

中图法分类号: TP18; TP391 文献标识码: A 文章编号: 1006-8961(2026)02-0541-15

论文引用格式: 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):0541-0555(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):0541-0555)[DOI:10.11834/jig.250326]

# 非空间配准的多模态目标检测决策融合策略

张荣, 姚亮, 张奕欣, 王翌骏, 张传一, 刘凡\*

河海大学计算机与软件学院, 南京 211100

**摘要:** 目的 多模态目标检测通过融合红外与可见光等多源传感器数据,有效提升了模型在复杂环境下的检测精度与鲁棒性。然而,现有方法普遍基于严格配准的多模态数据开展研究,不能直接适配实际应用中不同模态相机得到的非空间配准图像,在图像输入算法前仍需完成配准,这损失了实时性和灵活性。为此,提出一种非空间配准条件下的多模态目标检测任务,并设计了一种非空间配准的多模态目标检测决策融合方法。在数据层面,利用3种公共数据集模拟双光载荷拍摄得到的非空间配准图像对,在不引入额外标注成本的前提下,为新的任务提供了基准数据。在算法层面,设计了一种基于图结构的非空间配准决策融合方法。**方法** 首先,根据不同模态检测器的检测结果构建带权有向图,实现不同模态目标的图结构化表示;接着,利用图结构中目标间的相对位置关系,实现跨模态目标的自适应匹配;最终,对匹配成功的目标进行决策融合,并设计了模态迁移策略以实现多模态信息的高效互补。**结果** 在3个数据集上的实验结果表明,本文方法在非空间配准场景下较单模态检测器实现了最大10.03%的漏检率降幅。同时,该方法在配准数据集上同样适用,相较于多光谱行人检测Transformer(multi spectral pedestrian detection Transformer, MS-DETR)、动态自适应多光谱检测Transformer(dynamic adaptive multispectral detection Transformer, DAMSDet)等先进多模态目标检测方法,检测准确率提升了6.8%。**结论** 本文所提出的非空间配准多模态目标检测决策融合方法,能够很好地适应存在空间差异的实际场景,并且相比其他先进的多模态目标检测模型,有更高的准确率和鲁棒性。相关的代码与数据集将在此仓库公开:<https://github.com/1e12Leon/ProbDet>。

**关键词:** 目标检测;多模态目标检测;有向图;决策融合;非空间配准

## Non-spatial registration decision fusion for multimodal object detection

Zhang Rong, Yao Liang, Zhang Yixin, Wang Yijun, Zhang Chuanyi, Liu Fan\*

College of Computer Science and Software Engineering, Hohai University, Nanjing 211100, China

**Abstract: Objective** Multimodal object detection significantly improves detection accuracy and robustness in complex environments by integrating complementary information from multiple sensors such as visible light and infrared imaging, especially in harsh conditions such as low light, occlusion, smoke, or haze. However, a key limitation of current multimodal detection research is the high dependence on spatial registration data. Most existing methods assume that multimodal image pairs are aligned at the spatial level, which is usually true in manually processed datasets but is often difficult to guarantee in real-world scenarios. In practical deployment, such as autonomous driving, drone monitoring, and disaster rescue, multimodal sensors are typically installed on independent platforms with different perspectives, resolutions, and directions. Therefore, accurate registration under real-time conditions is difficult to achieve, frequently requiring expen-

收稿日期:2025-07-21;修回日期:2025-09-03;预印本日期:2025-09-10

\* 通信作者:刘凡 fanliu@hhu.edu.cn

基金项目:国家自然科学基金项目(62372155,62302149);江苏省基础研究计划项目(BK20250188)

Supported by: National Natural Science Foundation of China(62372155,62302149); Basic Research Program of Jiangsu Province, China (BK20250188)

sive hardware, manual calibration, or computationally expensive preprocessing. These limitations seriously affect the flexibility, scalability, and real-time performance of multimodal detection systems. To address this issue, this study defines a new research task: multimodal object detection under nonspatial registration conditions. This task explicitly eliminates the dependence on image spatial alignment, taking into consideration the conditions faced by multimodal perception systems in practical applications. The core goal is to realize accurate and robust target fusion of nonspatially registered multimodal image inputs without the need for explicit image spatial registration. **Method** To support this new task, this study proposes a decision-level fusion strategy for nonregistered multimodal object detection. This method can be directly operated on the basis of the output results of a single modal detector, without the need for image- or feature-level fusion, and does not rely on image registration. It mainly includes two contributions: the data and algorithm layers. At the data level, we have constructed a benchmark dataset that can simulate nonregistered inputs. Specifically, multi-spectral multi-scene and multi-resolution fusion dataset (M3FD), FLIR thermal dataset (FLIR), and low-light visible-infrared paired dataset (LLVIP), three widely used and originally registered multimodal datasets of infrared and visible light images, were selected. The differences in multimodal viewing angles and the inconsistent target numbers in reality were simulated by cropping image pairs. This process achieves data augmentation without introducing additional annotation information, providing a realistic and efficient testing platform for evaluating nonregistration detection methods without increasing annotation costs. At the algorithmic level, we propose a nonspatial registration decision fusion method based on graph structure. First, a weighted directed graph is constructed on the basis of the detection results of different modal detectors to obtain a structured representation of diverse modal targets. Each node in the diagram represents a detection target, and the edge weights between nodes are established, corresponding to the relative spatial positions and Euclidean distances between target nodes. Next, the target matching between different modalities is transformed into a graph structure matching problem for different modalities, utilizing the relative positional relationships between targets in the graph structure to achieve adaptive matching of cross-modal targets. Finally, for successfully matched target pairs, a Bayesian strategy based on probability ensemble and a weighted average method are used for confidence fusion and detection box fusion. At the same time, a modal transfer strategy was designed to attain efficient complementarity of multimodal information between two modalities for matching failed targets to compensate for missed detections and improve recall rates. **Result** We conducted extensive experiments on the constructed dataset and evaluated the performance of the method in nonregistered and registered scenarios. On the nonspatially registered M3FD, FLIR, and LLVIP datasets, our method achieved the lowest missed detection rates of 12.45%, 12.83%, and 2.67%, respectively. Compared with single-modal detectors, the proposed method reduced the maximum missed detection rate by 10.03%, fully demonstrating that even under nonspatial registration conditions, it can effectively utilize modal complementary information. On the M3FD, FLIR, and LLVIP datasets of spatial registration, this method also performed well, with  $AP_{50}$  reaching 87.0%, 84.4%, and 44.1%, respectively. In comparison with existing multimodal fusion methods such as MS-DETR and DAMSDet, it improved the detection accuracy by up to 6.8%. Experimental results show that the proposed method has good generalization ability on different datasets, can adapt to complex nonregistration situations, and exhibits strong adaptability and universality. **Conclusion** This article proposes a multimodal object detection decision fusion strategy suitable for nonspatial registration environments, which is in line with common real-world perception conditions. This method is based on graph matching and decision-level fusion and does not rely on image-level registration to achieve effective correlation and integration of multimodal information. Compared with existing methods, the method proposed in this paper significantly improves detection accuracy and recall in nonregistered scenarios and is applicable to registered data, providing a more universal solution for multimodal perception. The designed modal transfer mechanism further enhances the robustness of detection, allowing the complementary advantages between various modalities to be fully utilized. The code and dataset will be available at <https://github.com/1e12Leon/ProbDet>.

**Key words:** object detection; multimodal object detection; directed graph; decision fusion; non spatial registration

## 0 引言

多模态目标检测(multimodal object detection)(Brenner等,2023;El Sakka等,2025;Karim等,2023;李成龙等,2023)是指利用多种传感器或数据源(如可见光、红外、雷达、深度等图像)识别和定位目标的技术。与传统的单模态目标检测(如仅使用可见光图像)相比,多模态目标检测通过融合不同模态的信息,提升检测的准确性、鲁棒性和泛化能力,尤其是在复杂场景或单一模态数据不足的情况下。这种技术广泛利用于自动驾驶(Wei等,2022;Wang等,2024b;李昌财等,2024)、安全监控(Karim等,2023)、医疗影像(Tian等,2023;Ming等,2022;陈磊等,2025)和智能交通(李昌财等,2024;Xiang等,2024;杨丹和戴芳,2018)等领域。

以红外—可见光目标检测(别倩等,2023)为例,多模态目标检测的融合方式主要有3种策略:前期融合(Neves等,2023)、中期融合(Wan等,2024;Wang等,2024a)和后期融合(Chen等,2022;Lin等,2023)。前期融合一般在数据输入阶段就对不同模态的信息进行融合,然后将其作为一个整体输入同一目标检测网络进行后续操作;中期融合是在特征提取过程中进行模态融合,例如,在深度网络中分别提取可见光(red, green, blue, RGB)和红外(infrared, IR)特征,然后进行特征拼接或注意力(Di等,2024)加权融合;后期融合是一种决策级的融合,即利用不同检测器的预测结果融合信息。具体而言,首先对不同模态进行独立训练,然后在结果层采取加权平均、贝叶斯概率集成(Chen等,2022)策略融合结果。虽然上述策略均能取得可观的效果,但需要不同模态的图像严格配准(陈磊等,2025;Jia等,2021),即在空间、时间、内容和视角上严格对应。然而,在实际应用中,以DJI H20T相机(Li等,2022)为例,如图1(a)所示,即使两种模态的传感器同一时刻工作,也会由于传感器的位置差异导致图像无法完全对齐。因此,这些检测算法往往需要事先利用配准策略实现两种模态的图像对齐,这大大限制了算法的实用性。

考虑到这种不足,一些工作开始探究使用不同模态有轻微差异的算法。Ji等人(2025)提出特征互补感知网络,通过多级特征的像素级对齐增强可见

光—热成像(red green blue-thermal, RGB-T)模态间的语义关联性,实现高精度热—可见光目标检测;Zhang等人(2019)设计对齐区域卷积神经网络,其区域特征对齐模块通过学习偏移量自适应对齐弱配准多光谱区域特征,并引入动态特征重加权机制抑制无效模态特征,提升跨模态融合鲁棒性。然而,如图1(b)所示,当两种模态传感器的分辨率以及拍摄范围也存在差异时,采集到的图像会存在目标大小、数目的差异。这种情况下,上述工作仍然无法修正并实现精确检测。非配准的图像融合算法目前逐渐兴起,如多模态融合监督下的跨模态对齐感知算法(multimodal fusion-supervised cross-modality alignment perception, MulFS-CAP)通过可学习的模态字典补偿模态差异,并利用跨模态对齐感知机制生成相关性矩阵来实现特征对齐与融合。然而,此类融合算法主要关注不同模态图像间的空间轻微错位问题以及像素形变导致的像素级偏移,仍无法处理视野差异较大的情况。



(a) DJI H20T 相机

(b) 实拍示例

图1 可见光—红外模态实拍图

Fig. 1 Visible-infrared modal real shot

((a) DJI H20T camera; (b) real shooting example)

针对上述问题,本文重新审视了多模态目标检测任务。为满足实际应用场景,这一任务应该具备两种特征。1)数据同时性。不同模态的一对图像应在同一时刻采集,但无需在空间上完全对齐;2)模态互补性。不同模态的图像应能通过融合策略实现检测结果互补(互补不仅体现在目标位置,还应表现于目标置信度上)。基于此特性,本文提出一种新的多模态目标检测任务——非空间配准条件下的多模态目标检测,并以红外—可见光模态为例开展研究。具体研究工作如下:首先,基于现有的公共配准数据集,构造了一系列模拟非空间配准情况的多模态图

像以支撑该任务的训练与评估。接着,设计了一种基于带权有向图的匹配策略,通过构造不同模态图像中目标位置关系的有向图,实现两种模态目标的匹配。该策略无需输入数据空间对齐,根据不同模态目标的相对位置实现匹配。最后,提出一种非空间配准多模态决策融合方法。根据匹配结果,利用贝叶斯规则和加权平均融合匹配成功目标的置信度和检测框。获得更加精确的检测结果的同时,该策略还能在同一位置出现相同目标时大幅度提升置信度。

在 LLVIP (low-light visible-infrared paired dataset)、M3FD (multi-spectral multi-scene and multi-resolution fusion dataset) 和 FLIR (FLIR thermal dataset) 3 种数据集的非空间配准形式上的实验结果表明,本文方法能够有效实现非空间配准情况下的多模态目标检测,大幅度降低了目标漏检率。此外,本文方法在配准情况下的数据上依然适用,相比 DAMSDet 等多模态检测模型,精度有大幅度提升。

本文的主要贡献如下:1)提出一种新的多模态目标检测任务:非空间配准条件下的多模态目标检测。该任务无需事先对多模态图像配准,提升了多模态目标检测任务的应用价值。2)构造了一批非空间配准条件下的可见光—红外多模态目标检测数据集。数据集模拟了目标缺失、分辨率差异等非空间配准情况,可支撑该任务的后续研究。3)设计了一种基于图匹配的非空间配准多模态目标检测决策融合策略。该方法实现了决策级的目标自适应匹配与融合,在解决数据配准问题的同时,降低了漏检率,提升检测置信度。4)在3种公共数据集上的实验结果表明,本文方法较单模态检测器漏检率最大降低10.03%。本文方法也同样适配于配准数据集,相较于 DAMSDet 等先进方法提升了近7%精度。

## 1 相关工作

### 1.1 多模态目标检测数据集

多模态目标检测数据集主要包括3类:可见光—红外、可见光—深度图和可见光—雷达。

#### 1.1.1 可见光—红外数据集

这类数据集涵盖了可见光图像和热红外图像的信息,主要用于低照度环境中的行人检测(罗艳等,2022)、目标识别(郑胤等,2014)等任务。红外图像

可补偿可见光在夜间或复杂天气下感知能力下降的问题。为了实现有效融合,这类数据集通常采用严格的空配准和时间同步策略。如 LLVIP 数据集(Jia等,2021)专注于低光照行人检测,包含30 976对可见光和红外图像。该数据集通过几何和像素级配准,确保可见光与红外图像在空间上的严格对齐。M3FD数据集(Liu等,2022)是一个多类别的RGB-T目标检测数据集,包含约9 000对配准图像,来源于真实监控视频,具有较强的现实意义。所有图像经过严格配准,并提供高质量的边框和类别标注。RoadScene数据集(Xu等,2020)提供221对配准的RGB和热红外图像,主要采集于交通场景中。尽管该数据集未提供目标检测标注信息,但因其高质量的像素级配准图像,常用于图像对齐、模态融合与伪彩色合成等任务研究。

#### 1.1.2 可见光—深度图数据集

该类数据集融合了可见光图像与深度图信息,适用于三维场景理解、机器人感知、姿态估计和目标检测任务。如 VDT-2048 数据集(Song等,2023),一个包含2 048组RGB、深度图和红外图像的三模态数据集。该数据集主要面向多模态行人检测与识别任务,经过精确配准,适用于多模态融合建模。SUN RGB-D数据集(Song等,2015),包含超过10 000幅图像,是目前最广泛使用的室内RGB-D场景数据集之一,是研究RGB-D感知的重要基准数据集。TUM RGB-D数据集(Sturm等,2012),主要用于视觉里程计和同步定位与地图构建研究。

#### 1.1.3 可见光—雷达数据集

该类数据集主要面向自动驾驶等复杂感知任务,通过结合毫米波雷达和视觉图像,在雨、雾、夜间等极端天气条件下提高感知能力。雷达模态的优势在于抗干扰、抗遮挡以及可测速度与距离信息。如 STF (seeing through fog) 数据集(Bijelic等,2020),聚焦于恶劣天气下的多模态感知,数据采集于实际交通环境,包括雨、雾和夜晚等极端情况。模态之间实现了严格的空配准和时间对齐,常用于全天候目标检测、多模态融合建模等研究。nuScenes数据集(Caesar等,2020),一个大型多模态自动驾驶数据集,包含1 000个城市街景场景,总计约140万帧图像,是当前多模态3D目标检测和感知系统评估的标准基准之一。

## 1.2 多模态目标检测

多模态目标检测旨在利用各模态之间的互补特性,协同完成检测任务,主流方法主要分为前期融合和中期融合两种方式。前期融合主要在图像层面处理,例如,Zhang等人(2023)直接在通道层面拼接红外可见光图像,并采用辅助超分辨率学习,在兼顾检测精度与成本的同时,对多尺度目标进行高分辨率检测。中期特征融合的方法更加常见,例如Fang等人(2022)提出跨模态融合Transformer网络,在特征提取阶段学习长距离依赖关系并整合全局上下文信息。通过利用Transformer的自注意力机制,进行模态内与模态间的融合,显著提升多光谱物体检测的性能。Shen等人(2024)提出一种迭代交互机制,在Transformer模块之间共享参数,从而降低模型复杂度和计算成本,具有卓越的性能和更快的推理。Zhang等人(2021)提出在模态间和模态内注意力模块的指导下,进行动态加权和多光谱特征融合的学习,在较低的计算成本下显著提高检测精度。Fu等人(2024)提出多模态特征融合网络(long-range attention fusion network, LRAF-Net),首先对提取的可见光和红外双模态特征进行数据增强,再通过跨模态特征增强模块改善模态内特征表示。Zhang等人(2022)提出一种改进的去噪锚框检测Transformer(DETR with improved denoising anchor boxes, DINO),引入带噪声的标签训练去噪能力,并通过引入ground-truth anchor boxes和learnable queries结合的方式选择queries,提升性能,最后融合模态间的长距离依赖关系,提高检测性能。Cao等人(2023)提出一种基于通道切换和空间注意力机制的轻量化模型,在较少算力消耗下具备更优的效果。Xing等人(2024)提出MS-DETR(mixed supervision detection Transformer),设计了松耦合的融合策略,从多模态特征中独立稀疏采样一些关键点,并将其与自适应学习的注意力权重进行融合。随后,进一步提出一种基于实例感知的模态平衡优化策略,更好地对不同大小目标进行建模,可抵抗多模态图像间的错位影响。Guo等人(2025)提出DAMSDet(dynamic adaptive multispectral detection Transformer),包括模态竞争查询选择策略,以提供有用的先验信息;多光谱可变形交叉注意力模块,以适应性地采样和聚合红外和可见光图像的多语义级特征。此外,进一步采用了DETR的级联结构,以更好地挖掘互补信息。

然而,上述方法都需要各模态图像严格对齐,在实际应用中需要事先做配准,缺乏灵活性。因此,一些工作开始探究不同模态输入轻微偏差的情况,如Zhang等人(2025)提出基于移位窗口的级联对齐模块,以挖掘未对齐特征之间的空间长距离依赖关系,并级联对齐感应特征与参考特征。Wanchaitana-wong等人(2021)指出,现有研究多局限于轻微错位场景,并提出一种适用于大错位的多模态行人检测器。尽管这些方法在一定程度上缓解了多模态数据的轻微错位问题,但在应对拍摄范围、分辨率差距较大的多模态图像来说,仍难以实现精确检测。

## 2 方法

本文提出一种基于图匹配的非空间配准多模态目标检测决策融合方法,整体框架如图2所示,主要分为3个部分:非空间配准数据集构建、基于带权有向图的目标自适应匹配与基于贝叶斯公式和加权平均的决策融合。

### 2.1 问题定义

本文提出非空间配准条件下的多模态目标检测任务,定义为:在不依赖多模态图像空间对齐的前提下,直接对同一场景下同步采集但存在视角、分辨率或视野差异的多源传感器数据(如可见光与红外图像),通过自适应决策融合策略实现鲁棒目标检测的任务。

传统方法依赖特征图空间对齐,要求输入图像具备严格的几何一致性。弱配准方法虽然容忍轻微错位,但仍需通过可学习偏移量隐式实现特征级配准,无法解决视野不一致(如红外仅覆盖可见光局部区域)的极端场景。本任务需要一种空间对齐无关的融合范式,直接基于可见光与红外模态的目标检测结果构建跨模态关联,通过建模目标相对位置关系(而非绝对坐标),实现完全脱离图像级/特征级配准的自适应目标匹配。

给定未配准的红外—可见光图像对 $(I_{ir}, I_{RGB})$ ,分别输入预训练的单模态检测器,得到目标集合 $O_{ir} = \{o_i^{ir}\}_{i=1}^M$ 与 $O_{RGB} = \{o_j^{rgb}\}_{j=1}^N$ ,其中 $o = (b, s)$ 包含检测框坐标 $b \in \mathbf{R}^7$ 和置信度 $s \in [0, 1]$ 。本文目标是通过决策级融合生成鲁棒的目标集合 $O_{fused}$ ,为此需要解决以下问题:1)跨模态目标匹配:在未配准的条件

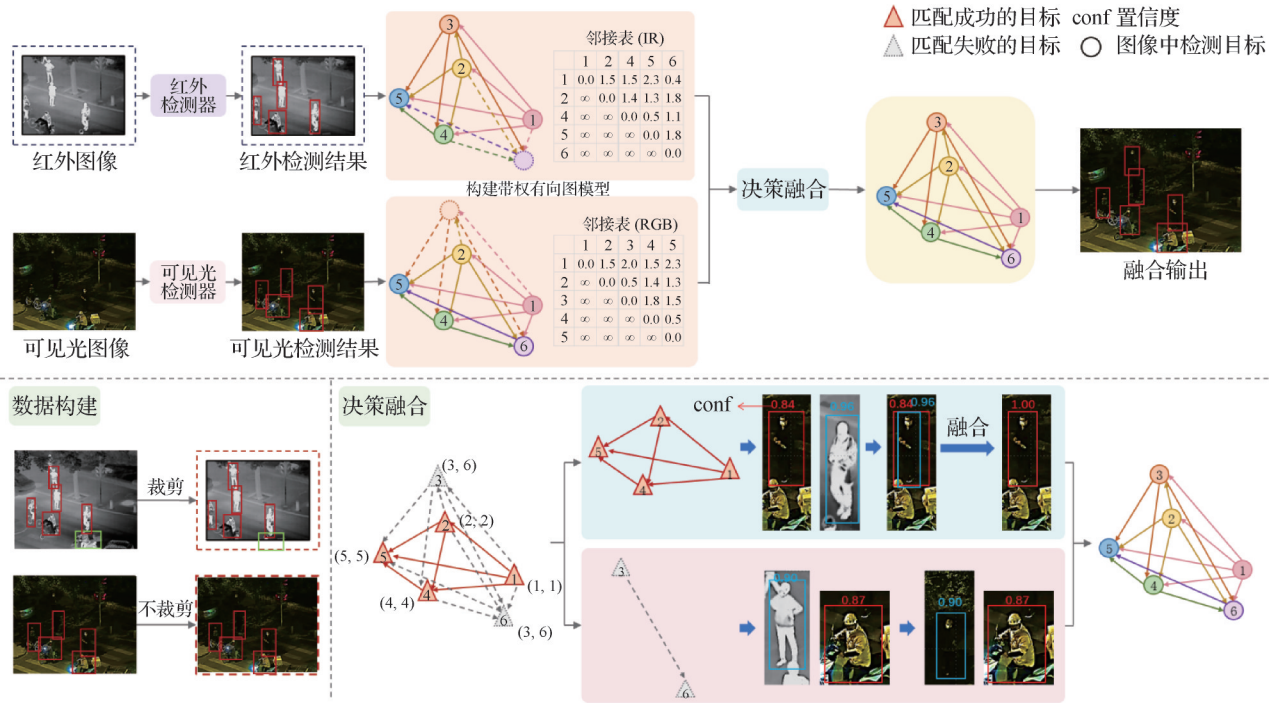


图2 非空间配准多模态目标检测决策融合总体框架图

Fig. 2 Overall framework diagram of decision fusion for non-spatial registration multimodal object detection

下,为 $O_{ir}$ 和 $O_{RGB}$ 构建有向图 $G_{ir}$ 与 $G_{RGB}$ ,通过图匹配算法求解结点间的对应关系矩阵 $P \in [0, 1]^{M \times N}$ ;2)自适应决策融合:对匹配结点对 $(o_i^{ir}, o_j^{RGB})$ ,基于贝叶斯公式融合置信度 $s_{fused} = B(s_i^{ir}, s_j^{RGB})$ ,并根据置信度加权融合检测框 $b_{fused} = F(s_i^{ir}, s_j^{RGB})$ ;3)冲突消解:对未匹配的结点,保留置信度高于阈值的结果。最终输出融合检测集合 $O_{fused} = \{(b_{fused}, s_{fused})\} \cup \{o_k | s_k > \tau\}$ 。

## 2.2 非空间配准数据集构建

本节为模拟真实场景中可见光与红外传感器在分辨率及视野范围的固有差异,对原始配准数据集的红外模态图像进行自适应比例裁剪,在不破坏中心目标分布的前提下,构建具有分辨率和目标位置及数量差异的双模态数据对。具体而言,针对LLVIP、M3FD、FLIR 3类数据集分别进行差异化比例裁剪:在红外图像四周对称裁减随机宽高像素(裁减量控制在原图尺寸15%~30%区间),使处理后红外图像分辨率与可见光模态形成显著差异但保持合理可比性(如可见光图像保持原始分辨率,红外图像调整为70%~85%尺度范围)。原始目标框标注需同步调整以适配裁剪后图像,并去除超出图像以外的标注。坐标修正公式为

$$(x'_{min}, x'_{max}) = (x_{min}, x_{max}) - \frac{\Delta W}{2} \quad (1)$$

$$(y'_{min}, y'_{max}) = (y_{min}, y_{max}) - \frac{\Delta H}{2} \quad (2)$$

式中, $(x_{min}, y_{min})$ 、 $(x_{max}, y_{max})$ 分别表示原始检测框左上与右下坐标; $(x'_{min}, y'_{min})$ 、 $(x'_{max}, y'_{max})$ 分别表示裁剪后检测框的左上与右下坐标。

## 2.3 基于带权有向图的目标自适应匹配

针对可见光—图像对中因传感器差异导致的目标空间错位问题,本研究提出一种基于带权有向图的跨模态匹配策略。

首先,基于目标中心到图像中心的欧氏距离分别对两模态目标进行空间排序。对于目标 $i$ ,其距离计算为

$$d_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \quad (3)$$

式中, $(x_c, y_c)$ 为图像中心坐标。将可见光与红外模态的目标分别按 $d_i$ 升序排列,得到有序集合 $V_R = \{v_1, v_2, \dots, v_m\}$ 和 $V_I = \{u_1, u_2, \dots, u_n\}$ 。为适配一对一匹配策略,截取前 $z = \min(m, n)$ 个目标构成候选集 $V'_R$ 与 $V'$ ,冗余目标划入匹配失败集合 $U_R$ 与 $U_I$ 。

进一步构建双模态带权有向图 $G_R = (V_R, q_R)$ 与 $G_I = (V_I, q_I)$ ,以建模目标间空间关系。图中结点表示目标,边集 $\varepsilon$ 由近邻目标指向远邻目标(若 $d_a < d_b$ 则存在有向边 $e_{ab}$ ),边权通过归一化欧氏距离量化,

具体为

$$q_{ab}^R = \sqrt{\left(\frac{x_a}{W_R} - \frac{x_b}{W_R}\right)^2 + \left(\frac{y_a}{H_R} - \frac{y_b}{H_R}\right)^2} \quad (4)$$

$$q_{ab}^I = \sqrt{\left(\frac{x_a}{W_I} - \frac{x_b}{W_I}\right)^2 + \left(\frac{y_a}{H_I} - \frac{y_b}{H_I}\right)^2} \quad (5)$$

式中,  $W_R, H_R$  和  $W_I, H_I$  分别为可见光和红外图像的宽高,  $(x_a, y_a)$  表示目标  $a$  的坐标,  $(x_b, y_b)$  表示目标  $b$  的坐标,  $q_{ab}^R$  表示在可见光图像中, 任意两目标结点  $a, b$  边上的权值,  $q_{ab}^I$  表示在红外图像中, 任意两目标结点  $a, b$  边上的权值。

匹配过程中, 遍历候选集结点对  $(v_i, u_j)$ , 若存在后继结点数量相等且边权比最相近的目标结点对, 则判定为跨模态匹配对。匹配失败的目标分别加入匹配失败集合  $U_R$  与  $U_I$ 。最终输出匹配成功集合  $M = \{(v_1, u_1), (v_2, u_2), \dots, (v_t, u_t)\} (t \leq z)$ , 及匹配失败集合  $U_R = \{v_{t+1}, \dots, v_m\}$ ,  $U_I = \{u_{t+1}, \dots, u_m\}$ 。

## 2.4 基于贝叶斯规则和加权平均的决策融合

### 2.4.1 基于贝叶斯规则的置信度融合

设  $N$  个独立检测器对同一目标输出类别  $c_k$  的置信度分别为  $P(c_k | s_1), P(c_k | s_2), \dots, P(c_k | s_N)$ , 其中  $s_i$  表示第  $i$  个检测器的观测数据。基于贝叶斯独立假设, 多模态联合概率可表示为:  $P(c_k | s_1, s_2, \dots, s_N) \propto P(c_k | s_1) \times P(c_k | s_2) \times \dots \times P(c_k | s_N)$ 。然而, 直接概率连乘易引发数值下溢 (累积概率值急剧衰减)。为提升计算稳定性并强化高置信度检测器的决策权重, 引入对数空间变换, 具体为

$$\ln P(c_k | s_1, s_2, \dots, s_N) = \sum_{i=1}^N \ln P(c_k | s_i) \quad (6)$$

接着, 通过指数运算与归一化映射回概率空间。具体为

$$P(c_k) = \frac{\exp\left(\sum_{i=1}^N \ln P(c_k | s_i)\right)}{\sum_{j=1}^K \exp\left(\sum_{i=1}^N \ln P(c_j | s_i)\right)} \quad (7)$$

式中,  $K$  为类别总数。本文针对可见光与红外双检测器 ( $N=2$ ) 场景, 利用式 (7) 实现跨模态置信度融合, 有效平衡数值稳定性与检测器可靠性。

### 2.4.2 基于加权平均的检测框融合与迁移

对于匹配成功目标集合  $M$  中的检测对  $(v_i, u_j)$ , 可以利用加权平均策略实现检测框融合。首先将红外检测框按分辨率比  $r_w = w_R/w_I$  与  $r_h = h_R/h_I$  ( $w_R, h_R$

和  $w_I, h_I$  分别表示可见光图像检测框的宽、高和检测框的宽、高) 映射至可见光模态空间, 并采用置信度加权融合, 具体为

$$w_i^{R'} = \frac{s_i^R}{s_i^R + s_i^I} \times w_i^R + \frac{s_i^I}{s_i^R + s_i^I} \times w_i^I \times r_w \quad (8)$$

$$h_i^{R'} = \frac{s_i^R}{s_i^R + s_i^I} \times h_i^R + \frac{s_i^I}{s_i^R + s_i^I} \times h_i^I \times r_h \quad (9)$$

式中,  $s_i^R$  和  $s_i^I$  分别为目标  $i$  在可见光与红外检测器的置信度分数,  $w_i^R$  和  $h_i^R$  分别表示目标  $i$  在可见光图像中检测框的宽和高,  $w_i^I$  和  $h_i^I$  分别表示目标  $i$  在红外图像中检测框的宽和高,  $w_i^{R'}$  和  $h_i^{R'}$  分别表示目标  $i$  加权平均后的融合检测框宽和高。若双模态预测类别不一致, 则计算检测框交并比  $IoU = B_R \cap B_I / B_R \cup B_I$  ( $B_R$  和  $B_I$  为可见光与红外模态下的检测框面积), 当  $IoU \geq 0.5$  时融合检测框并选取置信度较高类别; 否则分别输出双模态独立结果以降低误匹配风险。

对于匹配失败集合  $U_R$  与  $U_I$ , 通过坐标的变换实现跨模态的检测框迁移。以红外匹配失败目标为例, 其中心坐标  $(x_i, y_i)$  按原始裁剪量  $\Delta W, \Delta H$  平移至可见光空间, 具体为

$$(x_i', y_i') = \left(x_i + \frac{\Delta W}{2}, y_i + \frac{\Delta H}{2}\right) \quad (10)$$

式中,  $(x_i', y_i')$  表示第  $i$  个匹配失败的红外检测框按裁剪量迁移至可见光模态后的检测框中心坐标。

检测框尺寸同步按分辨率比缩放, 具体为

$$(w_i', h_i') = (w_i \times r_w, h_i \times r_h) \quad (11)$$

式中,  $w_i$  和  $h_i$  分别表示第  $i$  个匹配失败的红外检测框的宽和高,  $w_i'$  和  $h_i'$  分别表示第  $i$  个匹配失败的红外检测框按分辨率比缩放后的宽和高。

基于新中心坐标与尺寸重构检测框边界, 具体为

$$(x_{\min}, y_{\min}) = \left(x_i' - \frac{w_i'}{2}, y_i' - \frac{h_i'}{2}\right) \quad (12)$$

$$(x_{\max}, y_{\max}) = \left(x_i' + \frac{w_i'}{2}, y_i' + \frac{h_i'}{2}\right) \quad (13)$$

式中,  $(x_{\min}, y_{\min})$  表示第  $i$  个匹配失败的红外检测框迁移至可见光模态后的左边界与上边界坐标,  $(x_{\max}, y_{\max})$  表示第  $i$  个匹配失败的红外检测框迁移至可见光模态后的右边界与下边界坐标。

最终, 输出至可见光模态坐标系, 包含匹配成功目标的加权融合框与匹配失败目标的迁移框。

3 实验与结果分析

3.1 实验数据与设置

本文方法与检测器无关,仅对已有检测器输出的结果进行图结构匹配与融合,不依赖于特定检测架构。本节实验均以YOLOv7-L(you only look once version 7-Large)作为基础检测器进行,在训练阶段,本文采用LLVIP、M3FD、FLIR 3个数据集的原始红外和可见光样本,分别训练YOLOv7-L得到红外检测器和可见光检测器;在测试阶段,采用上述检测器,在3个公开数据集基础上构建的非空间配准测试集进行测试。具体数据配置如下:

LLVIP:面向低光行人检测的单类别(行人)数据集,包含12 025对训练图像与3 463对测试图像;M3FD:多类别交通场景数据集,划分为3 402对训练图像与354对测试图像;FLIR:热成像目标识别数据集,包含4 129对训练图像与1 013对测试图像。

实验基于PyTorch实现,使用Ubuntu18.04.3、Intel Xeon CPU E5-2678v3和单张NVIDIA GeForce RTX 2080Ti GPU。单模态检测器训练时,模型输入图像大小为640×640像素,批次大小为4,epoch为300。模型使用SGD(stochastic gradient descent)优化器,初始学习率和初始动量分别为0.01和0.937,权重衰减梯度为0.000 5。测试时,图像输入大小为640×640像素,批量大小为32。

3.2 评测指标

1)平均精度(average precision,AP)。基于PASCAL VOC(pattern analysis, statistical modeling and computational learning visual object classes)标准,以交并比(intersection over union, IoU)阈值0.5与0.75分别计算每个类别平均的AP值,记为AP<sub>50</sub>和AP<sub>75</sub>。

2)漏检率(false negatives per image, FNPI)。定义为图像漏检目标数与真实目标总数比值。具体为

$$FNPI = \frac{N_{FN}}{N_{GT}} \tag{14}$$

式中, N<sub>FN</sub>表示IoU阈值0.5下的漏检数量, N<sub>GT</sub>为标注目标总数。测试集整体性能以mFNPI量化,即每个类别上的平均漏检率,数值越低表明精度越高。

3.3 对比实验

本节在非空间配准数据上验证了本文提出的决策融合策略性能,并且与现有的多模态目标检测方法对比,探究了该策略的先进性。

3.3.1 非空间配准数据集的对比实验

本小节在本文构建的非空间配准数据集上进行了对比实验,以模拟双光载荷直接拍摄得到的图像直接输入检测器的情况,实验结果如表1所示。其中,图像输出类型表示利用本文方法对两模态图像进行融合后的输出图像类型。例如,输出图像类型为可见光,表示将决策融合后的结果迁移到可见光图像,相当于用红外信息增强可见光检测结果。实验结果表明:在单模态数据训练得到的基线检测器基础上,本文方法在LLVIP、M3FD与FLIR数据集上的平均漏检率分别实现了2.32%、10.03%与5.83%的最大降幅。这些结果得益于本文方法自适应匹配了不同检测器的输出结果,利用了不同模态的信息。

3.3.2 配准数据集的对比实验

本文方法在配准数据集上同样适用,与其他多模态目标检测方法以及单模态检测器对比结果如表2所示。

在M3FD数据集上,本文方法的AP<sub>50</sub>和AP<sub>75</sub>分别达到87.0%和56.8%,较单可见光模态检测器DINO提高13.7%和8.6%,较DAMSDet融合方法分别提高6.8%和0.8%;在FLIR数据集上AP<sub>50</sub>和AP<sub>75</sub>分别达到84.4%和44.1%,较单红外模态检测器

表1 本文方法有效性验证实验  
Table 1 Effectiveness verification experiment of the method

| 方法     | mFNPI/%     |              |              |             |              |              |
|--------|-------------|--------------|--------------|-------------|--------------|--------------|
|        | 图像输出类型:红外   |              |              | 图像输出类型:可见光  |              |              |
|        | LLVIP数据集    | M3FD数据集      | FLIR数据集      | LLVIP数据集    | M3FD数据集      | FLIR数据集      |
| YOLOv7 | 6.84        | 24.38        | 15.13        | 3.84        | 16.66        | 26.96        |
| 本文     | <b>4.52</b> | <b>14.35</b> | <b>12.83</b> | <b>2.67</b> | <b>12.45</b> | <b>21.13</b> |

注:加粗字体表示各列最优结果。

表 2 本文方法与现有先进方法在 LLVIP、M3FD 和 FLIR 数据集的对比  
Table 2 Comparison of this method with existing advanced methods on LLVIP, M3FD, and FLIR datasets

|      |                                    | 数据类型     | M3FD 数据集         |                  | FLIR 数据集         |                  | LLVIP 数据集        |                  |
|------|------------------------------------|----------|------------------|------------------|------------------|------------------|------------------|------------------|
| 融合阶段 | 模型                                 |          | AP <sub>50</sub> | AP <sub>75</sub> | AP <sub>50</sub> | AP <sub>75</sub> | AP <sub>50</sub> | AP <sub>75</sub> |
| 前期融合 | *Yolov5                            | IR       | —                | —                | 80.1             | —                | 96.5             | 76.4             |
|      | Yolov5                             | RGB      | —                | —                | 67.8             | 25.9             | 90.8             | 56.4             |
|      | DINO(Zhang 等, 2022)                | IR       | 58.8             | 36.1             | 80.6             | 42.7             | 96.6             | 70.6             |
|      | DINO(Zhang 等, 2022)                | RGB      | 73.3             | 48.2             | 70.9             | 25.9             | 91.6             | 58.0             |
|      | TarDAL <sub>CT</sub> (Liu 等, 2022) | IR + RGB | 68.3             | 45.4             | —                | —                | —                | —                |
|      | ICFNN(Zhang 等, 2020)               | IR + RGB | 69.0             | 40.8             | 74.9             | 35.1             | 95.5             | 72.4             |
|      | DenseFuse(Li 和 Wu, 2019)           | IR + RGB | 70.2             | 40.2             | 75.4             | 35.9             | 95.8             | 72.3             |
|      | VT(Heredia 等, 2025)                | IR + RGB | —                | —                | —                | —                | 95.8             | 64.7             |
| 中期融合 | CFT(Fang 等, 2022)                  | IR + RGB | 68.2             | 44.6             | 78.7             | 35.5             | 97.5             | 72.9             |
|      | ICAFusion(Shen 等, 2024)            | IR + RGB | 67.8             | 44.5             | 79.2             | 36.9             | —                | —                |
|      | GAFF(Zhang 等, 2021)                | IR + RGB | —                | —                | 72.9             | 32.9             | —                | —                |
|      | LRAF-Net(Fu 等, 2024)               | IR + RGB | —                | —                | 80.5             | —                | 97.9             | —                |
|      | CSAA(Cao 等, 2023)                  | IR + RGB | —                | —                | —                | —                | 94.3             | 66.6             |
|      | MS-DETR(Xing 等, 2024)              | IR + RGB | —                | —                | —                | —                | 97.9             | 76.3             |
|      | DAMSDet(Guo 等, 2025)               | IR + RGB | 80.2             | 56.0             | —                | —                | 97.9             | 79.1             |
| 决策融合 | 本文                                 | IR + RGB | <b>87.0</b>      | <b>56.8</b>      | <b>84.4</b>      | <b>44.1</b>      | <b>99.4</b>      | <b>92.4</b>      |

注:加粗字体表示各列最优结果。“—”表示原论文中未提供实验结果。\*指代开源网址:https://docs.ultralytics.com/zh/models/yolov5/

DINO 提高 3.8% 和 1.4%, 其中 AP<sub>50</sub> 较 LRAF-Net 提高 3.9%, AP<sub>75</sub> 较 ICAFusion 提高 7.2%; 在 LLVIP 数据集上, AP<sub>50</sub> 和 AP<sub>75</sub> 分别达到 99.4% 和 92.4%, 较单红外模态检测器 Yolov5 提高 2.9% 和 16%, 较 DAMSDet 提高 1.5% 和 13.3%。

3.4 消融实验

本节对决策融合的两策略、有向图匹配算法以及决策融合策略的有效性进行了消融实验, 结果分别如表 3—表 5 所示。

3.4.1 功能性验证实验

实验结果如表 3 所示, 单独采用检测框融合时, M3FD 与 FLIR 的 mFNPI 最低降至 12.69% 与 13.75%, 单独采用检测框迁移时, 对应降幅为 12.80% 与 14.75%。两个策略协同作用下, mFNPI 进一步降至 12.45%、12.83%, 这表明两种策略均对最终结果有积极贡献。其中, 检测框迁移主要针对跨模态匹配失败的目标, 通过坐标映射将检测结果

补充到另一模态中, 有效减少了漏检; 而检测框融合则利用双模态在同一位置的检测结果进行加权优化, 提升了定位精度和置信度。当两者协同时, 迁移策略首先弥补了漏检目标, 确保候选集合的完整性;

表 3 检测框融合与迁移策略功能性验证实验  
Table 3 Functional verification experiment of detection box fusion and migration strategy

| 检测框融合 | 检测框迁移 | mFNPI/%      |              |              |              |
|-------|-------|--------------|--------------|--------------|--------------|
|       |       | 图像输出类型: 可见光  |              | 图像输出类型: 红外   |              |
|       |       | M3FD         | FLIR         | M3FD         | FLIR         |
| —     | —     | 13.05        | 25.30        | 22.48        | 14.80        |
| √     | —     | 12.69        | 24.11        | 20.91        | 13.75        |
| —     | √     | 12.80        | 22.65        | 15.91        | 14.75        |
| √     | √     | <b>12.45</b> | <b>21.13</b> | <b>14.35</b> | <b>12.83</b> |

注:加粗字体表示各列最优结果。“√”表示实验验证的功能, “—”表示实验未验证的功能。

融合策略再对匹配成功的目标进行位置和置信度的优化,因此整体 mFNPI 降至最低。

3.4.2 不同图结构的有效性验证

实验结果如表4所示,通过对比分析有向图与无向图的性能差异,可以看出在 FLIR 和 LLVIP 数据集中,有向图的 mFNPI 最低达到 12.83% 和 2.67%,有向图在所有数据集上的平均漏检率均低于无向图,验证了方向约束在匹配过程中的有效性。这是因为有向图在构建边关系时,不仅利用了目标间的归一化距离,还引入了由近到远的空间指向性,从而在匹配时同时考虑距离与方位两种约束。该机制能够有效剔除那些虽然距离接近但方位差异较大的错误匹配对,提升匹配精度与融合结果的可靠性。相比之下,无向图缺乏方向信息,仅依赖距离进行匹配,在目标分布复杂或存在对称结构时更容易出现误匹配,导致检测性能下降。

表4 有向图有效性验证实验  
Table 4 Validity verification experiment of directed graph

| 方法  | mFNPI/%     |              |             |              |
|-----|-------------|--------------|-------------|--------------|
|     | 图像输出类型:可见光  |              | 图像输出类型:红外   |              |
|     | LLVIP       | FLIR         | LLVIP       | FLIR         |
| 无向图 | 3.48        | 25.49        | 6.35        | 17.55        |
| 有向图 | <b>2.67</b> | <b>21.13</b> | <b>4.52</b> | <b>12.83</b> |

注:加粗字体表示各列最优结果。

3.4.3 决策融合策略消融实验

为验证本文使用的决策融合策略的有效性,本节选取了其他置信度和检测框融合策略在 LLVIP 数据上进行消融实验,结果如表5所示。其中,置信度融合策略的最大值是指取两种模态在同一目标的最大置信度作为最终输出,平均值是指取二者的平均置信度;检测框融合策略中的最大值是指取两种模态在同一目标检测框的最大坐标值作为最终输出,平均值是指直接计算检测框坐标的平均值。

实验结果表明,本文使用的贝叶斯置信度融合策略与加权平均的检测框融合策略结合后模型取得了最佳的精度。这与概率统计学原理一致,当两个模态的检测器均在同一位置检测出同一目标时,模型应该更加确信此处存在该目标,并结合两种模态的检测框信息完成最后的位置优化。

表5 决策融合策略消融实验  
Table 5 Decision fusion strategy ablation experiment

| 置信度融合策略 | 检测框融合策略 | AP <sub>50</sub> /% |
|---------|---------|---------------------|
| 最大值     | 平均值     | 83.1                |
| 最大值     | 最大值     | 83.2                |
| 最大值     | 加权平均    | 83.3                |
| 平均值     | 平均值     | 82.7                |
| 平均值     | 最大值     | 82.9                |
| 平均值     | 加权平均    | 83.0                |
| 贝叶斯融合   | 平均值     | 83.2                |
| 贝叶斯融合   | 最大值     | 83.9                |
| 贝叶斯融合   | 加权平均    | <b>84.4</b>         |

注:加粗字体表示最优结果。

3.4.4 边权计算方式消融实验

为验证不同边权计算方式对非空间配准决策融合性能的影响,本节在 LLVIP 数据上进行消融实验,选取曼哈顿距离(L1距离)作为对比,如表6所示。

表6 边权计算方式消融实验  
Table 6 Ablation experiment on edge weight calculation

| 边权计算方式 | 图像输出类型 | mFNPI/%      |
|--------|--------|--------------|
| 曼哈顿距离  | 红外     | 15.13        |
|        | 可见光    | 21.42        |
| 欧氏距离   | 红外     | <b>12.83</b> |
|        | 可见光    | <b>21.13</b> |

注:加粗字体表示在图像输出类型相同时的最优结果。

从表6中可以看出,使用欧氏距离在两种模式上的 mFNPI 分别为 12.83% 与 21.13%,相比曼哈顿距离的 15.13% 和 21.42% 取得了更佳的漏检率,这表明直接计算两个目标之间的最短路径作为边权更有助于匹配目标,同时也验证了选取欧氏距离的合理性。其原因为非空间配准情况下,物理世界中目标之间的距离和相对位置不会改变,欧氏距离更加符合物理空间几何规律的距离度量,从而更适合辅助非空间配准下目标的匹配。

3.5 定性结果分析

3.5.1 决策融合策略可视化分析

本文在自主构建的非空间配准数据集上进行了决策融合策略的有效性验证,并将最终融合结果

与单模态检测结果对比,进行了可视化的效果展示。

如图3所示,与可见光图像相比,红外图像的拍摄范围更小,目标数量也相应减少。这种情况下,如图3(b)(d),决策融合策略能够融合可见光模态的漏检类别行人;又如图3(c),能够融合可见光漏检类别小汽车。特别在图3(a)烟雾遮挡严重的复杂

环境下,决策融合策略能够有效利用红外模态对可见光模态进行信息补充。此外,对于两模态共有的目标而言,经过决策融合,置信度有了大幅提升。以上结果均说明在非空间配准情况下,本文方法能够有效匹配并融合两模态中对应目标的检测框与置信度,同时综合各模态中的互补信息,较单检测器相比,有着更高的鲁棒性和准确性。

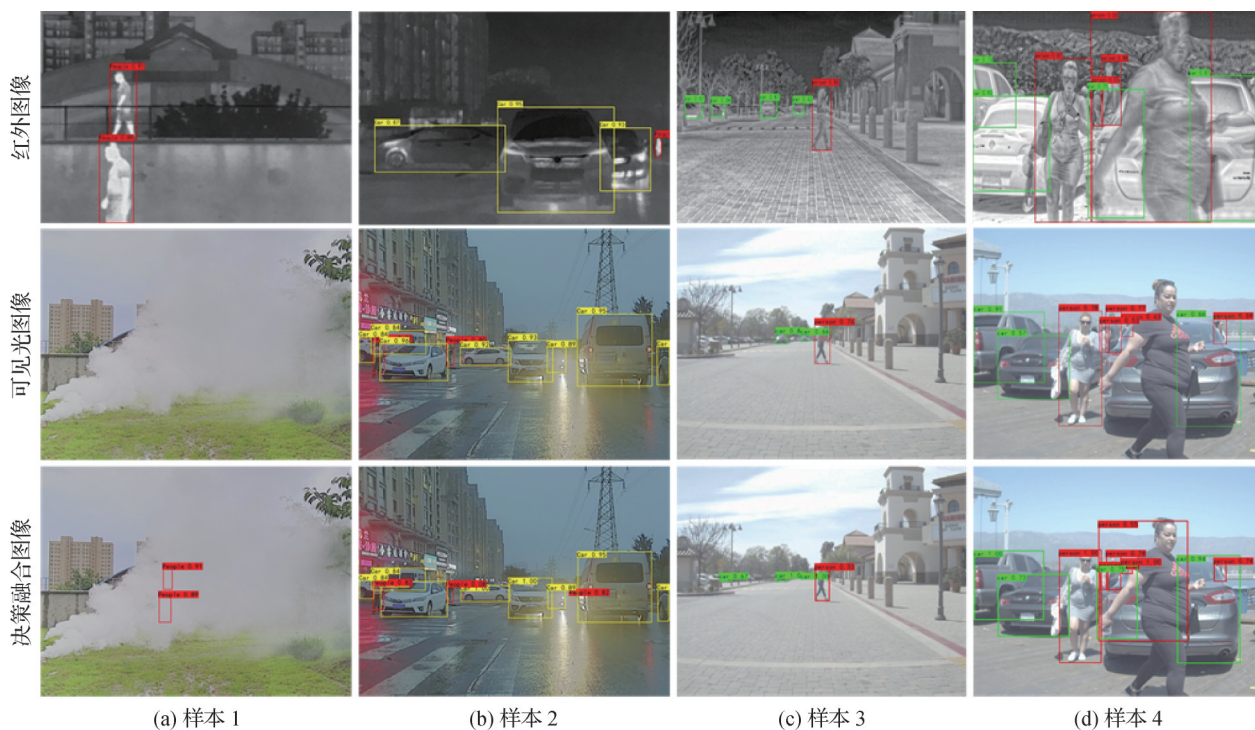


图3 本文方法在M3FD、FLIR数据集上的可视化结果

Fig. 3 Visualization results of the proposed method on M3FD and FLIR datasets  
((a) sample 1; (b) sample 2; (c) sample 3; (d) sample 4)

为进一步验证决策融合策略在极端非空间配准情况下的适用性,本文在M3FD数据集的基础上构造了视角范围、视角中心都有着明显差异的数据,并在此数据集上进行了实验,可视化结果如图4所示。在该组实验数据中,红外图像所覆盖的区域仅为可见光图像的右下部分。在图4(a)(b)展示的两组实验结果中,决策融合策略能够正确匹配可见光和红外模态的目标,图中放大区域显示,经过决策融合后的目标置信度更高,更加贴合实际情况。该实验结果表明,在极端的非空间配准情况下,本文方法依然能够正确匹配可见光与红外模态中的目标,且与单检测器相比,经过决策融合的目标置信度更高,充分体现了决策融合策略的普遍适用性。

### 3.5.2 真实拍摄的非空间配准数据上的鲁棒性分析

尽管本文方法在自主构建的非空间配准数据集上有着可观的正确性和适用性,但自主构建的非空间配准数据仅能模拟非配准环境,并不能起到完全代替的作用,因此,本文进一步在LLVIP-raw数据集(未经任何处理和标注的LLVIP数据集)上进行实验,可视化结果如图5所示。可以看出,决策融合图像中的目标完全包含了两模态中的目标,这得益于匹配策略的正确性。对比图中放大区域发现,相较于单检测器,融合后的目标置信度更高,体现了该方法在真实拍摄场景下的鲁棒性。

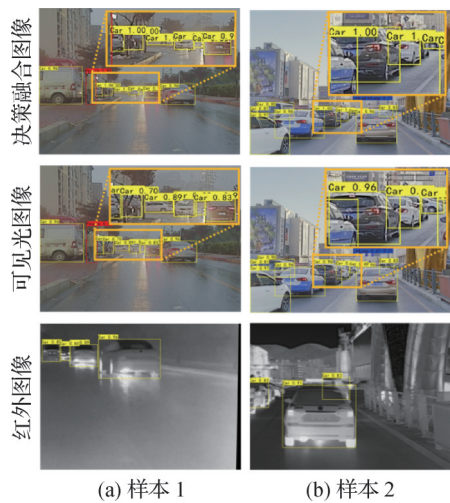


图4 本文方法在极端非空间配准情况下的可视化结果  
Fig. 4 Visualization results of proposed method in extreme non spatial registration scenarios ((a) sample 1; (b) sample 2)

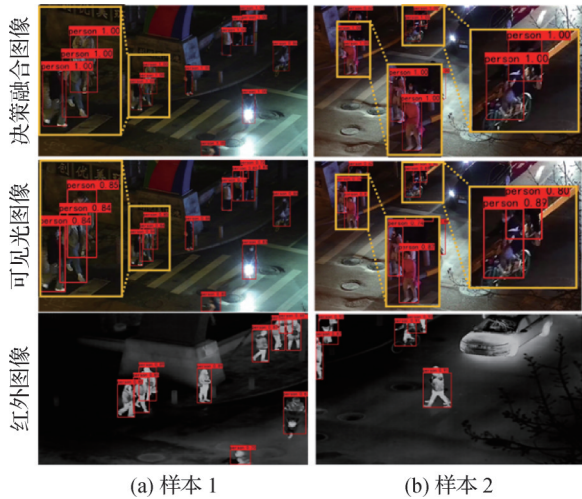


图5 本文方法在真实拍摄场景下的可视化结果  
Fig. 5 Visualization results of proposed method in real shooting scenes ((a) sample 1; (b) sample 2)

3. 5. 3 与先进融合方法的可视化对比

为进一步验证本文方法对其他多模态目标检测方法的性能领先,选取最新的 DAMSDet 模型作为对比,在配准数据集上的可视化比较如图 6 所示,在低照度强逆光环境下(图 6(a)),本文方法能够检测出 DAMSDet 漏检的自行车和小汽车类别,同时其他目标的置信度也更高。对于远距离目标(图 6(b)),DAMSDet 的检测精度大幅下降,但本文方法给出了高置信度,且数量更接近真实值的融合结果。在一般环境下(图 6(c)),当两种方法融合出的目标数量相同时,本文方法给出的置信度更接近真实值,这表明本文提出的非空间配准决策融合方法较 DAMS-

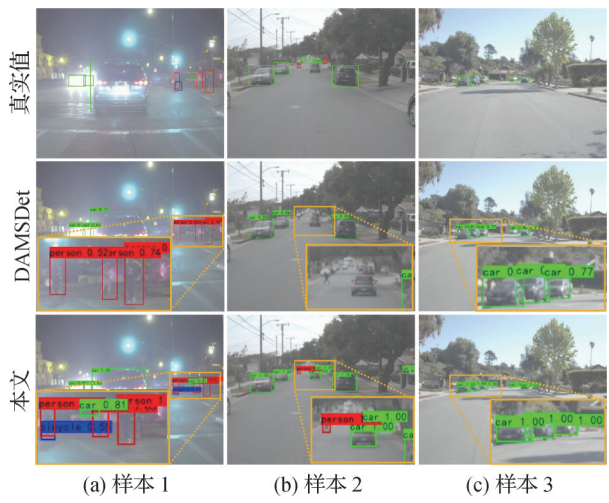


图6 本文方法与 DAMSDet 在 FLIR 数据集上可视化结果  
Fig. 6 Visualization results of the proposed method and DAMSDet on the FLIR dataset ((a) sample 1; (b) sample 2; (c) sample 3)

Det 有更高的准确性。

4 结 论

针对传统多模态目标检测依赖严格配准数据的问题,本文提出一种非空间配准条件下的多模态目标检测决策融合方法。首先,模拟了多模态图像分辨率差异等实际情况,构建了非空间配准场景的数据集;其次,设计了一种基于带权有向图的目标匹配策略,结合贝叶斯规则与加权平均的决策级融合方法,实现了未对齐模态间的目标自适应匹配与检测结果融合。实验表明,该方法在非空间配准数据上漏检率最大降幅 10.03%,在配准数据上也同样适用,AP<sub>50</sub> 相比现有模型最高提升 6.8%。

然而,本文方法只能适配同一时间拍摄且相近视角的非空间配准情况,对不同时段或极端视角差异场景仍需结合几何校正。注意到最新的 CVPR (Conference on Computer Vision and Pattern Recognition) 最佳论文提出一种前馈式 Transformer 网络 VGGT (Visual Geometry Grounded Transformer) 模型,该方法能够有效建模不同视角图像的关联点,未来将结合此技术,探索对复杂观测视角更加鲁棒的融合策略。同时,本文的图匹配思想也可用于非配准的图像融合网络设计中,例如,可采用超图结构更加准确地建模跨模态特征之间的关系,从而选择不同模态的共同区域进行融合。融合后的图像可传入检

测器,以促进检测精度。此外,本文仅探索了如何在不依赖空间对齐的前提下实现目标的匹配与置信度融合,在检测器的设计层面仍具备改进空间,未来可以针对模态具体特性,研发更加具有针对性和鲁棒性的检测器以进一步提升算法性能和实用价值。

## 参考文献(References)

- Bie Q, Wang X, Xu X, Zhao Q J, Wang Z, Chen J, et al. 2023. Visible-infrared cross-modal pedestrian detection: a summary. *Journal of Image and Graphics*, 28(5): 1287-1307 (别倩, 王晓, 徐新, 赵启军, 王正, 陈军, 等. 2023. 红外—可见光跨模态的行人检测综述. *中国图象图形学报*, 28(5): 1287-1307) [DOI: 10.11834/jig.220670]
- Bijelic M, Gruber T, Mannan F, Kraus F, Ritter W, Dietmayer K, et al. 2020. Seeing through fog without seeing fog: deep multimodal sensor fusion in unseen adverse weather//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, USA: IEEE: [DOI: 10.1109/CVPR42600.2020.01170]
- Brenner M, Reyes N H, Susnjak T and Barczak A L C. 2023. RGB-D and thermal sensor fusion: a systematic literature review. *IEEE Access*, 11: 82410-82442 [DOI: 10.1109/ACCESS.2023.3301119]
- Caesar H, Bankiti V, Lang A H, Vora S, Liong V E, Xu Q, et al. 2020. nuScenes: a multimodal dataset for autonomous driving//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, USA: IEEE: 11618-11628 [DOI: 10.1109/CVPR42600.2020.01164]
- Cao Y, Bin J C, Hamari J, Blasch E and Liu Z. 2023. Multimodal object detection by channel switching and spatial attention//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. Vancouver, Canada: IEEE: 403-411 [DOI: 10.1109/CVPRW59228.2023.00046]
- Chen L, Sun K C, Wu J J, Shi F and Shen D G. 2025. Full stack and full spectrum: exploration and application of artificial intelligence in medical imaging. *Journal of Image and Graphics*, 30(6): 2187-2206 (陈磊, 孙开聪, 吴交交, 石峰, 沈定刚. 2025. 全栈全谱: 医疗影像人工智能的探索与应用. *中国图象图形学报*, 30(6): 2187-2206) [DOI: 10.11834/jig.240449]
- Chen Y T, Shi J H, Ye Z L, Mertz C, Ramanan D and Kong S. 2022. Multimodal object detection via probabilistic ensembling//*Proceedings of the 17th European Conference on Computer Vision*. Tel Aviv, Israel: Springer: 139-158 [DOI: 10.1007/978-3-031-20077-9\_9]
- Di J, Guo W Q, Liu J Z, Ren L and Lian J. 2024. AMMNet: a multimodal medical image fusion method based on an attention mechanism and MobileNetV3. *Biomedical Signal Processing and Control*, 96: #106561 [DOI: 10.1016/j.bspc.2024.106561]
- El Sakka M, Ivanovici M, Chaari L and Mothe J. 2025. A review of CNN applications in smart agriculture using multimodal data. *Sensors*, 25(2): #472 [DOI: 10.3390/s25020472]
- Fang Q Y, Han D P and Wang Z K. 2022. Cross-modality fusion transformer for multispectral object detection [EB/OL]. [2025-07-21]. <https://doi.org/10.2139/ssrn.4227745>
- Fu H L, Wang S X, Duan P H, Xiao C Y, Dian R W, Li S T, et al. 2024. LRAF-Net: long-range attention fusion network for visible-infrared object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10): 13232-13245 [DOI: 10.1109/TNNLS.2023.3266452]
- Guo J J, Gao C Q, Liu F C, Meng D Y and Gao X B. 2025. DAMSDet: dynamic adaptive multispectral detection transformer with competitive query selection and adaptive feature fusion//*Proceedings of the 18th European Conference on Computer Vision*. Milan, Italy: Springer: 464-481 [DOI: 10.1007/978-3-031-73383-3\_27]
- Heredia Aguado E, Cabrera J J, Jiménez L M, Valiente D and Gil A. 2025. Static early fusion techniques for visible and thermal images to enhance convolutional neural network detection: a performance analysis. *Remote Sensing*, 17(6): #1060 [DOI: 10.3390/rs17061060]
- Ji Y Y, Song K C, Wen H W, Xue X T, Yan Y H and Meng Q G. 2025. UAV applications in intelligent traffic: RGBT image feature registration and complementary perception. *Advanced Engineering Informatics*, 63: #102953 [DOI: 10.1016/j.aei.2024.102953]
- Jia X Y, Zhu C, Li M Z, Tang W Q and Zhou W L. 2021. LLVIP: a visible-infrared paired dataset for low-light vision//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops*. Montreal, Canada: IEEE [DOI: 10.1109/ICCVW54120.2021.00389]
- Karim T, Mahayuddin Z R and Hasan M K. 2023. Singular and multimodal techniques of 3D object detection: constraints, advancements and research direction. *Applied Sciences*, 13(24): #13267 [DOI: 10.3390/app132413267]
- Li C C, Chen G, Hou Z X, Huang K and Zhang W. 2024. Survey of 3D object detection algorithms for autonomous driving. *Journal of Image and Graphics*, 29(11): 3238-3264 (李昌财, 陈刚, 侯作勋, 黄凯, 张伟. 2024. 自动驾驶中的三维目标检测算法研究综述. *中国图象图形学报*, 29(11): 3238-3264) [DOI: 10.11834/jig.230779]
- Li C L, Lu A D, Liu L and Tang J. 2023. Multi-modal visual tracking: a survey. *Journal of Image and Graphics*, 28(1): 37-56 (李成龙, 鹿安东, 刘磊, 汤进. 2023. 多模态视觉跟踪方法综述. *中国图象图形学报*, 28(1): 37-56) [DOI: 10.11834/jig.220578]
- Li H and Wu X J. 2019. DenseFuse: a fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5): 2614-2623 [DOI: 10.1109/TIP.2018.2887342]
- Li S, Qiao L H, Zhang Y M and Yan J. 2022. An early forest fire detection system based on DJI M300 drone and H20T camera//*Proceedings of 2022 International Conference on Unmanned Aircraft Systems (ICUAS)*. Dubrovnik, Croatia: IEEE [DOI: 10.1109/ICUAS4217.

- 2022.9836119]
- Lin Z W, Shen Y Q, Zhou S P, Chen S T and Zheng N N. 2023. MLF-DET: multi-level fusion for cross-modal 3D object detection//Proceedings of the 32nd International Conference on Artificial Neural Networks on Artificial Neural Networks and Machine Learning. Heraklion, Greece: Springer: 136-149 [DOI: 10.1007/978-3-031-44195-0\_12]
- Liu J Y, Fan X, Huang Z B, Wu G Y, Liu R S, Zhong W, et al. 2022. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: [DOI: 10.1109/CVPR52688.2022.00571]
- Luo Y, Zhang C Y, Tian Y H, Guo J and Sun J. 2022. An overview of deep learning based pedestrian detection algorithms. *Journal of Image and Graphics*, 27(7): 2094-2111 (罗艳, 张重阳, 田永鸿, 郭捷, 孙军. 2022. 深度学习行人检测方法综述. *中国图象图形学报*, 27(7): 2094-2111) [DOI: 10.11834/jig.200831]
- Ming Y, Dong X Y, Zhao J H, Chen Z F, Wang H and Wu N. 2022. Deep learning-based multimodal image analysis for cervical cancer detection. *Methods*, 205: 46-52 [DOI: 10.1016/j.ymeth.2022.05.004]
- Neves F S, Claro R M and Pinto A M. 2023. End-to-end detection of a landing platform for offshore UAVs based on a multimodal early fusion approach. *Sensors*, 23(5): #2434 [DOI: 10.3390/s23052434]
- Shen J F, Chen Y F, Liu Y, Zuo X, Fan H and Yang W K. 2024. ICA-Fusion: iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145: #109913 [DOI: 10.1016/j.patcog.2023.109913]
- Song K C, Wang J, Bao Y Q, Huang L M and Yan Y H. 2023. A novel visible-depth-thermal image dataset of salient object detection for robotic visual perception. *IEEE/ASME Transactions on Mechatronics*, 28(3): 1558-1569 [DOI: 10.1109/TMECH.2022.3215909]
- Song S R, Lichtenberg S P and Xiao J X. 2015. SUN RGB-D: a RGB-D scene understanding benchmark suite//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 567-576 [DOI: 10.1109/CVPR.2015.7298655]
- Sturm J, Engelhard N, Endres F, Burgard W and Cremers D. 2012. A benchmark for the evaluation of RGB-D SLAM systems//Proceedings of 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vilamoura-Algarve, Portugal: IEEE: 573-580 [DOI: 10.1109/IROS.2012.6385773]
- Tian Y, Xu Z C, Ma Y J, Ding W P, Wang R L, Gao Z H, et al. 2023. Survey on deep learning in multimodal medical imaging for cancer detection. *Neural Computing and Applications*, 37(26): 22239-22254 [DOI: 10.1007/s00521-023-09214-4]
- Wan B, Zhou X F, Sun Y Q, Wang T Y, Lyu C T, Wang S, et al. 2024. MFFNet: multi-modal feature fusion network for V-D-T salient object detection. *IEEE Transactions on Multimedia*, 26: 2069-2081 [DOI: 10.1109/TMM.2023.3291823]
- Wanchaitanawong N, Tanaka M, Shibata T and Okutomi M. 2021. Multi-modal pedestrian detection with large misalignment based on modal-wise regression and multi-modal IoU//Proceedings of the 17th International Conference on Machine Vision and Applications (MVA). Aichi, Japan: IEEE: 1-6 [DOI: 10.23919/MVA51890.2021.9511366]
- Wang J P, Su N, Zhao C H, Yan Y M and Feng S. 2024a. Multi-modal object detection method based on dual-branch asymmetric attention backbone and feature fusion pyramid network. *Remote Sensing*, 16(20): #3904 [DOI: 10.3390/rs16203904]
- Wang Y J, Abd Rahman A H, Nor Rashid F A and Md Razali M K. 2024b. Tackling heterogeneous Light Detection and Ranging-camera alignment challenges in dynamic environments: a review for object detection. *Sensors*, 24(23): #7855 [DOI: 10.3390/s24237855]
- Wei Z Q, Zhang F K, Chang S, Liu Y Y, Wu H C and Feng Z Y. 2022. MmWave radar and vision fusion for object detection in autonomous driving: a review. *Sensors*, 22(7): #2542 [DOI: 10.3390/s22072542]
- Xiang T, Zhao H X, Zhu F H, Chen Y Y and Lyu Y S. 2024. Self-aware adaptive alignment: enabling accurate perception for intelligent transportation systems//Proceedings of 2024 IEEE Intelligent Vehicles Symposium (IV). Jeju, Korea (South): IEEE: 595-601 [DOI: 10.1109/IV55156.2024.10588464]
- Xing Y H, Yang S, Wang S, Zhang S Z, Liang G Q, Zhang X W, et al. 2024. MS-DETR: multispectral pedestrian detection transformer with loosely coupled fusion and modality-balanced optimization. *IEEE Transactions on Intelligent Transportation Systems*, 25(12): 20628-20642 [DOI: 10.1109/TITS.2024.3450584]
- 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 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press: 12484-12491 [DOI: 10.1609/aaai.v34i07.6936]
- Yang D and Dai F. 2018. Improved ViBe algorithm for detection of moving objects. *Journal of Image and Graphics*, 23(12): 1813-1828 (杨丹, 戴芳. 2018. 运动目标检测的 ViBe 算法改进. *中国图象图形学报*, 23(12): 1813-1828) [DOI: 10.11834/jig.180304]
- Zhang H, Fromont E, Lefevre S and Avignon B. 2021. Guided attentive feature fusion for multispectral pedestrian detection//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 72-80 [DOI: 10.1109/WACV48630.2021.00012]
- Zhang H, Li F, Liu S L, Zhang L, Su H, Zhu J, et al. 2022. DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection [EB/OL]. [2025-07-21]. <https://arxiv.org/pdf/2203.03605.pdf>
- Zhang J Q, Lei J, Xie W Y, Fang Z M, Li Y S and Du Q. 2023. SuperYOLO: super resolution assisted object detection in multi-modal remote sensing imagery. *IEEE Transactions on Geoscience*

- and Remote Sensing, 61: #5605415 [DOI: 10.1109/TGRS.2023.3258666]
- Zhang L, Zhu X Y, Chen X Y, Yang X, Lei Z and Liu Z Y. 2019. Weakly aligned cross-modal learning for multispectral pedestrian detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5126-5136 [DOI: 10.1109/ICCV.2019.00523]
- Zhang Y, Liu Y, Sun P, Yan H, Zhao X L and Zhang L. 2020. IFCNN: a general image fusion framework based on convolutional neural network. Information Fusion, 54: 99-118 [DOI: 10.1016/j.inffus.2019.07.011]
- Zhang Y, Yang W, Xu C, Hu Q, Xu F and Xia G S. 2025. Mitigating the impact of prominent position shift in drone-based RGBT object detection [EB/OL]. [2025-07-13].  
<https://arxiv.org/pdf/2502.09311.pdf>
- Zheng Y, Chen Q Q and Zhang Y J. 2014. Deep learning and its new progress in object and behavior recognition. Journal of Image and Graphics, 19(2): 175-184 (郑胤, 陈权崎, 章毓晋. 2014. 深度

学习及其在目标和行为识别中的新进展. 中国图象图形学报, 19(2): 175-184) [DOI: 10.11834/jig.20140202]

## 作者简介

张荣, 女, 本科生, 主要研究方向为多模态目标检测。

E-mail: 2206060206@hhu.edu.cn

刘凡, 通信作者, 男, 教授, 主要研究方向为计算机视觉、多媒体分析与理解。E-mail: fanliu@hhu.edu.cn

姚亮, 男, 博士研究生, 主要研究方向为多模态环境感知、视觉基础模型。E-mail: yaoliang@hhu.edu.cn

张奕欣, 女, 本科生, 主要研究方向为多模态目标检测与识别。E-mail: 2206010406@hhu.edu.cn

王翌骏, 男, 硕士研究生, 主要研究方向为指代图像分割。

E-mail: wangyijun@hhu.edu.cn

张传一, 男, 副教授, 主要研究方向为图像处理与计算机视觉、多模态学习。E-mail: zhangchuanyi@hhu.edu.cn