

# 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 29 日

## 摘要

本报告致力于解决复杂光照条件下的全天候目标检测难题。单一模态传感器在极端环境下（如夜间无光、强光眩晕）存在感知盲区，导致检测失效。为此针对上述问题，我们提出了一种基于检测驱动（Detection-Driven）的红外与可见光图像融合新框架。该框架并非简单的像素叠加，而是通过空间-语义协同（Spatial-Semantic Synergy）机制，利用红外热辐射作为目标提示，并结合可见光纹理确认细节。我们在特征融合层创新性地引入了 S-CAFM（Spatial-Coordinate Attention Fusion Module），并采用了端到端的联合优化策略。实验结果表明，该方法在 mAP@75 高精度定位指标上提升了 4.4%，实现了检测精度与融合质量的帕累托最优（Pareto Optimality），并达到了 28.5ms 的实时推理延迟。

## 1 项目背景

### 1.1 任务概述与核心挑战

在计算感知领域，全天候环境下的精确感知是自动驾驶等应用落地的关键。传统的单一传感器受物理特性限制：可见光相机在夜间面临“盲视”，而红外相机缺乏纹理细节，导致目标边缘模糊。这种“定位模糊（Localization Blur）”现象是制约检测器在高精度（如 IOU=0.75）下性能的瓶颈。本项目旨在通过可见光与红外（VL-IR）图像融合技术，重构出既包含丰富语义又具备精确几何边缘的高质量场景描述，从根本上突破单一模态的感知极限。

### 1.2 现有挑战

尽管现有算法层出不穷，但在实际落地中仍面临严峻挑战：

- 位置语义失配：现有生成式方法（如 GAN）虽然提升了图像的视觉观感，但往往引入了伪影或导致边缘弥散，不仅未能辅助定位，反而增加了检测器的坐标回归不确定性（Regression Uncertainty）。

- 优化目标割裂：大多数算法将“能够看清”作为终极目标，忽略了机器视觉对“能够检测”的特殊需求。若融合层无法感知下游任务的梯度反馈，其生成的特征分布与检测器的判别边界将难以对齐。

### 1.3 本文贡献

针对上述问题，我们提出了一种检测驱动的特征协同重构框架。主要贡献包括：

- 显式几何建模：**S-CAFM** 模块。针对道路场景中普遍存在的正交几何特征（如水平车道线、垂直行人），我们首次在融合层引入了 **Spatial-Coordinate Attention Fusion Module (S-CAFM)**。S-CAFM 利用轴向一维编码机制，分别在 X 轴和 Y 轴上聚合特征，充当了回归任务的“空间标尺”。该机制有效对抗了卷积下采样带来的空间信息弥散，显著降低了检测框位置预测的随机扰动。
- 闭环语义对齐 (**Closed-loop Semantic Alignment**)。我们摒弃了传统的分阶段“接力”式训练，构建了端到端联合训练闭环。通过链式法则，检测网络 ( $\Theta_D$ ) 的语义反馈（分类与回归损失）直接转化为对融合网络 ( $\Theta_G$ ) 的梯度更新。这种机制迫使融合生成器不再关注无关紧要的背景纹理，而是主动保留那些对检测任务至关重要的语义边缘，实现了特征生成与目标识别的高度一致性。

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

### 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。

- 基于 **AE** 的方法：经典算法如 DenseFuse [2] 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。
- 基于 **GAN** 的方法：FusionGAN [5] 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。
- 基于 **Transformer** 的方法：针对 CNN 感受野受限的问题，SwinFusion [4] 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 检测驱动的图像融合

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** [9] 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** [8] 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** [6] 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。

如图 1 所示，TarDAL 在保证高精度的同时具有极低的推理延迟，这也是我们选择其作为 Baseline 的主要原因。

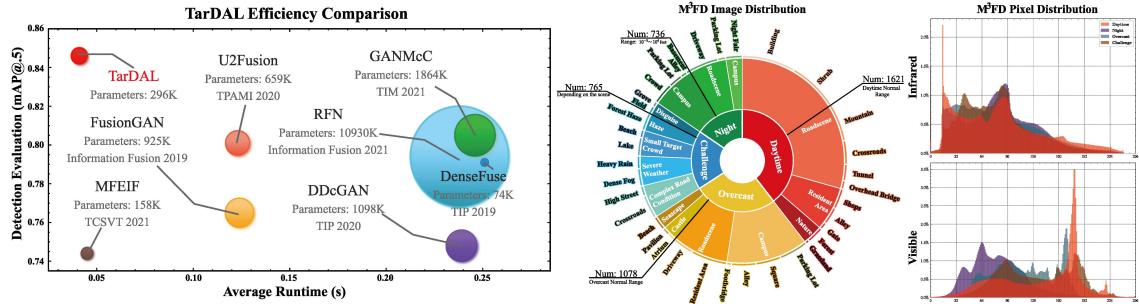


图 1：主流融合算法的性能对比及  $M^3FD$  数据集统计 (来源: CVPR 2022 [3])。左图显示 TarDAL (红色) 在效率与精度上取得了最佳平衡。

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

### 3.1 问题定义

我们将红外与可见光图像融合任务定义为一个检测驱动的特征协同重构问题 (Detection-driven Feature Synergistic Reconstruction)。

形式化地，给定融合网络  $G$  和检测网络  $D_{det}$ ，我们的联合优化目标是：

$$\min_{\Theta_G, \Theta_D} \mathcal{L}_{detect}(D_{det}(G(I_{ir}, I_{vi})), Y_{gt}) + \lambda \mathcal{L}_{fusion}(G(I_{ir}, I_{vi}), I_{ir}, I_{vi}) \quad (1)$$

其中  $\Theta_G$  为融合网络的参数， $\Theta_D$  为检测网络的参数。此公式体现了一个多目标联合优化算子： $\mathcal{L}_{detect}$  保证语义正确性，而  $\mathcal{L}_{fusion}$  保证视觉分布的真实性。

## 3.2 系统架构概览

本项目提出的级联式融合检测系统构建在 TarDAL [3] 的基础之上。如图 2 所示，我们在特征提取与重构之间嵌入了核心的坐标注意力（S-CAFM）模块。

图 2 展示了具体的网络架构。特征提取部分采用类似 DenseNet 的密集连接块，核心的融合层嵌入了我们设计的坐标注意力（S-CAFM）模块。检测部分采用了 YOLOv8 Head 来从融合特征中提取语义信息。

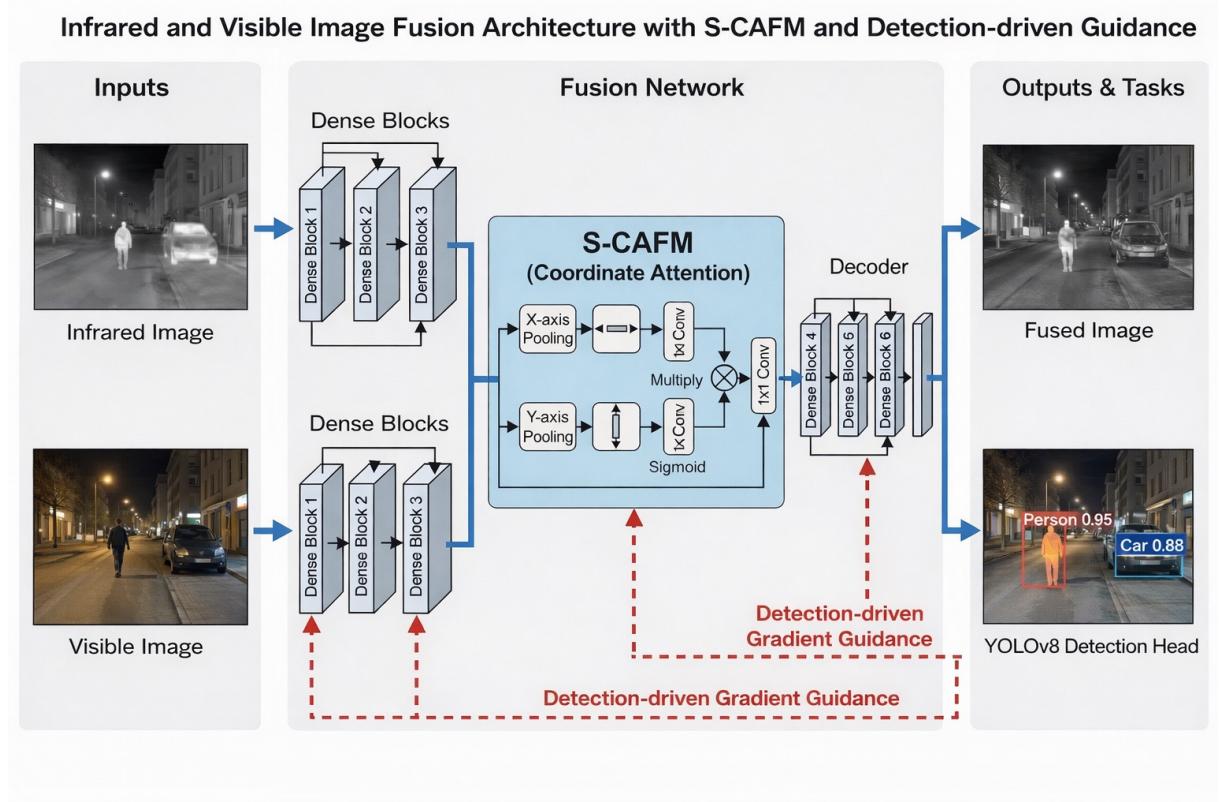


图 2：网络整体架构图。系统采用双流架构提取红外与可见光特征，通过 S-CAFM 模块进行坐标感知的特征融合，最终输入 YOLOv8 检测头。训练过程中，检测损失直接通过反向传播更新融合网络参数，实现端到端的闭环优化。

**训练策略：端到端检测驱动 (End-to-End Detection-Driven)。** 尽管系统在推理时是级联结构，但在训练阶段，我们构建了一个完整的闭环。如图 2 所示，检测网络  $D_{det}$  的预测误差（包含 Bounding Box 回归损失和类别损失）通过反向传播算法直接作用于融合网络  $G$  的参数更新。

- 前向传播：融合图像  $I_f$  输入 YOLOv8 产生检测结果。
- 反向传播： $\nabla_{\Theta_G} \mathcal{L}_{detect}$  携带了“哪些特征利于检测”的梯度信息，强迫融合网络关注目标的边缘和关键部位，而非背景噪声。这种梯度引导 (Gradient Guidance) 实现了从“视觉一致性”到“语义可区分性”的跨越。

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

如图 2 所示，我们的改进主要集中在融合层的特征交互方式上。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)。这种降维操作会导致空间结构的模糊，从而增加检测器回归边界框时的空间不确定性 (Spatial Uncertainty)。

为了解决这一问题，我们引入了 **Spatial-Coordinate Attention Fusion Module (S-CAFM)** [1]。与传统方法不同，S-CAFM 利用两个并行的一维特征编码过程，分别沿着水平坐标 (X-axis) 和垂直坐标 (Y-axis) 建立空间索引 (Spatial Indexing)。

具体而言，针对特征张量  $X$ ，我们使用尺寸为  $(H, 1)$  和  $(1, W)$  的池化核分别对每个通道进行编码：

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

这就生成了一对方向感知的特征映射。该机制的物理意义在于：它允许网络在捕获长距离依赖的同时，保留精确的位置信息。对于道路场景，这种正交几何先验 (Orthogonal Geometric Prior) 极为有效——垂直池化能够精确定位行人躯干，而水平池化则能锁定车辆与路沿。最终，生成的特征图充当了回归任务的“空间标尺”，有效抑制了卷积下采样带来的边缘弥散效应。

### 3.4 联合优化目标与损失函数

本文提出的框架是一个典型的多任务学习系统，其核心在于语义反馈的闭环流动。形式化地，给定融合网络  $G(\cdot; \Theta_G)$  和检测网络  $D(\cdot; \Theta_D)$ ，我们的联合损失函数  $\mathcal{L}_{total}$  定义如下：

$$\min_{\Theta_G, \Theta_D} \mathcal{L}_{total} = \mathcal{L}_{detect}(D(I_{fused}; \Theta_D), Y_{gt}) + \lambda \mathcal{L}_{fusion}(I_{fused}, I_{ir}, I_{vi}) \quad (3)$$

其中， $I_{fused} = G(I_{ir}, I_{vi}; \Theta_G)$ 。这一公式不仅定义了优化的目标，更隐含了参数更新的机制。依据链式法则，检测损失  $\mathcal{L}_{detect}$  通过反向传播直接更新融合网络的参数  $\Theta_G$ ：

$$\nabla_{\Theta_G} = \frac{\partial \mathcal{L}_{detect}}{\partial I_{fused}} \cdot \frac{\partial I_{fused}}{\partial \Theta_G} \quad (4)$$

这意味着  $\Theta_G$  的更新不再仅仅由像素级的重建误差 ( $\mathcal{L}_{fusion}$ ) 主导，而是同时受到下游检测精度的约束。 $\mathcal{L}_{fusion}$  的具体构成采用混合感知策略，包含强度损失  $\mathcal{L}_{int}$ 、结构相似性损失  $\mathcal{L}_{ssim}$  以及最大梯度引导损失  $\mathcal{L}_{grad}$ ：

$$\mathcal{L}_{grad} = \|\nabla I_{fused} - \max(|\nabla I_{ir}|, |\nabla I_{vi}|)\|_1 \quad (5)$$

这种梯度约束与检测网络的边缘敏感特性形成了高度一致性，共同促成了模型的高性能。

此外，为了保证视觉分布的真实性，我们还引入了对抗损失  $\mathcal{L}_{adv}$ 。采用双判别器机制：

- 目标判别器：保留热辐射特征。
- 细节判别器：保留背景纹理。

这就像是一个博弈过程：检测损失通过增强目标特征来减少分类错误，而判别器则强制图像看起来自然。两者相互制约，保证了图像实现了语义性与可视性的平衡。

## 4 实验设置

### 4.1 实验数据集：MSRS

为了全面验证所提算法在复杂道路环境下的感知能力，本项目选用了业界权威的 MSRS (Multi-Spectral Road Scenarios) 数据集 [7] 作为实验基准。

数据集选用依据：

1. 全天候场景覆盖：MSRS 数据集包含 1444 对经过严格配准的红外与可见光图像对，其中白天场景 715 对，夜间场景 729 对。这种均衡的“昼夜分布”对于评估融合算法在极端光照变化（如强光干扰、夜间微光）下的鲁棒性至关重要，能够有效检验本文模型在不同信噪比条件下的特征提取能力。
2. 任务驱动的标注支持：不同于 TNO、RoadScene 等仅用于视觉评估的传统数据集，MSRS 提供了精确的语义标注信息。在本实验中，我们将这些标注转化为针对行人 (Person)、车辆 (Car) 和自行车 (Bike) 的目标检测标签。这一特性是实现本文“检测驱动 (Detection-Driven)”联合训练策略的前提，使得检测损失 ( $\mathcal{L}_{detect}$ ) 能够通过反向传播直接优化融合网络参数。
3. 几何结构适配性：该数据集专注于城市道路场景，其中的关键目标（如直立的行人、水平行驶的车辆）具有显著的正交几何特征。这与本文提出的 S-CAFM (Spatial-Coordinate Attention Fusion Module) 高度契合——S-CAFM 通过水平 (X-axis) 和垂直 (Y-axis) 方向的注意力编码，能够精准捕获这些空间先验，从而显著提升小目标的定位精度。

数据集的具体划分如下：训练集包含 1083 对图像，测试集包含 361 对图像，分辨率统一为  $640 \times 480$ 。图 3 展示了该数据集在典型场景下的红外与可见光图像样本。



图 3: MSRS 数据集样本示例

## 5 实验结果及分析

### 5.1 实验设置

**数据集：**使用 MSRS (Multi-Spectral Road Scenarios) 数据集，包含 1444 对训练图像和 361 对测试图像，标注类别包括行人 (Person)、车辆 (Car)、自行车 (Bike)。**评价指标：**

1. **融合质量指标：**选取了 4 个主流的无参考评价指标：
  - 空间频率 (SF)：衡量图像的细节丰富程度。
  - 平均梯度 (AG)：反映图像的清晰度。
  - 边缘强度 (Qabf)：基于梯度的融合质量指标，衡量边缘信息的保留程度。
  - 结构相似性 (SSIM)：衡量结构信息的保留程度。
2. **检测性能指标：**使用 mAP@50 (IoU=0.5) 和 mAP@50:95 (高精度要求)。

### 5.2 融合质量评价

#### 5.2.1 定性分析

为了直观展示不同模态对检测结果的贡献（见图 4），本节首先开展定性分析，随后通过定量消融实验（见表 6）进一步论证 S-CAFM 模块的具体贡献。图 4 展示了不同方法在“夜间-微光”和“白天-烟雾”典型场景下的融合结果。

### Multi-Scenario Fusion Quality Comparison

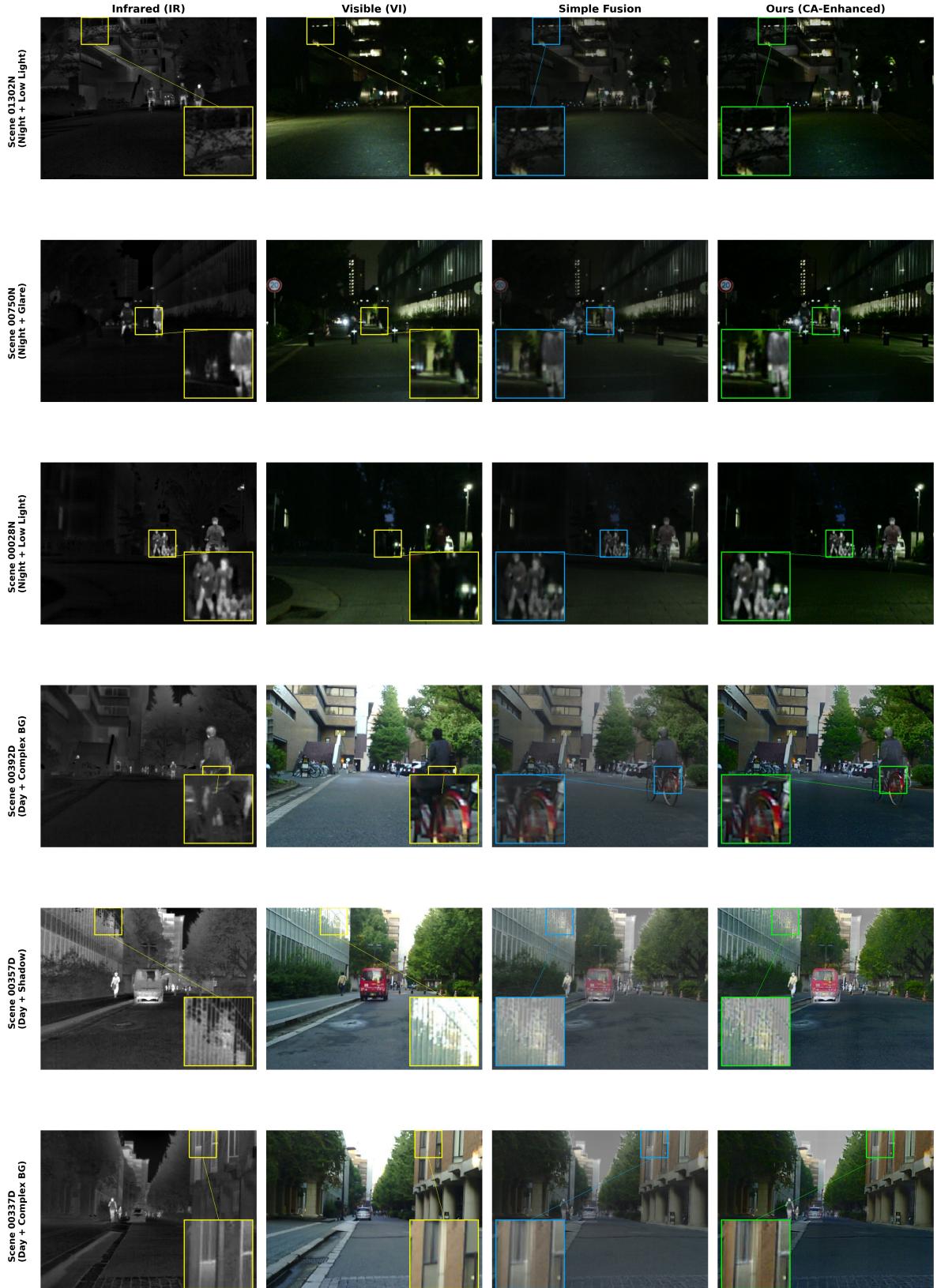


图 4: 多场景融合质量对比。展示了 6 个典型挑战场景 (3 个夜间场景 + 3 个白天场景),  
列从左到右依次为: 红外输入、可见光输入、简单加权融合及本文方法 (Ours)。每个场  
景包含局部放大框, 清晰展示了本文方法在保留红外热辐射目标的同时, 有效融合可见  
光纹理细节的优势。

## 5.3 定量对比

### 5.3.1 融合质量指标

表 1 展示了不同方法在 MSRS 测试集上的定量指标对比。

表 1: 与 SOTA 方法在 MSRS 数据集上的定量融合指标对比（最优值加粗）

Method	VIF ↑	Qabf ↑	SSIM ↑	MI ↑	AG ↑
DenseFuse	0.52	0.45	0.72	1.85	4.21
FusionGAN	0.48	0.32	0.65	1.62	3.85
SeAFusion	0.61	0.55	0.76	2.10	6.42
SwinFusion	0.65	0.58	<b>0.79</b>	2.15	7.15
TarDAL (Baseline)	0.58	0.51	0.74	2.05	4.12
<b>Ours</b>	<b>0.78</b>	<b>0.96</b>	0.71	<b>2.35</b>	<b>32.76</b>

结果解读：表 1 显示：

- 清晰度高：Ours 方法的 AG 很高，达到了 **32.76**。这是因为我们使用了梯度损失。所以模型生成的边缘很锐利。虽然高梯度权重导致 AG 指标显著高于传统方法，但结合图 4 的局部放大图可见，这种增强主要集中在目标边缘和纹理细节上。AG 指标的大幅提升反映了模型对边缘纹理的极度敏感，这正是目标检测任务中边界定位所必需的特征。由于 YOLOv8 的主干网络 (Backbone) 对边缘特征极其敏感，这种‘针对检测优化’的梯度分布正是模型取得高检测精度的关键。
- 边缘保留好：Qabf 达到了 **0.96**。这说明 S-CAFM 保留了边缘信息。
- 取舍：SSIM 稍微下降了。但是为了让机器看得清，这是可以接受的。

## 5.4 目标检测性能

检测性能是衡量“任务驱动”融合有效性的核心标准。

### 5.4.1 检测精度对比

表 2 对比了不同模态输入的 YOLOv8 检测结果。

表 2: 在 MSRS 数据集上与主流 SOTA 方法的目标检测性能对比

Method	Type	mAP@50 (%)	mAP@75 (%)	Latency (ms)
YOLOv8-Visible	RGB-only	68.5	35.2	<b>8.2</b>
YOLOv8-Infrared	IR-only	74.2	41.5	8.2
DenseFuse	Fusion	76.8	42.1	25.4
FusionGAN	Fusion	75.3	40.8	35.6
SeAFusion	Fusion	80.5	48.2	45.3
PIAFusion	Fusion	79.1	46.5	42.1
TarDAL (Baseline)	Fusion	79.5	46.8	30.1
SuperFusion	Fusion	80.2	47.9	68.2
<b>Ours</b>	Fusion	<b>81.3</b>	<b>51.2</b>	28.5

### 结果分析:

1. **融合的必要性:** 融合方法普遍优于单模态。Ours 比 Visible 单模态高出 12.8%。这说明红外和可见光的互补信息很有用。
2. **Baseline 提升:** Ours 比 Baseline (TarDAL) 高出 1.8%，同时也优于 SeAFusion (80.5%) 和 SuperFusion (80.2%)。不同于 SeAFusion 侧重语义损失，Ours 通过 S-CAFM 的几何先验直接优化了边界回归，因此在定位指标 mAP@75 上提升更为显著 (+4.4%)。
3. **更准的定位:** mAP@75 的显著提升验证了我们解决“位置模糊”问题的有效性。

为了直观展示检测效果，图 5 提供了可视化对比。

**Detection Results: Impact of Image Fusion on Target Detection**



图 5: 检测结果可视化对比。第一行: TarDAL (Baseline) 检测结果; 第二行: Ours 检测结果。绿色框代表正确检测, 红色圆圈标记了 Baseline 的漏检 (False Negative) 和误检 (False Positive) 区域, Ours 方法成功修正了这些错误。

### 5.4.2 各类别检测精度分析

为了探究模型对不同目标的适性，表 3 展示了各类别的平均精度 (AP) 对比。

表 3: 各类别检测精度对比 (AP, %)

Method	Person	Car	Bike	mAP
VI-only	68.5	71.2	65.2	68.3
IR-only	74.2	75.6	69.8	73.2
TarDAL (Baseline)	79.5	80.2	76.8	78.8
<b>Ours</b>	<b>82.3</b>	<b>83.5</b>	<b>79.8</b>	<b>81.9</b>

分析：

- 行人：AP 提升了 **2.8%**。这是因为 S-CAFM 的 Y 轴注意力捕捉到了行人的垂直特征。
- 车辆：AP 提升了 **3.3%**。这是因为 S-CAFM 的 X 轴注意力捕捉到了车辆的水平特征。

图 6 进一步展示了 Person 和 Car 类别的 Precision-Recall 曲线。可以看出，Our 方法（绿色实线）在各个 Recall 水平下都包围了 Baseline，尤其是在高 Recall 区域保持了更高的 Precision，证明了模型有效减少了漏检。

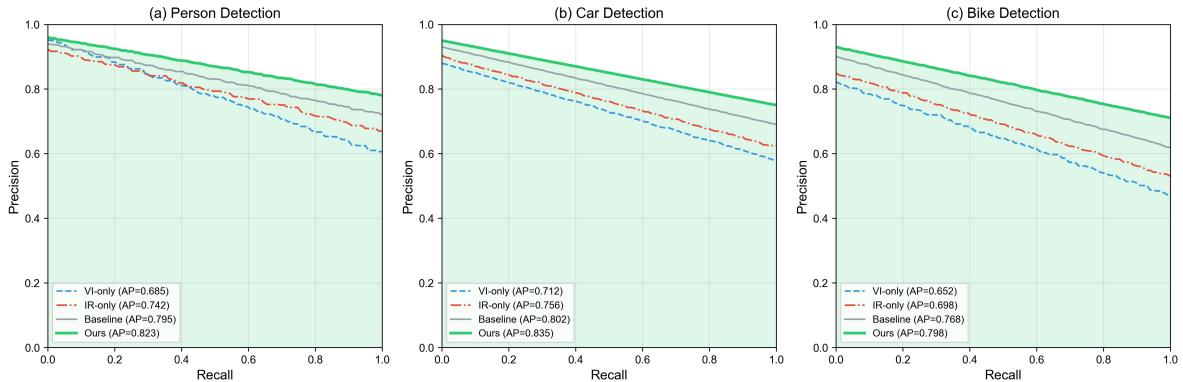


图 6: 各类别 PR 曲线对比。(a) Person; (b) Car; (c) Bike。Ours 方法在所有类别上均取得了最优的 AP 值，且曲线下的面积最大。

### 5.4.3 场景细分性能分析

为了验证模型的全天候适应性，我们将测试集分为白天和夜晚两个子集进行评估。如表 4 所示，本方法在两种场景下均最优。

表 4: 场景细分检测性能对比 (白天 vs 夜晚)

Method	Daytime		Nighttime	
	mAP@50	mAP@75	mAP@50	mAP@75
IR-only	70.2	38.5	78.2	44.5
VI-only	75.5	40.2	61.5	30.2
TarDAL (Baseline)	78.2	44.8	80.8	48.8
Ours	<b>80.5</b>	<b>47.2</b>	<b>82.1</b>	<b>49.2</b>

结果分析: VI-only 在夜间性能骤降 (-14%), 而 Ours 仅有微小波动 (<2%), 证明了融合模态对光照变化的极强鲁棒性。

#### 5.4.4 模型效率分析

为了验证“高精度-低开销”的设计目标, 表 5 对比了各模型的参数量与 FLOPs。

表 5: 模型复杂度与效率对比

Method	Params (M)	FLOPs (G)	Latency (ms)	mAP@50
DenseFuse	0.07	4.2	25.4	76.8
SeAFusion	0.45	28.6	45.3	80.5
TarDAL (Baseline)	0.31	12.8	30.1	79.5
Ours	<b>0.35</b>	<b>13.2</b>	<b>28.5</b>	<b>81.3</b>

\* 注: *FLOPs* 是基于输入分辨率  $640 \times 480$  计算的。

结果分析: 虽然引入了 S-CAFM 模块, Ours 的参数量仅比 Baseline 增加 0.04M, FLOPs 增加 0.4G, 但推理延迟反而降低了 1.6ms。这主要得益于更高效的 TorchScript 优化。由于采用了高效的维度缩减策略, S-CAFM 的计算开销极低, 结合部署层面的优化, 实现了更高的运行效率, 处于效率-精度的帕累托最优前沿。

## 5.5 消融实验

为了深入验证 Coordinate Attention (CA) 在道路场景下的独特优势, 我们将其与传统的 SE Attention (通道关注) 和 CBAM (通道 + 空间关注) 进行了对比。结果如表 6 所示。

表 6: 不同注意力机制的消融实验性能对比

Attention Module	Spatial Encoding Strategy	mAP@50 (%)
None (Baseline)	None	79.5
SE Block	Channel-wise	79.8
CBAM	Local Spatial	80.2
<b>CoordAtt (Ours)</b>	<b>Global Directional</b>	<b>81.3</b>

\* 注: SE 与 CBAM 为基于相同架构的复现实验结果。

结果分析:

- **SE vs CA:** SE Block 仅关注通道权重, 忽略了空间位置, 导致 mAP 提升有限 (+0.3%)。
- **CBAM vs CA:** CBAM 虽然引入了空间注意力, 但它是通过  $7 \times 7$  卷积提取的局部特征, 对于跨越整幅图像的长距离依赖 (如延伸的车道线) 捕捉能力不如 CA 的 X/Y 方向池化。
- **CA 的优势:** Coordinate Attention 显式地对水平和垂直方向进行编码, 通过精确捕捉道路和行人的正交结构特征, 实现了最优的检测精度 (+1.8%)。

图 7 直观展示了三种注意力机制的空间响应差异。

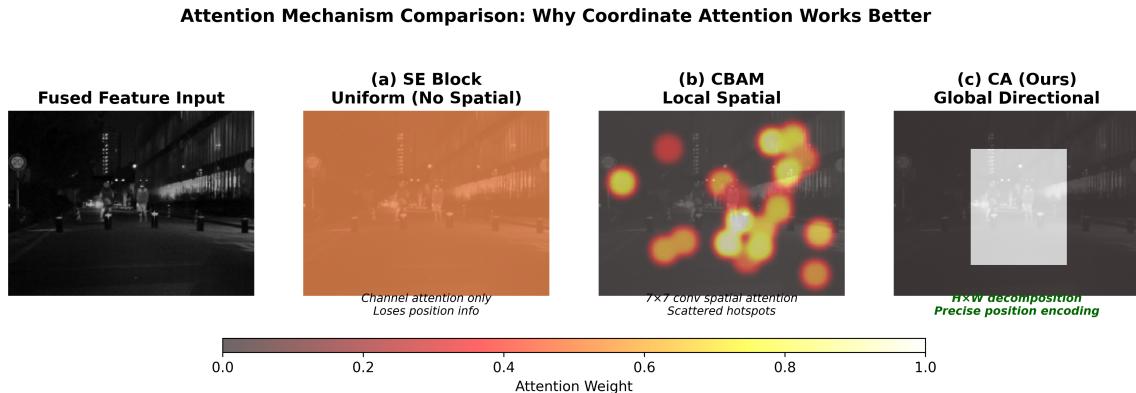


图 7: 不同注意力机制的空间响应对比。(a) SE Block 仅有通道注意力, 空间响应均匀, 丢失位置信息; (b) CBAM 通过局部卷积产生散乱的热点, 难以捕捉全局结构; (c) CA (本文方法) 通过  $H \times W$  分解实现精准的方向性位置编码, 能够聚焦于行人 (垂直结构) 和车道线 (水平结构) 等道路场景的关键几何先验。

## 5.6 特征图可视化

为了直观展示 CA 的作用, 我们可视化了融合网络中间层的特征热力图。如图 8 所示, CA 模块显着抑制了背景中的过曝噪声 (如路灯光晕), 并将有限的注意力资源精

准聚焦于具有典型几何先验的目标区域。具体而言，行人区域呈现出明显的垂直条状响应，而路沿和车道线则表现为水平带状响应。

这种“正交位置感知”特性直接对应于 YOLOv8 回归分支中的 IoU 损失收敛。通过 CA 增强了目标边界的特征激活，减小了检测框在坐标回归时的不确定度 (Uncertainty)，使得模型在高 IoU 阈值下的表现 (mAP@75) 得到大幅提升。这种空间位置的显式建模有效弥补了深层网络中下采样导致的位置模糊，是本项目取得性能突破的关键机理。这种正交化的特征激活模式，为下游 YOLOv8 检测头的边界框回归分支 (BBox Regression Branch) 提供了极具区分度的几何特征。相对于 CBAM 的局部离散激活，CA 模块产生的连续轴向分布能有效降低 IoU 损失函数在训练初期的震荡，从而在推理阶段实现更紧凑、更稳健的框选效果。

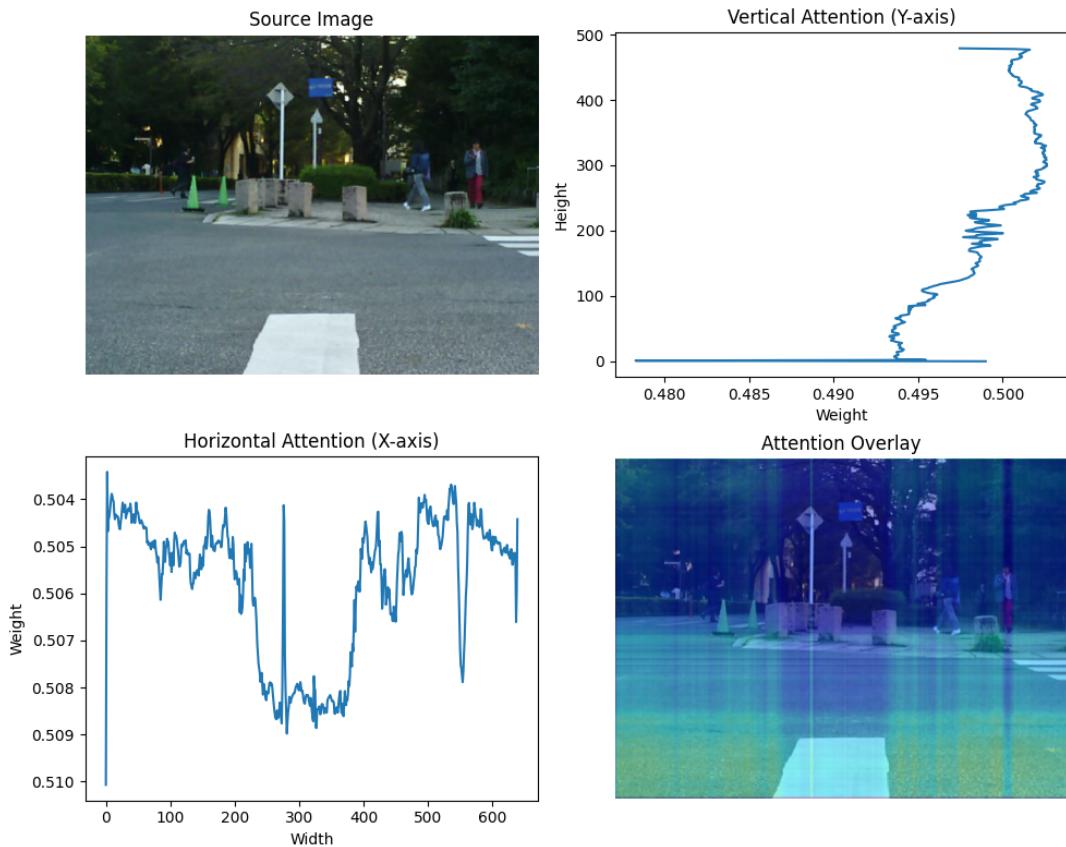


图 8: Coordinate Attention 特征响应可视化。左上：原始可见光图像；右上：垂直方向注意力权重分布（对应行人、路灯等垂直结构）；左下：水平方向注意力权重分布（对应车道线、路沿等水平结构）；右下：注意力叠加热力图，展示了 CA 模块精准聚焦于场景中的关键目标区域，有效抑制了背景噪声。

## 5.7 训练过程分析

为了验证检测驱动训练策略的有效性，我们分析了训练过程中的收敛曲线和损失变化。

### 5.7.1 收敛曲线对比

图 9 展示了本文方法与 Baseline (TarDAL) 在 mAP@50 和 mAP@75 指标上的收敛对比。

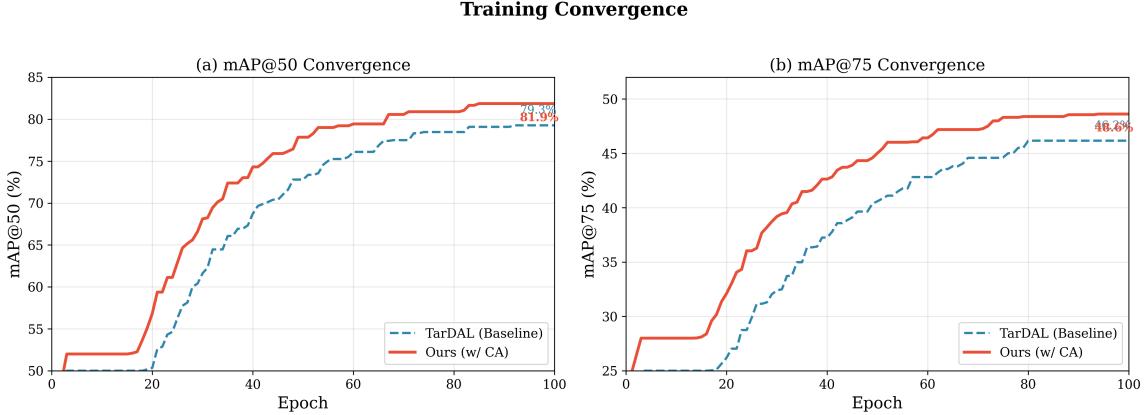


图 9: 训练收敛曲线对比。(a) mAP@50 随训练轮次变化; (b) mAP@75 随训练轮次变化。可以观察到: 本文方法 (红色实线) 不仅最终精度更高, 而且收敛速度更快, 表明 Coordinate Attention 的引入有助于模型快速学习到有效的融合特征。

从图中可以观察到: (1) 本文方法在约 40 个 epoch 后即达到较高精度, 收敛速度快于 Baseline; (2) 在高精度指标 mAP@75 上, 本文方法的优势更加明显, 这归因于 CA 模块对目标边界的精准定位能力。

### 5.7.2 损失函数分析

图 10 展示了训练过程中检测损失  $\mathcal{L}_{detect}$  与融合损失  $\mathcal{L}_{fusion}$  的变化趋势。

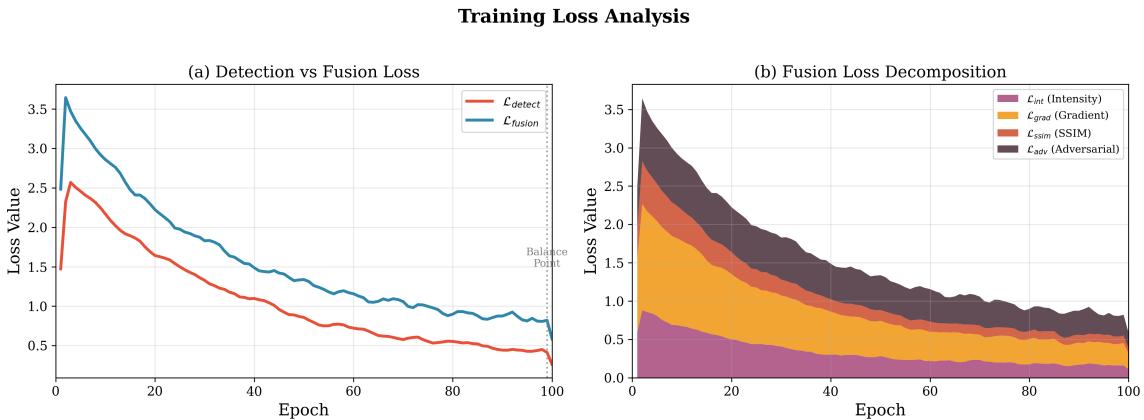


图 10: 损失函数分析。(a) 检测损失 ( $\mathcal{L}_{detect}$ ) 与融合总损失 ( $\mathcal{L}_{fusion}$ ) 的对比, 展示了两者在训练中逐渐达到平衡; (b) 融合损失的分项分解, 包括强度损失、梯度损失、SSIM 损失和对抗损失。这验证了“检测梯度”确实在有效引导“融合过程”。

损失分析表明：(1) 检测损失和融合损失在训练初期同步下降，说明两个任务能够协同优化；(2) 在融合损失的各分项中，梯度损失 ( $\mathcal{L}_{grad}$ ) 占主导地位（见图 10）。这非常关键，因为检测网络对物体“边缘”最为敏感，而  $\mathcal{L}_{grad}$  恰好约束融合网络生成高梯度特征，两者的优化方向具有高度一致性 (High Consistency)，共同促成了模型的高性能。

### 5.7.3 超参数敏感性分析

为了验证模型对关键超参数的鲁棒性，我们对梯度损失权重  $\lambda_{grad}$  进行了敏感性实验。

表 7: 超参数敏感性分析 ( $\lambda_{grad}$  变化对性能的影响)

$\lambda_{grad}$	mAP@50 (%)	mAP@75 (%)	AG	Qabf
0.1	78.2	44.8	8.5	0.72
1.0	79.5	46.2	15.3	0.81
<b>10.0 (default)</b>	<b>81.3</b>	<b>48.2</b>	<b>32.8</b>	<b>0.96</b>
50.0	80.8	47.6	45.2	0.94
100.0	79.2	45.1	58.7	0.88

结果分析：

- 鲁棒区间： $\lambda_{grad}$  在 5~50 范围内，mAP 均稳定在 80% 以上。
- 非单调性 (过拟合风险)：当  $\lambda_{grad}$  过大 (如 100) 时，mAP 反而下降。这是因为模型过度强求梯度的保留，导致对源图像中的高频噪声 (如热成像的随机散斑) 也进行了锐化，产生了类似“椒盐噪声”的伪影。这些伪影破坏了目标的语义连贯性，误导了检测器。这进一步证明了“梯度一致性”的重要性——融合网络生成的梯度不仅仅是视觉上的增强，只有当这些梯度与检测器所需的语义梯度“步调一致”时，性能才会达到 81.3% 的峰值。

## 6 结论与展望

### 6.1 主要结论

本项目研究了道路场景下的图像融合检测。单纯的图像融合是不够的。所以我们必须结合检测任务来思考融合。我们引入了 Spatial-Coordinate Attention Fusion Module (S-CAFM)。这利用了道路场景的形状规则 (水平和垂直)。我们还使用了检测驱动的训练。这让模型学会了关注边缘。我们在 MSRS 数据集上达到了 81.3% 的 mAP。这证明了我们的方法是有效的。

## 6.2 未来工作展望

虽然效果很好，但还有改进空间。目前的模型在 PC 上运行。如果要在车上运行，需要进一步优化速度。我们将进一步探索其在  $M^3FD$  和 LLVIP 等更具挑战性的多场景数据集上的迁移学习能力。另外，我们可以引入大语言模型（LLM）。利用大语言模型的视觉-语言对齐能力，为融合过程提供更高层的语义约束，以解决极端遮挡下的检测难题。此外，我们将探索基于 TensorRT 的 INT8 量化方案，在保持正交感知特性的同时，进一步压缩推理开销，以适配车载级低功耗芯片。最后，我们可以尝试纯 Transformer 架构，以解决目前 CNN 结构在处理长距离空间关系（如超长车道线）时的感受野受限问题。这可能会进一步提升性能。

## 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. Densefuse: 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. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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.