E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号:TP18;TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2026)09-3308-13

论文引用格式:Yu D, Zhang C J, Zhang X Y and Zheng X L. 2026. Modality reliability modeling for aerial RGB-IR object detection. Journal of Image and Graphics, 31(9):3308-3320(余东, 张淳杰, 张晓宇, 郑晓龙. 2026. 模态可靠性建模的航空遥感可见光—红外目标检测. 中国图象图形学报, 31(9):3308-3320)[DOI:10.11834/jig.260197]

模态可靠性建模的航空遥感可见光—红外目标检测

余东[1,2],张淳杰[1,2*],张晓宇[3],郑晓龙[4,5]

1. 北京交通大学计算机科学与技术学院信息科学研究所,北京　100044; 2. 北京交通大学计算机科学与技术学院视觉智能交叉创新教育部国际合作联合实验室,北京　100044; 3. 中国科学院信息工程研究所,北京　100190; 4. 中国科学院自动化研究所多模态人工智能系统全国重点实验室,北京　100190; 5. 中国科学院大学人工智能学院,北京　100049

摘　要:**目的**　航空遥感可见光与红外(red-green-blue and infrared, RGB-IR)目标检测中,不同模态对检测任务的贡献会随成像条件动态变化。现有方法虽然能在一定程度上利用条件信息调节模态融合,但对与模态可靠性直接相关的质量属性的显式建模仍然不足,难以根据成像条件变化自适应调节不同模态对检测任务的贡献。针对上述问题,提出一种基于模态可靠性建模的航空遥感RGB-IR目标检测方法,通过语义先验蒸馏引导检测网络学习检测导向的模态可靠性表征,并实现可见光与红外模态的自适应融合。**方法**　首先,构建面向无人机场景的模态质量属性描述数据集,对影响检测性能的关键成像因素进行结构化表达。然后,利用视觉语言模型对属性描述文本进行编码,形成与模态可靠性相关的语义先验,并通过训练阶段的蒸馏监督与属性监督,引导检测网络学习检测导向的模态可靠性表征。最后,从场景级全局可靠性和位置级局部空间可靠性两个层面联合建模可见光与红外模态的有效性,实现面向目标检测的动态自适应融合。**结果**　在DroneVehicle和VEDAI两个公开RGB-IR数据集上,所提方法均取得了较优性能。其中,在DroneVehicle上的mAP@0.5和mAP@0.5:0.95分别达到79.7%和53.7%;在VEDAI上分别达到67.1%和29.8%,并在夜间、弱光及复杂干扰场景下表现出更好的检测精度与鲁棒性。消融实验进一步验证了模态质量属性建模、语义先验蒸馏和全局—局部模态可靠性联合建模的有效性。**结论**　所提方法能够以较低开销将视觉语言模型的模态质量感知能力迁移至检测网络内部,在无需测试阶段额外引入大模型分支的条件下,有效建模复杂成像条件下的模态可靠性变化,提升航空遥感红外与可见光目标检测的精度与鲁棒性。

关键词:航空遥感;RGB-IR目标检测;视觉语言模型;模态可靠性;语义先验蒸馏;自适应融合

Modality reliability modeling for aerial RGB-IR object detection

Yu Dong[1,2], Zhang Chunjie[1,2*], Zhang Xiaoyu[3], Zheng Xiaolong[4,5]

1. Institution of Information Science, School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China; 2. Intelligence + X International Cooperation Joint Laboratory of MOE, School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China; 3. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100190, China; 4. State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; 5. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: **Objective** Aerial RGB-infrared (IR) object detection has received increasing attention in remote sensing

收稿日期:2026-04-14;**修回日期**:2026-06-03;**预印本日期**:2026-06-10

*通信作者:张淳杰　cjzhang@bjtu.edu.cn

基金项目:国家自然科学基金项目(62476021,72434005,62376265);中央高校基本科研业务费专项资金资助(2025JBZX062)

Supported by: National Natural Science Foundation of China(62476021,72434005,62376265); Fundamental Research Funds for the Central Universities(2025JBZX062)

because visible and IR images provide complementary information under complex imaging conditions. Visible images preserve rich texture, color, and structural details, whereas IR images are less sensitive to low illumination and can highlight thermal targets at night or in low-light scenes. However, the contribution of each modality changes with the imaging factors involved, such as illumination, exposure status, texture clarity, target-background contrast, background clutter, and artificial light interference. To date, existing methods have improved detection performance through cross-modal alignment, feature interaction, attention reweighting, and multiscale fusion. Nevertheless, most of them still focus on feature-level fusion and lack explicit modality reliability modeling. In other words, they are unable to sufficiently estimate how reliable each modality is under the current imaging condition, or how much each modality should contribute to detection. While some language-guided and condition-aware methods introduce semantic cues into multimodal detection, they usually rely on coarse scene descriptions, category-level prompts, or additional large-model branches during inference. These strategies are insufficient for characterizing modality quality attributes that are directly related to detection reliability. They may also increase the deployment burden on computation-constrained aerial platforms. To address these issues, this paper proposes a modality reliability modeling method for aerial RGB-IR object detection. This modeling method transfers modality quality perception from a vision-language model to the detector during training, thus enabling adaptive multimodal fusion without additional large-model inference cost. **Method** The proposed method consists of structured modality quality description, semantic prior distillation, and reliability-aware adaptive fusion. First, a modality quality attribute description dataset is constructed for UAV-oriented aerial scenes, thus providing structured supervision for modality reliability learning. Instead of using only category labels or coarse scene tags, the annotation scheme explicitly describes key imaging factors that affect detection performance in both modalities. For the RGB modality, the attributes are illumination condition, exposure status, texture clarity, artificial light interference, and background clutter. For the IR modality, the attributes are target-background contrast, boundary clarity, and background cleanliness. Second, a vision-language model is used to encode the modality quality descriptions and generate semantic priors related to RGB and IR reliability. These priors are used only during training. Here, the detector is guided to learn detection-oriented reliability representations through the combination of semantic distillation with attribute supervision. Thus, modality quality perception is internalized into the visual detection network. Third, a global-local adaptive fusion mechanism is designed on the basis of learned reliability representations. Global scene reliability captures the overall effectiveness of RGB and IR cues under the current imaging condition, while local spatial reliability further adjusts the modality contribution at different spatial positions. In this way, the detector can dynamically fuse RGB and IR features in accordance with scene- and region-level reliability. Given that semantic priors are used only during training, the proposed framework does not require additional text prompts, language branches, or large-model participation during inference. **Result** Experiments are conducted on DroneVehicle and VEDAI, two public aerial RGB-IR object detection datasets. On DroneVehicle, the proposed method achieves 79. 7% mAP@0. 5 and 53. 7% mAP@0. 5:0. 95. On VEDAI, it achieves 67. 1% mAP@0. 5 and 29. 8% mAP@0. 5:0. 95. The method also shows strong robustness in challenging scenarios, especially under nighttime, low-light, and complex interference conditions. Ablation studies further verify the effectiveness of modality quality attribute modeling, semantic prior distillation, and joint global-local modality reliability modeling. Additionally, the proposed method maintains good inference efficiency because no extra large-model branch is introduced during testing. **Conclusion** This paper presents a modality reliability modeling method for aerial RGB-IR object detection. To encode modality quality descriptions and provide semantic supervision, the proposed method uses a vision-language model only during training. The detector then learns reliability-aware representations through semantic distillation and attribute supervision. By jointly modeling global scene reliability and local spatial reliability, the detector can adaptively adjust the contributions of visible and IR modalities under varying imaging conditions. Finally, experimental results on DroneVehicle and VEDAI demonstrate that the proposed method improves detection accuracy and robustness, especially in nighttime, low-light, and cluttered scenes.

Key words: aerial remote sensing; RGB-IR object detection; vision-language model; modality reliability; semantic prior distillation; adaptive fusion

0 引 言

随着智能遥感与无人系统技术的快速发展，面向复杂环境的目标检测已成为航空遥感智能感知中的关键基础任务之一，在空域监测、目标搜索和边境巡检等应用中具有重要价值（张迎梅 等，2026；钱孟豪 等，2026）。受成像条件变化和任务场景多样性的影响，单一可见光或红外成像系统往往难以满足航空遥感任务对全天候、高鲁棒感知的需求。可见光图像具有较强的纹理与结构表达能力，但易受光照变化、夜间成像和恶劣天气干扰；红外图像能够表征目标热辐射特性，在弱光、复杂背景和伪装场景下具有独特优势（Yu等，2022）。由于两种模态在信息表达上具有较强互补性，可见光与红外联合目标检测已成为航空遥感领域提升复杂环境感知能力的重要研究方向（张荣 等，2026）。

近年来，围绕可见光与红外（red-green-blue and infrared，RGB-IR）目标检测，研究者在跨模态对齐、特征交互、注意力加权和多尺度融合等方面开展了大量研究。Sun等人（2022）基于DroneVehicle数据集提出了一种不确定性感知的跨模态检测框架，从模态不确定性加权的角度建模红外与可见光信息之间的互补关系。Yuan等人（2022）进一步提出一种结合平移、尺度和旋转变换的区域级对齐方法，并构建双流特征对齐检测框架，从位置、尺度和角度3个方面缓解跨模态空间偏差问题。随后，Yuan和Wei（2024）的工作通过跨模态交叉注意力学习校准且互补的融合表示，并利用自适应特征采样机制降低全局注意力计算开销；Chen等人（2024）则利用偏移引导的自适应特征对齐策略，进一步缓解跨模态空间错位带来的性能退化。总体来看，这类方法推动了RGB-IR目标检测从简单特征拼接向跨模态对齐、特征交互和互补表示学习发展，在复杂场景下取得了较好的检测性能。

然而，从建模机制上看，现有方法仍大多聚焦于视觉特征层面的融合策略设计，对“不同模态在当前样本中究竟有多可靠、应当贡献多少”这一问题缺乏显式建模。事实上，在航空遥感场景中，模态贡献并非静态不变，而是会随成像条件变化而动态波动。例如，在光照充足且纹理清晰的条件下，可见光模态通常更有利于精细目标识别；而在夜间、弱光或复杂背景干扰条件下，红外模态往往更具辨识优势（Yu等，2026）。若缺乏对这种模态可靠性变化的显式刻画，仅依赖固定融合结构或隐式注意力重加权，容易导致优势模态未被充分利用、劣势模态噪声被过度引入，从而制约复杂环境下的检测鲁棒性。

为提升模型对复杂环境变化的适应能力，近期部分研究开始在多模态检测中引入条件感知或语义引导机制。Chen等人（2023）的工作在可见光与热红外检测中引入光照引导的特征校正与融合模块，根据不同照明条件动态调节可见光与热红外模态在融合过程中的作用。Xiong等人（2026）则在高效多光谱检测框架中引入隐式光照估计，并通过零样本方式获取照明信息用于注意力聚合。在语义引导方面，Kim等人（2025）利用多光谱文本描述和思维链提示引导大语言模型进行跨模态推理，以缓解多光谱行人检测中的模态偏置；Wu等人（2026）利用大语言模型生成类别级文本描述，并提取其语义特征以指导跨模态语义与空间对齐；Chen等人（2025）为可见光与红外图像标注多种成像条件属性，并将其编码为文本提示，以实现条件感知的动态融合；Xiang等人（2026）则构造场景描述文本，并通过语义引导调制对视觉特征进行动态重标定。

上述研究表明，条件信息和语义信息有助于增强多模态检测中的环境感知与跨模态建模能力。然而，相比一般场景目标检测，航空遥感场景下目标通常具有尺度小、分布密集、背景复杂和成像条件变化显著等特点，光照、纹理清晰度、热对比和背景干扰等因素会更直接地影响不同模态对目标判别的有效性。因此，若仅依赖类别语义对齐、场景类别或光照条件等粗粒度提示，仍难以准确刻画当前样本中不同模态的实际可靠性。现有相关方法对与模态可靠性直接相关的模态质量属性缺乏显式建模，难以根据成像条件变化精细调节不同模态对检测任务的贡献。同时，部分方法仍需要在推理阶段额外引入文本提示、条件输入或大模型分支，这会增加航空端侧平台、机载设备和实时感知任务中的计算与部署负担。因此，如何围绕模态质量属性进行显式建模，借助视觉语言模型的模态质量感知能力学习与模态可靠性相关的检测导向表征，并将这种能力以低开销方式迁移至检测网络内部，是一个值得研究的问题。

针对上述问题，本文提出一种基于语义先验蒸馏的模态可靠性自适应融合方法，用于航空遥感场

景下的可见光与红外目标检测。本文并非在推理阶段直接引入大模型参与检测，而是通过训练阶段的语义先验建模与知识蒸馏，将视觉语言模型的模态质量感知能力迁移至检测网络内部。具体而言，本文首先构建面向无人机场景的模态质量属性描述数据集，对影响检测性能的关键成像因素进行结构化表征。随后利用视觉语言模型对模态质量属性描述文本进行编码，获得与模态质量属性对应的语义先验，并结合蒸馏监督与属性监督，引导检测网络学习与模态可靠性相关的检测导向表征。在此基础上，进一步联合建模场景级全局可靠性和位置级局部空间可靠性，实现面向目标检测的RGB-IR动态自适应融合。由于语义先验仅在训练阶段用于监督与蒸馏，测试阶段无需额外引入文本提示或大模型分支，因此该方法能够在保持部署效率的同时提升复杂环境下的多模态检测能力，为算力受限航空端侧平台上的高效多模态目标检测提供了一种可行方案。

本文主要贡献如下：1）面向航空遥感RGB-IR目标检测任务，构建模态质量属性描述数据集，提出一种基于语义先验蒸馏的模态可靠性自适应融合方法，实现对与模态可靠性直接相关质量属性的显式建模，并以低开销方式将视觉语言模型的模态质量感知能力迁移至检测网络内部。2）提出全局—局部模态可靠性联合建模策略，从场景级和位置级两个层面自适应评估不同模态的有效性，实现面向目标检测的动态融合。3）在DroneVehicle和VEDAI两个公开 RGB-IR 数据集上开展实验，结果表明，所提方法在检测精度、复杂环境鲁棒性及端侧部署友好性方面均表现出较好优势。

1 方　法

1.1 网络整体架构

本文提出一种基于语义先验蒸馏的模态可靠性自适应融合方法，用于RGB-IR目标检测。其整体框架如图1所示。

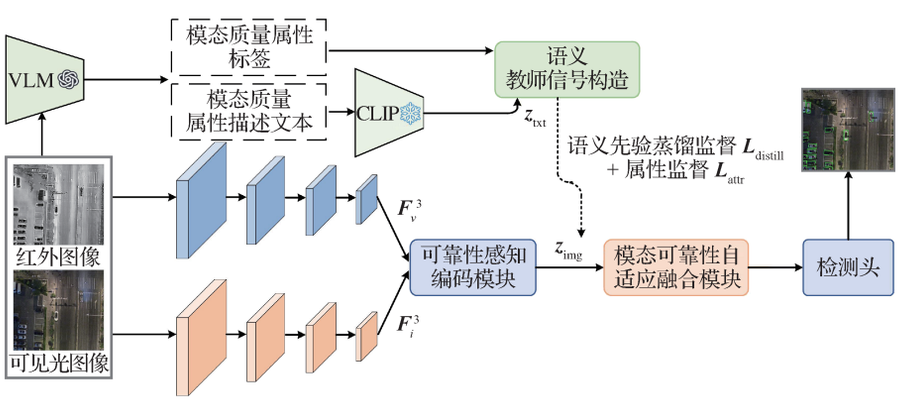


图1　所提方法整体框架

Fig. 1　Overall framework of the proposed method

给定一对配准的可见光图像与红外图像，分别记为$\boldsymbol{X}^v$和$\boldsymbol{X}^i$，首先通过双流主干网络提取两种模态的多尺度特征，具体为

$$\boldsymbol{F}_v^l = B_v^l(\boldsymbol{X}^v),\quad \boldsymbol{F}_i^l = B_i^l(\boldsymbol{X}^i) \tag{1}$$

式中，$l \in \{0,1,2,3\}$表示检测特征层级，B_v^l和B_i^l分别表示可见光与红外分支在第l个尺度上的特征提取。

随后，可靠性感知编码模块基于高层特征提取当前样本的图像端模态可靠性表征z_{img}，具体为

$$z_{\text{img}} = H_{\text{env}}(\boldsymbol{F}_v^3, \boldsymbol{F}_i^3) \tag{2}$$

式中，H_{env}表示可靠性感知编码模块。z_{img}在训练阶段受语义先验蒸馏与模态质量属性监督，在测试阶段作为模态可靠性评估与特征融合的控制信号，无需额外引入文本提示或大模型分支。

接着，模态可靠性自适应融合模块在各个检测尺度上联合建模场景级全局可靠性和位置级局部空间可靠性，对双模态特征执行动态融合，得到融合特征$\boldsymbol{F}_{\text{out}}^1$，具体为

$$\boldsymbol{F}_{\text{out}}^1 = \mathcal{F}_{\text{fuse}}^1(\boldsymbol{F}_v^1, \boldsymbol{F}_i^1, z_{\text{img}}) \tag{3}$$

式中，$\mathcal{F}_{\text{fuse}}^1(\cdot)$表示第$l$个尺度上的可靠性自适应融合函数。

最终，多尺度融合特征$\{\boldsymbol{F}_{\text{out}}^l\}_{l=0}^3$被送入检测头$\mathcal{D}_{\text{head}}$，输出目标类别预测$\boldsymbol{O}_{\text{cls}}$与边界框回归结果$\boldsymbol{O}_{\text{reg}}$，具体为

$$(\boldsymbol{O}_{\text{cls}}, \boldsymbol{O}_{\text{reg}}) = \mathcal{D}_{\text{head}}(\{\boldsymbol{F}_{\text{out}}^l\}_{l=0}^3) \tag{4}$$

1.2 基于语义先验蒸馏的模态质量学习

在实际应用中，若在无人机端侧平台直接引入大模型参与推理，将带来较大的计算与部署负担。为此，本文采用知识蒸馏范式，将视觉语言模型所具备的模态质量感知能力迁移至检测网络内部，引导模型在训练阶段学习与模态可靠性相关的检测导向表征，从而在测试阶段无需额外引入大模型分支，其结构如图2所示。为使蒸馏过程能够围绕航空遥感场景下影响模态有效性的关键成像因素展开，本文进一步构建了面向无人机场景的模态质量属性描述数据集。具体而言，对于每个训练样本，该数据集可提供可见光与红外模态对应的属性描述文本 $\boldsymbol{T}_v$ 和 $\boldsymbol{T}_i$，以及由各属性维度判定结果组成的结构化模态质量属性标签，记为 $\boldsymbol{q}=[\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_K]$。前者用于语义先验建模，后者用于图像端属性监督与结构化关系建模。其具体构建过程见第1.5节。

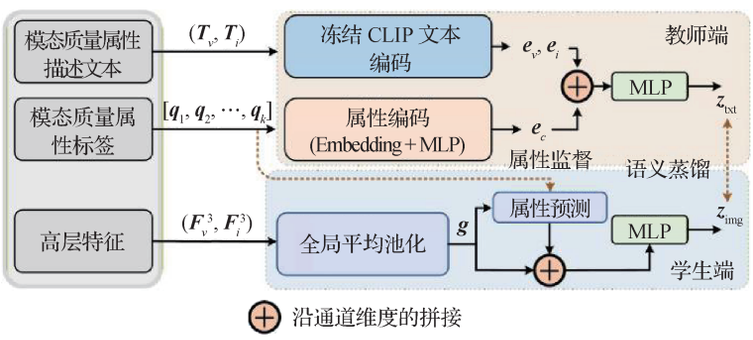


图2 语义先验蒸馏与模态质量学习模块

Fig. 2 Semantic prior distillation and modal quality learning module

1.2.1 语义教师信号构造

首先，针对模态质量属性描述文本 $\boldsymbol{T}_v$ 和 $\boldsymbol{T}_i$，本文采用预训练并冻结的CLIP（contrastive language-image pre-training）ViT-L/14@336px文本编码器 $\mathcal{E}_{\text{text}}$ 对其进行语义编码（Radford等，2021），得到对应的文本特征表示，具体为

$$\boldsymbol{e}_v=\mathcal{E}_{\text{text}}(\boldsymbol{T}_v),\quad \boldsymbol{e}_i=\mathcal{E}_{\text{text}}(\boldsymbol{T}_i) \tag{5}$$

式中，$\boldsymbol{e}_v$ 和 $\boldsymbol{e}_i$ 分别表示可见光与红外模态属性描述的文本语义表征。

进一步地，虽然模态质量属性描述文本可通过CLIP文本编码器形成连续语义特征，但该类特征主要反映整体文本语义，对各质量维度的离散取值缺乏显式约束。已有研究表明，属性级提示和属性级对齐能够弥补仅依赖全局图文表示的不足，增强模型对细粒度视觉属性的感知能力（Liu等，2024a）。因此，本文进一步对模态质量属性标签 $\boldsymbol{q}=[\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_K]$ 进行编码。其中，$\boldsymbol{q}_k$ 表示第 k 个属性的离散取值，$K=8$ 为属性总数。本文首先将每个属性值映射为可学习嵌入，并将所有属性嵌入在通道维度上进行拼接，得到多属性联合表示。随后，采用由两层线性映射和一个修正线性单元（rectified linear unit，ReLU）构成的多层感知机（multilayer perceptron，MLP），对拼接后的属性表示进行非线性聚合，生成属性编码向量 $\boldsymbol{e}_c$。具体表示为

$$\boldsymbol{e}_c=MLP([Emb_1(\boldsymbol{q}_1);Emb_2(\boldsymbol{q}_2);\cdots;Emb_K(\boldsymbol{q}_K)]) \tag{6}$$

式中，Emb_k 表示第 k 个属性的可学习嵌入映射。在获得文本特征 $\boldsymbol{e}_v$、$\boldsymbol{e}_i$ 以及属性编码 $\boldsymbol{e}_c$ 之后，将三者进行拼接，并通过MLP映射到统一语义空间，构造语义教师信号 $\boldsymbol{z}_{\text{txt}}$，具体为

$$\boldsymbol{z}_{\text{txt}}=MLP([\boldsymbol{e}_v;\boldsymbol{e}_i;\boldsymbol{e}_c]) \tag{7}$$

由此得到的语义教师信号包含了模态质量属性相关的先验知识。

1.2.2 图像端学生表示与语义蒸馏

为使模型能够仅依据输入图像学习与模态可靠性相关的检测导向表征，本文进一步构建图像端学生表示，并通过语义先验蒸馏与属性监督对其进行联合约束。整体上，该过程包括全局图像表征提取、属性预测、学生表示构造以及语义蒸馏4个步骤。

首先，本文从主干网络的高层特征 $\boldsymbol{F}_v^3$ 和 $\boldsymbol{F}_i^3$ 中提取全局表征，并在通道维度上进行拼接，得到联合图像表示 $\boldsymbol{g}$，具体为

$$\begin{aligned}&\boldsymbol{g}_v=GAP(\boldsymbol{F}_v^3),\quad \boldsymbol{g}_i=GAP(\boldsymbol{F}_i^3)\\&\boldsymbol{g}=[\boldsymbol{g}_v;\boldsymbol{g}_i]\end{aligned} \tag{8}$$

式中，$GAP(\cdot)$ 表示全局平均池化。该表示融合了可见光与红外模态的高层语义信息，为后续的模态质

量属性预测提供基础。

进一步地，为增强图像表示对关键成像因素的敏感性，本文引入显式属性监督。对于第j个模态质量属性，模型输出预测概率分布$\hat{\boldsymbol{p}}_j$，具体为

$$\hat{\boldsymbol{p}}_j = softmax(Head_j(\boldsymbol{g})),\ \ j = 1,\cdots,K \tag{9}$$

式中，$Head_j(\cdot)$表示第j个属性对应的预测头。通过这一设计，联合图像表示$\boldsymbol{g}$不仅包含双模态语义信息，还被进一步约束去感知与检测性能相关的关键成像质量因素。

随后，将全局特征$\boldsymbol{g}$与各属性的预测分布共同输入由两层线性映射和一个ReLU激活函数构成的MLP，构造图像端学生表示z_{img}，具体为

$$z_{\text{img}} = MLP([\boldsymbol{g};\hat{\boldsymbol{p}}_1;\hat{\boldsymbol{p}}_2;\cdots;\hat{\boldsymbol{p}}_K]) \tag{10}$$

在训练阶段，本文进一步利用语义教师信号z_{txt}对图像学生表示z_{img}进行蒸馏约束，使其向语义先验所蕴含的模态质量感知知识靠拢。对应的蒸馏损失定义为

$$L_{\text{distill}} = 1 - cos(z_{\text{img}}, z_{\text{txt}}) \tag{11}$$

式中，$cos(\cdot,\cdot)$用于衡量图像学生表示与语义教师信号在特征空间中的方向一致性。

与此同时，本文利用真实属性标签$\boldsymbol{q}_j$对各属性预测分支进行监督，其交叉熵损失定义为

$$L_{\text{attr}} = \sum_{j=1}^{K} CE(\hat{\boldsymbol{p}}_j, \boldsymbol{q}_j) \tag{12}$$

图像学生表示z_{img}通过语义先验蒸馏和属性监督学习视觉语言模型中的模态质量感知能力，从而形成与模态可靠性相关的检测导向表征，为后续的全局—局部模态可靠性建模与自适应融合提供支撑。

1.3 全局—局部模态可靠性联合建模与自适应融合

对于RGB-IR目标检测任务而言，跨模态融合的关键不在于简单叠加两种模态特征，而在于根据当前样本的成像条件和局部区域特性，准确判断不同模态在目标判别中的相对有效性。为此，本文从场景级全局可靠性和位置级局部空间可靠性两个层面联合建模模态有效性，并据此执行动态融合，其过程如图3所示。

1.3.1 全局可靠性与局部可靠性估计

场景级成像条件（如昼夜）会决定不同模态在当前样本中的整体有效性。为刻画这种由整体成像条件引起的模态差异，本文首先基于图像模态质量语义表征z_{img}在各检测尺度上预测双模态的场景级全局可靠性先验，具体为

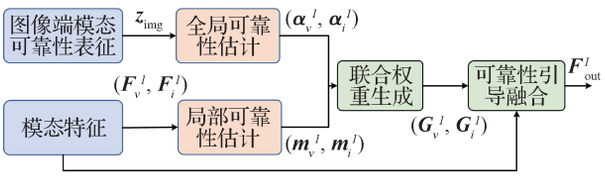


图3 全局—局部模态可靠性引导的自适应融合模块

Fig. 3 Global-local modality reliability-guided adaptive fusion module

$$[\boldsymbol{\alpha}_v^l, \boldsymbol{\alpha}_i^l] = softmax(MLP^l(z_{\text{img}})) \tag{13}$$

式中，$MLP^l(\cdot)$表示第l个检测尺度上的全局可靠性预测分支，它由两层线性映射和一个ReLU激活函数构成。$\boldsymbol{\alpha}_v^l$和$\boldsymbol{\alpha}_i^l$分别表示可见光与红外模态的全局可靠性权重。

然而，仅依赖场景级全局可靠性不足以刻画同一图像内不同空间位置上的模态优势差异。实际场景中，局部区域可能受到成像退化影响，导致不同模态在不同位置上的判别能力不一致。因此，在全局可靠性建模的基础上，本文进一步从位置级对模态有效性进行细粒度刻画，以估计局部空间可靠性。具体而言，本文首先对双模态特征进行通道压缩，得到紧凑的局部表示，具体为

$$\bar{\boldsymbol{F}}_v^l = \phi_v^l(\boldsymbol{F}_v^l),\ \ \bar{\boldsymbol{F}}_i^l = \phi_i^l(\boldsymbol{F}_i^l) \tag{14}$$

式中，$\phi_v^l(\cdot)$和$\phi_i^l(\cdot)$为1×1卷积操作，$\bar{\boldsymbol{F}}_v^l, \bar{\boldsymbol{F}}_i^l \in \mathbf{R}^{1\times H_l\times W_l}$。

随后，将双模态局部特征进行拼接，并映射为共享上下文特征，具体为

$$\boldsymbol{S}^l = \psi^l([\bar{\boldsymbol{F}}_v^l; \bar{\boldsymbol{F}}_i^l]) \tag{15}$$

式中，$\psi^l(\cdot)$表示局部上下文编码函数，它是由两层3×3卷积实现，用于提取当前位置周围的共享双模态上下文信息。

在此基础上，分别结合单模态局部特征与共享上下文信息，预测逐位置的局部可靠性响应，具体为

$$\boldsymbol{E}_v^l = \delta_v^l([\bar{\boldsymbol{F}}_v^l; \boldsymbol{S}^l]),\ \ \boldsymbol{E}_i^l = \delta_i^l([\bar{\boldsymbol{F}}_i^l; \boldsymbol{S}^l]) \tag{16}$$

式中，$\delta_v^l(\cdot)$和$\delta_i^l(\cdot)$由3×3卷积映射实现。再在模态维度上进行归一化，得到位置级局部空间可靠性图，具体为

$$[\boldsymbol{m}_v^l, \boldsymbol{m}_i^l] = softmax([\boldsymbol{E}_v^l, \boldsymbol{E}_i^l]) \tag{17}$$

式中，$\boldsymbol{m}_v^l$和$\boldsymbol{m}_i^l$分别表示第l个尺度上每个空间位置对应的可见光与红外局部可靠性权重，用于反映局部区域内两种模态对目标判别的相对贡献。

1.3.2　联合可靠性引导的动态融合

为同时利用全局和局部的可靠性信息，本文首先在各检测尺度上将场景级全局可靠性和位置级局部空间可靠性图进行联合，生成双模态的融合权重，具体为

$$\boldsymbol{G}_v^l = \boldsymbol{\alpha}_v^l \cdot \boldsymbol{m}_v^l,\quad \boldsymbol{G}_i^l = \boldsymbol{\alpha}_i^l \cdot \boldsymbol{m}_i^l \tag{18}$$

式中，$\boldsymbol{G}_v^l$和$\boldsymbol{G}_i^l$分别表示第l个尺度上可见光与红外模态的联合融合权重，用于综合表征两种模态在当前样本整体层面及具体空间位置上的相对有效性。在此基础上，利用联合权重对双模态特征进行动态加权融合，得到第l个尺度上的融合特征，具体为

$$\boldsymbol{F}_{\text{fuse}}^l = \boldsymbol{G}_v^l \odot \tilde{\boldsymbol{F}}_v^l + \boldsymbol{G}_i^l \odot \tilde{\boldsymbol{F}}_i^l \tag{19}$$

式中，$\odot$表示逐通道乘法。

最后，通过轻量卷积层对融合结果进行空间上下文细化，得到最终输出特征，具体为

$$\boldsymbol{F}_{\text{out}}^l = Conv_{3\times 3}(\boldsymbol{F}_{\text{fuse}}^l) \tag{20}$$

1.4　优化目标

本文采用联合优化策略对整体网络进行训练。网络的优化目标由目标检测损失、模态质量属性监督损失以及语义先验蒸馏损失3部分构成，整体损失函数定义为

$$L = L_{\text{det}} + \lambda_1 L_{\text{attr}} + \lambda_2 L_{\text{distill}} \tag{21}$$

式中，L_{det}为基础检测器的检测损失，L_{attr}表示模态质量属性监督损失，L_{distill}表示语义先验蒸馏损失，λ_1和λ_2为平衡各损失项的超参数。

1.5　模态质量属性描述数据集构建

本文构建了面向无人机场景的模态质量属性描述数据集，用于表征影响可见光与红外模态检测有效性的关键成像因素，并为后续的语义先验建模、属性监督和蒸馏监督提供数据基础。

以DroneVehicle和VEDAI数据集的配准可见光—红外图像对为基础，从检测任务需求出发，定义可见光模态与红外模态的质量属性体系。模态质量属性的选取主要依据其对目标可见性、目标—背景可分性和定位稳定性的影响。对于可见光图像，光照条件、曝光状态和纹理清晰度会影响目标外观结构、边缘细节和局部纹理表达，人工光干扰和背景杂波则增加误检与漏检风险。对于红外图像，目标—背景对比度影响热目标显著性，边界清晰度影响定位精度，背景热杂波会削弱红外响应的判别性。因此，本文将上述因素作为模态质量属性，用以表征不同成像条件下可见光与红外模态对检测任务的相对有效性。在此基础上，本文设计统一的属性判定提示模板，将图像对输入Qwen3-VL-Plus（Bai等，2025）多模态视觉语言模型，对两种模态进行属性分析与判定，并生成结构化属性标签及对应的自然语言描述文本。模态质量属性的构建流程如图4所示，各属性维度及其样本分布统计如图5所示。

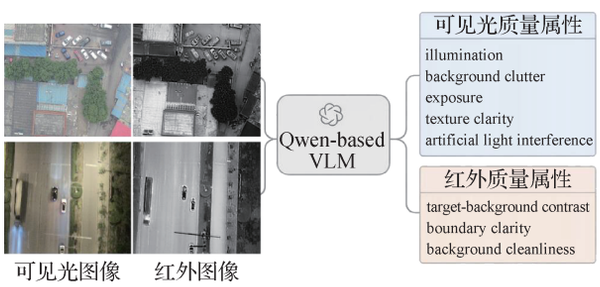


图4　模态质量属性描述数据构建示意图

Fig. 4　Illustration of the construction of modality quality attribute annotations

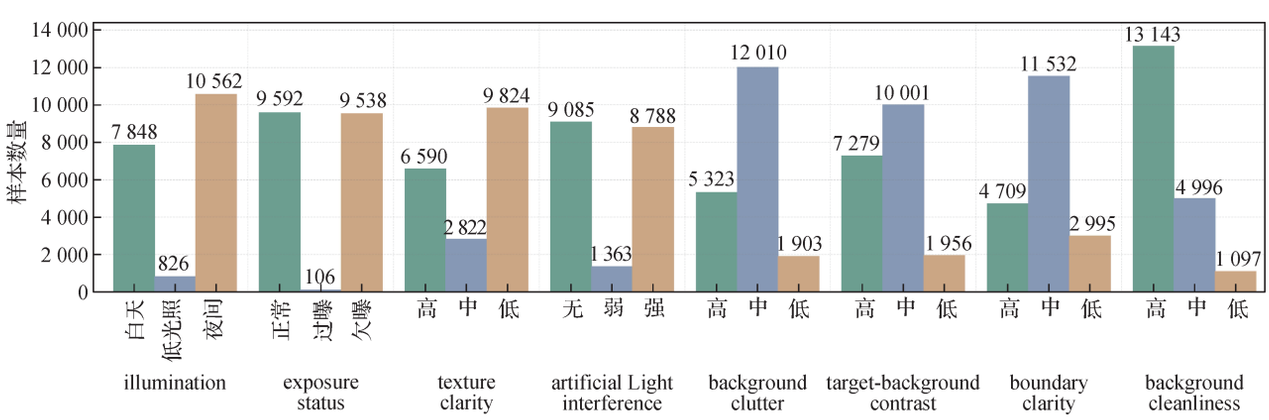


图5　DroneVehicle与VEDAI数据集上的模态质量属性分布统计

Fig. 5　Distribution of modality quality attributes on the DroneVehicle and VEDAI datasets

基于上述属性体系，本文进一步为每个样本生成两种形式的标注结果。第 1 种是结构化属性标签，即将各属性维度的判定结果编码为离散标签；第 2 种是自然语言属性描述，即将同一样本的属性判定结果组织为对应的文本描述。前者用于图像侧属性学习与监督，后者用于提取语义先验，并在训练阶段服务于后续的蒸馏约束。

2 实　验

2.1 实验配置与实现细节

本文实验基于 PyTorch 实现，并在 NVIDIA RTX 3090 GPU（24 GB 显存）上完成。为验证所提方法的有效性，在 DroneVehicle 和 VEDAI（Razakarivony 和 Jurie，2016）两个公开 RGB-IR 航空遥感数据集上开展实验。其中，DroneVehicle 是一个大规模无人机视角多模态目标检测数据集，包含配准的可见光与红外图像对，并提供 5 类车辆目标的有向边界框标注。VEDAI（vehicle detection in aerial imagery）是一个面向高分辨率航空影像车辆检测的数据集，同样提供 RGB-IR 配准图像及有向目标标注。本文分别按照两个数据集的标准划分或通用评测协议进行训练与测试。

在实现上，本文选用 S2ANet（single-shot alignment network）（Han 等，2022）作为基础检测框架，并采用预训练的 VMamba（Liu 等，2024b）作为可见光与红外分支的双流骨干网络，用于多尺度特征提取。在此基础上，本文引入所提出的模态可靠性自适应融合模块，对双模态特征进行可靠性建模与动态融合。

训练阶段采用 AdamW 优化器，初始学习率设为 1×10^{-4}，权重衰减系数设为 0.05，batch size 设为 2，总训练轮数为 12。数据增强方面，仅采用随机翻转操作，翻转概率设为 0.5。为避免异常标注对训练过程造成干扰，在预处理阶段对无效标注框予以剔除，超参数 $\lambda_1 = 0.1$，$\lambda_2 = 0.5$。

评测方面，本文采用平均精度均值（mean average precision，mAP）作为主要评价指标，并同时报告 mAP@0.5 和 mAP@0.5:0.95。前者对应 IoU 阈值为 0.5 时的检测结果，后者则在 0.5 至 0.95 范围内以 0.05 为步长对多个 IoU 阈值下的平均精度（average

表 1　各方法在 DroneVehicle 数据集上的检测性能比较

Table 1　Comparison of detection performance of different methods on the DroneVehicle dataset

/%

方法	AP@0.5						mAP@0.5:0.95	mAP@0.5
	模态	Car	Truck	Freight Car	Bus	Van		
RetinaNet	RGB	78.5	34.4	24.1	69.8	28.8	25.0	47.1
S^2ANet	RGB	80.0	54.2	42.2	84.9	43.8	31.4	61.0
Faster R-CNN	RGB	79.0	49.0	37.2	77.0	37.0	28.5	55.9
RoITransformer	RGB	61.6	55.1	42.3	85.5	44.8	32.9	61.6
Oriented R-CNN	RGB	80.1	53.8	41.6	85.4	43.3	32.7	60.8
RetinaNet	IR	88.8	35.4	39.5	76.5	32.1	30.4	54.5
S^2ANet	IR	89.9	54.5	55.8	88.9	48.4	40.4	67.5
Faster R-CNN	IR	89.4	53.5	48.3	87.0	42.6	40.1	64.2
RoITransformer	IR	89.6	51.0	53.4	88.9	44.5	41.2	65.5
Oriented R-CNN	IR	89.8	57.4	53.1	89.3	45.4	41.5	67.0
TSFADet	RGB + IR	89.2	72.0	54.2	88.1	48.8	44.6	70.4
C^2Former	RGB + IR	90.2	68.3	64.4	89.8	58.5	47.5	74.2
DMM	RGB + IR	90.4	**79.8**	**68.2**	89.9	68.6	52.1	79.4
LPANet	RGB + IR	90.4	78.0	65.0	89.5	65.4	–	77.7
本文	RGB + IR	**90.6**	79.6	67.9	**90.4**	**70.1**	**53.7**	**79.7**

注：加粗字体表示各列最优结果。“–”表示该指标未报告。

precision, AP)进行平均,用于更全面地衡量模型的检测精度与定位能力。

2.2 对比实验分析

2.2.1 DroneVehicle数据集检测结果分析

为验证所提方法在航空遥感RGB-IR目标检测任务中的有效性,本文在DroneVehicle数据集上与多种代表性方法进行了比较,结果如表1所示。对比方法包括单模态检测方法和多模态检测方法两类。单模态方法包括RetinaNet(Lin等,2017)、S^2ANet(Han等,2022)、Faster R-CNN(faster region-based convolutional neural network)(Ren等,2015)、RoITransformer(Ding等,2019)和Oriented R-CNN(Xie等,2021)。多模态方法包括TSFADet(two-stream feature alignment detector)(Yuan等,2022)、C^2Former(calibrated and complementary transformer)(Yuan和Wei,2024)、DMM(disparity-guided multi-spectral Mamba)(Zhou等,2025)以及LPANet(large language model guided progressive feature alignment network)(Wu等,2026)。

从整体结果可以看出,单模态红外方法的检测性能普遍优于单模态可见光方法。这说明在DroneVehicle所包含的夜间、弱光及复杂背景场景中,红外模态通常能够提供更加稳定的目标响应,对检测任务具有更强的鲁棒性。与此同时,多模态方法整体优于单模态方法,进一步表明可见光与红外模态在纹理结构信息与热目标显著性方面具有明显互补性,联合建模能够有效提升检测性能。

在此基础上,本文方法在DroneVehicle数据集上取得了53.7%的mAP@0.5:0.95和79.7%的mAP@0.5,均为表1中的最佳结果。在Car、Bus和Van 3个类别上的AP分别达到最佳值。Truck和Freight Car类别为次佳,这说明所提方法在多模态目标检测中具有更强的竞争力。

本文方法与具有竞争力的C^2Former以及DMM的可视化结果如图6所示,在蓝色和黄色虚线圆圈标出的区域中,本文方法在夜间、弱光和复杂场景下能够检测出更多目标实例,并获得更准确的定位结果。具体来看,在部分区域中,现有方法存在漏检、误检或定位不准确等问题,而本文方法能够更完整地识别成排车辆、细长车辆以及局部遮挡目标。该现象与表1中的定量结果保持一致,进一步说明本文方法不仅能够有效利用红外与可见光模态之间的互补信息,而且能够围绕复杂成像条件下的模态质量属性进行显式建模,并进一步根据模态可靠性变化自适应调节不同模态对检测任务的贡献,从而提升整体检测精度。

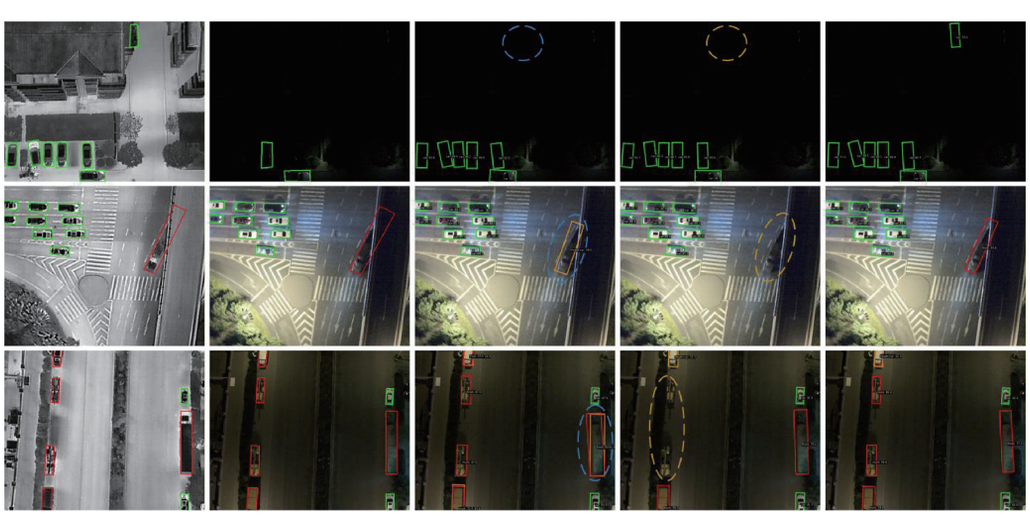


图6 检测结果可视化对比

Fig. 6 Visual comparison of detection results

2.2.2 VEDAI数据集检测结果分析

在VEDAI数据集上,本文进一步与若干代表性单模态方法及可比的多模态方法进行了对比实验,结果如表2所示。由于现有针对VEDAI的多模态融合检测研究大多基于水平边界框,而面向有向边界框设置的可比方法相对较少,因此本文主要选取若

干典型单模态检测器以及具有代表性的多模态方法进行比较。

从表2可以看出，本文方法在VEDAI数据集上取得了29.8%的mAP@0.5:0.95和67.1%的mAP@0.5，均优于现有多模态对比方法。其中，相比性能最优的对比方法DMM，分别提升了1.7%和1.4%。在类别指标上，本文方法在Car、Truck、Van和Boat类别上均取得了较优结果，说明所提方法对于VEDAI中尺度较小、外观差异细微的目标具有较好的识别能力。上述结果表明，所提方法不仅适用于复杂无人机场景，在以小目标为主的航空遥感场景中同样具有良好的有效性。这说明，相比固定的跨模态融合方式，围绕模态质量属性与模态可靠性进行显式建模，能够更有效地适应不同样本之间的成像差异，从而提升RGB-IR目标检测性能。

2.3 消融实验

为更清楚地分析所提方法中各组成部分的作用，本文在DroneVehicle数据集上开展消融实验，并通过逐步向基线模型中引入不同模块的方式，对各设计中影响整体检测性能的模块进行拆解分析。具

表2 各方法在VEDAI数据集上的检测性能比较

Table 2 Comparison of detection performance of different methods on the VEDAI dataset

/%

方法	模态	AP@0.5					mAP@0.5:0.95	mAP@0.5
		Car	Truck	Van	Boat	Plane		
RetinaNet	RGB	48.9	16.8	5.9	4.4	21.2	8.5	20.7
S^2ANet	RGB	74.5	47.3	32.5	16.7	7.1	18.9	44.5
Faster R-CNN	RGB	71.4	54.2	59.5	52.3	77.1	24.3	61.5
RoITransformer	RGB	77.3	56.1	60.2	56.7	**85.7**	27.6	65.4
Oriented R-CNN	RGB	77.6	59.7	60.9	60.1	84.0	29.4	66.4
RetinaNet	IR	44.2	15.3	7.2	4.0	33.4	8.0	18.7
S^2ANet	IR	73.0	39.2	32.3	13.9	12.0	17.2	40.0
Faster R-CNN	IR	71.6	49.1	57.0	35.6	71.6	21.6	55.4
RoITransformer	IR	76.1	51.7	**64.3**	46.9	83.3	24.6	60.5
Oriented R-CNN	IR	77.0	55.0	63.2	49.4	79.6	27.1	60.9
C^2Former	RGB+IR	76.7	52.0	48.0	43.3	47.0	22.9	55.6
DMM	RGB+IR	77.9	59.3	57.4	61.2	77.5	28.1	65.7
本文	RGB+IR	**80.1**	**63.4**	63.7	**61.9**	77.4	**29.8**	**67.1**

注：加粗字体表示各列最优结果。

表3 不同组成模块对模型性能影响的消融实验结果

Table 3 Ablation study results of the effects of different components on model performance

方法	属性监督	语义先验蒸馏	全局可靠性	局部空间可靠性	mAP@0.5/%
基线	–	–	–	–	75.6
+属性监督	√	–	–	–	76.3
+语义先验蒸馏	√	√	–	–	78.9
+全局可靠性	√	√	√	–	79.3
+局部空间可靠性	√	√	–	√	79.1
完整模型	√	√	√	√	**79.7**

注：加粗字体表示最优结果。“√”和“–”分别表示采用和未采用对应模块。

体而言，本文分别评估属性监督、语义先验蒸馏、全局可靠性建模和局部空间可靠性建模的贡献。表3给出了消融实验结果。可以看出，属性监督、语义先验蒸馏以及全局—局部模态可靠性建模均能够带来稳定的性能提升。其中，属性监督增强了模型对关键成像因素的感知能力，语义先验蒸馏进一步将视觉语言模型中的模态质量感知知识迁移到检测网络内部，而局部空间可靠性在全局可靠性基础上的进一步引入，则使模型能够从位置级角度更精细地调节不同模态的贡献。最终，完整模型取得了最佳检测性能。

2.4 属性维度有效性分析

为验证各模态质量属性的有效性，本文在DroneVehicle数据集上进行单属性移除实验。每次从完整属性体系中移除一个属性维度，并同步删除对应的文本描述、结构化标签及属性监督分支，其余设置保持不变。结果如图7所示，完整属性体系取得最高mAP@0.5，移除任一属性后性能均下降，说明各属性维度均对模态可靠性建模起到积极作用。其中，去除光照条件、边界清晰度和纹理清晰度带来的下降更明显，表明这些因素对复杂成像条件下的模态有效性判断影响较大。

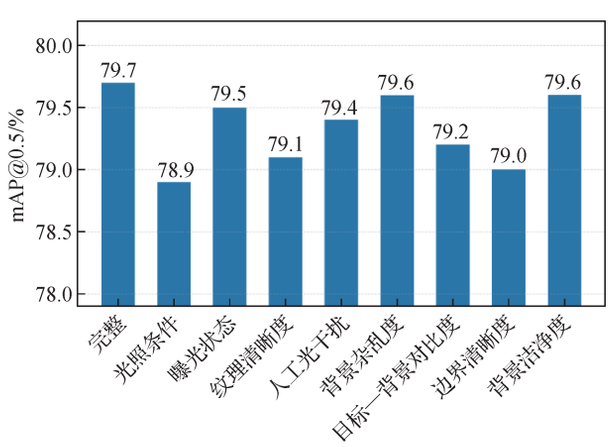


图7 不同模态质量属性维度对检测性能的影响

Fig. 7 Effect of different modality quality attribute dimensions on detection performance

2.5 超参数敏感性分析

为分析辅助损失权重对模型性能的影响，本文对λ_1和λ_2进行敏感性实验。首先固定$\lambda_2 = 0.5$调整λ_1，随后固定$\lambda_1 = 0.1$调整λ_2，结果如表4所示。可以看出，当$\lambda_1 = 0.1$，$\lambda_2 = 0.5$时模型取得最高mAP@0.5，当$\lambda_1 = 0$或$\lambda_2 = 0$时，性能分别下降至78.6%和78.3%，说明属性监督和语义先验蒸馏均有助于模态可靠性表征学习。随着λ_1或λ_2继续增大，性能也随之出现下降，表明过强的辅助约束会对检测主任务产生干扰。因此，本文最终设置$\lambda_1 = 0.1$，$\lambda_2 = 0.5$，以在检测性能和训练稳定性之间取得较好平衡。

表4 超参数敏感性分析结果

Table 4 Results of hyperparameter sensitivity analysis

分析对象	设置	mAP@0.5/%
λ_1	$\lambda_1 = 0, \lambda_2 = 0.5$	78.6
	$\lambda_1 = 0.1, \lambda_2 = 0.5$	79.7
	$\lambda_1 = 0.5, \lambda_2 = 0.5$	79.2
	$\lambda_1 = 1, \lambda_2 = 0.5$	78.9
λ_2	$\lambda_1 = 0.1, \lambda_2 = 0$	78.3
	$\lambda_1 = 0.1, \lambda_2 = 0.1$	78.5
	$\lambda_1 = 0.1, \lambda_2 = 0.5$	79.7
	$\lambda_1 = 0.1, \lambda_2 = 1$	78.8

2.6 计算复杂度分析

为验证所提方法的计算效率，本文以RGB-IR样本对作为输入，对可复现方法的计算复杂度和推理效率进行统计，并结合相关方法公开报告的浮点运算次数（floating point operations，FLOPs）进行对比，结果如表5所示。考虑到不同方法的实现细节、运行环境和代码可获得性存在差异，部分方法的参数量和推理时间难以在完全一致的条件下进行公平统计。因此，表5中对于未能统一统计的指标以“-”表示，并主要采用FLOPs作为计算复杂度对比指标。从表5可以看出，本文方法的FLOPs为135.8 G，低于语义引导方法LPANet的184.6 G，但高于DMM的102.4 G。相比DMM，本文方法引入了模态质量属性学习和全局—局部可靠性建模，因此计算量有所增加；但与LPANet等语义引导方法相比，本文方法

表5 不同方法的计算复杂度与推理效率比较

Table 5 Comparison of computational complexity and inference efficiency of different methods

方法	参数量/M	运行时间/s	FLOPs/G
DMM	89.97	0.068	102.4
LPANet	-	-	184.6
本文	99.30	0.073	135.8

注：“-”表示该指标未报告。

测试阶段无需文本编码器或大模型分支参与推理，因此计算复杂度更低。与此同时，在相同实验环境下，本文方法的推理时间为 0.073 s，与 DMM 的 0.068 s 接近，说明所提方法在引入模态可靠性建模的同时，仍保持了较低的额外推理开销。

3 结 论

针对航空遥感红外与可见光目标检测中不同模态贡献随成像条件动态变化、现有方法缺乏显式可靠性建模等问题，本文提出一种基于模态可靠性建模的航空遥感 RGB-IR 目标检测方法。通过构建面向无人机场景的模态质量属性描述数据，并结合视觉语言模型语义先验蒸馏，本文实现了对红外与可见光模态有效性的检测导向建模；在此基础上，进一步从场景级全局可靠性和位置级局部空间可靠性两个层面进行联合建模，实现了面向目标检测的动态自适应融合。实验结果表明，本文方法在 DroneVehicle 和 VEDAI 两个公开 RGB-IR 数据集上均取得了较好的检测性能，在夜间、弱光及复杂干扰场景下表现出较好的精度与鲁棒性。消融实验进一步说明，模态质量属性建模、语义先验蒸馏以及全局—局部联合建模均对性能提升具有积极作用。总体来看，围绕模态质量属性进行显式建模，能够提升多模态检测对复杂成像条件变化的适应能力，为航空遥感场景下高效、鲁棒的 RGB-IR 目标检测提供了一种可行思路。

参考文献(References)

Bai S, Cai Y X, Chen R Z, Chen K Q, Chen X H, Cheng Z S, et al. 2025. Qwen3-VL technical report [EB/OL]. [2026-04-13]. https://arxiv.org/pdf/2511.21631.pdf

Chen C, Bin K C, Hu T, Qi J H, Liu X Y, Liu T P, et al. 2025. Fusion meets diverse conditions: a high-diversity benchmark and baseline for UAV-based multimodal object detection with condition cues// Proceedings of 2025 IEEE/CVF International Conference on Computer Vision (ICCV). Honolulu, USA: IEEE: 27958-27967 [DOI: 10.1109/ICCV51701.2025.02595]

Chen C, Qi J H, Liu X Y, Bin K C, Fu R G, Hu X K, et al. 2024. Weakly misalignment-free adaptive feature alignment for UAVs-based multimodal object detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 26826-26835 [DOI: 10.1109/CVPR52733.2024.02534]

Chen K Y, Liu J Q and Zhang H. 2023. IGT: illumination-guided RGB-T object detection with transformers. Knowledge-Based Systems, 268: #110423 [DOI: 10.1016/j.knosys.2023.110423]

Ding J, Xue N, Long Y, Xia G S and Lu Q K. 2019. Learning RoI Transformer for oriented object detection in aerial images//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 2844-2853 [DOI: 10.1109/CVPR.2019.00296]

Han J M, Ding J, Li J and Xia G S. 2022. Align deep features for oriented object detection. IEEE Transactions on Geoscience and Remote Sensing, 60: #5602511 [DOI: 10.1109/TGRS. 2021.3062048]

Kim T, Chung S, Yeom D, Yu Y, Kim H G and Ro Y M. 2025. MSCoT-Det: language-driven multi-modal fusion for improved multispectral pedestrian detection. IEEE Transactions on Circuits and Systems for Video Technology, 35(5): 5006-5021 [DOI: 10.1109/TCSVT.2024.3524645]

Lin T Y, Goyal P, Girshick R, He K M and Dollár P. 2017. Focal loss for dense object detection//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 2999-3007 [DOI: 10.1109/ICCV.2017.324]

Liu X, Wu J M, Yang W F, Zhou X and Zhang T Z. 2024a. Multi-modal attribute prompting for vision-language models. IEEE Transactions on Circuits and Systems for Video Technology, 34(11): 11579-11591 [DOI: 10.1109/TCSVT.2024.3424566]

Liu Y, Tian Y J, Zhao Y Z, Yu H T, Xie L X, Wang Y W, et al. 2024b. VMamba: visual state space model//Proceedings of the 38th Conference on Neural Information Processing Systems. Vancouver, Canada: Neural Information Processing Systems Foundation, Inc. (NeurIPS): 103031-103063 [DOI: 10.52202/079017-3273]

Qian M H, Liu K, Zhang F B and Su B Y. 2026. Infrared small target detection with multi-branch perception and cross-layer semantic fusion. Journal of Image and Graphics, 31(8): 2942-2957 (钱孟豪, 刘奎, 章丰博, 苏本跃. 2026. 多分支感知与跨层语义融合的红外小目标检测. 中国图象图形学报, 31(8): 2942-2957 [DOI: 10.11834/jig.250448]

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

Razakarivony S and Jurie F. 2016. Vehicle detection in aerial imagery: a small target detection benchmark. Journal of Visual Communication and Image Representation, 34: 187-203 [DOI: 10.1016/j. jvcir. 2015.11.002]

Ren S Q, He K M, Girshick R and Sun J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks//Proceedings of the 29th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 91-99

Sun Y M, Cao B, Zhu P F and Hu Q H. 2022. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. IEEE Transactions on Circuits and Systems for Video Technology, 32(10): 6700-6713 [DOI: 10.1109/TCSVT.2022.3168279]

Wu W T, Li C L, Wang X and Luo B. 2026. Large language model guided progressive feature alignment for multimodal UAV object detection IEEE Transactions on Image Processing, 35: 6517-6529 [DOI:10.1109/TIP.2026.3702419]

Xiang X T, Zhou G Y, Wen Z X, Li W S, Niu B, Wang F, et al. 2026. SLGNet: synergizing structural priors and language-guided modulation for multimodal object detection [EB/OL]. [2026-04-13]. https://arxiv.org/pdf/2601.02249.pdf

Xie X X, Cheng G, Wang J B, Yao X W and Han J W. 2021. Oriented R-CNN for object detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 3500-3509 [DOI: 10.1109/ICCV48922.2021.00350]

Xiong Z X, Yao Z Y, Liu X, Zhao W Y, Cao J and Wu X K. 2025. Efficient multispectral object detection with attentive feature aggregation leveraging zero-shot implicit illumination guidance. Information Fusion, 118: #102939 [DOI: 10.1016/j.inffus.2025.102939]

Yu D, Lin S Z, Lu X F, Wang B, Li D W and Wang Y B. 2022. A multi-band image synchronous fusion method based on saliency. Infrared Physics and Technology, 127: #104466 [DOI: 10.1016/j.infrared.2022.104466]

Yu D, Tang Y P, Zhang C J, Wang W, Yang G D, Zheng X L, et al. 2026. IA^{3}GNN: imbalance-aware adaptive graph construction for multi-modal image fusion. IEEE Transactions on Multimedia, 28: 4747-4758 [DOI:10.1109/TMM.2026.3660161]

Yuan M X, Wang Y Y and Wei X X. 2022. Translation, scale and rotation: cross-modal alignment meets RGB-infrared vehicle detection// Proceedings of the 17th European Conference on Computer Vision - ECCV 2022. Tel Aviv, Israel: Springer: 509-525 [DOI: 10.1007/978-3-031-20077-9_30]

Yuan M X and Wei X X. 2024. C^{2}Former: calibrated and complementary transformer for RGB-infrared object detection. IEEE Transactions on Geoscience and Remote Sensing, 62: #5403712 [DOI: 10.1109/TGRS.2024.3376819]

Zhang R, Yao L, Zhang Y X, Wang Y J, Zhang C Y and Liu F. 2026. Non-spatial registration decision fusion for multimodal object detection. Journal of Image and Graphics, 31(2): 541-555 (张荣, 姚亮, 张奕欣, 王翌骏, 张传一, 刘凡. 2026. 非空间配准的多模态目标检测决策融合策略. 中国图象图形学报, 31(2): 541-555) [DOI: 10.11834/jig.250326]

Zhang Y M, Bao W T, Xiao Q, Yang Y, Wan W G, Luo Y T, et al. 2026. Selective attention-based for infrared small target detection. Journal of Image and Graphics, 31(3): 797-810 (张迎梅, 鲍王涛, 肖沁, 杨勇, 万伟国, 罗亦韬, 等. 2026. 基于选择性注意力的红外小目标检测. 中国图象图形学报, 31(3): 797-810) [DOI: 10.11834/jig.250313]

Zhou M H, Li T Y, Qiao C F, Xie D Y, Wang G Q, Ruan N J, et al. 2025. DMM: disparity-guided multispectral Mamba for oriented object detection in remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 63: #5404913 [DOI: 10.1109/TGRS.2025.3578309]

作者简介

余东,男,博士研究生,主要研究方向为图像融合和计算机视觉。E-mail:23115071@bjtu.edu.cn

张淳杰,通信作者,男,教授,主要研究方向为计算机视觉、图像分类。E-mail:cjzhang@bjtu.edu.cn

张晓宇,男,研究员,主要研究方向为多媒体分析与理解。E-mail:zhangxiaoyu@iie.ac.cn

郑晓龙,男,研究员,主要研究方向为多模态数据感知与理解、认知大模型与通用人工智能。E-mail:xiaolong.zheng@ia.ac.cn