面向城市场景的视觉多模态目标检测

一、赛题背景

随着智慧城市及智能安防技术的飞速发展，复杂环境下的全天候智能视觉感知需求日益迫切。在真实的城市场景中，夜间低照度、强逆光、烟雾弥漫、物体遮挡以及低纹理背景等极端条件频繁出现。在这些场景下，传统基于单一可见光（RGB）图像的目标检测算法性能会显著下降，难以满足高可靠性应用的需求，成为制约相关技术落地的关键瓶颈。然而，多传感器融合为解决上述痛点提供了新路径：红外（Infrared）信息则利用目标与背景的温差特性，在弱光甚至无光环境下突出关键目标区域；深度（Depth）信息能够提供稳定的几何结构约束，有效应对遮挡与尺度变化。通过融合可见光、热红外与深度三种互补模态，可显著提升关键目标检测的鲁棒性与可靠性。

基于以上背景与需求，本赛题旨在推动复杂城市场景下准确鲁棒的多模态目标检测算法发展，鼓励参赛者通过可见光、热红外与深度多源异构数据融合和检测理论方法的创新改进，有效突破单一模态感知在夜间、逆光及遮挡等极端环境下的局限性，提升算法的泛化性能。本赛题的研究成果不仅有助于促进多模态数据融合理论的深化，更可直接应用于全天候交通监控、恶劣天气自动驾驶导航、城市安防巡检及应急救援探测等关键领域，对改善社会公共安全、提升城市智能化管理水平具有重要的学术价值与现实应用意义。本赛题将提供包含可见光 RGB 图像、红外 Infrared 图像和深度 Depth 图像和的目标检测数据集。参赛队伍需要设计基于三模态数据输入的视觉目标检测算法，充分利用多源异构数据互补信息，实现复杂场景下精确的目标检测。

二、赛题应用场景

本赛题聚焦于复杂城市场景下的多源异构智能视觉感知。在城市各种场景中，低照度、目标遮挡及低纹理背景等极端条件频繁出现，导致传统单一可见光模态难以稳定工作：天空为背景的无人机检测、交通监控中人车目标纹理模糊；安防巡检面临复杂背景遮挡与远距离小目标检测难题；应急救援则在朦胧低光环境下完全丧失可见光感知优势。本赛题通过引入深度与热红外模态，利用深度信息的几何结构约束解决遮挡与尺度估计问题，借助热红外信息的温差特性在弱光或无光条件下突出关键目标，实现多源数据的优势互补。该方面研究可以提升系统在极端环境下的鲁棒性与准确性，其成果可服务于智慧交通、自动驾驶、公共安全及应急管理等关键领域，对推动城市智能化治理与社会安全保障具有重要的应用价值。

三、赛题任务

本赛题旨在解决复杂城市场景下单一模态感知能力受限的问题，要求参赛者设计并实现一种基于可见光、热红外与深度三模态数据融合的目标检测算法。

（一）任务描述

参赛队伍需利用赛题提供的空间对齐三模态图像数据，通过构建深度学习模型，有效提取并融合多源异构特征，实现对路口、公园、室内及空中等多场景下关键目标的精准定位与分类。如图 1 所示。

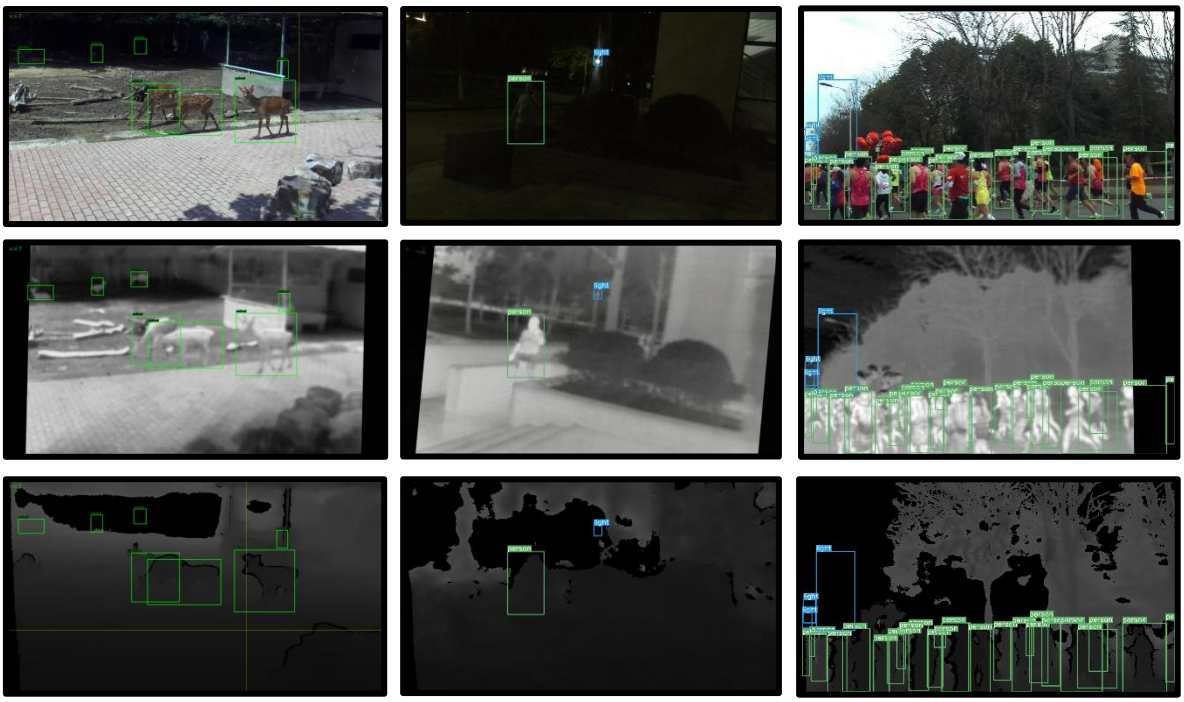

图 1 部分样例展示（第一行为可见光图像、第二行为红外图像、第三行为深度图像）

数据集中总共包含从“0”开始编号的 12 个类别：行人、船、动物、座椅、标识（包括路牌、标语、标志）、双轮车（包含自行车、双轮电动车）、四轮汽车、球、灯（包含路灯和室内的照明灯）、垃圾箱、无人机、三轮车），分别编号为：0: person, 1: boat, 2: animal, 3: seat, 4: sign, 5: bicycle, 6: car, 7: ball, 8: light, 9: garbage can, 10: uav, 11: tricycle。

（二）任务输入输出说明

数据集标签说明：目标在图像中的标签为对应目标的矩形边界框，格式为 [class_id, norm_center_x, norm_center_y, norm_w, norm_h]。其中 class_id 是类别编号，表示该目标属于哪个类别；norm_center_x 表示目标框中心点在图像宽度方向的位置，经过归一化处理后的值（取值范围为 0~1）；norm_center_y 表示目标框中心点在图像高度方向的位置，经过归一化处理的值（取值范围为 0~1）；norm_w 是目标框的宽度，占整张图像宽度的比例（归一化到 0~1）；norm_h 是目标框的高度，占整张图像高度的比例（归一化到 0~1）。

输入数据：处理赛题提供的空间对齐的三模态图像数据，包括可见光图像 (RGB)、红外图像 (Infrared)和深度图 (Depth)。

核心目标：对图像中的关键物体进行精准定位与分类。

输出要求：算法需输出目标的边界框（Bounding Box）坐标、置信度及对应的类别标签（Class Label），格式为[class_id, norm_center_x, norm_center_y, norm_w, norm_h, confidence]，其中 confidence 范围为[0,1]。

四、数据集及数据说明

（一）数据来源

本赛题提供的数据集为自行采集的数据，其中可见光 RGB 图像序列和 Depth 深度图像序列是使用 ZED 双目相机采集；采用 InfiRay LG6122 热成像相机获取红外图像。赛题所提供数据集在采集过程中涉及的受试者获得了同意。

（二）数据规模

本赛题提供的数据集包含训练集、初赛测试集、复赛测试集和半决赛测试集，每张图像都有包含多个目标的空间对齐的三模态图像（RGB、Depth 和 Infrared）。其中，训练集共包含带目标标注框的 2,000 组多模态图像，赛事初赛、复赛、半决赛阶段的测试集分别包含 1,000 组不同的测试数据。

示例数据可以通过以下链接获取，正式数据将在报名后开放下载。

https://pan.baidu.com/s/1FBMH8t-boH2j4YiRjrk9NQ，提取码: sxwu

（三）数据格式

RGB、Infrared 和 Depth 图像均是 PNG 格式图像。其中，可见光 RGB 图像为三通道 8 位无符号整型数据，像素取值范围为[0, 255]。红外 Infrared 图像存储为三通道 8 位无符号整型数据，各通道视觉特征高度一致，无实际彩色语义，本质为 3 个单通道热辐射灰度图像堆叠，像素取值范围为[0, 255]，像素数值越高表示目标温度越高，画面视觉效果越亮。深度图像为单通道 16 位无符号整型数据，理论取值范围为[0, 65535]，单位为毫米，深度相机的可感知距离范围约为 30 厘米到 20 米（即取值范围[0, 19999]）。深度图像素值为 0 或者过小代表无效深度区域，数值越大表示目标距离相机越远，数值越小表示目标距离相机越近。所有图像均未经过归一化处理。

每组图像中的所有物体的标签在对应的一个 TXT 文件中，文件每一行是目标在一帧图像中的类别、位置和尺寸[class_id, norm_center_x, norm_center_y, norm_w]。

五、算法设计要求

（一）算法类型

本赛题对算法类型不做严格限制，鼓励参赛者采用深度学习的视觉目标检测识别算法，如基于卷积神经网络或 Transformer 的视觉目标检测网络，也可以结合其他机器学习算法。

（二）创新性

本赛题鼓励参赛队伍提出创新算法框架或者在现有算法基础上做出创新改进，以

提高复杂场景下基于多源异构数据的视觉目标检测识别算法性能。例如，参赛队伍设计新的多源异构数据信息融合方法和视觉目标检测识别网络架构提升性能。

六、性能指标要求

本赛题使用经典目标检测性能指标 mAP@50-95 来评估提交方法的有效性。评测脚本将严格遵循以下逻辑计算模型得分，确保评估的公平性与准确性。mAP@50-95 衡量的是模型是在 IoU 阈值从 0.50 到 0.95 且步长为 0.05 的 10 个不同阈值下计算的 Average Precision（AP）的平均值，并对所有类别取平均。

1. 交并比 (Intersection over Union, IoU)

对于任意预测框B_{pred}和真实框B_{gt}，其重叠率（交并比）计算公式为：

$$\text{IoU} = \begin{cases} \frac{area(B_{pred} \cap B_{gt})}{area(B_{pred} \cup B_{gt})}, & if\ B_{pred} \neq \emptyset\ and\ B_{gt} \neq \emptyset \\ 0, & otherwise \end{cases},$$

其中$area(B_{pred} \cap B_{gt})$为两个框的交集面积，$area(B_{pred} \cup B_{gt})$是为两框并集面积。如果没有对应目标，那么 IoU 被赋值 0。

2. TP 与 FP 匹配策略

IoU thresholds: T = {0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95}

对于每个类别 c 和每个 IoU 阈值 t，判定逻辑如下：

（1）排序：将该类别下所有测试图像的预测框按置信度 (Confidence) 从高到低排序。

（2）匹配：依次遍历排序后的预测框：若该预测框与某个尚未被匹配的真实框满足 IoU≥t，则判定为 TP，并将该真实框标记为“已匹配”；若 IoU<t，或对应的真实框已被更高置信度的预测框匹配，则判定为 FP；若所有遍历结束后，未被任何预测框匹配的真实框记为 FN。

（3）累积统计：计算每一步的累积 TP 数量 TP(k)和累积 FP 数量 FP(k)，其中 k 为当前类遍历到的预测框索引。

3. 单类别 (c) 的 AP（Average Precision）在某个IoU阈值 t∈T 下的计算：

（1）计算原始 P-R 值：

对于每一个k，计算对应的召回率（Recall）和精确率（Precision）：

$$R_k = \frac{TP(k)}{N_{gt}},\ P_k = \frac{TP(k)}{TP(k)+FP(k)}$$

其中N_{gt}是该类别在所有测试图像中的真实目标总数。平均精度（AP）定义为精确率-召回率（P-R）曲线下的面积。本赛事采用 101 点插值法进行近似计算：即在召回率$R \in [0,1]$区间内均匀选取 101 个采样点(0.00, 0.01, …, 1)，召回率大于当前插值点 r 时，对应的最大精确率$P_{\text{interp}}(r)$，最终 AP 为这 101 个插值点的算术平均值：

$$AP_c(t) = \frac{1}{101} \sum_{r\in\{0,\ 0.01,\ldots,1.0\}} P_{\text{interp}}(\mathrm{r})$$

4. 计算每个 IoU 阈值下的 mAP:

对每个 IoU 阈值 t∈T，计算所有类别的平均 AP（其中C是类别总数）。

$$mAP(\mathrm{t}) = \frac{1}{C}\sum_{c=1}^{C} AP_c(t)$$

5. 对所有 10 个 IoU 阈值下的 mAP 取平均：

$$mAP@50-95 = \frac{1}{10}\sum_{t=0.50}^{0.95} mAP(t)$$

七、功能要求

（一）准确性

算法模型实现面向多源异构数据的视觉目标检测识别时，需要具备较高的准确性，确保在较为复杂的场景下也能准确识别目标

（二）鲁棒性

面对不同质量的 RGB 可见光图像、Depth 深度图像和 Infrared 红外图像数据，算法应能稳定运行，输出可靠的检测识别结果。在一个或者多个模态数据成像质量较差时，算法也不应出现大幅性能波动,保持对目标检测的准确性和稳定性。

八、开发环境

（一）软件环境

1. 编程语言

本赛题不强制编程语言的使用，但推荐使用 Python 语言的 3.8 及以上版本，因其丰富的数据处理、科学计算库和深度学习框架支持，如 NumPy、Pandas、Matplotlib 等用于数据处理和可视化的库、OpenCV 等图像数据处理的库、PyTorch 等深度学习框架库。

2. 深度学习框架

推荐使用 PyTorch，该框架在深度学习领域广泛应用，具有高效的计算性能和丰富的 API，便于模型的搭建、训练和部署。

3. 计算资源

本赛题不对参赛队伍使用的计算资源做限制，亦不提供算力支持。参赛者可使用本地工作站或云端计算平台进行开发和训练。

（二）硬件环境

本赛题对参赛队伍使用的 CPU 型号、内存大小、GPU 型号等硬件资源不做强制性要求，参赛队伍可根据自身需求自由配置。参赛者也可以使用云计算平台，例如，阿里云天池、腾讯云 TI-ONE、百度 AI Studio、火山引擎方舟平台、智谱 AI 开放平台等。

九、成绩评价

（一）输入数据格式要求

参赛者算法应能正确读取赛题数据，包括 RGB 图像、Infrared 图像与 Depth 图像，以及目标矩形边界框[class_id, norm_center_x, norm_center_y, norm_w, norm_h]的 TXT 文件。

（二）输出数据格式要求

算法对每个图像数据进行目标检测识别，要求每张测试图必须提交一个同名的预测 TXT 文件，文件中每一行表示一个目标在图像中的位置、置信度和类别信息：[class_id，norm_center_x, norm_center_y，norm_w，norm_h，confidence]。若某张图未检测到目标，也须提交一个空 TXT 预测文件，不允许缺失。此外，每张图最大预测框数量为 100，超过数量会按置信度截断；出现非法类别、非法坐标或置信度缺失，将导致该预测无效，不参与结果运算。最终将所有 TXT 文件打包为压缩包进行提交。

（三）成绩计算

成绩将根据算法输出结果与真实标注数据对比，依据性能评估指标 mAP@50-95 进行打分。

注：若参赛者提交结果低于赛事方公布的基线成绩，赛事方有权将其认定为无效成绩。

十、解题思路

（一）数据预处理

在数据训练过程中可以进行一些数据增强的预处理操作，例如对 RGB 图像、Infrared 图像和 Depth 图像进行归一化操作、进行图像旋转、缩放、翻转、裁剪等操作扩充训练数据集，增强模型的泛化能力。

（二）特征提取

参赛队伍应选取合适的深度学习网络结构提取 RGB 图像、Infrared 图像和 Depth 图像的特征，通过训练并结合机器学习方法如特征选择等实现高判别性和高鲁棒性的数据特征表示。

（三）模型训练

参赛队伍可以选择合适的深度学习框架（如 PyTorch）搭建模型，设置合理的训练参数，如学习率、迭代次数、批量大小等。在训练过程中，可自行划分验证集对模

型进行评估和调优，防止模型过拟合。

十一、赛题约束条件

（一）算法约束

❑ 禁止调用任何在线服务或 API，所有训练与推理须可在本地离线完成。

❑ 禁止采用手工标注测试集目标位置的方式生成结果，所有预测结果必须由算法模型自主推理输出，严禁任何形式的人工干预测试结果生成过程。

❑ 禁止将多个不同结构或训练阶段的模型进行简单集成，直接采用投票法、平均法等方式决策。

（二）数据使用约束

❑ 参赛者仅可使用赛事官方提供的训练数据集，严禁私自扩充外部数据。

❑ 允许使用 ImageNet、COCO、Objects365 等公开预训练权重。

❑ 参赛者仅可使用赛事官方统一提供的初赛、复赛、半决赛测试集数据进行算法预测，严禁私自对测试集进行标注、篡改，严禁将测试集数据用于模型训练环节。

❑ 赛事提供的所有测试数据集均为赛事专用，参赛者需严格遵守数据保密要求，禁止将数据集以任何形式（包括但不限于分享、上传至公开平台、转售、开源）泄露、传播或用于非本次赛事相关的商业、科研等用途，违反者立即取消参赛资格，并追究相关责任。

十二、参考资源

（一）文献资料

[1] Cao Y, Bin J, Hamari J, et al. Multimodal object detection by channel switching and spatial attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 403-411.

[2] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[3] Cheng C, Xu T, Wu X J, et al. EvaNet: towards More Efficient and Consistent Infrared and Visible Image Fusion Assessment[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2026.

[4] Tang Z, Xie Y, Xu T, et al. Learning Bi-Directional Fusion and Deformation-Sensitive Loss for RGB-T Tiny Object Detection[J]. Information Fusion, 2025: 103985.

[5] Zhu, X. F., Xu, T., Pan, Y., Gu, J., Li, X., Lu, J., ... & Kittler, J. Collaborating Vision, Depth, and Thermal Signals for Multi-Modal Tracking: Dataset and Algorithm. In The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2025.

（二）在线课程

1. Coursera 上的“DeepLearningSpecialization”课程，由吴恩达教授授课，涵盖了深度学习的多个关键领域，包括神经网络基础、卷积神经网络、循环神经网络等，课程内容丰富且理论性强。

2. Bilibili 上的“动手学深度学习 Pytorch 版”课程，讲解了如何使用 Python 语言 Pytorch 深度学习框架实现深度神经网络，课程内容的应用实践性强。

十三、提交要求

初赛、复赛：参赛队伍需提交测试集预测结果文件。

半决赛：参赛队伍需提交算法代码、预测结果文件、模型文件与环境、技术报告。具体提交内容如下：

（一）算法代码

参赛队伍应提交完整的算法代码，包括数据预处理、模型训练、预测推理等各个环节的代码。代码需使用符合规范的 Python 语言编写，具备清晰的注释和文档说明，以便评审人员理解和运行。

（二）测试结果文件、模型文件与环境

参赛队伍应提交测试的结果文件、训练好的模型文件（由于模型可能较大，请提供模型下载链接），并提供模型的加载和使用说明，包括所需的运行环境、依赖库等信息。模型文件应能够在指定的测试环境中正常运行并输出预测结果。

（三）技术报告

参赛队伍应提交详细的技术报告，内容包括算法设计思路、模型架构图、实验设置（如训练参数、数据增强方法等）、性能分析（对主要指标和次要指标的详细分析）以及算法的创新点和不足之处的分析。技术报告格式采用 PDF，页数不限。内容格式可参考“附件：算法赛题技术方案大纲”。

总决赛：提交内容及具体要求以组委会后续正式通知为准。

十四、其他说明

公平性：严禁任何形式的作弊行为，包括但不限于数据泄露、模型预训练数据与测试数据重叠、抄袭他人代码等。一经发现，立即取消参赛资格，并追究相关责任。

知识产权：参赛者提交的作品必须为原创，未在其他比赛中获奖或公开发表。比赛主办方有权对参赛作品进行展示、宣传等相关活动，但知识产权仍归参赛者所有。

十五、联系方式

赛题交流 QQ 群：1023687056

邮　箱：xuefeng.zhu@jiangnan.edu.cn

报名官网：www.aicomp.cn