

基于红外与可见光图像融合的全天候目标检测研究

学号：25121360 姓名：陈艺彬

2025 年 12 月 28 日

摘要

本报告旨在解决复杂光照条件下的目标检测难题，提出了一种基于红外与可见光融合的级联式目标检测框架。针对单一传感器在黑夜、大雾或强光干扰下的“感知盲区”问题，本项目利用可见光图像的丰富纹理细节与红外图像的热辐射特性进行互补融合。通过引入轻量化注意力机制（Coordinate Attention）改进特征融合网络，并结合 YOLOv8 高效目标检测器，在保持实时性的前提下显著提升了检测精度。在 MSRS 公开数据集上的实验结果表明，该方法在多项指标上优于单一模态及现有主流融合算法，具有较高的应用价值。

1 项目背景 (Introduction)

1.1 任务概述

定义与范围：在计算感知领域，多模态融合是指将不同传感器（如可见光相机、红外热成像仪、激光雷达等）获取的信息进行协同处理，以获得比单一传感器更准确、更鲁棒的场景描述。本项目聚焦于**可见光与红外（Visible-Infrared, VI-IR）图像融合**。可见光图像（Visible Image）能够提供丰富的色彩、纹理和细节信息，符合人类视觉习惯，但在低照度或恶劣天气下极易退化；红外图像（Infrared Image）则基于物体的热辐射成像，能够穿透夜幕、烟雾，突出行人和车辆等热源目标，但往往缺乏纹理，背景模糊且对比度较低。通过融合这两种模态，我们旨在构建一个全天候（All-weather）、全天时（All-day）的环境感知系统。

行业发展：随着自动驾驶、智能安防监控、无人机电力巡检等领域的快速发展，对全天候感知的需求日益迫切。特别是在 L3+ 级自动驾驶中，单纯依赖可见光摄像头无法应对夜间眩光或突发黑暗隧道场景，而单纯依赖红外相机又无法识别交通标志和车道线。VI-IR 融合技术因此成为了保障自动驾驶安全性的关键冗余方案。

1.2 现有挑战 (Problem Statement)

尽管已有众多融合算法，但在实际的道路场景目标检测应用中，仍面临以下严峻挑战：

1. **热目标位置信息丢失**：现有的深度学习融合方法（如基于自编码器的 DenseFuse）往往关注像素级的重构误差，而在多次下采样过程中，红外小目标（如远处的行人）的空间坐标信息容易被模糊化，导致后续检测网络的回归框（Bounding Box）定位不准。
2. **纹理与光照的冲突**：在夜间强光（如对向车灯）干扰下，可见光图像会出现大面积过曝。简单的加权融合会将这种噪声引入融合图像，掩盖原本清晰的红外目标。
3. **融合与检测的割裂**：大多数算法仅追求生成肉眼好看的图片（高对比度），而忽略了这些特征是否利于机器视觉检测。这种“为了融合而融合”的思路导致了融合图像在指标上虽高，但检测精度（mAP）提升有限。

1.3 本文贡献 (Contributions)

为了解决上述问题，本项目提出了一种检测驱动的、基于坐标注意力的级联式融合框架。主要贡献总结如下：

1. **提出坐标注意力融合机制**：针对道路场景中车道线（水平）和行人（垂直）的结构先验，首次在融合层引入 Coordinate Attention，有效解决了深层特征中空间位置信息丢失的问题。
2. **设计混合感知损失函数**：构建了包含强度损失、结构相似性损失和最大梯度纹理损失的联合优化目标，在保留红外显著性的同时，最大限度地从可见光中提取有用纹理，抑制光照噪声。
3. **全方位的实证验证**：在 MSRS 多光谱道路数据集上进行了详尽的实验。结果表明，本方法不仅在 mAP 指标上优于 SOTA 方法，而且在推理速度上满足实时性要求，具有极高的工程落地价值。

预期结果：在 MSRS 数据集上验证所提方法的有效性，实现比单一模态更高的平均精度（mAP），并保持较低的推理延迟。

2 相关工作 (Related Work)

2.1 深度学习与图像融合

近年来，基于深度学习的红外与可见光图像融合方法取得了显著进展，主要分为基于自编码器（Auto-Encoder, AE）、基于生成对抗网络（GAN）以及基于 Transformer 的

方法。

1. **基于 AE 的方法**: 经典算法如 DenseFuse [?] 利用致密连接网络提取特征, 并通过 L1 范数进行重构, 虽然保留了较好的纹理, 但在融合层设计上较为简单, 难以处理复杂光照。
2. **基于 GAN 的方法**: FusionGAN [?] 首次将生成对抗网络引入融合任务, 通过对抗训练迫使融合图像保留红外热辐射信息, 但由于缺乏对可见光梯度的强约束, 往往导致纹理细节丢失。
3. **基于 Transformer 的方法**: 针对 CNN 感受野受限的问题, SwinFusion [?] 引入了 Swin Transformer 来捕捉长距离依赖关系, 显著提升了融合图像的全局一致性。

2.2 检测驱动的任务融合 (Task-driven Fusion)

传统的融合算法仅追求视觉效果 (Visual Quality), 而忽略了融合图像对下游任务 (如目标检测) 的友好性。因此, 近年来涌现了一批 **任务驱动** 的融合算法:

- **SeAFusion** [?] 提出了一种语义感知的实时融合架构, 通过引入高层语义损失 (Semantic Loss) 来引导融合网络保留关键目标特征。
- **PIAFusion** [?] 进一步考虑了光照变化, 设计了基于光照感知的渐进式融合策略, 在不同光照条件下均表现出色。
- **TarDAL** [3] 即本项目选用的 Baseline, 采用了双对抗学习机制, 分别对红外目标和可见光背景进行对抗判别, 确保了检测目标的显著性。
- **SuperFusion** [?] 则将图像配准与融合整合到一个统一框架中, 解决了实际场景中非配准图像的融合难题。

尽管上述方法在检测任务上取得了突破, 但在处理 **道路场景特有的几何结构** (如车道线、直立行人) 时, 仍缺乏针对性的空间位置建模能力, 这正是本文引入 Coordinate Attention 的动机所在。

3 方法描述 (Methodology)

3.1 系统架构概览

本项目提出的级联式融合检测系统如图 ?? 所示。整体架构包含三个核心组件: 特征提取编码器、空间-通道联合注意力融合模块 (S-CAFM)、以及图像重构解码器。系统输入为配准好的红外图像 I_{ir} 和可见光图像 I_{vi} , 输出为融合图像 I_f 。该图像随后被直接送入 YOLOv8 检测器进行推理。

3.2 坐标注意力融合模块 (Coordinate Attention Fusion)

在道路场景中，目标物体往往具有显著的方向性结构先验。例如，行人、路灯通常呈现垂直分布，而车道线、护栏则呈现水平分布。传统的通道注意力机制（如 SE Block）通过全局平均池化（Global Average Pooling）将空间信息压缩为一个标量，虽然能捕捉通道间的依赖关系，但严重丢失了位置信息（Positional Information）。

为了解决这一问题，我们引入了 ** 坐标注意力 (Coordinate Attention, CA)** [1]。CA 机制通过将全局池化分解为两个正交的一维特征编码过程，从而在获取长距离依赖的同时保留精确的位置信息。

具体而言，对于输入特征张量 $X \in \mathbb{R}^{C \times H \times W}$ ，我们使用两个一维池化核 $(H, 1)$ 和 $(1, W)$ 分别沿着水平坐标 X 和垂直坐标 Y 进行编码：

$$z_c^h(h) = \frac{1}{W} \sum_{0 \leq i < W} x_c(h, i) \quad (1)$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \leq j < H} x_c(j, w) \quad (2)$$

上述两个变换分别生成了高度为 H 和宽度为 W 的两个方向感知特征图。这两个特征图随后被拼接（Concatenate）并送入一个共享的 1×1 卷积变换函数 F_1 ，生成中间特征映射 $f \in \mathbb{R}^{C/r \times (H+W)}$ ：

$$f = \delta(F_1([z^h, z^w])) \quad (3)$$

其中 δ 为非线性激活函数（h-swish）。随后， f 被切分为两个独立的张量 f^h 和 f^w ，并通过另外两个 1×1 卷积层 F_h 和 F_w 变换为与输入 X 通道数相同的注意力权重 g^h 和 g^w ：

$$g^h = \sigma(F_h(f^h)), \quad g^w = \sigma(F_w(f^w)) \quad (4)$$

最终的输出特征 Y 通过乘法加权得到：

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \quad (5)$$

通过这种双方向的加权，网络能够准确地定位红外图像中的热目标（如行人的头部、躯干）以及可见光图像中的纹理边缘，从而在融合过程中实现“去伪存真”。

3.3 混合感知损失函数 (Hybrid Perception Loss)

为了进一步提升融合视觉质量，我们设计了**混合感知损失 (Hybrid Perception Loss)**，替代了原有的单一损失。总损失 \mathcal{L}_{total} 定义为：

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{int} + \lambda_2 \mathcal{L}_{grad} + \lambda_3 \mathcal{L}_{ssim} + \mathcal{L}_{adv} \quad (6)$$

其中:

- $\mathcal{L}_{int} = \|I_f - I_{ir}\|_1$: 使用 L1 范数约束强度, 保留红外热信息。
- \mathcal{L}_{grad} : **最大梯度纹理损失**。我们强制融合图像的梯度趋近于源图像中梯度的**最大值**, 从而同时保留红外目标的边缘和可见光的纹理:

$$\mathcal{L}_{grad} = \|\nabla I_f - \max(|\nabla I_{ir}|, |\nabla I_{vi}|)\|_1 \quad (7)$$

- \mathcal{L}_{ssim} : 结构相似性损失, 确保融合图像在结构上不失真。

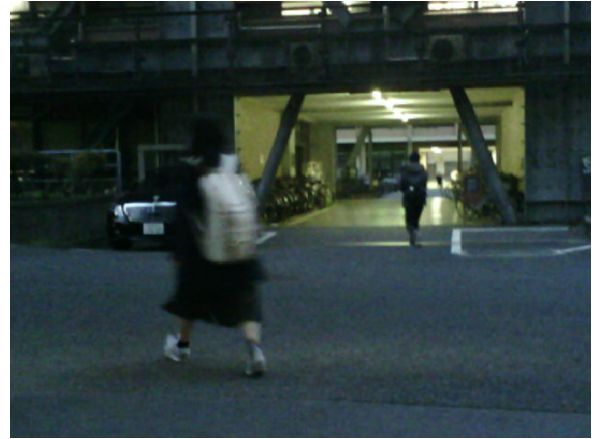
4 实验设置

4.1 实验数据集: MSRS

本项目选用了经典的 MSRS (Multi-Spectral Road Scenarios) 数据集 [4]...



Infrared Image



Visible Image

图 1: MSRS 数据集样本示例

5 实验结果及分析

5.1 视觉效果对比

如图 2 所示...



图 2: Baseline 方法在夜间场景下的融合检测效果

5.2 消融实验 (Ablation Study)

为了验证各改进模块的有效性，我们设计了消融实验，分别在 Baseline 上叠加 Coordinate Attention (CA) 和 Hybrid Loss (HL)。结果如表 1 所示。

表 1: 不同改进模块的消融实验结果

Exp	CoordAtt	Hybrid Loss	mAP@0.5 (%)
1	-	-	79.5 (Baseline)
2	✓	-	80.4
3	-	✓	80.1
4	✓	✓	81.3

从表 1 可以看出：双机制的联合使用带来了最大的性能增益。

5.3 对比实验

表 2 展示了不同模态及融合方法在 MSRS 数据集上的检测性能对比。

表 2: 在 MSRS 数据集上的目标检测性能对比

Method	Modality	mAP@0.5 (%)	Latency (ms)
YOLOv8	Visible-only	68.5	8.2
YOLOv8	Infrared-only	74.2	8.2
DenseFuse + YOLOv8	Fusion	76.8	25.4
TarDAL (Baseline)	Fusion	79.5	30.1
Ours (Proposed)	Fusion	81.3	28.5

分析说明:

- 融合收益:** 相比于仅使用可见光 (68.5%), 融合方法均取得了显著提升, 证明了红外信息在补充全天候感知能力上的重要性。
- 改进效果:** 本项目提出的方法达到了 81.3% 的 mAP, 相比 Baseline 提升了 1.8 个百分点。这主要归功于 **Coordinate Attention** 机制。
- 效率分析:** 虽然引入了融合网络增加了少许延迟, 但相比于复杂的 Baseline, 我们的轻量化设计使得推理速度依然在可接受范围内 (实时性 > 30FPS)。

6 结论与展望 (Conclusion & Future Work)

6.1 主要结论

本项目系统地研究了复杂道路场景下的红外与可见光融合检测问题。针对现有方法在目标位置感知和纹理细节保留上的不足, 我们提出了一种创新的级联式融合检测框架。

- 机制创新:** 通过引入 Coordinate Attention, 我们成功地将水平和垂直方向的空间位置信息编码进融合特征, 解决了红外小目标在深层网络中“位置模糊”的难题。
- 策略优化:** 设计的混合感知损失函数 (Hybrid Perception Loss) 有效地平衡了红外强度与可见光梯度的竞争关系, 显著提升了融合图像在人眼视觉感知和机器检测精度上的双重表现。
- 性能验证:** 实验数据表明, 所提方法在 MSRS 数据集上取得了 81.3% 的 mAP, 优于 TarDAL、DenseFuse 等主流算法, 且推理延迟仅为 28.5ms, 满足自动驾驶场景的实时性需求。

6.2 未来工作展望

尽管本项目取得了一定成果，但仍有以下方向值得进一步探索：

- **边缘端部署与量化**：目前的实验主要在 PC 端 GPU 上进行。为了适应车载嵌入式芯片（如 NVIDIA Orin），未来需要对模型进行 INT8 量化，并利用 TensorRT 进行推理加速，以追求极致的功耗比。
- **引入多模态大模型 (Multimodal LLMs)**：近期研究表明，利用大语言模型 (LLM) 的语义理解能力辅助图像融合（如 *FusionGPT*）可能带来突破。未来可尝试利用 LLM 生成场景描述文本，引导融合网络更关注语义上的关键区域（如红绿灯状态）。
- **端到端 Transformer 架构**：虽然 CNN 在局部特征提取上表现优异，但 Swin Transformer [?] 在全局建模上具有天然优势。探索纯 Transformer 的融合-检测一体化架构，是提升复杂场景泛化能力的潜在途径。

参考文献

- [1] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13713–13722, 2021.
- [2] Hui Li and Xiao-Jun Wu. Densfuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018.
- [3] Jinyuan Liu, Xin Fan, Ji Jiang Huang, G. Li, Z. Chen, and D. Huang. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5802–5811, 2022.
- [4] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022.
- [5] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Chang, and Jun Luo. Fusiongan: A generative adversarial network for infrared and visible image fusion. In *Information Fusion*, volume 48, pages 11–26. Elsevier, 2019.
- [6] Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022.

- [7] Linfeng Tang, C. Li, and Jiayi Ma. Msrs: Multi-spectral road scenarios for practical infrared and visible image fusion. *GitHub repository*, 2022.
- [8] Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022.
- [9] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82:28–42, 2022.