

中图法分类号: TP391.4 文献标识码: A 文章编号: 1006-8961(XXXX)XX-0001-32

论文引用格式: Chen Yu, Lu Jialiang, Li Xuewei, Li Xi. General Intelligence-Driven RGB-X Fusion: A Survey[J/OL]. Journal of Image and Graphics, XXXX, 1-32. DOI: 10.11834/jig.260229. (陈瑜, 卢嘉亮, 励雪巍, 李玺. 基于通用智能的RGB-X融合技术研究综述[J/OL]. 中国图象图形学报, XXXX, 1-32. DOI: 10.11834/jig.260229.) [DOI: 10.11834/jig.260229]

基于通用智能的RGB-X融合技术研究综述

陈瑜¹, 卢嘉亮², 励雪巍^{3*}, 李玺¹

1. 浙江大学计算机科学与技术学院, 浙江省杭州市 310013; 2. 浙江工商大学计算机科学与技术学院, 浙江省杭州市 310018; 3. 上海电机学院人工智能与数据科学学院, 上海 201306

摘要: 在自动驾驶、机器人和医疗检测等应用中, 单一RGB模态在夜间、低照度以及遮挡等环境下存在显著性能瓶颈, 限制其在实际场景中的鲁棒性。红外、深度、激光雷达等异源信息的引入有效缓解了这一问题, 但如何高效集成RGB与异源成分, 同时滤除无关信息, 成为限制RGB-X跨模态交互的关键。现有融合技术主要采用显式交互的方式, 即通过元素加、桥接或外接复杂融合模块的方式进行异构模态集成。近年来, 视觉基础模型(Vision Foundation Models, VFMs)和视觉语言模型(Vision Language Models, VLMs)等基础模型的出现, 为RGB-X融合技术提供了新的研究思路。相较于现有融合方法, 这类基础模型凭借较强的通用表征、跨模态理解与知识迁移能力, 为专用融合模型提供了新的设计策略与能力增强方式, 进一步拓展了RGB-X跨模态融合的研究路径。本文围绕这一前沿方向, 对基础模型驱动下的RGB-X跨模态交互技术进行系统综述。首先, 梳理RGB-X感知与融合技术的基本原理; 其次, 基于融合过程中信息交互所处的表示层级, 分别从模态级融合、特征级融合及表征级融合三个层面出发, 深入剖析现有通专结合的融合方法及各自技术特点; 随后, 在公开数据集上对代表性方法进行性能对比分析, 并讨论上述融合技术在自动驾驶、智能机器人、医疗影像等核心下游感知领域中的应用进展; 最后, 总结当前技术面临的挑战并展望未来发展方向。区别于现有综述主要关注传统显式融合策略, 本文聚焦通用智能驱动的RGB-X融合新范式, 系统梳理相关技术路线、代表方法及应用进展, 旨在为大模型时代的RGB-X跨模态交互研究提供技术参考与发展思路。本文的参考文献及文中涉及的算法、数据集与评估指标, 已整理汇总于<https://github.com/yuyuxiazhu-pixel/RGB-X-fusion-technologies>。

关键词: 自动驾驶; 跨模态交互; 跨模态感知; 视觉基础模型; RGB-X融合

General Intelligence-Driven RGB-X Fusion: A Survey

Chen Yu¹, Lu Jialiang², Li Xuewei^{3*}, Li Xi¹

1. College of Computer Science and Technology, Zhejiang University, Hangzhou 310013, China; 2. School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou 310018, China; 3. School of Artificial Intelligence and Data Science, Shanghai Dianji University, Shanghai 201306, China

Abstract: In applications such as autonomous driving, robotics, industrial inspection, and medical imaging, single RGB

收稿日期: 2026-04-21; 修回日期: 2026-09-22

* 通信作者: 励雪巍 xueweili@sdju.edu.cn

基金项目: 浙江省自然科学基金重大项目(LD24F020016); 国家杰出青年科学基金(62225605); 国家自然科学基金(62506225); 科创甬江2035(2025Z028); 上海市自然科学基金(24ZR1425600)

Supported by: Zhejiang Provincial Natural Science Foundation of China (LD24F020016); National Science Foundation for Distinguished Young Scholars (62225605); National Natural Science Foundation of China (62506225); Ningbo Science and Technology Special Projects (2025Z028); Natural Science Foundation of Shanghai (24ZR1425600)

modalities suffer from significant performance limitations under challenging conditions, including nighttime environments, low illumination, adverse weather, complex backgrounds, and occlusions, which restrict their robustness in real-world scenarios. The introduction of heterogeneous modalities, such as infrared, depth, LiDAR, event cameras, and other complementary sensing modalities, provides additional geometric, thermal, structural, or motion information to alleviate these limitations. However, the heterogeneity of RGB-X data also brings new challenges, including differences in sensing mechanisms, spatial resolution, representation characteristics, and modality reliability. Meanwhile, redundant, noisy, and even conflicting information may be introduced during cross-modal interaction. Therefore, effectively exploiting complementary information while suppressing redundant and unreliable information remains a critical challenge for RGB-X cross-modal interaction. Traditional fusion approaches mainly rely on explicitly designed interaction strategies, such as element-wise operations, feature concatenation, bridging structures, attention mechanisms, and additional complex fusion modules, to achieve heterogeneous modality integration. Although these approaches have achieved substantial progress, their performance generally depends on task-specific network architectures, carefully designed fusion strategies, and large-scale annotated data, which may limit their adaptability and generalization across different scenarios and sensing conditions. In particular, when the characteristics or reliability of input modalities change, conventional fusion architectures may require additional optimization or retraining to establish effective cross-modal representations. Recently, the emergence of foundation models (FMs), including Vision Foundation Models (VFM), Vision Language Models (VLMs), and Multimodal Large Language Models (MLLMs), has provided new perspectives for RGB-X fusion. Compared with conventional fusion methods, foundation models offer strong general-purpose representation, cross-modal understanding, semantic alignment, and knowledge transfer capabilities, providing new design strategies and capability enhancement approaches for specialized fusion models. By introducing the representations and prior knowledge acquired from large-scale pre-training, foundation models can participate in different stages of RGB-X interaction, including modality perception, feature extraction, cross-modal alignment, feature enhancement, semantic interaction, and high-level representation reasoning. This emerging paradigm not only provides new possibilities for improving the complementary utilization of heterogeneous modalities, but also expands RGB-X fusion from conventional task-specific feature interaction toward more generalizable and knowledge-driven multimodal perception. This paper presents a systematic survey of RGB-X cross-modal interaction techniques driven by foundation models. First, we review the fundamental principles and development of RGB-X perception and fusion technologies, focusing on the characteristics of heterogeneous modalities and the major challenges in cross-modal alignment, information redundancy, modality reliability, and uncertainty. Then, we introduce the basic concepts and capabilities of foundation models, including VFMs, VLMs, and MLLMs, and analyze their potential roles in RGB-X perception. According to the representation hierarchy at which cross-modal interaction occurs, we systematically review existing approaches from three levels: modality-level fusion, feature-level fusion, and representation-level fusion. Modality-level fusion focuses on incorporating heterogeneous modalities at the input or modality representation stage, feature-level fusion emphasizes the interaction and enhancement of intermediate representations, while representation-level fusion further exploits high-level semantic representations and knowledge provided by foundation models. For each level, representative methods are summarized and compared from the perspectives of model architecture, modality interaction mechanism, foundation model utilization, and fusion characteristics, providing a unified view of how general-purpose foundation models are integrated with specialized RGB-X fusion architectures. Subsequently, representative methods are quantitatively compared on widely used public benchmarks to analyze their performance differences under different multimodal configurations and perception tasks. The comparison further provides insights into the effectiveness of different foundation model integration strategies and their applicability to diverse RGB-X perception settings. Beyond benchmark evaluation, we further summarize the applications of foundation model-driven RGB-X fusion in key downstream perception scenarios, including autonomous driving, intelligent robotics, and medical imaging, highlighting how heterogeneous sensing and foundation model capabilities can jointly improve perception robustness, semantic understanding, and adaptability in practical environments. Based on the reviewed methods and experimental results, we further discuss several important challenges that remain unresolved, including efficient cross-modal alignment, modality reliability and uncertainty modeling, information redundancy and conflict suppression, computational efficiency, foundation model adaptation, and the lack of unified evaluation protocols. In

particular, how to effectively balance the generalization capabilities of foundation models with the task-specific requirements of RGB-X perception remains an important research problem. Future research may focus on dynamic modality interaction, uncertainty-aware fusion, efficient foundation model adaptation, knowledge-guided multimodal reasoning, lightweight architectures, and more flexible fusion mechanisms for resource-constrained real-world applications. Unlike existing surveys that mainly focus on traditional explicit fusion strategies or conventional multimodal architectures, this work focuses on the emerging paradigm of foundation model-driven RGB-X fusion and provides a comprehensive overview of its technical foundations, representation-level taxonomy, representative methods, quantitative comparisons, and practical applications. By establishing a systematic perspective on the interaction between foundation models and heterogeneous RGB-X modalities, this survey aims to clarify the current research landscape, identify key technical challenges, and provide valuable references and insights for future research on RGB-X cross-modal interaction in the era of foundation models. All references, algorithms, datasets, and evaluation metrics are summarized at <http://github.com/yuyuxiazhu-pixel/RGB-X-fusion-technologies>.

Key words: autonomous driving; cross-modal interaction; cross-modal perception; vision foundation models; RGB-X fusion

论文引用格式: 10. 11834/jig. 260229

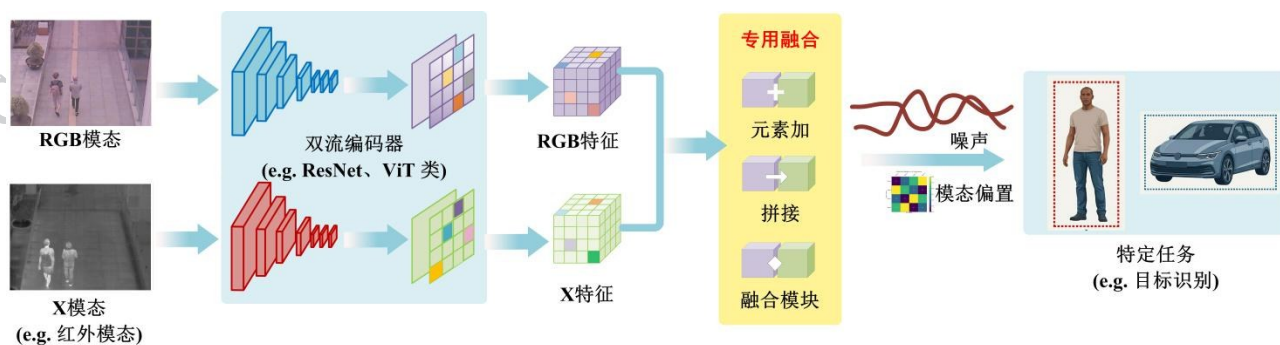
0 引言

随着人工智能技术的快速发展,视觉感知系统在自动驾驶(Ma 等, 2024b)、智能机器人(Qi 等, 2024)、医疗检测(Xu 等, 2025)以及精准农业(Akbar 等, 2024)等复杂应用场景中发挥着越来越重要的作用。传统计算机视觉方法主要依赖 RGB 图像进行目标检测、识别与场景理解。然而,在夜间、低照度、恶劣天气以及严重遮挡等复杂环境条件下,单一可见光模态往往难以提供稳定可靠的视觉信息,导致感知系统性能显著下降(Hollier 等, 1999)。为了提升视觉系统在复杂环境中的鲁棒性,自 2010 年代初期起,研究者逐渐将红外信息(Hwang 等, 2015)、深度信息(Silberman 等, 2011)及激光雷达信息(Geiger 等, 2011)等多源传感器信息引入感知任务(Zhu 等, 2025),并与 RGB 图像进行协同感知,形成 RGB-X 视觉感知框架(Yang 等, 2025a)。然而,由于不同模态在成像机理、数据分布及语义表达上的显著差异,简单的模态叠加难以充分发挥其互补优势,如何实现高效且鲁棒的跨模态融合,成为这一领域研究的核心问题。

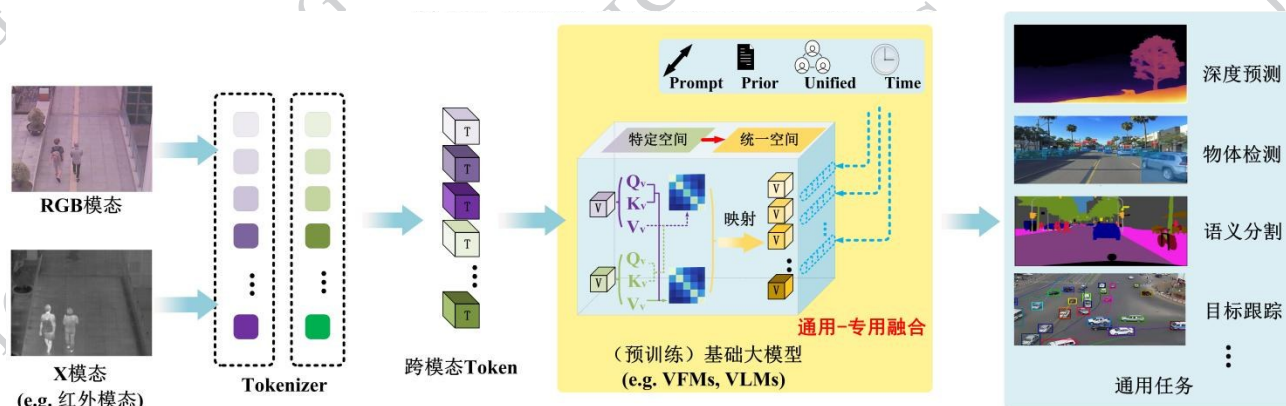
早期的跨模态融合机制主要基于 RGB 与 X 模态的显式融合策略,如图 1(a)所示,即通过元素加、桥接或专用融合模块实现异构信息的交互与对齐。这类方法主要利用当前输入中显式可观测的模态信息,并围绕特定模态组合和下游任务设计融合机制,

虽在单一场景下取得了较好效果,但其本质仍属于任务驱动的专用融合范式,在面对模态扩展、类别变化和场景迁移时,一般缺乏统一建模能力与泛化性能。

近年来,随着大规模预训练视觉基础模型(Vision Foundation Models, VFMs)与视觉语言模型(Vision Language Models, VLMs)的发展,新一代 VFMs(如 DINOv3(Siméoni 等, 2025))及开放世界视觉分割基础模型(如 SAM3(Carion 等, 2025))的不断涌现,跨模态感知研究逐渐突破传统针对特定任务设计专用网络的局限,步入通用智能引导专用网络的新范式。需要说明的是,本文所称“通用智能”并非广义的通用人工智能(Artificial General Intelligence, AGI)。此处特指以基础模型(Foundation Models, FMs)为技术载体所呈现的通用能力,具体包括:通过大规模预训练获得的跨任务、跨场景通用表征、跨模态语义理解、知识迁移及多模态推理等能力,以及上述能力对 RGB-X 融合的赋能作用。与传统方法主要利用当前输入中的显式信息不同,通用智能驱动的方法则进一步将预训练获得的通用知识先验引入融合过程,使模型能够结合已有知识对多模态观测进行深层次的语义解释,挖掘不同模态中未被表达的隐含关联,从而形成更加丰富、统一且具有迁移能力的跨模态表征。本文重点讨论三类与 RGB-X 融合密切相关的基础模型,即 VFMs、VLMs 和 MLLMs,其具体分类及基本原理将在第 1.1 节中进一步介绍。在此基础上,RGB-X 融合不再局限于依赖显式网络架构实现像素级或特征级信息交互,



(a) 基于专用任务融合范式的RGB-X感知技术路线



(b) 基于通用智能融合范式的RGB-X感知技术路线

((a)Conventional Task-Oriented Fusion Pipeline; (b) Generalized Foundation-Model-Based Fusion Pipeline)

图1 基于RGB-X(传统-新型)融合范式下的跨模态感知技术路线图

Fig. 1 Evolution of RGB-X Multimodal Perception Paradigms from Task-Specific to Generalized Fusion

还可借助基础模型的通用先验,并结合提示(Prompt)等调制机制,引导不同模态在更高层次的语义空间中实现信息关联、重组与对齐,从而提升对新类别和新场景的适应能力。具体而言,如图1(b)所示,异构模态经线性投影后转化为统一特征序列,并利用基础模型的自注意力或交叉注意力机制进行Token交互,最终通过任务头或解码器适配不同下游任务。由此可见,这种通用-专用结合的RGB-X交互方式为缓解传统专用感知网络中融合架构冗余及任务泛化能力受限的问题提供了新的技术路径。

在这一新型范式下,本文依据RGB-X融合发生的表示层级对现有技术路线进行系统梳理,具体分类如图2所示。整体来看,基础模型驱动的RGB-X融合主要可分为三类。其中,模态级融合方法(对应图2(a))主要发生于基础模型输入阶段,其核心是在进入模型之前对RGB与X模态信息进行组织与整合,使不同传感器数据以联合输入形式参与后续感知过程。特征级融合方法(图2(b))发生于基础

模型内部特征空间,其核心是在模型提取的中间特征表示上建立跨模态关联,通过特征交互促进不同模态信息的交换与互补增强。表征级融合方法(图2(c))则进一步作用于基础模型学习得到的隐空间或共享语义空间中,其核心并非直接在特征空间中进行模态特征交互,而是学习跨模态一致的高层语义表示,将RGB与X模态映射至统一表示域,并在该空间中实现跨模态语义对齐与信息融合。

尽管RGB-X跨模态融合技术在大模型背景下已发生显著的范式拓展,但现有综述(Ou等,2024; Hou等,2026a)仍主要停留在传统的显式融合范式上,围绕CNN(Song等,2023b)或Transformer等网络结构实现跨模态特征直接融合(Huang等,2022; Zhang等,2024a),尚未对通用智能驱动下的跨模态融合机制进行重新审视,形成了本领域亟待填充的研究空白。基于此,本文围绕基础模型在RGB-X专有融合机制的作用展开系统概述,重点分析通用智能如何通过Prompt调制、先验引导、统一表示及时

空建模等手段,推动跨模态融合机制的演进。

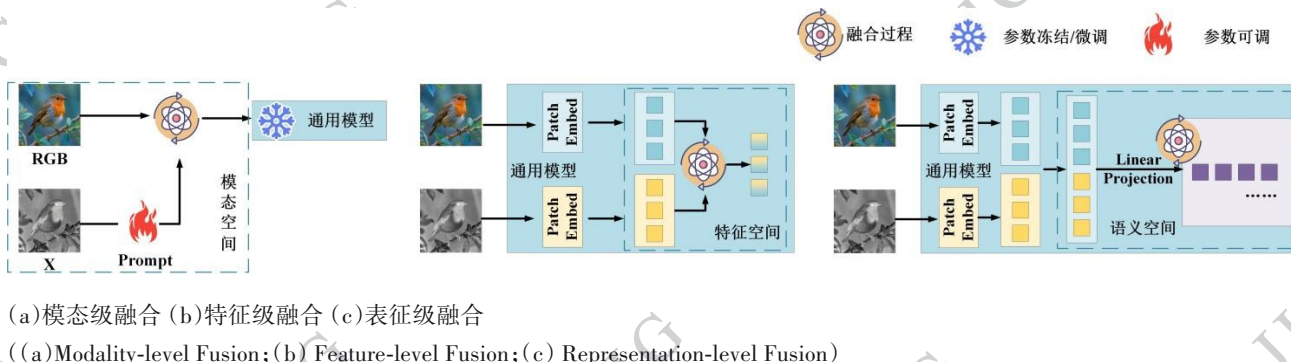


图2 通用智能驱动RGB-X融合的代表层级化分类示意图

Fig. 2 Representation Hierarchy-based Taxonomy of Foundation Model-driven RGB-X Fusion

1 前置知识

本章重点介绍 RGB-X 感知与融合的相关技术基础,旨在系统梳理支撑 RGB-X 融合发展的关键模型与技术。首先,围绕与 RGB-X 融合密切相关的基基础模型,依次对 VFMs、VLMs 和 MLLMs 的基本原理与核心能力进行介绍;随后,对传统跨模态融合技术进行系统回顾,梳理 RGB-X 数据在空间对齐、特征交互等方面的发展脉络。结合上述两个方面,为后文分析基础模型如何以不同功能赋能 RGB-X 融合及其技术范式演进奠定基础。

1.1 基础模型

随着大规模预训练技术的发展,基础模型逐渐成为人工智能研究的重要范式。通过海量数据预训练,基础模型能够学习具有较强泛化能力的通用表征,并迁移至多种下游任务。根据建模范围及能力侧重点,本文将相关基础模型分为三类:VFMs 侧重通用视觉表征学习,VLMs 侧重视觉与语言的跨模态语义关联,MLLMs 进一步融合大语言模型的指令理解与推理能力,三类模型共同为 RGB-X 融合提供技术支撑(如图3所示)。

1.1.1 VFMs

视觉基础模型通常通过大规模图像数据进行监督、自监督或弱监督预训练,以获得具有较强泛化能力的视觉表征,并迁移至不同下游视觉任务。代表性模型包括基于自监督学习的 DINOv2(Oquab 等, 2023)、MAE(He 等, 2022)和 EVA(Fang 等, 2023),以及面向通用视觉任务的 SAM(Kirillov 等, 2023)和

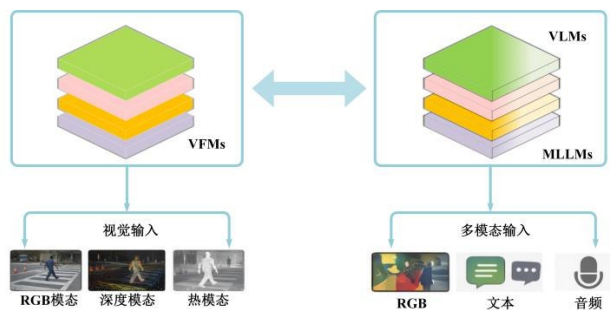


图3 视觉基础、视觉语言与多模态大语言模型示意图

Fig. 3 Architecture of VFMs, VLMs, and MLLMs

SEEM(Zou 等, 2023)。上述模型在检测、分割和识别等任务中表现出良好的迁移能力,为通用视觉表征学习提供了重要基础。

1.1.2 VLMs

VLMs 进一步将视觉与语言进行联合建模,通过图文对齐等方式建立跨模态语义关联。典型模型包括 CLIP(Radford 等, 2021)、ALIGN(Jia 等, 2021)和 BLIP(Li 等, 2022)等,其通过构建统一的视觉—语言表示空间,实现视觉信息与语言语义的有效关联,并为开放词汇理解、跨模态语义对齐及统一表示学习提供基础。

1.1.3 MLLMs

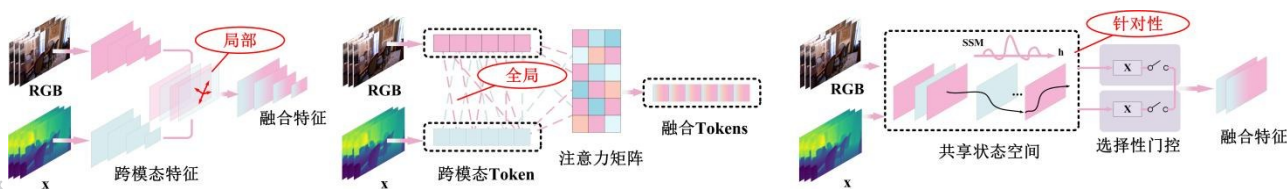
MLLMs 在 VLMs 视觉—语言跨模态建模的基础上,引入大语言模型的语言理解、指令跟随与推理能力,进一步拓展多模态模型的理解与交互能力。该类模型通常将视觉编码器与大语言模型结合,并通过多模态连接模块和指令微调将视觉信息映射至语言模型的语义空间。代表性模型包括 LLaVA(Liu 等, 2023a)、Qwen-VL(Bai 等, 2023)、InternVL(Chen 等, 2024)和 MiniGPT-4(Zhu 等, 2023a)等,可支持图

像理解、视觉问答、多轮交互及复杂推理等任务。

1.2 经典融合范式

在基础模型介入之前,RGB-X跨模态融合主要面向特定任务,通过监督或自监督学习实现不同模态信息的互补。随着融合机制的发展,相关方法逐

渐由局部特征交互扩展至全局建模,并进一步形成图模型、状态空间模型和自适应对齐等多种融合范式。综上,本节从局部交互、全局交互及其他融合方式三个层面对经典RGB-X融合方法进行概述。



(a)基于局部交互 (b)基于全局交互 (c)其他融合方法(e.g., Mamba架构)

((a) Local-based; (b) Global-based; (c) Other Fusion Methods)

图4 经典RGB-X融合方式

Fig. 4 Conventional RGB-X Fusion Frameworks

1.2.1 基于局部交互的融合方法

早期RGB-X融合主要基于CNN进行局部特征提取,并通过空间权重实现不同模态的特征融合,如图4(a)所示。针对固定权重适应性不足的问题,部分研究引入自适应权重分配机制(Liu等,2023c; Shao等,2025),并进一步采用辅助分支(Jiang等,2024b)和多分支结构(Fan等,2022)增强不同层级与模态间的信息交互。基于CNN的融合方法(Huang等,2026)具有较好的局部特征建模能力,但对长程依赖和全局语义关系的刻画有限。

1.2.2 基于全局交互的融合方法

Transformer通过全局注意力机制增强了长距离依赖建模能力,并逐渐应用于RGB-X融合任务,如图4(b)所示。现有研究主要通过跨模态注意力交互与共享单元(Hui等,2023; Jiang等,2024a; Li等,2025; Wang等,2026)、注意力对齐单元(Wang等,2025e; Zhang等,2024b)以及频域注意力机制(Jia等,2026; Lyu等,2026)实现跨模态特征交互。同时,动态注意力机制被用于自适应协调全局语义与局部特征(Xu等,2021; Wang等,2025a; Lu等,2025a; Song等,2025)。然而,基于自注意力的全局建模通常具有较高的计算开销,对模型效率和实际部署提出了更高要求。

1.2.3 其他融合方法

除CNN和Transformer外,近年来研究者进一步引入图模型、状态空间模型及自适应对齐等机制拓展RGB-X融合范式。例如,图模型通过节点与边建

模跨模态结构关系(Wu等,2024a)、基于Mamba的融合方法利用状态递归更新实现跨模态信息传递,在降低计算开销的同时增强长程依赖建模能力(Dong等,2025; Wan等,2025)、可变形对齐机制通过学习自适应偏移建立跨模态空间对应关系(Hu等,2025a)。此外,部分研究从模态可靠性和场景先验出发,引入狄利克雷分布刻画模态不确定性(Wang等,2023b),以及结合人类视觉机制(Wang等,2025b)和数字孪生环境状态信息(Guo等,2022; Zhang等,2020)增强融合过程。上述方法从结构关系、状态演化、空间对齐、模态可靠性和环境先验等角度拓展了传统融合机制,但由于建模机制差异较大,在通用性和跨场景稳定性方面仍存在提升空间。

2 现有方法

本章围绕大模型背景下RGB-X融合技术的发展展开,对相关技术路径进行系统梳理与归纳。在多源信息协同建模过程中,跨模态信息的融合位置逐渐由模型输入端扩展至内部特征空间以及高层语义表示空间,体现出基础模型参与RGB-X交互的层级演进。本章依次从模态级融合、特征级融合及表征级融合三类技术路线出发,结合典型下游感知场景对各类融合方式进行解析,分析其如何在不同层级实现信息组织、特征交互与语义统一,进而推动相关技术由传统显式特征融合范式向基础模型驱动的跨模态理解与协同表征演进。

2.1 模态级 RGB-X 融合

在此类方法中,融合过程发生于深层特征提取之前,即在基础模型输入端完成跨模态信息的组织与交互。传统方法主要通过模态拼接、加权等方式直接融合输入信息,而随着基础模型的发展,模态级融合进一步演进为基于 Prompt 的条件化输入,将异构模态或外部辅助信息编码为不同形式的 Prompt,以引导基础模型进行跨模态感知。在保持主干网络冻结或轻量微调的条件下,该机制能够以较低参数开销实现跨模态适配与条件激活。因此,本文依据 Prompt 对输入信息的组织方式及作用路径,将现有方法划分为联合式 Prompt 融合、并行式 Prompt 融合和多源式 Prompt 融合三类。

2.1.1 联合式 Prompt 融合

直接将 RGB 与异构模态信息进行联合 Prompt 编码是较为常见的前端融合机制,凭借较高的参数效率成为代表性的融合范式,其架构图参见图 5(a)。其核心目标是缓解 RGB 与异构模态之间初始语义对齐不足的问题,基本思想是在统一 Prompt 空间中共同编码不同模态信息,使基础模型能够利用共享提示建立跨模态关联,进而降低模态间的表示差异。

ProTrack (Yang 等, 2022) 较早将这一思路引入跨模态视觉架构,该模型通过构建提示框引导 RGB 与红外特征在统一表示空间进行集成,为联合式 Prompt 融合奠定了基础。随后,考虑到不同模态在开放场景下的表达能力并不恒定,单纯的静态联合编码容易忽略模态间的动态关联,后续研究逐步通过共享 Prompt 参数实现联合编码。例如, He (2024b) 等提出的工作通过共享部分 Prompt 参数并聚合类别表征,进一步强化了跨模态语义的一致性建模。这类方法本质上属于单阶段 Prompt 融合。为进一步降低不同语义层之间的信息冗余,并提升不同尺度特征的表达能力,部分研究进一步演化出分阶段、多尺度 Prompt 设计。以双驱动 Prompt 范式 (Luo 等, 2025) 及 M3PT 架构 (Wang 等, 2024a) 中的多阶段视觉 Prompt 为例,这些方法通过在编码及解码等不同阶段设置独立的视觉 Prompt,使模型在不同语义层级下学习差异化调制策略,在共享基础模型潜在知识的同时增强阶段间表达能力,更好地兼容各种下游任务的需求。此外,一些前沿架构 (Dong 等, 2024) 通过耦合层级化的异构特征,设计

出多尺度、多模态 Prompt,以更紧凑的形式融合不同表征层级的跨模态信息,显著提升了模型在复杂场景下的表示效率与鲁棒性。

以上联合式融合方法主要通过共享提示建立跨模态初始语义关联,为后续深层融合提供统一的语义基础。

2.1.2 并行式 Prompt 融合

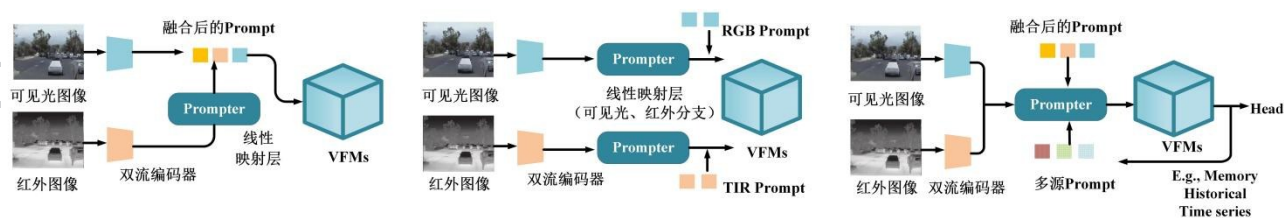
与联合式 Prompt 主要关注跨模态语义对齐不同,并行式 Prompt 更关注不同模态可靠性差异所带来的特征偏置问题。这种方式本质上侧重于保留异构模态在语义表达与结构敏感性上的特异性,其主张在 RGB 与异构模态链路上分别部署独立的提示分支 (如图 5(b)),通过解耦的提示路径实现对各自模态特征的针对性诱导。在这种设计下, Prompt 更多扮演着模态专属引导的角色,通过条件化调制缓解网络的模态偏置问题。目前,此类研究 (Tian 等, 2026; Luo, 2023) 已证明,这种分支式 Prompt 部署能有效提升模型对异构特征的表征精度。

值得注意的是,这种基于 Prompt 的隐式知识引导与显式对齐的知识蒸馏技术在原理上呈现互补的趋势。因此,近期部分前沿工作开始探究将这两种引导范式协同耦合,以充分利用不同模态的互补特征。例如双流 Prompt 跟踪框架 (Luo 等, 2024) 通过构建教师-学生结构并引入分层蒸馏策略,在实现跨模态知识迁移的同时显著压缩了模型参数。随后, LRPD 等模型 (Hu 等, 2025b; Liu 等, 2025b) 通过引入 LoRA 低秩适配技术 (Hu 等, 2021) 构建起 Prompt 驱动的蒸馏框架,这类方法不仅促进了模态特征的深度对齐,更体现了视觉 Prompt 与参数高效微调技术融合的发展趋势,为复杂感知任务提供了高效、低开销的适配方案。

上述并行方式的融合策略更加关注模态特性的保持与可靠性差异建模,因此在复杂环境下具有更好的适应能力。但与联合式融合类似地,其主要仍面向 RGB 与单一 X 模态信号,在开放场景下对可利用信息的范围有所限制。

2.1.3 多源式 Prompt

分析 2.1.1 及 2.1.2 节两种 Prompt 融合机制可知,这类方式一方面将注意力全部聚焦于 RGB 与单一 X 模态,可利用的信息仍主要受限于当前视觉观测。为突破这一限制,研究者开始将文本、时序信息及地理空间先验等外部辅助条件引入 Prompt 机制,



(a)联合式 Prompt 融合 (b)并行式 Prompt 融合 (TIR 代表红外)
(c)多源式 Prompt 融合
(a) Joint Prompt; (b) Parallel Prompt; (c) Multi-Source Prompt

图5 模态级 RGB-X 融合机制示意图

Fig. 5 Illustration of the Modality-level RGB-X Fusion Mechanism

使其在融合过程中承担更丰富的语义整合功能(对应技术原理见图 5(c))。由图可知,上述额外信息(如时间、文本、地理先验等)并不与 RGB 及 X 模态并列,而是主要作为条件先验或上下文信息参与 RGB-X 感知过程,用于补充原始视觉观测中缺失的时空与语义信息。由此, Prompt 由单一视觉模态调制工具进一步演变为融合多源先验的条件化信息载体。

例如,时间维(时序)信息的引入是其中重要的研究方向,可显著提升了模型在动态场景下的稳健性。例如, Wang 等(2026)提出时间维 Prompt 机制,通过将历史特征逐步引入当前 Token,构建跨时间的语义衔接,使模型能够利用前序状态引导当前预测。类似地, Zeng(2025)等利用历史目标位置作为结构化提示,引导模型聚焦关键空间区域,以提升跨模态信息的空间对齐与融合效果。 MoKA-HP(Wu 等, 2026)与 PromptTrack(Li 等, 2026)则进一步构建了流式时空 Prompt 方法,实现了对视频序列中的长程依赖关系进行显式建模。

另一方面,受到多模态大模型对文本、音频等多源信息的综合应用能力的启发, Prompt 融合机制进一步发展多源条件协同模式。此类方法不再孤立依赖视觉成分,而是将文本、掩码等多种提示形式协同注入。例如, Hou(2026b)等将动态生成的视觉、文本与掩码提示相结合,引导 SAM 模型完成复杂环境下的显著目标分割。随后, X-Prompt 框架的提出进一步拓展了该范式的边界,在保持基础模型参数冻结的前提下,将包括深度(Muturi 等, 2025)、事件流(Tang 等, 2025)以及跨域联合输入(Wu 等, 2024c; Hong 等, 2024)等多源信息统一编码为 Prompt 提示。此类工作通过构建统一提示适配机

制,在不同模态组合间共享调制策略,增强了模型在复杂场景下的泛化能力与可延展性。

总体来看,多源式 Prompt 融合机制主要通过 Prompt 等外部先验建立 RGB 与 X 模态间的初始语义关联,并进一步利用文本、时序及空间先验等辅助信息增强 RGB-X 感知能力,在保持基础模型主体冻结或少量参数微调的前提下,实现高效的跨模态适配与知识迁移。然而,这类方法更多承担外部语义调控作用,对跨模态深层交互与统一表示建模能力仍然有限;随着提示来源不断增加,不同来源信息间的一致性建模与动态可靠性评估亦是后续需要解决的重点问题。

2.2 特征级 RGB-X 融合

分析上节可知,模态级融合主要在输入端完成 RGB 与 X 模态的信息整合,具有结构简单、计算成本低等特点,但受限于原始输入空间,对高层语义关联和跨模态互补信息的挖掘有限。为进一步利用基础模型的预训练知识,研究者将其引入特征提取阶段,在中高层特征空间开展 RGB-X 信息交互与表示优化。根据基础模型在融合过程中的功能角色,现有方法主要分为三类:特征交互型、模态适配型和状态预测型融合。其中,特征交互型通过交叉注意力、特征校准等机制增强模态间信息交换;模态适配型利用 LoRA、Adapter 等参数高效微调策略调整特征空间,以适应不同传感器的数据分布;状态预测型则引入时序建模,通过学习跨时间的状态演化规律提升动态场景下的融合稳定性。以下分别对三类方法进行介绍。

2.2.1 特征交互型融合

特征交互型方法将 VFMs 作为语义先验或特征交互载体,参与 RGB-X 特征的对齐与融合(参考图 6

(a)),以缓解不同模态在成像机制、特征表达上的差异及由此产生的语义缺失与模态偏置。基于上述思想,早期研究主要探索利用VFM中蕴含的通用视觉先验辅助RGB-X特征交互。其中,SAM系列模型凭借其目标理解与像素级分割能力,成为该方向应用较为广泛的基础模型之一。例如,He等(2024c)率先将SAM引入RGB-T显著性检测任务,并通过引入HPTP(High-Resolution Transformer Pixel Extraction)等模块增强其跨模态特征提取与融合能力,为后续基于VFM的RGB-X融合提供了重要参考。类似地,Fang等(2024)提出基于SAM语义先验引导的RGB-T跨模态计数方法,通过SAM生成的语义图指导跨模态特征融合,实现前景背景区分与密度图生

成。FusionSAM(Li等,2025a)将SAM引入跨模态图像分割任务,通过潜在空间特征量化与融合提示生成模块,将传统黑盒训练转变为可控的提示引导机制,实现可控的像素级分割。MiLNet(Liu等,2025a)从视觉基础模型中迁移稳健的知识表达,并将其与任务专有模型进行特征传递,提升异构模态与可见光信息的集成能力。KAN-SAM(Li等,2025b)则将热成像特征作为引导提示注入SAM2(Ravi等,2024),并通过KAN(Kolmogorov-Arnold Network)适配器实现高效的多模态显著性检测。总体而言,上述方法将SAM系列架构作为视觉知识先验,通过增强模态特征表达来缓解异构模态间的语义缺失与特征不稳定问题。



(a)特征交互型融合 (b)模态适配型融合

((a) Feature-aware fusion; (b) Modality-aware fusion)

图6 特征级RGB-X融合机制示意图

Fig. 6 Framework of the Feature-level RGB-X Fusion Mechanism

随着视觉语言模型的发展,研究进一步引入跨模态语言知识作为高层语义约束,以弥补单一视觉先验在场景理解方面的不足。一方面,借鉴语言模型中的提示调优思想,ViPT(Zhu等,2023b)冻结预训练基础模型主干,将RGB-X中的X模态编码为可学习的Prompt Token,并注入Transformer的层间特征提取过程,实现基础模型内部的跨模态特征交互与选择。另一方面,部分研究利用VLMs的图文对齐能力引导跨模态特征学习。例如,TIETracker(Xia等,2025)利用CLIP的图文对齐能力提取文本特征作为语义先验,并在视觉特征可信度较低时进行跨模态对齐,从而提升感知准确度。Gao等(2025b)利用BLIP2(Li等,2023b)生成图像描述,并通过冻结权重的BERT(Alaparthi等,2020)提取语言特征,作为提示引导跨模态特征融合,实现视觉-语言联合驱动的目标跟踪。Ren等(2025)则利用BLIP的图文理解能力生成输入文本描述,并传递给SAM分割执行

器,实现跨模态感知。

需要指出的是,VFM的先验知识已不仅限于跨模态特征融合,更延伸至RGB-X数据构建与标注优化环节,从数据层面间接提升融合模型的泛化能力。例如,Salas等(2026)利用SAM2设计半自动标注流程,通过特征匹配实现RGB-T图像的几何对齐,兼顾标注效率与跨模态对齐精度,实现了基础模型知识由特征层向数据层的有效迁移。总体而言,这类特征交互型方法借助VFM与VLMs的视觉及语义先验增强多模态特征表达,促进模态间信息交互,但其主要依赖固定预训练表征,对复杂场景及模态分布变化的动态适应能力仍有限,泛化性能有待进一步提升。

2.2.2 模态适配型融合

与直接利用VFM固定先验的方法不同,模态适配型融合方法关注预训练知识与异构模态数据分布间的域偏移问题。由于VFM多基于大规模RGB

图像预训练,其特征表示难以充分适应红外、深度等模态在纹理与语义上的差异,因此该类方法通过参数高效的适配策略,使模型在保留原有视觉知识的同时,学习面向 RGB-X 任务的模态特定表征。为降低计算成本,现有研究多采用 Adapter、LoRA 等参数高效微调(PEFT)技术,在冻结主体网络的前提下仅优化少量参数完成适配。相较于先验引导方法,该范式不再依赖通用知识,而是通过任务驱动参数优化,使基础模型主动适应各异构模态的特征分布并参与跨模态调整,从而增强模态间的特征一致性与融合性能。

早期研究主要通过对基础模型进行微调与结构适配,提高其对异构模态数据的表征与融合能力。例如,Liu 等(2026)提出的 SAMSOD 模型针对 SAM 在 RGB-T 显著性检测中存在的双模态收敛不平衡与梯度冲突问题,通过单模态监督、梯度冲突缓解机制以及解耦适配器掩码等策略对 SAM 模型进行优化微调,从而提升跨模态特征融合效果。Lin 等(2025)提出 SAM-DAQ 模型,基于深度引导自适应查询,通过并行适配器与查询驱动时序模块实现 SAM2 的跨模态融合与时序建模。OpenRSS(Zhao 等,2024)提出了具有开放词汇分割能力的 RGB-T 语义分割框架,通过联合改进 SAM 视觉基础模型并引入 CLIP 视觉语言模型,实现了跨模态感知任务中的开放词汇识别。

然而,传统适配技术通常依赖于特定模型结构与任务设计,导致其迁移能力和通用性受到一定限制,难以作为即插即用的模块广泛应用于不同视觉基础模型中。为提升模型适配的灵活性与参数效率,研究人员进一步将 LoRA 技术引入 RGB-X 交互融合中,通过低秩参数更新增强 VFM 对 RGB-X 特征分布的适应能力,同时降低基础模型调整与迁移的成本。如 SAM-FuseNet(Zhu 等,2026)基于 SAM2 构建模态特定编码器,并通过 LoRA 实现参数高效微调,结合像素级空间注意力与通道重校准实现自适应跨模态融合,并采用双路径分割头提升预测质量。TASeg(Yu 等,2025)提出了一种基于文本感知的 RGB-T 分割架构,该模型通过 LoRA 参数高效微调技术适配视觉基础模型,并结合动态特征融合与 CLIP 文本语义对齐,解决了跨模态分割中特征融合与语义理解问题。类似地,AdaptRGB-T(Yu 等,2026)进一步探索了参数高效适配在 RGB-T 感知任

务中的应用。此外,也有研究采用知识蒸馏方式(Xing 等,2025),将经 LoRA 微调后的 VFMs 中的通用知识迁移至轻量级任务网络,实现模型压缩与效率提升。

2.2.3 状态预测型融合

前述特征交互、模态适配型融合方法主要针对当前观测条件下 RGB 与 X 模态之间的信息融合。然而,在真实动态环境中,由于目标运动、场景变化、传感器失效以及模态可靠性波动所引起的时序依赖关系仍缺乏有效刻画(Wang 等,2024b)。因此,进一步的研究开始探索使基础模型承担动态状态建模的角色,通过学习多模态观测背后的潜在环境状态及其演化规律,实现从静态特征融合向动态环境理解的扩展。具体而言,该类方法引入时空建模与世界模型(World Model)思想,将 RGB、热红外、深度以及雷达等多源传感器信息视为对同一物理世界状态的多视角观测,并通过学习潜在状态空间实现历史信息融合、未来状态预测以及任务决策支撑。与仅关注模态间静态对应关系的融合方法不同,状态预测型方法能够利用时间维度上的连续观测信息,捕获环境状态随时间变化的规律,从而提升复杂开放场景下 RGB-X 融合的稳定性和鲁棒性。

在状态预测型融合流程如图 7 所示,多源传感器信息首先通过基础模型进行特征编码,形成包含时序信息的动态状态表示,并进一步利用时序建模机制捕获跨模态特征随时间变化的演化规律(Ge 等,2024;Mazzaglia 等,2024)。与传统融合方法仅关注当前帧特征互补不同,动态状态建模方法利用历史观测信息刻画环境演化过程,从而增强模型在遮挡、噪声干扰以及模态可靠性变化情况下的鲁棒性。

进一步来看,状态预测型融合方法主要围绕两个层面发挥作用。首先,在环境状态理解层面,部分研究关注如何利用多模态历史观测信息推断未来环境变化趋势,其核心目标是学习具有时间连续性和物理一致性的动态状态表示。具体而言,模型通过构建跨模态共享状态空间,将不同传感器信息映射至统一的动态特征空间,并利用时序模型或世界模型学习潜在状态转移规律,实现对未来环境状态的预测。例如,Shang 等(2025)通过构建跨模态共享状态表示,实现不同传感器信息的联合编码,并进一步利用动态模型预测未来状态。在 RGB-D 视频场

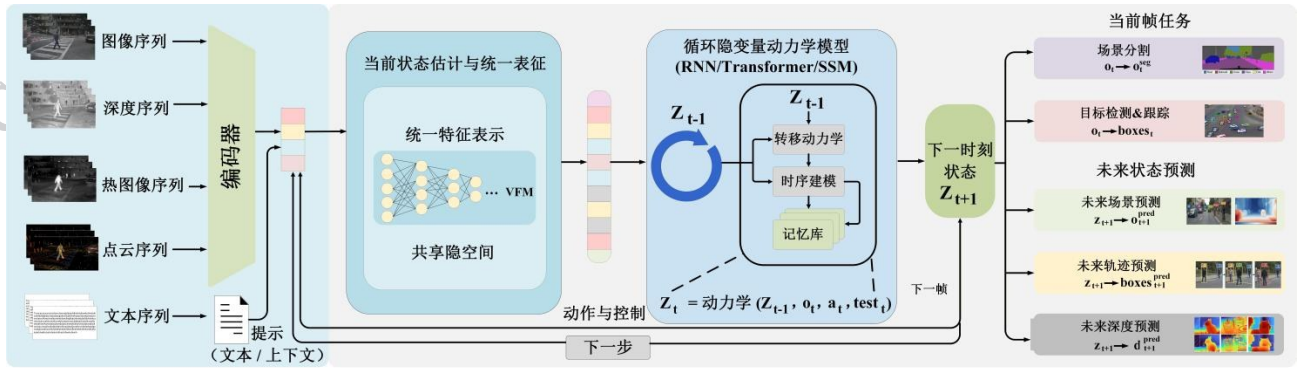


图7 (特征级融合中的)状态预测型RGB-X融合机制示意图

Fig. 7 Pipeline of State Prediction-driven RGB-X Fusion Mechanism within Feature-level Fusion

景(Guo等, 2026)以及自动驾驶场景(Zhang等, 2024d)中,相关方法通过融合摄像头、激光雷达等多源观测信息,学习环境状态演化规律,并进一步应用于未来轨迹预测(Bogdoll等, 2025; Hassan等, 2025)、物体交互关系推理以及场景变化预测(Liang等, 2025)。该类方法通过引入未来状态约束,使融合后的动态状态表示不仅能够描述当前环境信息,还能够保持跨时间的一致性,从而降低单帧观测噪声以及短时模态缺失对融合结果的影响。另一方面,随着状态预测型融合能力的发展,部分研究进一步探索如何利用动态状态表示服务于下游感知与决策任务。与显式预测未来状态不同,该类方法更加关注动态环境表征对具体任务性能的提升,即将世界模型视为一种结构化信息建模模块,通过整合不同时间阶段和不同传感器信息,为任务提供更加鲁棒的高层语义表示。目前,动态RGB-X融合已逐渐应用于目标检测(Zhou等, 2025a)、分类(Wang等, 2025d)、机器人规划(Zhen等, 2025)以及遥感分析等视觉感知任务中(Lu等, 2025b)。例如,在自动驾驶(Goff等, 2025)和机器人任务(Guo等, 2025)中,动态状态建模能够通过持续更新环境状态表示,提高模型在遮挡、传感器失效以及复杂运动条件下的感知与决策能力。随着具身智能的发展,动态RGB-X融合进一步与视觉-语言-动作模型(Vision-Language-Action, VLA)结合,例如多传感器VLA框架(Duan等, 2026)通过融合RGB、深度、触觉以及激光雷达等多源观测,并结合语言指令完成环境理解与动作规划,使RGB-X融合由传统感知增强进一步拓展至面向任务执行的闭环智能交互。

总体来看,状态预测型融合范式增强了基础模

型在RGB-X融合过程中的作用范围,但现有方法仍面临动态环境中的不确定性传播以及模型计算开销较高等问题。这进一步增加了系统的计算与存储负担。如何解决上述问题,实现可靠的动态世界状态表达,仍是未来RGB-X跨模态融合的重要研究方向。

2.3 表征级融合

在RGB-X跨模态感知任务中,由于不同传感器在成像机理、数据分布和语义表达方面存在差异,仅依靠输入端融合或特征空间交互仍难以充分消除跨模态表示偏移。因此,部分研究进一步探索表征级融合,将融合过程推进至基础模型学习得到的共享隐空间或语义空间,通过共享嵌入空间、跨模态对齐约束和统一编码机制,将RGB、深度、热红外等异构模态映射至统一表示域,从而在高层语义空间实现跨模态信息的统一表达与融合。这种方式有助于降低模态差异带来的分布偏移,并提升模型在复杂场景和未知条件下的泛化能力。然而,如何构建准确的跨模态共享语义空间,以及如何保证统一表示在跨域、未知场景和模态缺失等开放条件下的稳定性,仍是该类方法面临的主要问题。围绕上述问题,现有研究主要形成了两条技术路线:统一表征型融合,侧重于构建共享语义空间以实现跨模态语义一致性学习;开放泛化型融合,侧重于增强统一表示对跨域变化、模态缺失及未知场景的适应能力。

2.3.1 统一表征型融合

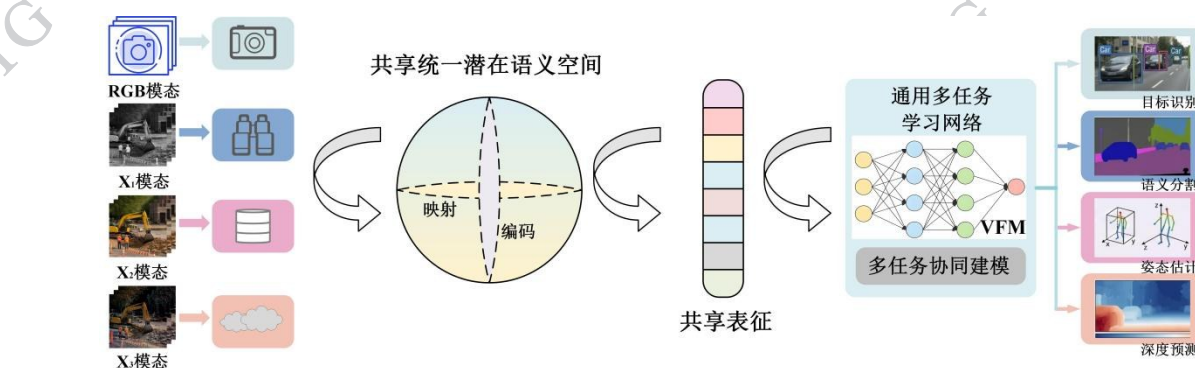
为消除RGB与异构模态在成像机理、数据分布及语义表达上的天然差异,统一表征型融合的核心目标是在基础模型构建的共享潜在空间中,将不同来源的模态信息映射至统一语义坐标系,实现跨模

态语义对齐与统一表征学习,其典型结构如图8(a)所示。

早期统一表征学习方法主要围绕共享表示空间的构建展开,其目标是建立潜在空间(Wu等,2024b),降低不同模态之间的分布差异,实现模态无关的特征表达。例如,CMX(Zhang等,2023a)通过双向特征校准与跨模态长程交互机制,实现了对多种RGB-X组合的统一建模。该方法通过建立公共潜在空间,使RGB与X模态能够在同一表示域内完成信息交互,有效缓解了由于成像机制差异导致的语义鸿沟问题。围绕这一思想,近年来大量研究(Qashqai等,2024;Feng等,2025a;Duan等,2025;Chen等,2025;Tuzlupinar等,2026;Hoang等,2026)进一步探索了不同模态组合下的统一表示构建方法。

随着统一表征学习的发展,一部分研究进一步关注如何获得更加准确、稳定的共享表示。例如,一些研究(Orfaig等,2026;Zhang等,2025a;Xuan等,2025)开始引入生成式模型和动态建模机制,将跨模态对齐过程转化为逐步去噪的迭代优化过程(Chen等,2023),通过渐进式优化增强异构特征的一致性。此外,部分轻量化架构(Li等,2025c)则进一步关注统一表示构建过程中的计算效率,在保持跨模态融合能力的同时降低模型复杂度。

总体来看,统一表征型融合通过构建共享表示空间,有效缓解了异构模态之间的分布差异。但如何在统一表示过程中兼顾模态共性与互补特征,并实现动态环境下的不确定性建模,仍是统一表征型融合方法未来需要进一步探索的方向。



(a)统一表征型融合 (b)开放泛化型融合

((a)Unified Representation-based Fusion; (b) Open-world Generalization-based Fusion)

图8 表征级RGB-X融合机制示意图

Fig. 8 Framework of the Representation-level RGB-X Fusion Mechanism

2.3.2 开放泛化型融合

在统一表征型融合构建跨模态共享语义空间的基础上,开放泛化型融合进一步关注如何使该统一表示适应任务变化、模态缺失、传感器退化及跨域分布偏移等复杂开放环境。由于固定共享空间难以保证未知条件下的稳定性能,近年研究开始探索增强统一表示的迁移能力与鲁棒性,以突破预设模态与任务限制(原理图参见图8(b))。该类方法并非简单改变融合结构,而是从模态泛化、任务泛化和整体泛化增强三个维度,提升模型在动态场景下的跨模态融合稳定性。

首先,针对模态条件变化导致统一表示退化的问题,相关研究主要探索面向模态泛化的融合机制。

例如,Hoang等(2026)针对未配准模态问题,通过构建更加灵活的跨模态表示机制,降低不同模态空间偏移带来的影响;Hu等(2025)和Tian等(2026)进一步研究缺失模态条件下的统一表示学习,使模型在部分传感器信息缺失时仍能够保持稳定的融合性能。此外,面向异构模态特性的动态调整策略也逐渐受到关注,相关方法(Deevi等,2024;Wang等,2025c;Cai等,2025;Zhang等,2026)通过根据输入模态分布动态调整融合过程,提高复杂环境下的模态适应能力。例如,MMPX模型(Wu等,2025a)通过设计基于Mamba的特征融合器以及提示器(Prompter),实现不同模态特征的自适应集成;Uni-SOT(Ma等,2026)提出秩自适应的模态适配机制,

不仅支持图像模态,同时能够扩展至视频模态对齐,为增量式模态扩展提供了新的解决思路(Hu 等, 2025c)。

其次,随着 RGB-X 应用场景不断扩展,研究进一步关注统一表示跨任务迁移能力不足的问题。例如,Choi 等(2025)和 Zhang 等(2025b)通过共享骨干网络结合任务特定解码器,实现多任务条件下的统一建模。然而,这类方法仍存在任务间关联利用不足的问题。进一步地,M-SpecGene(Zhou 等, 2025b)通过大规模自监督训练学习模态无关(modality-invariant)的潜在表示,并结合跨模态结构稀疏性约束和渐进式掩码策略,提高模型对不同任务场景的适应能力。与此同时,Yuan 等(2025)引入条件化调制机制,实现任务与模态联合适配;LoME(Zhang 等, 2026)利用低秩模态适配器和动态参数融合策略,使模型能够根据输入特征动态调整融合过程;ComPtr(Pang 等, 2025)则基于 Transformer 构建统一序列交互框架,通过一致性增强模块提取共享语义,并利用差异感知模块保留模态特有信息,实现面向多种密集预测任务的统一建模。

最后,随着通用人工智能的发展,部分研究进一步探索利用大模型的泛化能力增强统一表示鲁棒性。例如,混合专家模型(Mixture of Experts, MoE)通过动态路由机制,根据不同模态条件选择更加适合的专家网络,提高多模态表示能力(Tan 等, 2025);对比学习(Jamal 等, 2025)利用跨模态一致性约束,进一步增强共享表示的稳定性;测试时适应(Test-Time Adaptation, TTA)技术则通过调整模型对测试分布的响应,缓解模态分布变化导致的性能退化问题(Liu 等, 2023b)。此外,因果表征学习也逐渐被引入多模态泛化任务,通过挖掘跨模态稳定因素并减少环境相关偏差,提高模型面对未知场景时的表示鲁棒性(Wang 等, 2025)。需要指出的是,这些方法并非独立的融合机制,而是作为基础模型能力增强策略,与统一表示学习和动态建模框架相结合,从而进一步提升 RGB-X 融合系统在开放环境中的可靠性。

总体而言,开放泛化型融合突破了统一表示方法对固定模态和固定任务的依赖,但如何在保证跨模态一致性的同时保留不同传感器的独特信息,以及如何利用基础模型知识实现更加高效的不确定性建模与轻量化部署,仍是未来表征级 RGB-X 融合需

要进一步探索的问题。

3 基准评测

为评估不同 RGB-X 跨模态融合机制的有效性,本文围绕典型跨模态感知任务构建统一的评测视角。考虑到融合效果通常通过下游任务性能加以体现,选取显著目标检测、语义分割和目标跟踪作为代表性评测任务,分别介绍各任务所采用的数据集及评价指标,并汇总主流方法的性能表现,从融合机制的角度分析不同方法的性能差异及适用场景。

3.1 数据集

3.1.1 RGB-X 显著目标检测数据集

RGB-T: VT821(Wang 等, 2018)是首个公开的 RGB-T 显著目标检测数据集,后续 VT1000(Tu 等, 2020)和 VT5000(Tu 等, 2023)逐步扩大数据规模,后者包含 5000 对配准图像。VI-RGBT1500(Song 等, 2023a)针对充足光照、不均匀光照和低照度三种条件采集,用于评估模型在复杂光照场景下的检测性能。

RGB-D: STERE(Niu 等, 2012)、NLPR(Peng 等, 2014)和 NJU2K(Ju 等, 2014)是该领域较早的代表性数据集,样本量分别为 1000、1000 和 1985 对。此外,Piao 等(2019)、Fan 等(2020)及 ReDWeb-S(Liu 等, 2022)等通过实测采集构建。GIT(Ciptadi 等, 2013)、LFSD(Li 等, 2014)、DES(Cheng 等, 2014)和 SSD(Zhu 等, 2017)等小规模数据集常被用于模型泛化能力的评估。

3.1.2 RGB-X 语义分割数据集

RGB-T: MFNet(Ha 等, 2017)是首个面向城市驾驶场景的 RGB-T 语义分割数据集,包含 1569 对日间与夜间图像,标注 9 类道路目标。PST900(Shibakumar 等, 2020)面向机器人自主作业,包含 894 对图像,标注 5 类目标。Freiburg Thermal(Vertens 等, 2020)规模较大,包含 2 万余对图像,覆盖 13 个类别。RoadScene-seg(Yi 等, 2022)在 RoadScene 基础上人工标注扩充,包含 1326 对样本,标注 8 个类别,并按 10:1 划分训练、验证和测试集。RGB-D: NYUDv2(Silberman 等, 2012)包含 1449 对室内 RGB-D 图像,涵盖 894 个类别。SUN RGB-D(Song 等, 2015)包含 10335 对样本,提供大量 2D 多边形及 3D 物体框标注。Stanford 2D-3D-S(Armeni 等, 2017)规模更大,

包含 70496 张高清图像,覆盖 13 个类别。ScanNet (Dai 等, 2017) 是首个 RGB-D 视频数据集,包含超过 250 万帧,支持室内场景理解与 3D 重建研究。

3.1.3 RGB-X 目标跟踪数据集

RGB-T: GTOT (Li 等, 2016) 是首个 RGB-T 目标跟踪标准数据集,包含 50 个视频序列。随后 RGBT210 (Li 等, 2017) 和 RGBT234 (Li 等, 2019) 分别将规模扩展至 210 和 234 个序列,并标注多类挑战属性。VOT-RGBT2019 (Kristan 等, 2019) 是首个竞赛级 RGB-T 数据集,包含 60 个序列。LasHeR (Li 等, 2021) 为当前较大规模的数据集,包含 1224 对视频序列,总帧数超过 73 万帧。

RGB-D: PTB (Song 等, 2013) 是 RGB-D 目标跟踪领域较早的公开数据集,包含 100 个视频序列。STC (Xiao 等, 2018) 包含 36 个序列,CDTB (Lukezic 等, 2019) 包含 80 个序列,总帧数约 10 万帧,分辨率亦有提升。随后 DepthTrack (Yan 等, 2021)、RGBD1K (Zhu 等, 2023c) 和 ARKitTrack (Zhao 等, 2023) 等数据集进一步扩大了数据规模。DTTD (Feng 等, 2023) 将跟踪任务由二维扩展至三维,包含 103 个场景,每帧均具有像素级语义标注和姿态信息。

3.2 评测标准

3.2.1 RGB-X 显著目标检测评价指标

显著目标检测通常采用平均绝对误差 (Mean Absolute Error, MAE)、综合评价指标 (F-Measure, F_β)、结构评价指标 (S-Measure, S_m) 和增强配准指标 (E-Measure, E_m) 进行综合评价。其中,MAE 是用于评价基准显著图 S 与预测图 Y 之间的误差,数值越小表示预测结果越接近真实标注; F_β 综合考虑精度和召回率 (Zhao 等, 2019),用于评价预测结果的整体准确性; S_m 从区域结构和目标结构两个方面衡量预测结果与真实显著图之间的结构相似性 (Fan 等, 2017); E_m 则综合考虑像素级匹配和图像级统计信息,用于评价预测结果与真实标注之间的整体一致性。上述指标从像素误差、查准率与查全率、结构相似性以及整体一致性等不同角度反映显著目标检测性能。

3.2.2 RGB-X 语义分割评价指标

在语义分割任务中,像素准确率 (Accuracy, Acc) 和交并比 (Intersection over union, IoU) 是最核心的两个像素级评价指标。其中,Acc 表示预测正确的像素占全部像素的比例,用于反映整体像素分

类准确性;IoU 衡量预测区域与真实标注区域之间的重叠程度,是语义分割任务中最常用的评价指标之一。对于包含多个类别的数据集,通常进一步采用各类别 IoU 的平均值,即平均交并比 (mean IoU, mIoU),以综合评价模型对不同类别的分割能力。上述指标均为数值越高表示性能越优。

3.2.3 RGB-X 目标跟踪评价指标

经调研,现有跟踪任务的评价指标主要有五类,依次为精确率 (Precision Rate, PR)、成功率 (Success Rate, SR)、跟踪精度 (Accuracy, A)、跟踪鲁棒性 (Robustness, R) 及期望平均重叠率 (Expected Average Overlap, EAO)。其中,PR 主要衡量预测目标中心位置与真实位置之间的距离误差,SR 根据预测框与真实框之间的重叠程度评价跟踪成功情况;A 和 R 分别从目标定位精度和跟踪失败情况评价跟踪性能;EAO 综合考虑跟踪精度与鲁棒性,用于对跟踪器进行整体性能评价。不同数据集和评测协议所采用的具体指标及参数设置可能存在差异,本文按照各基准数据集的官方评测协议进行结果比较。

3.3 定量评估

本节选取显著目标检测、语义分割及目标跟踪任务进行算法的性能对比,其中,显著目标检测任务选择在数据集 VT821、VT1000 及 VT5000 上对 10 个基于大模型驱动的检测模型进行对比;语义分割任务选择在 RGB-D 类数据集 NYU Depth V2、SUN-RGBD 和 RGB-T 类数据集 MFNet Dataset、PST900 上对 22 个基于大模型驱动的分割模型进行对比;目标跟踪任务选择 GTOT、RGBT234 及 LasHeR 三个公开数据集,对 25 种基于大模型驱动的 RGB-X 目标跟踪方法进行性能比较。对上述三部分对比结果依次统计于表 1-3,其中在各表内,按照从上至下的顺序依次分三大类 (按照不同颜色区分),即模态级融合、特征级融合及表征级融合三种方法。

分析表 1 可知,采用不同融合层级的 RGB-X 显著目标检测方法均取得了较好的检测性能,但由于融合位置和信息建模方式的差异,其性能表现存在一定区别。其中,模态级融合方法主要在模型输入阶段完成 RGB 与 X 模态信息的组织与整合,能够直接利用不同传感器之间的互补特性,但由于融合发生于特征提取之前,对高层语义关系的建模能力相对有限,因此整体性能通常低于后续深层融合方法。特征级融合方法进一步作用于基础模型内部特征空

间,通过交叉注意力等机制促进 RGB 与异构模态之间的信息交互,在目标区域定位和细节边界建模方面具有更强优势。例如,KAN-SAM(Li 等,2025b)利用 SAM 提供的开放世界视觉先验,在特征交互过程中增强目标结构表达,相较同样利用 SAM2 作为驱动单元的 SAMSOD(Liu 等,2026),其 F_{max} 指标提升 1.5%,体现了特征级融合结合通用视觉先验对复杂目标感知的促进作用。相比之下,表征级融合方法进一步将 RGB 与 X 模态映射至共享语义空间,通过学习跨模态一致的高层表示实现语义对齐,因此在缓解模态分布差异和提升融合一致性方面具有优势。例如,ComPtr(Pang 等,2025)通过构建统一跨模态表示,在多个评价指标上取得了较优性能,表明基于共享语义空间的融合机制能够有效提升 RGB-X 显著目标检测方法的综合表现。

总结表 2 可知,采用不同融合层级的 RGB-X 语义分割方法在 RGB-T 与 RGB-D 等数据集上均展现出良好的性能,但不同融合机制在信息利用方式和适用场景方面存在差异。由于语义分割属于像素级密集预测任务,不仅需要充分利用不同模态提供的局部细节信息,还要求建立稳定的跨模态语义对应关系,因此融合层级对最终分割性能具有重要影响。其中,特征级融合方法主要作用于基础模型内部特征空间,通过跨模态特征交互增强边界、纹理等局部信息表达,在复杂目标区域具有较好的细节保持能力。例如,相关方法(Cai 等,2025)利用通用模

型提供的视觉先验,通过特征交互机制提升了 RGB-D 和 RGB-T 场景下的像素级预测能力。相比之下,表征级融合方法进一步在共享语义空间中实现 RGB 与异构模态的统一表示,通过降低模态分布差异增强跨场景语义一致性,因此在多模态条件下具有更强的泛化能力。典型代表如 CMX(Zhang 等,2023a)以及相关统一表示方法,能够同时支持 RGB-D 和 RGB-T 等多种融合场景,实现统一的像素级语义分割。此外,部分方法通过引入 Prompt 等参数高效适配策略,在已有融合层级基础上进一步增强基础模型能力,例如 CPAL-L(Liu 等,2025b)在 NYU Depth V2 和 SUN-RGBD 数据集上取得较优分割性能,表明轻量化适配机制能够有效提升基础模型在 RGB-X 任务中的迁移能力。

总体来看,语义分割任务的发展趋势表明,RGB-X 融合正在由局部特征交互逐渐向共享语义

表示学习方向演进,通过在更高层级建立跨模态一致性,为复杂场景下的稳定感知提供支持。

表 2 采用不同融合层级的 RGB-X 分割方法在 RGB-T 与 RGB-D 数据集上的性能比较

Table 2 Performance comparison of RGB-X segmentation methods employing different fusion levels on RGB-T and RGB-D datasets.

注:“-”代表文中未提供该指标推理结果,“-”表示该工作中未将对应数据集引入评测基准;自上而下按照不

同色度区域划分对应算法,分别对应“模态级融合(蓝)”、“特征级融合(紫)”及“表征级融合(绿)”。

分析表 3 可知,不同融合层级的 RGB-X 目标跟踪方法在不同数据集上的性能表现存在一定差异。需要说明的是,GTOT、RGBT234 和 LasHeR 的具体评价协议并不完全一致(参考式 8),其中 PR 采用的距离阈值存在差异,LasHeR 部分工作还进一步报告 NPR(Normalized Precision Rate)指标。因此,本文仅基于相同数据集、相同评价指标及评价协议对不同方法进行性能分析,不进行跨数据集的直接性能排序。

由于目标跟踪任务不仅需要融合 RGB 与 X 模态的互补信息,还要求在连续视频序列中保持目标身份一致性和状态稳定,因此融合方法需要同时考虑跨模态信息交互与时序关联建模。其中,特征级融合方法主要利用基础模型内部特征空间,通过跨模态特征交互增强目标区域表征,在短时间尺度下能够有效利用 RGB 与热红外信息互补,提高目标定位精度。相比之下,模态级融合方法通过在输入阶段组织多源信息,具有较好的结构灵活性和数据适应能力,但由于缺少深层语义交互,对于复杂背景和外观变化的建模能力相对有限。例如,GOLA-L(Shao 等,2026)通过引入视觉-语言联合 Prompt 增强基础模型的跨模态理解能力,在 RGBT234 数据集相同评价协议下取得了较优 PR 指标,相较 MoKA-HP(Wu 等,2026)提升约 1.5%,说明结合基础模型先验信息能够有效提升复杂场景下的目标判别能力。

另一方面,表征级融合方法通过构建共享语义空间增强 RGB 与 X 模态之间的一致性,在跨场景迁移和模态差异缓解方面具有潜力。然而,由于目标跟踪任务高度依赖连续帧之间的目标状态变化和身份关联,仅依靠静态统一表示仍难以充分建模动态

表 1 基于不同融合层级的 RGB-X 显著目标检测方法在 VT5000、VT1000 和 VT821 数据集上的性能比较

Table 1 Performance comparison of RGB-X salient object detection methods based on different fusion levels on the VT5000, VT1000, and VT821 datasets.

注：“-”代表文中未提供该指标推理结果。自上而下按照不同色度区域划分对应算法，分别对应“模态级融合(蓝)”、“特征级融合(紫)”及“表征级融合(绿)”。

识别算法	VT5000						VT1000						VT821					
	F _{avg}	F _{max}	F _w	M	E _m	S _m	F _{avg}	F _{max}	F _w	M	E _m	S _m	F _{avg}	F _{max}	F _w	M	E _m	S _m
HyPSAM (Hou 等, 2026b)	92.8	93.9	91.1	1.9	96.3	93.1	94.6	95.7	94.4	1.1	96.5	95.4	91.4	93.0	90.3	2.0	94.8	93.2
DPLNet (Dong 等, 2024)	-	82.8	-	3.8	91.6	87.9	-	88.1	-	2.2	95.1	92.8	-	81.0	-	4.3	90.8	87.8
KAN-SAM (Li 等, 2025b)	90.9	93.1	90.5	2.0	95.7	92.7	93.0	94.7	93.4	1.3	95.8	94.6	88.3	91.1	88.0	2.5	93.2	91.5
SAMSOD (Liu 等, 2026)	-	91.6	-	2.1	95.6	92.4	-	94.9	-	1.2	98.1	94.8	-	93.3	-	1.7	96.9	93.4
H RTP (He 等, 2024c)	-	91.8	-	1.7	95.1	91.1	-	96.3	-	1.1	98.3	95.5	-	95.0	-	1.7	97.1	93.9
DiffSOD (Song 等, 2025)	-	90.8	-	2.6	93.5	91.8	-	93.7	-	2.1	96.9	95.1	-	88.6	-	3.1	95.6	91.1
AiOSOD (Jia 等, 2023)	-	87.5	-	3.46	93.8	89.6	-	94.1	-	2.0	97.4	94.1	-	88.2	-	2.8	93.7	90.5
ComPtr (Pang 等, 2025)	-	91.9	88.3	2.21	96.3	92.8	-	95.4	92.7	1.5	98.0	95.1	-	90.7	87.5	2.2	95.0	92.4
M-SpecGene (Zhou 等, 2025b)	-	87.2	-	2.8	92.8	89.2	-	92.5	-	1.5	95.2	93.5	-	86.2	-	2.8	91.9	89.1
UniRGB-IR (Yuan 等, 2025)	-	84.9	-	2.7	93.5	90.6	-	89.4	-	1.8	94.3	93.9	-	80.6	-	3.9	89.5	88.1

运动过程,因此当前表征级融合方法在跟踪任务中的优势尚未完全体现。总体来看,RGB-X 目标跟踪的发展趋势正在由单纯的模态信息融合逐渐向结合时序建模、语义理解和动态状态预测的融合方向演进。

表 3 采用不同融合层级的 RGB-X 目标跟踪方法性能比较

Table 3 Performance Comparison of RGB-X Object Tracking Methods with Different Fusion Levels

注：“-”代表文中未提供该指标推理结果，“--”表示该工作中未将对应数据集引入评测基准；自上而

下按照不

同色度区域划分对应算法，分别对应“模态级融合(蓝)”、“特征级融合(紫)”及“表征级融合(绿)”。由于不

同数据集的 PR 评价协议存在差异,本文仅在相同数据集及相同评价协议下比较不同方法的性能,不进行跨数据集比较。

4 相关应用

基于通用智能驱动的跨模态融合技术近年来逐
© 中国图象图形学报版权所有

分割算法		发行日期	NYU Depth V2		SUN-RGBD		MFNet Dataset		PST900	
			mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU
DPLNet	(Dong 等, 2024)	2024	--	--	--	--	-	59.3	-	86.7
	GoPT(He, 2024b)	2024	67.4	54.3	64.5	52.3	77.0	57.7	-	81.5
CPAL-L	(Liu 等, 2025b)	2025	84.5	65.8	86.3	58.2	--	--	--	--
MM-SAM	(Xiao 等, 2024)	2024	-	81.4	-	81.2	-	75.9	--	--
TUNI	(Guo 等, 2025)	2025	--	--	--	--	--	--	-	87.3
MiLNet	(Liu 等, 2025a)	2025	--	--	--	--	73.2	60.1	92.99	85.11
	TASeg(Yu 等, 2025)	2025	--	--	--	--	-	77.6	-	86.09
SARTM	(Xing 等, 2025)	2025	--	--	--	--	-	60.03	-	89.88
FusionSAM	(Li 等, 2025a)	2025	--	--	--	--	-	64.4	--	--
SHIFNet	(Zhao 等, 2025)	2025	--	--	--	--	-	59.2	94.1	89.8
	SAM-FuseNet(Zhu 等, 2026)	2026	--	--	--	--	--	--	-	98.92
AdaptRGB-t	(Yu 等, 2026)	2026	--	--	--	--	85.4	77.6	-	86.09
CMX	(Zhang 等, 2023a)	2023	80.1	56.9	83.8	52.4	-	59.7	--	--
CSFNet	(Qashqai 等, 2024)	2024	--	--	--	--	-	59.98	--	--
UniMRSeg	(Zhao 等, 2025)	2025	--	--	-	53.2	--	--	--	--
	Keep balance(Cai 等, 2025)	2025	-	59.9	-	55.0	-	59.2	-	88.7
BiXFormer	(Chen 等, 2025)	2025	-	54.5	--	--	--	--	--	--
ComPtr	(Pang 等, 2025)	2025	--	--	-	52.8	--	--	--	--
MMPX	(Wu 等, 2025a)	2025	-	66.2	-	57.6	--	--	-	87.3
UniRGB-IR	(Yuan 等, 2025)	2025	--	--	--	--	--	--	94.3	82.8
RXSSNet	(Zhang 等, 2026)	2026	-	55.2	--	--	-	61.2	-	87.45
SGFNet-LoME	(Zhang 等, 2026)	2026	--	--	--	--	75.6	57.9	91.5	83.4

步拓展至遥感、医学、农业及具身智能等领域。本章围绕上述典型应用,梳理跨模态融合在不同数据形态与任务需求下的应用方式,并分析其发展趋势。

跟踪算法	发行时间	GTOT		RGBT234		LasHeR	
		PR (%)	SR (%)	PR (%)	SR (%)	PR (%)	SR (%)
ProTrack							
(Yang 等, 2022)	2022	80.8	68.9	78.6	58.7	53.8	42.0
MPLT(Luo 等, 2023)	2023	--	--	88.4	65.7	–	–
M3PT(Wang 等, 2024a)	2024	92.7	76.9	86.6	63.9	70.0	56.1
OneTracker							
(Hong 等, 2024)	2024	--	--	85.7	64.2	67.2	53.8
Tea-Stu Net							
(Luo 等, 2024)	2024	92.6	77.5	88.3	66.1	71.2	56.5
Un-Track							
(Wu 等, 2024b)	2024	--	--	83.7	61.8	66.7	53.6
CFBT(Zeng 等, 2024)	2024	--	--	89.9	65.9	73.2	58.4
LRPD-L(Hu 等, 2025b)	2025	--	--	–	–	80.5	63.8
RFFMNet							
(Zeng 等, 2025)	2025	93.4	76.8	86.5	62.9	68.9	54.1
VFPTTrack							
(Yang 等, 2025d)	2025	--	--	89.0	65.0	73.5	58.5
M ³ Track(Tang 等, 2025)	2025	92.1	77.2	84.5	63.0	64.0	51.1
GOLA-L							
(Shao 等, 2026)	2026	95.3	80.9	92.8	71.3	78.1	61.9
PromptTrack							
(Li 等, 2026)	2026	--	--	91.7	67.2	76.2	60.7
MoKA-HP							
(Wu 等, 2026)	2026	94.4	80.9	91.3	68.9	74.3	59.0
ViPT(zhu 等, 2023b)	2023	61.1	54.0	83.5	61.7	65.1	52.5
TIETracker							
(Xia 等, 2025)	2025	--	--	85.7	63.8	70.0	55.8
AMATrack							
(Ye 等, 2024)	2024	--	--	82.6	61.2	–	–
XTracker(Tan 等, 2025)	2025	--	--	87.8	65.4	73.1	58.7
CSTrack(Feng 等, 2025)	2025	--	--	94.0	70.9	75.6	60.8
LightFC-X							
(Li 等, 2025c)	2025	88.7	71.4	83.4	60.3	64.7	50.7
UASTrack							
(Wang 等, 2025c)	2025	93.3	78.9	87.6	65.1	71.1	57.0
APTrack(Hu 等, 2025d)	2025	--	--	–	–	74.1	58.9
Diff-MM							
(Xuan 等, 2025)	2025	--	--	94.8	71.2	77.9	62.4
STTrack(Hu 等, 2025c)	2025	--	--	89.8	66.7	76.0	60.3
UniSOT(Ma 等, 2026)	2026	--	--	87.4	65.0	68.4	54.8

4.1 遥感与对地观测应用

在复杂自然环境中,单一 RGB 模态难以全面刻画目标的物理属性与环境变化。遥感与对地观测场景具有天然的多模态特征,除可见光数据外,长波红外、合成孔径雷达(Synthetic Aperture Radar, SAR)及高光谱数据等多源信息能够从温度、植被反射率、结构形态及物质组成等维度提供互补信息。随着基础模型的发展,遥感领域开始依托其表征与泛化能力,构建面向多模态地表观测的统一表示框架,将 RGB 与多光谱等数据映射至共享特征空间进行协同建模,推动多种对地观测任务的迁移与泛化。

SatMAE(Cong 等, 2022)是较早将大规模自监督预训练引入遥感领域的视觉基础模型之一,验证了大规模视觉预训练对下游遥感任务的提升作用。随后,由 IBM 与 NASA 联合发布的遥感基础模型 Prithvi(Jakubik 等, 2023)进一步将多光谱与时间序列数据统一 Token 化,拓展了传统 RGB-X 数据处理范式。在此基础上,该团队针对不同应用场景相继提出系列模型,包括森林火灾评估模型 Prithvi Geo-Temporal(Kimura 等, 2024)、气象气候预测模型 Prithvi WxC(Schmude 等, 2024)以及农作物分类与土地覆盖变化模型 Prithvi 等(Nishikado 等, 2024)。这些模型通过统一建模多源物理观测信息,增强了遥感数据的时空表征与任务泛化能力。

在视觉基础模型的推动下,研究者进一步将通用多模态大模型引入遥感任务,提升遥感数据的智能解译能力。例如,首个针对遥感影像任务优化的视觉-语言模型 RemoteCLIP(Liu 等, 2024a)通过结合自然语言与卫星图像信息,提升遥感场景中的跨模态理解与语义检索能力,实现遥感图像的自动识别与语义描述。随后,多种视觉-语言遥感大模型进一步将 RGB-X 感知从目标识别拓展至复杂空间推理,应用于防灾减灾中的交互式损毁评估(Zhan 等, 2025)、全球尺度气象动态推演(Gao 等, 2025a)、地理空间规律的世界建模(Zhang 等, 2024c)以及城市演化与资源可持续规划(Kuckreja 等, 2024)等任务。

4.2 智慧医疗诊断应用

人体组织病理特征及深层代谢变化需要多源信息支撑,单一 RGB 模态难以满足这一需求。随着电磁波、微波等多种成像技术应用于现代医疗场景,逐渐形成了计算机断层扫描(Computed Tomography, CT)、核磁共振(Magnetic Resonance Imaging, MRI)、

超声及热红外成像等多种模态。这些异构数据结合 RGB 图像,能够从解剖结构、血流动力学及组织代谢等维度提供诊断信息。面对异构医学影像数据,传统依赖人工判读和规则的图像分析方法逐渐显现出效率瓶颈与认知局限。深度学习技术由此进入医学影像研究领域(Roth 等, 2015),并在 CNN 时代得到广泛应用。然而,医学模态间难以严格配准以及异构特征间的语义鸿沟,也限制了深度学习在医疗场景中的进一步应用。

随着基础模型的兴起,研究人员开始探索利用其跨域表征与对齐能力解决上述问题。由伦敦大学学院和莫菲尔德眼科团队开发的 RETFound 率先问世(Zhou 等, 2023),是医疗领域较早的人工智能基础模型之一,也是眼科领域的首个基础模型。RETFound 采用自监督学习架构,旨在降低 AI 模型对大规模标注数据的依赖,并已被多个研究团队应用于疾病监测和系统性疾病预测,验证了其泛化能力。随后,基础模型被进一步作为基座应用于医疗诊断,其中,MedSAM(Ma 等, 2024a)是首个扩展 SAM 的医学图像分割架构,通过在大规模医学图像数据集上微调,实现医学影像的像素级预测。在此基础上,面向特定临床场景的演进模型相继出现,包括用于内窥镜多模态手术导航的 Endo-FM(Wang 等, 2023c)、专注于放射影像时空对齐的 Rad-SAM(Khlaut 等, 2025)以及融合病理形态与基因组学信息的 STPath(Huang 等, 2025)等。这些基础模型通过在共享空间中对齐解剖特征、生化指标与影像表征,增强了医疗感知系统的跨模态病灶分析与演变推理能力。

另一方面,通用多模态大模型也逐步向医学场景迁移,通过发挥零样本或少样本泛化能力,结合图像与文本数据实现生物医学视觉问答和对话式诊断辅助。作为早期多模态医学大模型,LLaVA-Med(Li 等, 2023a)以多模态大模型 LLaVA 为基础,在包含 1500 万张生物医学图像-文本对的 PMC-15M 数据集(Zhang 等, 2023b)上进行微调,实现了视觉与文本的多模态融合,为后续医学多模态大模型 Med-Flamingo(Moor 等, 2023)、XrayGPT(Thawakar 等, 2024)和 MUSK(Xiang 等, 2025)等提供了基线。

4.3 精准农业管理应用

农田环境中的农业智能系统不仅需理解作物长势与病虫害分布,还需基于环境感知完成施肥、灌溉、喷药等实时决策与精准控制。早期研究多集中

于多光谱与RGB联合建模,如基于Transformer的多模态作物识别与产量预测模型,通过统一Token化编码不同光谱通道实现跨区域、跨品种迁移学习(Chemmalar等,2024)。随后,多模态预训练理念引入农业场景,以大模型作为底座支撑感知任务,部分研究(Sornapudi等,2024)利用大规模农田图像与环境传感器数据进行自监督训练,并迁移至病害检测(Ngugi等,2024)、杂草识别(Goyal等,2025)及生长阶段估计(Liu等,2024b;Shams等,2025)等下游任务,预训练-微调范式推动了农业智能向基础模型范式发展。

在此基础上,首个面向农业的通用视觉感知框架FoMo4Wheat(Han等,2025)被提出,该架构在大规模农业真实场景数据上进行自监督预训练,并通过轻量化适配完成多种下游任务,标志着农业视觉从“任务定制模型”向“领域基础模型”转变。后续研究(Srivastava等,2025;Nahian A等,2025;Knapen等,2025)沿袭并深化该架构,通过引入更大规模行业数据或更精细适配策略,进一步验证领域基础模型在作物表型分析与田间管理中的泛化能力。另一方面,通用多模态大模型能够同时处理视觉、文本等多种模态,利用先验知识理解作物图像与田间环境,并结合自然语言理解提供诊断与决策支持。AgriGPT-VL(Yang等,2025b)是首个专注视觉-语言双模态的农业大模型,其构建了目前规模最大的农业视觉-语言语料库Agri-3M-VL,实现多模态推理。随后,该团队提出统一语音、视觉与文本的三模态框架AgriGPT-Omni(Yang等,2025c),引入语音模态使模型直接理解农户口语指令并生成语音响应,为低资源地区包容性农业智能提供支撑。当前,多种下游任务(Nguyen等,2026)正引入视觉-文本建模思想,推动多模态农业大模型发展,这些模型逐步从实验室走向田间,在智能诊断、精准施药、产量预测等场景中展现应用潜力,为智慧农业规模化落地提供技术支撑。

4.4 具身智能交互应用

真实物理环境中运行的智能体(Feng,2025b)不仅需要理解场景,还需基于感知结果进行实时决策与动作执行。具身智能强调通过感知-决策-行动持续与环境交互,并刻画物质属性、几何结构与动态变化的趋势。将“Agent”思想引入具身智能的工作(Ahn等,2022)利用大语言模型进行任务规划,通过价值

函数判断任务可行性并驱动智能体执行。随后,谷歌在2022—2023年相继提出的RT-1(Brohan等,2022)、PaLM-E(Driess等,2023)及RT-2(Zitkovich等,2023)成为大模型驱动具身智能的代表性工作。RT-1基于130K+真实机器人演示数据实现从图像到动作Token的映射;PaLM-E和RT-2则将大模型能力迁移至机器人控制,实现从Web-scale VLM到物理世界执行的跨域策略迁移。

上述工作奠定基础模型驱动具身智能的发展方向,后续研究逐步进入规模化与统一建模阶段,多项开源模型尝试构建跨平台、跨任务的通用视觉-语言-行动模型。RT-X(O'Neill等,2024)首次构建跨机器人平台与操作任务的通用控制模型,实现跨机器人结构的策略迁移。OpenVLA(Kim等,2024)将视觉、语言与行动统一建模为通用Transformer框架,支持不同机器人形态的动作Token标准化表示。Octo(Octo等,2024)强调通用策略网络,通过跨任务、多源数据训练构建广泛操作能力。Mobile ALOHA(Fu等,2024)构建低成本全身遥操作系统,采集真实机器人数据并提升移动操作任务性能。

与此同时,研究开始探索智能体的持续学习与自主进化能力,使认知-决策-行动进一步拓展至环境预测与风险规避。谷歌RoboCat(Bousmalis等,2023)构建多任务自进化机器人基础模型,通过持续学习机制在执行新任务时生成训练数据并迭代优化,形成闭环自增强结构。在此基础上,世界模型理念进一步推动具身智能发展。Genie(Bruce等,2024)通过可交互环境模型实现未来状态预测;DPPO(Ren等,2024)将扩散策略与强化学习相结合优化机器人控制;Dreamer系列(Hafner等,2019;Hafner等,2020;Hafner等,2025)通过学习环境潜在动态实现基于预测的决策;Voyager类型终身学习模型(Wang等,2023a;He等,2024a;He等,2025)则探索开放环境下智能体的持续学习(Lyu等,2025)与自主探索。总体而言,近年具身智能研究正转向构建具备内部世界模拟、未来状态预测与数据自生成能力的持续演化系统,在真实物理环境中实现更稳健、安全的行为决策。

5 挑战与展望

由前文可知,RGB-X跨模态感知正逐步由传统
©中国图象图形学报版权所有

显式融合向基础模型驱动的统一感知范式发展。然而,相关研究仍处于快速发展阶段,在诸多方面仍存在潜在问题。故本章总结当前研究面临的主要挑战,并进一步展望未来发展方向。

5.1 现存挑战

挑战一:多模态数据资源不足与标注成本较高。

RGB-X 融合依赖多源传感器的协同采集,而不同模态之间的同步、标定和标注均会增加数据构建成本。目前公开 RGB-X 数据集在规模、模态类型和场景覆盖方面仍较为有限,难以满足基础模型对大规模、多样化多模态数据的需求;不同数据集之间的采集设备和标注标准差异也进一步增加了跨域泛化难度。

挑战二:跨模态数据对齐与配准困难。 RGB、红外、深度及 LiDAR 等传感器在成像机制、视场范围、空间分辨率和安装位置等方面存在显著差异,使不同模态难以实现天然的空间对应。现有方法通常依赖额外的标定和配准过程,这不仅增加了系统复杂度,也限制了跨模态信息的有效交互。

挑战三:跨模态统一表示能力仍有局限。 现有 RGB-X 方法仍多针对检测、分割和跟踪等特定任务进行适配,尚缺乏能够统一支持多任务、多模态感知的通用表示。同时,不同模态之间显著的信息差异也使共享语义空间中的稳定对齐与融合仍面临挑战。

5.2 未来展望

展望一:构建大规模多模态数据与自监督学习机制。

未来需要进一步扩大 RGB-X 数据规模、丰富模态类型与应用场景,并推动数据采集和标注标准化。同时,可结合自监督学习等技术,充分利用未标注多模态数据,降低人工标注成本,为基础模型学习通用跨模态知识提供数据基础。

展望二:探索基于 FM 的跨模态隐式对齐机制。 未来可充分利用基础模型在大规模预训练过程中形成的通用视觉结构和语义知识,通过共享特征空间等方式降低对严格像素级配准的依赖,使不同模态能够在存在空间偏差的情况下实现有效的信息交互。

6 结 论

本文围绕基础模型驱动的跨模态融合这一前沿方向,系统梳理了通用智能赋能 RGB-X 融合的研究进展,并依据融合发生的表示层级,将现有方法归纳为模态级融合、特征级融合和表征级融合三类典型技术路线。进一步探讨了基础模型在不同融合层级中如何促进 RGB 与异构模态之间的信息组织、特征交互与统一语义表示学习。随后,本文梳理了上述融合技术在自动驾驶、智能机器人、医学影像等下游应用中的发展情况,并针对不同下游任务,在统一数据集与评价协议下对代表性方法进行性能比较,进一步分析不同融合层级的性能特点。最后,围绕数据采集与标注、大模型高效适配、开放环境泛化以及动态场景理解等关键问题,总结当前 RGB-X 融合面临的主要挑战。本文旨在为基础模型时代的 RGB-X 跨模态融合研究提供系统的技术参考与未来发展思路。

参考文献(References)

- Ahn M, Brohan A, Brown N, Chebotar Y, Cortes O, David B, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances[EB/OL]. [2026-04-17].
<https://arxiv.org/pdf/2204.01691.pdf>
- Akbar J U M, Kamarulzaman S F, Muzahid A J M, Rahman M A and Uddin M. 2024. A comprehensive review on deep learning assisted computer vision techniques for smart greenhouse agriculture. IEEE Access, 12: 4485-4522. [DOI: 10.1109/ACCESS.2024.3349418]
- Al Nahian M J, Ghosh T, Sheikhi F and Maleki F. 2025. Agri-FM+: A Self-Supervised Foundation Model for Agricultural Vision//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, TN, USA: IEEE: 5511-5523.[DOI: 10.1109/CVPRW67362.2025.00548]
- Alaparthi S and Mishra M. 2020. Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey [EB/OL]. [2026-04-17].
<https://arxiv.org/pdf/2007.01127.pdf>
- Armeni I, Sax S, Zamir A R and Savarese S. 2017. Joint 2d-3d-semantic data for indoor scene understanding[EB/OL]. [2026-04-17].
<https://arxiv.org/pdf/1702.01105.pdf>
- Bai J Z, Bai S, Chu Y F, Cui Z Y, Dang K, Deng X D, et al. 2023.

- Qwen technical report[EB/OL]. [2026-04-17].
<https://arxiv.org/pdf/2309.16609.pdf>
- Bogdoll D, Yang Y, Joseph T, Yazgan M and Zollner J M. 2025. Muvo: A multimodal generative world model for autonomous driving with geometric representations//2025 IEEE Intelligent Vehicles Symposium (IV). Cluj-Napoca, Romania: IEEE: 2243-2250.[DOI: 10.1109/IV64158.2025.11097718]
- Bousmalis K, Vezzani G, Rao D, Devin C, Lee A X, Bauza M, et al. 2023. RoboCat: A self-improving generalist agent for robotic manipulation[EB/OL]. [2026-04-17].
<https://arxiv.org/pdf/2306.11706.pdf>
- Brohan A, Brown N, Carbajal J, Chebotar Y, Dabis J, Finn C, et al. 2022. RT-1: Robotics transformer for real-world control at scale [EB/OL]. [2026-04-17].
<https://arxiv.org/pdf/2212.06817.pdf>
- Bruce J, Dennis M, Edwards A, Parker-Holder J, Shi Y G, Hughes E, et al. 2024. Genie: Generative interactive environments [EB/OL]. [2026-04-17].
<https://arxiv.org/pdf/2402.15391.pdf>
- Cai J X, Su J Z, Li Q, Yang W J, Wang S, Zhao T S, et al. 2025. Keep the Balance: A Parameter-Efficient Symmetrical Framework for RGB+ X Semantic Segmentation//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, TN, USA: IEEE: 10587-10598. [DOI: 10.1109/CVPR52734.2025.00990]
- Carion N, Gustafson L, Hu Y T, Debnath S, Hu R, Suris D, et al. 2025. SAM 3: Segment anything with concepts [EB/OL]. [2026-04-17].
<https://arxiv.org/pdf/2511.16719.pdf>
- Chemmlar Selvi G, Charan H J and Kumar D. 2024. CropViT: a light-weight transformer model for crop disease detection//2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT). Vellore, India: IEEE: 1-6. [DOI: 10.1109/AIIoT58432.2024.10574729]
- Chen J L, Zheng X, Paudel D P, Van Gool L, Murase H and Deguchi D. 2025. BiXFormer: A Robust Framework for Maximizing Modality Effectiveness in Multi-Modal Semantic Segmentation [EB/OL]. [2026-04-17].
<https://arxiv.org/pdf/2506.03675.pdf>
- Chen X, Kang B, Zhu J W, Wang D, Peng H W and Lu H C. 2023. Unified sequence-to-sequence learning for single-and multi-modal visual object tracking[EB/OL]. [2026-04-17].
<https://arxiv.org/pdf/2304.14394.pdf>
- Chen Z, Wu J N, Wang W H, Su W J, Chen G, Xing S, et al. 2024. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle, WA, USA: IEEE: 24185-24198. [DOI: 10.1109/CVPR52733.2024.02283]
- Cheng Y P, Fu H Z, Wei X X, Xiao J J and Cao X C. 2014. Depth enhanced saliency detection method//Proceedings of international conference on internet multimedia computing and service. Xiamen, China: ACM: 23-27.[DOI: 10.1145/2632856.2632866]
- Choi Y H, Bong E J and Kee S C. 2025. LiDAR and Camera Fusion for Joint Depth Completion and Panoptic Segmentation Tasks in a Unified Network for 3D Semantic Segmentation//2025 IEEE Intelligent Vehicles Symposium (IV). Cluj-Napoca, Romania: IEEE: 889-896.[DOI: 10.1109/IV64158.2025.11097722]
- Ciptadi A, Hermans T and Rehag J. 2013. An In Depth View of Saliency/BMVC. Bristol: BMVA: 1-11. [DOI: 10.5244/C.27.112]
- Cong Y Z, Khanna S, Meng C L, Liu P, Rozi E, He Y T, et al. 2022. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. Advances in Neural Information Processing Systems, 35: 197-211.[DOI: 10.52202/068431-0015]
- Dai A, Chang A X, Savva M, Halber M, Funkhouser T and Nießner M. 2017. ScanNet: Richly-annotated 3d reconstructions of indoor scenes//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA: IEEE: 5828-5839. [DOI: 10.1109/CVPR.2017.261]
- Deevi S A, Lee C, Gan L, Nagesh S, Pandey G and Chung S J. 2024. Rgb-x object detection via scene-specific fusion modules//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, HI, USA: IEEE: 7366-7375.[DOI: 10.1109/WACV57701.2024.00720]
- Dong S H, Feng Y H, Yang Q, Huang Y, Liu D F and Fan H. 2024. Efficient multimodal semantic segmentation via dual-prompt learning//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, United Arab Emirates: IEEE: 14196-14203. [DOI: 10.1109/IROS58592.2024.10801872]
- Dong W H, Zhu H D, Lin S H, Luo X Y, Shen Y H, Guo G D, et al. 2025. Fusion-mamba for cross-modality object detection. IEEE Transactions on Multimedia, 27: 7392-7406. [DOI: 10.1109/TMM.2025.3599020]
- Driess D, Xia F, Sajjadi M S M, Lynch C, Chowdhery A, Ichter B, et al. 2023. PaLM-E: An embodied multimodal language model[EB/OL]. [2026-04-17].
<https://arxiv.org/pdf/2303.03378.pdf>
- Duan X M, Sun H M, Zhang Z M, Qin L X and Jia R S. 2025. CMFX: Cross-modal fusion network for RGB-X crowd counting. Neural Networks, 184: 107070.[DOI: 10.1016/j.neunet.2024.107070]
- Duan H C, Xue G S, Zhang J Y, Zhu M L and Zhao Z W. 2026. ATP-VLA: A robust grasping framework driven by temporal asymmetric multimodal fusion and representation isomorphism. IEEE Access, 14: 83044-83054. [DOI: 10.1109/ACCESS. 2026.3698160]
- Fan D P, Cheng M M, Liu Y, Li T and Borji A. 2017. Structure-

- measure: A new way to evaluate foreground maps//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4548-4557.[DOI: 10.1109/ICCV.2017.487]
- Fan D P, Gong C, Cao Y, Ren B, Cheng M M and Borji A. 2018. Enhanced-alignment measure for binary foreground map evaluation//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence: 698-704.[DOI: 10.24963/ijcai.2018/97]
- Fan D P, Lin Z, Zhang Z, Zhu M L and Cheng M M. 2021. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on Neural Networks and Learning Systems*, 32 (5): 2075-2089.[DOI: 10.1109/TNNLS.2020.2996406]
- Fan J Q, Wang F, Chu H Q, Hu X, Cheng Y F and Gao B Z. 2023. MLFNet: Multi-level fusion network for real-time semantic segmentation of autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 8(1): 756-767.[DOI: 10.1109/TIV.2022.3176860]
- Fang Y Q, Shi Y, Bei J and Ren T W. 2024. Semantic-guided RGB-thermal crowd counting with segment anything model//Proceedings of the 2024 International Conference on Multimedia Retrieval. Phuket, Thailand: ACM: 570-578. [DOI: 10.1145/3652583.3658108]
- Fang Y X, Wang W, Xie B H, Sun Q, Wu L, Wang X G, et al. 2023. EVA: Exploring the limits of masked visual representation learning at scale//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Vancouver, BC, Canada: IEEE: 19358-19369.[DOI: 10.1109/CVPR52729.2023.01855]
- Feng W Y, Zhao S Z, Pan C Y, Chang A, Chen Y C, Wang Z K, et al. 2023. Digital twin tracking dataset (dttdd): A new rgb+ depth 3d dataset for longer-range object tracking applications//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Vancouver, BC, Canada: IEEE: 3289-3298.[DOI: 10.1109/CVPRW59228.2023.00331]
- Feng X, Zhang D, Hu S, Li X, Wu M, Zhang J, et al. 2025a. CSTrack: Enhancing rgb-x tracking via compact spatiotemporal features[EB/OL].
<https://arxiv.org/pdf/2505.19434.pdf>
- Feng Y K, Li X W, Liu P W, et al. 2025b. User-customized intention understanding agent for planar scanned images. *Journal of Image and Graphics*, 30(1): 198-211 (冯弋珂, 励雪巍, 刘鹏伟, 等. 2025. 面向平面扫描图像的用户定制意图理解智能体. *中国图象图形学报*, 30(1): 198-211)
- Fu Z P, Zhao T Z and Finn C. 2024. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation[EB/OL].
<https://arxiv.org/pdf/2401.02117.pdf>
- Gao Y, Wu H, Shu R Q, Dong H S, Xu F, Chen R R, et al. 2025a. OneForecast: A Universal Framework for Global and Regional Weather Forecasting[EB/OL]. [2026-04-17].
<https://arxiv.org/pdf/2502.00338.pdf>
- Gao Z, Zhou D M, Cao J D, Liu Y S and Shan Q Q. 2025b. Enhanced RGBT tracking network with semantic generation and historical context. *IEEE Transactions on Instrumentation and Measurement*, 74: 1-17.[DOI: 10.1109/TIM.2025.3551143]
- Ge Z Q, Huang H Z, Zhou M Z, Li J C, Wang G M, Tang S L, et al. 2024. WorldGPT: Empowering llm as multimodal world model//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, VIC, Australia: ACM: 7346-7355.[DOI: 10.1145/3664647.3681488]
- Geiger A, Lenz P and Urtasun R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite//2012 IEEE conference on computer vision and pattern recognition. Providence, RI, USA: IEEE: 3354-3361.[DOI: 10.1109/CVPR.2012.6248074]
- Goff M, Hogan G, Hotz G, du Pare Locmaria A, Raczky K, Schäfer H, et al. 2025. Learning to drive from a world model//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, TN, USA: IEEE: 1964-1973. [DOI: 10.1109/CVPRW67362.2025.00184]
- Goyal R, Nath A, Niranjan U and Sharda R. 2025. Analyzing the performance of deep convolutional neural network models for weed identification in potato fields. *Crop Protection*, 188: 107035. [DOI: 10.1016/j.cropro.2024.107035]
- Guo C, Yang D D, Li C and Song P. 2022. Dual Siamese network for RGBT tracking via fusing predicted position maps. *The Visual Computer*, 38 (7): 2555-2567. [DOI: 10.1007/s00371-021-02131-4]
- Guo J, Ma X J, Wang Y K, Yang M, Liu H P and Li Q. 2026. Flowdreamer: A rgb-d world model with flow-based motion representations for robot manipulation. *IEEE Robotics and Automation Letters*, 11(3): 2466-2473.[DOI: 10.1109/LRA.2026.3653273]
- Guo Y J, Shi L X Y, Chen J Y and Finn C. 2025. Ctrl-world: A controllable generative world model for robot manipulation[EB/OL].
<https://arxiv.org/pdf/2510.10125.pdf>
- Ha Q S, Watanabe K, Karasawa T, Ushiku Y and Harada T. 2017. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Vancouver, BC, Canada: IEEE: 5108-5115.[DOI: 10.1109/IROS.2017.8206396]
- Hafner D, Lillicrap T, Ba J and Norouzi M. 2019. Dream to control: Learning behaviors by latent imagination[EB/OL].
<https://arxiv.org/pdf/1912.01603.pdf>
- Hafner D, Lillicrap T, Norouzi M and Ba J. 2020. Mastering atari with discrete world models[EB/OL]. [2026-04-17].
<https://arxiv.org/pdf/2010.02193.pdf>
- Hafner D, Pasukonis J, Ba J and Lillicrap T. 2025. Mastering diverse control tasks through world models. *Nature*, 640(8059): 647-653.

- [DOI: 10.1038/s41586-025-08744-2]
- Han B, Zhu C, Han D, Yu R, Cao S L, Wu J H, et al. 2025. FoMo4Wheat: Toward reliable crop vision foundation models with globally curated data[EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2509.06907.pdf>
- Hassan M, Stapf S, Rahimi A, Rezende P M B, Haghighi Y, Brüggemann D, et al. 2025. Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 22404-22415. [DOI: 10.1109/CVPR52734.2025.02087]
- He H L, Yao W L, Ma K X, Yu W H, Dai Y, Zhang H M, et al. 2024a. Webvoyager: Building an end-to-end web agent with large multimodal models//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: ACL: 6864-6890.[DOI: 10.18653/v1/2024.acl-long.371]
- He H L, Yao W L, Ma K X, Yu W H, Zhang H M, Fang T Q, et al. 2025. Openwebvoyager: Building multimodal web agents via iterative real-world exploration, feedback and optimization//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: ACL: 27545-27564.[DOI: 10.18653/v1/2025.acl-long.1336]
- He K M, Chen X L, Xie S N, Li Y H, Dollar P and Girshick R. 2022. Masked autoencoders are scalable vision learners//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New Orleans, LA, USA: IEEE: 16000-16009.[DOI: 10.1109/CVPR52688.2022.01553]
- He Q B. 2024b. Prompting multi-modal image segmentation with semantic grouping//Proceedings of the AAAI conference on artificial intelligence. 38(3): 2094-2102.[DOI: 10.1609/aaai.v38i3.27981]
- He S H and Shi L. 2024c. RGB-T salient object detection based on the segment anything model//2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE). Guangzhou, China: IEEE: 1105-1108.[DOI: 10.1109/CISCE62493.2024.10653167]
- Hoang D C, Tan P X, Nguyen A N, Ngo D H, Cao M D, Vu M Q, et al. 2026. MambaAlign: Alignment-aware state-space fusion for RGB-X industrial anomaly detection. Journal of Computational Design and Engineering, 13(1): 514-537.[DOI: 10.1093/jcde/qwaf143]
- Hollier M P, Rimell A N, Hands D S and Voelcker R M. 1999. Multimodal perception. BT Technology Journal, 17(1): 35-46.[DOI: 10.1023/A:1009666623193]
- Hong L Y, Yan S L, Zhang R R, Li W Y, Zhou X Y, Guo P X, et al. 2024. Onetracker: Unifying visual object tracking with foundation models and efficient tuning//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle, WA, USA: IEEE: 19079-19091. [DOI: 10.1109/CVPR52733.2024.01805]
- Hou J, Wen H F, Zhao Q, Ma J B, Li Z P, Wang S, et al. 2026a. A comprehensive survey on the research and development of RGB-T salient object detection. Computational Visual Media, 12(2): 273-303.[DOI: 10.26599/CVM.2025.9450501]
- Hou R C, Li X Y, Ren T W, Zhuo D M, Wu G S and Cao J D. 2026b. HyPSAM: Hybrid Prompt-driven Segment Anything Model for RGB-Thermal Salient Object Detection. IEEE Transactions on Circuits and Systems for Video Technology, 36(3): 2697-2712. [DOI: 10.1109/TCSVT.2025.3613770]
- Hu E J, Shen Y L, Wallis P, Allen-Zhu Z Y, Li Y Z, Wang S, et al. 2021. LoRA: Low-rank adaptation of large language models[EB/OL]. <https://arxiv.org/pdf/2106.09685.pdf>
- Hu G Y, Wang Z H, Li C L, Yuan D Z, He B and Tang J. 2025a. Missingness-aware prompting for modality-missing RGBT tracking. Journal of King Saud University Computer and Information Sciences, 37(6): 128.[DOI: 10.1007/s44443-025-00142-y]
- Hu L P, Wang F S, Chen F M, Sun F M and Li H J. 2025b. Breaking Alignment Barriers: TPS-Driven Semantic Correlation Learning for Alignment-Free RGB-T Salient Object Detection[EB/OL]. <https://arxiv.org/pdf/2512.21856.pdf>
- Hu Q K, Li Y C and Yu W B. 2025c. Exploiting Multimodal Prompt Learning and Distillation for RGB-T Tracking//Proceedings of the 2025 International Conference on Multimedia Retrieval. Chicago, IL, USA: ACM: 469-477.[DOI: 10.1145/3731715.3733332]
- Hu X T, Tai Y, Zhao X, Zhao C, Zhang Z Y, Li J, et al. 2025d. Exploiting multimodal spatial-temporal patterns for video object tracking//Proceedings of the AAAI Conference on Artificial Intelligence. 39(4): 3581-3589.[DOI: 10.1609/aaai.v39i4.32372]
- Hu X T, Zhong B N, Liang Q H, Shi L T, Mo Z Y, Tai Y, et al. 2025e. Adaptive perception for unified visual multi-modal object tracking. IEEE Transactions on Artificial Intelligence, 6(10): 2819-2829.[DOI: 10.1109/TAI.2025.3558482]
- Huang K L, Shi B T, Li X, Li X, Huang S Y and Li Y K. 2022. Multimodal sensor fusion for auto driving perception: A survey[EB/OL]. <https://arxiv.org/pdf/2202.02703.pdf>
- Huang T L, Liu T Y, Babadi M, Ying R and Jin W G. 2025. STPath: a generative foundation model for integrating spatial transcriptions and whole-slide images. npj Digital Medicine, 8(1): 659. [DOI: 10.1038/s41746-025-02020-3]
- Huang R M, Yu H, Zhang Y X, et al. 2026. RGB-D camouflaged object detection with state space model-guided multimodal fusion. Journal of Image and Graphics, 1-16 (黄荣梅, 余宏, 张永选, 等. 2026. 状态空间模型引导多模态融合的RGB-D伪装目标检测. 中国图象图形学报, 1-16[DOI: 10.11834/jig.260176])

- Hui T R, Xun Z Z, Peng F G, Huang J S, Wei X M, Wei X L, et al. 2023. Bridging search region interaction with template for rgb-t tracking//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Vancouver, BC, Canada: IEEE: 13630-13639.[DOI: 10.1109/CVPR52729.2023.01310]
- Hwang S, Park J, Kim N, Choi Y and Kweon I S. 2015. Multispectral pedestrian detection: Benchmark dataset and baseline//Proceedings of the IEEE conference on computer vision and pattern recognition. Boston, MA, USA: IEEE: 1037-1045. [DOI: 10.1109/CVPR.2015.7298706]
- Jakubik J, Roy S, Phillips C E, Fraccaro P, Godwin D, Zadrozny B, et al. 2023. Foundation models for generalist geospatial artificial intelligence[EB/OL]. [2026-04-17].
<https://arxiv.org/pdf/2310.18660.pdf>
- Jia C, Yang Y F, Xia Y, Chen Y T, Parekh Z, Pham H, et al. 2021. Scaling up visual and vision-language representation learning with noisy text supervision//International conference on machine learning. PMLR, 139: 4904-4916.
- Jia X Z, Zhao Z Q, Dongye C L and Zhang Z. 2023. All in one: RGB, RGB-D, and RGB-T salient object detection[EB/OL].
<https://arxiv.org/pdf/2311.14746.pdf>
- Jia D, Zhao C, Zhang H X, et al. 2026. Lightweight RGB-D semantic segmentation network incorporating frequency-domain guidance. *Journal of Image and Graphics*, 31(2): 479-498 (贾迪, 赵辰, 张华修, 等. 2026. 融合频域引导的 RGB-D 轻量级语义分割网络. *中国图象图形学报*, 31(2): 479-498) [DOI: 10.11834/jig.250212]
- Jamal M A, Mohareri O. Multi-modal contrastive masked autoencoders: A two-stage progressive pre-training approach for rgb-d datasets[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 17947-17957.[DOI: 10.1109/cvpr52734.2025.01672]
- Jiang M F, Ma J H, Chen J T, Wang Y M and Fang X. 2024a. PAT-Net: Patch-to-pixel attention-aware transformer network for RGB-D and RGB-T salient object detection. *Knowledge-Based Systems*, 291: 111597.[DOI: 10.1016/j.knosys.2024.111597]
- Jiang X R, Hou Y F, Tian H and Zhu L. 2024b. Mirror complementary transformer network for RGB-thermal salient object detection. *IET Computer Vision*, 18(1): 15-32.[DOI: 10.1049/cvi2.12221]
- Ju R, Ge L, Geng W J, Ren T W and Wu G S. 2014. Depth saliency based on anisotropic center-surround difference//2014 IEEE international conference on image processing (ICIP). Paris, France: IEEE: 1115-1119.[DOI: 10.1109/ICIP.2014.7025222]
- Khlaut J, Ferreres E, Tordjman D, Philippe H, Boeken T, Manceron P, et al. 2025. Radsam: Segmenting 3d radiological images with a 2d promptable model//International Conference on Medical Image Computing and Computer-Assisted Intervention. Daejeon, South Korea: Springer Cham: 435-444. [DOI: 10.1007/978-3-032-04965-0_41]
- Kim M J, Pertsch K, Karamcheti S, Xiao T, Balakrishna A, Nair S, et al. 2024. OpenVLA: An open-source vision-language-action model[EB/OL]. [2026-04-17].
<https://arxiv.org/pdf/2406.09246.pdf>
- Kimura D, Simumba N, Freitag M, Schmude J and Tatsubori M. 2024. Geospatial Sampling by Maximizing Information Entropy//2024 IEEE International Geoscience and Remote Sensing Symposium. Athens, Greece: IEEE: 7483-7487. [DOI: 10.1109/IGARSS53475.2024.10641065]
- Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C, Gustafson L, et al. 2023. Segment anything//Proceedings of the IEEE/CVF international conference on computer vision. Paris, France: IEEE: 4015-4026.[DOI: 10.1109/ICCV51070.2023.00371]
- Knapen R, de Wit A, Buyukkaya E, Petrou P, Paudel D, Janssen S, et al. 2025. Efficient and scalable crop growth simulations using standard big data and distributed computing technologies. *Computers and Electronics in Agriculture*, 236: 110392.[DOI: 10.1016/j.compag.2025.110392]
- Kristan M, Matas J, Leonardis A, Felsberg M, Pflugfelder R, Kämäräinen J K, et al. 2019. The seventh visual object tracking VOT2019 challenge results//Proceedings of the IEEE/CVF international conference on computer vision workshops. Seoul, Korea (South): IEEE: 2206-2241. [DOI: 10.1109/ICCVW.2019.00276]
- Kuckreja K, Danish M S, Naseer M, Das A, Khan S and Khan F S. 2024. Geochat: Grounded large vision-language model for remote sensing//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle, WA, USA: IEEE: 27831-27840.[DOI: 10.1109/CVPR52733.2024.02629]
- Li B, Peng F G, Hui T R, Wei X M, Wei X L, Zhang L J, et al. 2025a. RGB-T tracking with template-bridged search interaction and target-preserved template updating. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1): 634-649.[DOI: 10.1109/TPAMI.2024.3475472]
- Li C L, Cheng H, Hu S Y, Liu X B, Tang J and Lin L. 2016. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing*, 25(12): 5743-5756. [DOI: 10.1109/TIP.2016.2614135]
- Li C L, Xue W L, Jia Y Q, Qu Z C, Luo B, Tang J, et al. 2022a. LasHer: A large-scale high-diversity benchmark for RGBT tracking. *IEEE Transactions on Image Processing*, 31: 392-404.[DOI: 10.1109/TIP.2021.3130533]
- Li C L, Zhao N, Lu Y J, Zhu C L and Tang J. 2017. Weighted sparse representation regularized graph learning for RGB-T object tracking//Proceedings of the 25th ACM international conference on Multimedia. Mountain View, California, USA: ACM: 1856-1864. [DOI: 10.1145/3123266.3123289]
- Li C Y, Song D, Tong R F and Tang M. 2019. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pat-*

- tern Recognition, 85: 161-171. [DOI: 10.1016/j.patcog.2018.08.005]
- Li C Y, Wong C, Zhang S, Usuyama N, Liu H T, Yang J W, et al. 2023a. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day[EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2306.00890.pdf>
- Li D X, Xie W Y, Cao M X, Wang Y K, Zhang Y S, Fang L Y, et al. 2025b. FusionSAM: Visual Multi-Modal Learning with Segment Anything Model//Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2. New York, NY, United States: ACM: 1330-1340. [DOI: 10.1145/3711896.3736973]
- Li H F, Liu G B, Guo X, Zha Y F and Zhang P. 2026. PromptTrack: Streaming spatial-temporal prompt learning for RGB-T tracking. IEEE Transactions on Multimedia, 28: 5175-5189. [DOI: 10.1109/TMM.2026.3664973]
- Li J N, Li D X, Savarese S and Hoi S. 2023b. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the 40th International Conference on Machine Learning. Honolulu, Hawaii, USA: PMLR, 202: 19730-19742.
- Li J N, Li D X, Xiong C M and Hoi S. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation//International conference on machine learning. Baltimore, Maryland, USA: PMLR, 162: 12888-12900.
- Li N Y, Ye J W, Ji Y, Ling H B and Yu J Y. 2014. Saliency detection on light field//Proceedings of the IEEE conference on computer vision and pattern recognition. Columbus, OH, USA: IEEE: 2806-2813. [DOI: 10.1109/CVPR.2014.359]
- Li X Y, Hou R C, Ren T W and Wu G S. 2025c. Kan-sam: Kolmogorov-arnold network guided segment anything model for rgb-t salient object detection//2025 IEEE International Conference on Multimedia and Expo (ICME). Nantes, France: IEEE: 1-6. [DOI: 10.1109/ICME59968.2025.11209979]
- Li Y F, Wang B and Li Y. 2025d. Lightfc-x: Lightweight convolutional tracker for rgb-x tracking[EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2502.18143.pdf>
- Liang D K, Zhang D Y, Zhou X, Tu S F, Feng T R, Li X F, et al. 2025. Seeing the future, perceiving the future: A unified driving world model for future generation and perception[EB/OL]. <https://arxiv.org/pdf/2503.13587v1.pdf>
- Lin J, Zhou X F, Liu J Y, Cong R M, Zhang G D, Liu Z, et al. 2025. SAM-DAQ: Segment Anything Model with Depth-guided Adaptive Queries for RGB-D Video Salient Object Detection[EB/OL]. <https://arxiv.org/pdf/2511.09870.pdf>
- Liu F, Chen D L, Guan Z Q Y, Zhou X C, Zhu J L and Ye Q L. 2024a. Remoteclip: A vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 62: 1-16. [DOI: 10.1109/TGRS.2024.3390838]
- Liu H T, Li C Y, Wu Q Y and Lee Y J. 2023a. Visual instruction tuning//Advances in Neural Information Processing Systems 36. New Orleans, LA, USA: Curran Associates Inc: 36: 34892-34916.
- Liu J F, Liu H, Li X, Ren J L and Xu X H. 2025a. MiLNet: Multiplex interactive learning network for RGB-T semantic segmentation. IEEE Transactions on Image Processing, 34: 1686-1699. [DOI: 10.1109/TIP.2025.3544484]
- Liu N, Zhang N, Shao L and Han J W. 2022. Learning selective mutual attention and contrast for RGB-D saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44 (12): 9026-9042. [DOI: 10.1109/TPAMI.2021.3122139]
- Liu Y, Feng H K, Yue J B, Jin X L, Fan Y G, Chen R Q, et al. 2024b. Improving potato AGB estimation to mitigate phenological stage impacts through depth features from hyperspectral data. Computers and Electronics in Agriculture, 219: 108808. [DOI: 10.1016/j.compag.2024.108808]
- Liu Y, Wu P F, Wang M H and Liu J. 2025b. Cpal: Cross-prompting adapter with loras for rgb+ x semantic segmentation. IEEE Transactions on Circuits and Systems for Video Technology, 35 (6): 5858-5871. [DOI: 10.1109/TCSVT.2025.3536086]
- Liu Z Y, Wang X R, Fang X Y, Tu Z Z and Wang L B. 2026. SAMSOD: Rethinking SAM optimization for RGB-T salient object detection. IEEE Transactions on Multimedia. [DOI: 10.1109/TMM.2026.3654410]
- Liu Y, Zhang W, Zhao G, et al. Test-time adaptation for nighttime color-thermal semantic segmentation [J]. IEEE Transactions on Artificial Intelligence, 2023b, 5 (10): 4893-4904. [DOI: 10.1109/TAI.2023.3336611]
- Liu Z, Liu J Y, Zhang B Z, Ma L, Fan X and Liu R S. 2023c. PAIF: Perception-aware infrared-visible image fusion for attack-tolerant semantic segmentation//Proceedings of the 31st ACM international conference on multimedia. Ottawa, ON, Canada: ACM: 3706-3714. [DOI: 10.1145/3581783.3611928]
- Lu A D, Wang W Y, Li C L, Tang J and Luo B. 2025a. After: Attention-based fusion router for RGBT tracking. IEEE Transactions on Image Processing, 34: 4386-4401. [DOI: 10.1109/TIP.2025.3586467]
- Lu Y X, Wu B, Li Z D, Li K Q, Huang C Y, Wang H C, et al. 2025b. Remote Sensing-Oriented World Model[EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2509.17808.pdf>
- Lukezic A, Kart U, Kapyla J, Durmush A, Kamarainen J K, Matas J, et al. 2019. Cdtb: A color and depth visual object tracking dataset and benchmark//Proceedings of the IEEE/CVF international conference on computer vision. Seoul, Korea (South): IEEE: 10013-10022. [DOI: 10.1109/ICCV.2019.01011]
- Luo Y, Guo X Q and Li H. 2024. From two-stream to one-stream: Efficient RGB-T tracking via mutual prompt learning and knowledge distillation[EB/OL]. [2026-04-17].

- <https://arxiv.org/pdf/2403.16834.pdf>
- Luo Y, Guo X Q, Feng H and Ao L. 2023. RGB-T tracking via multi-modal mutual prompt learning[EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2308.16386.pdf>
- Luo Z Q, Lai J X, Chen R T, Yang J Y, Gao B B, Nie Q, et al. 2025. UMM: A Unified Multi-Modal Model for Low-Level Vision Tasks with Dual-Driven Prompting. 36th British Machine Vision Conference 2025. Sheffield, UK: BMVA: 392.
- Lyu F, Wang L, Li X, Zheng W S, Zhang Z, Zhou T, Hu F Y. 2025. Comprehensive survey of continual learning. *Journal of Image and Graphics*, 30(8):2599-2632 (吕凡,王亮,李玺,郑伟诗,张彰,周涛,胡伏原. 2025. 持续学习研究进展. *中国图象图形学报*, 30(8):2599-2632)[DOI: 10.11834/jig.240661]
- Lyu P F, Yu X S, Yeung P H, Wu C D and Rajapakse J C. 2026. Deep Fourier-embedded Network for RGB and Thermal Salient Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(4): 4226-4241. [DOI: 10.1109/TCSVT. 2025. 3629727]
- Lyu F, Wang L, Li X, Zheng W S, Zhang Z, Zhou T, Hu F Y. 2025. Comprehensive survey of continual learning. *Journal of Image and Graphics*, 30(8):2599-2632 (吕凡,王亮,李玺,郑伟诗,张彰,周涛,胡伏原. 2025. 持续学习研究进展. *中国图象图形学报*, 30(8):2599-2632)[DOI: 10.11834/jig.240661]
- Ma J, He Y T, Li F F, Han L, You C Y and Wang B. 2024a. Segment anything in medical images. *Nature Communications*, 15(1): 654.[DOI: 10.1038/s41467-024-44824-z]
- Ma Y C, Tang Y Y, Yang W F, Zhang T Z, Zhou X and Wu F. 2026. UniSOT: A Unified Framework for Multi-Modality Single Object Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2): 1554-1572. [DOI: 10.1109/TPAMI.2025. 3615714]
- Ma Y X, Wang T, Bai X Y, Yang H T, Hou Y N, Wang Y M, et al. 2024b. Vision-centric bev perception: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 10978-10997.[DOI: 10.1109/TPAMI.2024.3449912]
- Mazzaglia P, Verbelen T, Dhoedt B, Courville A and Rajeswar S. 2024. Genrl: Multimodal-foundation world models for generalization in embodied agents. *Advances in neural information processing systems*, 37: 27529-27555.[DOI: 10.52202/079017-0865]
- Moor M, Huang Q, Wu S, Yasunaga M, Dalmia Y, Leskovec J, et al. 2023. Med-Flamingo: A multimodal medical few-shot learner// *Proceedings of the 3rd Machine Learning for Health Symposium*. New Orleans, Louisiana, USA: PMLR, 225: 353-367.
- Muturi T W, Kyem B A, Asamoah J K, Owor N J, Dyzinela R, Danyo A, et al. 2025. Prompt-guided spatial understanding with RGB-D transformers for fine-grained object relation reasoning//*Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision Workshops*. Honolulu, HI, USA: IEEE: 5339-5347. [DOI: 10.1109/ICCVW69036.2025.00557]
- Ngugi H N, Ezugwu A E, Akinyelu A A and Abualigahet L. 2024. Revolutionizing crop disease detection with computational deep learning: a comprehensive review. *Environmental Monitoring and Assessment*, 196(3): 302.[DOI: 10.1007/s10661-024-12454-z]
- Nguyen Quoc K, Le Thi Thu L and Quach L D. 2026. A vision-language foundation model for leaf disease identification. *Expert Systems with Applications*, 299: 130084.[DOI: 10.1016/j.eswa. 2025.130084]
- Niu Y Z, Geng Y J, Li X Q and Liu F. 2012. Leveraging stereopsis for saliency analysis//2012 IEEE conference on computer vision and pattern recognition. Providence, RI, USA: IEEE: 454-461. [DOI: 10.1109/CVPR.2012.6247708]
- Octo Model Team, Ghosh D, Walke H, Pertsch K, Black K, Mees O, et al. 2024. Octo: An open-source generalist robot policy [EB/OL]. <https://arxiv.org/pdf/2405.12213.pdf>
- Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. 2023. DiNOv2: Learning robust visual features without supervision[EB/OL]. <https://arxiv.org/pdf/2304.07193.pdf>
- Orfaig E, Stainvas I and Bilik I. 2026. RGBX-DiffusionDet: a framework for multi-modal RGB-X object detection using DiffusionDet. *Pattern Recognition*, 172: 112460.[DOI: 10.1016/j.patcog.2025. 112460]
- Ou Z, Ying G, Zhang D W and Zheng Z L. 2024. A survey of rgb-depth object tracking. *Journal of Computer-Aided Design & Computer Graphics*, 36(11): 1673-1690.[DOI: 10.3724/sp.j.1089. 2024.2023-00537]
- O'Neill A, Rehman A, Maddukuri A, Gupta A, Padalkar A, Lee A, et al. 2024. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration0//2024 IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE: 6892-6903. [DOI: 10.1109/ICRA57147. 2024.10611477]
- Pang Y W, Zhao X Q, Zhang L H and Lu H C. 2025. ComPtr: Towards diverse bi-source dense prediction tasks via a simple yet general complementary transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10): 8613-8629.[DOI: 10.1109/TPAMI.2025.3578494]
- Peng H W, Li B, Xiong W H, Hu W M and Ji R R. 2014. RGBD salient object detection: A benchmark and algorithms//*European conference on computer vision*. Zurich, Switzerland: Springer Cham: 92-109.[DOI: 10.1007/978-3-319-10578-9_7]
- Piao Y R, Ji W, Li J J, Zhang M and Lu H C. 2019. Depth-induced multi-scale recurrent attention network for saliency detection//*Proceedings of the IEEE/CVF international conference on computer vision*. Seoul, Korea (South): IEEE: 7254-7263.[DOI: 10.1109/ICCV.2019.00735]
- Qashqai D, Mousavian E, Shokouhi S B and Mirzakuchaki S. 2024.

- CSFNet: A cosine similarity fusion network for real-time RGB-X semantic segmentation of driving scenes[EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2407.01328.pdf>
- Qi J, Ma L, Cui Z C and Yu Y S. 2024. Computer vision-based hand gesture recognition for human-robot interaction: a review. *Complex & Intelligent Systems*, 10(1): 1581-1606.[DOI: 10.1007/s40747-023-01173-6]
- Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//*Proceedings of the 38th International Conference on Machine Learning*. Virtual: PMLR, 139: 8748-8763.
- Ravi N, Gabeur V, Hu Y T, Hu R H, Ryali C, Ma T Y, et al. 2024. Sam 2: Segment anything in images and videos[EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2408.00714.pdf>
- Ren A Z, Lidard J, Ankile L L, Simeonov A, Agrawal P, Majumdar A, et al. 2024. Diffusion policy policy optimization[EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2409.00588.pdf>
- Ren G Y, Liu H Y, Lazarou M and Stathaki T. 2025. Multi-modal Segment Anything Model for Camouflaged Scene Segmentation//*Proceedings of the IEEE/CVF International Conference on Computer Vision*. Honolulu, HI, USA: IEEE: 19882-19892.
- Roth H R, Lee C T, Shin H C, Seff A, Kim L, Yao J H, et al. 2015. Anatomy-specific classification of medical images using deep convolutional nets//*2015 IEEE 12th international symposium on biomedical imaging (ISBI)*. Brooklyn, NY, USA: IEEE: 101-104. [DOI: 10.1109/ISBI.2015.7163826]
- Salas-Espinales A, Vázquez-Martín R and Mandow A. 2026. A SAM2-Driven RGB-T Annotation Pipeline with Thermal-Guided Refinement for Semantic Segmentation in Search-and-Rescue Scenes. *Modelling*, 7(2): 50.[DOI: 10.3390/modelling7020050]
- Savitha, Nishikado K, Sukumar A, Swain S and Vullam K. 2024. IBM-NASA Geo-Spatial Model Prithvi 100M for Predicting Agricultural Machinery Demand//*2024 International Conference on Cybernation and Computation (CYBERCOM)*. Dehradun, India: IEEE: 383-387.[DOI: 10.1109/CYBERCOM63683.2024.10803197]
- Schmude J, Roy S, Trojak W, Jakubik J, Civitarese D S, Singh S, et al. 2024. Prithvi wxc: Foundation model for weather and climate [EB/OL]. <https://arxiv.org/pdf/2409.13598.pdf>
- Shams M Y, Elmessery W M, Oraith A A T, Elbeltagi A, Salem A, Kumar P, et al. 2025. Automated on-site broiler live weight estimation through YOLO-based segmentation. *Smart Agricultural Technology*, 10: 100828.[DOI: 10.1016/j.atech.2025.100828]
- Shang Y, Zhang X, Tang Y Z, Jin L, Gao C, Wu W, et al. 2025. Roboscape: Physics-informed embodied world model [EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2506.23135.pdf>
- Shao Z K, Hu Y F, Fan B and Liu H M. 2025. PURA: Parameter Update-Recovery Test-Time Adaption for RGB-T Tracking//*Proceedings of the Computer Vision and Pattern Recognition Conference*. Nashville, TN, USA: IEEE: 22089-22098.[DOI: 10.1109/CVPR52734.2025.02057]
- Shao Z K, Hu Y F, Liu J Y, Fan B and Liu H M. 2026. Group orthogonal low-rank adaptation for RGB-T tracking//*Proceedings of the AAAI Conference on Artificial Intelligence*. 40(11): 8887-8895. [DOI: 10.1609/aaai.v40i11.37843]
- Shivakumar S S, Rodrigues N, Zhou A, Miller I D, Kumar V and Taylor C J. 2020. PST900: Rgb-thermal calibration, dataset and segmentation network//*2020 IEEE international conference on robotics and automation (ICRA)*. Paris, France: IEEE: 9441-9447. [DOI: 10.1109/ICRA40945.2020.9196831]
- Silberman N and Fergus R. 2011. Indoor scene segmentation using a structured light sensor//*2011 IEEE international conference on computer vision workshops (ICCV workshops)*. Barcelona, Spain: IEEE: 601-608.[DOI: 10.1109/ICCVW.2011.6130298]
- Silberman N, Hoiem D, Kohli P and Fergus R. 2012. Indoor segmentation and support inference from rgbd images//*European conference on computer vision*. Florence, Italy: Springer Berlin, Heidelberg: 746-760.[DOI: 10.1007/978-3-642-33715-4_54]
- Siméoni O, Vo H V, Seitzer M, Baldassarre F, Oquab M, Jose C, et al. 2025.DINOV3[EB/OL]. <https://arxiv.org/pdf/2508.10104.pdf>
- Simonyan K and Zisserman A. 2014. Two-stream convolutional networks for action recognition in videos//*Advances in Neural Information Processing Systems 27*. Montreal, Canada: Curran Associates Inc: 568-576.
- Song K C, Huang L M, Gong A J and Yan Y H. 2023a. Multiple graph affinity interactive network and a variable illumination dataset for RGBT image salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7): 3104-3118. [DOI: 10.1109/TCSVT.2022.3233131]
- Song K C, Zhao Y, Huang L M, Yan Y H and Meng Q G. 2023b. RGB-T image analysis technology and application: A survey. *Engineering Applications of Artificial Intelligence*, 120: 105919.[DOI: 10.1016/j.engappai.2023.105919]
- Song S R and Xiao J X. 2013. Tracking revisited using RGBD camera: Unified benchmark and baselines//*Proceedings of the IEEE international conference on computer vision*. Sydney, NSW, Australia: IEEE: 233-240.[DOI: 10.1109/ICCV.2013.36]
- Song S R, Lichtenberg S P and Xiao J X. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite//*Proceedings of the IEEE conference on computer vision and pattern recognition*. Boston, MA, USA: IEEE: 567-576. [DOI: 10.1109/CVPR. 2015. 7298655]
- Song X G, Tan Y P, Guo F Q, et al. 2025. Cross-modal feature fusion and detail-enhanced RGB-D salient object detection. *Journal of*

- Image and Graphics, 30(12):3838-3854 (宋 霄罡, 谭裕平, 郭富强, 等. 2025. 跨模态特征融合与细节信息增强的 RGB-D 显著目标检测. 中国图象图形学报, 30(12):3838-3854) [DOI: 10.11834/jig.240653]
- Sornapudi S and Singh R. 2024. Self-supervised backbone framework for diverse agricultural vision tasks[EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2403.15248.pdf>
- Srivastava M, Sisaudia V and Meena J. 2025. AgriTL-ViT: A Vision Transformer model with attention techniques for classification of plant leaf disease. Expert Systems with Applications, 294: 128793.[DOI: 10.1016/j.eswa.2025.128793]
- Tan Y D, Wu Z W, Fu Y Q, Zhou Z Y, Sun G L, Zamfir E, et al. 2025. Xtrack: Multimodal training boosts rgb-x video object trackers//Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu, Hawai'i, USA: IEEE: 5734-5744.
- Tang Z Y, Xu T Y, Wu X J and Kittler J. 2025. M3Track: Meta-Prompt for Multi-Modal Tracking. IEEE Signal Processing Letters, 32: 1705-1709.[DOI: 10.1109/LSP.2025.3558116]
- Thawakar O C, Shaker A M, Mullapilly S S, Cholakkal H, Anwer R M, Khan S, et al. 2024. Xraygpt: Chest radiographs summarization using large medical vision-language models//Proceedings of the 23rd workshop on biomedical natural language processing. Bangkok, Thailand: ACL: 440-448.[DOI: 10.18653/v1/2024.bionlp-1.35]
- Tian C, Zhou Z K, Yang C, Zhu G Q, Zhong F A and He Z Y. 2026. Modality-decoupled RGB-thermal object detector via query fusion//2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE: 8537-8541.[DOI: 10.1109/ICASSP55912.2026.11461564]
- Tu Z Z, Ma Y, Li Z, Li C L, Xu J M and Liu Y T. 2023. RGBT salient object detection: A large-scale dataset and benchmark. IEEE Transactions on Multimedia, 25: 4163-4176. [DOI: 10.1109/TMM.2022.3171688]
- Tu Z Z, Xia T, Li C L, Wang X X, Ma Y and Tang J. 2020. RGB-T image saliency detection via collaborative graph learning. IEEE Transactions on Multimedia, 22(1): 160-173. [DOI: 10.1109/TMM.2019.2924578]
- Tuzlupinar B, Amasyali M F and Karaca A C. 2026. X-Change: An RGB-Inspired Spectral-Aware Framework for Multi-Task and Explainable Change Captioning. IEEE Geoscience and Remote Sensing Letters, 23: 1-5.[DOI: 10.1109/LGRS.2026.3663866]
- Vertens J, Zörn J and Burgard W. 2020. Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas, NV, USA: IEEE: 8461-8468. [DOI: 10.1109/IROS45743.2020.9341192]
- Wan Z F, Zhang P P, Wang Y H, Yong S L, Stepputis S, Sycara K, et al. 2025. Sigma: Siamese mamba network for multi-modal semantic segmentation//2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Tucson, AZ, USA: IEEE: 1734-1744.[DOI: 10.1109/WACV61041.2025.00176]
- Wang G Z, Li C L, Ma Y P, Zheng A H, Tang J and Luo B. 2018. RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach//Chinese Conference on Image and Graphics Technologies. Beijing, China: Springer Singapore: 359-369.[DOI: 10.1007/978-981-13-1702-6_36]
- Wang G Z, Xie Y Q, Jiang Y F, Mandlekar A, Xiao C W, Zhu Y K, et al. 2023a. Voyager: An open-ended embodied agent with large language models[EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2305.16291.pdf>
- Wang H, Chen H M, Du P C, Si P J, Ji B F, Liu L P, et al. 2025a. Dynamic Enhancement Network With Probability Switch for RGB-T Tracking. IEEE Transactions on Instrumentation and Measurement, 74: 1-13.[DOI: 10.1109/TIM.2025.3633376]
- Wang H Z, Guo H, Chai X L, Mu B Y and Shao F. 2025b. Cognition-inspired dynamic feature integration network for RGB-D and RGB-T salient object detection. IEEE Transactions on Instrumentation and Measurement, 74: 1-16.[DOI: 10.1109/TIM.2025.3600718]
- Wang H, Xu T Y, Tang Z Y, Wu X J and Kittler J. 2025c. UASTrack: A Unified Adaptive Selection Framework with Modality-Customization in Single Object Tracking[EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2502.18220.pdf>
- Wang J P, Song W W, Chen H, Ren J C and Zhao H M. 2025d. Fus-Dreamer: Label-efficient remote sensing world model for multimodal data classification. IEEE Transactions on Geoscience and Remote Sensing, 63: 5702314. [DOI: 10.1109/TGRS.2025.3554862]
- Wang K P, Chen K K, Li C L, Tu Z Z and Luo B. 2025e. Alignment-Free RGB-T Salient Object Detection: A Large-scale Dataset and Progressive Correlation Network//Proceedings of the AAAI Conference on Artificial Intelligence, 39(7): 7780-7788. [DOI: 10.1609/aaai.v39i7.32838]
- Wang Q M, Bai Y Q and Song H X. 2024a. Middle fusion and multi-stage, multi-form prompts for robust RGB-T tracking. Neurocomputing, 596: 127959.[DOI: 10.1016/j.neucom.2024.127959]
- Wang Q W, Yin C, Song H C, Shen T and Gu Y F. 2023b. UTFNet: Uncertainty-guided trustworthy fusion network for RGB-thermal semantic segmentation. IEEE Geoscience and Remote Sensing Letters, 20: 1-5.[DOI: 10.1109/LGRS.2023.3322452]
- Wang S, Chen K, Yu M, et al. CDMRNet: multimodal meta-adaptive reasoning network with dynamic causal modeling and co-evolution of quantum states[J]. Scientific Reports, 2025, 15(1): 26370. [DOI: 10.1038/s41598-025-11237-x]
- Wang X F, Zhu Z, Huang G, Chen X Z, Zhu J G and Lu J W. 2024b. Drivedreamer: Towards real-world-drive world models for autonomous driving//European conference on computer vision. Milan, Italy: Springer Cham: 55-72. [DOI: 10.1007/978-3-031-73195-7]

- 2_4]
- Wang X, Zhang D W, Yang C Z, et al. 2026. Three-branch multistage fusion network for RGB-T object tracking. *Journal of Image and Graphics*, 31(4): 1142-1155 (王炫, 张大伟, 阳诚砖, 等. 2026. 面向RGB-T目标跟踪的三分支多阶段融合网络. *中国图象图形学报*, 31(4): 1142-1155) [DOI: 10.11834/jig.250256]
- Wang Z, Liu C, Zhang S T and Dou Q. 2023c. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train// *International conference on medical image computing and computer-assisted intervention*. Vancouver, BC, Canada: Springer Cham: 101-111. [DOI: 10.1007/978-3-031-43996-4_10]
- Wu P F, Liu Y, Gao H and Liu J. 2025. MMPX: Multi-modal Mamba Prompter to Large Vision Foundation Model for RGB-X Semantic Segmentation//2025 IEEE International Conference on Multimedia and Expo (ICME). Nantes, France: IEEE: 1-6. [DOI: 10.1109/ICME59968.2025.11209345]
- Wu Y H, Jia T, Chang X Y, Wang H and Chen D Y. 2024a. RGB-T saliency detection based on multi-scale modal reasoning interaction. *IEEE Transactions on Instrumentation and Measurement*, 73: 1-15. [DOI: 10.1109/TIM.2024.3419115]
- Wu Z W, Zheng J L, Ren X X, Vasluianu F A, Ma C, Paudel D P, et al. 2024b. Single-model and any-modality for video object tracking//*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Seattle, WA, USA: IEEE: 19156-19166. [DOI: 10.1109/CVPR52733.2024.01812]
- Wu Z X, Chen S, Wang D H and Zhu S Z. 2026. MoKA-HP: Motion-aware KAdaptation with historical prompts for efficient and robust RGB-t tracking. *Neurocomputing*, 665: 132163. [DOI: 10.1016/j.neucom.2025.132163].
- Xia W D, Mao X L, Wu W, Zhu C Z and Li F F. 2025. TIETracker: A CLIP-based RGB-T Tracking via Feature Interaction and Semantic Enhancement//2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hangzhou, China: IEEE: 9677-9682. [DOI: 10.1109/IROS60139.2025.11246825]
- Xiang J X, Wang X Y, Zhang X M, Xi Y H, Eweje F, Chen Y J, et al. 2025. A vision - language foundation model for precision oncology. *Nature*, 638(8051): 769-778. [DOI: 10.1038/s41586-024-08378-w]
- Xiao A R, Xuan W H, Qi H L, Xing Y, Yokoya N and Lu S J. 2024. Segment anything with multiple modalities [EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2408.09085.pdf>
- Xiao J J, Stolkin R, Gao Y Q and Leonardis A. 2018. Robust fusion of color and depth data for RGB-D target tracking using adaptive range-invariant depth models and spatio-temporal consistency constraints. *IEEE Transactions on Cybernetics*, 48(8): 2485-2499. [DOI: 10.1109/TCYB.2017.2740952]
- Xing D, Zhu X X, Zhou W, Lin Q K, Yang H and Wang Y Q. 2025. Segment any RGB-thermal model with language-aided distillation [EB/OL]. <https://arxiv.org/pdf/2505.01950.pdf>
- Xu H, Ma J Y, Le Z L, Jiang J J and Guo X J. 2020. FusionDn: A unified densely connected network for image fusion//*Proceedings of the AAAI conference on artificial intelligence*. New York, New York, USA: AAAI: 12484-12491. [DOI: 10.1609/aaai.v34i07.6936]
- Xu J T, Lu K G and Wang H. 2021. Attention fusion network for multi-spectral semantic segmentation. *Pattern Recognition Letters*, 146: 179-184. [DOI: 10.1016/j.patrec.2021.03.015]
- Xu Y W, Khan T M, Song Y and Meijering E. 2025. Edge deep learning in computer vision and medical diagnostics: a comprehensive survey. *Artificial Intelligence Review*, 58(3): 93. [DOI: 10.1007/s10462-024-11033-5]
- Xuan S Y, Li Z C and Tang J H. 2025. Diff-MM: Exploring Pre-trained Text-to-Image Generation Model for Unified Multi-modal Object Tracking[EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2505.12606.pdf>
- Yan S, Yang J Y, Käpylä J, Zheng F, Leonardis A and Kämäräinen J K. 2021. DepthTrack: Unveiling the Power of RGBD Tracking//*Proceedings of the IEEE/CVF international conference on computer vision*. Montreal, QC, Canada: IEEE: 10725-10733. [DOI: 10.1109/ICCV48922.2021.01055]
- Yang B Q, Li J X and Zeng T. 2025a. A review of environmental perception technology based on multi-sensor information fusion in autonomous driving. *World Electric Vehicle Journal*, 16(1): 20. [DOI: 10.3390/wevj16010020]
- Yang B, Chen Y K, Feng L F, Zhang Y, Xu X, Zhang J Y, et al. 2025b. AgriGPT-VL: Agricultural vision-language understanding suite[EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2510.04002.pdf>
- Yang B, Feng L F, Chen Y K, Zhang Y, Zhang J Y, Xu X, et al. 2025c. AgriGPT-Omni: A Unified Speech-Vision-Text Framework for Multilingual Agricultural Intelligence[EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2512.10624.pdf>
- Yang H T, Zhong B N, Liang Q H, Zhu Z R, Zheng Y Z and Li N. 2025d. Robust RGB-T tracking via learnable visual Fourier prompt fine-tuning and modality fusion prompt generation [EB/OL]. <https://arxiv.org/pdf/2509.19733.pdf>
- Yang J Y, Li Z, Zheng F, Leonardis A and Song J K. 2022. Prompting for multi-modal tracking//*Proceedings of the 30th ACM international conference on multimedia*. Lisboa, Portugal: ACM: 3492-3500. [DOI: 10.1145/3503161.3547851]
- Ye P, Xiao G and Liu J. 2024. AMATrack: A unified network with asymmetric multimodal mixed attention for RGBD tracking. *IEEE Transactions on Instrumentation and Measurement*, 73: 1-11. [DOI: 10.1109/TIM.2024.3421435]

- Yi S, Li J J, Liu X and Yuan X S. 2022. CCAFFMNet: Dual-spectral semantic segmentation network with channel-coordinate attention feature fusion module. *Neurocomputing*, 482: 236-251.[DOI: 10.1016/j.neucom.2021.11.056]
- Yu M, Cui T, Chu Q T, Song W J, Yang Y and Yue Y F. 2025. TASeg: Text-aware RGB-T Semantic Segmentation based on Fine-tuning Vision Foundation Models//2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hangzhou, China: IEEE: 4136-4141. [DOI: 10.1109/IROS60139.2025.11246766]
- Yu M, Yue Y F, Yang Y and Fu M Y. 2026. AdaptRGB-t: Adaptive RGB-t semantic segmentation via efficient parameter-tuning with textual guidance. *Neurocomputing*, 664: 132060.[DOI: 10.1016/j.neucom.2025.132060]
- Yuan M X, Cui B, Zhao T Y, Wang J Y, Fu S, Yang X, et al. 2025. Unirgb-ir: A unified framework for visible-infrared semantic tasks via adapter tuning//Proceedings of the 33rd ACM International Conference on Multimedia. Dublin, Ireland: ACM: 2409-2418. [DOI: 10.1145/3746027.3754806]
- Zeng D S, Luo H T, Li J C and Gao P Y. 2025. RGB-T Tracking via Region Filtering-Fusion and Prompt Learning//International Conference on Autonomous Unmanned Systems. Shenyang, China: Springer Singapore: 148-158. [DOI: 10.1007/978-981-96-3552-8_15]
- Zeng Z R, Liu X T, Sun M, Wang H Y and Liu J. 2024. Cross fusion RGB-T tracking with bi-directional adapter[EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2408.16979.pdf>
- Zhan Y, Xiong Z T and Yuan Y. 2025. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221: 64-77.[DOI: 10.1016/j.isprsjprs.2025.01.020]
- Zhang J M, Liu H Y, Yang K L, Hu X X, Liu R P and Stiefelhagen R. 2023a. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 24 (12): 14679-14694. [DOI: 10.1109/TITS.2023.3300537]
- Zhang P Y, Wang D and Lu H C. 2024a. Multi-modal visual tracking: Review and experimental comparison. *Computational Visual Media*, 10(2): 193-214.[DOI: 10.1007/s41095-023-0345-5]
- Zhang S, Huang J M, Tang W B, Tian L L, Wei Y and Liu J. 2025a. Multi-modal Salient Object Detection via a Unified Diffusion Model//2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad, India: IEEE: 1-5.[DOI: 10.1109/ICASSP49660.2025.10887966]
- Zhang S, Xu Y B, Usuyama N, Xu H, Bagga J, Tinn R, et al. 2023b. Large-Scale Domain-Specific Pretraining for Biomedical Vision-Language Processing[EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2303.00915v1.pdf>
- Zhang T L, He X Y, Jiao Q, Zhang Q and Han J G. 2024b. AMNet: Learning to align multi-modality for RGB-T tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 34 (8): 7386-7400.[DOI: 10.1109/TCSVT.2024.3377471]
- Zhang W, Cai M X, Zhang T, Zhuang Y and Mao X R. 2024c. Earth-GPT: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1-20.[DOI: 10.1109/TGRS.2024.3409624]
- Zhang X C, Ye P, Peng S Y, Liu J and Xiao G. 2020. DSiamMFT: An RGB-T fusion tracking method via dynamic Siamese networks using multi-layer feature fusion. *Signal Processing: Image Communication*, 84: 115756.[DOI: 10.1016/j.image.2019.115756]
- Zhang Y J, Choi S Y and Hong S. 2025b. Memory-efficient cross-modal attention for RGB-X segmentation and crowd counting. *Pattern Recognition*, 162: 111376. [DOI: 10.1016/j.patcog.2025.111376]
- Zhang Y M, Gong S, Xiong K X, Ye X Q, Li X F, Tan X, et al. 2024d. BEVWorld: A multimodal world simulator for autonomous driving via scene-level BEV latents[EB/OL]. [2026-04-17]. <https://arxiv.org/pdf/2407.05679.pdf>
- Zhang Y S, Xie W Y, Hui T L, Li D X, Zhang J Q, Lei J, et al. 2026a. LoME: LoRA-Driven Multimodal Extractor for RGB-X Vision Tasks. *IEEE Transactions on Circuits and Systems for Video Technology*, 36 (2): 1487-1500. [DOI: 10.1109/TCSVT.2025.3601340]
- Zhang Z W, Zhuang Y, Liu Y S and Zhang X T. 2026b. A Novel RGB-X Semantic Segmentation Network With Cross-Modal Feature Reweighting and Local-Global Feature Aggregation. *Pattern Recognition*, 177: 113334.[DOI: 10.1016/j.patcog.2026.113334]
- Zhao G Q, Huang J J, Yan X Y, Wang Z J, Tang J W, Ou Y J, et al. 2024. Open-vocabulary RGB-Thermal semantic segmentation//European Conference on Computer Vision. Milan, Italy: Springer Cham: 304-320.[DOI: 10.1007/978-3-031-72904-1_18]
- Zhao H J, Chen J S, Wang L J and Lu H C. 2023. Arkittrack: a new diverse dataset for tracking using mobile RGB-D data//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada: IEEE: 5126-5135.[DOI: 10.1109/CVPR52729.2023.00496]
- Zhao J Y, Teng F, Luo K, Zhao G Q, Li Z Y, Zheng X, et al. 2025a. Unveiling the potential of Segment Anything Model 2 for RGB-thermal semantic segmentation with language guidance[EB/OL]. <https://arxiv.org/pdf/2503.02581.pdf>
- Zhao K, Gao S H, Wang W G and Cheng M M. 2019. Optimizing the F-measure for threshold-free salient object detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 8849-8857.[DOI: 10.1109/ICCV.2019.00894]
- Zhao X Q, Pang Y W, Yu C Y, Zhang L H, Lu H C, Lu S J, et al.

- 2025b. UniMRSeg: Unified Modality-Relax Segmentation via Hierarchical Self-Supervised Compensation[EB/OL]. [2026-04-17].
https://arxiv.org/pdf/2509.16170.pdf
- Zhen H Y, Sun Q, Zhang H X, Li J Y, Zhou S Y, Du Y L, et al. 2025. Tesseract: learning 4d embodied world models[EB/OL]. [2026-04-17].
https://arxiv.org/pdf/2504.20995.pdf
- Zhou J M, Zhu Q, Wang Y N, Feng M T, Liu J, Huang J N, et al. 2025a. A state space model for multiobject full 3-D information estimation from RGB-D images. *IEEE Transactions on Cybernetics*, 55(5): 2248-2260.[DOI: 10.1109/TCYB.2025.3548788]
- Zhou K L, Yang F Q, Wang S X, Wen B H, Zi C D, Chen L S, et al. 2025b. M-SpecGene: Generalized Foundation Model for RGBT Multispectral Vision//Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu, Hawai'i, USA: IEEE: 7861-7872.
- Zhou Y K, Chia M A, Wagner S K, Ayhan M S, Williamson D J, Struyven R R, et al. 2023. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981): 156-163.[DOI: 10.1038/s41586-023-06555-x]
- Zhu Z J, Zhang L, Li P, et al. 2025. Text image super-resolution based on structure-aware enhancement and cross-modal fusion. *Journal of Image and Graphics*, 30(5): 1364-1376 (朱仲杰, 张磊, 李沛, 等. 2025. 结构感知增强与跨模态融合的文本图像超分辨率. *中国图象图形学报*, 30(5): 1364-1376)[DOI: 10.11834/jig.240559]
- Zhu C B and Li G. 2017. A three-pathway psychobiological framework of salient object detection using stereoscopic technology//Proceedings of the IEEE International Conference on Computer Vision Workshops. Venice, Italy: IEEE: 3008-3014. [DOI: 10.1109/ICCVW.2017.355]
- Zhu C Y, Wang J R, Zhang L L, Liang J, Su Q X and Li B H. 2026. SAM-FuseNet: Segment Anything Guided Multimodal Fusion for RGB - Thermal Aerial Robotic Perception. *IEEE Transactions on Geoscience and Remote Sensing*, 64: 1-12.[DOI: 10.1109/TGRS.2025.3648127]
- Zhu D Y, Chen J, Shen X Q, Li X and Elhoseiny M. 2023a. MiniGPT-4: Enhancing vision-language understanding with advanced large language models[EB/OL]. [2026-04-17].
https://arxiv.org/pdf/2304.10592.pdf
- Zhu J W, Lai S M, Chen X, Wang D and Lu H C. 2023b. Visual prompt multi-modal tracking//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Vancouver, BC, Canada: IEEE: 9516-9526. [DOI: 10.1109/CVPR52729.2023.00918]
- Zhu X F, Xu T Y, Tang Z Y, Wu Z C, Liu H D, Yang X, et al. 2023c. RGBD1K: A large-scale dataset and benchmark for RGB-D object tracking//Proceedings of the AAAI Conference on Artificial Intelligence. 37(3): 3870-3878. [DOI: 10.1609/aaai.v37i3.25500]
- Zitkovich B, Yu T H, Xu S C, Xu P, Xiao T, Xia F, et al. 2023. RT-2: Vision-language-action models transfer web knowledge to robotic control//Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR, 229: 2165-2183.
- Zou X Y, Yang J W, Zhang H, Li F, Li L J, Wang J F, et al. 2023. Segment everything everywhere all at once//Advances in Neural Information Processing Systems 36. New Orleans, LA, USA: Curran Associates Inc: 19769-19782.

作者简介

陈瑜,女,博士,主要从事计算机视觉研究,E-mail: 323107343@qq.com

卢嘉亮,男,本科生,主要从事计算机视觉研究,E-mail: 1368323066@qq.com

励雪巍,通信作者,男,副教授,主要研究领域为计算机视觉与人工智能,如全景场景理解,知识蒸馏,场景图生成,扩散模型等,E-mail:xuweili@sdju.edu.cn

李玺,男,教授,主要研究领域为人工智能,计算机视觉,机器学习,模式识别,数据挖掘等,E-mail:xilizju@zju.edu.cn