

超图驱动的三位一体语义框架在零售商品检测中的应用

曹志钢¹, 刘立波^{1, 2}, 刘倩¹, 邓箴¹

(1. 宁夏大学信息工程学院, 宁夏 银川 750021; 2. 宁夏“东数西算”人工智能与信息安全重点实验室, 宁夏 银川 750021)

摘要: 零售自助结账系统的运行效率直接影响消费者购物体验。现有系统在提取小商品的细粒度关键特征方面存在困难, 同时在严重遮挡场景中空间推理能力不足, 限制了检测性能的提升。为解决上述问题, 本文在YOLOv12的基础上设计超图驱动的三位一体语义框架。该框架首先通过构建的跨层语义收集模块, 结合跨层特征注意力融合子模块(CFAF, Cross-Layer Feature Attention Fusion)与风车卷积(PSConv, Pinwheel-Shaped Convolution), 实现跨层级语义互补与深度网络中细粒度特征的高效保留, 从而增强对小商品细粒度特征的捕捉能力。其次, 引入语义精炼模块, 其核心混合动态感知的超图计算(HDP-HGC, Hybrid Dynamic Perception Hypergraph Computation)能有效提升了系统在严重遮挡场景下的识别性能。最后, 设计语义映射模块, 搭建了非相邻层语义交互通道, 三个模块协同构成完整的三位一体语义框架。我们在公开数据集RPC、FinVolution以及真实场景下自建数据集Mine进行了系统评估。实验结果表明, 本文算法在平均检测精度上达到了90.94%、92.04%和69.61%, 领先目前大多主流模型, 特别是在小商品识别和严重遮挡结算场景中, 表现出显著优势, 在保持较高识别准确率的同时有效控制了模型参数。此外, 本文方法将漏检和误检概率降低了约30%和50%, 进一步验证了其在复杂零售环境下的稳定性与实用价值。

关键词: 商品检测; 小商品; 遮挡; 超图; 三位一体语义框架

中图分类号: TP18

doi: 10.11959/j.issn.2096-3750.XXXX.

Application of Hypergraph-Driven "Trinity" Semantic Framework in Retail Commodity Detection

CAO ZHIGANG¹, LIU LIBO^{1,2}, LIU QIAN¹, DENG ZHEN¹

1. School of Information Engineering, Ningxia University, NingXia Yinchuan, 750021

2. Ningxia Key Laboratory of Artificial Intelligence and Information Security for Channelling Computing Resources from the East to the West, Ningxia University, NingXia Yinchuan, 750021

Abstract: The efficiency of retail checkout systems directly affects the shopping experience. Current systems have difficulty capturing key details of Small-sized merchandise and lack spatial reasoning in crowded scenes, which limits detection performance. To address the aforementioned issues, this paper develops a hypergraph-driven trinity semantic framework based on YOLOv12. First, the framework uses a cross-layer semantic gathering module. It combines the Cross-Layer Feature Attention Fusion (CFAF) submodule with Pinwheel-Shaped Convolution (PSConv). This process enables complementary semantics across different network layers. It also effectively preserves fine-grained features in the deep network. As a result, the model's ability to capture detailed features of small commodities is significantly improved. Next, a semantic refinement module is introduced. Its core component, Hybrid Dynamic Perception Hypergraph Computation

收稿日期: 2025-09-19; 修回日期: 2025-11-11

通信作者: 邓箴, dengzhen@nxu.edu.cn

基金项目: 2024年宁夏回族自治区重点研发计划(引才专项)(No. 2024BEH04026); 国家自然科学基金(No. 62262053); 宁夏科技创新领军人才项目(No. 2022GKLRLX03)

Foundation Items: The 2024 Key Research and Development Plan (Talent Introduction) Project of Ningxia (No. 2024BEH04026), The National Natural Science Foundation of China (No. 62262053), The National Natural Science Foundation of China (No. 62262053) (No. 2022GKLRLX03)

(HDP-HGC), effectively enhances the system's recognition performance in heavily occluded scenarios. Finally, a semantic mapping module is designed to establish semantic interaction pathways between non-adjacent layers. These three modules work collaboratively to form a complete trinity semantic framework. We conducted systematic evaluations on the public datasets RPC, FinVolution, and our self-built real-world dataset Mine. Experimental results show that our method achieved average detection accuracies of 90.94%, 92.04%, and 69.61%, respectively—surpassing most mainstream models. In particular, it demonstrates significant advantages in small-object recognition and heavily occluded checkout scenarios, maintaining high accuracy while effectively controlling model parameters. Furthermore, our approach reduces missed detections and false detections by approximately 30% and 50%, respectively, further verifying its stability and practical value in complex retail environments.

Key words: Retail merchandise detection, Small-sized merchandise, Occlusion, Hypergraph, Trinity semantic framework

0 引言

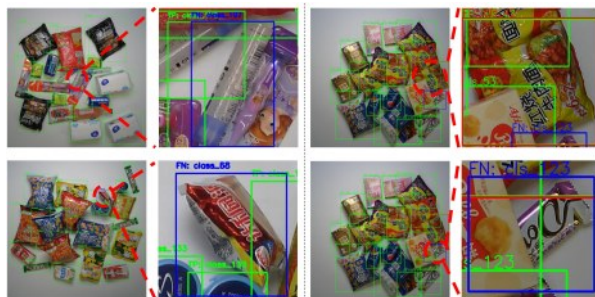
近年来, 基于深度学习的目标检测技术^{[1][2][3]}因其高效的自动检测能力, 在零售商品识别中得到了广泛应用。现有的目标检测模型可分为两类^[4]。一类是以 Faster RCNN^{[5][6]}为代表的两阶段目标检测, 此类模型虽具有较高的检测精度, 但存在计算成本较高的不足; 另一类为 YOLO^{[7][8]}为代表的单阶段目标检测, 其优势在于可实现快速的端到端推理适合实时部署, 但在检测精度上通常略低于两阶段模型。尽管这些模型在通用场景中已均取得显著成效, 但将其直接应用到零售自助结算实际场景中, 仍面临两个关键性挑战。

首先, 通用检测方法应用在零售自助结算场景中, 应对小尺寸商品检测问题尤为突出。许多零售商品(例如口香糖、小面包、胶水等)在图像中像素占比极低, 外感细节稀少且易受细节压缩影像, 常导致漏检。其次, 在零售商品实际结算过程中, 商品往往会相互覆盖、叠放或部分被其他物品遮挡, 常用的单物体检测策略缺乏对物体间空间拓扑关系的显式建模, 限制了模型的空间推理能力, 从而引发零售商品误检与漏检问题。这两种局限性共同制约了整个自助结算系统的准确识别率和运行可靠性(图1), 从而严重影响消费者的购物体验。

1 相关工作

目前, 国内外学者针对小物体检测, 主流方法通常从特征融合与特征提取两个方向入手。在特征融合层面, 其核心思路是将模型从不同层次或不同模态中提取到的特征, 通过特定策略组合

成更具鲁棒性的特征, 以强化对小物体的表征能力。例如, 孙^[9]等人提出一种多输入处理模



小尺寸商品结算漏检事例

遮挡结算场景漏检和误检事例

图1 小商品和遮挡结算场景漏检或误检事例。蓝色检测框代表漏检事例, 红色检测框代表误检事例

块, 通

过融合原始帧、光流图像与背景抑制图像, 在小目标检测性能上实现了突破。周^[10]等人引入ECA注意力机制, 让模型能够自适应地融合不同通道的信息, 突出对小目标检测有用的特征, 抑制无用或噪声特征。雷^[10]等人使用GSConv轻量化模块替换标准卷积, 结合尺度/空间/任务三重注意力机制, 增强小目标定位能力。在特征提取方面, 其核心目标是通过调整模型训练策略或引入新型技术方案, 优化模型从原始数据中提取有效特征的能力。例如, 吴^[12]等人采用正交扩张卷积以扩大感受野, 同时保持小目标的高频细节表达。冯^[13]等人引入轻量化模块SGConv来避免下采样中的细节信息丢失, 从而增强对小目标特征的表征能力。尽管上述研究在两方面均做出了重要贡献, 但仍存在一定局限。在特征融合方面, 现有方法多局限于相邻层级之间的融合, 未能充分挖掘深层与浅层特征之间的语义互补关系; 在特征提取方面, 深层网络对小目标细粒度特征的表达能力仍有不足, 具体表现在两方面: 一是方法多依赖复杂的卷积堆叠或注意力机制, 导致计算开销显著增大; 另一方面, 方法普遍

采用对称填充和固定感受野，容易忽略小目标的边缘信息，从而导致关键特征丢失，影响小目标检测精度。

遮挡严重的目标检测任务中，研究者通常从特征增强的角度出发，提出了一系列改进方法。例如，叶^[14]等人通过引入多维注意力机制以抑制背景干扰，同时设计前景增强模块（FEM）以强化前景特征表达，并结合通道混合策略优化上下文信息交互，从而更好地应对密集遮挡场景。罗^[15]等人设计

空间注意力驱动的对抗网络，通过三阶段竞争训练，动态生成强针对性的遮挡样本，迫使检测网络学习遮挡不变特征来提高遮挡目标的检测精度。然而，上述方法依赖于注意力机制或复杂的特征融合策略，不仅计算开销大，且难以部署在轻量级设备，更忽略了目标之间的空间拓扑关系（如相对位置等）。由于缺乏对空间结构信息的显式建模，在严重遮挡环境下，模型的空间推理能力有限，容易导致漏检与误检。

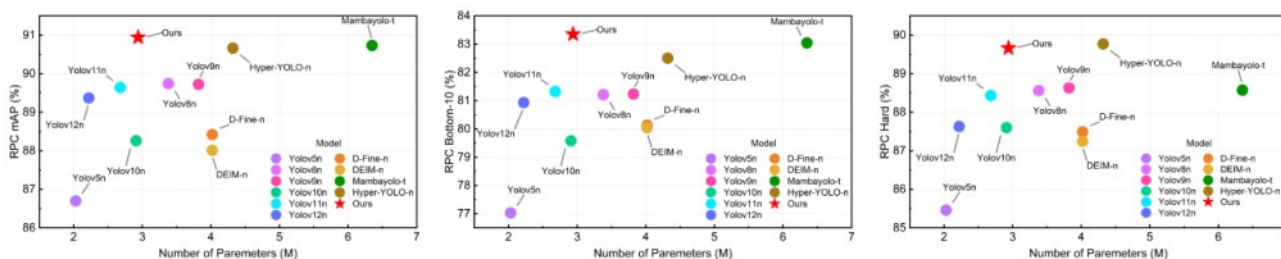


图2 在RPC数据集上与其他先进方法的对比。红星表示我们模型的性能，彩色圆圈代表其他主流检测模型。

针对上述研究的局限性，本文提出了一种基于超图驱动的三位一体语义框架，用于改进零售商品目标检测任务。图2展示了本文模型在RPC数据集上的检测效果，表明所提出框架在多种复杂结账场景中均表现出优异且稳健的检测性能。本文的主要贡献如下：

1) 设计了跨层语义收集模块。通过融合跨层特征注意力融合子模块和风车卷积，在实现深层与浅层语义互补的同时，高效保留深层网络中的特征细粒度信息，从而有效增强对小商品细粒度特征的捕捉能力，提升小商品的检测效果。

2) 搭建了语义精炼模块。其核心混合动态感知超图计算子模块，动态结合商品间的视觉特征和空间位置的双相似度量来增强对遮挡商品的感知分类能力，随后利用超图卷积进行高阶语义聚合，增强了在严重遮挡的结账场景中的检测鲁棒性。

3) 三位一体语义框架。该框架在前两模块的基础上引入一个语义映射模块，将增强后的语义信息再映射回各原始检测层，实现全局上下文之间语义交互。三个模块最终形成一个“收集-精炼-映射”的闭环语义增强机制，最终全面提升小商品和遮挡严重实际场景下的商品目标检测效果。

2 方法

YOLOv12^[16]是一种以注意力机制为核心的实

时目标检测架构，旨在克服传统YOLO模型严重依赖卷积神经网络（CNN）架构的局限性。其核心创新在于使用区域注意力（Area Attention）在保持较大感受野的同时，将计算复杂度降低到原来的四分之一。此外，YOLOv12还采用FlashAttention^[17]来缓解内存访问瓶颈，去除位置编码以简化模型结构。在现实零售结账场景中，小商品与严重遮挡情况下的目标检测精度较低。为此，我们在YOLOv12的基础上设计了一种由三位一体语义框架（图3）为核心的新方法。该框架通过跨层语义聚合模块、语义精炼模块与语义映射模块的协同作用，显著提升了对小尺寸商品及严重遮挡场景的检测精度，有效降低了结账过程中的漏检率与误检率。

2.1 跨层语义收集模块

现有的单阶段目标检测模型普遍采用特征金字塔网络（FPN^[18]）作为其主干网络（Backbone）架构。该框架通过多层下采样操作构建多尺度特征金字塔以处理不同大小的目标。然而，这种方式主要依赖于相邻层级特征间的简单交互，难以有效建立深层特征与浅层特征之间复杂且显式的语义互补。同时，多次下采样操作不可避免地压缩了有限的输入信息，导致小商品的细粒度特征在深层逐渐丢失。这些固有的局限性严重削弱了检测模型对小商品关键特征的识别能力。为解决这些局限性，我们

首先提出了CFAF模块，以加强深层与浅层特征之间的语义交互与信息融合。该模块首先通过上/下

采样操作完成两个不同层级的特征图分辨率对齐，得到初始跨层特征 $X_1 \in P^H \times W \times C_1$ 和 $X_2 \in P^H \times W \times C_2$ 。

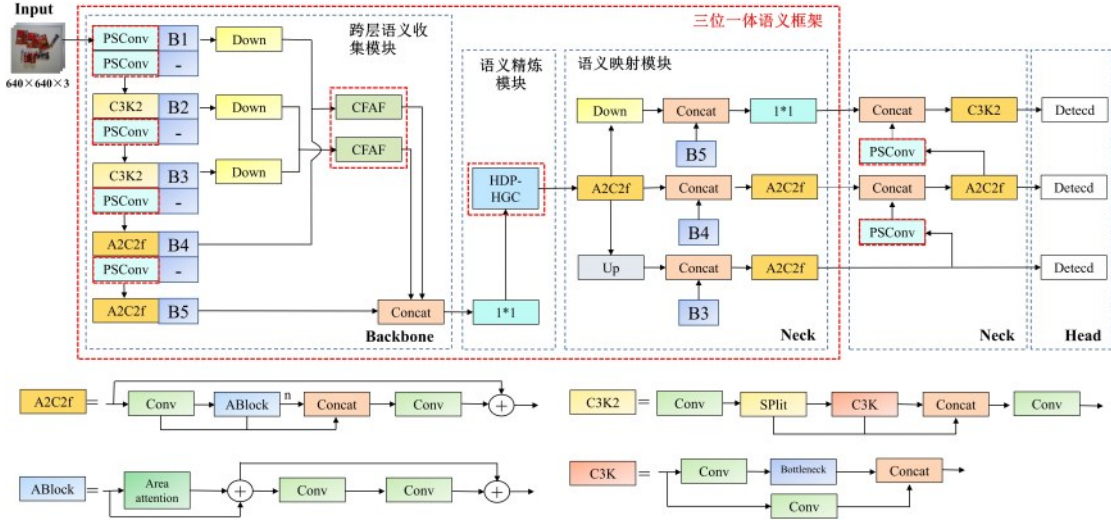


图3 模型的整体架构。红色框内的内容代表本文工作。

随后，将两者分别输入一组LGAttention^[19]模块（包含局部与全局注意力分支），以增强局部细节与全局上下文信息的提取能力。此外为了保留原始特征信息，两原始特征图还会按元素相加生成基础融合特征。接着，将局部增强特征、全局增强特征与基础融合特征拼接，并通过一个混合聚合网络（MixNet^[25]），进一步强化特征间的语义融合。经过CFAF模块处理，深层与浅层特征实现了有的互补与融合。该模块的整体计算过程如下。

$$\begin{cases}
 F_1 = \text{Cat}(\text{LGAttention}_{p=2}(X), \text{LGAttention}_{p=4}(X)) \\
 F_2 = \text{Cat}(\text{LGAttention}_{p=2}(Y), \text{LGAttention}_{p=4}(Y)) \\
 F_3 = \text{Conv}_3(\text{Conv}_1(X) + \text{Conv}_2(Y)) \\
 F_4 = \text{Cat}(F_1, F_2, F_3) \\
 \text{CAFA} = \text{MixNet}(F_4)
 \end{cases}
 \quad (1)$$

其中 $p=2$ 和 $p=4$ 代表局部和全局特征注意力，MixNet代表混合聚合网络模块。

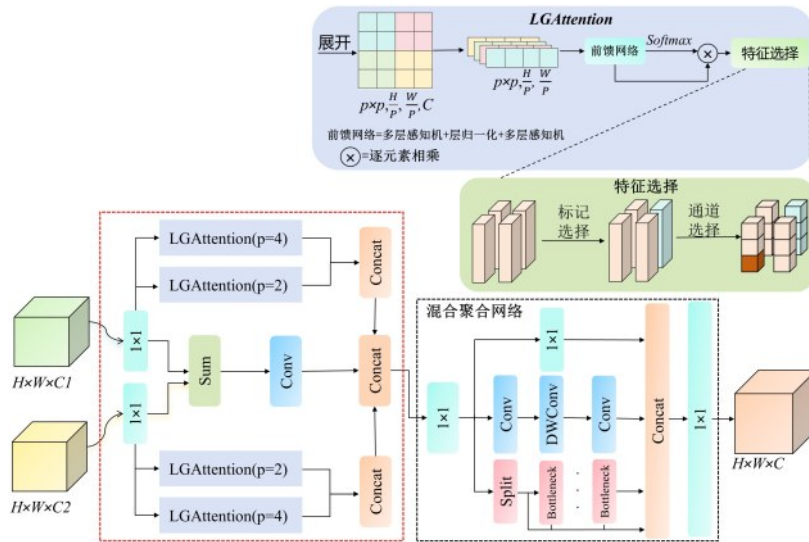


图4 CFAF模块示意图。红色框代表LGAttention增强特征模块，黑色框代表Mixnet模块的工作原理

LGAttention模块首先将输入特征 $X \in P^H \times W \times C$ 划分为 $P \times P$ 非重叠块，经平均池化压缩空间维度后

通过两层前馈网络生成中间特征 $F \in P^{\frac{H}{P} \times \frac{W}{P} \times D}$ ，显著降低计算复杂度的同时初步提取局部或全局信

息。随后引入可学习任务 ε ，通过余弦相似度计算生成空间掩码 $M \in [0,1]$ 进行特征调制，有效筛选关键空间标记以增强目标聚焦能力。为了强化跨通道特征交互，利用可学习上三角矩阵 $T \in P^{D \times D}$ 实现通道重组，最终经双线性插值恢复原分辨率，完成局部或全局特征增强。具体公式如下。

$$\begin{cases} F_{\text{mask}} = F \odot \text{Clamp}\left(\frac{F \cdot \varepsilon}{\|F\| \cdot \|\varepsilon\|}, 0, 1\right) \\ X_{\text{out}} = \text{Interpolate}(F_{\text{mask}} \cdot T, H, W) \cdot W_3 \end{cases} \quad (2)$$

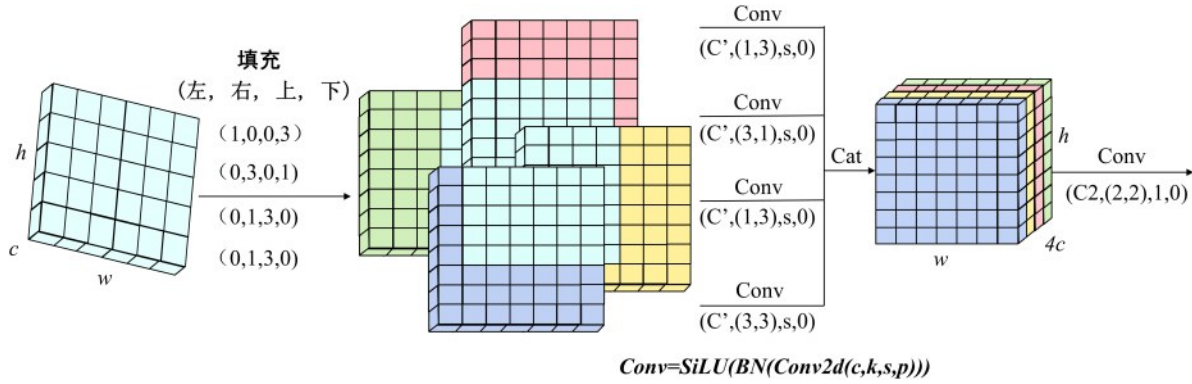


图5 PSConv模块示意图

为了降低小商品的细粒度特征在深层易丢失问题，我们又引入了PSConv^[21]（图5）替代标准卷积。首先对特征图 $X \in P^{h \times w \times c}$ 的左、右、上和下方进行不对称填充。这种填充引入了方向偏差，促使模型从不同方向捕捉上下文信息，从而显著增强模型的感受野。具体公式如下。

$$\begin{cases} X_1^{(h',w',c')} = \text{SiLU}\left(\text{BN}\left(X_{P(1,0,0,3)}^{(h,w,c)} \otimes W_1^{(1,3,c')}\right)\right) \\ X_2^{(h',w',c')} = \text{SiLU}\left(\text{BN}\left(X_{P(0,3,0,1)}^{(h,w,c)} \otimes W_2^{(3,1,c')}\right)\right) \\ X_3^{(h',w',c')} = \text{SiLU}\left(\text{BN}\left(X_{P(0,1,3,0)}^{(h,w,c)} \otimes W_3^{(1,3,c')}\right)\right) \\ X_4^{(h',w',c')} = \text{SiLU}\left(\text{BN}\left(X_{P(3,0,1,0)}^{(h,w,c)} \otimes W_4^{(3,1,c')}\right)\right) \end{cases} \quad (3)$$

其中 $P(1,0,0,3)$ 表示左、右、上、下方向的填充像素数量。 $\otimes$ 表示卷积算子， $W_1^{(1,3,c')}$ 表示第一个分支的卷积和参数，并且 $c' = \frac{c}{4}$ 表示分组通道数。

然后，为了增强特征的多样性，在交错卷积处理之后，我们沿着通道维度拼接特征结果，并通过一个 2×2 卷积层对拼接后的特征向量进行特征压缩，最后利用归一化和 SiLU 激活函数处理，保持训练的稳定性。该过程如式如下。

$$X' = \text{Cat}(X_1^{(h',w',c')}, \dots, X_4^{(h',w',c')}) \quad (4)$$

其中 $\odot$ 表示逐元素相乘，Interpolate 表示插值上采样操作。

混合聚合网络由三个并行的卷积分支组成。第一个分支使用 1×1 卷积在通道维度上校准特征。第二个分支使用深度可分离卷积，在保持较低参数数量的同时，扩展空间感受野。第三个分支结合了 C2f^[20] 结构，以促进多尺度特征融合。这些分支沿着通道维度进行拼接，使网络能够有效地捕捉通道依赖关系、空间细节和多尺度语义表示。

$$X_{\text{out}} = \text{SiLU}\left(\text{BN}\left(X' \otimes W^{(2,2,c_2)}\right)\right) \quad (5)$$

CFAF 和 PSConv 作为跨层语义收集模块的核心模块，两者协同工作，实现了小目标特征的层级互补与融合，有效缓解小目标细粒度特征损失，显著提升了对小商品的检测性能。

2.2 语义精炼模块

零售自助结账场景中，商品的高密度、复杂堆叠模式易引发遮挡问题，而传统的卷积网络无法有效地对遮挡商品之间的高阶关系进行建模。

针对商品遮挡问题，我们提出了一个语义精炼模块。其核心 HDP-HGC（图6）创新地融合商品外观特征与空间位置关系的双重模式，来增强对遮挡商品的感知分类能力。具体来说，首先对输入特征执行特征归一化和空间归一化操作，分别生成特征序列 X_{fea} 和 X_{spa} 。对特征序列 X_{fea} ，使用余弦相似度来生成特征相似度 S_{fea} ；对空间序列 X_{spa} ，我们先计算其坐标点之间的欧几里得距离，然后通过高斯核变换得到空间相似度 S_{spa} 。两者之间我们采用一种由自适应平均池化(AdaptiveAvgPool)、全连接层和 Sigmoid 激活函数构成的动态网络，实时生成融合权重 α ，并加权得到混合相似度特征 S_{mix} 。具体

公式如下。

$$\begin{cases} S_{fea} = X_{fea} \cdot X_{fea}^T \\ D_{spa} = \|X_{spa1} - X_{spa1}\|_2 \\ S_{spa} = e^{-\frac{D_{spa}^2}{2\sigma^2}} \\ S_{mix} = \alpha S_{fea} + (1 - \alpha) S_{spa} \end{cases}, \quad (6)$$

其中 D_{spa} 表示坐标归一化后节点之间的欧几里得距离, σ 在空间相似性计算中控制高斯核函数的衰减率, 并且 $\alpha \in (0, 1)$ 表示一个权重。

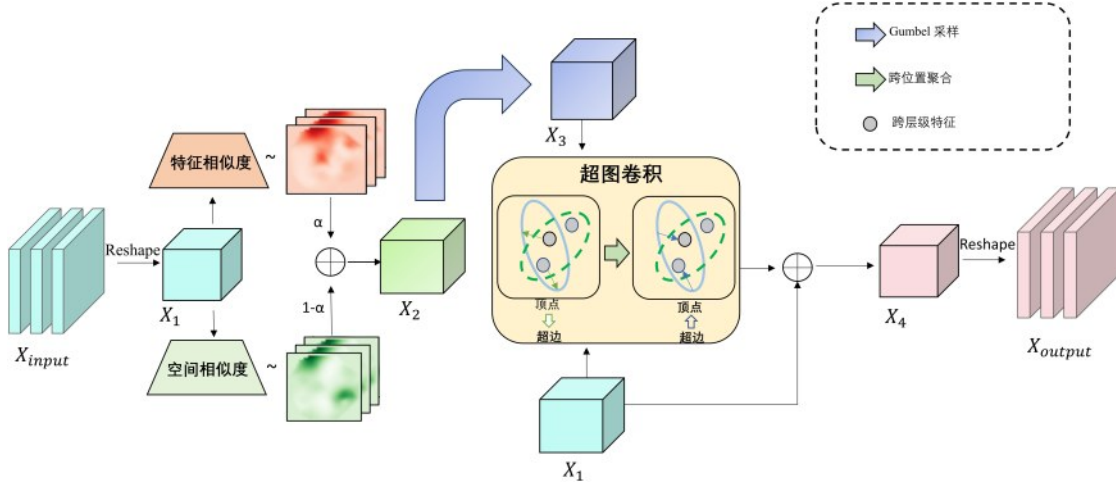


图6 HDP-HGC 模块结构示意图

传统的超图计算通常依赖于人工设计的启发式规则或预定义的相似度阈值来构建超图结构, 这种方法往往难以自适应地捕获数据中复杂的高阶关系, 且对参数设置敏感。受^[22]启发, 我们先对输入的混合相似度特征进行归一化处理, 并将其转换为概率分布。随后, 将该分布转换到对数几率空间, 以提高数值稳定性。接着采用重参数化技巧, 通过向对数几率表示中注入独立采样的 Gumbel 噪声, 得到一个松弛且可微的分布^[23]。这种高效超图构建过程, 可有效地将离散的超边采样问题转化为一个连续且可微的优化任务, 使得梯度能够通过噪声注入进行传播。具体公式如下。

$$H = \sigma'(\log(\text{logit}(S_{mix})) + G), \quad (7)$$

其中 $\text{logit}(S_{mix}) = \log \frac{S_{mix}}{1 - S_{mix}}$ 表示对数几率函数, σ' 是 sigmoid 激活函数, $G \sim \text{Gumbel}(0, 1)$ 从独立同分布的 Gumbel 分布中采样得到。

进一步的, 通过超图卷积和残差连接生成超图特征, 然后经过批量归一化和 Sigmoid 激活函数处理, 以生成具有空间拓扑关系的高阶语义特征。该过程公式如下。

$$X' = \text{HyperConv}(X_{flat}, H) + X_{flat}, \quad (8)$$

$$X_{HAD-HGC} = \text{SiLU}(\text{BN}(X')), \quad (9)$$

其中 HyperConv 可以看作是一个两阶段的信息传递过程, 即节点到超边和超边到节点的聚合, 分别表示为 $h_j = f_{v \rightarrow e}(\{x_k\})$ 和 $x_i = f_{e \rightarrow v}(\{h_k\})$ ^[24]。这两个阶段都可看作是两种均值聚合函数, 即在第一阶段的信息传递中, 每个超边内的节点 $\{x_k\}$ 通过 $f_{v \rightarrow e}$ 形成超边 h_j , 而在第二阶段, $f_{e \rightarrow v}$ 则用于更新节点表示 x_i 。

总体而言, 语义精炼模块通过引入特征与空间的双重相似性度量, 并采用基于 Gumbel 噪声的可微超图构建方法, 实现了对遮挡商品间高阶关系的自适应建模。该模块增强了模型在遮挡结账场景下对商品的感知与分类能力, 为零售自助结账系统中的商品识别问题提供了有效的解决方案。

2.3 三位一体语义框架

传统目标检测模型在特征提取与融合过程中普遍存在语义割裂问题。以 YOLO 系列为代表的单阶段检测器为例, 其采用的 FPN-PAN 结构仅通过简单的下采样与上采样操作实现相邻层级之间的语义交互, 却忽视了非相邻层之间的长距离语义关联, 导致空间位置上的高阶语义一致性受到破坏。因此, 模型难以有效捕捉远距离目标之间的语义依赖

关系，这一局限性严重制约了模型在多尺度目标感知中的统一表征能力。针对当前零售结账场景中中小商品检测困难以及严重遮挡条件下商品识别能力不足的双重挑战，本小节在2.1节所述跨层语义收集

模块与2.2节的语义精炼模块基础上，进一步引入语义映射模块。这三个模块协同工作，共同构成一个基于超图的三位一体语义框架（图7）。

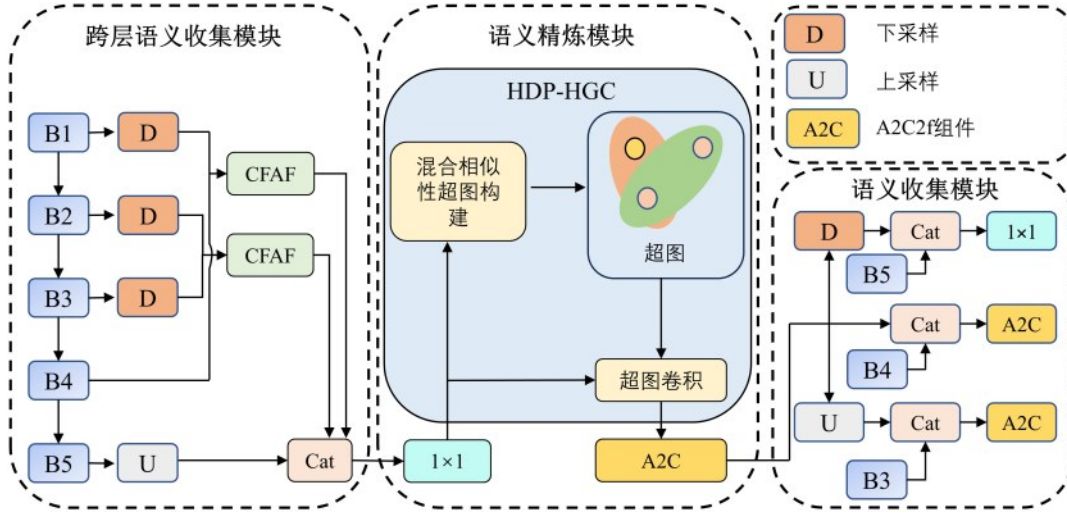


图7 三位一体语义框架结构示意图

语义映射模块是在基线检测模型的颈部网络进行改进。具体而言，采用上采样与下采样操作协同调整特征图分辨率，并通过跨层连接实现各尺度特征与原始层级的信息融合^[25]。该设计使每个空间位置在保留局部细节感知能力的同时，充分融合全局上下文语义线索，形成空间-语义协同增强的特征表示。该过程公式如下。

$$N_3, N_4, N_5 \xleftarrow{\text{mapping}} \phi(X_1, B_3), \phi(X_2, B_4), \phi(X_3, B_5), \quad (10)$$

其中 ϕ 表示特征融合函数，该函数执行逐元素求和或通道拼接，这一操作是通过调整 X_i 的分辨率大小，使其与目标特征图的维度相匹配来实现的。 B_3, B_4 和 B_5 表示模型主干部分中的各原始检测层。 N_3, N_4, N_5 表示网络颈部的三个输出，分别对应小型、中型和大型目标的检测。

纵观三位一体整体框架，我们可以看出，零售商品结账图像首先通过跨层语义收集模块实现多尺度特征的层次互补与融合。其次由语义精炼模块的超图建模机制，在特征空间内建立高阶语义关联并强化空间推理能力。最后将精炼后的特征映射各原始检测层，实现全局上下文之间的语义交互。三个模型协同工作，最终形成一个端到端的统一优化框架。最终实现对零售结账场景中中小商品检测困难与严重遮挡条件下商品识别能力不足两大核心挑战的协同优化，显著提升了模型在复杂实际场景下的综

合性能。该过程公式如下。

$$\begin{cases} X_{\text{Collect}} = \{B_5, \text{CFAF}(B_4, B_1), \text{CFAF}(B_2, B_3)\} \\ X_{\text{Refinement}} = f(X_{\text{HAD-HGC}}(X_{\text{Collect}})) \\ N_3, N_4, N_5 \xleftarrow{\text{mapping}} \phi(X_{\text{Refinement}}, B_i) \end{cases}, \quad (11)$$

其中 $\{B_5, \text{CFAF}(B_4, B_1), \text{CFAF}(B_2, B_3)\}$ 表示基础特征图与CFAF增强后的特征图， i 表示原始层数3,4,5。

3 实验与结果分析

3.1 数据集

本实验主要采用了三个数据集，其一为南京旷视研究院发布的大规模零售商品数据集RPC^[26]。该数据集包含83,739张图像，涵盖200个细粒度商品类别，其中包括53,739张样本图像和30,000张结账图像。结账图像是通过将商品进行多角度摆放随机组合的。这些图像分为简单、中等、困难三个难度等级，旨在模拟真实的结账场景。简单等级的图像包含3-5种商品类型，数量为3-10个，摆放相对分散，旋转变形轻微，遮挡较少。中等等级包含5-8种商品类型，数量为10-15个，摆放相对集中，旋转变形中等，遮挡程度中等。困难等级包含8-10种商品类型，数量为15-20个，摆放非常集中，旋转变形严重，遮挡明显。该数据集按8:1:1的比例随

机划分为训练集、验证集和测试集，每个难度等级在这些子集里均匀分布。另外，为了展示本文模型的泛化能力，我们还从第六届“FinVolution Cup”图像算法竞赛中选取了另一个数据集，即智能零售柜商品识别数据集（我们在后续实验部分统一称FinVolution）。该数据集包含 5,422 张图像，涉及 113 个商品类别，划分为 3,796 张训练图像、1,084 张验证图像和 542 张测试图像。我们对每张图像的边界框数量进行了分析，将包含 10 个以上边界框的图像标记为困难等级，这类图像共有 934 张，占数据集总数的 17.22%。为进一步验证所提模型在真实零售环境下的泛化性和鲁棒性，我们自主构建了实拍 Mine 数据集。该数据集的图像均严格模拟真实零售结算场景采集图像，共包含 31 个种类零售商品的 900 张高质量图像。数据集涵盖了不同商品数量、不同商品像素占比和不同遮挡程度的典型案例，可基本覆盖零售商品结算过程中的各类实际检测需求。借鉴上述 RPC 数据集划分标准，Mine 数据集的难度等级划分参照 RPC 标准，也分为简单、中等、困难三个等级，以确保评估的全面性与可比性。具体展示如图 8 所示。

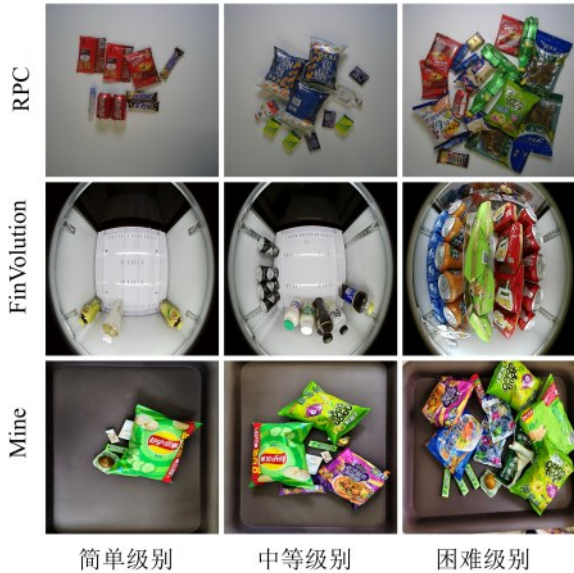


图 8 RPC、FinVolution 和 Mine 数据集

3.2 实验设置与评估指标

本文的实验设置如下。所有实验均在 Ubuntu 22.04 操作系统上进行，基于使用开源深度学习框架 PyTorch2.2.2 构建的网络模型。硬件配置包括 NVIDIA GeForce RTX 3090GPU 和 Xeon Silver 4210R CPU，软件环境由 Python3.10.14 和

CUDA12.1 组成。在训练过程中，输入图像大小统一设置为 640×640 像素，批量大小设置为 64。采用带动量的随机梯度下降（SGD）优化器，学习率为 0.01，动量值为 0.937，权重衰减为 0.0005。

本实验使用六个关键指标来评估网络性能，分别是参数数量（Params）、计算成本（GFLOPs）、模型内存占用（Model size）、平均精度均值（mAP）、表现最差的十个类别的平均准确率（Bottom-10，我们观察到在我们的实验中，所有 Bottom-10 类别都是小商品，如口香糖、铅笔和巧克力。）以及在困难级别下的平均检测准确率（Hard）。该过程公式如下。

$$\left\{ \begin{array}{l} Precision = \frac{TP}{TP + FP} \\ Recall = \frac{TP}{TP + FN} \\ AP = \int_0^1 P(R) dR \\ mAP = \frac{\sum_{i=1}^N AP_i}{N} \end{array} \right. \quad (12)$$

3.3 消融实验

为了评估本文模型的有效性，我们在 RPC 数据集上进行了消融研究，以考察每个模块的单独贡献。实验结果如表 1 所示。本文实验均在 3.2 节中规定的相同设置下进行，以确保实验一致性和公平性。

通过对表 1 实验数据分析可知，基线模型在引入跨层语义收集模块后，mAP 和 Hard 指标均显著提升，尤其是在 Bottom-10 指标上，从 80.93 提高至 82.78。这一实质性的提升得益于模块中 CFAF 与 PSConv 模块的协同作用。该模块不仅实现了对小商品特征的跨层互补与语义增强，还有效扩大了模型的感受野，从而缓解了细粒度特征在深层卷积过程中易被削弱的问题。当单独为基线模型增加语义精炼模块时，其 mAP 和 Bottom-10 指标分别提升了 0.35 和 0.78，特别是 Hard 上显著提高了 1.32。这是因为语义精炼模块中的 HDP-HGC 模块能够自适应融合遮挡结账场景下的视觉特征和空间位置相似度以构建动态超图。此类超图能对精准构建遮挡场景中的高阶空间和拓扑关系，提高模型的空间推理能力。进一步地，当语义精炼模块和跨层语义收集模块协同工作时，模型平均检测精度得到了显著提升

至90.8。这一结果充分验证了两个模块在语义增强与结构建模方面具有良好的互补性，并能在复杂场景下发挥最大的协同效果。

表1 不同策略下基线模型性能的比较

Model	Improvement Module			Params(M)	GFLOPs(G)	Model size(MB)	mAP(%)	Bottom-10(%)	Hard(%)
	a	b	c						
Baseline				2.2	5.8	4.6	89.37	80.93	87.63
1				2.35	6.8	5	90.12	82.78	88.57
2				2.4	6.9	5	89.72	81.71	88.95
3				2.5	7.5	5.3	90.80	83.04	89.33
Ours	-	-		2.94	8.7	6.2	90.94	83.35	89.66

其中, $\sqrt{}$ 表示使用该模块,-表示未使用该模块,a表示跨层级语义收集模块,b表示语义细化模块,c表示统一框架。因此,在“Ours”一栏中只需要勾选c。

最后在前两个模块的基础上进一步引入语义映射模块,三个模块协同工作共同构成完整的三位一体语义框架。实验结果表明,相较于基线模型,该模型的mAP、bottom-10和Hard指标均明显提升,具体提升幅度分别为1.57、2.42和2.03。这一突破的核心原因在于整体框架不仅集成了跨层语义收集模块的细粒度特征增强的优势和语义精炼模块的空间拓扑结构建模能力,更通过语义映射模块将双重优化的语义信息反向映射至原始检测层,有效弥补了原框架中缺少非相邻层之间语义交互的不足,使不同层级特征能够实现更全面的信息流通与整合,最终推动整体检测性能的系统性提升。

综上,三位一体语义框架能应对小商品检测和遮挡严重的复杂结账场景,为零售商品结账检测领域提供一种全新的端到端的解决方案。

3.4 对比实验

相同实验条件下,我们将本文方法与多种先进目标检测模型开展对比实验,其中包括YOLOv8和YOLOv12为代表的YOLO系列模型,还包含基于DETR^[27]的最新模型D-FINE和DEIM^[28]模型。此外,为适配资源受限边缘设备的部署要求,我们将YOLOv12的深度缩放因子从0.5降至0.33。具体对比结果如表2所示。

表2 主流检测模型的性能对比

Model	Params (M)	GFLOPs (G)	Model size (MB)	RPC			FinVolution			Mine		
				mAP (%)	Bottom- 10 (%)	Hard (%)	mAP (%)	Bottom- 10 (%)	Hard (%)	mAP	Bottom- 3(%)	Hard
Yolov5n	2.03	5	4.46	86.7	77.03	85.46	83.74	56.15	84.01	60.37	42.29	50.31
Yolov8n	3.38	9.2	6.7	89.74	81.21	88.56	91.21	77.72	91.11	67.23	50.18	57.92
Yolov9n	3.82	17	5.81	89.72	81.23	88.63	90.04	76.43	89.95	67.06	50.04	57.96
Yolov10n	2.91	9.4	5.5	88.26	79.58	87.60	90.15	74.3	90.02	66.33	48.77	56.94
Yolov11n	2.68	6.8	5.3	89.64	81.32	88.43	90.25	77.68	88.45	66.97	48.77	57.84
Yolov12n	2.22	5.8	4.6	89.37	80.93	87.63	90.21	54.53	89.6	66.44	48.89	57.48
Hyper-yolo-n	4.32	12.4	8.6	90.66	82.5	89.77	92.15	79.43	91.94	68.97	51.88	60.09
Mambayolo-t	6.35	15.3	37.1	90.73	83.04	88.47	92.55	81.03	92.34	68.74	51.34	59.84
D-Fine	4.02	7.57	48.2	88.4	80.13	87.49	90.81	77.61	90.75	67.58	48.94	58.28
DEIM-n	4.02	7	48.2	88	80.04	87.24	89.11	77.04	89.01	67.89	49.35	58.18
Ours	2.94	8.7	6.2	90.94	83.35	89.66	92.08	81.17	91.88	69.41	53.24	60.97

加粗值表示每个类别中的最佳结果

由表二可知,,本文模型在RPC数据集上的参数量和计算量略高于YOLOv11和YOLOv12等,但mAP指标显著优于这些方法。此外,在模拟小商

品检测场景的Bottom-10指标中,本文方法的mAP指标达到83.35,明显优于其他对比模型,充分说明其在小目标检测任务上具有较强的鲁棒性。在模

拟严重遮挡结账场景下的 Hard 指标上, 本文模型的 mAP 指标达到 89.66, 仅次于 Hyper-YOLO。但值得注意的是, Hyper-YOLO 的参数量、计算量和模型大小分别比本文模型高出 47%、43% 和 38%, 高额的参数量和复杂的计算量使其难以部署在资源受限的边缘设备中。总体而言, 本文模型在计算效率与检测性能之间实现了较好平衡, 尤其在小目标和高遮挡场景下表现优异。在 FinVolution 数据集上, 本文模型的 mAP 和 Hard 指标略低于 MambaYOLO 与 Hyper-YOLO, 但仍显著优于其他主流模型, 并在 Bottom-10 指标上超越了所有对比方法。在自建数据集 Mine 上, 本文提出的模型在三个检测指标上均优于对比模型。尤其在小目标占比高的 Bottom-3 和严重遮挡的 Hard 两个等级场景中, 本文模型表现出显著优势, 比基准模型提高了 8.8% 和 6.1%, 同时也均优于当前主流目标检测模型, 这些结果进一步验证了该模型在小商品检测及严重遮挡结账场景下的优势。综上所述, 本文模型在复杂零售结算环境中展现出良好的实际应用潜力, 能够有效提升真实场景下对小商品和遮挡严重目标的检测精度。

3.5 可视化实验

为了直观地验证本文模型的可行性和优越性, 我们在 RPC 数据集上进行了可视化实验。首先, 对有效感受野贡献分析^[29]方法进行可视化, 并与基线模型进行比较。在可视化结果中, 绿色阴影区域代表模型的有效感受野区域, 面积越大表明全局感知能力越强。颜色强度反映了逐像素的梯度贡献, 颜色越深对应模型提取物体关键特征的能力越强。如图 9 所示, 与基线模型相比, 本文模型在网络更深层展现出更大的感受野和更强的特征提取能力, 有助于在小商品检测中更准确地提取关键特征, 从而显著提升模型对小商品的检测性能。

其次, 在 RPC 数据集上进行了预测边界框的可视化实验。我们随机选择了代表两种不同难度级别的结账图像进行对比测试 (图 10)。基线模型存在明显缺陷, 在结账图像 1 和结账图像 2 中, 它未能检测到属于类别 189 和类别 123 的产品。而在结账图像 3 和结账图像 4 中, 它分别将正确检测到的类别 41 和类别 131 的物品误分类为类别 60 和类别 130。相比之下, 本文模型在这些图像中没有出现漏检或误分类的情况。在 Mine 数据集上的对比实

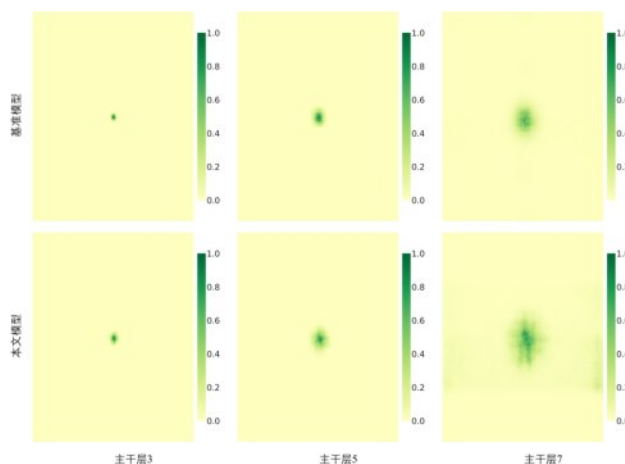


图9 模型感受野分析可视化对比

验进一步表明, 尽管基线模型与本文模型均存在一定程度的漏检与误检现象, 但本文模型在错误数量上显著更少, 表现出更优的检测稳定性。例如, 检测图三显示, 基线模型漏检了被严重遮挡的类别 1 牛奶和类别 15 软糖, 并将颜色相同的类别 8 旺仔牛奶误检为类别 19 罐头, 而在检测图四中, 其未能识别出类别为 21 的小商品口香糖。相比之下, 本文模型在这些困难场景中错误的漏检和误检框数量减少明显。这一结果进一步说明, 本文模型在面对小尺寸商品时具备更强的细粒度特征提取能力, 并在遮挡严重的复杂结算环境下中展现出更可靠的空间推理能力。

为了进行定量评估本文模型和基准模型的漏检和误检数量, 我们分别对 RPC 和 Mine 数据集的每张图片进行个体目标检测分析, 统计不同结算等级的漏检和误检数量。从图 11 可以看出, 在 RPC 数据集上, 相比基线模型的 139 次、179 次和 306 次漏检, 本文模型将这些数量降低至 94 次、122 次和 216 次, 分别降低了 32.37%、31.84% 和 30.07%。同样, 基线模型还记录到 191 次、250 次和 428 次误检, 而本文模型将这些数字显著减少到 92 次、128 次和 228 次, 相应地分别减少了 51.83%、48.80% 和 46.73%。在 Mine 数据集上同样观察到了本文模型在减少漏检与误检上的显著效果, 漏检和误检数相较于基准模型下降了约 30% 和 50%。这些结果充分证明了该模型在应对不同复杂度的零售结账场景时的优越性。

最后, 我们使用 Grad-CAM^[30]方法在 RPC 和 Mine 数据集上都进行了对比热图可视化分析 (图 12)。RPC 数据集上, 基线模型表现出弥散的热响

应，在简单、中等和困难的结账场景中都无法捕捉到关键特征。例如，在袋装零食场景中，热力图响应错误大多集中在包装边缘。相比之下，本文模型展现出了精确的注意力，热力图覆盖与目标语义紧密对齐。由图12可知，Mine数据集上，基线模型的热力图呈现出明显的注意力模糊或背景干扰问题，对密集、小尺寸或遮挡商品的响应不够明确。

相比之下，本文模型的热力图则表现出清晰的响应区域与更强的语义感知能力，能准确突出有效目标，显示出优异的实际场景适应能力。综上所述，热力图对比清晰地表明本文模型所具备的全局上下文建模能力，使其能够精准定位目标并有效抑制干扰，最终在各种复杂的结账场景中均展现出卓越的适应性与可靠性。

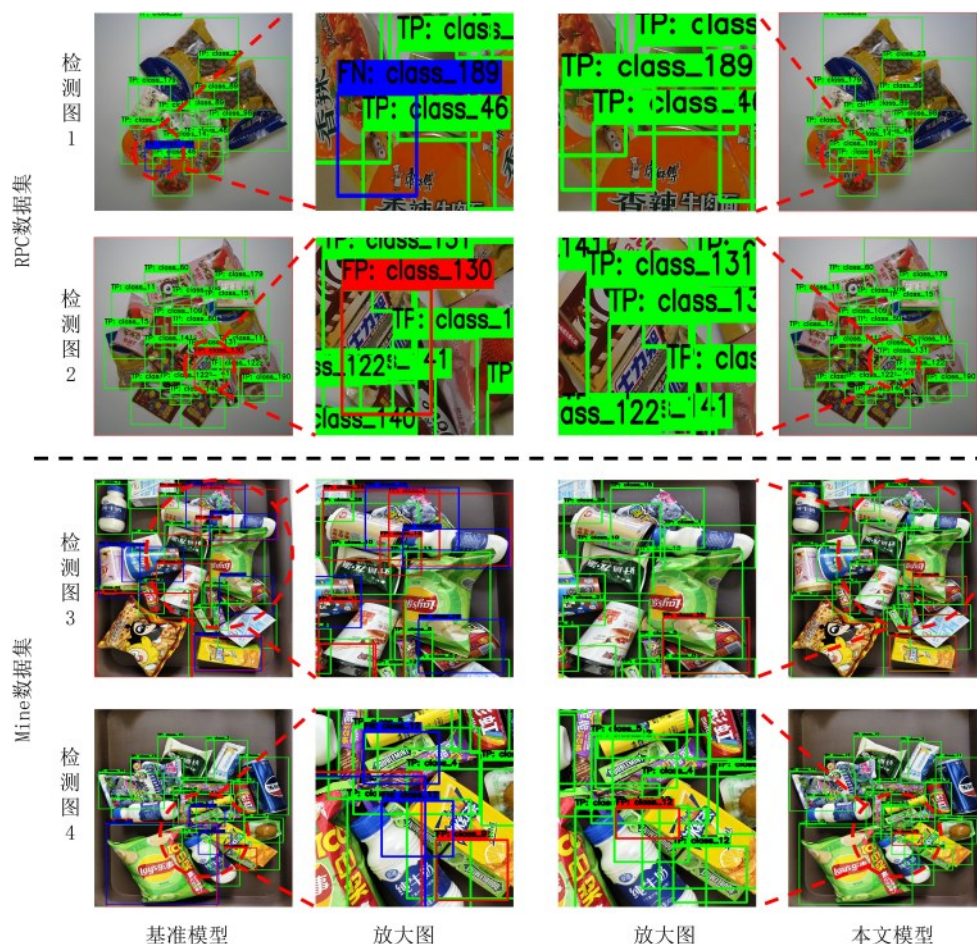


图10 RPC和Mine数据集检测效果可视化对比，绿色检测框表示正确检测到的商品，蓝色检测框表示漏检的商品，红色检测框表示误检的商品。

4 结论

本文提出了一种超图驱动的三位一体语义框架。该框架在YOLOv12的基础上，首先结合跨层特征注意力融合子模块和风车卷积以构建跨层语义收集模块，实现跨层级语义互补与深度网络中细粒度特征的高效保留，显著增强了模型对小商品细粒度特征的捕捉能力；其次，设计了基于混合动态感知超图计算的语义精炼模块，有效提升了模型对高阶关系的建模能力；最后，在上述两模块的基础上

增设语义映射模块，实现了非相邻层之间的语义交互。三个模块联动协同工作，共同构成了完整的三位一体语义框架。实验结果证明，本文提出的方法在保持轻量化的同时，使小商品和遮挡严重结账场景下的检测性能均优于大多最新检测模型。此外，与基线模型相比，漏检和误检概率分别降低30%和50%，进一步证明本文方法对小商品和遮挡严重商品检测的优越性。虽然本文所提模型已在检测精度与轻量化部署间实现较好平衡，但是该模型还有一定改进空间，未来在提高检测速度的同时进一步优

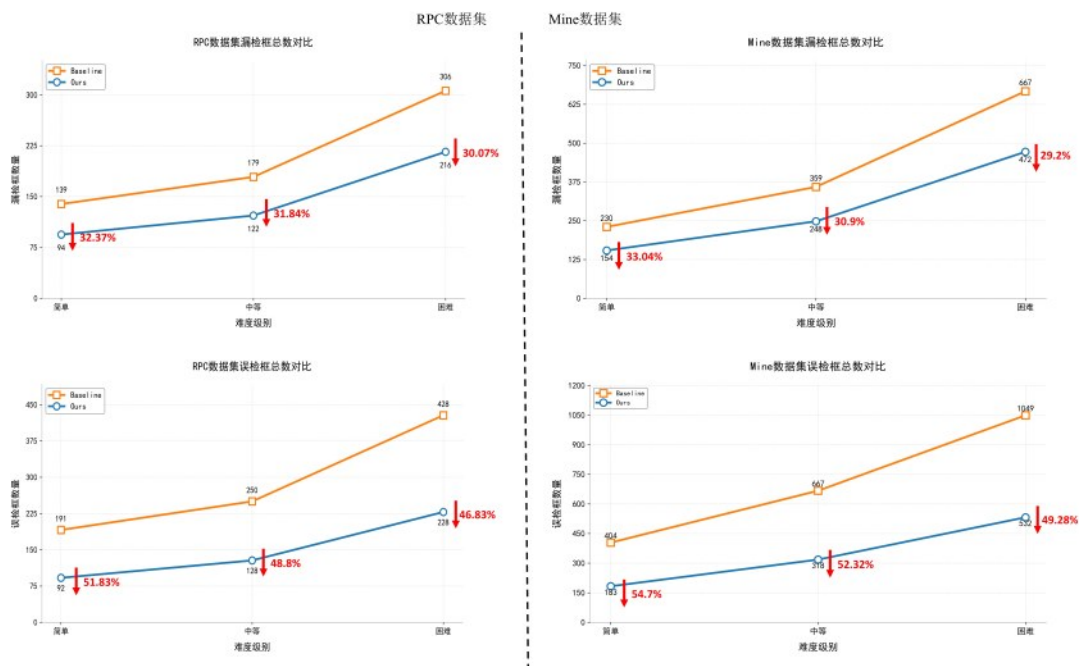


图 11 RPC和Mine数据集中三种级别漏检和误检数对比

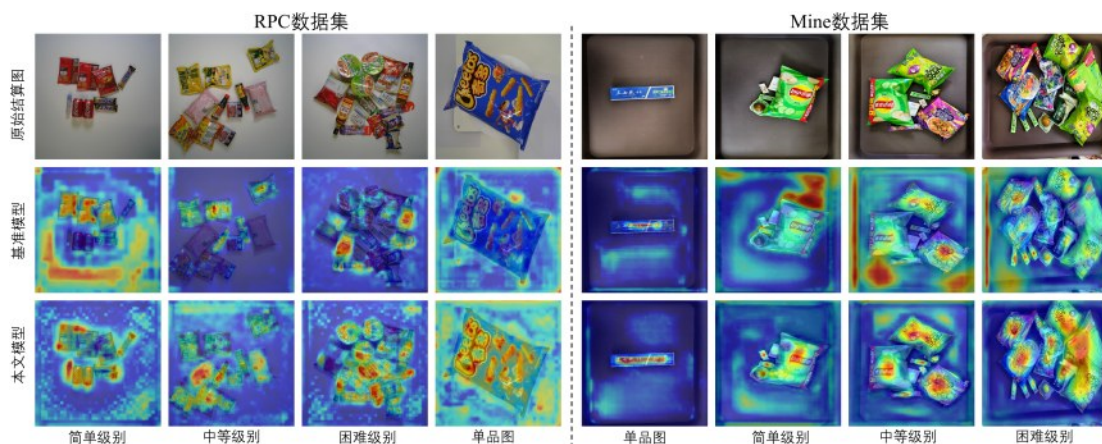


图 12 模型热力图可视化对比

化检测速度，最终为零售结账系统的智能化、高效化与场景化升级提供更坚实的技术支撑。

参考文献:

- [1] Janiesch C., Zschech P., Heinrich K. Machine Learning and Deep Learning[J]. Electronic Markets, 2021, 31(3): 685 - 695.
- [2] Zou Z. X., Chen K. Y., Shi Z. W., Guo Y. H., Ye J. P. Object Detection in 20 Years: A Survey[J]. Proceedings of the IEEE, 2023, 111(3): 257 - 276.
- [3] Wu P., Zhang Z., Peng X., Wang R. Deep Learning Solutions for Smart City Challenges in Urban Development[J]. Scientific Reports, 2024, 14(1): 5176.
- [4] 张静,农昌瑞,杨智勇.基于卷积神经网络的目标检测算法综述[J].兵器装备工程学报,2022,43(06):37-47.
- [5] 夏楠,袁海兵,吴俊.基于改进Faster RCNN的汽车零件密封圈装配检测研究[J].计算机应用与软件,2025,42(06):93-99.
- [6] 王伟辉,信泽阳,车清论,等.基于改进Faster RCNN模型的冬枣缺陷检测方法[J].农业工程学报,2024,40(22):283-289.
- [7] 吕晓华,魏铭辰,刘立波.基于位置可学习视觉中心机制的零售商品检测方法[J].物联网学报,2023,7(04):142-152.
- [8] 马超伟,张浩,马新明,等.基于改进YOLOv8的轻量化小麦病害检测方法[J].农业工程学报,2024,40(05):187-195.
- [9] Sun S., Mo B., Xu J., Li D., Zhao J., Han S. Multi-YOLOv8: An Infrared Moving Small Object Detection Model Based on YOLOv8 for Air Vehicle[J]. Neurocomputing, 2024, 588: 127685.
- [10] 周昶雯,宋强,张月.基于改进YOLOv8的凝视雷达小目标检测算法[J].信号处理,2025,41(05):853-866.
- [11] 雷帮军,余翱,吴正平,等.改进YOLOv8n的无人机航拍小目标检测算法[J].现代电子技术,2025,48(03):26-34.DOI:10.16652/j.

- issn.1004-373x.2025.03.005.
- [12] Wu S., Lu X., Guo C., Guo H. MV-YOLO: An Efficient Small Object Detection Framework Based on Mamba[J]. IEEE Transactions on Geoscience and Remote Sensing, 2025, 63: 5632814.
 - [13] 冯庆胜,付明雨,姚泽圆,等. 基于改进YOLOv8的轨道小尺度异物入侵算法研究[J]. 现代电子技术, 2025, 48(11): 174-179. DOI: 10.16652/j.issn.1004-373x.2025.11.027.
 - [14] Ye J., Zhang Y., Li P., Guo Z., Zeng S., Wei T. Real-Time Dense Small Object Detection Model for Floating-Litter Detection and Removal on Water Surfaces[J]. Marine Pollution Bulletin, 2025, 218: 118189.
 - [15] 罗偲,李凯扬,吴吉花,等. 基于对抗注意力机制的水下遮挡目标检测算法[J]. 计算机工程, 2024, 50(10): 313-321.
 - [16] Tian Y., Ye Q., Doermann D. YOLOv12: Attention-Centric Real-Time Object Detectors. arXiv preprint: arXiv:2502.12524, 2025.
 - [17] Dao T., Fu D., Ermon S., Rudra A., Re, C. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness[C]//Advances in Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc, 2022, 35: 16344 - 16359.
 - [18] Lin T. Y., Dollár P., Girshick R., He K., Hariharan B., Belongie, S. Feature Pyramid Networks for Object Detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE Computer Society, 2017: 2117-2125.
 - [19] Xu S., Zheng S., Xu W., Xu R., Wang C., Zhang J., Guo, L. HEF-Net: Hierarchical Context Fusion Network for Infrared Small Object Detection[C]//2024 IEEE International Conference on Multimedia and Expo (ICME). Singapore: IEEE, 2024: 1 - 6.
 - [20] Varghese R., Sambath, M. YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness[C]//2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). Chennai, India: IEEE, 2024: 1-6.
 - [21] Yang J., Liu S., Wu J., Su X., Hai N., Huang, X. Pinwheel-Shaped Convolution and Scale-Based Dynamic Loss for Infrared Small Target Detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania, USA: AAAI Press, 2025, 39: 9202 - 9210.
 - [22] Bai S., Zhang F., Torr P. H. Hypergraph Convolution and Hypergraph Attention[J]. Pattern Recognition, 2021, 110: 107637.
 - [23] Cai D., Song M., Sun C., Zhang B., Hong S., Li, H. Hypergraph Structure Learning for Hypergraph Neural Networks[C]//Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI-22). Vienna, Austria: IJCAI Press, 2022: 1923 - 1929.
 - [24] Chien E., Pan C., Peng J., Milenkovic, O. You Are AllSet: A Multi-set Function Framework for Hypergraph Neural Networks [C]//Proceedings of the Tenth International Conference on Learning Representations (ICLR 2022). Virtual: ICLR Organization, 2022.
 - [25] Feng Y., Huang J., Du S., Ying S., Yong J. H., Li Y., Gao Y. Hyper-YOLO: When Visual Object Detection Meets Hypergraph Computation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(4): 2388-2401.
 - [26] Wei X. S., Cui Q., Yang L., Wang P., Liu L. Q., Yang J. RPC: A Large-Scale and Fine-Grained Retail Product Checkout Dataset[J]. Science China Information Sciences, 2022, 65(9): 1 - 2.
 - [27] Carion N., Massa F., Synnaeve G., Usunier N., Kirillov A., Zagoruyko, S. End-to-End Object Detection with Transformers[C]//Proceedings of the European Conference on Computer Vision (ECCV). Virtual: Springer, 2020: 213 - 229.
 - [28] Huang S., Lu Z., Cun X., Yu Y., Zhou X., Shen, X. DEIM: DETR with Improved Matching for Fast Convergence [C]//Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR). Nashville, Tennessee, USA: IEEE, 2025: 15162 - 15171.
 - [29] Ding X., Zhang X., Han J., Ding G. Scaling Up Your Kernels to 31×31 : Revisiting Large Kernel Design in CNNs[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE Computer Society, 2022: 11963 - 11975.
 - [30] Narkhede J. (2024). Comparative evaluation of post-hoc explainability methods in AI: LIME, SHAP, and Grad-CAM[C]//2024 4th International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSSES), Guilin, China: IEEE, 2024: 826-830.