E-Commerce Letters 电子商务评论, 2025, 14(11), 1533-1541
Published Online November 2025 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2025.14113592

Hans汉斯

# 面向电商无人配送复杂场景的感知模型研究

**刘予莘**

贵州大学大数据与信息工程学院，贵州 贵阳

收稿日期：2025年10月9日；录用日期：2025年10月29日；发布日期：2025年11月20日

## 摘　要

**在电商高速发展与用工成本上升的背景下，无人配送具备降本增效、全天候运行与无接触服务等优势，但其规模化落地仍受限于复杂环境下的稳健感知。环境感知是保障无人配送车辆安全通行与高效决策的关键，其准确性直接决定配送车对动态场景的理解能力。基于Transformer的RT-DETR依托全局注意力与端到端检测实现了较高效率与精度，但在无人配送典型场景中的多尺度目标与频繁遮挡下，仍存在特征融合与遮挡鲁棒性不足。为此，本文提出面向电商无人配送的RT-DETR改进方案：在骨干网络关键层嵌入自适应聚焦全局上下文注意力模块，通过动态调节感受野增强多尺度表征，从而提升对小目标与遮挡体的可分辨性；并在FPN/PAN中引入指数移动平均增强的跨维度注意力机制，以更稳健地建模长程依赖并优化跨层特征融合。实验结果表明，改进模型在Udacity自动驾驶数据集上实现mAP@50提升25.6%、mAP@50-95提升13%，验证了方法在电商无人配送典型场景中的迁移性与应用价值。**

## 关键词

**无人配送，无人驾驶，目标检测，RT-DETR，AFGC，EMA**

# Research on Perception Models for Complex Scenarios in E-Commerce Unmanned Delivery

**Yushen Liu**

College of Big Data and Information Engineering, Guizhou University, Guiyang Guizhou

Received: October 9, 2025; accepted: October 29, 2025; published: November 20, 2025

## Abstract

**Against the backdrop of rapid e-commerce development and rising labor costs, unmanned delivery offers advantages, such as cost reduction, efficiency improvement, 24/7 operation, and contactless**

文章引用: 刘予莘. 面向电商无人配送复杂场景的感知模型研究[J]. 电子商务评论, 2025, 14(11): 1533-1541.
DOI: 10.12677/ecl.2025.14113592

**services. However, its large-scale deployment remains constrained by robust perception in complex environments. Environmental perception is crucial for ensuring the safe passage and efficient decision-making of unmanned delivery vehicles, with its accuracy directly determining the vehicle's ability to understand dynamic scenes. While the Transformer-based RT-DETR achieves high efficiency and accuracy through global attention and end-to-end detection, it still suffers from insufficient feature fusion and occlusion robustness when dealing with multi-scale objects and frequent occlusions in typical unmanned delivery scenarios. To address these issues, this paper proposes an improved RT-DETR model tailored for e-commerce unmanned delivery. An adaptive global context attention module is embedded into key layers of the backbone network to enhance multi-scale representation by dynamically adjusting the receptive field, thereby improving discernibility for small objects and occluded targets. Additionally, an exponential moving average-enhanced cross-dimensional attention mechanism is introduced into the FPN/PAN to more robustly model long-range dependencies and optimize cross-layer feature fusion. The experimental results demonstrate that the improved model achieved a 25.6% increase in mAP@50 and a 13% improvement in mAP@50-95 on the Udacity autonomous driving dataset, validating the transferability and application value of the proposed method in typical e-commerce unmanned delivery scenarios.**

## Keywords

**Unmanned Delivery, Autonomous Driving, Object Detection, RT-DETR, AFGC, EMA**



## 1. 引言

在电商末端无人配送加速落地的背景下[1]，配送小车需在复杂且多样化的作业环境中稳定运行，其系统的可靠性与安全性首先取决于环境感知能力。作为感知中枢，目标检测必须在实时约束下对行人、非机动车、车辆等多样目标进行精准识别与定位[2]。在遮挡频繁、光照变化显著、目标尺度长尾且目标密度波动的无人配送场景中，检测的精度与稳定性直接决定系统对潜在风险的前瞻性预判与安全冗余配置，因此持续提升目标检测性能是推动无人配送规模化与常态化运营的关键前提。

尽管深度学习显著推动了目标检测技术的发展，但在无人配送场景中仍存在诸多挑战：小目标检测受限于分辨率约束，密集遮挡易导致特征丢失，而动态环境对实时性能的要求依然严苛。为应对这些问题，学界提出了多种改进方案。例如基于 Transformer 的 DETR [3]模型通过全局注意力机制增强长程依赖建模，YOLO [4]系列则侧重轻量化设计以优化检测速度。然而传统方法仍难以有效平衡精度与速度，且纯 Transformer 架构的高计算成本制约了其在实时系统中的实用性。因此，开发更高效的解决方案已成为迫切需求。

作为基于 Transformer 的先进实时检测器，RT-DETR [5]融合了 CNN 的局部特征提取能力与 Transformer 的全局建模优势。该模型采用混合编码器架构与动态匹配策略，在保持端到端检测特性的同时显著提升推理速度。相较于传统方法，RT-DETR 通过自适应特征选择机制增强多尺度融合效率，在 Udacity 自动驾驶数据集等基准测试中实现精度与速度的协同提升，为自动驾驶感知提供了可行的技术路径。

尽管 RT-DETR 性能优异，但在复杂场景中仍存在两个关键局限：(1) 对遮挡目标与小物体的特征表达能力不足；(2) 特征金字塔内部的跨尺度信息交互效率有待提升。针对这些问题，本文提出针对性改进

方案：首先，在骨干网络关键层嵌入 AFGCAttention [6]模块，通过自适应感受野机制增强多尺度特征聚焦能力；其次，将 FPN/PAN 中的 RepC3 模块替换为由 EMA [7]改进的 EMA_attentionC3 结构，利用跨维度注意力与时序特征增强优化信息融合效果。

本文的主要贡献可归纳如下：

(1) 我们采用 AFGC 注意力模块，通过动态权重分配机制，增强骨干网络对遮挡目标与小目标的特征提取能力。

(2) 我们引入 EMA-Attention C3 结构，将基于 EMA 的时间建模与空间－通道双维度注意力结合，以提升特征金字塔中的长程依赖建模能力。

(3) 我们提出的改进方案，强化对遮挡、尺度变化与背景干扰的适应性，从而提升整体检测稳定性。

实验结果表明，改进后的模型性能方面优于原始 RT-DETR 模型。

## 2. 相关工作

### 2.1. 无人配送场景中的目标检测技术发展

无人配送面向复杂且多样化的作业环境，目标呈现长尾、小尺度与遮挡频发等特征[8]。两阶段方法以 Faster R-CNN [9]为代表，定位精度较高但推理开销大；单阶段方法以 YOLO 系列为代表，速度占优但在密集与遮挡场景下小目标易漏检。Transformer 检测如 DETR 及其改进依托全局建模提升复杂场景鲁棒性，但计算成本较高，车载侧实时部署受限。RT-DETR 在端到端范式下压缩时延并兼顾精度，更契合无人配送的实时需求，但在多尺度特征融合与遮挡鲁棒性方面仍有提升空间。

### 2.2. Transformer 在目标检测中的应用

Transformer 模型在自然语言处理领域的成功，推动了其在计算机视觉中的应用。DETR 首次将 Transformer 架构引入目标检测任务，采用编码器－解码器结构实现端到端检测，摒弃了传统方法中常用的锚框和非极大值抑制等组件。后续研究通过不同方向进行优化：如 DINO-DETR [10]则借助对比学习策略提升检测性能。然而，纯 Transformer 结构仍存在计算开销大的问题。为此，研究者开发了结合 CNN 优势的混合模型，例如通过分层设计提升效率的 Swin Transformer [11]，在保持全局建模能力的同时显著改善了计算效率。

### 2.3. 目标检测模型的优化方法

实时目标检测对自动驾驶系统至关重要。为提升模型效率，研究者提出了多种轻量化设计与结构优化方案。例如 YOLO 系列通过精简网络深度与宽度实现高速检测。而 RT-DETR 创新性地融合了 CNN 的局部特征提取能力与 Transformer 的全局建模优势，通过混合编码器架构与动态匹配策略实现实时检测。然而，该模型在复杂配送场景下的多尺度特征融合性能仍有提升空间。

## 3. 方法

本研究基于 Ultralytics 提出的 RT-DETR-l 检测框架，构建了面向无人配送场景的改进模型。RT-DETR 作为一种结合 Transformer 编码器－解码器结构与轻量 CNN 骨干的实时检测器，具备强大的端到端检测能力。该框架采用分层特征提取结构，输出 P3 (1/8)、P4 (1/16)和 P5 (1/32)三个分辨率层级的特征图，通过多尺度聚合进行特征融合与优化，最终经解码器生成检测结果。

在改进后的架构中，我们对骨干网络和特征融合网络进行了针对性增强：首先，在骨干网络关键阶段嵌入 AFGC 注意力模块，通过频域感知机制增强全局语义表征，提升对小目标与遮挡物体的感知能力；

其次，将特征融合网络中的标准 RepC3 模块替换为 EMA-Attention C3 模块，该模块集成轻量级多头注意力与门控融合机制，有效建模空间和通道维度的长程依赖关系。这些改进模块被部署在特征融合流程的 P3~P5 多个阶段，从而实现更高效的多尺度表征学习。值得注意的是，所有增强措施在保持原有模型轻量化和实时性特点的同时，显著提升了检测精度。完整架构如图 1 所示。

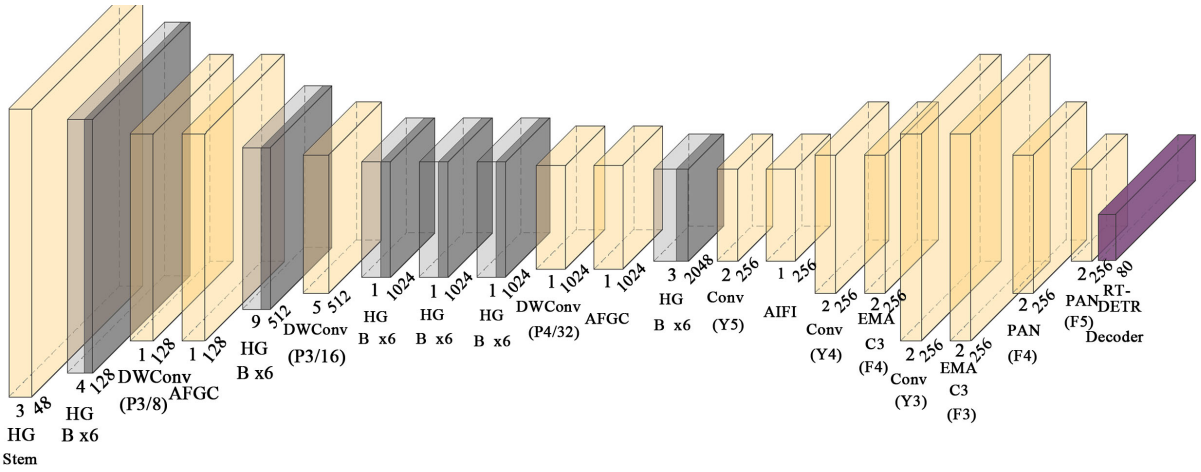



**Figure 1.** Improved model architecture. This model represents the overall architecture of the enhanced RT-DETR: the backbone network adopts a hybrid convolutional structure and incorporates the AFGC module to enhance frequency-aware global contextual information; the neck network integrates the EMA-Attention C3 module within the PAN structure to achieve multi-scale feature fusion; the final detection is performed end-to-end by the RT-DETR decoder

**图 1.** 改进后的模型架构。该模型为改进版 RT-DETR 的整体架构：骨干网络采用混合卷积结构并嵌入 AFGC 模块，以增强频率感知的全局上下文信息；颈部网络在 PAN 结构中集成 EMA-Attention C3 模块，实现多尺度特征融合；最终检测由 RT-DETR 解码器以端到端方式完成

### 3.1. AFGC Attention

AFGC (Adaptive Fine-Grained Channel，自适应细粒度通道)注意力模块是一个轻量且高效的机制，用于更精确地建模通道间的长程依赖。与传统的通道注意力不同，AFGC 在全局池化特征上引入可学习的一维卷积，以自适应地为不同通道分配重要性。该卷积操作使网络能够动态捕获跨通道的细粒度上下文交互，从而提升特征判别力。

在我们的实现中，AFGC 模块首先对输入进行全局平均池化以获得按通道的描述子；随后将其送入 1D 卷积层以提取局部模式，并与另一条通过投影得到的通道注意力分支进行融合，融合方式为可学习的门控混合函数。最终的注意力权重经由 sigmoid 激活得到，并用于重加权原始输入。其计算流程见公式(1)：

$$\text{Output} = X \cdot \sigma\left(\text{Conv1D}\left(\text{Mix}\left(f_1, f_2\right)\right)\right) \tag{1}$$

其中 $f_1$ 和 $f_2$ 分别表示通过卷积运算和矩阵交互得到的两个通道特征表示。最终加权输出结果由门控混合机制计算生成，并用于对原始输入特征图 X 进行重标定。

AFGC 模块部署于 RT-DETR 主干网络的中高层特征提取阶段，用于增强中尺度和高语义层特征的通道注意力响应能力。在 RT-DETR 中，Backbone 提供多尺度特征用于后续融合，AFGC 的引入可在每个尺度内部实现更精准的通道选择，从而提升检测目标的特征可分性。

AFGC 注意力模块的优势在于其动态通道加权机制，该机制使模型能够捕捉跨通道的细粒度上下文交互。这种设计显著提升了模型在复杂交通场景下检测小目标与遮挡物体的性能。此外，模块的轻量化

设计确保了其在实时检测任务中的高效性。

综上所述，AFGC 注意力模块通过融合全局平均池化、一维卷积与门控混合机制，显著增强了特征图的表达能力。该模块在保持计算效率的同时，有效提升了特征判别力——特别是在处理多尺度目标与复杂交通环境时表现突出，为实时目标检测任务提供了显著优势。

### 3.2. EMA-Attention C3

为增强 RT-DETR 原有的特征聚合结构，我们设计了一种新型轻量模块 EMA-Attention C3，用于替代 FPN/PAN 颈部网络中的传统 RepC3 模块。该模块受分组交互与多尺度上下文建模思想启发，引入了跨维度注意力机制。

其核心组件 EMA-Attention 单元将通道分组，并在空间与通道维度上实施并行注意力计算。该单元通过水平与垂直池化捕获方向感知的上下文信息，结合门控激励与 softmax 引导的重加权机制，显著增强全局与局部特征的交互效能。辅以分组归一化与基于卷积的特征优化管道，进一步提升了特征判别能力。

EMA-Attention C3 模块采用双路径结构设计：其中一条路径通过堆叠的 EMA-Attention 单元进行特征处理，另一条路径则保留残差连接以实现快捷传播。这种结构在不过度增加计算开销的前提下，既改善了信息流动效率，又实现了动态特征重校准功能。

EMA-Attention C3 模块的整体计算过程可抽象为以下公式(2)：

$$\text{Output} = \text{Conv}_3\left(\sum_{i=1}^{n}\text{EMA}\left(\text{Conv}_1\left(x\right)\right)\right) + \text{Conv}_2\left(x\right) \tag{2}$$

其中 $\text{Conv}_1$， $\text{Conv}_2$ 和 $\text{Conv}_3$ 为逐点卷积操作，EMA 表示在通道分组内执行的注意力运算。

EMA-Attention C3 被部署于 RT-DETR 的多尺度特征融合路径中，用以替代原有的 RepC3 模块。在原始 RT-DETR 架构中，FPN/PAN 负责自底向上融合不同分辨率的语义信息，但在面对复杂场景中尺度不均与长距离依赖建模能力不足的局限时，该结构存在感受野不充分的问题。引入 EMA-Attention C3 后，不仅增强了尺度间的上下文感知，还通过横纵维度的方向感知池化补全了局部与全局特征的耦合能力，从而优化了模型的检测表现。

## 4. 实验

### 4.1. 数据集

本研究使用的数据集为 Udacity 自动驾驶汽车数据集(由 Roboflow 提供)，专为自动驾驶感知任务设计。该数据集包含 15,000 张图像，共计 97,942 个标注边界框，涵盖车辆、行人、交通信号灯、交通标志、骑行等 11 类常见交通相关目标。其中约 1,720 张图像为负样本(即不包含可检测目标)，此类样本有助于提升模型在无目标场景中的鲁棒性。

原始图像分辨率为 1920 × 1200。Roboflow 同时提供了下采样版本(512 × 512)，该版本兼容 YOLO、SSD、Mask R-CNN 和 MobileNet 等主流目标检测模型。本实验选用固定小规模版本——这是一个经过预划分的轻量化子集，支持快速训练与结构对比，被广泛用于模型验证任务。

所有标注均经过 Roboflow 团队人工校验，确保标注精度。数据集采用 YOLO 标注格式进行组织，按照 8:1:1 的比例划分为训练集、验证集和测试集。训练过程中采用了标准数据增强策略，包括随机水平翻转、尺度抖动和亮度扰动。

### 4.2. 实验结果

为评估所提改进方法的有效性，我们基于 Udacity 自动驾驶汽车数据集开展了定量与定性结果分析。

图 2 展示了各类指标在 100 个训练周期内的学习曲线。损失函数(包括 GIoU 损失、分类损失和 L1 损失)持续下降，表明模型收敛稳定。同时，精确率、召回率与 mAP 等精度指标均呈现稳步上升趋势。最终取得的 mAP@0.5 = 76.8%与 mAP@50-95 = 39.2%结果，体现了模型在定位精度与分类质量方面的良好平衡。

如图 3 所示，精确率－召回率曲线反映了模型在各类别上的检测性能。改进后的模型在多数类别中表现出高精确度，AP 值最高达到 0.938。所有类别的综合 mAP@0.5 指标达到 76.8%，验证了模型在处理复杂多类别检测任务方面的强大能力。

图 4 展示了验证集中的代表性检测结果。该模型能准确识别多种目标类别，包括车辆以及不同状态(如红、绿、黄灯)的交通信号灯。在变化的照明条件与道路纹理下，模型仍保持稳定的预测性能，证明了其在真实驾驶环境中的鲁棒性。

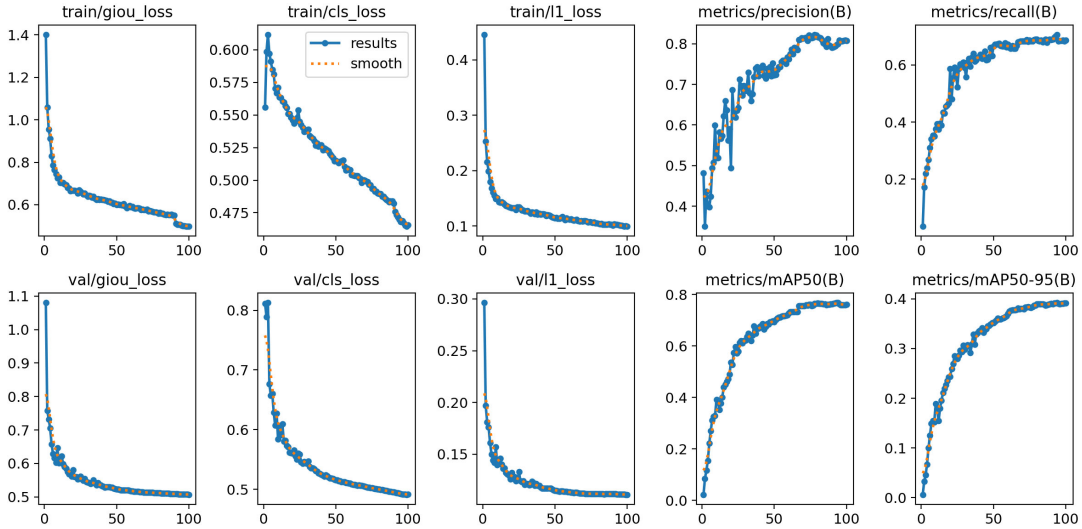



**Figure 2.** Loss and accuracy metrics training curves
**图 2.** 损失与精度指标训练曲线

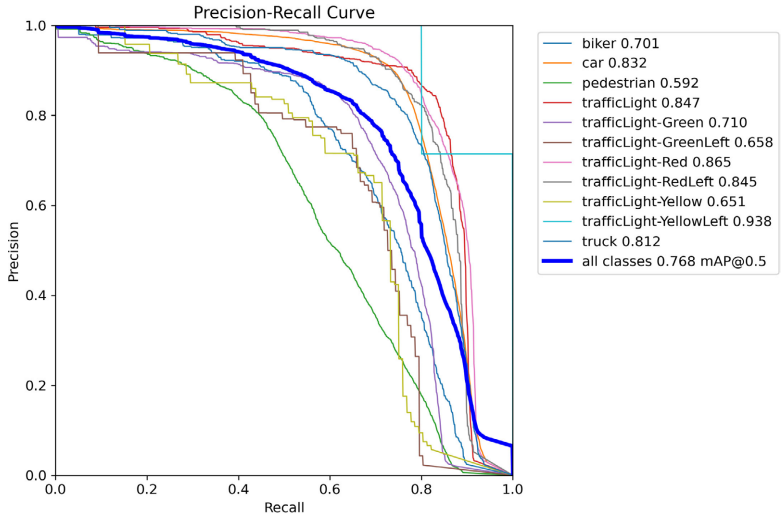



**Figure 3.** Precision-recall curves for each category
**图 3.** 各类别的精确率－召回率曲线

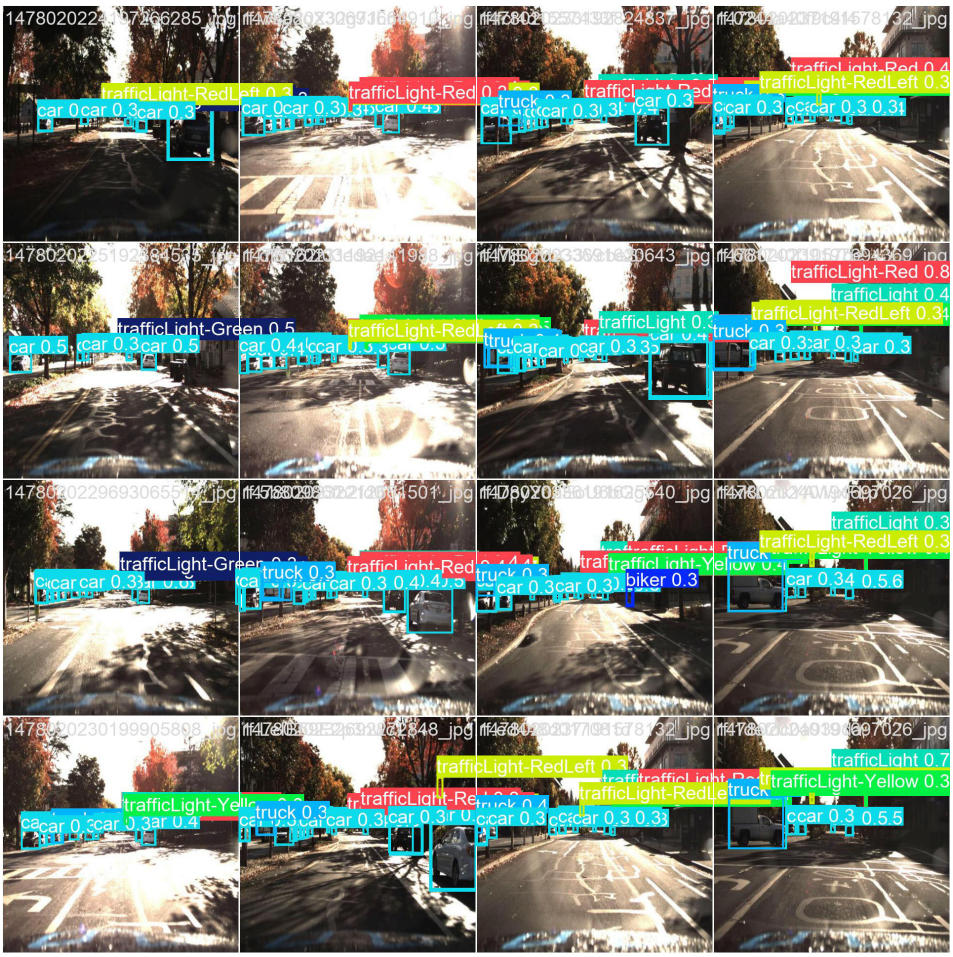
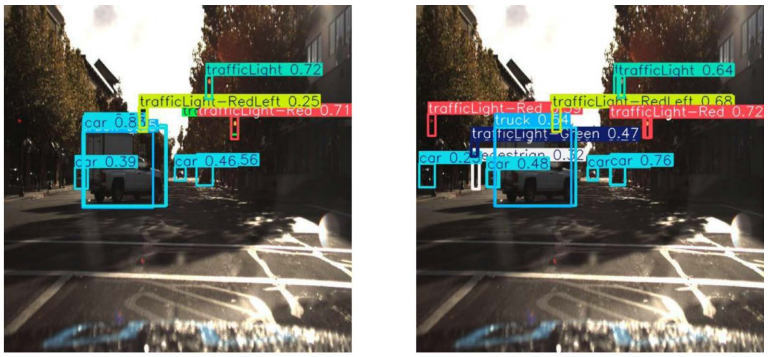


**Figure 4.** Visualization of qualitative detection results
**图 4.** 定性检测结果可视化

图 5 改进前后效果对比。左图为基线模型，存在对远处/小目标漏检、框偏移与置信度偏低等问题；右图为改进后的结果，交通灯与车辆目标检测更完整，误检减少，遮挡与光照变化下的小目标置信度明显提升，体现出在复杂道路场景中的整体鲁棒性与精度优势。



改进前　　改进后

**Figure 5.** Comparison of effects before and after improvement
**图 5.** 改进前后效果对比

#### 4.2.1. 对比实验

基于 Udacity 自动驾驶汽车数据集，我们对改进的 RT-DETR 模型与基准模型进行了对比研究。实验结果表明，改进模型在精确率、召回率和平均精度均值等指标上均有显著提升。当采用“l”尺度训练 100 个周期时，改进后的模型 mAP@0.5 指标提升 25.6%，mAP@50-95 指标提升 13%，其性能表现明显优于基准模型(具体数据见表 1)。

**Table 1.** Detection performance comparison between the original RT-DETR-l model and the improved RT-DETR-l model
**表 1.** 原始 RT-DETR-l 模型与改进 RT-DETR-l 模型检测性能对比

| Model | Evaluation Indicator (100 Epochs)/% | | | |
|---|---|---|---|---|
| | mAP50 | mAP50-95 | precision | recall |
| RT-DETR-l | 51.2 | 26.2 | 64.0 | 50.4 |
| 改进后 | 76.8 | 39.2 | 80.0 | 70.7 |

#### 4.2.2. 消融实验

为深入分析各模块的独立贡献，我们通过逐步引入 EMA-Attention C3 与 AFGC 模块进行了消融实验，并观测性能指标的变化。

实验结果表明(详见表 2)，两个模块均对精度提升产生正向影响，而组合使用更能获得协同增强效果，体现了其在特征增强方面的互补性。具体而言，EMA-Attention C3 对 mAP@0.5 指标提升贡献更为显著，而 AFGC 模块则对 mAP@50-95 指标增益作用更大，这凸显了其在多尺度目标建模方面的有效性。

**Table 2.** Performance comparison of models with different improvement methods
**表 2.** 不同改进方法的模型性能对比

| Configuration | Evaluation Indicator (100 Epochs)/% | |
|---|---|---|
| | mAP50 | mAP50-95 |
| Baseline RT-DETR | 51.2 | 26.2 |
| + EMA-Attention C3 | 61.8 | 31.2 |
| + AFGC | 57.4 | 33.5 |
| + EMA-Attention C3 + AFGC | 76.8 | 39.2 |

## 5. 结论

本文围绕电商末端无人配送的近场多目标检测需求，基于 RT-DETR 提出感知增强型改进模型。通过在骨干网络引入 AFGC 注意力以强化频域感知与全局语义建模，并在特征金字塔中采用 EMA-Attention C3 优化跨层融合与上下文建模，模型在不显著增加参数量与时延的前提下，显著提升了中小尺寸与遮挡目标的检测性能。

在产业层面，该模型有助于无人配送车辆在多样化、半结构化且强动态的作业环境中实现更稳健的环境感知。检测精度与稳定性的提升，可以增强系统对潜在风险的前瞻性预判与避障响应，降低远程接管与异常中断，进而提升履约成功率与时效。轻量高效的设计更契合边缘算力与能耗约束，有助于延长续航并降低整机成本，为大规模车队化运营提供可复制的感知底座。此外，所提模块具备良好的可插拔性，可与现有调度、定位与多传感器融合系统无缝衔接，有望为电商无人配送的规模化落地与长期运维带来可观的成本与安全收益[12]。

## 参考文献

[1] 无人配送在国内商业化的现状、挑战及建议[J]. 智能网联汽车, 2020(2): 60-67.

[2] 王世峰, 戴祥, 徐宁, 等. 无人驾驶汽车环境感知技术综述[J]. 长春理工大学学报(自然科学版), 2017, 40(1): 1-6.

[3] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S. (2020) End-to-End Object Detection with Transformers. In: *European Conference on Computer Vision*, Springer International Publishing, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13

[4] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/cvpr.2016.91

[5] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., *et al*. (2024) DETRs Beat YOLOs on Real-Time Object Detection. 2024 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 17-18 June 2024, 16965-16974. https://doi.org/10.1109/cvpr52733.2024.01605

[6] Sun, H., Wen, Y., Feng, H., Zheng, Y., Mei, Q., Ren, D., *et al*. (2024) Unsupervised Bidirectional Contrastive Reconstruction and Adaptive Fine-Grained Channel Attention Networks for Image Dehazing. *Neural Networks*, **176**, Article ID: 106314. https://doi.org/10.1016/j.neunet.2024.106314

[7] Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., *et al*. (2023) Efficient Multi-Scale Attention Module with Cross-Spatial Learning. 2023 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Rhodes Island, 4-10 June 2023, 1-5. https://doi.org/10.1109/icassp49357.2023.10096516

[8] 伍景琼, 陈子伟, 岑明睿, 等. 无人机配送模式及关键技术研究综述[J]. 交通信息与安全, 2025, 43(3): 112-127.

[9] Ren, S., He, K., Girshick, R., *et al*. (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149.

[10] Zhang, H., Li, F., Liu, S., *et al*. (2022) DINO: DETR with Improved Denoising Anchor Boxes for End-to-End Object Detection.

[11] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., *et al*. (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 10012-10022. https://doi.org/10.1109/iccv48922.2021.00986

[12] 伍景琼, 奠然, 字太升, 等. 无人机配送研究: 关于技术、效益、应用的系统综述[J/OL]. 交通运输系统工程与信息, 1-21. https://link.cnki.net/urlid/11.4520.u.20250905.0958.008, 2025-10-18.