测现状

与自然图像检测相比，SAR 目标检测长期缺少大规模、标准化、可复现的公共基准。现有代表性数据集如 SAR-AIRcraft、AIR-SARShip、SSDD 和 HRSID 等[212-215]，在推动早期研究方面发挥了重要作用，但普遍存在规模较小、类别有限、场景复杂度不足以及统计口径不统一等问题。不同数据源在图像分辨率、目标定义、标注格式与训练验证划分上的差异，也使得方法间的直接比较常常缺乏可比性。

因此，SAR 目标检测所面临的第一个基础性缺口并不是某个单点算法模块，而是缺少一个覆盖更广目标类别、更统一评测协议且便于社区复现的大规模基准。只有先解决“评什么、怎么评、在什么数据上评”这一问题，后续模型改进的真实贡献才能被更清晰地辨别。

四、问题分析与本章定位

在实际训练流程中，SAR 检测模型通常仍以自然图像预训练权重作为初始化。然而，持续预训练和领域适配的相关研究已经表明，预训练收益高度依赖于中间领域的过渡设计与目标任务的一致性[32,236-237]。对于 SAR 目标检测而言，这一过程至少同时包含两个相互耦合的鸿沟：一是 RGB 反射成像与 SAR 散射成像之间的领域差距，二是“仅预训练骨干”与“下游使用完整检测器”之间的模型差距。现有工作往往默认这两类差距可以通过一次微调同时解决，但这一假设在 SAR 场景中并不充分。

基于上述分析，本章的研究定位并非再提出一个局部改进的 SAR 检测器，而是围绕 SAR 目标检测补齐“统一基准 + 连续迁移”两块基础能力。一方面，SARDet-100K 用于建立规范的评测环境。另一方面，MSFA 通过输入适配、领域过渡与完整检测器迁移三层设计，系统缩小自然图像到 SAR 目标检测的跨域鸿沟。这一章也由此为后续统一多模态检测和跨模态预训练提供数据与迁移基础。

第三节 SAR 目标检测新基准数据集

一、当前现状

SAR 图像通常由卫星捕获，并且有大量的低分辨率 SAR 图像可用，通常地面采样距离 (GSD) 为 10 米 × 10 米或更大。诸如 Sentinel-1[238] 等平台提供了对这些图像的访问，这些图像提供了各种地球物理场所（如城市、山脉、河流和耕地）的宏观视图。这使得它们对于场景分类任务特别有利。然而，这些图像固有的低分辨率限制了它们描绘较小物体的精细细节的能力，例如船舶、汽车和飞机。相反，高分辨率 SAR 图像提供更详细的信息，但需要大量的硬件资源。此外，这些高清图像通常包含敏感信息，因此不适合公开发布。而且，获取高分辨率 SAR 数据集可能非常昂贵，对其可访问性构成重大挑战。

许多研究团队经常遇到预算限制，这限制了他们获取大量且多样化的高分辨率 SAR 数据集的能力。这些财务约束不仅限制了可以覆盖的地理区域范围，还会影响可以访问的数据源的多样性。因此，这些团队提供的数据集通常缺乏多样性，尤其是在光谱波段、极化和分辨率等方面。从研究人员的角度来看，在如此小而同质的数据集上评估模型可能会引入偏差，并导致不公平的性能比较。

表 3.1 SARDet-100K 数据集的图像和实例级别统计信息。*: 原始数据集被裁剪成 512 × 512 的图像块。

	图片				实例				
数据集	Train	Val	Test	ALL	Train	Val	Test	ALL	实例/图
AIR_SARShip 1*[213]	438	23	40	501	816	33	209	1,058	2.11
AIR_SARShip 2[213]	270	15	15	300	1,819	127	94	2,040	6.80
HRSID[215]	3,642	981	981	5,604	11,047	2,975	2,947	16,969	3.03
MSAR*[235]	27,159	1,479	1,520	30,158	58,988	3,091	3,123	65,202	2.16
SADD[239]	795	44	44	883	6,891	448	496	7,835	8.87
SAR-AIRcraft*[212]	13,976	1,923	2,989	18,888	27,848	4,631	5,996	38,475	2.04
ShipDataset[47]	31,784	3,973	3,972	39,729	40,761	5,080	5,044	50,885	1.28
SSDD[214]	928	116	116	1,160	2,041	252	294	2,587	2.23
OGSOD[240]	14,664	1,834	1,833	18,331	38,975	4,844	4,770	48,589	2.65
SIVED[241]	837	104	103	1,044	9,561	1,222	1,230	12,013	11.51
SARDet-100K	94,493	10,492	11,613	116,598	198,747	22,703	24,023	245,653	2.11

表 3.2 SARDet-100K 数据集来源信息。GF-3: 高分三号，S-1: Sentinel-1。目标类别 S: 船舶，A: 飞机，C: 汽车，B: 桥梁，H: 港口，T: 坦克。

数据集	目标	分辨率	波段	极化	卫星	协议
AIR_SARShip[213]	S	1,3m	C	VV	GF-3	-
HRSID[215]	S	0.5∼3m	C/X	HH, HV, VH, VV	S-1B,TerraSAR-X,TanDEMX	GNU General Pub.
MSAR[235]	A, T, B, S	≤ 1m	C	HH, HV, VH, VV	HISEA-1	CC BY-NC 4.0
SADD[239]	A	0.5∼3m	X	HH	TerraSAR-X	-
SAR-AIRcraft[212]	A	1m	C	Uni-polar	GF-3	CC BY-NC 4.0
ShipDataset[47]	S	3∼25m	C	HH, VV, VH, HV	S-1,GF-3	-
SSDD[214]	S	1∼15m	C/X	HH, VV, VH, HV	S-1,RadarSat-2,TerraSAR-X	Apache2.0
OGSOD[240]	B, H, T	3m	C	VV/VH	GF-3	-
SIVED[241]	C	0.1,0.3m	Ka,Ku,X	VV/HH	Airborne SAR synthetic slice	-

二、SARDet-100K 基准数据集

为了应对上述挑战，本章对公开 SAR 目标检测数据集进行了系统梳理，并从中筛选、收集了 10 个高质量数据源。这些数据由不同国家和机构发布，覆盖了多种成像条件、空间分辨率、极化方式和目标类型，且在类别定义上可以被统一整合。表 3.2 给出了各来源数据集的基本信息。

在此基础上，本章以“统一评测协议”为目标，对所有数据源执行严格标准化处理。表 3.1 展示了标准化后的子数据集统计。最终构建得到的 SARDet-100K 共包含 116,598 张图像和 245,653 个实例，覆盖飞机、船舶、汽车、桥梁、坦克和港口六类典型目标。无论从图像规模还是类别覆盖来看，SARDet-100K 都超过此前常用的小规模 SAR 检测数据集，并将该方向推进到接近 COCO[25] 级别

的研究范式。

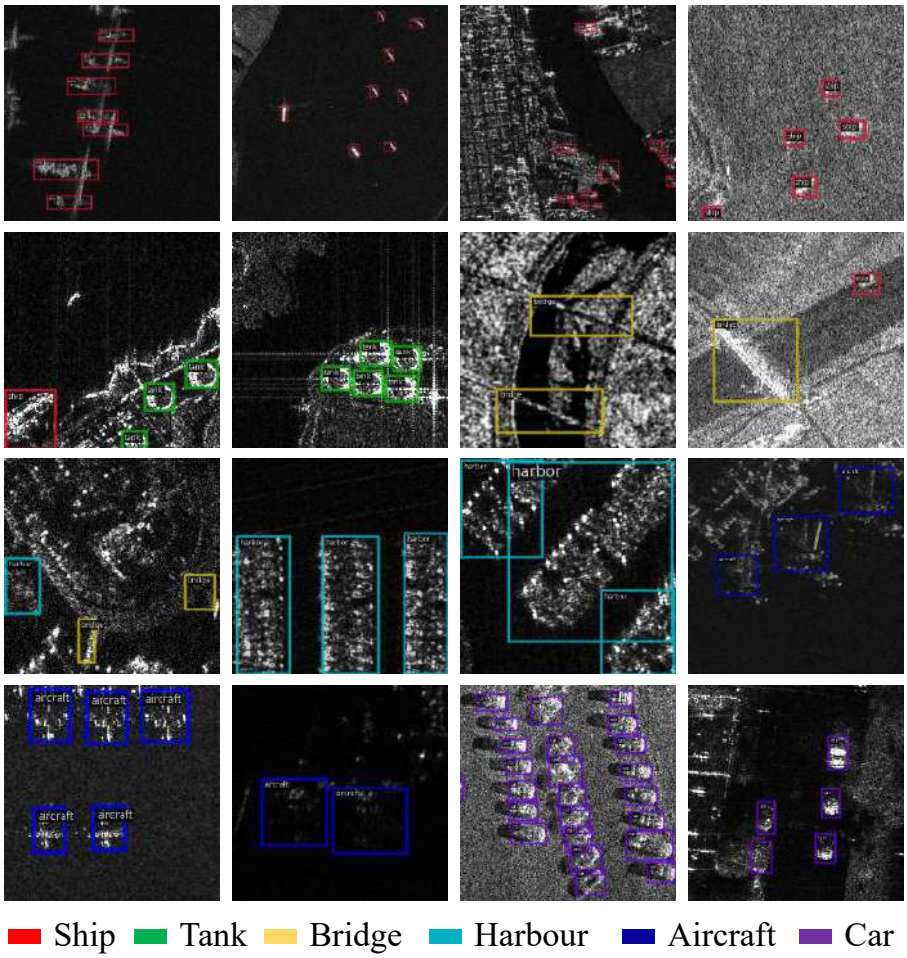
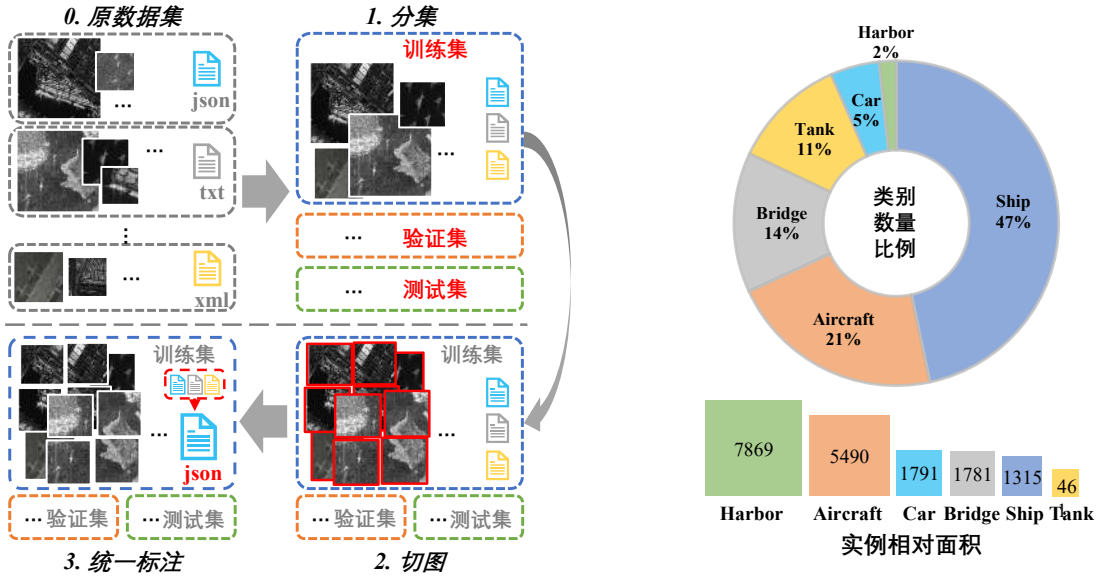


图 3.2 SARDet-100K 数据集样本图像可视化。

图 3.2 展示了 SARDet-100K 中具有代表性的样本。可以看到，不同子数据集之间在分辨率、成像条件、目标密度和背景复杂度上都存在明显差异，这也是构建统一基准必须首先解决的问题。

具体而言，如图 3.3(a) 所示，若原始数据集已经提供了固定的训练集、验证集和测试集划分，则本章沿用其官方协议，否则统一按照 8:1:1 的比例重新划分。对于分辨率高于 1000×1000 的图像，本章执行大图切片，以避免下采样后目标过小导致关键信息丢失。针对 AIR-SARShip 1、MSAR 和 SAR-AIRcraft 等数据集，本章将原图裁剪为 512×512 的图像块，并设置 200 像素重叠。最后，所有标注均被转换为 COCO 格式，以消除不同数据集之间在标注规范和评测接口上的差异。

图 3.3(b) 进一步给出了类别级统计信息。可以看到，六类目标在数量和尺度



(a) (b)

图 3.3 **(a)** SARDet-100K 数据集标准化流程，包括数据划分、大图像切片与标注格式统一。**(b)** SARDet-100K 中各类别实例占比与平均实例面积。

上存在明显不均衡，这更接近真实 SAR 应用场景，也使得 SARDet-100K 更适合作为大规模训练和稳健评测的统一平台。总体而言，SARDet-100K 并非简单拼接已有数据，而是通过标准化处理将多源异构数据转化为可用于系统研究的统一基准，为后续跨域预训练与模型分析提供了可靠基础。

第四节 滤波增强预训练框架

最近的一些研究[233,235,242-243] 已经说明成熟的手工特征和专门的网络模块设计能够提高 SAR 目标检测性能。然而，这些工作中的大多数都依赖于默认的 ImageNet 预训练方法，因此忽略了预训练的自然场景数据集与微调的 SAR 数据集之间存在的明显领域差距。此外，他们也未能解决骨干网络和整个检测框架之间存在的模型差距。为了解决这些局限性，本章提出了一种名为带有滤波增强的多阶段 (MSFA) 预训练框架的新框架。本章的框架从数据输入、领域过渡和模型迁移的角度应对挑战。MSFA 包含两个核心设计：滤波增强输入和多阶段预训练策略。

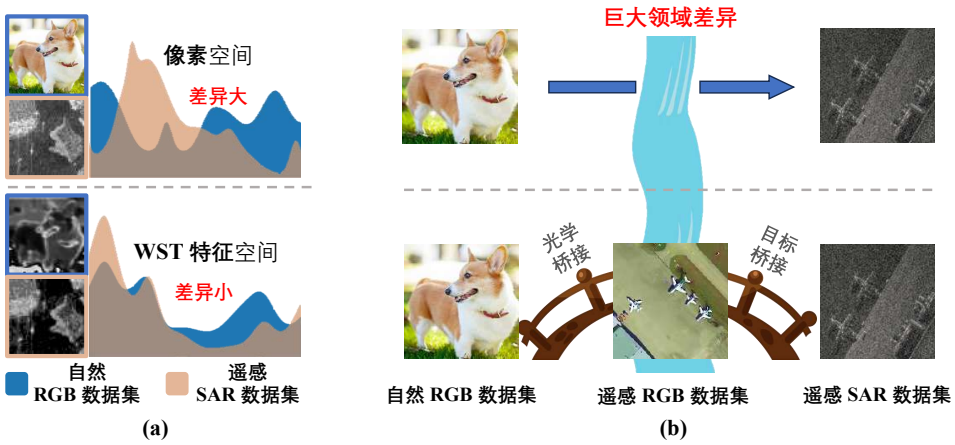


图 3.4 自然 RGB 数据集和遥感 SAR 数据集之间存在的明显领域差距。**(a)** 展示了 WST 特征空间缩小了领域差距。**(b)** 遥感 RGB 数据集充当了有效的领域过渡桥梁，促进更平滑的领域转移。

一、滤波增强输入

结合前述相关研究可知，许多手工特征描述符都依赖精心设计的滤波器来提取鲁棒表示。这些特征可以作为从原始图像中导出的增强信息。因此，使用此类特征作为原始像素数据的辅助信息。数据 x 的滤波增强特征 M 可以概括地定义为：

$$M_i^x = T_i(x), i \in \{HOG, Canny, Haar, WST, GRE\}. \tag{3.1}$$

其中，T_i 是预定义的变换。从 ResNet[22] 中的信息残差设计中汲取灵感，本章构建了检测模型的滤波增强输入 Inp，方法是将原始灰度 SAR 图像 x 与生成的滤波增强特征 M_i^x 连接起来，如下所示：

$$Inp = \text{concat}(x, M_i^x). \tag{3.2}$$

通过将原始数据输入从异构像素空间转换为同构滤波增强特征空间，可以大大缩小不同图像域之间的领域差距，如图 3.4(a) 所示。

本章重点考察了 HOG、Canny、Haar-like、WST 和 GRE 等典型手工特征。它们虽然形式各异，但都包含一个共同特征：通过显式滤波操作突出边缘、纹理、结构或多尺度散射信息，从而减弱原始 SAR 像素空间中噪声和散斑带来的干扰。其中，HOG 更关注局部梯度结构，Canny 与 GRE 更强调边缘响应，Haar-like 有助于编码局部对比模式，而 WST 同时保留了鲁棒的低频轮廓和具有判别性的高

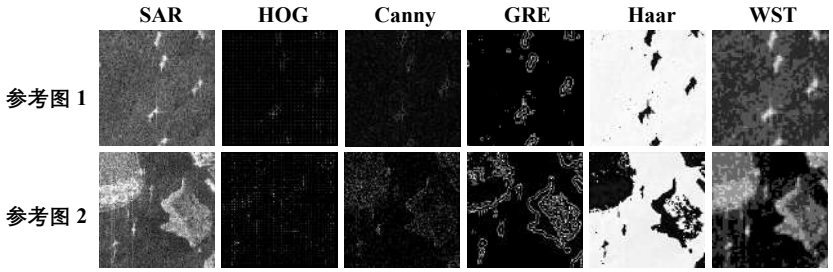


图 3.5 SAR 图像上不同手工特征的可视化。为便于展示，相关特征经过平均池化并以单通道形式呈现。

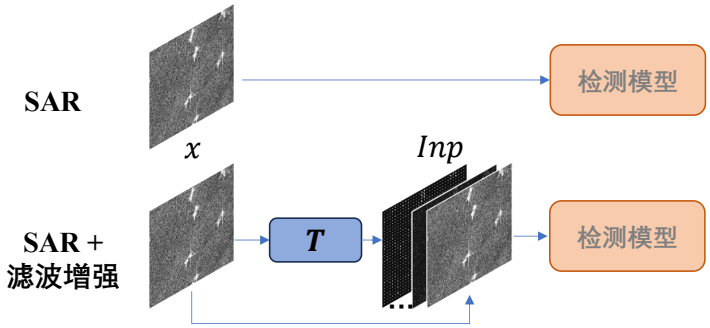


图 3.6 滤波增强输入的构造方式。

频信息。

图 3.5 给出了这些特征在 SAR 图像上的响应形式。可以看出，滤波后的特征图相比原始灰度图更突出目标边界、局部结构与尺度线索，这为后续的跨域迁移提供了更稳定的输入表示。基于这一观察，本章采用将原始 SAR 图像与滤波增强特征拼接的方式构造检测器输入，如图 3.6 所示。

二、多阶段预训练

本章将传统的预训练模式公式化为：

$$B = \mathrm{Train}_{cls}(B_\theta)(D_{IN}), \tag{3.3}$$

$$A = \mathrm{Train}_{det}(A_B)(D_{SAR}). \tag{3.4}$$

函数 $\mathrm{Train}_t(a)(b)$ 表示使用任务 t 在数据集 b 上训练模型 a，并返回训练后的模型。其中 t 是训练任务，$t \in \{cls, det\}$，cls 代表分类，det 代表检测。B 表示骨干模型，A 是完整的检测模型。传统上，预训练阶段会随机初始化骨干模型 B_θ，并在 ImageNet 数据集 D_{IN} 上进行训练（如公式 (3.3) 所示）。然后使用从检测模型 A_B 初始化预训练骨干网络的检测模型在 SAR 数据集 D_{SAR} 上进行微调（如公式 (3.4) 所示）。

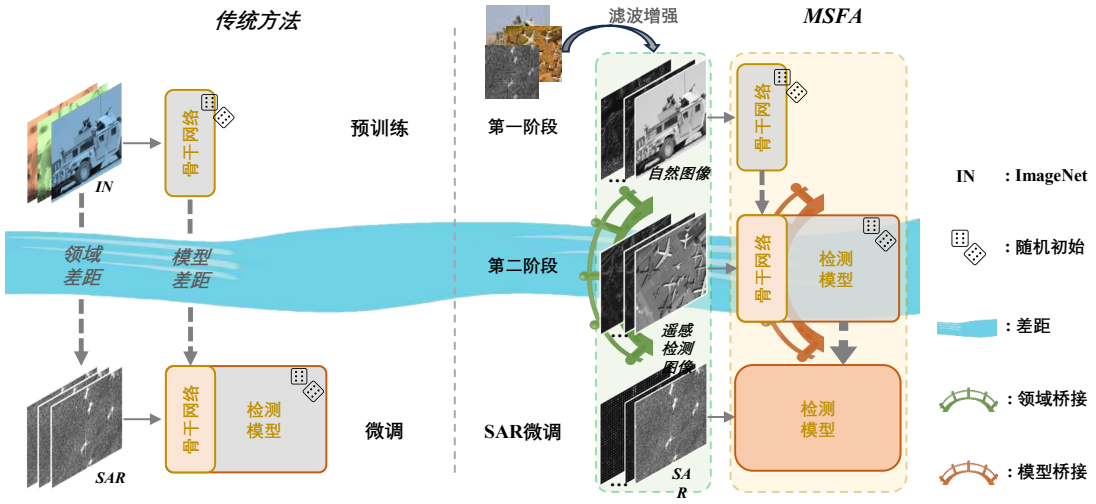


图 3.7 传统 ImageNet 预训练和本章提出的带有滤波增强的多阶段 (MSFA) 预训练框架的概念图。

本章提出的多阶段预训练策略，作为替代方案，可以如图公式 (3.3)、(3.5)、(3.6) 所示。

$$A' = \text{Train}_{det}(A_B)(D_{RS}). \tag{3.5}$$

$$A = \text{Train}_{det}(A_{A'})(D_{SAR}). \tag{3.6}$$

其中在公式 (3.5) 中增加了一个额外的第二阶段预训练。

本章建议利用大规模光学遥感数据集 D_{RS} 作为检测预训练以用于领域过渡。该数据集由光学模态图像组成，这些图像在下游 SAR 数据集中也共享相似的物体形状、尺度和类别。这一特性充当了 ImageNet 中自然图像的光学分布与 SAR 遥感图像中物体分布之间有价值的桥梁。通过利用这种第二阶段预训练，可以在一定程度上缩小领域差距，如图 3.4(b) 所示。

三、MSFA

最后，本章提出的 MSFA 框架集成了滤波增强输入与多阶段预训练，如图 3.7 所示。本章的 MSFA 框架用于缓解在自然图像上进行预训练和在 SAR 图像检测上进行微调之间存在的领域和模态差距。

通过引入滤波增强输入，本章利用成熟的手工特征描述符来提取对噪声鲁棒的特征。这也使其能够有效地将预训练和微调图像的异构图像域转换为同构特征域。通过将输入数据统一到一致的特征域中，本章解决了不同类型图像之

表 3.3 预训练与微调阶段的主要超参数设置。Cls. 表示分类，Det. 表示检测，B.S. 表示批大小，L.R. 表示学习率。

任务 / 模型	数据集	优化器	批大小	学习率	轮次
Cls. 预训练	ImageNet	AdamW	512	1e-8	100
Det. 预训练	DOTA	AdamW	16	1e-4	12
Det. 预训练	DIOR	AdamW	16	1e-4	12
Det. 预训练	SARDet-100K	AdamW	16	1e-4	12
Det. 预训练	SSDD	AdamW	32	2.5e-4	12
Det. 预训练	HRSID	AdamW	32	2.5e-4	12
DETR	DOTA/SARDet-100K	AdamW	16	1e-4	150
Deformable-DETR	DOTA/SARDet-100K	AdamW	16	2e-4	50
Dab-DETR	DOTA/SARDet-100K	AdamW	16	1e-4	50
Sparse-RCNN	DOTA/SARDet-100K	AdamW	16	2.5e-5	12

间存在的差异。因此，它增强了跨领域知识的对齐和可迁移性。此外，多阶段训练的结合包括利用额外的、大规模的光学遥感数据集进行检测预训练。该数据集充当领域桥梁，连接了 ImageNet 自然图像的领域与 SAR 遥感图像的领域。因此，它进一步缩小了领域差距，促进了两个领域之间更平滑的过渡。更重要的是，MSFA 框架第二阶段的检测预训练也可以充当模型桥梁。它允许对整个检测框架进行全面训练，而不仅仅是专注于骨干网络，使得整个检测框架得到良好的初始化，从而在 SAR 检测微调中取得较好性能。

第五节 实验与分析

一、实现细节

对于 ImageNet 预训练，本章采用 ImageNet-1K 上 100 个 epoch 的标准骨干预训练策略，并沿用 MMPretrain[244] 的默认配置。对于第二阶段检测预训练，本章主要在 DOTA 上训练，并用 DIOR 作为对照。考虑到 DOTA 原始图像分辨率差异较大，本章遵循[245] 的多尺度数据切分策略，将原图分别缩放到 0.5×、1.0× 和 1.5×，再裁剪为 1024 × 1024 的图块，图块间重叠 500 像素，以减少边界截断对目标的破坏。DIOR 则统一缩放到 800 × 800。SARDet-100K、SSDD 和 HRSID 上的微调也统一采用 800 × 800 输入与 0.5 概率的随机翻转。

所有实验主要基于 MMPretrain[244] 与 MMDetection[246] 实现，并在 8 张 RTX-3090 GPU 上完成。表 3.3 总结了主要超参数配置。总体上，本章保持了尽可能统一的训练范式，以保证不同预训练策略和不同检测框架之间的比较公平性。

二、主要实验

（一）滤波增强输入

Input	**mAP** ↑	mAP_{50} ↑
SAR (as RGB)	50.2	83.0
SAR+Canny	50.7	83.6
SAR+Hog	50.7	83.5
SAR+Haar	50.6	83.4
SAR+WST	**51.1**	83.9
SAR+GRE	50.6	83.8
SAR+Hog+Haar+WST	**51.1**	**84.0**

表 3.4 使用 Faster R-CNN 和 ResNet50 作为检测模型时，不同滤波增强输入的比较。

Domain	PCC ↑
Pixel Space	0.394
Canny Space	0.992
Hog Space	0.995
Haar Space	0.990
WST Space	**0.996**
GRE Space	0.984

表 3.5 ImageNet 和 SARDet-100K 在 RGB 和手工特征空间上的皮尔逊相关系数 (PCC)。

在本章提出的 MSFA 方法框架内，为了研究和评估滤波增强输入的影响，对前述各类传统特征描述符进行了实验。表 3.4 中的结果表明，结合这些手工特征能够提高检测器性能。此外，本章的分析显示，将图像像素转换到手工特征空间后，可以缩小 ImageNet 和 SARDet-100K 之间的分布差距，这一点在两者输入之间的皮尔逊相关系数 (PCC) 上表现得较为明显，如表 3.5 所示。这进一步说明了所提出方法在弥合自然图像与 SAR 图像领域差距、提升预训练知识迁移效率方面的有效性。

值得注意的是，小波散射变换 (WST) 特征在比较中表现较好。这种优势不仅可以归因于其在缩小领域差距方面的作用，还可以归因于其提取丰富的多尺度信息的能力。这些信息通过减轻噪声和保留与对象相关的细节，充当了有效的辅助特征。然而，使用多种滤波增强特征不会带来进一步稳定的性能提升。可能是现有的 WST 已经捕获了有效对象检测所需的必要信息，而结合额外的特征并不会提供大量额外的有益信息。

由于 WST 的出色性能，将其用作 MSFA 方法中的默认滤波增强输入。

（二）多阶段预训练

为了评估所提出的多阶段预训练方法的有效性，本章保持输入模态一致，并使用各种预训练策略在 SARDet-100K 数据集上微调检测模型。作为基线，表 3.6 中实验 1 采用单通道 SAR 数据作为输入，在 ImageNet 上预训练骨干网络模型 100 个 epoch，然后直接在 SARDet-100K 数据集上微调检测器（遵循广泛使用的默认设置）。除了基线之外，本章还进一步引入光学遥感检测数据集（如 DOTA[134] 和 DIOR[247]）进行第二阶段检测预训练，以构建从自然图像到遥感图像、再到 SAR 图像的连续迁移路径。在第二阶段预训练之后，本章分别考察“仅迁移骨干”和“迁移完整检测框架”两种设置。

表 3.6 中实验 2、4、6 和 8 的结果体现了两阶段预训练方法的优势。值得注意的是，即使是规模相对较小的 DIOR 数据集也显示出比基线（实验 1 和 5）更高的性能。这一观察结果揭示了在 SAR 检测的预训练阶段减少领域差距的重要性。

然而，DIOR 数据集预训练不如更大规模的 DOTA 数据集有效（实验 2 与 4，实验 6 与 8）。这种比较强调了预训练规模对结果的影响。DOTA 数据集具有更大的规模，并且平均实例面积与 SARDet-100K 相似，因此提供了更全面和信息更丰富的预训练，从而提高了后续微调阶段的性能。

实验 3 和 4 与实验 7 和 8 之间的比较表明，在本章设置下预训练整个框架相比仅预训练骨干网络能够取得更好结果，突出了模型差距对 SAR 检测性能的影响。

（三）MSFA 的泛化性

为了评估所提出的 MSFA 的有效性和泛化性，本章使用各种检测器和骨干网络进行了实验，如图 3.8(a) 和 3.8(b) 所示。在不同的框架（包括单阶段[18-19,222]、两阶段[17,154,248] 和端到端[249-251]）以及不同的骨干网络（包括 ResNets[22]、ConvNexts[225]、VANs[226] 和 Swin-Transformer[72]）中，都观察到了稳定的性能提升。这为本章提出的方法的有效性和适用性提供了实验证据。此外，如图 3.8(b) 所示，随着骨干网络规模的扩大，性能也稳定提升，这表明本章提出的方法具有良好的可扩展性。

值得注意的是，MSFA 方法的设计在开发时就考虑了灵活性、泛化性和广泛的适用性。因此，该方法可以无缝集成到大多数现有模型中，而无需进行任何

表 3.6 使用 Faster-RCNN 和 ResNet50 作为检测模型时，不同预训练策略的比较。

序号	模型输入	预训练			mAP ↑
		多阶段	数据集	网络成分	
1	SAR（原始像素）	✗	ImageNet	骨干网络	49.0
2		✓	ImageNet + DIOR	整体框架	49.5
3		✓	ImageNet + DOTA	骨干网络	49.3
4				整体框架	**50.2**
5	SAR+WST（滤波增强）	✗	ImageNet	骨干网络	49.2
6		✓	ImageNet + DIOR	整体框架	50.1
7		✓	ImageNet + DOTA	骨干网络	49.6
8				整体框架	**51.1**

修改。

（四）与近期方法的比较

本章比较了各种近期方法，包括通用目标检测模型[17,154,248,252-255] 以及 SAR 目标检测模型[233,235,256-263]。本章评估了它们在 SSDD 和 HRSID 数据集上的性能，这些数据集是常用的 SAR 目标检测基准。为了利用 VAN[226] 骨干网络较好的效率和性能（如图 3.8(b) 所示），本章采用经典的 Faster R-CNN 检测框架，并使用轻量级的 VAN-B (参数量 26.6M) 骨干网络作为检测模型。表 3.7 中展示的结果表明，MSFA 方法在所比较的方法中取得了更高结果。具体而言，MSFA 在 SSDD 数据集上实现了 97.9% 的 mAP@50，在 HRSID 数据集上实现了 83.7% 的 mAP@50。值得注意的是，在比较的 SAR 检测方法中，本章的方法是唯一开源的方法。

三、分析与对比实验

（一）相似计算预算下的预训练收益

除了常规精度比较之外，本章进一步关心一个更具实践意义的问题：MSFA 的收益是否仅仅来自更长训练时间，还是确实来自更合理的迁移路径。为此，本章在相似计算预算下比较了“直接微调”和“先进行第二阶段检测预训练再微调”两种策略。

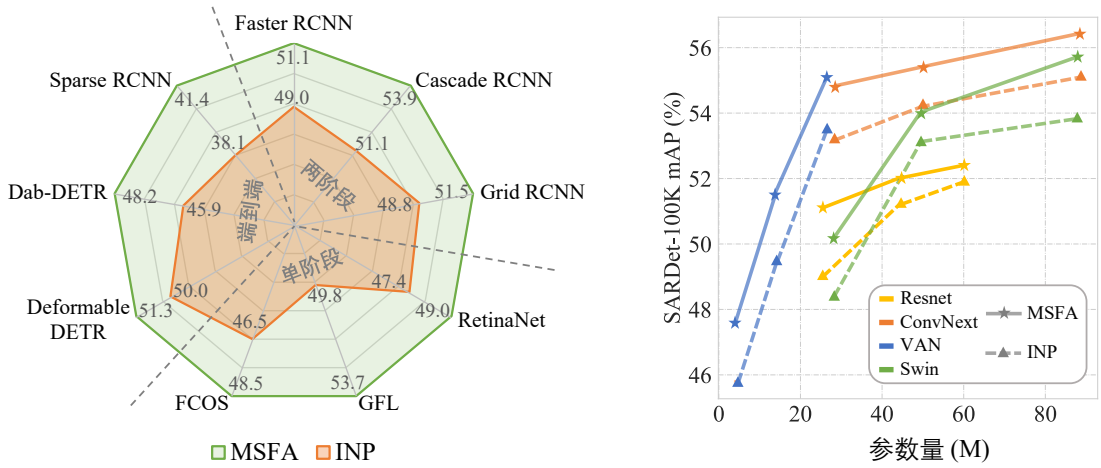


图 3.8 MSFA 在不同检测框架 **(a)** 和不同骨干网络 **(b)** 上的泛化性。模型在 SARDet-100K 数据集上进行了微调和测试。INP：仅在骨干网络上进行传统的 ImageNet 预训练。

表 3.7 提出的 MSFA 与先前方法在 SSDD 和 HRSID 数据集上的比较。

	检测器	开源	年份	mAP_{50} ↑	
				SSDD	HRSID
常规检测器	Grid R-CNN[248]	✓	2019	88.9	79.4
	Faster R-CNN[17]	✓	2015	89.7	80.7
	Cascade R-CNN[154]	✓	2019	90.5	81.3
	Free-Anchor[254]	✓	2019	91.0	81.8
	Double-Head R-CNN[255]	✓	2020	91.1	<u>82.1</u>
	PANET[261]	✓	2018	91.2	81.6
	DCN[252]	✓	2017	92.3	<u>82.1</u>
SAR 检测器	NNAM[256]	✗	2019	79.8	-
	DCMSNM[257]	✗	2018	89.6	-
	ARPN[258]	✗	2020	89.9	81.8
	DAPN[259]	✗	2019	90.6	81.8
	HR-SDNet[260]	✗	2020	90.8	82.5
	SER Faster R-CNN[262]	✗	2018	91.5	81.5
	FBR-Net[263]	✗	2020	94.1	-
	NRENet[264]	✗	2024	94.6	75.6
	CenterNet++[233]	✗	2021	95.1	-
	CRTransSar[235]	✗	2022	97.0	-
	SARATR-X[265]	✗	2024	<u>97.3</u>	80.3
Faster R-CNN + VAN-B		✓	2023	92.9	81.8
MSFA (Faster R-CNN + VAN-B)		✓	2024	**97.9**(+5.0)	**83.7**(+1.9)

表 3.8 DOTA 和 DIOR 的数据集统计比较。* 表示经过多尺度预处理。

数据集	图片	实例	类别数	图片大小	平均实例区域
DOTA*	68,324	1,058,641	15	1024×1024	5,021
DIOR	23,463	192,518	20	800×800	12,726

表 3.9 在相似计算预算下，MSFA 与直接微调性能的比较。

模型	ImageNet 预训练	MSFA 轮数	微调轮数	总轮数	总步数	mAP
Faster-RCNN[17]	✓	12	24	36	16.1k	54.5
Faster-RCNN[17]	✓	0	36	36	17.7k	52.8

对于多阶段预训练的第二阶段，本章主要使用大规模光学遥感检测数据集 DOTA，并额外采用 DIOR 进行对照。表 3.8 表明，DOTA 在图像规模、实例数量以及目标尺度分布上都更接近大规模检测预训练的需求，因此更适合作为从自然图像迁移到 SAR 检测的“中间域桥梁”。

如表 3.9 所示，在总训练轮数相同、总迭代步数甚至更少的条件下，MSFA 仍然比直接微调高出 1.7 个百分点。这一结果说明，性能提升并非简单来自训练更久，而是来自“ImageNet 分类预训练 → 光学遥感检测预训练 → SAR 检测微调”这一连续迁移链路本身。换言之，MSFA 通过中间域桥接和检测器级迁移，让预训练知识以更适合下游任务的形式被吸收。

（二）数据集难度与泛化分析

为了进一步验证 SARDet-100K 作为大规模研究基准的价值，本章比较了不同骨干网络在 SARDet-100K、SSDD 和 HRSID 上的表现，如图 3.9 所示。

可以看到，现代检测模型在 SSDD 和 HRSID 上的性能已经相对饱和，而在 SARDet-100K 上仍存在明显的模型间性能差距。这说明前两个数据集更适合做方法验证，而 SARDet-100K 更能真实反映大模型容量、迁移策略和检测框架设计的差异。另一方面，在较小数据集上，模型规模增大有时会因过拟合而带来退化。而在 SARDet-100K 上，模型通常仍能从更大容量中持续获益。这进一步说明本章构建的数据集更适合作为大规模 SAR 检测研究和基础模型迁移研究的公共平台。

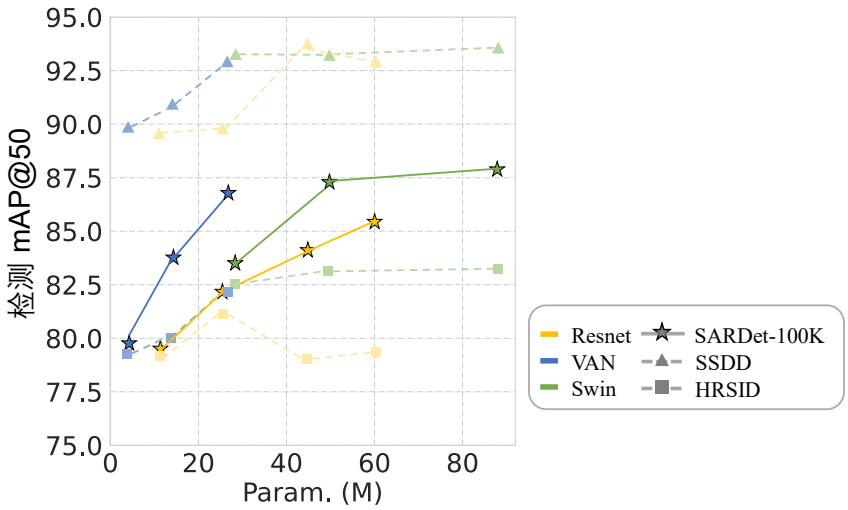


图 3.9 在 SARDet-100K、SSDD 和 HRSID 上评估不同骨干网络的检测性能，检测器统一为 Faster-RCNN[17]。

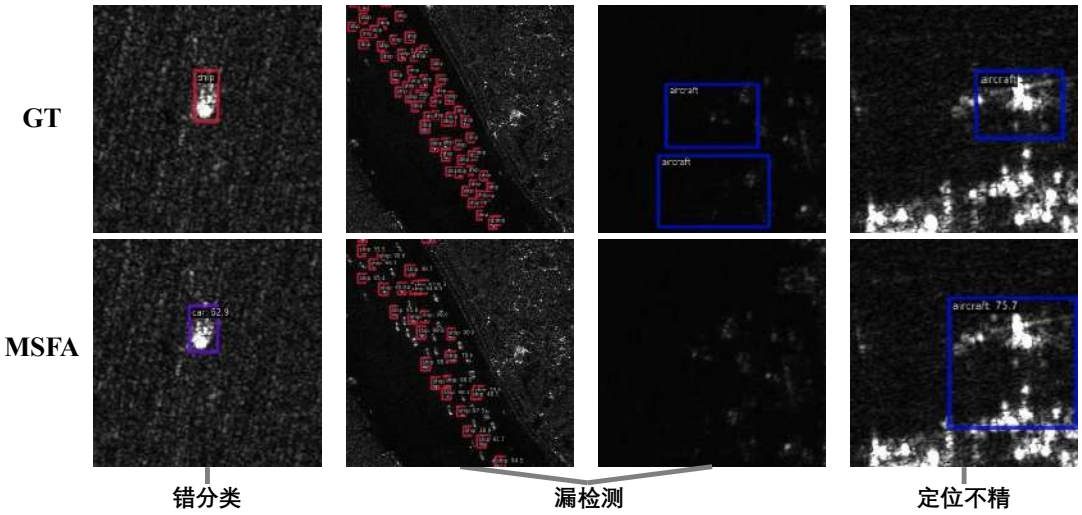


图 3.10 失败案例可视化，包括错误分类、漏检和定位不准确等典型情形。

（三）失败案例可视化分析

图 3.10 表明当前方法仍然面临若干困难情形。当 SAR 图像缺少足够的纹理细节或上下文信息时，模型可能出现错误分类。当目标尺寸很小且分布密集时，部分实例仍可能被漏检。而在模糊、低分辨率或质量较差的图像中，定位精度也更容易下降。这些现象说明，本章工作主要解决的是预训练与微调之间的领域差距和结构差距问题，而针对极端小目标、复杂形变目标和超低质量图像的专门建模仍值得在更强检测器中继续加强。

第六节 本章小结

本章围绕“大规模 SAR 检测基准构建”与“跨域连续预训练”两个关键问题展开研究。首先，本章构建了接近 COCO 级别的大规模多类别 SAR 检测基准 SARDet-100K，通过统一划分协议、图像切片策略和标注格式，提升了该方向的可比性与研究上限。随后，本章提出多阶段滤波增强预训练框架 MSFA，从滤波增强输入、光学遥感桥接预训练以及检测器级迁移三个层面共同缩小自然图像与 SAR 图像之间的差距。实验结果表明，MSFA 能够提升检测精度，并在不同框架、不同骨干以及相似计算预算条件下表现出稳定收益。这一章为全文后续的统一多模态检测与跨模态预训练研究提供了关键的数据基础与迁移基础。

本章工作仍存在一定局限。SARDet-100K 虽然扩大了 SAR 检测数据规模和类别覆盖，但仍难以穷尽不同传感器、极化方式、分辨率和地理区域带来的全部分布差异。未来可进一步结合半监督预训练、跨传感器域泛化和自动化数据配方搜索，增强 SAR 基础检测模型在开放场景中的适应性。

第四章　基于混合专家系统的异构多模态检测统一架构

随着多传感器协同观测逐渐成为遥感系统的重要形态，面向单一模态、单一数据集独立训练检测器的传统范式已难以满足复杂场景中的统一部署需求。现有方法通常忽视跨模态共享知识，导致模型在表达能力、训练成本和部署效率之间难以兼顾。为此，本章提出异构多模态遥感目标检测任务（Multi-Modal Datasets and Multi-Task Remote Sensing Object Detection, M2Det），旨在利用单一模型处理任意传感器模态下的水平框或旋转框检测任务。围绕这一任务，本章构建了统一基准，并提出统一检测框架 **SM3Det**（**S**ingle **M**odel for **M**ulti-Modal datasets and **M**ulti-Task object **Det**ection）。SM3Det 采用网格级稀疏混合专家（MoE）骨干网络，在保留模态特性的同时学习跨模态共享表示，同时引入一致性与同步优化机制，以缓解不同模态和不同任务之间的学习难度差异。实验结果表明，在本章构建的统一评测设置下，SM3Det 相比分别训练的多模型方案取得了更好的整体性能与泛化表现。

第一节　引言

第四章对应绪论中的架构层挑战，即如何在异构多模态条件下实现统一表示学习与“分而治之”的协同优化。前两章分别从单模态骨干能力和 SAR 数据/预训练基础上补齐了遥感检测的两个底座，但真实应用场景往往同时包含 RGB、SAR、红外等不同传感器，以及水平框、旋转框等不同检测任务。如果仍沿用“一个模态一个模型、一个任务一套系统”的烟囱式方案，不仅难以共享跨模态知识，也会带来极高的训练、部署和维护成本。

遥感目标检测[8-9,266-269]通常涉及采用不同成像机制的多类传感器，从而产生多样化的数据模态。传统上，检测模型是针对与单一模态和预定义格式检测任务相关联的特定数据集单独开发的[270-272]，如图 4.1(b) 所示。这种方式忽略了统一遥感场景中潜在的共享语义，也难以适应未来多传感器协同观测平台的需求。另一方面，既有多源目标检测方法[273-275]大多依赖稀缺且难以长期满足的空间配准图像对及其配准算法[276-277]，并且通常只服务于单一标注格式的检测任务，如图 4.1(a) 所示。

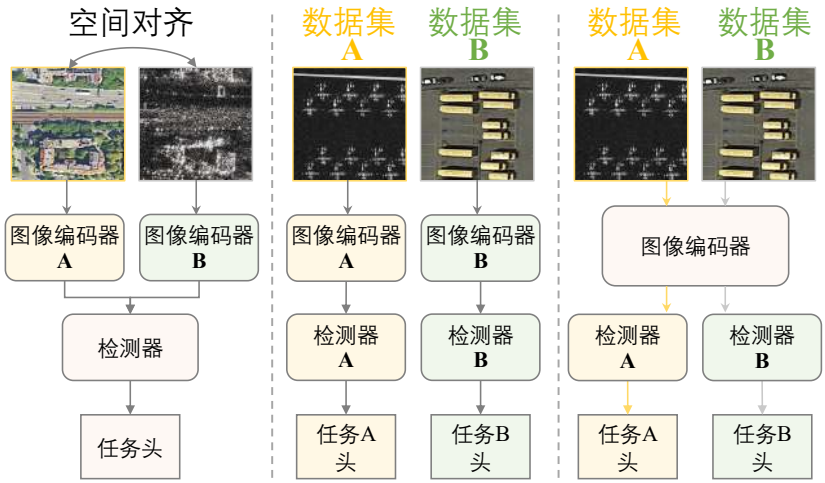


图 4.1 任务对比：(a) 空间配准多模态，(b) 传统单数据集，以及 (c) M2Det。M2Det 旨在利用统一模型检测任意模态的图像，并处理多种检测任务。

因此，本章将问题进一步推进为：能否构建一个无需空间配准样本、能够统一处理多种模态与多种检测格式的单模型框架？为此，本章定义了异构多模态遥感目标检测（Multi-Modal Datasets and Multi-Task Remote Sensing Object Detection, M2Det）任务。M2Det 旨在对任意给定图像完成目标检测，无论其输入模态为何，也无论其下游任务采用水平边界框还是定向边界框，如图 4.1(c) 所示。该任务不仅具有学术意义，也具有明确的工程价值，有望成为低空经济、空天平台协同感知等场景中的关键基础能力。

M2Det 任务虽然与多数据集目标检测[52,54]和多任务学习[55,278]密切相关，但又明显难于这两类经典设定。在传统多数据集检测中，不同数据集大多共享相近的视觉概念，联合训练往往只是在语义空间内整合不同风格或领域的数据。而在遥感场景中，RGB[279-280]、SAR[214,281]、红外[282]乃至多光谱图像[163]对应着完全不同的成像机制与噪声模型，如图4.6所示。与此同时，遥感数据集还常同时包含水平框与旋转框等异构标注形式[247,279,281-282]，这进一步放大了模型学习与优化的复杂度。

这些差异主要体现在以下两个层面，并直接阻碍传统统一模型的有效学习：

1）表示能力受限：在多个任务和模态间共享相同参数的密集模型，其表示能力可能面临限制，因为单一参数集难以有效拟合每个数据集固有的不同分布。

2）优化不一致：不同模态和任务的学习难度差异可能导致模型各部分优化

速率或优化方向的不同步。这种不一致性可能导致相互冲突的优化结果，对模型实现不同损失目标的能力产生不利影响。

为应对这些挑战，本章首先整合 SARDet-100K（第三章中提出的数据集）、DOTA[279]和 DroneVehicle[282]，构建覆盖 SAR、光学和红外模态的统一基准。随后提出单模型多模态多任务检测框架 SM3Det，从**模型架构**与**模型优化**两个层面协同解决表示拥塞与优化冲突问题：

模型架构：本章提出将一种即插即用的网格级稀疏混合专家（MoE）架构集成到骨干网络中，使模型能够同时捕捉共享知识和模态特定的表示。与以往使用硬编码、图像级路由的多数据集目标检测模型[52,283]不同，本章的方法引入了具有动态路由的网格级专家。这些专家对空间网格特征进行操作，允许模型在网格级别自适应地处理信息，这对于目标检测任务至关重要。

模型优化：本章引入动态子模块优化（DSO）机制，以实现模型优化的一致性和同步性。它能根据定制策略自适应地调整网络不同组件的学习率。DSO 通过平衡相对收敛速度并保证优化方向的一致性，适应了不同任务和模态间变化的学习复杂性。与传统技术（主要修改损失权重或梯度，通常缺乏对特定网络子模块的精确操控或效率低下）不同，本章的 DSO 在保持优化效率的同时提供了细粒度的控制。

实验结果表明，统一的 SM3Det 在本章所比较的模态数据集上相比分别训练的多模型方案取得了更优的整体结果。更重要的是，本章说明“统一建模”并不必然以牺牲模态特异性为代价，通过合理的架构与优化策略，单模型同样可以兼顾共享知识、模态差异和部署效率。这也为后续章节进一步讨论多模态预训练与跨模态对齐奠定了统一模型基础。贡献总结如下：

- 本章引入了一项新任务：遥感领域中使用统一检测模型的多模态数据集与多任务目标检测。
- 本章提出了 SM3Det 模型，通过从模型架构和模型优化角度提供创新解决方案，应对 M2Det 任务的挑战。
- 在所建立的基准数据集上进行的实验和分析表明，本章提出的单一模型能够在统一设置下取得较好的整体性能，并在多个模态上相比分别训练的模型表现出稳定收益。

第二节 相关研究与问题分析

一、多数据集检测与统一标签学习

多数据集目标检测旨在利用多个数据源共同学习更具泛化性的检测知识。已有研究表明，多数据集训练能够提升视觉模型的鲁棒性与数据效率[284-287]。在目标检测领域，DA 网络通过领域特定注意力缓解不同数据集间的分布差异[52,288]，Universal-RCNN 利用跨数据集图结构建模共享知识[53]，Unidet 则进一步尝试构建统一标签空间并讨论采样策略的重要性[54]。这些工作说明，联合训练并非简单拼接数据，而需要在共享概念与数据集特性之间建立平衡。

然而，现有多数据集检测大多建立在视觉概念相近、成像机制一致的数据前提下。对于遥感场景，不同数据集不仅意味着不同类别分布，更意味着不同传感器模态、不同噪声模型和不同标注形式。因此，传统多数据集检测中的“统一标签 + 共享骨干”思路，不足以直接解决异构遥感场景中的统一建模问题。

二、异构遥感多模态检测

多源遥感目标检测已经证明，不同模态之间存在可利用的互补信息[273-275]。但这类方法通常依赖严格空间配准的图像对以及相应的配准算法[276-277]，更适合成对输入下的融合检测，而不适合真实部署场景中“任意时刻仅有单一模态可用”的情况。与此同时，遥感数据集还常同时包含水平框和旋转框两类检测格式[247,279,281-282]。这意味着问题不再只是跨模态融合，而是要在没有配准样本的条件下，统一处理多模态输入和多任务输出。

从这个角度看，M2Det 并不是已有多源检测问题的简单延伸，而是把“多数据集”“多模态”和“多任务”三种异质性同时叠加到单一模型之中。由此带来的挑战，不仅体现在表征分布更复杂，也体现在不同模态和任务对模型容量与优化路径的需求明显不同。

三、多任务优化与训练冲突

多任务学习涉及使用单一模型学习多个目标，通常具有多个任务头和损失函数。在多任务学习中，已经开发了各种策略[55-56,289-290]来解决任务不平衡问题并优化学习成果。GradNorm[55]侧重于通过调整每个任务损失函数的梯度大小，来纠正反向传播过程中的梯度不平衡。像多梯度下降算法[289]之类的方法对梯度反向传播采用帕累托优化，但由于需要额外的梯度计算，它们可能效率低下。与 GradNorm 类似，DWA[52]也使用任务损失来评估收敛速度，但它动态调整每个

任务损失的权重。不确定性损失[56]则采用了一种不同的方法，将同方差不确定性纳入加权损失函数中。这些方法为多任务协同训练提供了重要启发，但它们大多把冲突建模为“不同损失之间的关系”，较少进一步考虑网络内部不同子模块在不同模态和任务条件下的收敛节奏差异。

对于 M2Det 而言，这一问题会被进一步放大。模态差异不仅导致不同任务头的学习难度不一致，也会使共享骨干、路由模块和检测头之间的优化步调产生偏移。因此，仅靠损失重加权往往不足以稳定训练过程，还需要一种面向子模块层级的协同优化机制。

四、混合专家模型

混合专家模型（MoE[291-292]）利用多个专家网络来提供丰富的特征。稀疏 MoE[293]进一步引入了稀疏性，使得能够在不明显增加计算复杂度的情况下扩展模型规模。在多任务学习中，稀疏 MoE 使得不同的专家网络能够学习到不同的判别性特征。大多数基于稀疏 MoE 的多任务方法[294-295]都基于 Transformer 架构，将专家集成到视觉 Transformer 的骨干块中，以便在推理过程中选择性地激活不同的路径。在多数据集学习中，近期的工作[283]在视觉 Transformer 中采用 MoE，将图像级特征路由到特定的专家。

但对于异构遥感多模态检测，现有 MoE 应用仍存在两个不足：其一，路由粒度往往停留在图像级或块级，难以充分匹配检测任务对局部空间结构的需求。其二，很多方法仍依赖硬编码的数据集身份或简单的全局路由，不利于在同一幅图像内部同时建模共享知识与模态特性。相比之下，本章更关注网格级稀疏路由，让专家在空间特征层面完成共享与分化。

五、问题分析与本章定位

综上，现有工作分别从多数据集检测、多源融合、多任务学习和 MoE 建模等角度提供了有价值的基础，但尚未形成一套适用于空间异构、多模态、多任务遥感检测的统一方案。M2Det 的关键难点并不只是“能否用一个模型覆盖多个数据集”，而是“如何在统一模型中同时保留共享知识、模态特性与优化稳定性”。基于这一判断，本章将研究重点落在“统一架构 + 协同优化”上：前者通过网格级稀疏 MoE 在共享与专有之间建立结构性平衡，后者通过动态子模块优化缓解不同模态、不同任务之间的收敛失衡。SM3Det 也由此成为全文从单模态感知走向统一多模态检测能力的关键章节。

第三节 方法

一、任务定义与方法概述

所提出的 M2Det 任务旨在利用一个统一的模型来检测来自任意模态的图像中的目标，处理各种预定义的检测任务，例如水平边界框和旋转边界框。该任务的重要性体现在多种实际应用中，包括低空经济[296-297]、空中监视[298-299]、对地观测[300-301]以及其他研究领域[302-303]。例如，配备了 M2Det 模型的平台可以充分利用可用的多模态数据，同时受益于简化的版本控制以及多传感器的无缝集成，而无需在设备上更新模型。这降低了工业应用中的模型维护成本。此外，在一个小批量内于单个模型中处理不同模态的图像，可以最大限度地发挥 GPU 的并行计算能力，从而提高边缘设备的计算和能源效率。

整体网络架构遵循多任务学习模型的经典设计[52,54]。它由一个相对较重的、共享的特征空间组件（骨干网络）和多个相对轻量的、独立的特征空间组件（任务头）组成。骨干网络负责联合表示学习，其大部分参数是共享的，从而保证了参数效率。轻量级的任务头则被分离出来，以适应不同的特征和任务学习。然而，如第一节所述，模态和任务之间的差异可能会降低此类经典多任务模型的性能。为了解决这个问题，本章提出了 SM3Det 模型，它由两部分组成：

模型架构：一个稀疏的 MoE 骨干网络，其中的专家在网格级别上对多模态数据集中图像的局部图像特征进行激活。

模型优化：一种高效的动态子模块优化机制，用于处理跨多个任务和模态的不同学习难度和优化不一致性问题。

二、网格级 MoE

以往的多数据集目标检测方法[53-54]利用密集模型来挖掘数据集之间的共享概念，以增强联合知识表示。在多模态遥感图像的情况下，这种联合知识同样存在[281]，尽管可能不那么显式，例如跨模态的常见弱线索，如形状和尺度。然而，由于固有的模态和任务差异，使用跨多个任务和模态采用相同参数的密集模型可能会导致特征/表示空间的拥塞，最终降低模型的表达能力。因此，有必要探索那些既能利用跨模态的联合知识，又能允许每个模态进行独立的表示学习以防止特征空间干扰的方法。

受稀疏 MoE 网络[293]（以其稀疏性和高容量为特点）成功经验的启发，本章提出利用 MoE 来解决 M2Det 任务。对于基于 Transformer 的骨干网络[72,304]，

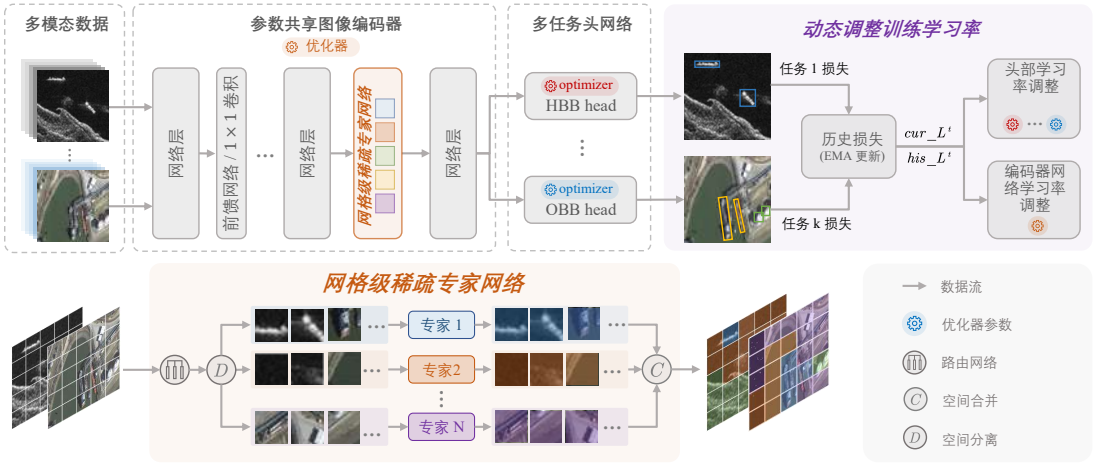


图 4.2 SM3Det 模型示意图。“HBB head”：水平边界框检测头，“OBB head”：旋转边界框检测头。

本章将 MoE 专家集成到前馈网络组件中。对于现代 CNN[83,226,245]，它们通常使用 1×1 卷积[305]进行特征交互或维度缩减/扩展，本章引入稀疏专家来增强这些层。与以往基于 Transformer 的检测器将整个图像的特征通过单个专家路由的方式不同[283]，本章的设计允许专家在骨干网络内的局部网格特征上操作。这种方法确保了专家能够处理跨模态的相似空间模式，从而促进共享表示学习。同时，多个专家捕获跨模态的不同模式，实现独立的表示学习。具体来说，对于深层图像特征中位于第 i 行和第 j 列的局部空间输入特征 x_{ij}，经过 MoE 层后的输出特征 $f_{MoE}(x_{ij})$ 为：

$$f_{MoE}(x_{ij}) = \sum_{n=1}^{N} G_n(x_{ij}) \cdot Conv_n^{1\times1}(x_{ij}), \tag{4.1}$$

$$G(x_{ij}) = \mathrm{TOP}_k\left(\mathrm{Softmax}\left(\frac{E^T W x_{ij}}{\tau \|W x_{ij}\| \|E\|}\right)\right), \tag{4.2}$$

其中，N 是专家总数，G 是门控函数，$Conv_n^{1\times1}$ 是第 n 个 1×1 卷积专家。每个专家在矩阵 E 中都有一个表示嵌入。输入特征 x 首先通过矩阵 W 进行变换。然后，Wx 的乘积与矩阵 E 中的每个专家嵌入进行比较以计算相似度。该比较通过除以 Wx 和 E 的范数乘积进行归一化，保证了尺度不变性。相似度分数通过 $Softmax$ 函数，转换为概率分布。这意味着门控函数为每个专家分配一个概率，表明其与输入特征 x 的相关性。最后，TOP_k 运算符选择概率最高的前 k 个专家。它通过将 $Softmax$ 概率分配给前 k 个专家来重新加权每个专家，并将其余专家

设置为零。这一步使模型稀疏化，只关注一小部分专家，从而降低了计算复杂度，并增强了模型处理不同任务和模态的表达能力。

总之，$f_{MoE}(x_{ij})$ 是前 k 个专家输出的加权和。权重由门控函数 G 决定，该函数为每个局部特征动态选择最相关的专家。MoE 在骨干模型中创建了一个更稀疏的特征空间。通过关注局部模式，模型可以独立学习以建模多种模态和局部目标模式。本章的设计有效地解决了特征空间拥挤的挑战，并增强了模型的表达能力。

在实际实现中，为了充分利用预训练的骨干网络权重，通过复制相应的预训练 1×1 卷积层的权重来初始化新增专家的权重，然后再进行下游模型的微调，确保在微调开始时所有专家都能被均匀地选择。对于任务头，本章保持简洁，并遵循[52,54]中任务头的现有设计。

三、动态子模块优化 (DSO)

在多模态、多数据集和多任务目标检测任务中，一个主要挑战是跨模态和任务的不同学习难度[55-56]。这种变化可能导致不同步的优化速率和不一致的优化方向[306]，从而在不同损失函数之间引发目标冲突。为了解决这个问题，本章提出动态子模块优化机制来管理跨任务和模态的不同学习难度。

DSO 将每个任务头的损失作为指标，用于判断当前每个任务的收敛速率和网络的整体优化方向，并相应地调整学习率。具体来说，一种策略是针对每个任务头子模块（非共享权重）的学习率进行调整，以平衡每个任务的相对收敛速率。另一种策略是针对骨干网络子模块（共享权重）的学习率进行调整，以确保优化方向的一致性。

本章将任务 t 在第 i 次迭代的训练损失记为 $cur\_L_i^t$。每个任务的损失都维护一个指数移动平均值作为平滑后的历史统计量，记为 $his\_L_i^t$，即：

$$his\_L_i^t = \alpha \cdot cur\_L_i^t + (1-\alpha) \cdot his\_L_{i-1}^t. \tag{4.3}$$

对于头子模块的学习率调整，本章使用 his_L 与 cur_L 的比值作为任务 t 在第 i 次迭代收敛速率的倒数：

$$w_i^t = \frac{his\_L_i^t}{cur\_L_i^t}. \tag{4.4}$$

使用带温度参数 θ 的 $Softmax$ 函数来重新加权相应网络任务头的学习率，

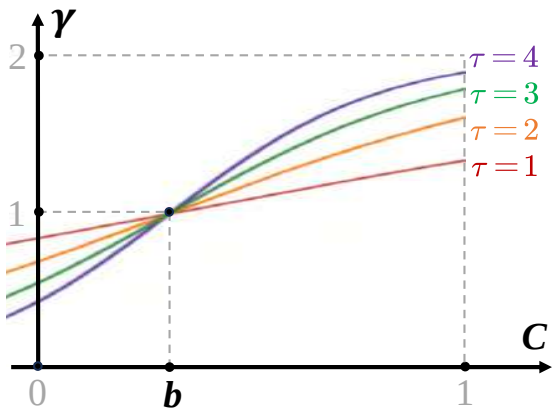


图 4.3 不同温度 (τ) 下的重加权曲线。

旨在平衡每个任务的收敛速度。任务 t 在训练迭代 i 时的重加权因子 λ_i^t 表示为：

$$\lambda_i^t = \frac{T \cdot e^{w_i^t/\theta}}{\sum_k^T e^{w_i^k/\theta}}, \tag{4.5}$$

其中 T 是任务总数。因此，相对较大的 $cur\_L_i^t$ 值表明任务 t 收敛较快，导致 w_i^t 较小，从而获得较低的重加权因子 λ_i^t，以防止其过快收敛。相反，较小的 $cur\_L_i^t$ 值会导致较大的 λ_i^t。这种策略确保了在整个训练过程中，每个任务的收敛速率保持平衡。

对于骨干子模块的学习率调整，重加权基于每个损失的历史一致性。为了衡量训练收敛一致性，本章基于 cur_L 和 his_L 定义了一个一致性分数 C。具体来说，cur_L 和 his_L 首先通过函数 P 转换为概率分布，该函数使用简单的 $Softmax$ 函数：

$$P(L) = Softmax(L). \tag{4.6}$$

接下来，计算 Kullback-Leibler 散度 D_{KL}，以评估当前每个任务的损失是否与其历史值保持稳定和一致：

$$C = 1 - D_{KL}\left(\, P(cur\_L) \,\|\, P(his\_L) \,\right) \tag{4.7}$$

$$= 1 - \sum_t^T P(cur\_L^t) \cdot \log \frac{P(cur\_L^t)}{P(his\_L^t)}, \tag{4.8}$$

因此 C 的取值范围为 $(-\infty, 1]$。C 越大表示当前迭代损失的相对值与历史值越相似，表明当前批次的样本使网络更新趋于稳定。在这种情况下，需要提高学习

率以使网络更快收敛。相反，C 越小表示不稳定，表明与之前的平均状态相比，当前样本使得某些任务变得更难学习，而其他任务变得更容易。如果网络在这种情况下过于激进地更新共享权重，网络将朝着当前迭代中较难任务的方向进行优化，这可能损害较容易的任务。因此，网络应谨慎更新以降低学习率。

为了平衡这一点，本章提出使用以下策略动态地重新加权共享权重的骨干网络：

$$\gamma_i = 2 \cdot Sigmoid((C - b) \cdot \tau) \tag{4.9}$$

$$= \frac{2}{1 + e^{-(C-b)\cdot\tau}}. \tag{4.10}$$

标量因子 2 确保 sigmoid 函数后的重加权值在 (0, 2) 范围内。b 是超参数偏置，可解释为重加权阈值，即当 C 等于 b 时，重加权值为 1。τ 是用于值敏感性调整的温度。不同温度下的重加权曲线以及 b 与 C 之间的关系如图 4.3所示。

第四节 实验与分析

为了训练和评估用于多模态多任务检测（M2Det）任务的模型，本章通过合并三个检测数据集构建了一个新的统一基准：SARDet-100K（合成孔径雷达）、DOTA-v1.0[279]（光学）和 DroneVehicle[282]（红外）。本章将该组合数据集称为 SOI-Det。在主要结果和消融研究中，除非另有说明，否则默认使用 ConvNext-T 作为骨干网络。

一、SOI-Det 基准与评测协议

SOI-Det 同时覆盖了模态差异和标注差异。具体而言，SARDet-100K 包含 6 个类别，采用水平边界框（HBB）标注。DOTA 包含 15 个类别，采用定向边界框（OBB）标注。DroneVehicle 包含 5 个红外车辆类别，也采用 OBB 标注。三者在成像机理、目标分布和任务形式上都存在明显差异，因此 SOI-Det 不仅是一个多数据集联合训练基准，更是一个真正的异构多模态、多任务统一检测基准。

考虑到直接合并完整 SARDet-100K 会造成明显的数据规模不平衡，本章在联合训练中仅使用其子集 HRSID、MSAR、SADD、OGSOD 和 SIVED。该子集包含 47,097 张训练图像和 4,481 张测试图像。训练期间，每个批次从 SARDet-100K 子集、DOTA 和 DroneVehicle 中按 2:1:1 的比例均匀采样，以减弱数据规模差异对共享骨干学习的干扰。

在评测指标方面，本章对各数据集均报告 IoU 阈值 0.5 下的 AP（@50）、IoU 阈值 0.75 下的 AP（@75），以及 IoU 从 0.5 到 0.95 的平均精度均值（mAP，%）。同时，本章报告三个数据集的总体 mAP，以评估统一模型的整体性能。

二、实现细节

所有模型均在各自对应的训练集上进行微调。对于 SARDet-100K 和 DroneVehicle，模型在测试集上评估，对于 DOTA，模型在验证集上评估。单数据集训练时，模型使用 AdamW 优化器训练 12 个轮次。多模态联合训练时，为保证公平比较，本章使总迭代次数与各单数据集训练迭代次数之和保持同量级。

在统一架构中，本章共享骨干网络，并为不同的数据集和任务使用独立任务头。具体来说，来自 SARDet-100K 的特征被送入 GFL 头，以适应水平框检测，来自 DOTA 和 DroneVehicle 的特征则分别送入两个独立的 O-RCNN[7] 头，以适应定向框检测。在主要实验中，初始学习率设为 1e-4，权重衰减设为 0.05，训练使用 8 张 RTX 3090 GPU，每张 GPU 的批大小为 4。本章报告的 FLOPs 统一基于 800×800 输入计算。

三、主要结果

本章将提出的 SM3Det 模型的性能与单个数据集训练、简单联合训练以及三种可适用于此任务的近期方法进行了比较：采用分区头的 UniDet[54]、在 ConvNext-T 骨干网络中实现的 DA 网络[52]，以及在 UniDet 上实现的不确定性损失[56]。主要结果见表 4.1。

可以观察到，对三个多模态数据集进行简单的联合训练，即仅仅合并数据集，使用一个共享骨干网络和独立任务头的模型，并采用随机数据采样策略，会导致性能下降。这一现象突显了此任务相较于通用目标检测中的多数据集训练的挑战性，在通用目标检测中，简单联合训练通常会提升单个数据集的性能[52,54]。以往方法，如 UniDet[54]、DA[52] 和不确定性损失[56]，仅略超基线。相比之下，本章提出的 SM3Det 模型将整体 mAP 性能从 48.23 提升至 50.20，提高了 1.97 个 mAP。值得注意的是，轻量版 SM3Det（仅包含 DSO 而无 MoE 结构）也在该设置下取得了较好的比较结果。

为了评估 SM3Det 的泛化能力，本章在不同的骨干网络和检测器上评估其性能。如图 4.4 所示，SM3Det 在多种现代卷积骨干网络（包括 ConvNext[83]、VAN[226]、第二章中提出的 LSKNet 和 PVT-v2[304]）上相比分别训练模型取得更

表 4.1 在 SOI-Det 数据集（SARDet-100K + DOTA + DroneVehicle）上的模型性能比较。提出的 SM3Det 模型优于单个模型和其他比较模型。

模型	FLOPs	参数量	测试集	**mAP**	@50	@75
多模型	403G	126M	Overall	48.23	79.39	51.26
GFL	131G	36M	SARDet-100K	57.31	87.44	61.99
O-RCNN	136G	45M	DOTA	45.31	77.70	46.45
O-RCNN	136G	45M	DroneVehicle	46.09	74.78	52.79
简单联合训练	403G	66M	Overall	47.05	77.56	50.11
			SARDet-100K	53.46	84.11	57.29
			DOTA	45.18	76.37	46.78
			DroneVehicle	44.99	73.28	51.50
DA +ConvNext-T	403G	66M	Overall	48.37	79.76	51.66
			SARDet-100K	53.86	84.93	58.09
			DOTA	46.23	78.47	47.58
			DroneVehicle	48.21	77.43	56.16
UniDet (Partitioned)	403G	66M	Overall	48.47	79.55	52.01
			SARDet-100K	53.81	84.70	57.43
			DOTA	46.49	78.28	48.59
			DroneVehicle	47.99	77.17	55.74
Uncertainty loss	403G	66M	Overall	48.79	79.99	52.50
			SARDet-100K	53.43	84.81	57.41
			DOTA	46.94	78.73	49.08
			DroneVehicle	48.78	77.96	56.88
SM3Det (仅用 DSO)	403G	66M	Overall	49.40	80.19	52.93
			SARDet-100K	58.54	88.59	62.67
			DOTA	46.18	77.86	47.95
			DroneVehicle	48.09	77.09	56.20
SM3Det	487G	178M	Overall	**50.20**	**80.68**	**53.79**
			SARDet-100K	60.64	89.94	65.06
			DOTA	46.47	77.88	48.24
			DroneVehicle	48.87	77.99	56.90

高结果。该模型还展现出随模型尺寸增加而合理扩展的能力。本章还使用不同的检测器评估 SM3Det。由于光学数据集（DOTA）和红外数据集（DroneVehicle）都涉及 OBB 回归任务，本章在模型中使用相同的头部网络结构。相比之下，对于涉及 HBB 回归任务的 SAR 数据集（SARDet-100K），本章实现了一个标准的水平目标检测头。图 4.5 展示了单阶段（RetinaNet[307]、GFL[222] 和 S^2ANet[21]）和两阶段（F-RCNN[17]、Cascade F-RCNN[154]、O-RCNN[7] 和 RoI-Transformer[308]）

检测器组合上对 SM3Det 的评估。结果表明，在所比较的检测器组合中，SM3Det 相比分别训练模型均取得更高结果。

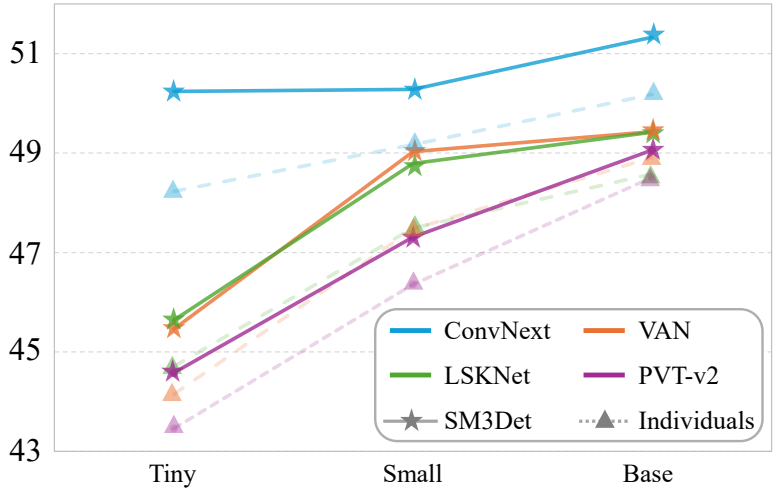


图 4.4 SM3Det 在不同骨干网络上的表现。

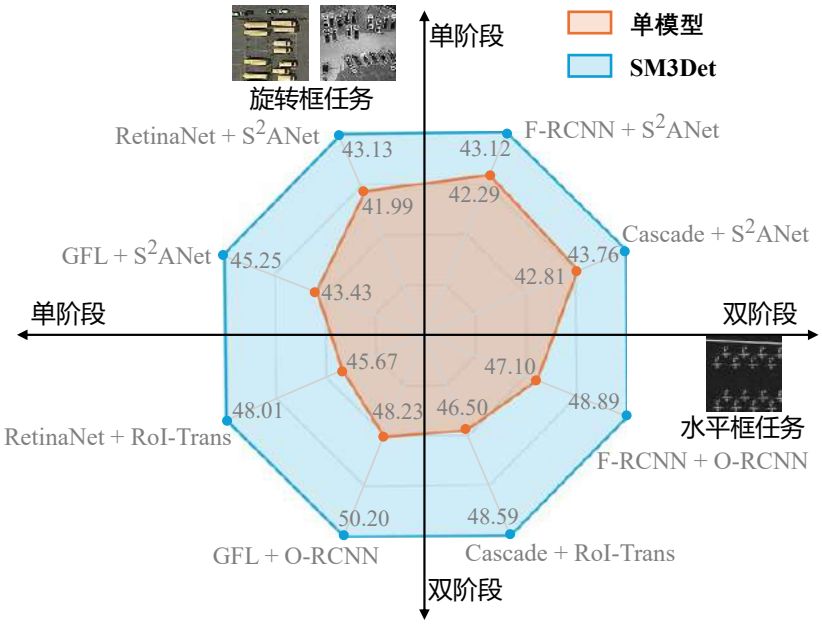


图 4.5 SM3Det 在不同检测器头上的表现。

四、分析与对比实验

（一）专家数和 top-k 数

在稀疏 MoE 架构中，添加的专家数 (N) 和 top-k 值对模型的性能和效率起着重要作用。增加 N 通常会增强模型的表示能力，而更高的 top-k 值允许将更专

表 4.2 具有不同专家数和 top-*k* 选择配置的 MoE 骨干网络实验。除最后两列外，为验证效率，专家仅应用于最后两个阶段的偶数索引层。*N*：添加的专家数量。k：激活的专家数量。兼顾性能与计算效率的较优配置为 8 个专家，top-*k* 值为 2。

MoE (*N*, k)	无 MoE	2, 2	4, 2	6, 2	8, 2	10, 2	8, 1	8, 2	8, 3	8, 2 图像级	8, 2 网格级
FLOPs (G)	403	469	469	469	469	469	403	469	531	487	487
参数量 (M)	66	82	113	142	174	205	174	174	174	178	178
mAP	48.51	48.94	49.11	49.13	**49.31**	49.24	49.05	**49.31**	49.13	48.25	**50.20**
@50	79.70	80.25	80.10	79.74	**80.26**	80.18	79.72	**80.26**	79.98	79.10	**80.68**
@75	51.78	52.01	52.13	52.76	52.84	52.79	52.30	52.84	**52.77**	51.31	**53.79**

业的知识应用于每个输入。然而，这些增强是以更大的模型尺寸、增加的计算复杂度以及可能需要更多训练数据以确保每个专家得到充分训练为代价的。因此，选择合适的专家数和 top-*k* 值对于在模型性能和计算效率之间取得平衡至关重要。表 4.2 中的结果强调了在稀疏 MoE 架构中调整专家数和 top-*k* 值的重要性。结果表明，就平衡性能和计算效率而言，这种稀疏 MoE 架构的较优配置是 8 个专家，且采用 top-2 专家激活。此配置有助于模型从不同输入中学习，同时不会引入不必要的复杂性或导致过拟合。

（二）图像级 MoE 对比网格级 MoE

在表 4.2 中，网格级 MoE 的表现优于图像级 MoE，这表明网格级专家在捕捉多模态图像中不同物体的空间变化方面更具优势。通过在更精细的空间粒度上处理特征，专家能更好地关注物体定位，这使得网格级 MoE 更适用于目标检测任务。

（三）网格级专家激活行为分析

本章可视化了经过良好调优的 ConvNext-T 骨干网络最后三个阶段中每个网格区域的选择结果。在此可视化中，每个方格代表该阶段相应的感受野，由不同专家处理的局部深度特征用不同颜色表示。图 4.6 展示了 top-1 选定的专家。对于 RGB 和红外图像，出现了一个一致的模式：专家 1 主要处理前景物体，而专家 3 在所有三个阶段都专注于背景区域块。相比之下，SAR 图像的情况更为复杂。特别是在第 4 阶段，有三个专家（专家 1、专家 4 和专家 6）负责处理背景区域，而专家 1 也处理船只物体。

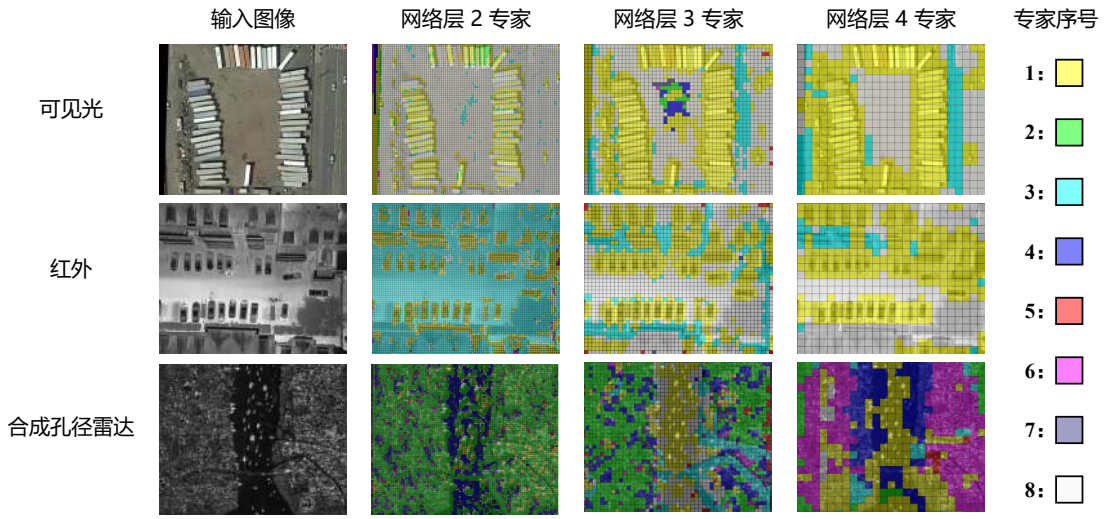


图 4.6 在 SAR、RGB 和红外图像上，经过良好调优的骨干网络最后三个阶段中网格专家激活的可视化。每个方格代表给定阶段的感受野，不同颜色表示由不同专家处理的局部网格区域。此图仅显示每个网格的 top-1 选定专家。每个专家专门处理独特的局部模式和语义。

（四）DSO 超参数

本章对提出的 DSO 方法的每个组成部分，以及其两个关键超参数的敏感性进行了消融研究。结果总结在表 4.3 中。省略对头部或骨干网络的学习率调整会导致性能下降。偏置参数 b 和温度 τ 动态调整学习率，以适应不同的任务和模态难度。具体来说，b 作为重新加权的平衡点，意味着当计算的一致性分数等于 b 时，重新加权因子为 1。当温度固定为 3 时，偏置值 $b = 0.4$ 在学习率调整中取得了良好的平衡。值得注意的是，b 的变化并未明显影响性能，表明该方法对偏置变化具有鲁棒性。关于温度 τ，它影响网络头部和骨干网络学习率调整机制中的重新加权曲线。较大的值会导致更急剧、更敏感的调整。温度 $\tau = 3$ 在稳定性和响应性之间提供了较好平衡。总之，$\tau = 3$ 和 $b = 0.4$ 的组合在该实验设置下取得了较好性能，有效地管理了跨不同任务和数据集的学习率调整。

表 4.3 关于 DSO 方法在不同温度 (τ) 和偏置 (b) 下的实验。DSO 对偏置 b 不敏感。

τ, b	3, 0.3	3, 0.4	3, 0.5	3, 0.6	2, 0.4	3, 0.4	4, 0.4	无 DSO	无头子模块	无骨干子模块
mAP	50.14	**50.20**	50.07	50.03	49.92	**50.20**	50.03	49.47	49.86	50.11
@50	80.61	**80.68**	80.66	80.61	80.55	**80.68**	80.44	80.33	80.53	80.66
@75	**53.81**	53.79	54.00	53.98	53.56	53.79	53.79	52.98	53.44	53.70

表 4.4 空间 MoE 在不同骨干阶段位置的实验。“None”表示无 MoE 层，“Even”表示仅在偶数索引层加入 MoE，“All”表示该阶段所有层都加入 MoE。每个 MoE 层包含 8 个专家，并采用 top-2 选择。

阶段 1	阶段 2	阶段 3	阶段 4	FLOPs	参数量	**mAP**	@50	@75
None	None	None	None	403G	66M	48.51	79.70	51.78
None	None	None	Even	422G	132M	48.85+(0.34)	80.07(+0.37)	51.86+(0.08)
None	None	Even	Even	469G	174M	49.31+(0.80)	80.26(+0.56)	52.84+(1.06)
None	Even	Even	Even	487G	178M	**49.53+(1.02)**	**80.47+(0.77)**	**53.06+(1.28)**
Even	Even	Even	Even	506G	179M	49.47+(0.96)	80.33+(0.63)	52.98+(1.20)
All	All	All	All	572G	249M	49.30 +(0.79)	80.23+(0.53)	53.03+(1.25)

（五）MoE 层位置

除了专家数量与 top-*k* 之外，MoE 层插入在骨干网络的哪些阶段同样会影响模型表现。为此，本章进一步研究了在 ConvNext 骨干不同阶段引入 MoE 层的效果，结果如表 4.4 所示。

结果表明，将 MoE 主要布置在后面三个阶段的偶数索引层最为有效，在仅带来适度计算开销的同时，使整体 mAP 提升了 1.02 个百分点。这说明更深层的语义特征更适合进行专家分工，而过度密集地引入 MoE 则可能带来额外优化负担，削弱其收益。

（六）跨模态专家路由行为分析

为了进一步理解 SM3Det 如何在共享知识与模态特性之间取得平衡，本章分析了网格级专家在不同模态中的激活行为。图 4.7、图 4.8 和图 4.9 分别展示了 DOTA、DroneVehicle 和 SARDet-100K 图像上最后三个阶段的专家选择结果，其中每个网格显示其 top-1 专家。

从图中可以看到，RGB 与红外图像在局部结构和目标区域上存在一定共享的专家模式，而 SAR 图像则会激活更具独特性的专家组合，尤其在深层阶段表现更为明显。这说明网格级 MoE 并非简单把不同模态“硬分开”，而是在共享与专门化之间形成了更细粒度的动态路由。

除定性可视化外，本章还统计了三种数据集上各专家的总体参与程度，如图 4.10 所示。

统计结果与 SM3Det 的设计目标相一致：一部分专家在三种模态上都有较高参与度，承担跨模态共享表示学习的职责，另一部分专家则主要服务于某一特定模态，负责捕捉该模态独有的视觉模式。尤其值得注意的是，SAR 图像通

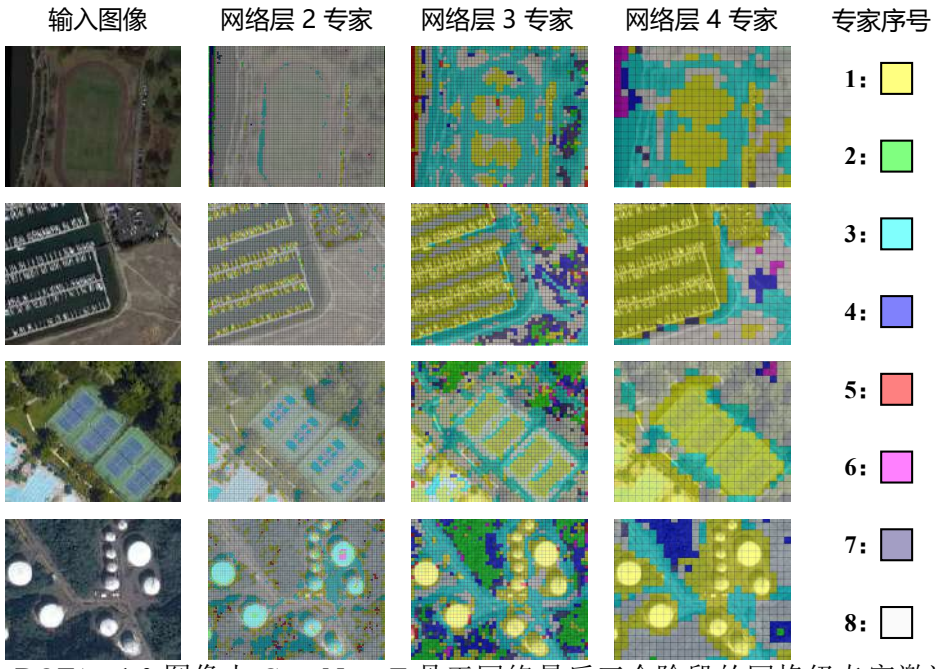


图 4.7 DOTA-v1.0 图像上 ConvNext-T 骨干网络最后三个阶段的网格级专家激活可视化。不同颜色表示不同专家处理的局部区域。

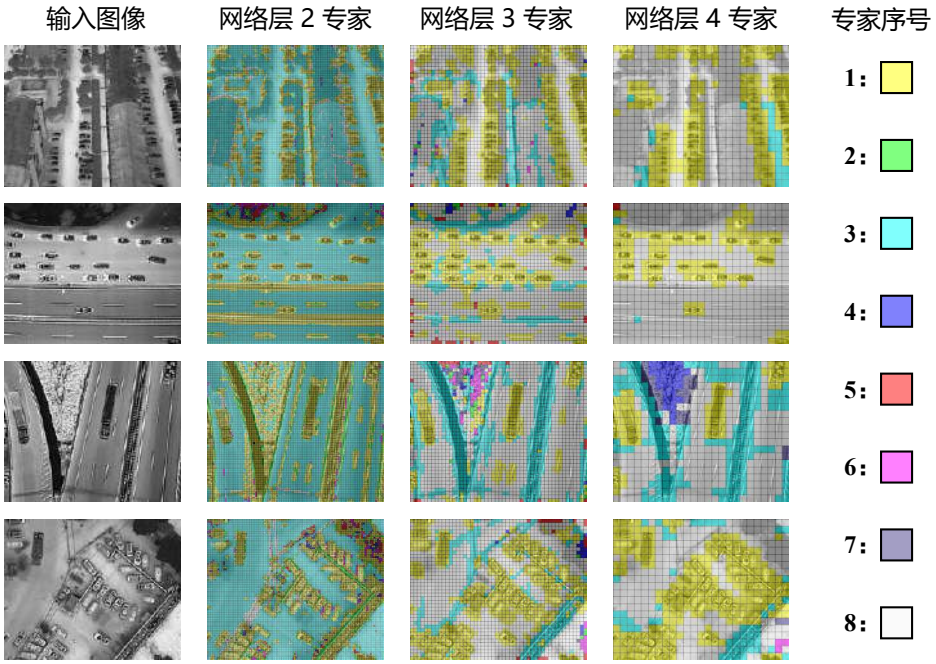


图 4.8 DroneVehicle 图像上 ConvNext-T 骨干网络最后三个阶段的网格级专家激活可视化。不同颜色表示不同专家处理的局部区域。

常会激活与另外两种模态明显不同的一组专家，而红外与 RGB 之间则存在更高的重叠度，这与它们在视觉表征上的相对接近性是一致的。

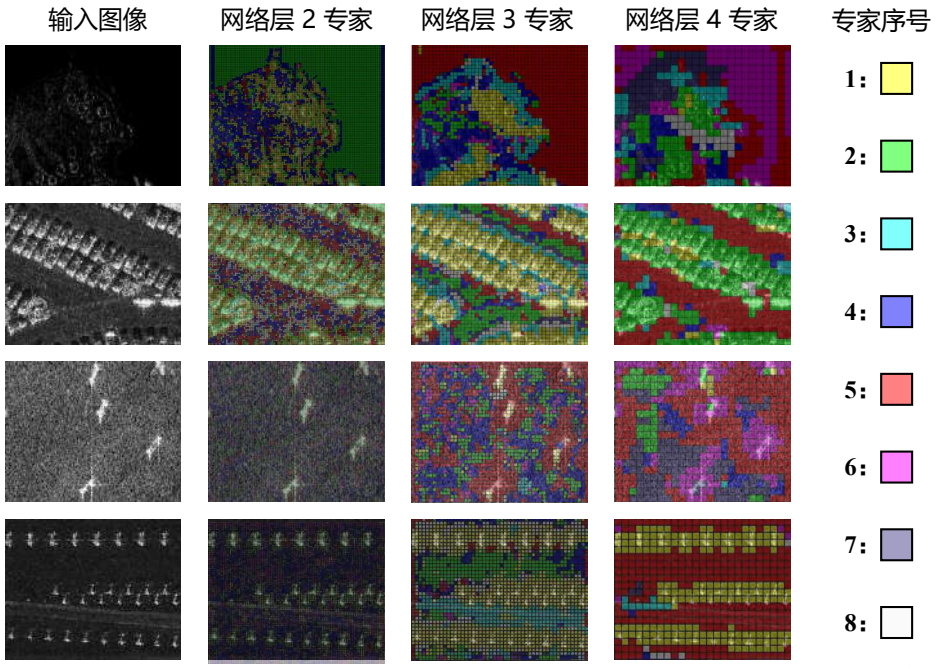


图 4.9 SARDet-100K 图像上 ConvNext-T 骨干网络最后三个阶段的网格级专家激活可视化。不同颜色表示不同专家处理的局部区域。

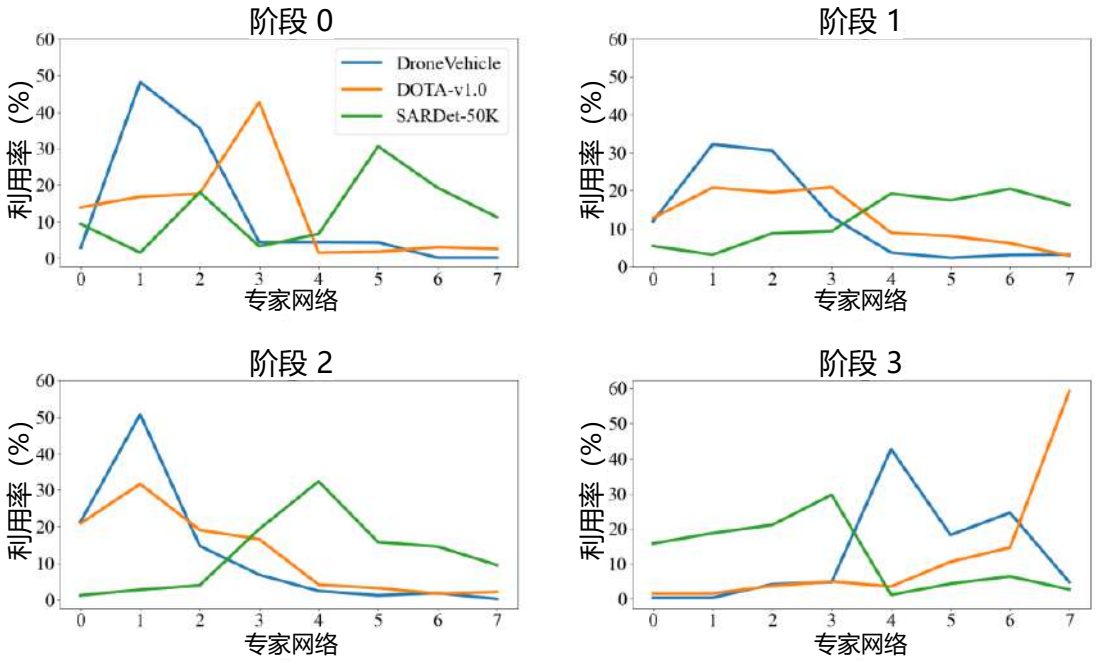


图 4.10 跨 SARDet-100K、DOTA 和 DroneVehicle 数据集的专家参与统计。一部分专家在多模态间共享激活，另一部分专家则表现出明显的模态专属性。

第五节 本章小结

总之，本章介绍了一项新的遥感领域多模态数据集与多任务目标检测任务。为了解决这个问题，本章开发了 SM3Det 模型，该模型集成了网格级 MoE 方法和动态子模块优化机制。实验和分析表明，SM3Det 具备较好的性能和泛化能力，也说明采用“共享骨干能力 + 稀疏专家分工 + 协同优化”的思路，可以缓解异构多模态遥感检测中的知识割裂与训练冲突问题。

本章工作仍存在一定局限。动态子模块优化主要依据不同任务损失轨迹和相对收敛状态调整子模块学习率，尚未显式建模多任务梯度之间的冲突方向与冲突强度。而 GradNorm、PCGrad 等梯度级方法虽然能够提供更直接的冲突刻画，但在大规模多模态检测训练中会引入额外反向传播或梯度存储开销。未来可探索更高效的梯度冲突近似估计方法，并进一步研究专家路由、损失平衡和检测头优化之间的联合调度机制。

第五章 语言引导的遥感基础模型预训练与跨模态对齐

第一节 引言

本章对应绪论中范式层挑战，核心问题是如何利用语言和高层语义推动遥感基础模型从“视觉特征学习”走向“语义引导的感知基础能力构建”。前几章已经分别从空间感知骨干、SAR 数据与迁移、异构多模态统一架构三个层面展开研究。但这些工作仍主要回答“模型结构如何适配复杂遥感场景”。若预训练目标仍停留在底层重建、图像级对比或下游微调阶段的隐式对齐，模型获得的表征往往难以同时满足细粒度定位、复杂空间关系理解和异构模态稳定协同的需求。

数十年来的认知神经科学研究表明，人类视觉感知并非单纯的自底向上过程，而是由底层输入、高层知识和任务目标共同调节形成的动态系统。现代计算机视觉从卷积网络到掩码图像建模（MIM）[26]与对比学习[27-28]，大多沿着“先学习视觉统计，再形成语义理解”的路径发展[72,309]。然而，高层理解、注意力与先验知识同样会反向调制底层感知，即使在视觉皮层较早阶段，认知目标也会主动筛选和重构输入信号[310-312]。这提示我们：对于复杂遥感场景，语言和任务语义不应只作为模型输出端的解释工具，也可以在预训练阶段直接参与视觉表征塑造。

从这一视角看，本章包含两个递进问题。第一个问题面向单一视觉骨干：如何让视觉语言模型中的高层语义监督直接作用于遥感视觉编码器，使其获得更适合检测、分割、变化检测等密集感知任务的表征？为此，本章提出遥感基础模型视觉指令预训练 ViTP。ViTP 将视觉 Transformer 骨干嵌入视觉语言模型中，通过领域化视觉指令遵循目标进行端到端持续预训练，并引入视觉鲁棒学习（Visual Robustness Learning, VRL），在稀疏视觉词元条件下增强语义保持能力。

第二个问题面向异构多模态遥感：当 RGB、SAR、红外等模态来自不同成像机制时，如何避免在下游微调阶段同时学习“模态对齐”和“任务检测”所造成的优化耦合？第四章的 SM3Det 已经从统一架构层面缓解了多模态部署和参数共享问题，但模态语义对齐仍需要更前置、更稳定的预训练机制。为此，本

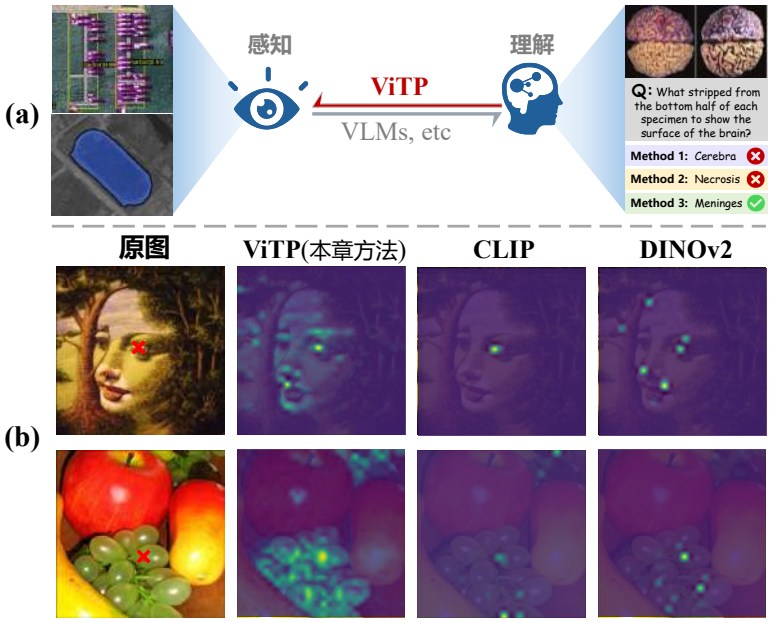


图 5.1 (a) 现代计算机视觉中感知与理解的协同关系。ViTP 补足了一条由高层理解反向塑造底层感知的关键路径。(b) 给定查询块（红色叉号）的自注意力激活图。ViTP 能够聚焦于与高层语义相关的细粒度目标部件。

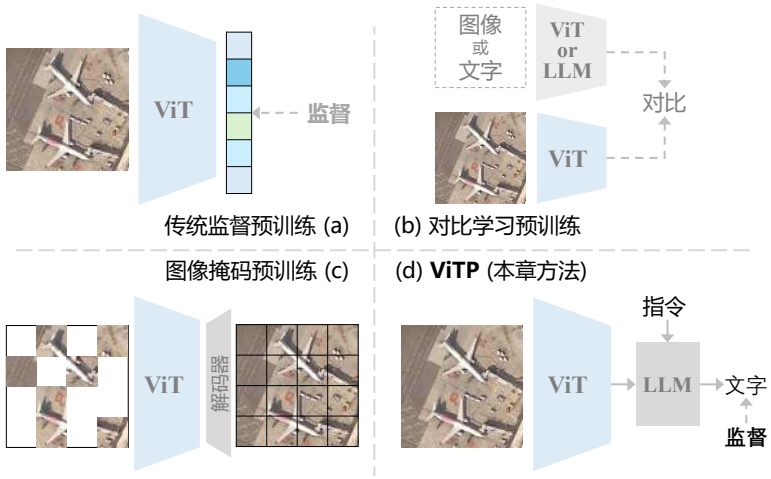


图 5.2 视觉 Transformer（ViT）基础模型预训练范式对比。ViTP 通过视觉指令遵循目标，将特定领域的高层语义监督直接注入视觉骨干。

章进一步提出语言引导的遥感异构多模态预训练 BabelRS。BabelRS 利用共享语言概念作为跨模态语义锚点，通过概念共享指令对齐（CSIA）和层级视觉语义退火（LVSA）将异构模态对齐前移到预训练阶段，从而减少微调阶段的梯度冲突和数值不稳定。

因此，ViTP 与 BabelRS 是本章中围绕“语言如何参与遥感基础模型预训练”

的两个层次：ViTP 回答语言如何塑造视觉骨干，BabelRS 回答语言如何组织异构模态。前者提供语义引导的感知基础能力，后者在此基础上进一步将语言作为跨模态对齐的公共坐标系，使本论文从单模态感知能力、数据迁移、统一架构进一步推进到语言驱动的基础模型预训练与对齐。

本章的主要贡献概括如下：

- 提出遥感基础模型视觉指令预训练 ViTP，通过端到端指令遵循目标将高层语义监督注入视觉 Transformer 骨干，并利用 VRL 提升视觉表征的鲁棒性与数据效率。
- 提出语言引导的遥感异构多模态预训练 BabelRS，通过概念共享指令对齐和层级视觉语义退火，在不依赖严格空间配对样本的条件下实现跨模态早期语义对齐。
- 在通用视觉、光学遥感、SAR 遥感、多模态检测和鲁棒性分析等实验中验证两类方法的作用，说明语言既可以塑造单模态感知骨干，也可以稳定组织异构多模态表征。

第二节 相关研究与问题分析

一、通用视觉预训练与遥感基础模型

通用视觉基础模型的发展，大体经历了由有监督预训练向自监督预训练演进的过程。早期方法主要依赖 ImageNet[22-23]等大规模标注数据，通过分类监督学习具有语义区分性的表征。为摆脱对人工标注的依赖，对比学习和掩码图像建模逐渐成为两条代表性路线：前者以 MoCo[57]、BYOL[58]和 DINOv2[27]为代表，强调不同增强视图之间的语义一致性。后者以 MAE[26]、SimMIM[313]和 UM-MAE[314]为代表，通过重建被遮蔽区域来学习上下文相关的视觉结构。这些方法推动了视觉预训练的发展，但其基本逻辑仍然是由底层视觉统计逐步归纳高层语义。

遥感领域也围绕专业场景开展了基础模型探索。由于遥感数据在尺度分布、成像模态和目标类别上明显不同于自然图像，预训练策略逐渐朝领域化方向演进[31,33,269,315]。SAMRS[316]、MSFA[317]等方法利用标注遥感数据开展监督式或半监督式预训练。RingMo[29]、SatMAE[30]和 Scale-MAE[31]等方法以 MIM 为基础，面向遥感影像中的密集小目标、多尺度结构和多模态观测特点进行定制。GeoRSCLIP[318]、RemoteCLIP[32]、CACo[59]和 SkySense[33]等工作则分别从图文

对比和图像对比方向探索遥感表征学习。这些研究提升了遥感基础表征质量，但多数仍主要关注“如何从视觉信号中组织语义”，对“高层语义如何反向塑造视觉骨干”讨论不足。

二、视觉指令学习与领域持续预训练

以 LLaVA 为代表的视觉指令微调范式推动了 Gemini、InternVL 和 Qwen-VL 等视觉语言模型的发展[34,61-62,319]。这类方法通常将视觉编码器输出投影到大语言模型嵌入空间，再利用图像-文本指令对进行监督，使模型具备问答、描述和推理能力。该范式增强了多模态理解能力，但在许多实现中，视觉编码器往往保持冻结或仅被轻度更新，主要被优化的部分仍是语言模型及其投影模块。也就是说，现有视觉指令学习更多是在利用感知提升推理，而非利用推理改善感知。

持续预训练为专业领域迁移提供了重要路径。Gururangan 等人[236]指出，在目标领域数据上继续预训练能够提升模型的领域适配能力。类似思想随后被引入视觉领域：CSPT[237]在 ImageNet 预训练基础上继续利用遥感数据执行 MIM 目标。TOV[320]通过课程学习策略逐步引入专业数据。RemoteCLIP[32]从 CLIP[28]初始化并在遥感图文数据上继续训练。第三章中的 MSFA 也通过多阶段桥接缩小自然图像与 SAR 数据之间的域鸿沟。然而，现有持续预训练大多仍服务于分类、重建或对比目标，尚未充分利用视觉语言模型的指令遵循损失来直接优化遥感视觉骨干。

三、异构遥感多模态融合与语言对齐

多模态遥感目标检测最早主要围绕空间配准样本展开。对于 RGB-红外、光学-SAR 等成对输入，特征级融合能够直接利用不同传感器的互补信息，因此成为主流路线[33,282]。相关研究引入光照感知、自适应融合、差分模态校准、区域级对齐和交叉注意力等机制[48-51,321-323]，在严格配准条件下取得了较好结果。但这一路线的核心假设是可以稳定获得空间对应的多传感器样本，而真实遥感系统常受到平台、时间、视角和成像机制差异影响，大规模高质量配对数据并不容易获得。

在通用多模态学习中，语言已被证明是组织异构表征的有效语义锚点。CLIP[28]通过对比学习建立共享嵌入空间，ImageBind[36]将多种感知模态绑定到统一表征空间，LanguageBind[37]则进一步将语言置于中心位置，使红外、视频等模态能够与冻结的大语言模型语义空间对齐。类似思想也被扩展到分子、蛋

白质等专业领域[63,324-325]。这些工作说明：即便不同模态在像素统计和成像机制上差异明显，只要语义概念可以共享，就有可能借助语言建立高层表征对齐。但现有语言对齐方法大多强调全局语义一致性，或默认可以获得成对样本，仍难以直接满足遥感密集检测所需的多尺度空间表征。

四、问题分析与本章定位

综上，语言引导的遥感基础模型预训练需要同时解决两个层次的缺口。第一，视觉语言模型能否不仅服务于高层理解，还能反向塑造专业领域视觉骨干，使其更适合小目标、密集目标、SAR 散射结构和变化区域等复杂感知任务？第二，在 RGB、SAR、红外等异构模态之间缺乏严格配对样本时，语言能否作为共享语义坐标系，将跨模态对齐从下游微调阶段前移到预训练阶段，并缓解晚期对齐中的优化冲突？

本章据此形成递进式方案：ViTP 首先利用视觉指令遵循目标和 VRL，让高层语义监督直接参与遥感视觉骨干优化。BabelRS 进一步把语言概念作为异构模态之间的公共锚点，通过 CSIA 和 LVSA 实现跨模态早期对齐。二者共同构成本论文在范式层的回答：遥感基础模型不仅需要结构上的统一和数据上的迁移，也需要在预训练目标中引入语言语义，使模型在进入下游任务之前具备更好的领域感知能力和跨模态组织能力。

第三节　遥感基础模型视觉指令预训练 ViTP 方法

与传统视觉预训练主要依赖重建或对比目标不同（如图5.2），ViTP 直接借助现代视觉语言模型的理解能力，以“视觉指令遵循”为核心训练目标驱动骨干学习。具体而言，ViT 需要提取能够支持大语言模型正确回答图像相关指令的视觉特征。为使该过程更好地服务于专业领域下游任务，本章进一步设计了面向领域数据的“数据配方”，并在训练阶段引入视觉鲁棒学习（VRL）这一词元级正则化机制。经过 ViTP 预训练后，得到的 ViT 骨干可作为统一的基础表示，迁移到检测、分割和变化检测等多类下游任务中。

一、视觉指令遵循目标

ViTP 的核心假设是：若视觉骨干在训练过程中持续接受高层语义任务的牵引，就能够学到更符合下游感知需求的特征表示。基于这一假设，本章将预训练构造为一个视觉指令遵循任务。其基本流程如图 5.3所示：输入图像首先经由

ViT 编码器得到一组视觉词元，随后这些词元被映射到 LLM 的嵌入空间，并与指令文本的词元序列拼接。LLM 在联合序列上生成回答，而回答损失则端到端地回传至视觉骨干。这样，高层理解任务的监督便可以直接作用于 ViT 的特征学习过程。

本章采用持续预训练策略[236]，从一个已经具备通用视觉语言能力的 VLM 出发，再在精心构建的领域化指令数据上继续训练。这样既能继承通用模型已有的视觉语言知识，又能降低从零开始训练的计算成本。设原始数据集为 $\mathcal{D}_{\text{raw}} = \{(I, Q, R)\}$，其中 I 表示图像，Q 表示文本查询或指令，R 表示对应的目标回答。经过预处理后，训练数据被记为 $\mathcal{D} = \{(x_i, x_t, y^\star)\}$，其中 x_i、x_t 和 $y^\star$ 分别表示图像词元、文本词元以及目标回答词元。

（一）ViT 特征提取与投影

给定输入图像 $I \in \mathbb{R}^{H\times W\times 3}$，ViT 骨干首先将其划分为若干非重叠图像块，并将每个图像块线性映射为词元表示。随后，这些块词元与分类词元共同输入一系列 Transformer 模块[326]，得到输出视觉词元序列 $x_i' = \{t_1, t_2, \ldots, t_N\}$，其中 N 为序列长度。为了与 LLM 的词元嵌入空间对齐，本章进一步引入一个轻量级投影层（例如两层 MLP），将 x_i' 映射为最终的图像词元 x_i。

（二）指令遵循词元连接

对于每幅图像，其对应的文本查询 Q 会通过 LLM 分词器转换为文本词元序列 x_t。随后，投影后的图像词元 x_i 与文本词元 x_t 进行拼接，形成统一的多模态输入序列。为保留视觉词元的空间位置信息以及文本词元的顺序信息，本章在二者的嵌入上加入位置编码。最终输入到 LLM 的序列表示为：

$$S_{llm} = [\mathrm{PE}(x_i); \mathrm{PE}(x_t)] \tag{5.1}$$

其中 $\mathrm{PE}(\cdot)$ 表示为词元嵌入加入位置编码，$[;]$ 表示序列拼接。

（三）基于 LLM 的监督

组合序列 S_{llm} 由 LLM 处理。此时，LLM 在指令 x_t 的约束下对来自 x_i 的视觉信息进行语义解释，并以自回归方式生成输出序列 O。整个系统的训练目标，是最小化生成结果 O 与目标回答 $y^\star$ 之间的差异。例如，当指令要求识别目标类别或描述其空间位置时，$y^\star$ 可以是相应的结构化文本描述。在预训练过程中，输出损失的梯度会反向传播到投影层和 ViT 骨干，因此视觉表征能够直接受到

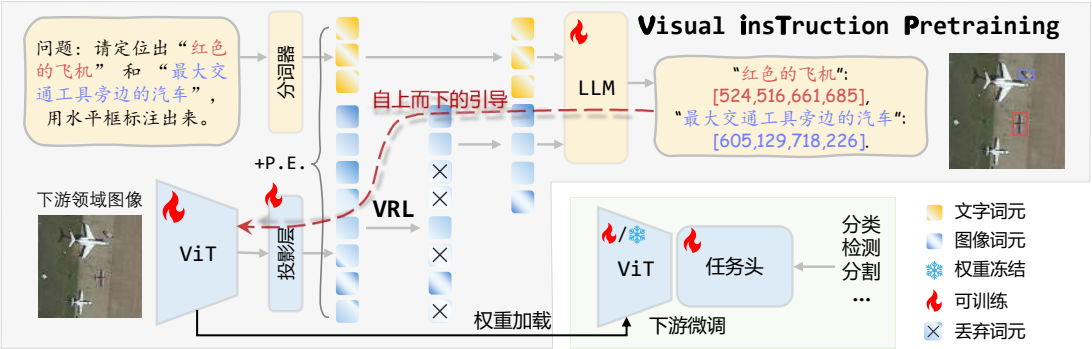


图 5.3 ViTP 框架示意图。该方法将视觉 Transformer（ViT）骨干嵌入视觉语言模型中，并通过领域化指令遵循目标与视觉鲁棒学习（VRL）进行持续预训练，从而把高层语义监督注入视觉骨干，得到的权重随后用于初始化各类下游感知模型。

高层语义监督的塑造。本章采用标准监督微调（SFT）目标，即最小化目标序列的负对数似然：

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = \mathbb{E}\left[-\log P_\theta\left(y^\star \mid S_{llm}\right)\right], \tag{5.2}$$

其中 $(x_i, x_t, y^\star) \sim \mathcal{D}$，$P_\theta$ 是由整个模型 θ 参数化的文本序列上的概率分布。

二、视觉鲁棒学习

为促使模型学习更鲁棒、更具语义浓度的视觉表征，本章引入视觉鲁棒学习（VRL）作为预训练阶段的正则化机制。如图 5.3所示，VRL 在投影后的图像词元 x_i 与文本词元 x_t 拼接之前，随机丢弃其中相当比例的视觉词元。该操作在位置编码加入之后执行，从而保证 LLM 仍然能够感知保留词元的原始空间位置。对应的训练目标写为：

$$\mathcal{L}_{\mathrm{VRL}}(\theta) = \mathbb{E}\left[-\log P_\theta\left(y^\star \mid [\mathcal{C}_r(\mathrm{PE}(x_i)); \mathrm{PE}(x_t)]\right)\right], \tag{5.3}$$

其中 $\mathcal{C}_r$ 表示按比例 r 随机丢弃词元的采样操作。对于任意序列 $\mathcal{S}$，其形式化定义为：

$$\mathcal{C}_r(\mathcal{S}) \sim \left\{X \subseteq \mathcal{S} \;\middle|\; |X| = \lceil (1-r)\cdot|\mathcal{S}| \rceil\right\}, \tag{5.4}$$

其中子集 X 在保持原始顺序的前提下均匀随机选取，且 $r \in [0,1)$。这种稀疏观测机制迫使 ViT 在较少的视觉词元中编码更完整的信息，因为模型必须依赖部分输入恢复整体视觉语义。由此，视觉骨干会倾向于学习更分布式、更稳健的注意力模式。与此同时，较高比例的词元丢弃（例如 $r = 0.75$）还能降低显存占用并提升训练效率，从而增强 ViTP 的可扩展性。

三、预训练数据集配方

随着 VLM 快速发展，面向专业领域的图文指令数据集不断涌现。ViTP 的效果在很大程度上取决于预训练语料的质量与组成。基于此，本章在构建数据配方时遵循以下四项原则：

1. **规模与多样性：**数据集必须规模大且多样，包含代表目标领域的广泛视觉概念、场景和对象。
2. **模态覆盖：**数据必须涵盖下游任务中预期的所有成像模态。例如，如果下游任务使用遥感合成孔径雷达（SAR）图像，预训练混合数据中应包含此类模态数据，以确保模型学习到特定模态的特征。
3. **任务能力对齐：**预训练数据中的指令遵循任务应培养下游所需的能力。例如，对于下游的目标检测任务，在预训练期间包含视觉定位和细粒度 VQA，可以增强骨干的定位和空间理解能力。
4. **保持通用性：**应该包含一定比例的通用领域自然图像（例如，来自公开 VLM 数据集的）。特定领域数据可能在多样性上有限，添加通用数据可以减轻过拟合，防止模型失去理解广泛视觉模式的基础能力。

四、下游微调

完成 ViTP 预训练后，本章将得到的 ViT 骨干提取出来，作为下游模型的初始化主干。具体而言，预训练 ViT 被接入标准的 ViT-Adapter[327]框架，并与任务相关的预测头组合，在目标数据集上进行端到端微调。这样做可以充分利用 ViTP 阶段获得的语义化表征，使模型在专业任务上实现更快适应和更优性能。

五、实现细节

从头训练 VLM 通常需要经历视觉-语言对比预训练、投影器对齐和大规模指令微调等多个阶段，计算代价很高。为降低这一成本，本章直接从公开可用的高容量 VLM 初始化。具体而言，本章采用 InternVL-2.5[35]作为起点，其视觉部分使用定制的 ViT-Large 骨干[61]，语言部分采用 Qwen2[328]。所有预训练实验均在 8 张 NVIDIA A40（48 GB）GPU 上进行，全局批量大小为 128。下游微调统一在 8× NVIDIA RTX3090（24 GB）GPU 上完成。ViTP 预训练使用 AdamW 优化器，学习率设置为 2e-5，并采用余弦衰减调度。遥感领域模型默认训练 8,000 步，而通用领域模型训练 16,000 步。为进一步提升训练吞吐量，本章在 Vision Transformer 的自注意力层中集成了 Flash-Attention[329]。

第四节 语言引导的遥感异构多模态预训练 BabelRS 方法

ViTP 说明了语言指令监督可以反向塑造遥感视觉骨干，但异构多模态遥感场景还面临另一类更难的问题：RGB、SAR、红外等传感器由不同物理机制产生，视觉统计差异很大，即使统一检测架构已经能够共享部分参数，模态对齐仍常常被推迟到下游微调阶段完成。BabelRS 进一步沿用“语言作为语义枢纽”的思想，把跨模态对齐前移到预训练阶段，使不同模态先围绕共享语言概念形成一致表征，再进入检测任务优化。

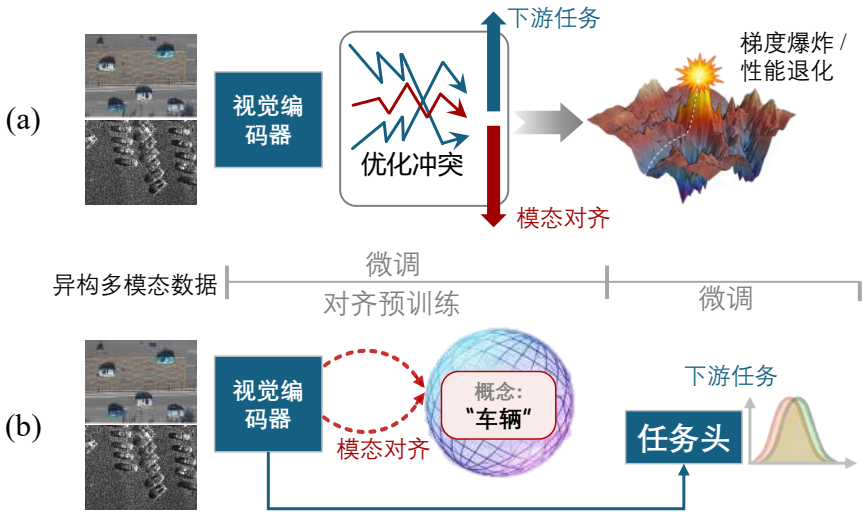


图 5.4 异构多模态遥感检测中两类学习范式的概念对比：(a) 晚期对齐；(b) 语言引导的早期对齐。晚期对齐在微调阶段同时处理模态对齐与任务优化，容易引发梯度冲突和训练不稳定。BabelRS 通过早期语义对齐将二者解耦，从而提升优化稳定性与泛化能力。

如图 5.4所示，现有方法通常采用“晚期对齐”策略：先用通用单模态预训练初始化骨干，再在微调阶段同时学习模态对齐和检测任务。当 SAR 散射信号、RGB 反射信号和红外热辐射信号被强行放在同一阶段耦合优化时，模态对齐目标与检测目标容易相互干扰。BabelRS 的核心思想是把跨模态语义统一前移到预训练阶段，使下游微调更专注于检测任务本身。

现有面向空间异构遥感数据的统一检测框架，通常在微调阶段同时承担模态对齐与任务学习两个目标。BabelRS 通过引入语言引导的预训练策略，在训练流程上显式解耦这两个目标。如图 5.5所示，整个框架由概念共享指令对齐和层级视觉语义退火两部分组成。

其中，概念共享指令对齐负责将不同遥感模态的视觉表示映射到共享语言空间，从而在不依赖空间配对样本的条件下实现隐式跨模态对齐。层级视觉语

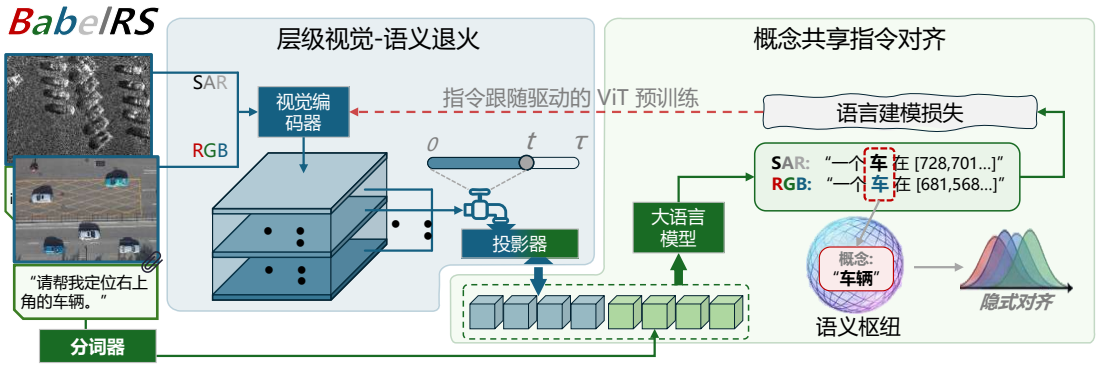


图 5.5 BabelRS 框架示意图。概念共享指令对齐借助指令遵循目标将异构遥感模态映射到共享语言语义空间。层级视觉语义退火则渐进式引入多尺度视觉特征，以弥合语言语义与密集检测之间的粒度差距。

义退火则通过逐步整合多尺度视觉特征，缓解语言级全局语义与密集检测需求之间的粒度差异。

一、概念共享指令对齐

设 $\mathcal{M}$ 表示包含 K 种遥感模态的集合（例如 RGB、SAR 和红外）：

$$\mathcal{M} = \{m_1, m_2, \ldots, m_K\}.$$

不同于依赖配对样本 (x^{rgb}, x^{sar}) 的既有方法，BabelRS 考虑一组彼此不相交的多模态遥感数据集 $\mathcal{D}$：

$$\mathcal{D} = \{\mathcal{D}^{m_1}, \mathcal{D}^{m_2}, \ldots, \mathcal{D}^{m_K}\},$$

其中每个数据集 $\mathcal{D}^{m_k} = \{(x_i^{m_k}, q_i^{m_k}, r_i^{m_k})\}$ 由图像、指令和回答三元组组成。这里，q_i 表示与图像 x_i 对应的自然语言问题或任务指令，r_i 表示相应的文本回答。这类指令-回答数据可以描述图像中的目标类别、空间关系或场景属性。

BabelRS 的核心假设是：尽管不同模态的像素分布 $P(x^m)$ 和成像机制差异明显，但它们的语义解释可以通过共享语言概念来表达。也就是说，存在一组模态相关映射，能够将异构观测统一到共同的语义概念 C 上，例如

$$f(x^{rgb}) \to C, \quad g(x^{sar}) \to C.$$

在这一设定下，$f(x^{rgb})$ 和 $g(x^{sar})$ 虽然来源于不同模态，却可以在诱导出的语义空间中实现隐式对齐。为将这一设想形式化，BabelRS 引入预训练大语言模型 Φ 作为语义枢纽，并使用模态共享视觉编码器 $E_{\mathcal{M}}$ 提取视觉特征，再将其投影到

Φ 的输入嵌入空间。对于任意模态图像 x 及其配套文本 $\{q, r\}$，优化目标写为：

$$\mathcal{L}_{\text{align}} = -\sum_{j=1}^{|r|} \log P_{\Phi}(r_j \mid q, r_{<j}, E_{\mathcal{M}}(x)).$$

通过要求异构模态产生一致的语言描述，视觉编码器被逐步牵引到共享语义流形之上。具体实现时，BabelRS 采用指令遵循范式[34]，将视觉词元与指令和回答对应的文本词元拼接起来，并只在回答词元上施加语言建模损失。这样设计的好处在于：它不仅将视觉表示投影到统一语言空间，还借助指令任务中天然包含的组合、推理和概念约束，促使视觉编码器学习信息更密集、语义更稳定的表征。更重要的是，这一对齐过程在进入下游检测之前独立完成，因此可以降低微调阶段的优化冲突。

二、层级视觉语义退火机制

尽管语义对齐为跨模态学习提供了基础，但密集目标检测还要求骨干具备良好的多尺度空间解析能力。大多数视觉语言模型只在最后一层与语言空间对齐，而最后一层更偏向全局语义，难以直接满足检测任务对局部结构和多尺度细节的需求。若简单聚合所有中间层特征，又容易破坏原有预训练分布并带来训练不稳定。为此，BabelRS 提出**层级视觉语义退火**（LVSA）机制，在保持预训练分布稳定的前提下，逐步把多尺度视觉特征引入语言对齐过程。设 $\mathcal{V}$ 表示来自 ViT 编码器的 L 个层级特征集合：

$$\mathcal{V} = \{F_l\}_{l=1}^{L}, \quad F_l \in \mathbb{R}^{H \times W \times C}.$$

BabelRS 从中选取一个待融合的层集合 $\mathcal{S} \subseteq \{1, \ldots, L\}$，并保证最后一层 $L \in \mathcal{S}$。为了使模型从单尺度表征平滑过渡到多尺度表征，引入与训练步数 t 及退火时长 τ 相关的融合系数 $\alpha(t)$：

$$\alpha(t) = \min\left(\frac{t}{\tau}, 1\right).$$

据此，融合特征 $\tilde{F}$ 由最后一层特征 F_L 与所选中间层特征的平均值进行动态插值得到：

$$\tilde{F} = (1 - \alpha(t))F_L + \alpha(t)\left(\frac{1}{|\mathcal{S}|}\sum_{l \in \mathcal{S}} F_l\right).$$

在训练初期，模型主要依赖最后一层输出，以尽可能保留原有预训练分布，随着训练推进（$t \to \tau$），较浅层和中间层特征被逐步纳入，从而在不过度扰动表

征空间的情况下增强空间定位能力。通过这种渐进式设计，下游检测器所需的多尺度特征不仅具备更好的空间分辨率，也保留了较强的语义一致性。

三、任务特定微调

完成预训练后，得到的编码器会被迁移到异构多模态目标检测任务中进行微调。与既有方法在微调阶段继续引入额外对齐模块不同，BabelRS 采用更简洁的设计：共享骨干负责承载跨模态公共表征，不同模态使用各自的检测头，训练时通过跨数据集随机采样实现联合优化。总损失写为各任务损失的求和：

$$Loss_{\text{total}} = \sum_{n} Loss_n.$$

由于跨模态对齐已在预训练阶段完成，微调阶段无需再引入辅助对齐项，因而优化可以更专注于检测性能本身。

四、调和模态 mAP (H-mAP)

多模态遥感数据集通常存在明显的跨模态类别不平衡。设 $\mathcal{C}_m$ 为模态 m 对应的类别集合。在 SOI-Det 基准中，$|\mathcal{C}_{RGB}|$ 往往大于 $|\mathcal{C}_{SAR}|$ 和 $|\mathcal{C}_{IR}|$。如果直接采用全局平均精度

$$mAP = \frac{1}{|\mathcal{C}_{total}|} \sum_{c \in \mathcal{C}_{total}} AP_c,$$

其中 $\mathcal{C}_{total} = \bigcup_{m \in \mathcal{M}} \mathcal{C}_m$，那么评价结果会天然偏向类别数更多的 RGB 模态。也就是说，一个模型即便在 SAR 或红外上表现较差，只要在 RGB 上得分较高，整体 mAP 仍可能看起来不错。为更客观地反映跨模态均衡性能，本章提出调和模态 mAP。首先分别计算各模态的 mAP：

$$mAP_m = \frac{1}{|\mathcal{C}_m|} \sum_{c \in \mathcal{C}_m} AP_c, \quad \forall m \in \mathcal{M}$$

再通过调和平均定义

$$H\text{-}mAP = \frac{|\mathcal{M}|}{\sum_{m \in \mathcal{M}} \frac{1}{mAP_m}}.$$

之所以选择调和平均，是因为它对低值更敏感，能够体现“短板效应”。与算术平均允许某一模态的高性能弥补另一模态的失败不同，调和平均会惩罚单一模态性能过低的情况。若存在某一模态的 $mAP_m \to 0$，则整体 $H\text{-}mAP \to 0$。因此，该指标更适合评价异构多模态检测模型是否真正具备均衡泛化能力。

第五节 实验与分析

本章实验围绕同一条主线展开：语言首先以指令监督的形式塑造视觉骨干，使其获得更适合遥感密集感知任务的基础表征，随后，语言进一步作为共享语义坐标系，将 RGB、SAR 和红外等异构模态的对齐过程前移到预训练阶段。基于这一逻辑，实验分为两组互补证据。第一组实验评估 ViTP，重点回答“语言指令监督能否提升视觉骨干的跨任务迁移能力”，因此覆盖通用视觉、光学遥感、SAR 遥感、分割、变化检测、训练效率、数据效率和鲁棒性等维度。第二组实验评估 BabelRS，重点回答“语言语义能否支撑异构模态的稳定早期对齐”，因此围绕 SOI-Det 基准、优化稳定性、中间层语义融合和退火调度展开。

两组实验的评价对象有所不同，但共同服务于本章核心论点：语言不仅是基础模型的高层交互接口，也可以在预训练阶段分别承担“塑造视觉骨干”和“组织异构模态”的作用。因而，本章将 ViTP 的视觉骨干迁移结果作为基础证据，再以 BabelRS 的多模态对齐结果进一步验证语言引导预训练在复杂遥感场景中的扩展价值。

一、ViTP 实验设置与数据集

下文首先说明 ViTP 所使用的预训练语料和下游微调数据集。预训练数据用于提供语言指令监督和领域语义先验，下游数据集则用于检验这些语义先验能否迁移到检测、分割和变化检测等遥感感知任务。

（一）ViTP 预训练数据构成

在 ViTP 预训练阶段，本章使用一组公开可得、类型多样的图文指令数据集，并严格遵循前文“预训练数据集配方”中给出的构建原则。

通用领域。通用领域预训练数据集如表 5.1所示。其中，ShareGPT4V[330]、DVQA[331]、ChartQA[332]、AI2D[333]、DocVQA[334]和 GeoQA+[335]主要提供多样化的视觉问答（VQA）监督，覆盖复杂推理、图表理解、文档理解和地理空间问答等能力，SynthDoG-EN[336]则补充了光学字符识别（OCR）相关监督。除问答外，ShareGPT4V 还提供视觉定位（VG）样本，用于增强空间定位能力，GAIA[337]与 Million-AID[338]则补充图像描述（CAP）和分类（CLS）任务，从而维持模型在通用视觉语言理解上的基础能力。整体而言，这一数据配方为模型提供了多任务、多场景和多语义粒度的训练信号，有助于形成稳健的通用视觉表征。

遥感领域。遥感领域预训练数据集如表 5.2所示。Million-AID[338]、GAIA[337]、LevirCCcaptions[339]、VHM[340]、RSVQA[341]和 FIT_RS[342]构成了遥感视觉指令语料的主体，覆盖场景理解、图像描述、属性判断和问答等多类任务。针对 SAR 模态，本章进一步引入 ISPRS_SAR[343]和 SAR Sentinel-1&2[344]，并基于这些数据构造与目标识别、场景判读和属性分析相关的问答样本，以强化模型对 SAR 成像特性的适应能力。与此同时，GeoChat[345]、DIOR-RSVG[346]、RSVG[347]和 VRSBench[348]主要提供视觉定位（VG）样本，用于加强空间理解与目标定位能力。最后，mini-InternVL[349]以低采样率保留在数据混合中，用于维持通用视觉语言理解能力，避免模型过度专化于遥感模式。整体来看，这一配方使 ViTP 能够学习到兼具模态适应性与任务相关性的遥感视觉表征。

（二）ViTP 下游微调数据集

自然图像检测与分割。自然场景中的目标检测与分割是计算机视觉中的基础任务，也是衡量通用视觉理解能力的重要基准。目标检测旨在对日常图像中的多个目标进行定位与分类，而分割任务则进一步提供像素级掩码，实现对目标区域的精确刻画。这类任务广泛服务于自动驾驶、增强现实等典型应用场景。为评估 ViTP 在通用视觉任务中的适用性，本章选用两个互补的自然图像基准：COCO[25]用于目标检测与实例分割，ADE20K[350]用于语义分割。遵循已有工作的常用设置[26-27]，本章在 COCO 任务上采用 Mask R-CNN[351]作为默认框架，训练周期为 12 个 epoch，并使用标准数据增强策略，在 ADE20K 语义分割任务

表 5.1 通用领域预训练数据集构成。

数据集	规模	采样率	任务
ShareGPT4V[330]	767k	1	VQA,VG
DVQA[331]	200k	1	VQA
ChartQA[332]	18k	1	VQA
AI2D[333]	12k	1	VQA
DocVQA[334]	10k	1	VQA
GeoQA+[335]	72k	1	VQA
SynthDoG-EN[336]	30k	1	OCR
GAIA[337]	33K	0.2	CAP
Million-AID[338]	920k	0.01	CAP,CLS

表 5.2 遥感领域预训练数据集构成。

数据集	规模	采样率	任务
Mini-InternVL[349]	1394k	0.03	CAP,VQA,OCR
RSVQA[341]	100k	0.1	VQA
FIT_RS[342]	100k	0.1	VQA
GeoChat[345]	64k	2	VG
VRSBench[348]	38k	5	VG
RSVG[347]	5.5k	10	VG
DIOR-RSVG[346]	27k	8	VG
ISPRS_SAR[343]	1.5k	1	CLS
SAR_Sentinel-1&2[344]	16k	1	CLS
VHM[340]	223k	1	CAP,VQA,CLS
LevirCCcaptions[339]	50k	0.5	CAP
GAIA[337]	33k	1	CAP
Million-AID[338]	920k	0.05	CAP,CLS

上，采用 UperNet[184]并按照 160k 次迭代的训练方案进行优化。评价指标方面，COCO 上的检测与分割任务报告标准 COCO 指标（mAP@0.5:0.95），ADE20K 语义分割任务报告平均交并比（mIoU）。

遥感目标检测。遥感目标检测[7,279,352-354]旨在从航空或卫星影像中识别并精确定位感兴趣目标，例如车辆、船舶和桥梁等。这一任务在城市规划、灾害监测等应用中具有重要意义。其主要挑战包括目标尺度变化大、朝向任意、分布密集以及背景复杂等。为评估本章模型在不同遥感目标检测场景下的适用性，本章在 DIOR[247]、DIOR-R[355]、DOTA-v2.0[279]、SARDet-100K、RSAR[356]和 SSDD[46]等数据集上开展实验，并使用标准 COCO 评价指标，即平均精度均值（mAP）衡量模型性能。遵循已有工作的常用做法[33,317,356]，本章对定向目标检测任务采用 Oriented R-CNN[7]作为默认检测器，对水平目标检测任务采用 Cascade R-CNN[357]。在训练与测试阶段，本章将所有光学数据集（DIOR、DIOR-R 和 DOTA-v2）统一缩放至 1024×1024 像素。对于原始尺寸较大的 DOTA-v2 数据集，本章采用单尺度设置，将图像裁剪为 1024×1024 的子图，并设置 200 像素重叠区域。SARDet-100K 和 RSAR 两个 SAR 数据集则统一缩放至 800×800 像素。

遥感语义分割。遥感语义分割[10,46,358]旨在将图像中的每个像素划分到预定义

表 5.3 下游任务数据集概览。

任务	数据集	模态	标注格式	图像数	类别数
通用领域					
检测与分割	COCO[25]	RGB	多边形	118,287	80
	COCO[25]	RGB	水平框	118,287	80
	ADE20K[350]	RGB	掩码	20,210	150
遥感领域					
目标检测	DIOR[247]	RGB	水平框	23,463	20
	DIOR-R[355]	RGB	定向框	23,463	20
	DOTA-v2.0[279]	RGB	定向框	11,268	18
	SARDet-100K	SAR	水平框	116,598	6
	SSDD[46]	SAR	水平框	1,160	1
	RSAR[356]	SAR	定向框	95,842	6
语义分割与实例分割	iSAID[358]	RGB	掩码	2,806	15
	LoveDA[165]	RGB	掩码	5,987	7
	UAVid[166]	RGB	掩码	5,510	8
	SSDD[46]	SAR	多边形	1,160	1
变化检测	SVCD[359]	RGB	掩码	16,000	2
	WHU[360]	RGB	掩码	11,456	2
	LEVIR-CD[361]	RGB	掩码	10,192	2
	S2Looking[198]	RGB	掩码	5,000	2

类别中，例如森林、水体、城市区域等地表覆盖类型，或特定类别目标。这种像素级理解对于环境监测、城市规划和资源管理等任务至关重要。本章在 iSAID[358]、LoveDA[165]、UAVid[166]和 SSDD[46]等常用遥感数据集上评估 ViTP 的分割性能。遵循已有工作的常用设置[29,33]，本章对语义分割任务（iSAID、LoveDA 和 UAVid）采用 UperNet[184]作为默认分割器，对实例分割任务（SSDD）采用 Mask R-CNN[351]。评价指标方面，语义分割任务报告平均交并比（mIoU），实例分割任务报告平均精度（AP）。

遥感变化检测。遥感变化检测关注通过不同时相的观测结果识别并刻画目标或地物状态的变化。这种双时相分析是城市扩张监测、森林砍伐追踪和城市化分析等应用的重要基础。该任务的主要挑战包括不同时相之间光照和大气条件差异带来的干扰、图像精确配准的要求，以及如何从复杂差异中区分真正有意义的语义变化。为评估 ViTP 在该任务上的能力，本章在四个广泛使用的公开数据集上开展实验，即 SVCD[359]、WHU-CD[360]、LEVIR-CD[361]和 S2Looking[198]。遵循常见做法，本章采用一个基于孪生 UperNet 的简洁框架[184]作为默认变化检

表 5.4 端到端微调预训练范式比较。结果展示了在 COCO val 上使用 Mask R-CNN（1x 调度）的目标检测（AP_{box}）和实例分割（AP_{mask}）性能，以及在 ADE20K 上使用 UperNet[184]（160k 次迭代）的语义分割（mIoU）性能。所有模型均使用 ViT-Large 骨干。

预训练范式	模型	COCO val (AP_{box})	COCO val (AP_{mask})	ADE20K (mIoU)
监督	IN. Cls.[23]	49.3	43.9	49.9
	DeiT III[362]	48.7	41.1	52.0
对比学习	CLIP[28]	51.3	-	52.2
	MoCov3[363]	49.3	44.0	49.1
	DINOv2[27]	53.4	<u>46.8</u>	<u>55.0</u>
掩码图像重建	BEiT[364]	53.3	47.1	53.3
	MAE[26]	53.3	**47.2**	53.6
	4M[365]	<u>53.7</u>	46.4	53.4
理解	**ViTP**	**53.9**	46.7	**55.8**

测器，并以 F1-Score 作为主要评价指标。

二、ViTP 下游迁移结果

下文从通用视觉和遥感两个方向系统评估 ViTP 的下游迁移能力，并进一步分析其训练效率、数据效率与鲁棒性。实验结果将与现有代表性预训练方法进行对比，其中较优结果以**粗体**表示，次优结果以<u>下划线</u>标记。

（一）通用视觉任务

为验证 ViTP 范式的有效性，本章在两种严格设置下与现有代表性预训练方法进行比较：一是端到端微调，二是冻结骨干后的特征评估。所有比较均统一使用 ViT-Large 骨干，以保证结果的可比性。

如表 5.4所示，ViTP 在通用视觉基准上同样表现出较强竞争力：在 COCO 目标检测（AP_{box} 53.9）和 ADE20K 语义分割（mIoU 55.8）上均取得较优结果，在 COCO 实例分割上也保持了较好性能（AP_{mask} 46.7）。与有监督预训练、对比学习和掩码图像建模等主流范式相比，ViTP 在不牺牲通用性的前提下体现出更好的下游适配能力。

为进一步衡量所学特征本身的质量，本章在冻结 ViT 骨干的条件下，在

表 5.5 使用 UperNet[184]架构在冻结 ViT 骨干情况下，ADE20K 上的语义分割性能（mIoU）。

模型	ViT	规模	ADE20K (mIoU)
SigLIP[366]	So/14	400M	35.2
SigLIP2[367]	So/16	400M	35.3
LLaVA-SigLIP[368]	So/14	400M	39.9
COMP-SigLIP[369]	So/14	400M	49.5
AIMv2[370]	L/14	300M	<u>51.5</u>
COMP-AIMv2[369]	L/14	300M	51.0
ViTP	L/16	300M	**51.9**

ADE20K 上执行语义分割评估。如表 5.5所示，ViTP（L/16）达到 51.9 mIoU，在所比较方法中取得较高结果，也优于多种基于 SigLIP 的架构。

ViTP 在端到端微调和冻结特征评估两种设置下均取得较优结果，说明基于 VLM 理解目标构造的监督，可以有效重塑 Vision Transformer 的表示空间。这一结果也从侧面表明，高层语义理解不仅服务于推理任务，也能够反向提升底层感知质量。

（二）遥感目标检测

表 5.6和表 5.7给出了光学遥感数据集上的目标检测结果。可以看到，ViTP 在 DIOR、DIOR-R、DOTA-v2.0 和 xView 上均取得了较优性能。其中，DOTA-v2.0 包含大量小目标和密集场景，是对空间解析能力要求很高的基准。ViTP 在该数据集上达到 60.23 mAP，高于 BillionFM（58.69）等比较方法，说明其在复杂空间关系建模和任意方向目标识别方面具有优势。值得注意的是，SkySense[33]所需的预训练开销远高于 ViTP，其 GPU 时长超过后者 17 倍，但 ViTP 仍取得更高结果。这表明，基于指令遵循的预训练能够以较低计算代价学习到更贴合遥感检测任务的光学表征。

表 5.8进一步展示了 ViTP 在 SAR 目标检测上的表现。相比光学图像，SAR 数据具有散斑噪声强、散射机理复杂、纹理线索弱等特点，因此对表征学习提出了更高要求。在 SARDet-100K 上，ViTP 取得 59.7 mAP，优于此前较强的 SARATR-X（57.3）及其他检测器。在 RSAR 上，ViTP 达到 72.31 mAP，同样取

表 5.6 光学遥感数据集上的目标检测性能（mAP，%）

模型	DIOR	DIOR-R	模型	DOTA-v2
GASSL[371]	67.40	65.65	RetinaNet[307]	46.68
SatMAE[30]	70.89	62.30	F-RCNN[372]	47.31
RingMo[29]	75.90	-	FCOS[19]	48.51
CACO[59]	66.91	64.10	ATSS[373]	49.57
SSL4EO[374]	64.82	61.23	SASM[140]	44.53
CMID[375]	75.11	66.37	S2ANet[21]	49.86
RVSA[376]	73.22	70.96	KLD[377]	47.26
SatLas[378]	74.10	67.59	O-RepPoints[379]	48.95
GFM[60]	72.84	67.67	RoT Trans.[380]	52.81
ScaleMAE[31]	73.81	70.20	O-RCNN[7]	53.28
MA3E[381]	-	71.82	GGHF[382]	57.17
Sel-MAE[383]	78.70	71.75	DCFL[384]	57.66
SkySense[33]	<u>78.73</u>	<u>74.27</u>	BillionFM[385]	<u>58.69</u>
ViTP	**79.80**	**75.08**	**ViTP**	**60.23**

表 5.7 xView[386]遥感数据集上的目标检测性能（mAP，%）

模型	骨干网络	xView mAP
YOLO11-x[387]	CSPNet*	8.4
RetinaNet[307]	ResNet-50	9.7
DETR[20]	swin-B	10.6
MAE[26]	ViT-L	13.3
RVSA[376]	ViT-L	15.0
SeCo[388]	ResNet-50	17.2
FCOS[19]	ResNet-50	17.2
CACO[59]	ResNet-50	17.2
GASSL[371]	ResNet-50	17.7
MTP[389]	InternImage-XL	18.2
CtxMIM[390]	swin-B	18.8
MTP[389]	ViT-L	<u>19.4</u>
ViTP	ViT-L	**26.6**

表 5.8 SAR 数据集上的目标检测性能（mAP，%）

模型	SARDet-100K	模型	RSAR
DETR[20]	31.8	Def. DETR[249]	46.62
Sparse RCNN[250]	38.1	RetinaNet[307]	57.67
Dab-DETR[251]	45.9	ARS-DETR[391]	61.14
FCOS[19]	46.5	R3Det[8]	63.94
Grid RCNN[392]	48.8	LLMRotate[393]	64.1
GFL[222]	49.8	ReDet[394]	64.71
Defor. DETR[249]	50.0	O-RCNN[7]	64.82
MSFA[317]	53.7	S2ANet[21]	66.47
DenoDet[395]	55.4	RoI-Trans.[380]	66.95
DenoDetv2[396]	56.4	SatMAE[30]	67.99
SARATR-X[397]	57.3	RemoteCLIP[32]	69.18
ViTP	**59.7**	**ViTP**	**72.31**

得了较优结果。上述结果说明，ViTP 不仅能在通用视觉与光学遥感场景中保持优势，也能够有效适应成像机理差异明显的 SAR 模态。

（三）遥感语义分割

表 5.9和表 5.10表明，ViTP 在遥感语义分割任务上同样具有稳定优势。无论是光学分割数据集 iSAID、LoveDA、UAVid，还是 SAR 模态下的 SSDD，ViTP 都取得了新的较优结果，且在 UAVid 与 SSDD 上的提升尤为明显。这说明，ViTP 通过指令遵循任务引入的高层语义监督，能够帮助 ViT 骨干更好地理解细粒度区域结构，从而为像素级分类提供更有判别力的语义表示。

（四）遥感变化检测

如表 5.11所示，ViTP 在 SVCD、WHU、LEVIR-CD 和 S2Looking 四个变化检测基准上均取得了较优性能。无论是变化类型更丰富的 SVCD，还是以建筑变化为主的 LEVIR-CD 与 WHU-CD，ViTP 都展现出稳定收益。这表明，经过视觉指令预训练后的骨干不仅对单帧语义内容具有较好的理解能力，也更能适应双时相场景中的细粒度变化建模。

表 5.9 光学遥感数据集上的语义分割性能（mIoU，%）

模型	iSAID	LoveDA	模型	UAVid
SeCo[388]	57.20	43.63	CANet[193]	63.50
DenseCLIP[398]	59.23	49.58	MP-Former[399]	63.67
SatMAE[30]	62.97	-	ABCNet[178]	63.80
CACo[59]	64.32	48.89	DecoupleNet[400]	65.80
RVSA[376]	64.49	52.44	CoaT[194]	65.80
RSSFormer[401]	65.55	52.43	UNetFormer[10]	67.80
ScaleMAE[31]	65.77	-	MaskFormer[402]	68.54
GASSL[371]	65.95	48.76	LSKNet	70.00
CMID[375]	66.21	-	Segmenter[179]	70.20
TOV[320]	66.24	49.70	RSSFormer[401]	70.69
RingMo[29]	67.20	-	DeepLabv3+[403]	71.33
SatLas[378]	68.71	-	SegFormer[183]	71.44
Sel-MAE[383]	-	53.92	DenseCLIP[398]	71.54
LSKNet	-	<u>54.00</u>	PSPNet[180]	71.71
SkySense[33]	<u>70.91</u>	-	OCRNet[404]	<u>71.84</u>
ViTP	**71.14**	**54.28**	**ViTP**	**73.39**

表 5.10 SSDD 数据集（SAR 模态）上的目标检测与实例分割性能

模型	AP_{box}	AP^{50}_{box}	AP^{75}_{box}	AP_{mask}	AP^{50}_{mask}	AP^{75}_{mask}
BoxInst[405]	44.76	83.75	44.11	34.10	71.16	27.27
Mask2Former[406]	53.40	78.45	67.02	56.52	85.10	69.48
InstaBoost[407]	54.77	87.85	58.54	58.95	89.05	71.57
CondInst[408]	57.89	92.53	67.40	50.31	90.46	54.80
SAM-Seg[99]	62.41	94.32	75.38	59.46	92.79	72.17
CATNet[409]	64.66	<u>96.46</u>	79.81	64.11	<u>96.35</u>	77.87
HQ-ISNet[410]	65.58	95.48	80.76	<u>64.75</u>	95.26	<u>81.70</u>
RSP-Query[411]	66.50	95.80	81.81	64.57	95.97	81.67
SCNet[412]	<u>67.25</u>	95.75	<u>83.38</u>	62.66	94.75	76.53
ViTP	**70.80**	**97.80**	**86.60**	**65.90**	**96.80**	**81.80**

表 5.11 SVCD、WHU、LEVIR-CD 和 S2Looking 数据集上的变化检测性能（F1 分数，%）

模型	SVCD	WHU	LEVIR	S2Looking
Scale-MAE[31]	-	-	86.60	50.20
SeCo[388]	-	-	88.40	66.00
CACo[59]	-	-	89.20	65.90
GASSL[371]	-	-	89.60	66.30
SatMAE[30]	-	-	90.00	65.00
SatMAE++[413]	-	-	90.70	56.40
CGNet[196]	-	-	92.01	64.33
Changer[45]	-	-	92.06	67.08
DiFormer[195]	-	-	92.15	66.31
Changen2[414]	-	-	92.20	69.10
SkySense[33]	-	-	92.58	-
CLNet[415]	92.10	-	90.00	-
SRCDNet[416]	92.94	87.40	-	-
GCD-DDPM[417]	94.93	92.54	90.96	-
CDContrast[418]	95.11	-	-	-
DDPM-CD[419]	95.62	92.65	90.91	-
DMNet[420]	95.93	-	-	-
SNUnet[205]	96.20	83.49	88.59	63.19
BIT[12]	-	83.98	89.31	63.76
BiFA[421]	-	94.37	90.69	-
SGSLN[422]	96.24	94.67	91.93	-
RSP[115]	96.81	-	90.93	-
SAAN[423]	97.03	-	91.41	-
SiamixFormer[424]	97.13	-	91.58	-
TransUNetCD[425]	97.17	93.59	91.11	-
RDPNet[426]	97.20	-	91.20	-
SDACD[427]	97.34			
WNet[428]	97.56	91.25	90.67	-
ChangeMamba[429]	-	92.55	90.16	-
RS-Mamba[430]	-	92.79	89.77	-
ChangeFormer[11]	-	93.04	91.11	63.39
CDMamba[431]	-	93.76	90.75	67.08
LSKNet	-	92.06	92.27	67.52
RVSA[376]	97.78	94.07	92.52	-
ChangeCLIP[432]	97.89	94.82	92.01	-
P2V-CD[433]	98.42	92.38	91.94	-
ViTP	**98.63**	**94.98**	**92.67**	**69.89**

三、ViTP 消融与鲁棒性分析

为了更全面地理解 ViTP 中各个组件与超参数的作用，本章进一步开展了一系列消融与分析实验，重点考察以下几个方面：（1）数据配方的有效性；（2）预训练步数对性能的影响；（3）VRL 中视觉词元丢弃比例的作用；（4）语言模型

表 5.12 数据配方消融实验。RSAR 上的结果体现了本章数据构建策略的作用。

预训练范式	RSAR mAP
无多样性数据	52.6
无 SAR 数据	52.5
无定位数据	53.0
无通用数据	52.3
全数据	**54.6**

规模对骨干学习的影响。

上述分析统一基于遥感领域预训练得到的 ViTP 模型，并在 RSAR 这一具有挑战性的 SAR 有向目标检测基准上进行评估。选择 RSAR 作为分析平台主要出于两点考虑：其一，SAR 模态具有区别于光学影像的成像特点，是检验模型跨领域泛化能力的理想测试场，其二，有向目标检测同时要求精确定位与细粒度识别，能够更敏感地反映骨干表征质量。除非特别说明，相关实验均在 RSAR 验证集上微调，并在测试集上报告结果。

此外，本章还从微调角度比较了 ViTP 与其他预训练范式在训练效率、数据效率以及图像退化鲁棒性方面的差异，以进一步揭示该预训练范式的优势来源。

（一）数据配方

ViTP 的预训练语料按照第 三、节中的原则构建。表 5.12展示了不同数据组成对性能的影响，其中“无多样性数据”表示移除 Million-AID、GAIA、Levir-CC、VHM、VRSBench 和 RSVG 等多样化遥感指令数据。可以看到，无论移除多样化遥感语料、SAR 专属数据、定位数据还是通用数据，RSAR 上的性能都会下降。其中，去除 SAR 数据会削弱模型对特定模态的适应能力，去除定位数据则直接影响空间定位性能，而缺失通用数据会导致模型更容易过拟合于狭窄的专业模式，并带来较为明显的退化。这说明，多样化、通用化、领域化和任务导向的数据配比对于获得较好性能缺一不可。

（二）预训练步数的影响

预训练时长是影响下游性能的重要因素。如图 5.6所示，模型在 RSAR 上的 mAP 会随着预训练步数增加而稳步提升，说明更长时间的指令驱动训练有助于骨干逐渐形成更优的表征。但当训练进行到约 8k 步后，性能增益开始趋于饱和。

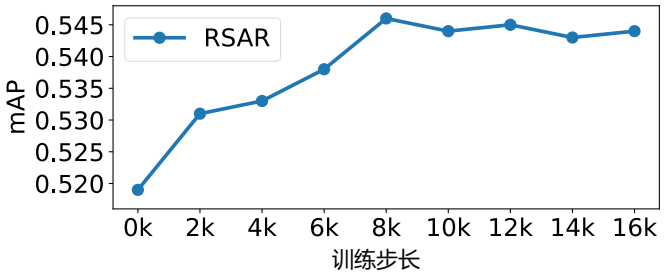


图 5.6 预训练时长的影响。RSAR mAP 随着预训练步数的增加而提高，在约 8k 步时趋于饱和。

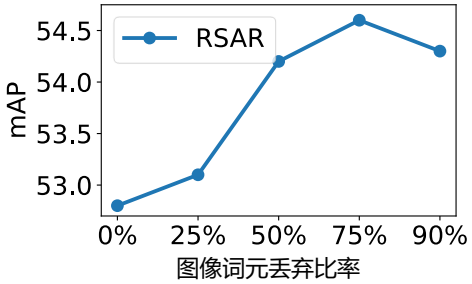


图 5.7 VRL 丢弃比例对性能的影响。ViTP 在 RSAR 上于 75% 词元丢弃率时达到峰值。

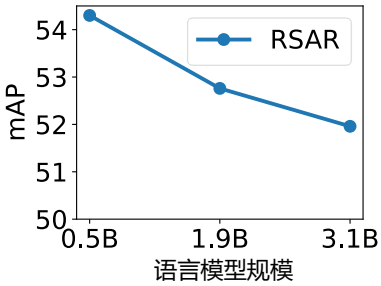


图 5.8 LLM 规模对性能的影响。过大的 LLM 会削弱 ViTP 在 RSAR 上的表现。

因此，本章将主要模型的默认预训练时长设置为 8k 步，以兼顾性能与计算成本。在 8× A40 GPU 上完成这一步骤约需 23 小时，也反映出 ViTP 较高的训练效率。

（三）VRL 词元丢弃比例的影响

视觉鲁棒学习（VRL）的关键超参数，是训练时随机丢弃视觉词元的比例。该实验进一步分析这一比例的影响。适当的词元丢弃能够迫使 ViT 在部分观测条件下恢复完整语义，从而学习到更鲁棒、更具压缩性的表征，但若丢弃比例过高，则可能遮蔽掉维持基本视觉理解所必需的信息，进而使训练目标过于困难。

如图 5.7所示，不同的词元丢弃比例会影响 RSAR 上的检测性能。当丢弃比例为 75% 时，模型取得较优结果，mAP 由不使用 VRL 时的 52.8 提升至 54.6，说明这一设置能够在信息保留与正则化强度之间取得较好平衡。较低的丢弃比例（0% 至 50%）提供的约束仍然偏弱，而即使在 90% 的极高丢弃率下，模型仍可达到 54.3 mAP，表明强正则化对该范式依然具有积极作用。

表 5.13 预训练效率与性能对比。时间估计基于 8× A40 GPU 上的预训练过程。ViTP 以较低计算成本取得了具有竞争力的性能。

	方法	小时	DIOR-R	iSAID
掩码图像重建	RVSA[376]	250	70.96	64.49
	Scale-MAE[31]	60	70.20	65.77
对比学习	RemoteCLIP[32]	100	70.20	62.53
	SkySense[33]	400	74.27	70.91
	ViTP	**23**	**75.08**	**71.14**

（四）语言模型规模的影响

该实验还考察了 LLM 规模对 ViTP 性能的影响。如图 5.8所示，随着 LLM 参数规模增大，RSAR 上的下游性能反而持续下降。这一现象与 Masked Autoencoders[26]中“轻量化解码器更有利于骨干学习”的观察具有一致性。一个可能的解释是：当语言模型过强时，它会在一定程度上“补偿”视觉骨干表征不足，从而减弱 ViT 必须主动学习高质量模态特征的压力。对于 SAR 这类成像机理复杂、纹理线索有限的模态而言，这种补偿效应尤其可能抑制视觉骨干学习到真正有效的专业表征。

（五）训练效率

ViTP 的一个重要特点在于较高的预训练效率，这降低了专业领域基础模型开发的门槛。如表 5.13所示，ViTP 在 8 张 A40 GPU 上可在 1 天内完成预训练，同时取得较优下游性能。具体而言，ViTP 相较于较高效的 MIM 方法 Scale-MAE 快约 **2.6 倍**，相较于计算开销更大的对比学习方法 SkySense 则快超过 **17 倍**。

（六）数据效率

该实验通过在 RSAR 训练集的不同比例子集（2% 至 20%）上微调 ViTP，评估其低数据条件下的表现。为充分释放各方法的建模潜力，该实验中的微调训练统一延长至 36 个 epoch。如表 5.14所示，ViTP 在不同数据比例下均高于 SatMAE[30]和 RemoteCLIP[32]。随着可用训练数据减少，ViTP 的相对收益更加明显。例如，在仅使用 2% 训练数据时，ViTP 仍可达到 46.98 mAP，而 SatMAE 和 RemoteCLIP 分别为 37.90 和 34.78 mAP。更值得注意的是，仅用 20% 数据微调

表 5.14 RSAR 基准上的数据效率。与掩码图像重建和对比学习相比，ViTP 在低数据下取得了更高的比较结果。

模型	100%	20%	10%	5%	2%
RemoteCLIP[32]	69.18	63.14 ↓6.04	57.10 ↓12.08	47.25 ↓21.93	34.78 ↓34.40
SatMAE[30]	67.99	61.36 ↓6.63	55.24 ↓12.75	50.76 ↓17.23	37.90 ↓30.09
ViTP	**72.31**	**67.07 ↓5.24**	**61.68 ↓10.63**	**56.42 ↓15.89**	**46.98 ↓25.33**

的 ViTP（67.07 mAP）就已达到或超过许多全数据训练方法的水平。这表明，指令驱动预训练带来的高层语义先验，有助于模型在有限样本条件下实现更有效的泛化。

（七）模型鲁棒性

尽管多数研究基准中的图像经过了较为理想化的整理，但真实遥感影像往往会受到云雾遮挡、传感器噪声和压缩伪影等因素影响。为评估 ViTP 对这类退化的鲁棒性，本章在 REOBench[434]的 DIOR-R 子基准上，对 12 种常见图像损坏进行了系统测试。为与 REOBench 基线设置保持一致，本章在 DIOR-R 上重新训练 ViTP，输入尺寸设为 800×800，且不使用测试时增强。如表 5.15所示，标准 ViTP 在多种损坏类型下都表现出优于 MIM 和对比学习基线的鲁棒性。其平均 mAP 达到 69.00，且相对于干净图像的性能下降（Δ_{TP}）较小，为 4.37。进一步比较“无 VRL”的 ViTP 可以发现，即便不使用 VRL，ViTP 也已具备较好的鲁棒性，加入 VRL 后，平均性能由 67.13 进一步提升到 69.00 mAP，同时 Δ_{TP} 由 4.78 下降至 4.37。这说明，视觉指令预训练与 VRL 的结合，有助于模型学习更稳健的视觉表征。

（八）自注意力图可视化

为定性分析 ViTP 骨干学习到的表征，本章进一步可视化了模型内部的自注意力分布，并将其与 CLIP[28]、DINOv2[27]预训练得到的骨干进行比较。

如图 5.9所示，本章选取一个查询块（红色叉号），并可视化其对应的自注意力激活区域，以考察模型在处理局部信息时关注的空间范围。实验中也可以

表 5.15 常见图像损坏下的模型鲁棒性对比。相较于掩码图像重建和对比学习基线，ViTP 在 REOBench 基准（DIOR-R，mAP）上表现出更好的性能保持能力：其在所比较损坏类型上取得较高平均性能，且相对于干净基线的性能下降（Δ_{TP}）较小。视觉鲁棒学习（VRL）的引入进一步增强了这种鲁棒性。

	模型	净图	亮度对比	云遮	压缩伪影	图像缺块	高斯模糊	高斯噪声	雾	动态模糊	旋转	椒盐噪声	尺度	位移	**平均**	$\Delta_{\text{TP}}\downarrow$
MIM	SatMAE[30]	62.30	56.84	57.86	55.80	58.36	55.38	58.44	59.34	56.92	56.60	53.76	51.58	60.90	56.82	5.49
	ScaleMAE[31]	70.20	64.80	65.98	62.50	64.46	62.58	63.82	66.10	63.08	63.44	60.50	53.08	68.26	63.22	6.98
	RVSA[376]	70.96	60.59	65.02	61.58	64.60	62.35	62.87	63.98	62.88	64.04	56.61	55.97	69.69	62.51	8.45
	SatMAE++[413]	65.20	59.44	61.02	60.30	59.88	59.66	61.06	61.72	59.56	59.14	58.64	48.48	64.70	59.47	5.73
CL	RemoteCLIP[32]	70.20	66.52	66.62	63.84	65.40	63.62	63.68	66.76	62.66	63.52	59.16	57.42	68.64	63.99	6.21
	GeoRSCLIP[318]	69.80	66.12	65.34	65.34	64.96	63.62	62.90	66.04	62.02	62.68	56.04	57.40	68.10	63.38	6.42
	ViTP 无 VRL	71.91	69.11	67.25	66.20	66.87	67.67	67.47	67.12	65.94	65.54	66.87	65.23	70.34	67.13	4.78
	ViTP	**73.37**	**70.56**	**69.54**	**67.74**	**70.58**	**69.05**	**68.97**	**70.85**	**67.86**	**67.10**	**67.23**	**66.64**	**71.87**	**69.00**	**4.37**

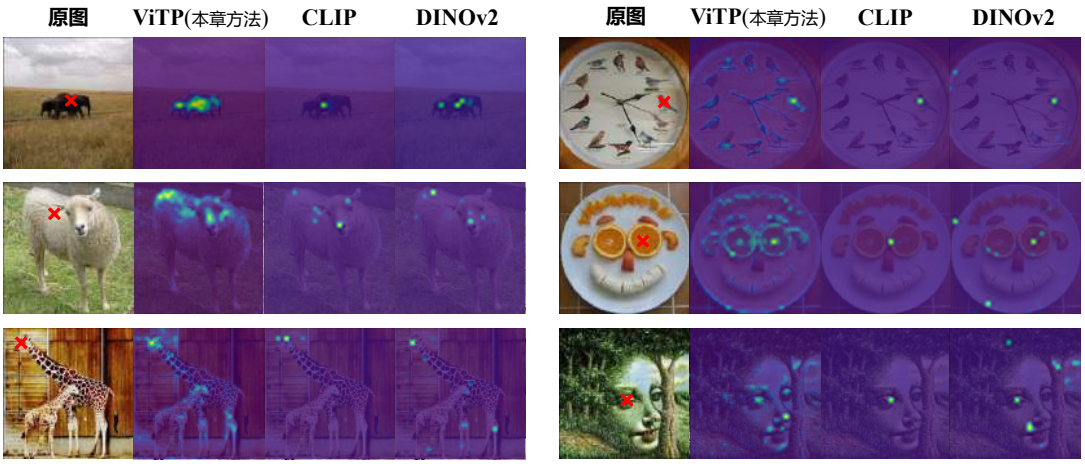


图 5.9 给定查询块（红色叉号）时第三层 ViT 自注意力激活图的比较。三种模型分别采用 ViTP（本章方法）、CLIP 和 DINOv2 进行预训练。左侧为一般自然图像，右侧为需要更高层次语义理解的抽象概念图像。

观察到已有工作提到的“伪影注意力”现象[435-436]，即某些层会产生较为弥散的注意力分布。综合比较后，三种方法在第三层给出的注意力图相对清晰、可解释性较好，因此本部分仅展示该层结果。

从大象、绵羊、长颈鹿等示例可以看到，ViTP 会将注意力扩展到图像中语义相关但空间上并不相邻的区域，这一现象在抽象概念图像上更为明显。总体来看，这些结果说明，ViTP 增强了 ViT 骨干的高层语义组织能力，使其不再仅依赖局部纹理，而能够按照概念和语义关联整合视觉信息。需要说明的是，该可视化主要采用较浅层注意力图，目的是获得相对稳定的局部语义观察，不能替代更严格的可解释性因果分析。

上述结果说明，语言指令监督可以在单一视觉骨干层面带来较稳定的迁移收益。接下来，实验进一步转向异构多模态场景，考察当输入来自不同成像机制且缺少严格空间配对时，语言是否还能作为共享语义枢纽，使跨模态对齐在预训练阶段提前发生。

四、BabelRS 实验设置与数据集

在 ViTP 实验验证语言指令预训练对视觉骨干的塑造作用之后，下文进一步考察语言引导早期对齐在异构多模态检测中的作用。实验重点包括三方面：首先比较 BabelRS 与微调阶段对齐方法、通用预训练方法和遥感预训练方法的检测结果，其次分析训练损失、梯度范数和自动混合精度条件下的优化稳定性，最

后通过中间层融合策略和退火调度消融，验证 LVSA 设计的必要性。

（一）预训练数据集

预训练语料的组成如表 5.16所示。BabelRS 汇集了多种大规模遥感视觉语言数据集，以确保在模态和任务两个维度上都具备较广的语义覆盖。Million-AID[338]、LevirCC[339]、VHM[340]、RSVQA[341]和 FIT_RS[342]主要提供场景理解、目标识别和属性推理相关的指令数据，GAIA[337]则进一步通过气象与多光谱图像扩展了语义覆盖范围。为增强 SAR 模态的语义对齐能力，本章引入 SARLang[437]这一面向 SAR 场景的大规模视觉问答数据集，重点补充位置感知查询和细粒度目标语义。对于红外模态，本章使用 MMRS-1M[438]中的红外样本，以引入热特征相关监督。GeoChat[345]、DIOR-RSVG[346]和 VRSBench[348]则提供视觉定位标注，用于加强语言与空间区域、目标范围之间的对齐。所有数据集在进入训练前都进行了统一预处理和概念标准化，以保证数据格式一致、概念命名统一（例如“桥”“港口”“船舶”等）。此外，Mini-InternVL[349]以较低采样率被保留在语料中，用于维持通用视觉语言理解能力，并避免模型过度偏向专业遥感概念。总体而言，这一语料设计使 BabelRS 能够在无需空间配对样本的前提下学习到兼具模态无关性和任务相关性的表示。

（二）微调数据集

本章在第四章中提出的 SOI-Det 基准上评估异构多模态目标检测，该基准结合了来自三种异构传感模态的数据集：SAR、光学和红外。该基准包括用于 SAR 图像的 SARDet-100K、用于光学航空图像的 DOTA-v1.0[279] 以及用于红外车辆检测的 DroneVehicle[282]。这些数据集共同涵盖了多样化的目标类别、成像分辨率和标注格式，包括水平边界框和定向边界框。

（三）实现细节

BabelRS 采用工程化视觉语言模型框架进行语言引导预训练。考虑到从头训练 VLM 通常需要多阶段流程和大量计算资源，BabelRS 直接从 InternVL-2.5 1B[35]初始化，其中视觉端采用 InternViT 变体，语言端采用 Qwen2[439]。对于 LVSA 中的特征集合 $\mathcal{V}$，BabelRS 选取 ViT-Large 的第 3、9、18 层及最后一层参与融合，这一设置参考了本章 ViTP 部分和 Perception Encoder[440]中的经验。

所有预训练实验均在 8 张 NVIDIA A40（48 GB）GPU 上进行，全局批量大

表 5.16 BabelRS 所用语言引导预训练数据集构成，包括数据规模、采样率与任务类型（VQA：视觉问答；VG：视觉定位；CLS：分类）。

数据集	规模	采样率	任务
Mini-InternVL[349]	1394k	0.01	VQA
RSVQA[341]	100k	1	VQA
FIT_RS[342]	100k	0.2	VQA
GeoChat[345]	64k	1	VG
VRSBench[348]	38k	1	VG
DIOR-RSVG[346]	27k	1	VG
VHM[340]	223k	1	VQA
LevirCC[339]	50k	0.2	Caption
GAIA[337]	33k	1	Caption
Million-AID[338]	920k	0.03	Caption,CLS
MMRS-1M[438]	52k	1	VQA
SARLang[437]	1126k	0.6	VQA

小为 128，预训练学习率设为 2e-5。微调阶段沿用相同硬件配置和统一训练流程。标准的数据增强、采样策略与归一化协议遵循第四章中的设定，优化器采用 AdamW，学习率为 5e-5，权重衰减为 0.05，每张 GPU 的批量大小为 4。评估方面，本章报告 IoU = 0.5 时的 AP（AP@50）、IoU 从 0.5 到 0.95 平均得到的 mAP，以及本章提出的 H-mAP，用于衡量跨模态性能的均衡性。

五、BabelRS 下游迁移结果

表 5.17给出了 SOI-Det 基准上的异构多模态目标检测结果。需要强调的是，对比方法大多把优化重点放在微调阶段，即通过额外的对齐项或正则项缓解跨模态差异，而 BabelRS 则把主要改进放在预训练阶段，并在微调时仅采用简单联合训练。从结果看，BabelRS 在主要评价指标上取得了较优性能，且提升并非局限于某一单一模态。尤其是在 SAR 和红外数据集上，BabelRS 相较于晚期对齐方法表现出更明显的优势，这说明预训练阶段的早期语义对齐对弱模态和难模态具有更直接的帮助。

总体而言，这些结果表明 BabelRS 学习到了一种更均衡、更具跨模态一致性的表示。较高的 H-mAP 进一步说明，其改进不是由某一主导模态“拉高”整体得分，而是源于不同模态上更加协调的性能提升。

表 5.18进一步比较了 BabelRS 与不同预训练策略在 SOI-Det 上的表现。为

表 5.17 SOI-Det 基准上的异构多模态目标检测性能。对比方法大多侧重微调阶段优化，而 BabelRS 聚焦于预训练阶段优化，并采用简洁的联合微调策略。BabelRS 在整体上取得较优性能。

模型	测试集	AP@50	**mAP**	**H-mAP**
微调阶段优化				
简单联合训练[53]	Overall	77.56	47.05	
	SARDet-100K	84.11	53.46	47.57
	DOTA	76.37	45.18	
	DroneVehicle	73.28	44.99	
DA[52]	Overall	79.76	48.37	
	SARDet-100K	84.93	53.86	49.23
	DOTA	78.47	46.23	
	DroneVehicle	77.43	48.21	
UniDet[54]	Overall	79.55	48.47	
	SARDet-100K	84.70	53.81	49.24
	DOTA	78.28	46.49	
	DroneVehicle	77.17	47.99	
Uncertainty loss[56]	Overall	79.99	48.79	
	SARDet-100K	84.81	53.43	49.57
	DOTA	78.73	46.94	
	DroneVehicle	77.96	48.78	
SM3Det	Overall	80.68	50.20	
	SARDet-100K	89.94	60.64	51.31
	DOTA	77.88	46.47	
	DroneVehicle	77.99	48.87	
预训练阶段优化				
BabelRS (本章方法)	Overall	**81.32**	**51.57**	
	SARDet-100K	91.70	63.30	**53.02**
	DOTA	77.73	46.96	
	DroneVehicle	79.63	51.32	

保证公平性，所有方法均采用 ViT-Large 骨干，并使用相同的简单联合训练协议进行微调。可以看到，CLIP[28]、RemoteCLIP[32]等图文对齐方法在密集检测任务上表现较弱，说明仅依赖最后一层全局语义对齐，难以满足目标检测所需的空间解析能力。通用自监督方法，如 MAE[26]、BEiT[364]、BEiTv2[441]和 DINOv2[27]，在面对 SAR 和红外等异构模态时也存在明显泛化不足。即便是 SatMAE[30]、

表 5.18 不同预训练策略在 SOI-Det 上的比较。所有模型均使用 ViT-Large 骨干和相同的微调协议。BabelRS 在主要指标上取得较优性能。

方法	测试集	AP@50	**mAP**	H-mAP
晚期对齐 *(通用骨干)*				
CLIP [28]	Overall	65.21	36.12	
	SARDet-100K	72.70	42.00	37.12
	DOTA	62.67	33.56	
	DroneVehicle	63.83	36.74	
MAE [26]	Overall	72.36	42.84	
	SARDet-100K	70.64	39.48	42.54
	DOTA	73.35	43.43	
	DroneVehicle	71.46	45.11	
BEiT [364]	Overall	76.50	44.59	
	SARDet-100K	81.90	50.10	44.94
	DOTA	76.39	43.14	
	DroneVehicle	70.34	42.35	
BEiTv2 [441]	Overall	77.63	44.67	
	SARDet-100K	78.50	46.70	45.35
	DOTA	78.01	43.36	
	DroneVehicle	75.48	46.14	
DINOv2 [27]	Overall	74.79	45.02	
	SARDet-100K	72.74	41.38	44.34
	DOTA	76.53	46.23	
	DroneVehicle	72.04	45.74	
晚期对齐 *(遥感骨干)*				
RemoteCLIP [32]	Overall	66.37	37.17	
	SARDet-100K	73.90	43.30	38.30
	DOTA	63.57	34.36	
	DroneVehicle	65.74	38.27	
SatMAE [31]	Overall	74.49	42.51	
	SARDet-100K	79.70	47.90	43.03
	DOTA	73.86	40.84	
	DroneVehicle	70.12	41.07	
ScaleMAE [31]	Overall	74.09	42.52	
	SARDet-100K	78.80	46.50	43.15
	DOTA	73.22	41.01	
	DroneVehicle	71.07	42.30	
早期对齐				
BabelRS (本章方法)	Overall	**81.32**	**51.57**	
	SARDet-100K	91.70	63.30	**53.02**
	DOTA	77.73	46.96	
	DroneVehicle	79.63	51.32	

ScaleMAE[31]等面向遥感场景的预训练方法，也没有显式解决异构模态之间的语义对齐问题，因此模态对齐只能在微调阶段隐式发生，导致优化过程更加复杂。

相比之下，BabelRS 在主要指标上整体高于这些晚期对齐方法。该结果说

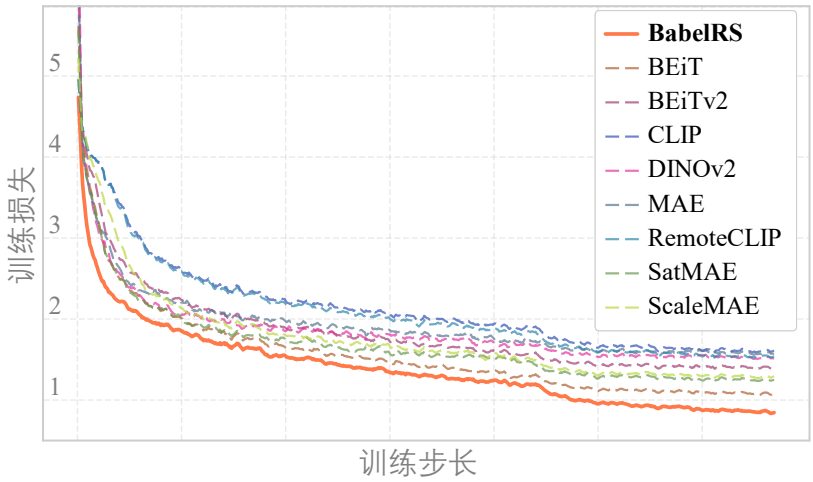


图 5.10 相同微调协议下的训练损失曲线。晚期对齐方法收敛缓慢且波动明显，而 BabelRS 从更低的初始损失出发并保持平滑收敛。

表 5.19 SOI-Det 上自动混合精度训练下的检测性能。多种晚期对齐方法出现明显数值不稳定，而 BabelRS 保持稳定并取得更优结果。

预训练方法	AP@50	mAP	H-mAP
MAE[26]	NaN	NaN	NaN
BEiTv2[441]	NaN	NaN	NaN
DINOv2[27]	NaN	NaN	NaN
ScaleMAE[31]	NaN	NaN	NaN
CLIP[28]	64.58	35.62	36.68
RemoteCLIP[32]	65.97	36.82	37.72
SatMAE[30]	70.49	39.45	40.00
BEiT[364]	75.74	43.82	44.35
BabelRS (本章方法)	**79.13**	**50.17**	**51.52**

明，语言引导的早期语义对齐是可行的，并且对于异构多模态遥感检测具有较好的适配性。

六、BabelRS 消融与鲁棒性分析

（一）优化稳定性和训练动态

除准确率外，下文还重点分析优化稳定性，因为这正是提出 BabelRS 的核心动机之一。图 5.10比较了在相同优化设置下的微调损失曲线。可以看到，晚期对齐方法普遍存在收敛缓慢、波动较大甚至发散的问题，相比之下，BabelRS 从更低的初始损失出发，并在整个训练过程中保持平滑稳定的收敛过程。这组

结果支持了本章的核心观点：将模态对齐前移到预训练阶段，有助于把跨模态对齐与下游任务优化解耦，从而减轻训练过程中的梯度干扰。

为进一步分析异构多模态检测器的优化稳定性，本章在自动混合精度（AMP）设置下观察训练动态。由于 AMP 会在前向和反向传播中降低数值精度，因此可以视作检验优化鲁棒性的“压力测试”。在这一训练条件下，BabelRS 稳定性优势更加清晰。如表 5.19所示，MAE、BEiTv2、DINOv2 和 ScaleMAE 等晚期对齐方法在 AMP 下出现严重数值不稳定，最终无法收敛，即便勉强可训练的方法，也普遍伴随性能退化。

相比之下，BabelRS 在 AMP 训练下依然保持稳定，并在主要指标上取得较优结果。考虑到 AMP 是大规模训练中降低显存开销、提升吞吐量的常用设置，这一稳定性对于实际部署具有重要意义。

图 5.11对比了晚期对齐基线与 BabelRS 在相同 AMP 配置下的训练损失和梯度范数轨迹，进一步揭示了这一差异。可以看到，DINOv2、MAE、BEiTv2 和 ScaleMAE 等方法存在明显的梯度峰值，并伴随剧烈的损失波动，部分实验最终以 NaN 发散。这种不稳定性，本质上来自模态对齐目标与检测目标在微调阶段的强耦合，它们在异构模态联合训练中容易引发严重梯度冲突。

与之相对，BabelRS 在整个训练过程中都保持了平滑的损失曲线和受控的梯度范数。由于模态语义统一已经在预训练阶段完成，微调时模型面对的是更加一致的特征分布，因此即便在较低数值精度下也仍然保持良好鲁棒性。

这些结果表明，BabelRS 在 AMP 下的稳定性提升并非偶然现象，而是早期语义对齐机制带来的直接收益。

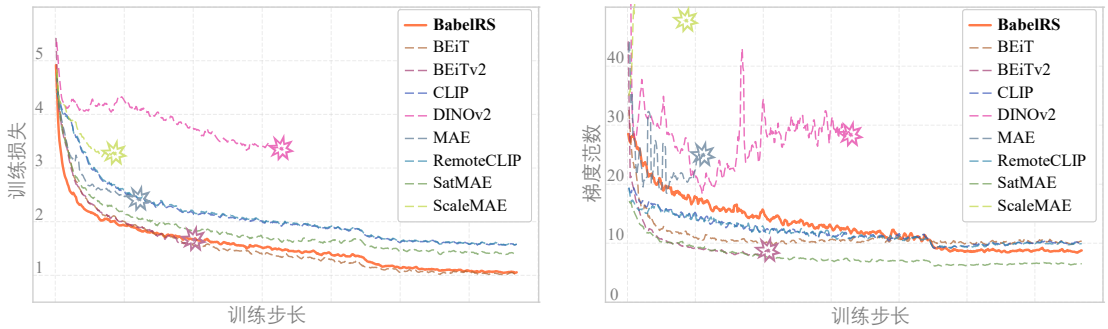


图 5.11　AMP 训练下的损失与梯度范数轨迹。晚期对齐方法出现尖锐的梯度范数峰值和不稳定的损失波动，而 BabelRS 保持受控的梯度变化与平滑收敛，体现出更好的数值稳定性。

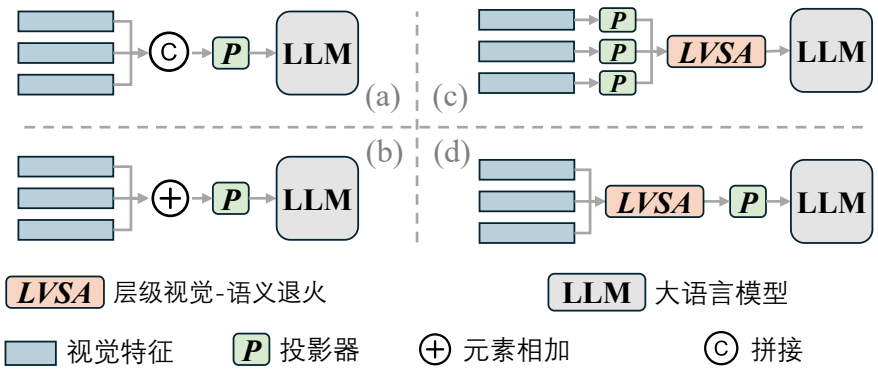


图 5.12 中间层特征合并策略对比：(a) 特征拼接，(b) 逐元素求和，(c) 带 LVSA 的逐层投影器，以及 (d) BabelRS 提出的基于 LVSA 的合并与共享投影器。

表 5.20 图 5.12所示不同中间层特征合并策略的检测性能。

配置	mAP	H-mAP
基准	49.33	50.67
(a)	50.25	51.60
(b)	50.31	51.55
(c)	49.88	50.92
(d) 本章方法	51.57	53.02

（二）与其他合并策略的比较

图 5.12与表 5.20比较了不同的中间层 ViT 特征融合策略。基线方法遵循原始 InternVL 设计，仅将最后一层特征送入投影器。实验表明，简单的中间层特征拼接（配置 (a)）或逐元素相加（配置 (b)）只能带来有限收益，而为每一层单独设置投影器（配置 (c)）则会引入额外复杂度，并削弱训练稳定性。相比之下，BabelRS 提出的基于 LVSA 的融合策略（配置 (d)）通过渐进式整合多尺度特征，并配合共享投影器，在主要指标上取得较优结果。这说明，受控而平滑的多尺度引入方式，比一次性堆叠中间层特征更适合语言引导预训练。

（三）预训练步数和退火调度

图 5.13分析了预训练时长和 LVSA 退火调度的影响。左图显示，检测性能在约 6k 步附近出现短暂回落，本章认为这主要是因为中间层特征在尚未充分适配之前被过早引入。随着训练继续进行，mAP 和 H-mAP 均稳步上升，并在约 20k 步附近趋于饱和，继续延长训练只带来轻微波动，说明此时已接近过拟合边界。

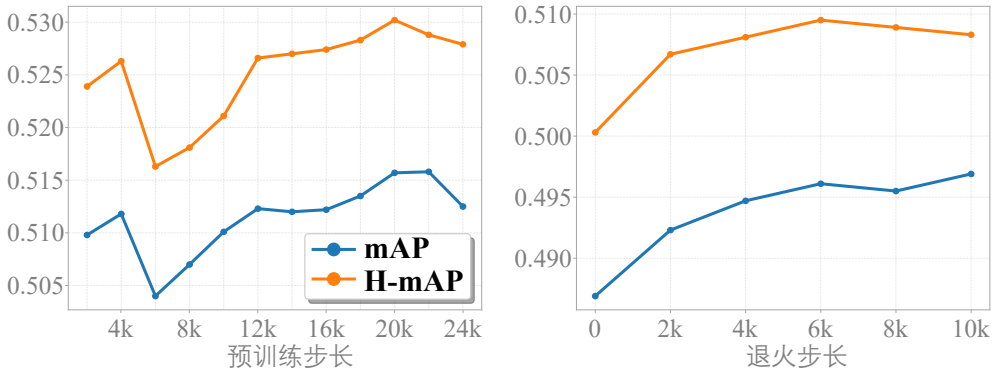


图 5.13 预训练时长（左）与 LVSA 退火调度（右）对 SOI-Det 性能的影响。

右图进一步给出了退火参数 τ 的影响。为提高实验效率，相关模型均在 AMP 设置下进行微调。可以看到，随着 τ 从 0 增大到 6k 步，mAP 和 H-mAP 持续提升，说明更平滑的多尺度特征引入过程有利于跨模态表示学习。当 $\tau = 6\text{k}$ 时性能达到峰值，继续增大则收益有限。

第六节 本章小结

本章围绕语言和高层语义如何服务于遥感基础模型预训练展开研究，并将 ViTP 与 BabelRS 组织为一个递进整体。ViTP 面向视觉骨干本身，利用视觉语言模型的指令遵循目标反向塑造遥感视觉 Transformer，使预训练目标从底层视觉重建或图像级对齐进一步走向面向下游感知任务的语义引导，BabelRS 面向异构多模态条件，将语言作为跨模态语义枢纽，把 RGB、SAR 和红外等模态的对齐过程前移到预训练阶段，从而缓解晚期对齐中常见的优化耦合与数值不稳定问题。二者共同回应了绪论中“范式层”挑战：遥感基础模型不仅需要有效骨干、数据迁移和统一架构，也需要能够利用语言语义进行领域化塑造和跨模态组织的预训练机制。

ViTP 的数据配方虽然体现了多样化、定位、SAR 和通用数据的重要性，但受算力限制，尚未系统分析更大规模数据、不同数据难度和不同领域组成对性能的影响，BabelRS 则主要在 SOI-Det 设置下检验了语言早期对齐的稳定性，未来仍需在更多传感器、更多任务形式和更大规模预训练语料上进一步检验其泛化边界。

第六章 总结与展望

第一节 工作总结

本文围绕“复杂遥感场景下多模态感知基础模型”这一核心问题展开研究，形成一条沿着感知层、架构层与范式层逐步推进的技术路线：先解决基础表征如何适应复杂空间组织与有限监督条件，再解决异构模态与异构任务如何在统一框架下协同建模，最后进一步回答高层语义如何在预训练阶段主导专业视觉学习，并支撑稳定的跨模态对齐。围绕这三层彼此衔接的挑战，本文形成了如下系统性研究结论。

在**感知层**上，本文主要聚焦复杂空间先验建模、SAR 数据基础设施建设以及跨域迁移学习三个相互耦合的问题。针对遥感场景中目标识别高度依赖广域上下文、传统骨干网络难以灵活适配空间先验的问题，本文提出了先验引导的轻量化选择性大核网络 LSKNet。该方法通过大核分解与动态空间选择机制，在可控计算开销下实现了广域上下文与局部判别信息的协同建模，从而提升了骨干对复杂空间组织的适应能力。进一步地，面向 SAR 目标检测长期存在的数据规模不足、类别覆盖有限和评测标准不统一等瓶颈，本文构建了大规模 SAR 目标检测基准 SARDet-100K，并提出多阶段滤波增强预训练框架 MSFA。前者为 SAR 目标检测研究提供了标准化、可扩展的数据底座，后者则通过手工稳健特征、光学遥感桥接预训练与检测器级迁移相结合的方式，有效缓解了自然图像向 SAR 图像迁移中的双重域鸿沟。整体而言，这部分工作从骨干建模、数据基准到迁移机制三个方面共同回应了感知层挑战，为复杂遥感场景下的基础表征学习奠定了方法与数据基础。

在**架构层**上，本文重点回应异构多模态数据“烟囱式”孤立和任务碎片化的问题。真实遥感应用往往同时涉及 RGB、SAR、红外等不同传感器，以及水平框、旋转框等不同检测形式，若仍延续单模态、单任务分别建模的方案，将难以支撑基础模型所需的共享表示学习与高效部署。为此，本文提出了基于混合专家系统的异构多模态检测统一架构 SM3Det。该框架通过网格级稀疏混合专家、动态路由机制以及一致性同步优化，在保持模态差异和任务差异的同时实

现跨模态知识共享与协同学习，从而将原本割裂的多模态检测问题统一到同一建模框架中。该工作表明，复杂遥感场景中的统一建模并不意味着抹平模态差异，而是在共享表示与专门能力之间建立可调和的结构关系，这为后续更高层次的多模态基础模型研究提供了架构支撑。

在**范式层**上，本文进一步针对传统“自下而上”预训练范式的局限，以及异构多模态学习中后期对齐易引发优化冲突的问题，探索语言和高层语义在专业视觉学习中的主导作用。本文将视觉指令预训练 ViTP 与语言引导跨模态早期对齐 BabelRS 统一纳入“语言引导的遥感基础模型预训练与跨模态对齐”这一研究主题：前者将视觉 Transformer 骨干嵌入视觉语言模型中，通过指令遵循目标实现高层语义对底层感知的反向塑造，并借助视觉鲁棒学习增强模型在低样本、分布偏移和图像退化条件下的稳定性与泛化能力，后者以共享语言语义空间为锚点，通过概念共享指令对齐与层级视觉语义退火，将跨模态语义统一前移到预训练阶段，在不依赖严格空间配对样本的条件下实现更稳定的跨模态早期对齐。该部分从“语言监督塑造专业视觉骨干”推进到“语言锚点连接异构传感器语义”，说明面向复杂遥感场景的基础模型不仅需要统一架构，更需要与专业任务需求相匹配的预训练机制。

综上，本文围绕感知层、架构层与范式层三个层面，系统回答了复杂遥感场景多模态感知基础模型构建中的关键问题，并形成了一条由基础表征学习、专业数据基准建设、统一多模态建模到语言引导预训练优化逐步递进的技术路线。相关方法在多个公开基准上取得了较优结果，表明本文所建立的研究框架在理论上具有较强的一致性与层次性，在方法上具有较好的可迁移性与可扩展性，也为后续通用遥感基础模型的发展提供了可复用的研究范式。

第二节 研究展望

尽管本文已经围绕复杂遥感场景的多模态感知基础模型进行了较为系统的探索，但面向真实开放环境的遥感智能系统，仍有若干值得持续深入的研究方向。

1. 面向低资源场景的遥感数据基础设施仍需持续完善。当前 SAR、多光谱、红外等模态的数据资源仍不均衡，尤其在大规模、多类别、高质量标注方面依然存在明显短板。未来可进一步探索半监督、弱监督与无监督学习在 SAR 检测中的应用，充分利用海量未标注遥感数据，同时可加强多光谱等关键模态的大

规模检测基准建设，为更广义的多模态遥感基础模型提供统一的数据底座。

2. 跨域迁移与统一预训练需要从“经验式桥接”走向“自动化生成”。本文在 MSFA 与 ViTP 中观察到，合理的中间域桥接与高质量指令数据对模型性能至关重要，但这些过程仍较依赖人工经验与领域知识。未来可进一步研究基于大语言模型或多模态大模型的自动数据构建机制，自动生成高质量的问答、描述、定位和跨模态语义监督，以提升遥感领域预训练的可扩展性与可复用性。

3. 统一多模态建模仍需覆盖更丰富的传感器与任务形态。本文重点关注了 RGB、SAR 和红外图像，但真实遥感应用还广泛涉及多光谱、高光谱、视频序列、三维点云以及多航过时序观测。未来可探索将统一建模框架拓展到更多传感器类型，并进一步覆盖语义分割、实例分割、变化检测、视频目标检测及时空联合理解等任务，推动“统一检测器”向“统一遥感感知模型”演进。

综上所述，复杂遥感场景的多模态感知基础模型研究仍处于快速发展阶段。随着数据基础设施、语言监督机制、统一架构设计与大模型技术的持续进步，构建具备跨模态统一感知、跨任务迁移与跨场景泛化能力的通用遥感基础模型，将成为未来遥感智能研究的重要方向。

参考文献

[1] PENG B, PENG B, ZHOU J, et al. Scattering model guided adversarial examples for SAR target recognition: Attack and defense[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-17. DOI: 10.1109/tgrs.2022.3213305.

[2] BRAUN A. Radar satellite imagery for humanitarian response. Bridging the gap between technology and application[D]. Universität Tübingen, 2019.

[3] WEGMULLER U, WIESMANN A, STROZZI T, et al. ENVISAT ASAR in disaster management and humanitarian relief[C]. IEEE International Geoscience and Remote Sensing Symposium. 2002: 2282-2284. DOI: 10.1109/igarss.2002.1026519.

[4] FROLIND P O, GUSTAVSSON A, LUNDBERG M, et al. Circular-aperture VHF-band synthetic aperture radar for detection of vehicles in forest concealment[J]. IEEE Transactions on Geoscience and Remote Sensing, 2012, 50(4): 1329-1339. DOI: 10.1109/tgrs.2011.2166081.

[5] RAMADAN T M, ONSI H M. Use of ERS-2 SAR and Landsat TM images for geological mapping and mineral exploration of Sol Hamid Area, South Eastern Desert, Egypt[J]. Egyptian Journal of Remote Sensing and Space Sciences, 2003.

[6] IVANOV A Y, GERIVANI H, EVTUSHENKO N V. Characterization of natural hydrocarbon seepage in the South Caspian Sea off Iran using satellite SAR and geological data[J]. Marine Georesources & Geotechnology, 2020, 38(5): 527-538. DOI: 10.1080/1064119x.2019.1600175.

[7] XIE X, CHENG G, WANG J, et al. Oriented R-CNN for object detection[C]. ICCV. 2021: 3500-3509. DOI: 10.1109/iccv48922.2021.00350.

[8] YANG X, YAN J, FENG Z, et al. R3det: Refined single-stage detector with feature refinement for rotating object[C]. AAAI. 2021: 3163-3171. DOI: 10.1609/aaai.v35i4.16426.

[9] DAI Y, ZOU M, LI Y, et al. DenoDet: Attention as Deformable Multi-Subspace Feature Denoising for Target Detection in SAR Images[J]. IEEE Transactions on Aerospace and Electronic Systems, 2025, 61(2): 4729-4743. DOI: 10.1109/taes.2024.3507786.

[10] WANG L, LI R, ZHANG C, et al. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 190: 196-214. DOI: 10.1016/j.isprsjprs.2022.06.008.

[11] BANDARA W G C, PATEL V M. A transformer-based siamese network for change detection[C]. IEEE International Geoscience and Remote Sensing Symposium. 2022: 207-210. DOI: 10.1109/igarss46834.2022.9883686.

[12] CHEN H, QI Z, SHI Z. Remote sensing image change detection with transformers[J]. IEEE

Transactions on Geoscience and Remote Sensing, 2022, 60: 1-14. DOI: 10.1109/tgrs.2021.3095166.

[13] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]. CVPR. 2005: 886-893. DOI: 10.1109/cvpr.2005.177.

[14] CANNY J. A computational approach to edge detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, PAMI-8(6): 679-698. DOI: 10.1109/tpami.1986.4767851.

[15] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features [C]. CVPR. 2001: I-511-I-518. DOI: 10.1109/cvpr.2001.990517.

[16] MALLAT S. Group invariant scattering[J]. Communications on Pure and Applied Mathematics, 2012, 65(10): 1331-1398. DOI: 10.1002/cpa.21413.

[17] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[C]. NeurIPS. 2017: 1137-1149. DOI: 10.1109/tpami.2016.2577031.

[18] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal Loss for Dense Object Detection[C]. ICCV. 2017: 2999-3007. DOI: 10.1109/iccv.2017.324.

[19] TIAN Z, SHEN C, CHEN H, et al. Fcos: Fully convolutional one-stage object detection[C]. ICCV. 2019: 9626-9635. DOI: 10.1109/iccv.2019.00972.

[20] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]. ECCV. 2020: 213-229. DOI: 10.1007/978-3-030-58452-8_13.

[21] HAN J, DING J, LI J, et al. Align Deep Features for Oriented Object Detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-11. DOI: 10.1109/tgrs.2021.3062048.

[22] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. CVPR. 2016: 770-778. DOI: 10.1109/cvpr.2016.90.

[23] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]. ICLR. 2021.

[24] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database [C]. CVPR. 2009: 248-255. DOI: 10.1109/cvpr.2009.5206848.

[25] LIN T Y, MAIRE M, BELONGIE S J, et al. Microsoft COCO: Common Objects in Context [C]. ECCV. 2014: 740-755. DOI: 10.1007/978-3-319-10602-1_48.

[26] HE K, CHEN X, XIE S, et al. Masked autoencoders are scalable vision learners[C]. CVPR. 2022: 15979-15988. DOI: 10.1109/cvpr52688.2022.01553.

[27] OQUAB M, DARCET T, MOUTAKANNI T, et al. Dinov2: Learning robust visual features without supervision[J]. Trans. Mach. Learn. Res., 2024.

[28] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]. ICML. 2021: 8748-8763.

[29] SUN X, WANG P, LU W, et al. RingMo: A Remote Sensing Foundation Model With Masked Image Modeling[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023,

61: 1-22. DOI: 10.1109/tgrs.2022.3194732.

[30] CONG Y, KHANNA S, MENG C, et al. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery[C]. NeurIPS. 2022: 197-211. DOI: 10.52202/068431-0015.

[31] REED C J, GUPTA R, LI S, et al. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning[C]. ICCV. 2023: 4065-4076. DOI: 10.1109/iccv51070.2023.00378.

[32] LIU F, CHEN D, GUAN Z, et al. Remoteclip: A vision language foundation model for remote sensing[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-16. DOI: 10.1109/tgrs.2024.3390838.

[33] GUO X, LAO J, DANG B, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery[C]. CVPR. 2024: 27662-27673. DOI: 10.1109/cvpr52733.2024.02613.

[34] LIU H, LI C, WU Q, et al. Visual instruction tuning[C]. NeurIPS. 2023: 34892-34916. DOI: 10.52202/075280-1516.

[35] CHEN Z, WANG W, CAO Y, et al. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling[J]. CoRR, 2024. DOI: 10.48550/ARXIV.2412.05271.

[36] GIRDHAR R, EL-NOUBY A, LIU Z, et al. Imagebind: One embedding space to bind them all[C]. CVPR. 2023: 15180-15190. DOI: 10.1109/cvpr52729.2023.01457.

[37] ZHU B, LIN B, NING M, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment[C]. ICLR. 2024.

[38] 徐丹青, 吴一全. 光学遥感图像目标检测的深度学习算法研究进展[J]. 遥感学报, 2024, 28(12): 3045-3073. DOI: 10.11834/jrs.20243166.

[39] 袁一钦, 李浪, 姚西文, 等. 光学遥感图像目标检测数据集综述[J]. 遥感学报, 2023, 27(12): 2671-2687. DOI: 10.11834/jrs.20233457.

[40] 张磊, 张永生, 于英, 等. 遥感图像倾斜边界框目标检测研究进展与展望[J]. 遥感学报, 2022, 26(9): 1723-1743. DOI: 10.11834/jrs.20210247.

[41] 成飞飞, 付志涛, 黄亮, 等. 深度学习在光学和 SAR 影像融合研究进展[J]. 遥感学报, 2022, 26(9): 1744-1756. DOI: 10.11834/jrs.20211136.

[42] 张良培, 何江, 杨倩倩, 等. 数据驱动的多源遥感信息融合研究进展[J]. 测绘学报, 2022, 51(7): 1317-1337. DOI: 10.11947/j.AGCS.2022.20220171.

[43] 罗汝, 赵凌君, 何奇山, 等. SAR 图像飞机目标智能检测识别技术研究进展与展望[J]. 雷达学报 (中英文), 2024, 13(2): 307-330. DOI: 10.12000/JR23056.

[44] 付琨, 卢宛萱, 刘小煜, 等. 遥感基础模型发展综述与未来设想[J]. 遥感学报, 2024, 28(7): 1667-1680. DOI: 10.11834/jrs.20233313.

[45] FANG S, LI K, LI Z. Changer: Feature interaction is what you need for change detection [J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-11. DOI: 10.1109/tgrs.2023.3277496.

[46] ZHANG T, ZHANG X, LI J, et al. SAR ship detection dataset (SSDD): Official release

and comprehensive data analysis[J]. Remote Sensing, 2021, 13(18): 3690. DOI: 10.3390/rs13183690.

[47] WANG Y, WANG C, ZHANG H, et al. A SAR dataset of ship detection for deep learning under complex backgrounds[J]. remote sensing, 2019, 11(7): 765. DOI: 10.3390/rs11070765.

[48] LI C, SONG D, TONG R, et al. Illumination-aware faster R-CNN for robust multispectral pedestrian detection[J]. Pattern Recognition, 2019, 85: 161-171. DOI: 10.1016/j.patcog.2018.08.005.

[49] ZHANG L, LIU Z, ZHU X, et al. Weakly aligned feature fusion for multimodal object detection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025: 4145-4159. DOI: 10.1109/TNNLS.2021.3105143.

[50] YUAN M, WANG Y, WEI X. Translation, scale and rotation: Cross-modal alignment meets RGB-infrared vehicle detection[C]. ECCV. 2022: 509-525. DOI: 10.1007/978-3-031-20077-9_30.

[51] YUAN M, WEI X. C^2former: Calibrated and complementary transformer for rgb-infrared object detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-12. DOI: 10.1109/tgrs.2024.3376819.

[52] WANG X, CAI Z, GAO D, et al. Towards universal object detection by domain attention [C]. CVPR. 2019: 7281-7290. DOI: 10.1109/cvpr.2019.00746.

[53] XU H, FANG L, LIANG X, et al. Universal-rcnn: Universal object detector via transferable graph r-cnn[C]. AAAI. 2020: 12492-12499. DOI: 10.1609/aaai.v34i07.6937.

[54] ZHOU X, KOLTUN V, KRÄHENBÜHL P. Simple multi-dataset detection[C]. CVPR. 2022: 7561-7570. DOI: 10.1109/cvpr52688.2022.00742.

[55] CHEN Z, BADRINARAYANAN V, LEE C Y, et al. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks[C]. ICML. 2018: 793-802.

[56] KENDALL A, GAL Y, CIPOLLA R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics[C]. CVPR. 2018: 7482-7491. DOI: 10.1109/CVPR.2018.00781.

[57] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning[C]. CVPR. 2020: 9726-9735. DOI: 10.1109/cvpr42600.2020.00975.

[58] GRILL J B, STRUB F, ALTCHÉ F, et al. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning[C]. NeurIPS. 2020.

[59] MALL U, HARIHARAN B, BALA K. Change-aware sampling and contrastive learning for satellite images[C]. CVPR. 2023: 5261-5270. DOI: 10.1109/cvpr52729.2023.00509.

[60] MENDIETA M, HAN B, SHI X, et al. Towards geospatial foundation models via continual pretraining[C]. ICCV. 2023: 16760-16770. DOI: 10.1109/iccv51070.2023.01541.

[61] CHEN Z, WU J, WANG W, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks[C]. CVPR. 2023. DOI: 10.48550/ARXIV.2312.14238.

[62] BAI S, CHEN K, LIU X, et al. Qwen2. 5-vl technical report[J]. CoRR, 2025. DOI: 10.

48550/ARXIV.2502.13923.

[63] ZHOU B, LI L, WANG Y, et al. UNIALIGN: Scaling Multimodal Alignment within One Unified Model[C]. CVPR. 2025: 29644-29655. DOI: 10.1109/cvpr52734.2025.02760.

[64] ZHAO Q, MA Y, LYU S, et al. Embedded Self-Distillation in Compact Multibranch Ensemble Network for Remote Sensing Scene Classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-15. DOI: 10.1109/tgrs.2021.3126770.

[65] ZHAO Q, LYU S, LI Y, et al. MGML: Multigranularity Multilevel Feature Ensemble Network for Remote Sensing Scene Classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(5): 2308-2322. DOI: 10.1109/tnnls.2021.3106391.

[66] ZHANG H, WU C, ZHANG Z, et al. ResNeSt: Split-Attention Networks[C]. CVPRW. 2022: 2735-2745. DOI: 10.1109/cvprw56347.2022.00309.

[67] LIU J J, HOU Q, CHENG M M, et al. Improving Convolutional Networks With Self-Calibrated Convolutions[C]. CVPR. 2020: 10093-10102. DOI: 10.1109/cvpr42600.2020.01011.

[68] LI X, WANG W, HU X, et al. Selective kernel networks[C]. CVPR. 2019: 510-519. DOI: 10.1109/cvpr.2019.00060.

[69] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]. NeurIPS.

[70] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[C]. ICLR. 2021.

[71] WANG D, ZHANG Q, XU Y, et al. Advancing Plain Vision Transformer Towards Remote Sensing Foundation Model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-15. DOI: 10.1109/tgrs.2022.3222818.

[72] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. CVPR. 2021: 9992-10002. DOI: 10.1109/iccv48922.2021.00986.

[73] ZHANG C, WANG L, CHENG S, et al. SwinSUNet: Pure Transformer Network for Remote Sensing Image Change Detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-13. DOI: 10.1109/tgrs.2022.3160007.

[74] PANBOONYUEN T, JITKAJORNWANICH K, LAWAWIROJWONG S, et al. Transformer-Based Decoder Designs for Semantic Segmentation on Remotely Sensed Images[J]. Remote Sensing, 2021, 13(24): 5100. DOI: 10.3390/rs13245100.

[75] WANG X, CHEN G, QIAN G, et al. Large-scale Multi-modal Pre-trained Models: A Comprehensive Survey[J]. Machine Intelligence Research, 2023, 20(4): 447-482. DOI: 10.1007/s11633-022-1410-8.

[76] WANG W, XIE E, LI X, et al. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions[C]. ICCV. 2021: 548-558. DOI: 10.1109/iccv48922.2021.00061.

[77] WU Y H, LIU Y, ZHAN X, et al. P2T: Pyramid Pooling Transformer for Scene Understanding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 12760-12771. DOI: 10.1109/tpami.2022.3202765.

[78] RANFTL R, BOCHKOVSKIY A, KOLTUN V. Vision Transformers for Dense Prediction [C]. ICCV. 2021: 12159-12168. DOI: 10.1109/iccv48922.2021.01196.

[79] YAN H, LI Z, LI W, et al. ConTNet: Why not use convolution and transformer at the same time?[J]. CoRR, 2021.

[80] ZHENG S, LU J, ZHAO H, et al. Rethinking Semantic Segmentation From a Sequence-to-Sequence Perspective With Transformers[C]. CVPR. 2021: 6877-6886. DOI: 10.1109/cvpr46437.2021.00681.

[81] LUO W, LI Y, URTASUN R, et al. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks[C]. NeurIPS. 2017.

[82] FAN D P, JI G P, XU P, et al. Advances in Deep Concealed Scene Understanding[J]. Visual Intelligence, 2023, 1(1). DOI: 10.1007/s44267-023-00019-6.

[83] LIU Z, MAO H, WU C Y, et al. A convnet for the 2020s[C]. CVPR. 2022: 11966-11976. DOI: 10.1109/cvpr52688.2022.01167.

[84] DING X, ZHANG X, HAN J, et al. Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs[C]. CVPR. 2022: 11953-11965. DOI: 10.1109/cvpr52688.2022.01166.

[85] LIU S, CHEN T, CHEN X, et al. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity[C]. ICLR. 2023.

[86] GAO S, LI Z Y, HAN Q, et al. RF-Next: Efficient Receptive Field Search for Convolutional Neural Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022: 1-19. DOI: 10.1109/tpami.2022.3183829.

[87] GUO M H, LU C, LIU Z N, et al. Visual Attention Network[J]. Computational Visual Media, 2023, 9(4): 733-752. DOI: 10.1007/s41095-023-0364-2.

[88] GUO M H, LU C Z, HOU Q, et al. SegNeXt: Rethinking Convolutional Attention Design for Semantic Segmentation[C]. NeurIPS. 2022.

[89] HOU Q, LU C Z, CHENG M M, et al. Conv2Former: A Simple Transformer-Style ConvNet for Visual Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 8274-8283. DOI: 10.1109/tpami.2024.3401450.

[90] GUO M H, XU T, LIU J J, et al. Attention Mechanisms in Computer Vision: A Survey[J]. Computational Visual Media, 2022, 8(3): 331-368. DOI: 10.1007/s41095-022-0271-y.

[91] HU J, SHEN L, SUN G. Squeeze-and-Excitation Networks[C]. CVPR. 2018: 7132-7141. DOI: 10.1109/cvpr.2018.00745.

[92] HU J, SHEN L, ALBANIE S, et al. Gather-Excite: Exploiting Feature Context in Convolutional Neural Networks[C]. NeurPIS. 2018: 9423-9433.

[93] CAO Y, XU J, LIN S, et al. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond[C]. ICCVW. 2019: 1971-1980. DOI: 10.1109/iccvw.2019.00246.

[94] LI Y, LI X, YANG J. Spatial Group-wise Enhance: Enhancing Semantic Feature Learning in CNN[C]. ACCV. 2023: 316-332. DOI: 10.1007/978-3-031-26348-4_19.

[95] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional Block Attention Module[C]. ECCV. 2018: 3-19. DOI: 10.1007/978-3-030-01234-2_1.

[96] PARK J, WOO S, LEE J Y, et al. BAM: Bottleneck Attention Module[C]. British Machine Vision Conference. 2018: 147.

[97] SRIVASTAVA S, SHARMA G. Omnivec: Learning robust representations with cross modal sharing[C]. Winter Conference on Applications of Computer Vision. 2024: 1225-1237. DOI: 10.1109/wacv57701.2024.00127.

[98] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]. ECCV. 2020: 213-229. DOI: 10.1007/978-3-030-58452-8_13.

[99] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[C]. ICCV. 2023: 3992-4003. DOI: 10.1109/iccv51070.2023.00371.

[100] WANG W, XIE E, LI X, et al. PVT v2: Improved baselines with Pyramid Vision Transformer[J]. Computational Visual Media, 2022, 8(3): 415-424. DOI: 10.1007/s41095-022-0274-8.

[101] ZHANG X, TIAN Y, XIE L, et al. Hivit: A simpler and more efficient design of hierarchical vision transformer[C]. ICLR. 2023.

[102] XU Y, ZHANG Q, ZHANG J, et al. Vitae: Vision transformer advanced by exploring intrinsic inductive bias[C]. NeurIPS. 2021: 28522-28535.

[103] YU H, TIAN Y, YE Q, et al. Spatial transform decoupling for oriented object detection[C]. AAAI. 2024: 6782-6790. DOI: 10.1609/aaai.v38i7.28502.

[104] YANG B, BENDER G, LE Q V, et al. CondConv: Conditionally parameterized convolutions for efficient inference[C]. NeurIPS. 2019: 1305-1316.

[105] CHEN Y, DAI X, LIU M, et al. Dynamic convolution: Attention over convolution kernels [C]. CVPR. 2020: 11027-11036. DOI: 10.1109/cvpr42600.2020.01104.

[106] LI X, WANG W, HU X, et al. Selective kernel networks[C]. CVPR. 2019: 510-519. DOI: 10.1109/cvpr.2019.00060.

[107] ZHU X, HU H, LIN S, et al. Deformable ConvNets V2: More Deformable, Better Results [C]. CVPR. 2019: 9300-9308. DOI: 10.1109/cvpr.2019.00953.

[108] DAI J, QI H, XIONG Y, et al. Deformable Convolutional Networks[C]. ICCV. 2017: 764-773. DOI: 10.1109/iccv.2017.89.

[109] LIU Z, MAO H, WU C Y, et al. A ConvNet for the 2020s[C]. CVPR. 2022: 11966-11976. DOI: 10.1109/cvpr52688.2022.01167.

[110] YU W, LUO M, ZHOU P, et al. MetaFormer is Actually What You Need for Vision[C]. CVPR. 2022: 10809-10819. DOI: 10.1109/cvpr52688.2022.01055.

[111] HENDRYCKS D, GIMPEL K. Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units[J]. CoRR, 2016.

[112] XIE X, CHENG G, WANG J, et al. Oriented R-CNN for Object Detection[C]. ICCV. 2021: 3500-3509. DOI: 10.1109/iccv48922.2021.00350.

[113] HAN J, DING J, LI J, et al. Align Deep Features for Oriented Object Detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-11. DOI: 10.1109/tgrs.2021.3062048.

[114] YANG X, LIU Q, YAN J, et al. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object[C]. AAAI. 2021: 3163-3171. DOI: 10.1609/aaai.v35i4.16426.

[115] WANG D, ZHANG J, DU B, et al. An empirical study of remote sensing pretraining[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-20. DOI: 10.1109/tgrs.2022.3176603.

[116] LONG Y, XIA G S, LI S, et al. On Creating Benchmark Dataset for Aerial Image Interpretation: Reviews, Guidances, and Million-AID[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 4205-4230. DOI: 10.1109/jstars.2021.3070368.

[117] CHEN S B, WEI Q S, WANG W Z, et al. Remote Sensing Scene Classification via Multi-Branch Local Attention Network[J]. IEEE Transactions on Image Processing, 2022, 31: 99-109. DOI: 10.1109/tip.2021.3127851.

[118] ZHANG X, AN W, SUN J, et al. Best representation branch model for remote sensing image scene classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 9768-9780. DOI: 10.1109/jstars.2021.3114404.

[119] LI B, GUO Y, YANG J, et al. Gated recurrent multiattention network for VHR remote sensing image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-13. DOI: 10.1109/tgrs.2021.3093914.

[120] YANG Y, NEWSAM S. Bag-of-visual-words and spatial extensions for land-use classification[C]. Proceedings of the international conference on advances in geographic information systems. 2010: 270-279. DOI: 10.1145/1869790.1869829.

[121] XIA G S, HU J, HU F, et al. AID: A benchmark data set for performance evaluation of aerial scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(7): 3965-3981. DOI: 10.1109/tgrs.2017.2685945.

[122] CHENG G, HAN J, LU X. Remote sensing image scene classification: Benchmark and state of the art[J]. Proceedings of the IEEE, 2017, 105(10): 1865-1883. DOI: 10.1109/jproc.2017.2675998.

[123] ZHANG G, XU W, ZHAO W, et al. A multiscale attention network for remote sensing scene images classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 9530-9545. DOI: 10.1109/jstars.2021.3109661.

[124] HE N, FANG L, LI S, et al. Skip-Connected Covariance Network for Remote Sensing Scene Classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(5): 1461-1474. DOI: 10.1109/tnnls.2019.2920374.

[125] LIU C, DAI H, WANG S, et al. Remote Sensing Image Scene Classification Based on Multidimensional Attention and Feature Enhancement.[J]. IAENG International Journal of Computer Science, 2023.

[126] WANG S, GUAN Y, SHAO L. Multi-Granularity Canonical Appearance Pooling for Remote Sensing Scene Classification[J]. IEEE Transactions on Image Processing, 2020, 29: 5396-5407. DOI: 10.1109/tip.2020.2983560.

[127] BI Q, QIN K, ZHANG H, et al. Local Semantic Enhanced ConvNet for Aerial Scene Recognition[J]. IEEE Transactions on Image Processing, 2021, 30: 6498-6511. DOI: 10.1109/tip.2021.3092816.

[128] WANG S, REN Y, PARR G P, et al. Invariant Deep Compressible Covariance Pooling for Aerial Scene Categorization[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(8): 6549-6561. DOI: 10.1109/tgrs.2020.3026221.

[129] ZHAO Z, LI J, LUO Z, et al. Remote sensing image scene classification based on an enhanced attention module[J]. IEEE Transactions on Geoscience and Remote Sensing Letters, 2021, 18(11): 1926-1930. DOI: 10.1109/lgrs.2020.3011405.

[130] LI F, FENG R, HAN W, et al. High-Resolution Remote Sensing Image Scene Classification via Key Filter Bank Based on Convolutional Neural Network[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(11): 8077-8092. DOI: 10.1109/tgrs.2020.2987060.

[131] WANG W, SUN Y, LI J, et al. Frequency and spatial based multi-layer context network (FSCNet) for remote sensing scene classification[J]. International Journal of Applied Earth Observation and Geoinformation, 2024, 128: 103781. DOI: 10.1016/j.jag.2024.103781.

[132] DONG Z, GU Y, LIU T. UPetu: A Unified Parameter-efficient Fine-tuning Framework for Remote Sensing Foundation Model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-13. DOI: 10.1109/tgrs.2024.3382734.

[133] LIU Z, WANG H, WENG L, et al. Ship Rotated Bounding Box Space for Ship Extraction From High-Resolution Optical Satellite Images With Complex Backgrounds[J]. IEEE Transactions on Geoscience and Remote Sensing Letters, 2016, 13(8): 1074-1078. DOI: 10.1109/lgrs.2016.2565705.

[134] XIA G S, BAI X, DING J, et al. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images[C]. CVPR. 2018: 3974-3983. DOI: 10.1109/cvpr.2018.00418.

[135] SUN X, WANG P, YAN Z, et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 184: 116-130. DOI: 10.1016/j.isprsjprs.2021.12.004.

[136] ZHIRUI W, SUN X. SAR-AIRcraft-1.0: High-resolution SAR Aircraft Detection and Recognition Dataset[Z]. https://radars.ac.cn/web/data/getData?dataType=SARDataset_en. 2023.

[137] LYU C, ZHANG W, HUANG H, et al. RTMDet: An Empirical Study of Designing Real-Time Object Detectors[J]. CoRR, 2022. DOI: 10.48550/ARXIV.2212.07784.

[138] GUO Z, LIU C, ZHANG X, et al. Beyond Bounding-Box: Convex-hull Feature Adaptation for Oriented and Densely Packed Object Detection[C]. CVPR. 2021: 8788-8797. DOI: 10.1109/cvpr46437.2021.00868.

[139] LANG S, VENTOLA F, KERSTING K. DAFNe: A One-Stage Anchor-Free Deep Model for Oriented Object Detection[J]. CoRR, 2021.

[140] HOU L, LU K, XUE J, et al. Shape-adaptive selection and measurement for oriented object

detection[C]. AAAI. 2022: 923-932. DOI: 10.1609/aaai.v36i1.19975.

[141] DAI L, LIU H, TANG H, et al. AO2-DETR: Arbitrary-Oriented Object Detection Transformer[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(5): 2342-2356. DOI: 10.1109/tcsvt.2022.3222906.

[142] YANG X, YAN J, MING Q, et al. Rethinking Rotated Object Detection with Gaussian Wasserstein Distance Loss[C]. ICML. 2021: 11830-11841.

[143] YANG X, YANG X, YANG J, et al. Learning High-Precision Bounding Box for Rotated Object Detection via Kullback-Leibler Divergence[C]. NeurIPS. 2021: 18381-18394.

[144] YANG X, YANG J, YAN J, et al. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects[C]. ICCV. 2019: 8231-8240. DOI: 10.1109/iccv.2019.00832.

[145] LI Y, MAO H, GIRSHICK R, et al. Exploring plain vision transformer backbones for object detection[C]. ECCV. 2022: 280-296. DOI: 10.1007/978-3-031-20077-9_17.

[146] DING J, XUE N, LONG Y, et al. Learning RoI Transformer for Oriented Object Detection in Aerial Images[C]. CVPR. 2019: 2844-2853. DOI: 10.1109/cvpr.2019.00296.

[147] XU Y, FU M, WANG Q, et al. Gliding Vertex on the Horizontal Bounding Box for Multi-Oriented Object Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(4): 1452-1459. DOI: 10.1109/tpami.2020.2974745.

[148] WANG J, YANG W, LI H C, et al. Learning Center Probability Map for Detecting Objects in Aerial Images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(5): 4307-4323. DOI: 10.1109/tgrs.2020.3010051.

[149] YANG X, YAN J. Arbitrary-Oriented Object Detection with Circular Smooth Label[C]. ECCV. 2020: 677-694. DOI: 10.1007/978-3-030-58598-3_40.

[150] HAN J, DING J, XUE N, et al. ReDet: A Rotation-equivariant Detector for Aerial Object Detection[C]. CVPR. 2021: 2785-2794. DOI: 10.1109/cvpr46437.2021.00281.

[151] CHENG G, YAO Y, LI S, et al. Dual-Aligned Oriented Detector[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-11. DOI: 10.1109/tgrs.2022.3149780.

[152] CHENG G, WANG J, LI K, et al. Anchor-Free Oriented Proposal Generator for Object Detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-11. DOI: 10.1109/tgrs.2022.3183022.

[153] YANG X, ZHOU Y, ZHANG G, et al. The KFIoU Loss for Rotated Object Detection[C]. ICLR. 2023.

[154] CAI Z, VASCONCELOS N. Cascade R-CNN: High quality object detection and instance segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(5): 1483-1498. DOI: 10.1109/tpami.2019.2956516.

[155] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results[Z].

[156] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results[Z].

[157] PAN X, REN Y, SHENG K, et al. Dynamic Refinement Network for Oriented and Densely Packed Object Detection[C]. CVPR. 2020: 11204-11213. DOI: 10.1109/cvpr42600.2020.01122.

[158] MING Q, ZHOU Z, MIAO L, et al. Dynamic Anchor Learning for Arbitrary-Oriented Object Detection[C]. AAAI. 2021: 2355-2363. DOI: 10.1609/aaai.v35i3.16336.

[159] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. CVPR. 2016: 770-778. DOI: 10.1109/cvpr.2016.90.

[160] GAO S H, CHENG M M, ZHAO K, et al. Res2Net: A New Multi-scale Backbone Architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(2): 652-662. DOI: 10.1109/tpami.2019.2938758.

[161] WOO S, DEBNATH S, HU R, et al. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders[C]. CVPR. 2023: 16133-16142. DOI: 10.1109/cvpr52729.2023.01548.

[162] WANG L, LI R, WANG D, et al. Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images[J]. Remote Sensing, 2021, 13(16): 3065. DOI: 10.3390/rs13163065.

[163] For PHOTOGRAMMETRY T I S, (ISPRS) R S. 2D Semantic Labeling Contest - Potsdam [Z]. https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx. 2022.

[164] ISPRS. 2D Semantic Labeling - Vaihingen[Z]. https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx. 2022.

[165] WANG J, ZHENG Z, MA A, et al. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation[C]. NeurIPS Datasets and Benchmarks. 2021.

[166] LYU Y, VOSSELMAN G, XIA G S, et al. UAVid: A semantic segmentation dataset for UAV imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 165: 108-119. DOI: 10.1016/j.isprsjprs.2020.05.009.

[167] TONG X Y, XIA G S, LU Q, et al. Land-cover classification with high-resolution remote sensing images using transferable deep models[J]. Remote Sensing of Environment, 2020, 237: 111322. DOI: 10.1016/j.rse.2019.111322.

[168] LI R, DUAN C, ZHENG S, et al. MACU-Net for Semantic Segmentation of Fine-Resolution Remotely Sensed Images[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5. DOI: 10.1109/lgrs.2021.3052886.

[169] ROMERA E, ALVAREZ J M, BERGASA L M, et al. ERFNet: Efficient residual factorized convnet for real-time semantic segmentation[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(1): 263-272. DOI: 10.1109/tits.2017.2750080.

[170] LI G, YUN I, KIM J, et al. DABNet: Depth-wise Asymmetric Bottleneck for Real-time Semantic Segmentation[C]. BMVC. 2019.

[171] YU C, WANG J, PENG C, et al. BiSeNet: Bilateral segmentation network for real-time semantic segmentation[C]. ECCV. 2018: 334-349. DOI: 10.1007/978-3-030-01261-8_20.

[172] ZHENG X, HUAN L, XIA G S, et al. Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 170: 15-28. DOI: 10.1016/j.isprsjprs.2020.09.019.

[173] LI R, ZHENG S, ZHANG C, et al. Multiattention network for semantic segmentation of fine-resolution remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-13. DOI: 10.1109/tgrs.2021.3093977.

[174] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation[C]. CVPR. 2019: 3141-3149. DOI: 10.1109/cvpr.2019.00326.

[175] ORŠIĆ M, ŠEGVIĆ S. Efficient semantic segmentation with pyramidal fusion[J]. Pattern Recognition, 2021: 107611. DOI: 10.1016/J.PATCOG.2020.107611.

[176] HU P, PERAZZI F, HEILBRON F C, et al. Real-time semantic segmentation with fast attention[J]. IEEE Robotics and Automation Letters, 2021, 6(1): 263-270. DOI: 10.1109/lra.2020.3039744.

[177] ZHUANG J, YANG J, GU L, et al. ShelfNet for fast semantic segmentation[C]. ICCVW. 2019: 847-856. DOI: 10.1109/iccvw.2019.00113.

[178] LI R, ZHENG S, ZHANG C, et al. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 181: 84-98. DOI: 10.1016/j.isprsjprs.2021.09.005.

[179] STRUDEL R, GARCIA R, LAPTEV I, et al. Segmenter: Transformer for semantic segmentation[C]. ICCV. 2021: 7242-7252. DOI: 10.1109/iccv48922.2021.00717.

[180] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]. CVPR. 2017: 6230-6239. DOI: 10.1109/cvpr.2017.660.

[181] SRINIVAS A, LIN T Y, PARMAR N, et al. Bottleneck transformers for visual recognition [C]. CVPR. 2021: 16514-16524. DOI: 10.1109/cvpr46437.2021.01625.

[182] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]. ECCV. 2018: 833-851. DOI: 10.1007/978-3-030-01234-2_49.

[183] XIE E, WANG W, YU Z, et al. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers[C]. NeurIPS. 2021: 12077-12090.

[184] XIAO T, LIU Y, ZHOU B, et al. Unified Perceptual Parsing for Scene Understanding[C]. ECCV. 2018: 432-448. DOI: 10.1007/978-3-030-01228-1_26.

[185] ZHOU Z, RAHMAN SIDDIQUEE M M, TAJBAKHSH N, et al. UNet++: A Nested U-Net Architecture for Medical Image Segmentation[C]. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. 2018: 3-11. DOI: 10.1007/978-3-030-00889-5_1.

[186] KIRILLOV A, GIRSHICK R, HE K, et al. Panoptic feature pyramid networks[C]. CVPR. 2019: 6392-6401. DOI: 10.1109/cvpr.2019.00656.

[187] ZHENG Z, ZHONG Y, WANG J, et al. Foreground-aware relation network for geospatial

object segmentation in high spatial resolution remote sensing imagery[C]. CVPR. 2020: 4095-4104. DOI: 10.1109/cvpr42600.2020.00415.

[188] MA A, WANG J, ZHONG Y, et al. FactSeg: Foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-16. DOI: 10.1109/tgrs.2021.3097148.

[189] CHEN J, LU Y, YU Q, et al. Transunet: Transformers make strong encoders for medical image segmentation[J]. CoRR, 2021.

[190] WANG J, SUN K, CHENG T, et al. Deep High-Resolution Representation Learning for Visual Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3349-3364. DOI: 10.1109/tpami.2020.2983686.

[191] WANG L L, LUI S S, CHAN R C. The past and future of mapping the biomarkers of psychosis[J]. Current Opinion in Behavioral Sciences, 2022, 43: 1-5. DOI: 10.1016/j.cobeha.2021.06.007.

[192] SUN L, ZOU H, WEI J, et al. Semantic Segmentation of High-Resolution Remote Sensing Images Based on Sparse Self-Attention and Feature Alignment[J]. Remote Sensing, 2023, 15(6): 1598. DOI: 10.3390/rs15061598.

[193] YANG M Y, KUMAAR S, LYU Y, et al. Real-time Semantic Segmentation with Context Aggregation Network[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 178: 124-134. DOI: 10.1016/j.isprsjprs.2021.06.006.

[194] XU W, XU Y, CHANG T, et al. Co-scale conv-attentional image transformers[C]. ICCV. 2021: 9961-9970. DOI: 10.1109/iccv48922.2021.00983.

[195] LIN H, HANG R, WANG S, et al. DiFormer: A Difference Transformer Network for Remote Sensing Change Detection[J]. IEEE Geoscience and Remote Sensing Letters, 2024, 21: 1-5. DOI: 10.1109/lgrs.2024.3359220.

[196] HAN C, WU C, GUO H, et al. Change guiding network: Incorporating change prior to guide change detection in remote sensing imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 8395-8407. DOI: 10.1109/jstars.2023.3310208.

[197] CHEN H, SHI Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection[J]. Remote Sensing, 2020, 12(10): 1662. DOI: 10.3390/rs12101662.

[198] SHEN L, LU Y, CHEN H, et al. S2Looking: A satellite side-looking dataset for building change detection[J]. Remote Sensing, 2021, 13(24): 5094. DOI: 10.3390/rs13245094.

[199] DAUDT R C, LE SAUX B, BOULCH A. Fully convolutional siamese networks for change detection[C]. 2018 25th IEEE international conference on image processing (ICIP). 2018: 4063-4067. DOI: 10.1109/ICIP.2018.8451652.

[200] LIU Y, PANG C, ZHAN Z, et al. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model[J]. IEEE Geoscience and Remote Sensing Letters, 2021, 18(5): 811-815. DOI: 10.1109/lgrs.2020.2988032.

[201] HAN C, WU C, GUO H, et al. HANet: A hierarchical attention network for change detection with bi-temporal very-high-resolution remote sensing images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 3867-3878. DOI: 10.1109/jstars.2023.3264802.

[202] CHEN H, LI W, SHI Z. Adversarial instance augmentation for building change detection in remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-16. DOI: 10.1109/tgrs.2021.3066802.

[203] ZHANG C J, LIU J W. Change detection with incorporating multi-constraints and loss weights[J]. Engineering Applications of Artificial Intelligence, 2024, 133: 108163. DOI: 10.1016/j.engappai.2024.108163.

[204] ZHANG C, YUE P, TAPETE D, et al. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 166: 183-200. DOI: 10.1016/j.isprsjprs.2020.06.003.

[205] FANG S, LI K, SHAO J, et al. SNUNet-CD: A densely connected Siamese network for change detection of VHR images[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5. DOI: 10.1109/lgrs.2021.3056416.

[206] HAN C, WU C, DU B. HCGMNET: A hierarchical change guiding map network for change detection[C]. IEEE International Geoscience and Remote Sensing Symposium. 2023: 5511-5514. DOI: 10.1109/igarss52108.2023.10283341.

[207] HAN C, WU C, HU M, et al. C2F-SemiCD: A Coarse-to-Fine Semi-Supervised Change Detection Method Based on Consistency Regularization in High-Resolution Remote-Sensing Images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-21. DOI: 10.1109/tgrs.2024.3370568.

[208] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks[C]. ICCV. 2017: 764-773. DOI: 10.1109/iccv.2017.89.

[209] MUHAMMAD M B, YEASIN M. Eigen-CAM: Class Activation Map using Principal Components[C]. 2020 International Joint Conference on Neural Networks (IJCNN). 2020: 1-7. DOI: 10.1109/ijcnn48605.2020.9206626.

[210] SUN G C, LIU Y, XIANG J, et al. Spaceborne synthetic aperture radar imaging algorithms: An overview[J]. IEEE Geoscience and Remote Sensing Magazine, 2022, 10(1): 161-184. DOI: 10.1109/mgrs.2021.3097894.

[211] TSOKAS A, RYSZ M, PARDALOS P M, et al. SAR data applications in earth observation: An overview[J]. Expert Systems with Applications, 2022, 205: 117342. DOI: 10.1016/j.eswa.2022.117342.

[212] ZHIRUI W, YUZHUO K, XUAN Z, et al. SAR-AIRcraft-1.0: High-resolution SAR aircraft detection and recognition dataset[J]. J. Radars, 2023.

[213] XIAN S, ZHIRUI W, YUANRUI S, et al. AIR-SARShip-1.0: High-resolution SAR ship detection dataset[J]. J. Radars, 2019.

[214] ZHANG T, ZHANG X, LI J, et al. SAR ship detection dataset (SSDD): Official release and comprehensive data analysis[J]. Remote Sensing, 2021, 13(18): 3690. DOI: 10.3390/rs13183690.

[215] WEI S, ZENG X, QU Q, et al. HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation[J]. IEEE Access, 2020, 8: 120234-120254. DOI: 10.1109/access.2020.3005861.

[216] MOREIRA A, PRATS-IRAOLA P, YOUNIS M, et al. A tutorial on synthetic aperture radar [J]. IEEE Geoscience and remote sensing magazine, 2013, 1(1): 6-43. DOI: 10.1109/mgrs.2013.2248301.

[217] LI W, WEI Y, LIU T, et al. Self-supervised learning for sar atr with a knowledge-guided predictive architecture[J]. CoRR, 2023. DOI: 10.48550/ARXIV.2311.15153.

[218] SONG S, XU B, YANG J. SAR target recognition via supervised discriminative dictionary learning and sparse representation of the SAR-HOG feature[J]. Remote Sensing, 2016, 8(8): 683. DOI: 10.3390/rs8080683.

[219] QI S, MA J, LIN J, et al. Unsupervised ship detection based on saliency and S-HOG descriptor from optical satellite images[J]. IEEE geoscience and remote sensing letters, 2015: 1451-1455. DOI: 10.1109/LGRS.2015.2408355.

[220] LI W, HE M, ZHANG S. A new Canny-based edge detector for SAR image[C]. Congress on Image and Signal Processing. 2008: 211-215. DOI: 10.1109/cisp.2008.686.

[221] LIU H, JEZEK K. Automated extraction of coastline from satellite imagery by integrating Canny edge detection and locally adaptive thresholding methods[J]. International journal of remote sensing, 2004, 25(5): 937-958. DOI: 10.1080/0143116031000139890.

[222] LI X, LV C, WANG W, et al. Generalized focal loss: Towards efficient representation learning for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022: 1-14. DOI: 10.1109/tpami.2022.3180392.

[223] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]. CVPR. 2016: 779-788. DOI: 10.1109/cvpr.2016.91.

[224] CHEN Y, YUAN X, WU R, et al. YOLO-MS: Rethinking Multi-Scale Representation Learning for Real-time Object Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(6): 4240-4252. DOI: 10.1109/tpami.2025.3538473.

[225] LIU Z, MAO H, WU C Y, et al. A convnet for the 2020s[C]. CVPR. 2022: 11966-11976. DOI: 10.1109/cvpr52688.2022.01167.

[226] GUO M H, LU C, LIU Z N, et al. Visual Attention Network[J]. Computational Visual Media, 2023, 9(4): 733-752. DOI: 10.1007/s41095-023-0364-2.

[227] CHEN L, LUO R, XING J, et al. Geospatial transformer is what you need for aircraft detection in SAR Imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-15. DOI: 10.1109/tgrs.2022.3162235.

[228] ZHOU K, ZHANG M, WANG H, et al. Ship detection in SAR images based on multi-scale feature extraction and adaptive feature fusion[J]. Remote Sensing, 2022, 14(3): 755. DOI:

10.3390/rs14030755.

[229] ZHANG P, XU H, TIAN T, et al. SEFEPNet: Scale expansion and feature enhancement pyramid network for SAR aircraft detection with small sample dataset[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 3365-3375. DOI: 10.1109/jstars.2022.3169339.

[230] ZHANG T, ZHANG X, KE X. Quad-FPN: A novel quad feature pyramid network for SAR ship detection[J]. Remote Sensing, 2021, 13(14): 2771. DOI: 10.3390/rs13142771.

[231] ZHAO Y, ZHAO L, LI C, et al. Pyramid attention dilated network for aircraft detection in SAR images[J]. IEEE Geoscience and Remote Sensing Letters, 2021, 18(4): 662-666. DOI: 10.1109/lgrs.2020.2981255.

[232] WANG J, XIAO H, CHEN L, et al. Integrating weighted feature fusion and the spatial attention module with convolutional neural networks for automatic aircraft detection from SAR images[J]. Remote Sensing, 2021, 13(5): 910. DOI: 10.3390/rs13050910.

[233] GUO H, YANG X, WANG N, et al. A CenterNet++ model for ship detection in SAR images [J]. Pattern Recognition, 2021, 112: 107787. DOI: 10.1016/j.patcog.2020.107787.

[234] ZHOU X, WANG D, KRÄHENBÜHL P. Objects as points[J]. CoRR, 2019.

[235] XIA R, CHEN J, HUANG Z, et al. CRTransSar: A visual transformer based on contextual joint representation learning for SAR ship detection[J]. Remote Sensing, 2022, 14(6): 1488. DOI: 10.3390/rs14061488.

[236] GURURANGAN S, MARASOVIĆ A, SWAYAMDIPTA S, et al. Don't stop pretraining: Adapt language models to domains and tasks[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8342-8360. DOI: 10.18653/v1/2020.acl-main.740.

[237] ZHANG T, GAO P, DONG H, et al. Consecutive pre-training: A knowledge transfer learning strategy with relevant unlabeled data for remote sensing domain[J]. Remote Sensing, 2022, 14(22): 5675. DOI: 10.3390/rs14225675.

[238] CONTRIBUTORS S 1. Sentinel-1 - Missions.[Z]. https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-1. 2024.

[239] ZHANG P, XU H, TIAN T, et al. SEFEPNet: Scale expansion and feature enhancement pyramid network for SAR aircraft detection with small sample dataset[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 3365-3375. DOI: 10.1109/jstars.2022.3169339.

[240] WANG C, RUAN R, ZHAO Z, et al. Category-oriented Localization Distillation for SAR Object Detection and A Unified Benchmark[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-14. DOI: 10.1109/tgrs.2023.3291356.

[241] LIN X, ZHANG B, WU F, et al. SIVED: A SAR Image Dataset for Vehicle Detection Based on Rotatable Bounding Box[J]. Remote Sensing, 2023, 15(11): 2825. DOI: 10.3390/rs15112825.

[242] JIN Y, DUAN Y. Wavelet scattering network-based machine learning for ground penetrat-

ing radar imaging: Application in pipeline identification[J]. Remote Sensing, 2020, 12(21): 3655. DOI: 10.3390/rs12213655.

[243] ZHANG J, XING M, XIE Y. FEC: A feature fusion framework for SAR target recognition based on electromagnetic scattering features and deep CNN features[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(3): 2174-2187. DOI: 10.1109/tgrs.2020.3003264.

[244] CONTRIBUTORS M. Openmmlab's pre-training toolbox and benchmark.[Z]. https://github.com/openmmlab/mmpretrain. 2024.

[245] LI Y, HOU Q, ZHENG Z, et al. Large Selective Kernel Network for Remote Sensing Object Detection[C]. ICCV. 2023: 16748-16759. DOI: 10.1109/iccv51070.2023.01540.

[246] CHEN K, WANG J, PANG J, et al. MMDetection: Open mmlab detection toolbox and benchmark[J]. CoRR, 2019.

[247] LI K, WAN G, CHENG G, et al. Object detection in optical remote sensing images: A survey and a new benchmark[J]. ISPRS, 2020, 159: 296-307. DOI: 10.1016/j.isprsjprs.2019.11.023.

[248] LU X, LI B, YUE Y, et al. Grid r-cnn[C]. CVPR. 2019: 7355-7364. DOI: 10.1109/cvpr.2019.00754.

[249] ZHU X, SU W, LU L, et al. Deformable DETR: Deformable Transformers for End-to-End Object Detection[J]. arXiv, 2021.

[250] SUN P, ZHANG R, JIANG Y, et al. Sparse r-cnn: End-to-end object detection with learnable proposals[C]. CVPR. 2021: 14449-14458. DOI: 10.1109/cvpr46437.2021.01422.

[251] LIU S, LI F, ZHANG H, et al. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR[C]. ICLR. 2022.

[252] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks[C]. ICCV. 2017: 764-773. DOI: 10.1109/iccv.2017.89.

[253] WANG J, CHEN K, YANG S, et al. Region proposal by guided anchoring[C]. CVPR. 2019: 2960-2969. DOI: 10.1109/cvpr.2019.00308.

[254] ZHANG X, WAN F, LIU C, et al. Freeanchor: Learning to match anchors for visual object detection[C]. NeurIPS. 2019: 147-155.

[255] WU Y, CHEN Y, YUAN L, et al. Rethinking classification and localization for object detection[C]. CVPR. 2020: 10183-10192. DOI: 10.1109/cvpr42600.2020.01020.

[256] CHEN C, HE C, HU C, et al. A deep neural network based on an attention mechanism for SAR ship detection in multiscale and complex scenarios[J]. IEEE Access, 2019, 7: 104848-104863. DOI: 10.1109/access.2019.2930939.

[257] JIAO J, ZHANG Y, SUN H, et al. A densely connected end-to-end neural network for multiscale and multiscene SAR ship detection[J]. IEEE Access, 2018, 6: 20881-20892. DOI: 10.1109/access.2018.2825376.

[258] ZHAO Y, ZHAO L, XIONG B, et al. Attention receptive pyramid network for ship detection in SAR images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and

Remote Sensing, 2020, 13: 2738-2756. DOI: 10.1109/jstars.2020.2997081.

[259] CUI Z, LI Q, CAO Z, et al. Dense attention pyramid networks for multi-scale ship detection in SAR images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(11): 8983-8997. DOI: 10.1109/tgrs.2019.2923988.

[260] WEI S, SU H, MING J, et al. Precise and robust ship detection for high-resolution SAR imagery based on HR-SDNet[J]. Remote Sensing, 2020, 12(1): 167. DOI: 10.3390/rs12010167.

[261] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]. CVPR. 2018: 8759-8768. DOI: 10.1109/cvpr.2018.00913.

[262] LIN Z, JI K, LENG X, et al. Squeeze and excitation rank faster R-CNN for ship detection in SAR images[J]. IEEE Geoscience and Remote Sensing Letters, 2019, 16(5): 751-755. DOI: 10.1109/lgrs.2018.2882551.

[263] FU J, SUN X, WANG Z, et al. An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(2): 1331-1344. DOI: 10.1109/tgrs.2020.3005151.

[264] MA W, YANG X, ZHU H, et al. NRENet: Neighborhood removal-and-emphasis network for ship detection in SAR Images[J]. International Journal of Applied Earth Observation and Geoinformation, 2024, 131: 103927. DOI: 10.1016/j.jag.2024.103927.

[265] YANG W, HOU Y, LIU L, et al. SARATR-X: A Foundation Model for Synthetic Aperture Radar Images Target Recognition[J]. CoRR, 2024. DOI: 10.48550/ARXIV.2405.09365.

[266] DAI Y, PAN P, QIAN Y, et al. Pick of the Bunch: Detecting Infrared Small Targets Beyond Hit-Miss Trade-Offs via Selective Rank-Aware Attention[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-15. DOI: 10.1109/tgrs.2024.3458896.

[267] ZHANG W, JIAO L, LI Y, et al. Laplacian Feature Pyramid Network for Object Detection in VHR Optical Remote Sensing Images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-14. DOI: 10.1109/tgrs.2021.3072488.

[268] ZHOU J, XIAO C, PENG B, et al. DiffDet4SAR: Diffusion-based aircraft target detection network for SAR images[J]. IEEE Transactions on Geoscience and Remote Sensing Letters, 2024, 21: 1-5. DOI: 10.1109/lgrs.2024.3386020.

[269] LI W, YANG W, LIU T, et al. Predicting gradient is better: Exploring self-supervised learning for SAR ATR with a joint-embedding predictive architecture[J]. ISPRS Journal of the Photogrammetry and Remote Sensing, 2024, 218: 326-338. DOI: 10.1016/j.isprsjprs.2024.09.013.

[270] LI Y, LI X, DAI Y, et al. LSKNet: A Foundation Lightweight Backbone for Remote Sensing [J]. International Journal of Computer Vision, 2025, 133(3): 1410-1431. DOI: 10.1007/s11263-024-02247-9.

[271] YANG X, YAN J, MING Q, et al. Rethinking Rotated Object Detection with Gaussian Wasserstein Distance Loss[C]. ICML. 2021: 11830-11841.

[272] DAI Y, WU Y, ZHOU F, et al. Attentional local contrast networks for infrared small target detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(11): 9813-9824. DOI: 10.1109/tgrs.2020.3044958.

[273] LIU J, CHEN H, WANG Y. Multi-source remote sensing image fusion for ship target detection and recognition[J]. Remote Sensing, 2021, 13(23): 4852. DOI: 10.3390/rs13234852.

[274] ZHANG H, HUANG S L, KURUOGLU E E. HGR Correlation Pooling Fusion Framework for Recognition and Classification in Multimodal Remote Sensing Data[J]. Remote Sensing, 2024, 16(10): 1708. DOI: 10.3390/rs16101708.

[275] ZHANG Z, ZHANG L, WU J, et al. Optical and Synthetic Aperture Radar Image Fusion for Ship Detection and Recognition: Current state, challenges, and future prospects[J]. IEEE Geoscience and Remote Sensing Magazine, 2024, 12(4): 132-168. DOI: 10.1109/mgrs.2024.3404506.

[276] DEVARAJ C, SHAH C A. Automated geometric correction of Landsat MSS L1G imagery [J]. IEEE Geoscience and Remote Sensing Letters, 2014, 11(1): 347-351. DOI: 10.1109/lgrs.2013.2257677.

[277] AHAMED T, TIAN L, JIANG Y, et al. Tower remote-sensing system for monitoring energy crops; image acquisition and geometric corrections[J]. Biosystems engineering, 2012, 112(2): 93-107. DOI: 10.1016/j.biosystemseng.2012.03.003.

[278] ZHANG Y, YANG Q. A survey on multi-task learning[J]. IEEE transactions on knowledge and data engineering, 2022, 34(12): 5586-5609. DOI: 10.1109/tkde.2021.3070203.

[279] XIA G S, BAI X, DING J, et al. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images[C]. CVPR. 2018: 3974-3983. DOI: 10.1109/cvpr.2018.00418.

[280] SUN X, WANG P, YAN Z, et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery[J]. ISPRS, 2022, 184: 116-130. DOI: 10.1016/j.isprsjprs.2021.12.004.

[281] LI Y, LI X, LI W, et al. SARDet-100K: Towards Open-Source Benchmark and ToolKit for Large-Scale SAR Object Detection[C]. NeurIPS. 2024.

[282] SUN Y, CAO B, ZHU P, et al. Drone-based RGB-Infrared Cross-Modality Vehicle Detection via Uncertainty-Aware Learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 6700-6713. DOI: 10.1109/tcsvt.2022.3168279.

[283] JAIN Y, BEHL H, KIRA Z, et al. DAMEX: Dataset-aware Mixture-of-Experts for visual understanding of mixture-of-datasets[C]. NeurIPS. 2023: 69625-69637. DOI: 10.52202/075280-3049.

[284] KAPIDIS G, POPPE R, VELTKAMP R C. Multi-dataset, multitask learning of egocentric vision tasks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 6618-6630. DOI: 10.1109/tpami.2021.3061479.

[285] ZHAO X, SCHULTER S, SHARMA G, et al. Object detection with a unified label space from multiple datasets[C]. ECCV. 2020: 178-193. DOI: 10.1007/978-3-030-58568-6_11.

[286] YAN K, CAI J, ZHENG Y, et al. Learning from multiple datasets with heterogeneous and

partial labels for universal lesion detection in CT[J]. IEEE Transactions on Medical Imaging, 2021, 40(10): 2759-2770. DOI: 10.1109/tmi.2020.3047598.

[287] YANG G, MANELA J, HAPPOLD M, et al. Hierarchical deep stereo matching on high-resolution images[C]. CVPR. 2019: 5510-5519. DOI: 10.1109/cvpr.2019.00566.

[288] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. CVPR. 2018: 7132-7141. DOI: 10.1109/cvpr.2018.00745.

[289] SENER O, KOLTUN V. Multi-task learning as multi-objective optimization[C]. NeurIPS. 2018: 525-536.

[290] GUO M, HAQUE A, HUANG D A, et al. Dynamic task prioritization for multitask learning [C]. ECCV. 2018: 282-299. DOI: 10.1007/978-3-030-01270-0_17.

[291] JACOBS R A, JORDAN M I, NOWLAN S J, et al. Adaptive mixtures of local experts[J]. Neural computation, 1991, 3(1): 79-87. DOI: 10.1162/neco.1991.3.1.79.

[292] JACOBS R A, JORDAN M I. Learning piecewise control strategies in a modular neural network architecture[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1993, 23(2): 337-345. DOI: 10.1109/21.229447.

[293] SHAZEER N, MIRHOSEINI A, MAZIARZ K, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer[J]. arXiv, 2017.

[294] CHEN T, CHEN X, DU X, et al. Adamv-moe: Adaptive multi-task vision mixture-of-experts[C]. ICCV. 2023: 17300-17311. DOI: 10.1109/iccv51070.2023.01591.

[295] YANG Y, JIANG P T, HOU Q, et al. Multi-Task Dense Prediction via Mixture of Low-Rank Experts[C]. CVPR. 2024: 27927-27937. DOI: 10.1109/cvpr52733.2024.02638.

[296] JIANG Y, LI X, ZHU G, et al. 6G Non-Terrestrial networks enabled low-altitude economy: Opportunities and challenges[J]. arXiv, 2023.

[297] HUANG C, FANG S, WU H, et al. Low-Altitude Intelligent Transportation: system architecture, infrastructure, and key technologies[J]. Journal of Industrial Information Integration, 2024, 42: 100694. DOI: 10.1016/j.jii.2024.100694.

[298] AVOLA D, CINQUE L, DI MAMBRO A, et al. Low-altitude aerial video surveillance via one-class SVM anomaly detection from textural features in UAV images[J]. Information, 2021, 13(1): 2. DOI: 10.3390/info13010002.

[299] BOZCAN I, KAYACAN E. Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance[C]. 2020 IEEE International Conference on Robotics and Automation. 2020: 8504-8510. DOI: 10.1109/icra40945.2020.9196845.

[300] LI D, WANG M, DONG Z, et al. Earth observation brain (EOB): An intelligent earth observation system[J]. Geo-spatial information science, 2017, 20(2): 134-140. DOI: 10.1080/10095020.2017.1329314.

[301] ANDERSON K, RYAN B, SONNTAG W, et al. Earth observation in service of the 2030 Agenda for Sustainable Development[J]. Geo-spatial Information Science, 2017, 20(2): 77-96. DOI: 10.1080/10095020.2017.1333230.

[302] KHAN A, YANMAZ E, RINNER B. Information merging in multi-UAV cooperative search

[C]. 2014 IEEE international conference on robotics and automation. 2014: 3122-3129. DOI: 10.1109/icra.2014.6907308.

[303] JENSEN O B. Drone city–power, design and aerial mobility in the age of “smart cities” [J]. Geographica Helvetica, 2016, 71(2): 67-75. DOI: 10.5194/gh-71-67-2016.

[304] WANG W, XIE E, LI X, et al. Pvt v2: Improved baselines with pyramid vision transformer [J]. Computational Visual Media, 2022, 8(3): 415-424. DOI: 10.1007/s41095-022-0274-8.

[305] LIN M. Network in network[J]. arXiv, 2013.

[306] NAKANO A, CHEN S, DEMACHI K. Cross-task consistency learning framework for multi-task learning[J]. CoRR, 2021.

[307] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal Loss for Dense Object Detection[C]. ICCV. 2017: 2999-3007. DOI: 10.1109/iccv.2017.324.

[308] DING J, XUE N, LONG Y, et al. Learning RoI Transformer for Oriented Object Detection in Aerial Images[C]. CVPR. 2019: 2844-2853. DOI: 10.1109/cvpr.2019.00296.

[309] HUBEL D H, WIESEL T N. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex[J]. The Journal of physiology, 1962, 160(1): 106-154. DOI: 10.1113/jphysiol.1962.sp006837.

[310] KASTNER S, DE WEERD P, DESIMONE R, et al. Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI[J]. Science, 1998, 282(5386): 108-111. DOI: 10.1126/science.282.5386.108.

[311] RAO R P, BALLARD D H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects[J]. Nature neuroscience, 1999, 2(1): 79-87. DOI: 10.1038/4580.

[312] MUCKLI L, DE MARTINO F, VIZIOLI L, et al. Contextual feedback to superficial layers of V1[J]. Current Biology, 2015, 25(20): 2690-2695. DOI: 10.1016/j.cub.2015.08.057.

[313] XIE Z, ZHANG Z, CAO Y, et al. Simmim: A simple framework for masked image modeling [C]. CVPR. 2022: 9643-9653. DOI: 10.1109/cvpr52688.2022.00943.

[314] LI X, WANG W, YANG L, et al. Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality[J]. CoRR, 2022. DOI: 10.48550/ARXIV.2205.10063.

[315] SU Z, ZHANG J, WANG L, et al. Lightweight pixel difference networks for efficient visual representation learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 14956-14974. DOI: 10.1109/tpami.2023.3300513.

[316] WANG D, ZHANG J, DU B, et al. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model[C]. NeurIPS. 2023: 8815-8827. DOI: 10.52202/075280-0385.

[317] LI Y, LI X, LI W, et al. Sardet-100k: Towards open-source benchmark and toolkit for large-scale sar object detection[C]. NeurIPS. 2024: 128430-128461. DOI: 10.52202/079017-4079.

[318] ZHANG Z, ZHAO T, GUO Y, et al. Rs5m and georsclip: A large scale vision-language

dataset and a large vision-language model for remote sensing[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-23. DOI: 10.1109/tgrs.2024.3449154.

[319] TEAM G, ANIL R, BORGEAUD S, et al. Gemini: a family of highly capable multimodal models[J]. CoRR, 2023. DOI: 10.48550/ARXIV.2312.11805.

[320] TAO C, QI J, ZHANG G, et al. TOV: The original vision model for optical remote sensing image understanding via self-supervised learning[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023: 4916-4930. DOI: 10.1109/JSTARS.2023.3271312.

[321] GUAN D, CAO Y, YANG J, et al. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection[J]. Information Fusion, 2019, 50: 148-157. DOI: 10.1016/j.inffus.2018.11.017.

[322] ZHOU K, CHEN L, CAO X. Improving multispectral pedestrian detection by addressing modality imbalance problems[C]. ECCV. 2020: 787-803. DOI: 10.1007/978-3-030-58523-5_46.

[323] ZHOU M, LI T, QIAO C, et al. Dmm: Disparity-guided multispectral mamba for oriented object detection in remote sensing[J]. IEEE Transactions on Geoscience and Remote Sensing, 2025, 63: 1-13. DOI: 10.1109/tgrs.2025.3578309.

[324] XIAO T, CUI C, ZHU H, et al. Molbind: Multimodal alignment of language, molecules, and proteins[J]. CoRR, 2024. DOI: 10.48550/ARXIV.2403.08167.

[325] DAI S, JIANG S, YANG Y, et al. Babel: A scalable pre-trained model for multi-modal sensing via expandable modality alignment[C]. ACM Conference on Embedded Networked Sensor Systems. 2025: 240-253. DOI: 10.1145/3715014.3722068.

[326] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]. NeurIPS.

[327] CHEN Z, DUAN Y, WANG W, et al. Vision Transformer Adapter for Dense Predictions [C]. ICLR. 2023.

[328] BAI J, BAI S, CHU Y, et al. Qwen technical report[J]. arXiv, 2023.

[329] DAO T, FU D Y, ERMON S, et al. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness[C]. NeurIPS. 2022: 16344-16359. DOI: 10.52202/068431-1189.

[330] CHEN L, LI J, DONG X, et al. Sharegpt4v: Improving large multi-modal models with better captions[C]. ECCV. 2025: 370-387. DOI: 10.1007/978-3-031-72643-9_22.

[331] KAFLE K, PRICE B, COHEN S, et al. Dvqa: Understanding data visualizations via question answering[C]. CVPR. 2018: 5648-5656. DOI: 10.1109/cvpr.2018.00592.

[332] MASRY A, LONG D X, TAN J Q, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning[C]. ACL. 2022: 2263-2279. DOI: 10.18653/V1/2022.FINDINGS-ACL.177.

[333] KEMBHAVI A, SALVATO M, KOLVE E, et al. A diagram is worth a dozen images[C]. ECCV. 2016: 235-251. DOI: 10.1007/978-3-319-46493-0_15.

[334] MATHEW M, KARATZAS D, JAWAHAR C. Docvqa: A dataset for vqa on document images[C]. Winter conference on applications of computer vision. 2021: 2199-2208. DOI:

10.1109/wacv48630.2021.00225.

[335] CAO J, XIAO J. An augmented benchmark dataset for geometric question answering through dual parallel text encoding[C]. International conference on computational linguistics. 2022: 1511-1520.

[336] KIM G, HONG T, YIM M, et al. Ocr-free document understanding transformer[C]. ECCV. 2022: 498-517. DOI: 10.1007/978-3-031-19815-1_29.

[337] ZAVRAS A, MICHAIL D, ZHU X X, et al. GAIA: A global, multi-modal, multi-scale vision-language dataset for remote sensing image analysis[J]. IEEE Geoscience and Remote Sensing Magazine, 2026, 14(2): 36-63. DOI: 10.1109/mgrs.2025.3650613.

[338] LONG Y, XIA G S, LI S, et al. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid[J]. IEEE Journal of selected topics in applied earth observations and remote sensing, 2021, 14: 4205-4230. DOI: 10.1109/jstars.2021.3070368.

[339] LIU C, ZHAO R, CHEN H, et al. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-20. DOI: 10.1109/tgrs.2022.3218921.

[340] PANG C, WENG X, WU J, et al. Vhm: Versatile and honest vision language model for remote sensing image analysis[C]. AAAI. 2025: 6381-6388. DOI: 10.1609/aaai.v39i6.32683.

[341] LOBRY S, MARCOS D, MURRAY J, et al. RSVQA: Visual question answering for remote sensing data[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(12): 8555-8566. DOI: 10.1109/tgrs.2020.2988782.

[342] LUO J, PANG Z, ZHANG Y, et al. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding[J]. CoRR, 2024. DOI: 10.48550/ARXIV.2406.10100.

[343] ISPRS. Classification of Multiscale Marine Phenomenon in SAR Images[Z]. isprs2024tc1. Accessed: 2025-08-22. 2025.

[344] TIWARI P. Sentinel-1&2 Image Pairs (SAR & Optical)[Z]. sentinel12-image-pairs-segregated-by-terrain/code?datasetId=1201791. Accessed: 2025-08-22. 2025.

[345] KUCKREJA K, DANISH M S, NASEER M, et al. Geochat: Grounded large vision-language model for remote sensing[C]. CVPR. 2024: 27831-27840. DOI: 10.1109/cvpr52733.2024.02629.

[346] ZHAN Y, XIONG Z, YUAN Y. Rsvg: Exploring data and models for visual grounding on remote sensing data[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-13. DOI: 10.1109/tgrs.2023.3250471.

[347] SUN Y, FENG S, LI X, et al. Visual grounding in remote sensing images[C]. ACMMM. 2022: 404-412. DOI: 10.1145/3503161.3548316.

[348] LI X, DING J, ELHOSEINY M. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding[C]. NeurIPS. 2024: 3229-3242. DOI: 10.52202/

079017-0106.

[349] GAO Z, CHEN Z, CUI E, et al. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance[J]. Visual Intelligence, 2024, 2(1). DOI: 10.1007/s44267-024-00067-6.

[350] ZHOU B, ZHAO H, PUIG X, et al. Scene parsing through ade20k dataset[C]. CVPR. 2017: 5122-5130. DOI: 10.1109/cvpr.2017.544.

[351] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]. ICCV. 2017: 2980-2988. DOI: 10.1109/iccv.2017.322.

[352] LI Y, LI X, LI Y, et al. SM3Det: A Unified Model for Multi-Modal Remote Sensing Object Detection[C]. AAAI. 2026: 6717-6725. DOI: 10.1609/aaai.v40i8.37603.

[353] SU Z, LIU L, MÜLLER M, et al. Rapid Salient Object Detection With Difference Convolutional Neural Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(10): 9061-9077. DOI: 10.1109/tpami.2025.3583968.

[354] ZHOU J, LIU Y, PENG B, et al. MaDiNet: Mamba Diffusion Network for SAR Target Detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(11): 10787-10800. DOI: 10.1109/tcsvt.2025.3574657.

[355] CHENG G, WANG J, LI K, et al. Anchor-free oriented proposal generator for object detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-11. DOI: 10.1109/tgrs.2022.3183022.

[356] ZHANG X, YANG X, LI Y, et al. Rsar: Restricted state angle resolver and rotated sar benchmark[C]. CVPR. 2025: 7416-7426. DOI: 10.1109/cvpr52734.2025.00695.

[357] CAI Z, VASCONCELOS N. Cascade R-CNN: Delving Into High Quality Object Detection [C]. CVPR. 2018: 6154-6162. DOI: 10.1109/cvpr.2018.00644.

[358] WAQAS ZAMIR S, ARORA A, GUPTA A, et al. isaid: A large-scale dataset for instance segmentation in aerial images[C]. CVPRW. 2019: 28-37.

[359] LEBEDEV M, VIZILTER Y V, VYGOLOV O, et al. Change detection in remote sensing images using conditional adversarial networks[J]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2018, XLII-2: 565-571. DOI: 10.5194/isprs-archives-xlii-2-565-2018.

[360] JI S, WEI S, LU M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(1): 574-586. DOI: 10.1109/tgrs.2018.2858817.

[361] CHEN H, SHI Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection[J]. Remote sensing, 2020, 12(10): 1662. DOI: 10.3390/rs12101662.

[362] TOUVRON H, CORD M, JÉGOU H. Deit iii: Revenge of the vit[C]. ECCV. 2022: 516-533. DOI: 10.1007/978-3-031-20053-3_30.

[363] CHEN* X, XIE* S, HE K. An Empirical Study of Training Self-Supervised Vision Transformers[C]. ICCV. 2021: 9620-9629. DOI: 10.1109/iccv48922.2021.00950.

[364] BAO H, DONG L, PIAO S, et al. Beit: Bert pre-training of image transformers[C]. ICLR. 2022.

[365] MIZRAHI D, BACHMANN R, KAR O, et al. 4m: Massively multimodal masked modeling [J]. NeurIPS, 2023, 36: 58363-58408.

[366] ZHAI X, MUSTAFA B, KOLESNIKOV A, et al. Sigmoid loss for language image pre-training[C]. ICCV. 2023: 11941-11952. DOI: 10.1109/iccv51070.2023.01100.

[367] TSCHANNEN M, GRITSENKO A, WANG X, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features[J]. CoRR, 2025. DOI: 10.48550/ARXIV.2502.14786.

[368] LI B, ZHANG Y, GUO D, et al. Llava-onevision: Easy visual task transfer[J]. arXiv, 2025.

[369] CHEN Y, MENG L, PENG W, et al. Comp: Continual multimodal pre-training for vision foundation models[J]. arXiv, 2025. DOI: 10.48550/ARXIV.2503.18931.

[370] FINI E, SHUKOR M, LI X, et al. Multimodal autoregressive pre-training of large vision encoders[C]. CVPR. 2025: 9641-9654. DOI: 10.1109/cvpr52734.2025.00901.

[371] AYUSH K, UZKENT B, MENG C, et al. Geography-aware self-supervised learning[C]. ICCV. 2021: 10161-10170. DOI: 10.1109/iccv48922.2021.01002.

[372] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]. NeurIPS: vol. 39: 6. 2017: 1137-1149. DOI: 10.1109/tpami.2016.2577031.

[373] ZHANG S, CHI C, YAO Y, et al. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection[C]. CVPR. 2020: 9756-9765. DOI: 10.1109/cvpr42600.2020.00978.

[374] WANG Y, BRAHAM N A A, XIONG Z, et al. SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in Earth observation [Software and Data Sets][J]. IEEE Geoscience and Remote Sensing Magazine, 2023, 11(3): 98-106. DOI: 10.1109/mgrs.2023.3281651.

[375] MUHTAR D, ZHANG X, XIAO P, et al. Cmid: A unified self-supervised learning framework for remote sensing image understanding[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-17. DOI: 10.1109/tgrs.2023.3268232.

[376] WANG D, ZHANG Q, XU Y, et al. Advancing Plain Vision Transformer Toward Remote Sensing Foundation Model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-15. DOI: 10.1109/tgrs.2022.3222818.

[377] YANG X, YANG X, YANG J, et al. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence[C]. NeurIPS. 2021: 18381-18394.

[378] BASTANI F, WOLTERS P, GUPTA R, et al. Satlaspretrain: A large-scale dataset for remote sensing image understanding[C]. ICCV. 2023: 16726-16736. DOI: 10.1109/iccv51070.2023.01538.

[379] LI W, CHEN Y, HU K, et al. Oriented reppoints for aerial object detection[C]. CVPR. 2022: 1819-1828. DOI: 10.1109/cvpr52688.2022.00187.

[380] DING J, XUE N, LONG Y, et al. Learning RoI transformer for oriented object detection in aerial images[C]. CVPR. 2019: 2844-2853. DOI: 10.1109/cvpr.2019.00296.

[381] LI Z, HOU B, MA S, et al. Masked angle-aware autoencoder for remote sensing images[C]. ECCV. 2025: 260-278. DOI: 10.1007/978-3-031-73242-3_15.

[382] HUANG Z, LI W, XIA X G, et al. A general Gaussian heatmap label assignment for arbitrary-oriented object detection[J]. IEEE Transactions on Image Processing, 2022, 31: 1895-1910. DOI: 10.1109/tip.2022.3148874.

[383] WANG F, WANG H, WANG D, et al. Harnessing massive satellite imagery with efficient masked image modeling[C]. ICCV. 2025: 6935-6947. DOI: 10.1109/iccv51701.2025.00652.

[384] XU C, DING J, WANG J, et al. Dynamic coarse-to-fine learning for oriented tiny object detection[C]. CVPR. 2023: 7318-7328. DOI: 10.1109/cvpr52729.2023.00707.

[385] CHA K, SEO J, LEE T. A billion-scale foundation model for remote sensing images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024: 1-17. DOI: 10.1109/jstars.2024.3401772.

[386] LAM D, KUZMA R, MCGEE K, et al. xview: Objects in context in overhead imagery[J]. CoRR, 2018.

[387] KHANAM R, HUSSAIN M. Yolov11: An overview of the key architectural enhancements [J]. CoRR, 2024. DOI: 10.48550/ARXIV.2410.17725.

[388] YAO T, ZHANG Y, QIU Z, et al. Seco: Exploring sequence supervision for unsupervised representation learning[C]. AAAI. 2021: 10656-10664. DOI: 10.1609/aaai.v35i12.17274.

[389] WANG D, ZHANG J, XU M, et al. MTP: Advancing remote sensing foundation model via multi-task pretraining[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, 17: 11632-11654. DOI: 10.1109/jstars.2024.3408154.

[390] ZHANG M, LIU Q, WANG Y. CtxMIM: Context-enhanced masked image modeling for remote sensing image understanding[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2025, 21(12): 1-22. DOI: 10.1145/3769084.

[391] ZENG Y, CHEN Y, YANG X, et al. ARS-DETR: Aspect Ratio-Sensitive Detection Transformer for Aerial Oriented Object Detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-15. DOI: 10.1109/tgrs.2024.3364713.

[392] LU X, LI B, YUE Y, et al. Grid r-cnn[C]. CVPR. 2019: 7355-7364. DOI: 10.1109/cvpr.2019.00754.

[393] LI Q, CHEN Y, SHU X, et al. A simple aerial detection baseline of multimodal language models[C]. IEEE International Geoscience and Remote Sensing Symposium. 2025: 6833-6837. DOI: 10.1109/igarss55030.2025.11242725.

[394] HAN J, DING J, XUE N, et al. Redet: A rotation-equivariant detector for aerial object detection[C]. CVPR. 2021: 2785-2794. DOI: 10.1109/cvpr46437.2021.00281.

[395] DAI Y, ZOU M, LI Y, et al. Denodet: Attention as deformable multi-subspace feature denoising for target detection in sar images[J]. IEEE Transactions on Aerospace and Electronic

Systems, 2025, 61(2): 4729-4743. DOI: 10.1109/taes.2024.3507786.

[396] NI K, ZOU M, LI Y, et al. DenoDet V2: Phase-Amplitude Cross Denoising for SAR Object Detection[C]. AAAI. 2026: 8142-8150. DOI: 10.1609/aaai.v40i10.37761.

[397] LI W, YANG W, HOU Y, et al. SARATR-X: Towards building a foundation model for SAR target recognition[J]. IEEE Transactions on Image Processing, 2025, 34: 869-884. DOI: 10.1109/tip.2025.3531988.

[398] RAO Y, ZHAO W, CHEN G, et al. Denseclip: Language-guided dense prediction with context-aware prompting[C]. CVPR. 2022: 18061-18070. DOI: 10.1109/cvpr52688.2022.01755.

[399] ZHANG H, LI F, XU H, et al. Mp-former: Mask-piloted transformer for image segmentation [C]. CVPR. 2023.

[400] LU W, CHEN S B, SHU Q L, et al. DecoupleNet: A Lightweight Backbone Network With Efficient Feature Decoupling for Remote Sensing Visual Tasks[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-13. DOI: 10.1109/tgrs.2024.3465496.

[401] XU R, WANG C, ZHANG J, et al. RSSFormer: Foreground Saliency Enhancement for Remote Sensing Land-Cover Segmentation[J]. IEEE Transactions on Image Processing, 2023, 32: 1052-1064. DOI: 10.1109/tip.2023.3238648.

[402] CHENG B, MISRA I, SCHWING A G, et al. Masked-attention mask transformer for universal image segmentation[C]. CVPR. 2022: 1280-1289. DOI: 10.1109/cvpr52688.2022.00135.

[403] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]. ECCV. 2018: 833-851. DOI: 10.1007/978-3-030-01234-2_49.

[404] YUAN Y, CHEN X, WANG J. Object-contextual representations for semantic segmentation [C]. ECCV. 2020: 173-190. DOI: 10.1007/978-3-030-58539-6_11.

[405] TIAN Z, SHEN C, WANG X, et al. Boxinst: High-performance instance segmentation with box annotations[C]. CVPR. 2021: 5439-5448. DOI: 10.1109/cvpr46437.2021.00540.

[406] CHENG B, MISRA I, SCHWING A G, et al. Masked-attention mask transformer for universal image segmentation[C]. CVPR. 2022: 1280-1289. DOI: 10.1109/cvpr52688.2022.00135.

[407] FANG H S, SUN J, WANG R, et al. Instaboost: Boosting instance segmentation via probability map guided copy-pasting[C]. ICCV. 2019: 682-691. DOI: 10.1109/iccv.2019.00077.

[408] TIAN Z, SHEN C, CHEN H. Conditional convolutions for instance segmentation[C]. ECCV. 2020: 282-298. DOI: 10.1007/978-3-030-58452-8_17.

[409] LIU Y, LI H, HU C, et al. Catnet: context aggregation network for instance segmentation in remote sensing images[J]. CoRR, 2021.

[410] SU H, WEI S, LIU S, et al. HQ-ISNet: High-quality instance segmentation for remote sensing imagery[J]. Remote Sensing, 2020, 12(6): 989. DOI: 10.3390/rs12060989.

[411] CHEN K, LIU C, CHEN H, et al. RSPrompter: Learning to prompt for remote sensing in-

stance segmentation based on visual foundation model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-17. DOI: 10.1109/tgrs.2024.3356074.

[412] VU T, KANG H, YOO C D. Scnet: Training inference sample consistency for instance segmentation[C]. AAAI. 2021: 2701-2709. DOI: 10.1609/aaai.v35i3.16374.

[413] NOMAN M, NASEER M, CHOLAKKAL H, et al. Rethinking transformers pre-training for multi-spectral satellite imagery[C]. CVPR. 2024: 27811-27819. DOI: 10.1109/cvpr52733.2024.02627.

[414] ZHENG Z, ERMON S, KIM D, et al. Changen2: Multi-temporal remote sensing generative change foundation model[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(2): 725-741. DOI: 10.1109/tpami.2024.3475824.

[415] ZHENG Z, WAN Y, ZHANG Y, et al. CLNet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery[J]. ISPRS Journal of the Photogrammetry and Remote Sensing, 2021, 175: 247-267. DOI: 10.1016/j.isprsjprs.2021.03.005.

[416] LIU M, SHI Q, MARINONI A, et al. Super-resolution-based change detection network with stacked attention module for images with different resolutions[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-18. DOI: 10.1109/tgrs.2021.3091758.

[417] WEN Y, MA X, ZHANG X, et al. GCD-DDPM: A generative change detection model based on difference-feature-guided DDPM[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-16. DOI: 10.1109/tgrs.2024.3381752.

[418] WANG J, ZHONG Y, ZHANG L. Change detection based on supervised contrastive learning for high-resolution remote sensing imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-16. DOI: 10.1109/tgrs.2023.3236664.

[419] BANDARA W G C, NAIR N G, PATEL V M. DDPM-CD: Denoising diffusion probabilistic models as feature extractors for remote sensing change detection[C]. IEEE/CVF Winter Conference on Applications of Computer Vision. 2025: 5250-5262. DOI: 10.1109/WACV61041.2025.00513.

[420] LI X, YAN L, ZHANG Y, et al. ESR-DMNet: Enhanced super-resolution-based dual-path metric change detection network for remote sensing images with different resolutions[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-15. DOI: 10.1109/tgrs.2024.3362895.

[421] ZHANG H, CHEN H, ZHOU C, et al. Bifa: Remote sensing image change detection with bitemporal feature alignment[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-17. DOI: 10.1109/tgrs.2024.3376673.

[422] ZHAO S, ZHANG X, XIAO P, et al. Exchanging dual-encoder–decoder: A new strategy for change detection with semantic guidance and spatial localization[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-16. DOI: 10.1109/tgrs.2023.3327780.

[423] GUO H, SU X, WU C, et al. SAAN: Similarity-aware attention flow network for change detection with VHR remote sensing images[J]. IEEE Transactions on Image Processing, 2024, 33: 2599-2613. DOI: 10.1109/tip.2024.3349868.

[424] MOHAMMADIAN A, GHADERI F. SiamixFormer: A fully-transformer Siamese network with temporal fusion for accurate building detection and change detection in bi-temporal remote sensing images[J]. International Journal of Remote Sensing, 2023, 44(12): 3660-3678. DOI: 10.1080/01431161.2023.2225228.

[425] LI Q, ZHONG R, DU X, et al. TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-19. DOI: 10.1109/tgrs.2022.3169479.

[426] CHEN H, PU F, YANG R, et al. RDP-Net: Region Detail Preserving Network for Change Detection.[J]. arXiv, 2022.

[427] LIU J, XUAN W, GAN Y, et al. An end-to-end supervised domain adaptation framework for cross-domain change detection[J]. Pattern Recognition, 2022, 132: 108960. DOI: 10.1016/j.patcog.2022.108960.

[428] TANG X, ZHANG T, MA J, et al. WNet: W-shaped hierarchical network for remote-sensing image change detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-14. DOI: 10.1109/tgrs.2023.3296383.

[429] CHEN H, SONG J, HAN C, et al. ChangeMamba: Remote sensing change detection with spatiotemporal state space model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-20. DOI: 10.1109/tgrs.2024.3417253.

[430] ZHAO S, CHEN H, ZHANG X, et al. Rs-mamba for large remote sensing image dense prediction[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-14. DOI: 10.1109/tgrs.2024.3425540.

[431] ZHANG H, CHEN K, LIU C, et al. CDMamba: Incorporating local clues into mamba for remote sensing image binary change detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2025, 63: 1-16. DOI: 10.1109/tgrs.2025.3545012.

[432] DONG S, WANG L, DU B, et al. ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning[J]. ISPRS Journal of the Photogrammetry and Remote Sensing, 2024, 208: 53-69. DOI: 10.1016/j.isprsjprs.2024.01.004.

[433] LIN M, YANG G, ZHANG H. Transition is a process: Pair-to-video change detection networks for very high resolution remote sensing images[J]. IEEE Transactions on Image Processing, 2023, 32: 57-71. DOI: 10.1109/tip.2022.3226418.

[434] LI X, TAO Y, ZHANG S, et al. REOBench: Benchmarking Robustness of Earth Observation Foundation Models[J]. CoRR, 2025. DOI: 10.48550/ARXIV.2505.16793.

[435] DARCET T, OQUAB M, MAIRAL J, et al. Vision transformers need registers[J]. arXiv, 2024.

[436] JIANG N, DRAVID A, EFROS A, et al. Vision Transformers Don't Need Trained Registers [J]. arXiv, 2025. DOI: 10.48550/ARXIV.2506.08010.

[437] WEI Y, XIAO A, REN Y, et al. SARLANG-1M: A Benchmark for Vision-Language Modeling in SAR Image Understanding[J]. IEEE Transactions on Geoscience and Remote Sensing, 2026, 64: 1-20. DOI: 10.1109/tgrs.2026.3652099.

[438] ZHANG W, CAI M, ZHANG T, et al. EarthGPT: A universal multimodal large language model for multisensor image comprehension in remote sensing domain[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-20. DOI: 10.1109/tgrs.2024.3409624.

[439] BAI J, BAI S, CHU Y, et al. Qwen technical report[J]. arXiv, 2023.

[440] BOLYA D, HUANG P Y, SUN P, et al. Perception encoder: The best visual embeddings are not at the output of the network[J]. arXiv, 2025. DOI: 10.48550/ARXIV.2504.13181.

[441] PENG Z, DONG L, BAO H, et al. Beit v2: Masked image modeling with vector-quantized visual tokenizers[J]. CoRR, 2022. DOI: 10.48550/ARXIV.2208.06366.

作者简介及攻读博士期间研究成果

作者简介

攻读博士学位期间，在相关领域共发表或录用学术论文 18 篇。其中，以第一作者身份在 IJCV、NeurIPS、ICML、ICCV、AAAI 等计算机视觉与人工智能领域的国际顶级期刊和会议（CCF-A 类）上发表论文 5 篇（含一篇 ESI 高被引/热点论文，一篇 NeurIPS Spotlight 与一篇 AAAI Oral 报告）。此外，在 CVPR、ICLR、IEEE TGRS 等顶级学术会议及期刊上合作发表多篇高水平论文。相关研究成果在学术界与工业界产生了广泛影响，谷歌学术（Google Scholar）总引用量突破 1800 次，其中最高单篇论文引用量超过 1000 次。

在学术研究之余，积极投身开源社区建设。GitHub 公开代码仓库累计获得 Star 数超过 2000 个。作为国产开源框架 NK-JittorCV 的核心贡献者，其相关工作已被广泛应用于遥感图像分析、目标检测与图像分类等实际任务中。此外，博士期间先后参与 8 次国内外算法大赛，以第一完成人身份斩获 2022 年第二届 Jittor 人工智能挑战赛冠军，以及 2022 年首届粤港澳大湾区国际算法算例大赛亚军。同时，致力于推动该领域学术评测基准的发展，受邀担任 PRCV 2024 SARDet 竞赛，以及 2025 第七届全球校园人工智能算法精英大赛（大规模 SAR 图像多类别有向目标检测算法）赛题出题人。

发表/录用论文

[1] **Yuxuan Li (李宇轩)**，Qibin Hou，Zhaohui Zheng，Ming-Ming Cheng，Jian Yang，and Xiang Li. “Large selective kernel network for remote sensing object detection.” In Proceedings of the IEEE/CVF international conference on computer vision. 2023. (ICCV，CCF-A，Google 学术引用 1000 余次)

[2] **Yuxuan Li (李宇轩)**，Xiang Li，Yimain Dai，Qibin Hou，Li Liu，Yongxiang Liu，Ming-Ming Cheng，and Jian Yang. “LSKNet: A Foundation Lightweight Backbone for Remote Sensing.” International Journal of Computer Vision，2025:

1410-1431. (IJCV，CCF-A，Google 学术引用 170 余次)

[3] **Yuxuan Li (李宇轩)**, Xiang Li, Weijie Li, Qibin Hou, Li Liu, Ming-Ming Cheng, and Jian Yang. “Sardet-100k: Towards open-source benchmark and toolkit for large-scale sar object detection.” Advances in Neural Information Processing Systems. 2024: 128430-128461. (NeurIPS，CCF-A，Google 学术引用 130 余次，NeurIPS Spotlight 论文)

[4] **Yuxuan Li (李宇轩)**, Yuming Chen, Yunheng Li, Ming-Ming Cheng, Xiang Li and Jian Yang. “Unifying Heterogeneous Multi-Modal Remote Sensing Detection Via Language-Pivoted Pretraining.” International Conference on Machine Learning. 2026. (ICML，CCF-A，已接收.)

[5] **Yuxuan Li (李宇轩)**, Xiang Li, Yunheng Li, Yicheng Zhang, Yimian Dai, Qibin Hou，Ming-Ming Cheng，and Jian Yang. “Sm3det: A unified model for multi-modal remote sensing object detection.” In Proceedings of the AAAI Conference on Artificial Intelligence. 2026. (AAAI，CCF-A，Google 学术引用 30 余次，AAAI Oral 论文)

[6] **Yuxuan Li (李宇轩)**，Lingfeng Yang，and Xiang Li. “APF-GAN: Exploring asymmetric pre-training and fine-tuning strategy for conditional generative adversarial network.” Computational Visual Media. 2024 (CVM，CCF-B，期刊影响因子 18.3)

[7] Chenxu Wang，**Yuxuan Li (李宇轩)**，Yunheng Li，Xiang Li，Jingyuan Xia, and Qibin Hou. “ SLIP-RS: Structured-Attribute Language-Image Pre-Training for Remote Sensing Object Detection.” International Conference on Machine Learning. 2026. (ICML，CCF-A，已接收.)

[8] Yunheng Li，**Yuxuan Li (李宇轩)**，Quansheng Zeng，Wenhai Wang，Qibin Hou, and Ming-Ming Cheng. “ Unbiased Region-Language Alignment for Open-Vocabulary Dense Prediction” In Proceedings of the IEEE/CVF international conference on computer vision. 2025. (ICCV，CCF-A)

[9] Xin Zhang, Xue Yang, **Yuxuan Li (李宇轩)**, Jian Yang, Ming-Ming Cheng, and

Xiang Li. “Rsar: Restricted state angle resolver and rotated sar benchmark.” In Proceedings of the Computer Vision and Pattern Recognition Conference. 2025. (CVPR，CCF-A)

[10] Chenxu Wang，Chunyan Xu，Xiang Li，**Yuxuan Li (李宇轩)**，Xu Guo，Ziqi Gu，and Zhen Cui. “Multi-clue consistency learning to bridge gaps between general and oriented object in semi-supervised detection.” In Proceedings of the AAAI Conference on Artificial Intelligence. 2025. (AAAI，CCF-A)

[11] Shengdong Han，Shangdong Yang，Xin Zhang，**Yuxuan Li (李宇轩)**，Xiang Li，Jian Yang，Ming-Ming Cheng，and Yimian Dai. “DISTA-Net: Dynamic Closely-Spaced Infrared Small Target Unmixing.” In Proceedings of the IEEE/CVF international conference on computer vision. 2025. (ICCV，CCF-A)

[12] Qun Dai，Chunyang Yuan，Yimian Dai，**Yuxuan Li (李宇轩)**，Xiang Li，Kang Ni，Jianhui Xu，Xiangbo Shu，and Jian Yang. “MoCoLSK: Modality Conditioned High-Resolution Downscaling for Land Surface Temperature.” IEEE Transactions on Geoscience and Remote Sensing. 2025.

[13] Yimian Dai，Peiwen Pan，Yulei Qian，**Yuxuan Li (李宇轩)**，Xiang Li，Jian Yang，and Huan Wang. “Pick of the bunch: Detecting infrared small targets beyond hit-miss trade-offs via selective rank-aware attention.” IEEE Transactions on Geoscience and Remote Sensing. 2024.

[14] Yimian Dai，Minrui Zou，**Yuxuan Li (李宇轩)**，Xiang Li，Kang Ni，and Jian Yang. “Denodet: Attention as deformable multi-subspace feature denoising for target detection in sar images.” IEEE Transactions on Aerospace and Electronic Systems. 2024.

[15] Weijie Li，Wei Yang，Tianpeng Liu，Yuenan Hou，**Yuxuan Li (李宇轩)**，Zhen Liu，Yongxiang Liu，and Li Liu. “Predicting gradient is better: Exploring self-supervised learning for SAR ATR with a joint-embedding predictive architecture.” ISPRS Journal of Photogrammetry and Remote Sensing. 2024.

[16] Ni，Kang，Minrui Zou，**Yuxuan Li (李宇轩)**，Xiang Li，Kehua Guo，Ming-Ming

Cheng，and Yimian Dai. “DenoDet V2: Phase-Amplitude Cross Denoising for SAR Object Detection.” In Proceedings of the AAAI Conference on Artificial Intelligence. 2026. (AAAI，CCF-A)

[17] Yuan，Xinbin，Zhaohui Zheng，**Yuxuan Li (李宇轩)**，Xialei Liu，Li Liu，Xiang Li，Qibin Hou，and Ming-Ming Cheng. “Strip R-CNN: Large strip convolution for remote sensing object detection.” In Proceedings of the AAAI Conference on Artificial Intelligence. 2026. (AAAI，CCF-A)

[18] Danyang Li，Tianhao Wu，Bin Lin，Zhenyuan Chen，Yang Zhang，**Yuxuan Li (李宇轩)**，Ming-Ming Cheng，Xiang Li. “WOW-Seg: A Word-free Open World Segmentation Model.” The International Conference on Learning Representations. 2026. (ICLR，CCF-A)

竞赛及获奖荣誉

[1] 博士生国家奖学金，2025

[2] 第二届“吉林一号”杯卫星遥感应用青年创新创业大赛“高分辨率遥感数据道路提取研究”赛题，三等奖，2025

[3] 第二届“吉林一号”杯卫星遥感应用青年创新创业大赛“基于卫星影像的数字产品设计”赛题，三等奖，2025

[4] 南开大学 2023-2024 学年度研究生公能奖学金，2024

[5] 第二届“吉林一号”杯卫星遥感应用青年创新创业大赛“基于高分辨率卫星影像的耕地变化检测”赛题，三等奖，2024

[6] 第三届 Jittor 人工智能挑战赛，三等奖，2023

[7] 第三届中国图象图形学报研究生学术论坛，突出报告奖，2023

[8] 第二届 Jittor 人工智能挑战赛，冠军，2022

[9] 首届粤港澳大湾区国际算法算例大赛，亚军，2022

[10] 第五届开源创新大赛-开源任务挑战赛，团体一等奖，2022

参与项目

[1] 国家自然科学基金联合基金重点项目（U24A20330）

[2] 国家自然科学基金国际 (地区) 合作与交流项目（62361166670）

[3] 国家杰出青年科学基金（62206134）

[4] 国家自然科学基金面上项目（62225604）

学术服务

[1] 会议审稿人, NeurIPS, 2023/2024/2025

[2] 会议审稿人, International Conference on Machine Learning (ICML), 2026

[3] 会议审稿人, International Conference on Learning Representations (ICLR), 2025

[4] 会议审稿人, AAAI Conference on Artificial Intelligence (AAAI), 2025

[5] 会议审稿人, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024/2025

[6] 会议审稿人, IEEE/CVF International Conference on Computer Vision (ICCV), 2023/2024

[7] 会议审稿人, European Conference on Computer Vision (ECCV), 2024

[8] 会议审稿人, Pattern Recognition and Computer Vision (PRCV), 2025

[9] 期刊审稿人, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)

[10] 期刊审稿人, International Journal of Computer Vision (IJCV)

[11] 期刊审稿人, ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS)

[12] 期刊审稿人, IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)

[13] 期刊审稿人, IEEE Transactions on Geoscience and Remote Sensing (TGRS)

致谢

文末搁笔，思绪如潮。溯往昔，十七岁负笈南洋，后转徙英伦，游学数载，终得归棹国门，入南开攻读博士学位。岁月如淘，霜雪染履，所幸赤子之心未泯，少年之诚犹存。回望此番学术行旅，论文所承载者，非独技术难题之求解，实乃凝结师长之化育、同道之扶持、家人之守望，以及于困顿、试错与坚持中所获之成长与蜕变。

初入南开曾历两载蛰伏，实乃学术生涯之阴翳时刻。彼时转向易辙，步履维艰，对坐同窗，常有难望项背之惶。挣扎徘徊之际，幸得杨健恩师赐予明灯，李翔恩师携我披荆斩棘，方于学术之途觅得立锥之地。程明明恩师待我视如己出，扫平横逆、鼎力襄助，得蒙诸位前辈亲炙，实乃三生之幸。自入组以来，诸位恩师始终以谨严之治学态度、宏阔之研究视野、高度之责任之心引我前行。方向之抉择、选题之凝练，抑或关键之推敲、文辞之打磨，皆倾注耐心，悉心指授。先生们非独授我以研学之法，更教我以诚实、克制且坚定之姿态，直面科研之未知与漫长。每逢迷惘迟疑，导师之点拨总能令我重拾信念、校准航向。得蒙悉心栽培，诚为莫大之幸，亦将化作我未来学术道路上之不绝资粮。同谢 PCALab 与 ReductLab 之诸位老师同窗。博士岁月皆赖诸君陪伴与扶持。讨论时之思想碰撞、合作中之默契相投、撰稿时之反复切磋，皆令我真切体悟集体之暖意与力量。感谢同门师友于研究思路、工程实现与论文反馈上所倾注之援手。与诸位并肩同行之日，令孤寂漫长之科研长路，平添诸多踏实与明亮。

至深之感谢，当予吾之家人。叩谢父母始终如一之理解、包容与守护。二老或许未必尽知余所研之学理，却始终无条件地信我、持我；于余焦虑时予慰藉，于余疲惫时赐力量。兼谢族中尊长及诸位亲朋，多年来厚爱深情，始终如一。木有根而枝茂，亲族之和睦与期许，令余深感根脉之所系，虽身在远方而心有归处。亦感谢怿彤妹妹长情陪伴，隔屏相望之每一次笑语，皆为心灵之疗愈。家人之爱，乃余抵御千钧重压之底气，亦是余跬步不休之动力。

回顾数载求学之路，亦深感造化之垂怜与命运之眷顾。一路行来，虽经坎坷，然每至绝境，总能逢山开路，遇水搭桥，履险如夷。待行过万重深壑，轻舟终抵彼岸，回首是云开月明，万象更新；前路是朝阳映雪，坦途如砥。恰于此时，雨点不期而至，如甘霖骤降，沁人心脾，洗却经年铅华，乃是岁月之褒奖，令过往之困顿，尽化作笔底之余香。何其有幸，能于岁熟丰登之时，承此甘霖之恩，此段学旅，由是无憾。

末了，当谢一路前行之自己。谢自己于受挫之时未曾轻言放弃，于结果反复之际仍愿耐心求证。于漫长孤独之积淀中，依然葆有对未知之好奇、对学术之敬畏与对未来之期许。博士生涯终将落幕，然其所馈赠者，非止一纸学位，而是一种更为沉静、坚韧之心性。未来无论身处何方，皆当珍视此段岁月之磨砺，继续以认真、谦逊与热爱之姿，奔赴下一程山海。

谨以此文，献给所有关心、帮助与陪伴过我的人。