IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 August 2025, accepted 9 October 2025, date of publication 14 October 2025, date of current version 20 October 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3621105

RESEARCH ARTICLE

RetailDet: An Efficient Fusion Attention Network for Joint Product and Vacancy Identification in Smart Retail

BIDONG CHEN[1,2], (Member, IEEE), LINGUI LI[3], YAPENG WANG[1], (Member, IEEE), RUI PEDRO PAIVA[2], (Senior Member, IEEE), YUANDA LIN[4], XU YANG[1], (Member, IEEE), AND HAN ZHU[1,2], (Member, IEEE)

[1]Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR 999078, China
[2]Department of Informatics Engineering, University of Coimbra, 3004-516 Coimbra, Portugal
[3]School of Modern Information Industry, Guangzhou College of Commerce, Guangzhou, Guangdong 511363, China
[4]Whale TV Pte. (Singapore) Ltd., Fuzhou, Fujian 350001, China

Corresponding author: Yapeng Wang (yapengwang@mpu.edu.mo)

This work was supported in part by Macao Polytechnic University through Submission Control Code fca.6ae3.81a9.5 under Grant RP/FCA-04/2022; and in part by the Foundation for Science and Technology (FCT), I.P., within the scope of the research unit UID/00326 - Centre for Informatics and Systems of the University of Coimbra.

ABSTRACT Dense retail scenes pose unique challenges for both product detection and vacancy identification, where items are tightly packed on shelves and spaces between products carry critical inventory management implications. However, the existing detection frameworks, including state-of-the-art (SOTA) YOLO models, face limitations in feature fusion, particularly when integrating multi-level features and interpreting complex shelf arrangements. To address these challenges, we propose the RetailDet detection architecture and establish the RPV1K retail scene dataset. Our main contributions include: 1) A multi-attention module architecture that effectively fuses RGB and depth information for both product and vacancy detection; 2) A novel implicit gradient regulation (IGR) mechanism is proposed to address the redundant decision modules and feature modularization problems in traditional detection models. The proposed IGR mechanism dynamically regulates the gradient flow, optimizing the feature fusion path only during the training phase and not participating in calculations during the inference phase, thereby reducing computational overhead and enhancing model generalization performance; and 3) the RPV1K dataset, which is specifically tailored for retail detection tasks featuring product-vacancy co-occurrence scenarios. Experimental results show RetailDet significantly outperforms existing methods on the RPV1K dataset. Compared with YOLOv11-nano, RetailDet-nano achieves 10.6% mAP improvement while reducing inference latency by 17.2 ms. RetailDet-large achieves 85.2% precision with 20.79 M parameters. These innovations provide strong technical support for automatic shelf management and inventory replenishment in smart retail environments. Project resources can be obtained from https://github.com/bilychen88/RetailDet

INDEX TERMS Smart retail, object detection, feature fusion, attention mechanism, implicit gradient regulation, shelf management.

I. INTRODUCTION

The rapid advancement of deep learning has spurred continuous innovation in computer vision, providing robust support for emerging fields such as robotics and smart retail. In the autonomous new retail supermarket scenario, the robot system based on dense target detection algorithm plays a key role in core technologies such as automatic shelf inventory auditing and product recognition [1]. The technological synergy thereby

The associate editor coordinating the review of this manuscript and approving it for publication was Yeon-Ho Chung.

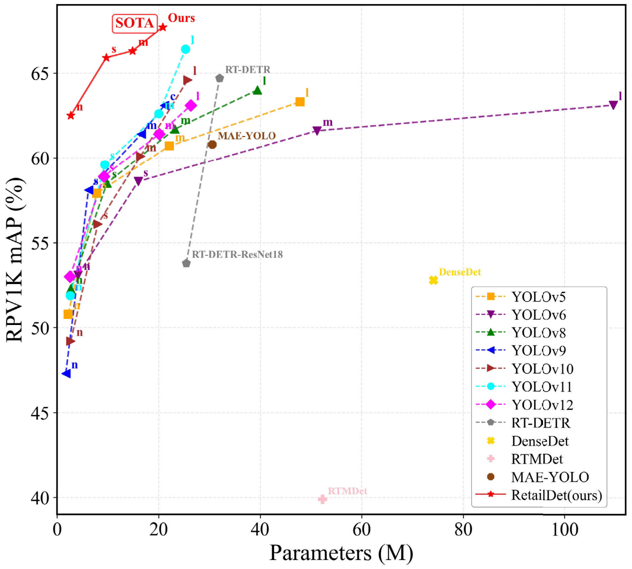


FIGURE 1. Comparison with other SOTA Methods series methods on the RPV1K.

accelerates the practical deployment of unattended commercial environments.

Object detection, as a fundamental task in computer vision, demonstrates its crucial importance across diverse application domains. In autonomous driving, it enables real-time identification of vehicles, pedestrians, and traffic signs [1]. In industrial settings, it facilitates production line defect detection, while in smart cities, it supports intelligent traffic monitoring and urban environment perception [2], [3]. The technology extends to specialized fields such as agricultural automation for crop counting and harvesting [4], and aerial surveillance for crowd density estimation [5]. Notably in retail environments, object detection enables automated shelf management and inventory control through precise product identification and recognition [1], [6], [8].

The rapid digital transformation of the retail industry has given rise to an unprecedented demand for intelligent solutions in product detection, identification, and inventory management. To meet these emerging needs, researchers and industry organizations have developed several comprehensive datasets specifically tailored for the retail environment. These datasets include SKU-110K [7], RP2K [9], Shop-Scale [10], the Grocery Store Dataset [11], RPC [12], and Unitail [13]. These datasets cover a large number of densely arranged products in various retail scenarios, providing a rich resource for comprehensive research. However, their diverse product categories and complex display layouts also pose significant technical challenges for us. In autonomous retail scenarios, detection systems face multiple challenges, such as accurately identifying inventory gaps, quantifying replenishment needs, and evaluating product layout to develop optimal replenishment strategies [14]. Complex retail environments present numerous unique technical challenges for object detection tasks. Constantly changing lighting conditions alter product appearances and affect detection stability, while different viewing angles cause significant variations in target object sizes. Additionally, environmental background elements such as ceilings and floors introduce noise features that distract models from densely arranged products, increasing the difficulty of identifying shelf products. Existing detection methods perform poorly when addressing these retail-specific challenges. Mainstream models including Faster RCNN [15] and its Swin Transformer-based variants [16], RT-DETR [17], SSD-512 [18], and multiple versions of the YOLO series exhibit significantly decreased accuracy and frequent false positives when handling densely arranged products and lighting fluctuations. This stems from their feature fusion, multi-scale feature processing, and attention mechanisms not being adequately adapted to retail scenarios' unique characteristics.

Recent advances in related research have inspired our approach in four key aspects. First, the dynamic collaborative adversarial framework of DCADAN demonstrated how adaptive training architectures enhanced cross-domain feature alignment, informing our design of cross-stage feature interaction modules for multi-scale learning in retail [19]. Second, PhyFnoNet's physics-informed neural operator highlighted the effectiveness of integrating physical constraints into neural networks, motivating our exploration of physics-guided feature regularization to handle spatial constraints in shelf layouts [20]. Third, MILADNet's multimodal imitation learning emphasized the importance of multi-sensor fusion, motivating our exploration of RGB-D bimodal input to clarify spatial hierarchies in shelf scenes [21]. Fourth, AOD-Net's lightweight attention mechanism and Cross-Stage Partial Convolution (CSPC) structure inspired our context-aware attention design and lightweight backbone optimization [22]. Based on these insights, we propose innovative solutions to address the main challenges in retail scenarios: the extremely dense distribution of product targets and the significant imbalance between product and vacancy categories, as illustrated in Fig. 1. Our main contributions are as follows:

- We construct the RPV1K retail scenario-specific dataset, which contains 1,001 high-resolution shelf images of real retail environments, all with accurate annotations of products and vacancy. Focusing on product-vacancy co-occurrence scenarios, this dataset covers real retail features such as dense arrangements, multi-view angles, and lighting variations. It provides targeted benchmark support for product detection and vacancy recognition tasks in the retail field, filling the gap of existing datasets in dual-category joint detection scenarios.
- We propose the RetailDet detection architecture improved based on YOLOv11. Its backbone network integrates multiple attention modules to capture contextual dependencies, and the detection head optimizes the Feature Pyramid Network (FPN) through an adaptive attention mechanism and multi-scale feature fusion,

significantly enhancing the model's ability to learn multi-scale features.
- We design the Multi-scale Unified and Detail Fusion Pyramid Network (MUD-FPN), which enhances the feature representation of small objects and large background area through the separation and optimization of multi-scale features. Additionally, we innovatively introduce the cross-stage feature cross-interaction block (C2fCIB) without explicit output connections. This module, via the implicit regularization and gradient modulation mechanism (IGR), significantly improves the detection performance of small objects while maintaining computational efficiency. It only participates in gradient regulation during the training phase without increasing overhead in the inference phase.
- We develop a novel RGB-D bimodal input backbone network, which integrates RGB and depth information using a two-stream fusion strategy (by concatenating channels after generating depth maps via DepthPro). It effectively enhances the model's ability to detect vacancy, provides a targeted solution to the challenge of spatial hierarchy recognition in complex retail environments, and breaks through the limitations of traditional single-modal input in spatial relationship modeling.

The remainder of this paper is organized as follows: Section II reviews related work in object detection and retail applications; Section III presents the proposed methodology; Section IV describes experimental results and analysis; Finally, Section V concludes the paper with future directions.

II. RELATED WORK

A. YOLO SERIES AND SOTA METHODS

The development of the object detection field is closely related to the evolution of the YOLO series. Since Redmon et al. proposed the groundbreaking single-stage detection framework YOLOv1 in 2016, the YOLO series has been driving the progress of detection technology through continuous innovation [14]. YOLOv2 [23] introduced the anchor mechanism to optimize positioning performance, and YOLOv3 [24] adopted a multi-scale prediction strategy to improve detection adaptability. YOLOv4 [25] achieved performance breakthroughs through the CSPNet backbone network and innovative training strategies, while YOLOv5 [26] was widely used in engineering practice through the focus module and improved PANet feature fusion. YOLOv6 [27] innovatively designed RepBlock and introduced a decoupled head structure, YOLOv7 [28] proposed the E-ELAN framework to enhance feature extraction capabilities, and YOLOv8 [29] further expanded the performance boundaries through task decoupling design. The new generation of YOLO series has shown stronger scene adaptability. YOLOv9 [30] introduced a tiny variant to meet the lightweight requirements, YOLOv10 [31] and YOLOv11 [32] improved performance through feature fusion and network structure optimization respectively, and YOLOv12 [33] further promoted the advancement of detection technology through its innovative framework design.

In the research of special detection technologies for retail scenarios, various innovative methods continue to emerge. RT-DETR [17] first introduced the Transformer architecture into the field of real-time detection. Its lightweight attention mechanism effectively improved the feature expression capability, which was particularly suitable for processing complex spatial relationships between products. DenseDet [34] proposed a novel approach that optimized the feature extraction process through a dense connection strategy, significantly improving the detection reliability in dense display scenarios. The task-aware feature alignment mechanism proposed by RTMDet [35] specifically solved the deformation and occlusion problems in product detection. MAE-YOLO [36] innovatively combined the masked autoencoder pre-training strategy with the object detection architecture, enhancing the model's adaptability to complex retail environments.

However, in the face of challenges such as dense placement of commodities, frequent occlusion and illumination changes in actual retail scenes, existing methods still have potential for improvement in terms of balancing real-time performance and detection reliability. Based on this, we propose the RetailDet architecture, which improves the detection performance in complex scenarios while maintaining efficient reasoning through the RGB and depth feature fusion mechanism and the cross-stage feature interaction strategy without output connection.

B. SMART RETAIL DETECTION

In recent years, detecting visually similar objects in dense scenarios has achieved significant breakthroughs, showcasing substantial application potential across diverse domains. In the agricultural sector, Zhang et al. introduced an anchor-free YOLOX variant that mitigates feature confusion via long-range autonomous aerial vehicle (AAV) sampling, integrating spatial attention mechanisms and multi-scale feature aggregation strategies [4]. This method demonstrates superior performance in handling densely distributed agricultural targets, particularly in scenarios with high visual similarity and dynamic environmental conditions. For autonomous driving systems, He et al. proposed orthogonal attention modules to address overlapping vehicle features, enhancing detection robustness in complex traffic scenarios [37]. Their approach incorporates multi-dimensional feature decomposition and adaptive feature alignment, improving detection accuracy for occluded and densely packed vehicles.

In retail object detection research, Goldman et al. created the SKU-110K dataset targeting dense retail scenarios [7]. Their approach combines a Jaccard index evaluation strategy with an expectation-maximization merging mechanism to tackle the challenges of crowded retail scenes. Building upon this foundation, Peng et al. introduced the RP2K dataset, which enhanced fine-grained product recognition through its comprehensive category annotations [9]. This advancement enabled more sophisticated automated retail applications. These retail-focused architectures have effectively addressed retail specific challenges where traditional methods show limitations. Such challenges include severe occlusion, varying illumination conditions, and high product similarity. Furthermore, their solutions have advanced retail object detection technology by effectively handling practical issues, including reflective surfaces, complex product stacking patterns, and perspective distortions.

In the field of retail product detection, researchers have contributed diverse datasets and innovative detection models, which have continuously promoted the rapid development of this field. Klasson et al. developed a grocery store dataset that contains more than 10,000 real scene images with fine-grained annotations, covering 80 product categories [11]. These images are systematically captured under different shelf layouts and viewing conditions, providing an important basis for developing and evaluating retail detection systems. At the same time, Chen et al. introduced the Unitail benchmark with an innovative multimodal approach, which greatly advanced retail automation technology [13]. They adopted a complex matching mechanism to combine product detection algorithms with text recognition functions. The Unitail architecture shows remarkable adaptability in a variety of retail environments, from traditional supermarket shelves to modern autonomous stores. Through advanced feature extraction techniques, it effectively addresses key practical challenges such as product rotation, partial occlusion, and brand changes. Existing datasets have common shortcomings, including limited image clarity, incomplete coverage of product categories, a lack of joint labels for products and vacancies, insufficient diversity of perspectives, and limited coverage of lighting variations. Our newly created dataset, based on high-definition images, improves product category coverage, supplements the aforementioned joint labels, incorporates rich perspectives, and includes diverse data on handling lighting variations, thus providing more comprehensive support for enhancing model performance.

Retail vacancy detection has evolved significantly from traditional rule-based approaches to sophisticated computer vision systems. Early work by Papakiriakopoulos et al. focused on heuristic-based decision support systems, establishing the basic principles for automated retail management [38]. Later developments explored vision-based approaches combined with advanced deep learning techniques to improve accuracy. Pietrini et al. created an innovative shelf management architecture using RetinaNet, which achieved outstanding performance in product recognition through improved feature extraction and classification methods [39]. Milella et al. advanced the field by using depth cameras and 3D point cloud methods for shelf availability estimation, introducing a novel perspective for spatial analysis in retail environments [40]. Current challenges in practical applications include identifying densely packed and visually similar products, handling complex shelf architectures with different layouts and configurations, and addressing the class imbalance between products and vacancies. These technological challenges underscore the demand for more robust and adaptable detection systems.

While existing studies have made progress in retail product detection, three critical challenges remain unaddressed: single-category detection limitations, dense object distribution complexities, and insufficient attention concentration on key targets where visual noise from ceiling and floor regions often interferes with detection accuracy. To address these challenges, we propose RetailDet, a comprehensive detection framework that effectively handles dense product arrangements, varying illumination conditions, and environmental distractions in retail scenarios.

III. DATASET AND PROPOSED METHOD

A. RPV1K DATASET

To build a high-quality RPV1K dataset, we establish a comprehensive workflow covering data collection, annotation, and review procedures. In the data collection phase, four master's degree holders utilize smartphones (35+ megapixels) to capture 1,001 shelf images across 20 supermarkets, with balanced distribution among China (350), Southeast Asia (350), and Europe (301). This geographically diverse data collection comprehensively captures the product display characteristics of supermarkets in different regions, enhancing the diversity of scenarios and authenticity of the dataset. For the annotation process, we utilize the X-AnyLabeling tool to annotate products and vacancies using bounding boxes defined by their top-left and bottom-right coordinates. We establish a review team consisting of eight members (four PhD holders and four master's degree holders), organized into four pairs, each comprising one PhD holder and one master's degree holder. The annotation workflow begins with initial annotation performed by four master's degree holders, followed by the development of a pre-annotation model for generating initial labels and a first-round detailed review. The data is then divided into four equal parts (250 images per group) for pixel-level precise review of bounding boxes by each group, ensuring that the bounding boxes accurately encompass entire products or vacancies, particularly for side-shot images. Subsequently, a cross-group verification process is implemented where each group reviews the work of other three groups. This iterative review process continues until no further corrections are needed for product or vacancy labels. Through our comprehensive quality control mechanism incorporating human collaboration, cross-verification, and iterative optimization, combined with standardized tools and annotation protocols, we ensure the high quality and reliability of the RPV1K dataset annotations. The acquisition process of the RPV1K dataset is shown in Fig. 2.

As shown in Fig. 3, the RPV1K dataset consists of 1,001 images. These images are collected from a variety of retail environments with different resolutions. In these scenes, the objects are arranged closely, while the vacancies are not only

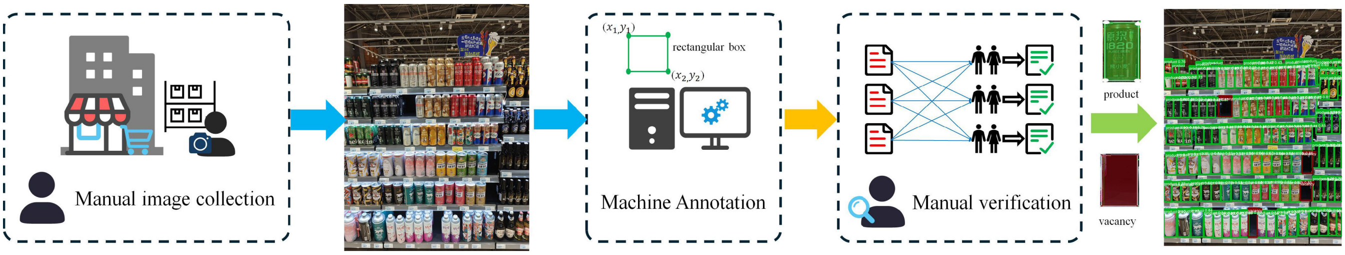


FIGURE 2. Steps for generating labels for the RPV1K dataset.

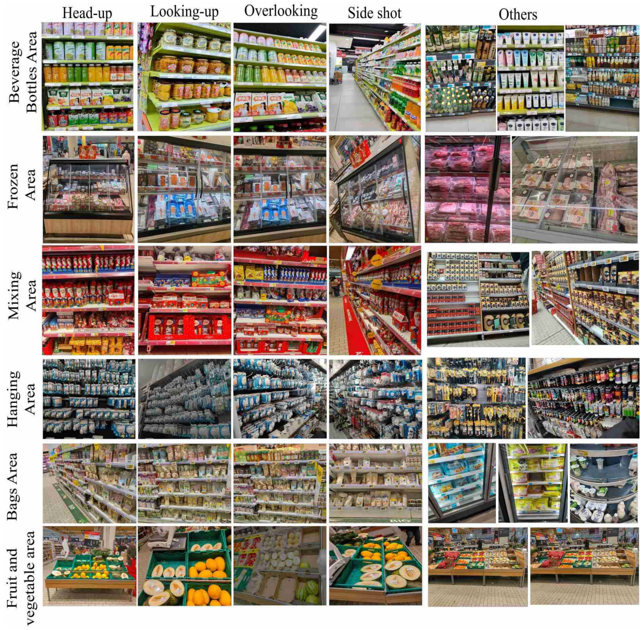


FIGURE 3. The retail products and Vacancy 1,001 (RPV1K) dataset covers 6 representative shelf scenes such as beverage bottles and frozen products. The images are taken from multiple angles, including common viewpoints such as horizontal, upward, downward, and side views.

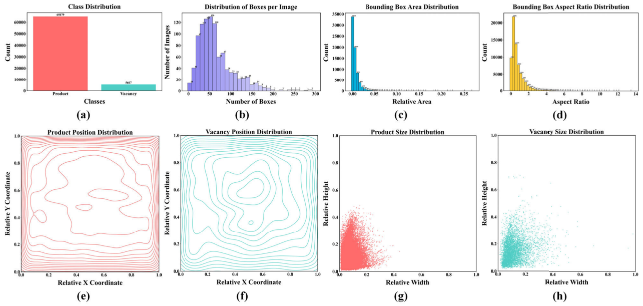


FIGURE 4. Statistical analysis of RPV1K (retail products and vacancies 1,001) dataset.

of different sizes and shapes, but also unevenly distributed and relatively scattered. In addition, the occlusion of surrounding objects further increases the difficulty of identifying vacancies. Fig. 4 presents comprehensive statistical analyses of the RPV1K dataset. Examining the category distribution shown in Fig. 4 (a) reveals significant class imbalance: the product class contains 68,597 instances, while the vacancy class includes 5,687 instances, yielding a product-to-vacancy ratio of approximately 12:1. This imbalanced ratio authentically reflects real-world retail scenarios where shelf vacancies naturally occur due to customer purchases. According to Fig. 4 (b), the distribution of boxes per image follows a right-skewed pattern, mainly concentrated between 20 and 120 instances. Analysis of the bounding box in Fig. 4 (c) demonstrates long-tail characteristics with a predominance of small-area targets. The aspect ratio distribution illustrated in Fig. 4 (d) also shows long-tail properties, primarily ranging between 0 and 4. The spatial distribution patterns are visualized through heatmaps in Fig. 4 (e) and (f). These visualizations depict the spatial density patterns across relative X and Y coordinates for both product and vacancies, highlighting their distinct distribution characteristics. The inter-class comparisons shown in Fig. 4 (g) and (h) reveal that while most targets are small in size, the vacant class tends to exhibit more irregular shapes and larger dimensions compared to the product class.

Previous retail datasets, such as SKU-110k, Grocery Store Dataset, and Locount [41], focus solely on product detection. In contrast, RPV1K innovatively adopts a dual-category detection paradigm. Our dataset not only provides precise annotations for densely distributed products but also includes bounding box annotations and category labels for vacancies, thus covering a wider range of real-world out-of-stock scenarios.

B. RETAILDET ARCHITECTURE

YOLOv11 is a relatively new version in the YOLO series. It has performed well in object detection tasks and has received high praise. However, in retail application scenarios, YOLOv11 still faces many severe challenges: the vacancy and product categories are seriously unbalanced, the scale changes significantly, and the spatial relationship is intricate. These factors greatly limit the feature expression ability of the model. To address these challenges, we develope RetailDet, a specialized detection model for retail environments, by incorporating multiple advanced algorithm modules based on the YOLOv11 architecture. Our model integrates both the MUD-FPN module and the efficient poly kernel initialization block (PKIB) [42] attention mechanism. Combining the strengths of YOLOv10 and YOLOv11, our experiments on the RPV1K dataset have demonstrated superior detection performance. Fig. 5 illustrates the overall architecture of the RetailDet.

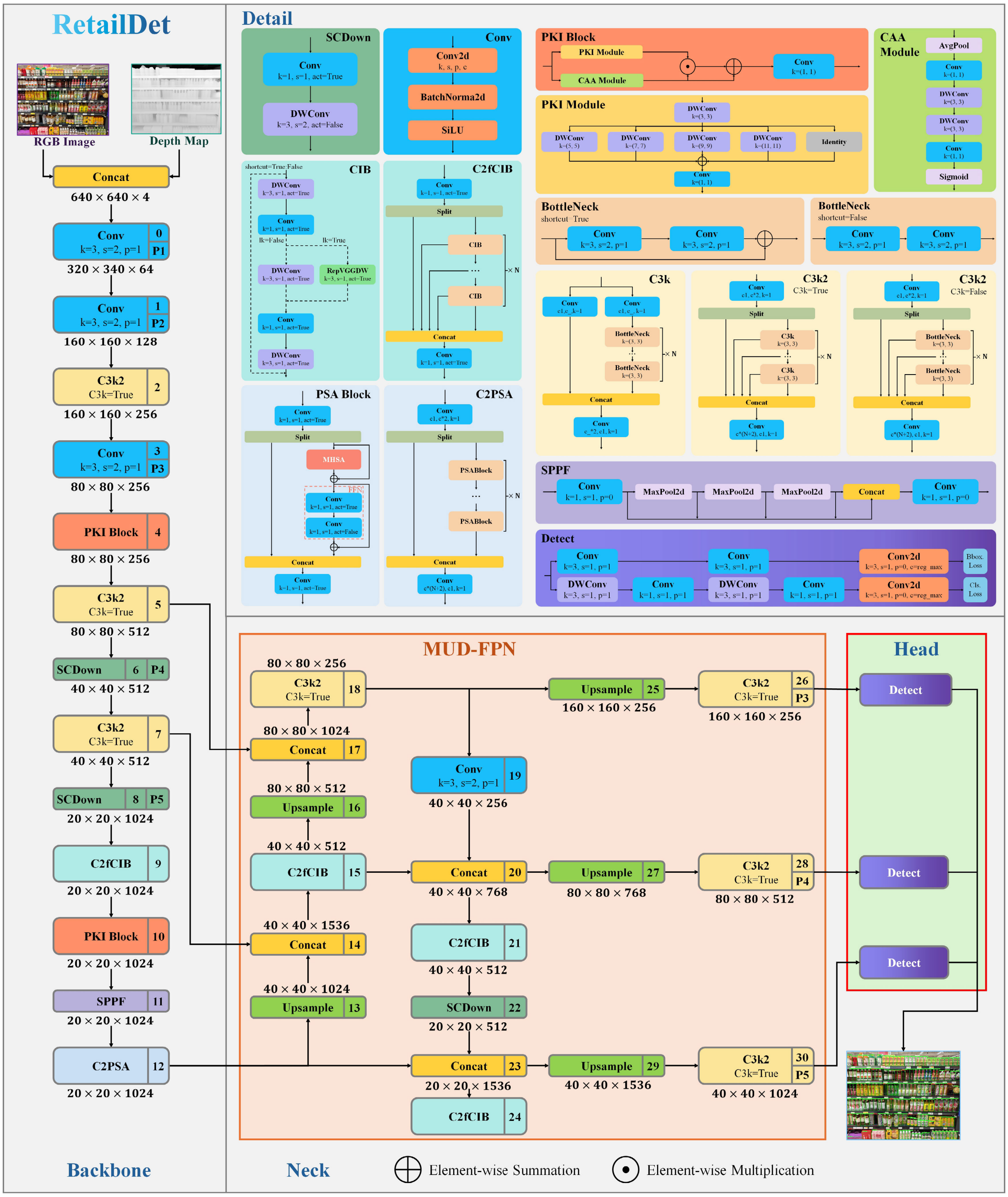


FIGURE 5. RetailDet architecture.

1) RGB-D BIMODAL

Traditional detection methods based on RGB images have difficulty in effectively handling spatial hierarchical relationships and occlusion scenarios [43]. Especially in shelf images, when products and vacancies coexist in the same area, detection results often favor products with prominent features, while missing vacancies due to their similar pixel characteristics, leading to detection confusion and false negatives. To address this challenge, we propose integrating depth information with RGB data as the backbone network input to enhance detection performance.

The RGB-D bimodal input process consists of two sequential stages. In the first stage, we estimate the depth map from the input RGB image. Fig. 6 shows the effect of a depth map generated using DepthPro [44]. With this depth map, we can more clearly observe the empty locations on the shelf (circled areas in the figure), and the discrimination between products and vacancies becomes more remarkable. Let $\mathbf{I}_{\text{rgb}} \in \mathbb{R}^{H \times W \times 3}$ be the input RGB color image, where H and W denote the height and width of the image (both set to 640 pixels), and 3 represents the RGB channels. Using the pre-trained DepthPro model, we infer the depth information

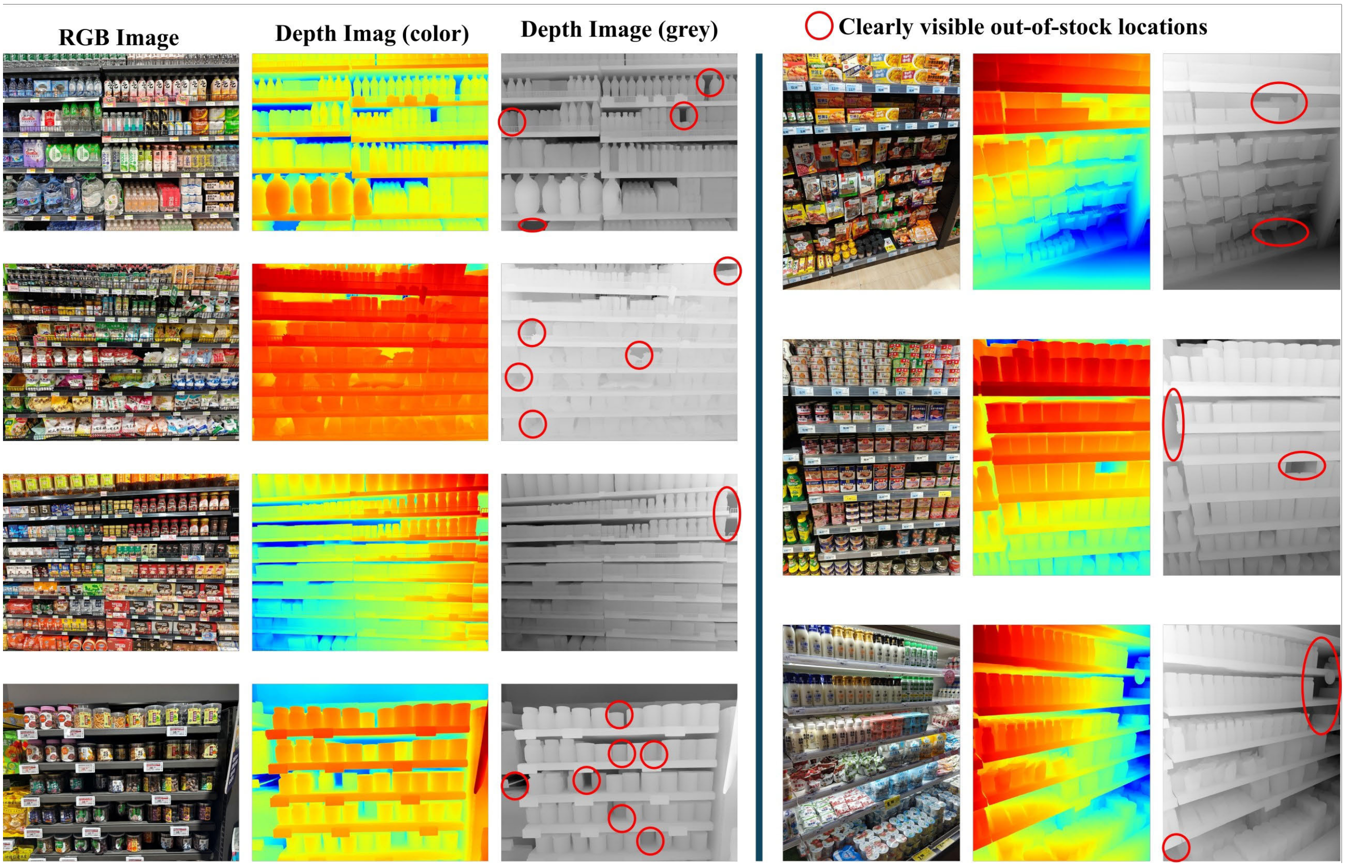


FIGURE 6. The effect of RGB-D bimodal input data.

from $\mathbf{I}_{\text{rgb}}$ to obtain a single-channel depth grayscale map $\mathbf{I}_{\text{depth}} \in \mathbb{R}^{H\times W\times 1}$. The DepthPro model is realized through its encoder-decoder architecture, and it can be written as:

$$\mathbf{I}_{\text{depth}} = \mathcal{F}_{\text{DepthPro}}(\mathbf{I}_{\text{rgb}}) \tag{1}$$

where $\mathcal{F}_{\text{DepthPro}}$ represents the mapping function learned by the DepthPro model, which encodes the RGB image information into a single-channel depth representation.

In the second stage, we perform channel concatenation on the three-channel RGB color image and the single-channel depth map, which can be written as:

$$\mathbf{I}_{\text{Concat}} = \text{Concat}(\mathbf{I}_{\text{rgb}}, \mathbf{I}_{\text{depth}}) \tag{2}$$

where $\mathbf{I}_{\text{Concat}} \in \mathbb{R}^{640\times 640\times 4}$ represents the fused feature map.

This sequential processing ensures that the depth information is accurately estimated before concatenation, allowing the model to effectively utilize both RGB features and depth information. We use the pre-trained weights of DepthPro for inference to generate the corresponding depth maps for our dataset. Considering the practical constraints and real-time performance requirements of edge devices, we implement this two-stage channel fusion strategy to fuse RGB and depth information, keeping the input size as 640 × 640 × 4.

2) DETAIL MODULE

In retail shelf scenarios, vacancies exhibit significant variations in both scale and appearance. Compared to product instances, vacancies present more complex characteristics due to their irregular shapes and contextual dependencies. The challenge is further compounded by class imbalance and limited vacancy samples, which constrains the model's feature learning capability and increases detection difficulty. To address these challenges, we introduce multiple advanced attention enhancement modules, as shown in Fig. 5.

As shown in Fig. 5, the poly kernel initialization (PKI) [42], [45] module uses a parallel processing multi-scale depthwise separable convolution mechanism (PPMDSConv) to handle scale changes, which can be expressed by

$$\text{PKI}_{\text{multi}} = \text{Conv}_{1\times 1}\left(\underbrace{\sum_{k_i\in\{3,5,7,9,11\}} \text{DWConv}_{k_i}(\mathbf{X}) \oplus \mathbf{X}}_{\text{PPMDSConv}}\right) \tag{3}$$

where $\sum_{k_i\in\{3,5,7,9,11\}} \text{DWConv}_{k_i}(\mathbf{X})$ indicates parallel execution of depthwise separable convolutions with kernel sizes 3, 5, 7, 9, and 11 on the input $\mathbf{X}$, $\oplus\mathbf{X}$ corresponds to the identity branch (preserving original input features), $\text{Conv}_{1\times1}$ integrates features after multi-branch fusion and outputs the multi-scale fused feature $\text{PKI}_{\text{multi}}$.

The context anchor attention (CAA) [42] module optimizes image features by modeling long-range inter-pixel correlations. It first compresses the spatial dimension of the input feature $\mathbf{X}$ via average pooling (AvgPool), then applies a 1×1 convolution ($\text{Conv}_{1\times1}$) to adjust the feature channels. After that, it excavates long-range pixel dependencies through multiple layers of 3×3 depthwise separable convolutions ($\text{DWConv}_{3\times3}$). Subsequently, another 1×1 convolution and the activation function σ (such as Sigmoid) are used to generate attention weights $\text{CAA}(\mathbf{X})$. This enhances the feature quality of key areas in the global feature map, improving the model's detection robustness in complex backgrounds.The process of the CAA module can be expressed by the following formula:

$$\begin{aligned}\text{CAA}(\mathbf{X}) = \sigma\,(&\text{Conv}_{1\times1}\,(\text{DWConv}_{3\times3} \\ &(\text{DWConv}_{3\times3}\,(\text{Conv}_{1\times1}\,(\text{AvgPool}(\mathbf{X})))))\end{aligned} \quad (4)$$

Assume the input feature is $\mathbf{X}$, the output of the PKI Module is $\text{PKI}(\mathbf{X})$, and the output of the CAA module is $\text{CAA}(\mathbf{X})$. The formula for PKI Block is expressed as:

$$\text{PKIB}(\mathbf{X}) = \text{Conv}_{1\times1}\,(\text{CAA}(\mathbf{X}) \oplus (\text{PKI}(\mathbf{X}) \odot \text{CAA}(\mathbf{X}))) \quad (5)$$

Here, $\odot$ denotes element-wise multiplication, and $\oplus$ denotes residual addition. In summary, through the design of multi-scale feature fusion and contextual attention modeling, PKIB effectively addresses the detection instability issues caused by scale variations and complex contextual information in object detection.

As an upgraded version of the C2f (Convolution Block with Two Features) architecture, the cross-stage local connection (C3k2) maintains the lightweight advantages of feature separation and fusion while introducing stackable C3 modules to replace standard 3×3 convolutions. This design enables dynamic receptive field adjustment and enhanced multi-scale feature representation. While preserving computational efficiency, it significantly improves model adaptability to complex scenarios, particularly benefiting multi-scale object detection tasks where feature diversity is crucial.

Aiming at the limitations of the original parallel spatial attention (PSA) architecture, the parallel spatial attention convolution block (C2PSA) achieves dual optimizations through the strategies of multi-level PSA block stacking and channel splitting. The progressive feature extraction mechanism helps integrate local details and global semantic information. The selective channel attention mechanism reduces computational complexity while preserving shallow feature flow integrity. These improvements enhance the model's context modeling ability in complex scenarios. This design also achieves a good balance between computational efficiency and feature expression.

3) BACKBONE MODULE

There are three key designs in RetailDet's backbone network, as shown in Fig. 5. Firstly, C2f is replaced with C3k2 to achieve a dynamic receptive field, and PSA is upgraded to C2PSA for optimizing multi-scale context modeling. Secondly, the spatial pyramid pooling-fast (SPPF) module is embedded to facilitate multi-scale feature extraction and improve shelf product detection accuracy. Thirdly, PKIBs are deployed at the P_3 and P_5 feature layers. At P_3, PKIB enhances small object features through multi-scale convolutions and CAA. At P_5, it integrates global information using large receptive field convolutions and strengthens complex scenario modeling through CAA.

The P_4 layer possesses unique attributes as it naturally fuses local information from shallow layers and global information from deep layers. In MUD-FPN, P_4 layer features are fused with those from P_3 and P_5 layers, strengthening its feature representation ability. Therefore, adding PKIB at P_4 is not beneficial and may lead to feature redundancy, increased feature fusion complexity, and more complicated gradient calculations, affecting gradient update stability and weight value learning. Given input feature $\mathbf{I}_{\text{Concat}} \in \mathbb{R}^{640\times640\times4}$ after RGB-D concatenation, this backbone architecture derives three output paths.

First Output Path (outputting $\mathcal{O}_5$ via node N_5). The input $\mathbf{I}_{\text{Concat}}$ first undergoes the Conv_{P1} convolution operation, with the kernel size (k), stride (s), and padding (p) set as $k = 3$, $s = 2, p = 1$, yielding $\mathbf{F}_0 \in \mathbb{R}^{320\times320\times64}$.

Subsequently, $\mathbf{F}_0$ is fed into Conv_{P2} (configured as $k = 3$, $s = 2$, $p = 1$) to generate $\mathbf{F}_1 \in \mathbb{R}^{160\times160\times128}$. Next, $\mathbf{F}_1$ is processed by the $\text{C3k2}_{\text{True}}$ module. Notably, for the C3k2 module (regardless of mode $=$ True or False), it only modifies the channel number while keeping the (H) and (W) unchanged. Thus, $\mathbf{F}_2 \in \mathbb{R}^{160\times160\times256}$ is obtained. Subsequently, $\mathbf{F}_2$ undergoes convolution via Conv_{P3} ($k = 3$, $s = 2, p = 1$) to produce $\mathbf{F}_3 \in \mathbb{R}^{80\times80\times256}$. After processing by PKIB, $\mathbf{F}_4 \in \mathbb{R}^{80\times80\times256}$ is output. Finally, $\mathbf{F}_4$ passes through $\text{C3k2}_{\text{True}}$, and the output $\mathcal{O}_5$ is defined as:

$$\mathcal{O}_5 = \text{C3k2}_{\text{True}}(\mathbf{F}_4) \in \mathbb{R}^{80\times80\times512} \quad (6)$$

where $\mathbf{F}_i$ denotes the feature map transmitted from node N_i in the network.

Second Output Path (through node N_7, output $\mathcal{O}_7$). $\mathcal{O}_5$ is first input into the selective channel down-sampling (SCDown) module of node N_6 (with kernel size $k = 3$, stride $s = 2$, and padding $p = 1$), resulting in a tensor of dimension $\mathbb{R}^{40\times40\times512}$. Then, this result is passed to the $\text{C3k2}_{\text{True}}$ module for processing, which preserves the spatial dimensions H and W. The output $\mathcal{O}_7$ is defined as:

$$\mathcal{O}_7 = \text{C3k2}_{\text{True}}(\text{SCDown}(\mathcal{O}_5)) \in \mathbb{R}^{40\times40\times512} \quad (7)$$

The SCDown module is a key component for multi-scale feature extraction and spatial dimension compression. The downsampling operation with a stride of 2 reduces feature map resolution and expands the receptive field. This operation enhances contextual awareness of large-scale targets while reducing computational overhead. The stacked three convolutional layers enable efficient hierarchical feature extraction. The SiLU activation function introduces nonlinear transformations to enhance network expressiveness. With 512 channels maintained for both input and output, the module seamlessly connects to C2PSA feature processing. This channel configuration enables efficient feature fusion and supports MUD-FPN multi-scale feature extraction. The design balances computational efficiency and feature extraction capabilities, making it suitable for product and vacancy detection in retail scenarios.

This design achieves a balance between computational efficiency and feature extraction capabilities, making it optimal for product and vacancy detection in retail scenarios.

Third Output Path (through node N_{12}, output $\mathcal{O}_{12}$). $\mathcal{O}_7$ passes through the SCDown module ($k = 3, s = 2, p = 1$) to obtain feature map $\mathbf{F}_8 \in \mathbb{R}^{20\times20\times1024}$. Subsequently, the feature is processed through C2fCIB, PKIB, and SPPF modules (none of which alter the feature map height and width). Finally, the C2PSA module processes $\mathbf{F}_{11}$ to produce output $\mathcal{O}_{12}$.

$$\mathcal{O}_{12} = \text{C2PSA}(\mathbf{F}_{11}) \in \mathbb{R}^{20\times20\times1024} \tag{8}$$

where $\mathbf{X}$ represents the feature input, and the definitions of the key components are as follows:

$$\text{Conv}(\mathbf{X}) = \text{Conv2D}(\mathbf{X}, k = 3, s = 2, p = 1) \tag{9}$$

$$\text{C3k2}_{\text{mode}}(\mathbf{X}) = \begin{cases} \text{C3k2}(\mathbf{X}), & \text{if mode} = \text{True} \\ \text{C3k2}(\mathbf{X}) + \mathbf{X}, & \text{if mode} = \text{False} \end{cases} \tag{10}$$

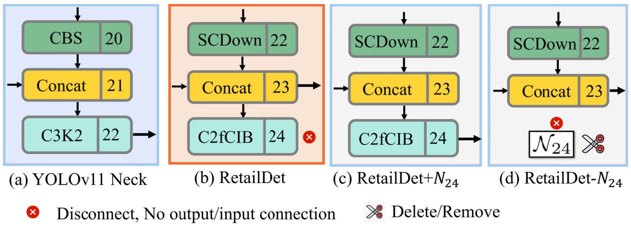
$$\text{SCDown}(\mathbf{X}) = \text{DWConv}(\text{Conv}(\mathbf{X})) \tag{11}$$

This multi-scale feature hierarchy $\{\mathcal{O}_5, \mathcal{O}_7, \mathcal{O}_{12}\}$ provides robust representations at high (80 × 80 × 512), medium (40 × 40 × 512), and low (20 × 20 × 1024) resolutions, effectively supporting subsequent detection tasks. Among them, DWConv belongs to depthwise separable convolutions. Introducing this module into the detection head can greatly reduce computational complexity and significantly decrease the parameter quantity.

4) MUD-FPN MODULE

The traditional FPN structures typically employ three main feature levels where P_3, P_4, and P_5 correspond to spatial resolutions of $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ of the input image, specifically 80×80, 40×40, and 20×20. Modern detection models, with YOLO series being a prime example, have made significant progress in handling dense small object scenarios. However, their conventional FPN structures still show limitations in retail scenes where significant scale differences and complex distributions are present.

Specifically, there is severe imbalance between product and vacancy targets. Object scales span significantly from small to large, accompanied by complex spatial patterns of high density and discrete distribution. Furthermore, there are inherent feature similarities between the two object categories, and the similarity exhibits strong spatial dependence, presenting an extreme distribution from highly similar to clearly distinguishable local areas.The 80 × 80 highest-resolution feature map of conventional FPN fails to provide fine-grained feature representation for accurately distinguishing small-scale targets with similar features. In highly dense areas, low-resolution feature maps easily cause feature confusion and information loss, affecting detection accuracy. The larger-scale feature maps P_4 and P_5 lose excessive spatial details, while the highest-resolution feature map P_3 lacks a sufficient receptive field to capture target contextual information, significantly decreasing cross-scale object detection performance.



(a) YOLOv11 Neck (b) RetailDet (c) RetailDet+N_{24} (d) RetailDet-N_{24}

Disconnect, No output/input connection Delete/Remove

FIGURE 7. Connection patterns of the N_{24} module: IGR and conventional structures. (a) Original YOLOv11 neck structure; (b) IGR structure adopted by RetailDet (C2fCIB attention mechanism module not shown in output); (c) Structure with N_{24} retained and explicit output connections; (d) Structure with N_{24} removed.

To address these challenges, we propose the MUD-FPN architecture, which steadily improves model performance while reducing model parameters. Specifically, we adopt the FPN feature hierarchical extraction approach from YOLOv10 and incorporate the cross-scale feature fusion design from YOLOv11. We replace the C2f in YOLOv10's Neck with C3k2 to enhance the model's feature extraction capability. Unlike traditional FPN that builds multi-scale feature representations through top-down feature fusion, we extend feature map resolution to higher dimensions (160 × 160, 80 × 80, 40 × 40) through progressive upsampling to enhance spatial details and receptive fields. This is achieved by embedding Upsample and C3k2 modules at nodes N_{18}, N_{20}, and N_{23} in YOLOv10's Neck. This design utilizes C3k2 modules to optimize post-upsampling feature quality, effectively reconstructing details lost during downsampling while enhancing feature representation and suppressing interpolation artifacts from simple upsampling. The dual-peak design combines FPN's basic feature extraction with high-resolution feature enhancement, enabling MUD-FPN to maintain multi-scale object detection capabilities while significantly improving dense small object detection, effectively addressing the challenge of non-uniform scale object detection.

Notably, we introduce an implicit gradient regulator. We use the output of node N_{23} as input to N_{24}, while N_{24} has no output connection. N_{24} is a C2fCIB attention mechanism module that has no explicit output connections in the model structure, thus not affecting the detection head output during

Algorithm 1 RetailDet Architecture

```
Require: RGB image and Depth image
Ensure: Multi-scale feature maps $P_3, P_4, P_5$
    Stage 1: Backbone module
    RGB-D channel fusion,    ▷ $640 \times 640 \times 4 \leftarrow 640 \times 640 \times 3$ and $640 \times 640 \times 1$
    $\mathbf{F}_{in} \leftarrow$ Concat(RGB, Depth)
    for each input tensor in backbone do
        $\mathcal{O}_5 \leftarrow$ C3K2(PKIB(Conv(C3K2(Conv(Conv($\mathbf{F}_{in}$))))))
        $\mathcal{O}_7 \leftarrow$ C3K2(SCDown($\mathcal{O}_5$))
        $\mathcal{O}_{12} \leftarrow$ C2PSA(SPPF(PKIB(C2fCIB(SCDown($\mathcal{O}_7$)))))
    end for
    Stage 2: MUD-FPN module
    $N_{15} =$ C2fCIB(Concat(Upsampling($\mathcal{O}_{12}$), $\mathcal{O}_7$))
    $N_{16} =$ Upsampling($N_{15}$)
    $N_{19} =$ Conv(C3K2(Concat($N_{16}$, $\mathcal{O}_5$)))
    $N_{20} =$ Concat($N_{15}, N_{19}$)
    $N_{22} =$ SCDown(C2fCIB($N_{20}$))
    Stage 3: Head module
    $P_3 \leftarrow$ C3K2(Upsample(C3K2(Concat($\mathcal{O}_5, N_{16}$))))
    $P_4 \leftarrow$ C3K2(Upsample(Concat($N_{15}, N_{19}$))
    $P_5 \leftarrow$ C3K2(Upsample(Concat($\mathcal{O}_{12}, N_{22}$)))
    return $P_3, P_4, P_5$
```

inference, but functions as an implicit gradient regulator influencing gradient loss updates during training.

To validate our design, we conducted a series of structural comparison experiments, as shown in Fig.7. Fig.7 (a) shows the Neck structure of the YOLOv11 model, where N_{22} directly outputs to the detection head. Through ablation experiments comparing Fig.7 (a) and (c), we verify the necessity of replacing N_{22} with N_{24}; comparison experiments between Fig.7 (c) and (d) demonstrate N_{24}'s impact on model performance; finally, comparison experiments between Fig.7 (b) and (c) validate the effectiveness of the IGR design. Experimental results show that N_{24} with IGR design structure plays a crucial role in balancing performance and efficiency.

We summarize the core algorithmic process of RetailDet architecture as pseudocode, as shown in Algorithm 1.

C. COMPREHENSIVE LOSS FUNCTION DESIGN

Following the RGB-D bimodal architecture design, we propose a comprehensive loss function based on YOLOv11 to optimize model performance for retail product and vacancy detection. The loss function addresses both classification accuracy and localization precision through three complementary components.

1) IOU

IoU (Intersection over Union) [46], [47], also referred to as the Jaccard index, serves as a fundamental metric in object detection. It is used to quantify the degree of overlap between a predicted bounding box and a ground-truth bounding box. Mathematically, the IoU is calculated as the ratio of the intersection area of the two boxes to their union area. Let the predicted bounding box be

$$B_{pred} = (x_1^p, y_1^p, x_2^p, y_2^p) \tag{12}$$

and the ground-truth bounding box be

$$B_{gt} = (x_1^g, y_1^g, x_2^g, y_2^g) \tag{13}$$

where (x_1, y_1) and (x_2, y_2) represent the top-left and bottom-right coordinates of the boxes respectively.

The intersection area A_{inter} can be calculated as follows:

$$\begin{aligned} x_{left} &= \max(x_1^p, x_1^g), & y_{top} &= \max(y_1^p, y_1^g), \\ x_{right} &= \min(x_2^p, x_2^g), & y_{bottom} &= \min(y_2^p, y_2^g) \end{aligned} \tag{14}$$

If $x_{right} > x_{left}$ and $y_{bottom} > y_{top}$, then

$$A_{inter} = (x_{right} - x_{left}) \times (y_{bottom} - y_{top}) \tag{15}$$

otherwise $A_{inter} = 0$.

The area of the predicted box is

$$A_{pred} = (x_2^p - x_1^p) \times (y_2^p - y_1^p) \tag{16}$$

and the area of the ground-truth box is

$$A_{gt} = (x_2^g - x_1^g) \times (y_2^g - y_1^g) \tag{17}$$

The union area is defined as:

$$A_{union} = A_{pred} + A_{gt} - A_{inter} \tag{18}$$

The IoU formula is given by

$$\mathrm{IoU} = \frac{A_{\mathrm{inter}}}{A_{\mathrm{union}}} \tag{19}$$

The value of IoU ranges from 0 to 1. A value of 0 indicates that there is no overlap between the two boxes, while a value of 1 implies that the predicted box and the ground-truth box are exactly the same. In practical object detection scenarios, if the IoU between a predicted detection and a ground-truth box exceeds a pre-defined threshold (commonly 0.5 or 0.75), the detection is usually considered correct.

2) CLASSIFICATION LOSS

The classification loss measures the accuracy of the model's prediction of the object categories, using the Binary Cross Entropy (BCE) loss function [48]. Let $\hat{y}_{ij}$ be the model's predicted score for the j-th category of the i-th sample, and y_{ij} be the corresponding ground-truth label. The classification loss $\mathcal{L}_{Cls}$ is

$$\begin{aligned} \mathcal{L}_{Cls} = \frac{1}{N_{pos}} \sum_{i=1}^{N_{pos}} \sum_{j=1}^{C} -[y_{ij} \log(\sigma(\hat{y}_{ij})) \\ + (1 - y_{ij}) \log(1 - \sigma(\hat{y}_{ij}))] \end{aligned} \tag{20}$$

where N_{pos} is the number of positive samples, C is the total number of categories, and $\sigma(x) = \frac{1}{1+e^{-x}}$ is the Sigmoid activation function.

3) BOUNDING BOX REGRESSION LOSS

To achieve precise object localization, particularly important for densely packed retail shelves, we combine two complementary components: the Complete Intersection over Union (CIoU) loss and the Distribution Focal Loss (DFL) [49], [50].

a: CIoU LOSS

The CIoU loss takes into account the overlapping area, the distance between the centers, and the aspect ratio consistency between the predicted box and the ground-truth box. Let $\hat{b}_i$ be the i-th predicted box and b_i be the corresponding ground-truth box. The CIoU loss $\mathcal{L}_{CIoU}$ (or $\mathcal{L}_{Bbox}$) is

$$\mathcal{L}_{CIoU} = \mathcal{L}_{Bbox} = 1 - \mathrm{CIoU}(\hat{b}_i, b_i) \tag{21}$$

where $\mathrm{CIoU}(\hat{b}_i, b_i)$ is the CIoU value of the i-th predicted box and the ground-truth box.

b: DFL LOSS

The DFL loss further optimizes the accuracy of bounding box prediction by modeling the distribution of box coordinates. Let $\hat{d}_i$ be the distribution of the i-th predicted box and d_i be the corresponding target distribution. The DFL loss is

$$\mathcal{L}_{\mathrm{DFL}} = \frac{1}{N_{\mathrm{pos}}} \sum_{i=1}^{N_{\mathrm{pos}}} \sum_{k=1}^{4} \left[\mathrm{CE}(\hat{d}_i^k, t_l^k) \cdot w_l^k + \mathrm{CE}(\hat{d}_i^k, t_r^k) \cdot w_r^k \right] \tag{22}$$

where k represents the four directions (left, top, right, bottom) of the bounding box, t_l^k and t_r^k are the left and right boundaries of the target distribution respectively, w_l^k and w_r^k are the corresponding weights, and $\mathrm{CE}(x, y) = -\sum_j y_j \log(x_j)$ is the cross-entropy (CE) loss.

4) TOTAL LOSS

To balance the contribution of each component, we combine these losses with equal weights. The total loss $\mathcal{L}_{total}$ is

$$\mathcal{L}_{total} = \lambda_{Cls} \cdot \mathcal{L}_{Cls} + \lambda_{CIoU} \cdot \mathcal{L}_{CIoU} + \lambda_{DFL} \cdot \mathcal{L}_{DFL} \tag{23}$$

Among them, λ_{Cls} is set to 0.5, λ_{CIoU} is set to 7.5, and λ_{DFL} is set to 1.5 to ensure balanced optimization of classification and regression accuracy. For YOLOv8 and higher models, as well as RetailDet, we use the comprehensive loss function L_{total}.

D. PERFORMANCE ASSESSMENT INDICATORS

Precision is a measure of the accuracy of positive predictions made by the classifier. It is the ratio of true positive predictions to the total number of positive predictions made. Precision is especially important when the cost of a false positive is high. It can be defined as:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{24}$$

True positives (TP) are the number of samples correctly identified as the positive class, while true negatives (TN) are the number of samples correctly identified as the negative class. False positives (FP) represent the misclassification of negative class samples as positive, and false negatives (FN) occur when positive class samples are not correctly identified.

Recall (also known as sensitivity) measures the ability of a classifier to find all relevant samples, calculated as the ratio of the number of true positive samples to the total number of actual positive samples. Recall is particularly important when it is necessary to minimize false negatives. The higher the recall, the lower the probability that the model will miss a true positive sample, which is defined as:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{25}$$

F1 score (F1) is a single metric that combines both precision and recall to measure the overall performance of a classifier. It is the harmonic mean of precision and recall, giving equal weight to both metrics. F1 is particularly useful when the class distribution is imbalanced, and there is a need to balance between precision and recall. The F1 can be calculated as:

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{26}$$

Average Precision (AP) is a basic metric used in object detection to evaluate the performance of single-class detection. It is calculated as the area under the precision-recall curve (AUC-PR), which quantifies the overall performance of the detector at different confidence thresholds. The calculation process involves determining the true positives of each predicted bounding box based on its IoU with the ground truth box, and then calculating the maximum precision value at different recall levels. The AP score is derived from the average of these precision values. An effective detector should maintain both high precision and high recall, which is reflected in a higher AP value, defined as:

$$\mathrm{AP} = \sum_n (r_{n+1} - r_n) \cdot \max_{\{i | r_i \geq r_{n+1}\}} p_i \tag{27}$$

where r_n and r_{n+1} are adjacent recall values, and p_i represents the maximum precision value for recall values greater than or equal to r_{n+1}.

Mean Average Precision (mAP) extends AP to multi-class detection scenarios by averaging the AP scores of all classes. It provides a single numeric metric to evaluate the overall performance of a detector across different object categories. For each class, the AP value is calculated independently following the same procedure as the single-class AP, and then the values are averaged to obtain the mAP. This metric is particularly useful when comparing different detection models. A higher mAP value indicates a better overall detection performance across all classes. It is defined as:

$$\mathrm{mAP} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{AP}_i \tag{28}$$

where N is the total number of classes and AP_i represents the Average Precision for the i-th class.

IV. EXPERIMENTAL ANALYSIS

A. EXPERIMENTAL PARAMETER SETTINGS

In the experiments of RetailDet, we select YOLOv11 as the baseline model to achieve an optimal balance between precision and efficiency. The model is trained using the Adam optimizer [52] for 700 epochs, with a weight decay of 1e-4, an initial learning rate of 1e-3, and a final learning rate of 1e-4, implementing a warm-up strategy. For data augmentation, we employ a composite strategy comprising Mosaic augmentation, geometric transformations (including translation and scaling), horizontal flipping, RandAugment (with color jittering, contrast/brightness adjustments), random erasing, and full-image cropping. Following the initial training settings of DenseDet and RTMDet, we maintain the input image resolution at 640 × 640 to ensure consistency across models. All models are trained on an AMD Ryzen Threadripper PRO S97WX CPU and 3 NVIDIA A6000 GPUs. We evaluate our model using the mean metrics of the COCO format [53], including mAP_{50} (IoU = 0.50), mAP_{75} (IoU = 0.75), mAP (IoU = .50:.05:.95), Precision, Recall, and F1, to assess the detection capability under different scenarios.

B. EXPERIMENTAL RESULTS

1) COMPARATIVE EXPERIMENTAL RESULTS

Table 1 shows the performance comparison of YOLO models of various scales and current SOTA object detection models on the RPV1K dataset. RetailDet achieves SOTA performance of the same generation at all scales (including nano, small, medium, and large). RetailDet (medium) achieves 66.3% in terms of mean average precision (mAP), surpassing many similar models. RetailDet (small) achieves 82.5% in terms of mAP_{50}, leading the same scale models and even surpassing many large YOLO models. RetailDet-large achieves the current SOTA performance with a mAP of 67.7%, significantly surpassing all YOLO variant models, including the YOLOv12-large model (63.1%) and the YOLOv11-large model (66.4%). It also achieves significant improvements in mAP_{50} (83.3%) and mAP_{75} (74.60%), greatly exceeding previous SOTA models. The model achieves the highest precision of 85.2%, surpassing the YOLOv11-large (84.8%) and the YOLOv12-large (83.7%), while maintaining a competitive recall rate of 76.2%. The F1 of 80.4% demonstrates its robust overall performance, significantly outperforming mature models such as the YOLOv8-large (78.08%) and RT-DETR (78.8%), indicating its superior ability in handling false positives and false negatives. This can be attributed to the combined effects of multiple attention mechanisms in the backbone network, the special scale transformation structure of MUD-FPN, and the IGR of $\mathbf{N}_{24}$.

In terms of computational efficiency, RetailDet (large) demonstrates remarkable advantages while maintaining excellent detection performance. The model only requires 20.79 million parameters, representing an 81% reduction compared to the YOLOv6-large (109.56 million parameters) and a 47% reduction compared to the YOLOv8-large (39.43 million parameters). Its inference latency of 25.1 milliseconds is significantly lower than YOLOv8 through YOLOv12 and other models. Compared with specialized detection models, RetailDet's mAP is 14.9% higher than that of DenseDet, while using only 28% of its parameters (20.79 M vs. 74.14 M). Compared with RT-DETR-L, the parameters are reduced by 35.0% (20.79 M vs. 31.99 M), and the latency is reduced by 14.63% (25.1 ms vs. 29.4 ms). The model has a computational volume of 140.8 GFLOPs, with moderate computational requirements, achieving the best balance between performance and efficiency, suitable for edge computing deployment.

From the ablation experimental data of RetailDet and $\mathbf{N}_{24}$ modules in Table 1, we can verify the significant role of the MUD-FPN module. Compared with the baseline model RetailDet-large (mAP_{75} = 74.6%), the RetailDet-$\mathbf{N}_{24}$ version without the module shows a decreased mAP_{75} of 74.38%, indicating that the absence of the module reduces detection accuracy in high IoU (IoU = 0.75) scenarios. In contrast, RetailDet+$\mathbf{N}_{24}$ version with the module achieves an increased mAP_{75} of 74.74%, surpassing the baseline model. This comparative analysis demonstrates that the MUD-FPN module enhances the model's detection capability for target boundary positioning and multi-scale objects through optimized feature fusion. The module serves as a crucial component for improving high-precision detection performance in complex scenarios, playing an essential role in maintaining and enhancing key detection metrics.

2) RESULTS VISUALIZATION AND GRAPHIC ANALYSIS

Fig. 8 visualizes the comprehensive performance analysis from three perspectives, highlighting that all scales of RetailDet(nano/small/medium/large) exhibit comprehensive SOTA advantages. For ease of comparison, we present the left figure, i.e., Fig. 1, together with the other two figures. In the left-side latency-accuracy trade-off analysis and middle-side parameter-accuracy comparison, every scale of RetailDet occupies the optimal frontier of the efficiency-performance curve. Whether the nano scale with minimal parameters, the small scale with balanced latency, or the medium/large scales, they all achieve superior detection accuracy at the lowest corresponding computational cost, surpassing YOLO variants, RT-DETR, and other models. Taking the large-scale RetailDet-L as an example, the right-side training metrics visualization reveals convergence characteristics across 700 epochs: precision and recall steadily improve to 85.2% and 76.2%, respectively. Throughout training and validation, the model maintains low-stable loss values, while the smooth convergence of mAP_{50} (83.3%) and mAP (74.60%) curves further proves its robust learning ability and generalization performance. These visualizations collectively validate that all RetailDet scales achieve an excellent balance in accuracy, computational efficiency, and model complexity, showcasing unparalleled SOTA dominance.

3) ABLATION STUDY

To verify the effectiveness of each component, we conduct a comprehensive ablation study on the RPV1K dataset. Table 2 presents results from four different configurations: baseline model (Case1), and progressive incorporation of RGB-D

TABLE 1. Comprehensive comparison of YOLO variants and other SOTA object detectors on the RPV1K test set (Grouped by scales; - indicates no scale breakdown available. Swin-T/B indicates the model uses Swin Transformer as the backbone network).

Model	Scales	mAP (%)	mAP_{50} (%)	mAP_{75} (%)	Precision (%)	Recall (%)	F1 (%)	Latency (ms)	Params (M)	FLOPs (G)
YOLOv5	nano	50.8	72.7	57.26	75.9	63.9	69.38	8.7	2.18	5.8
YOLOv6	nano	53.1	74.6	59.59	80.3	65.5	72.14	10.6	4.15	11.5
YOLOv8	nano	52.3	74.1	59.36	78.5	65.6	71.53	13.9	2.68	6.8
YOLOv9	tiny	47.3	68.6	51.87	71.7	63.6	67.40	36.8	1.73	6.4
YOLOv10	nano	49.2	68.5	55.87	73.5	60.3	66.25	35.1	2.7	8.2
YOLOv11	nano	51.9	73.6	58.65	78.5	65.1	71.23	33.9	2.58	6.3
YOLOv12	nano	53.0	75.2	60.56	80.9	66.5	72.99	14.3	2.55	6.3
Faster RCNN SwinT [16]	tiny	48.6	67.6	55.8	48.6	55.0	52.0	52	44.75	93.3
Cascade RCNN SwinT [54]	tiny	50.8	67.4	57.5	50.8	56.1	53.3	41.4	72.6	121
RetailDet(Ours)	**nano**	**62.5**	**81.0**	**69.8**	**84.9**	**72.0**	**77.92**	**16.7**	2.71	14.8
YOLOv5	small	57.9	78.3	65.40	81.5	69.6	75.08	9.9	7.81	18.7
YOLOv6	small	58.6	78.5	65.44	82.1	71.8	76.61	11.6	15.97	42.8
YOLOv8	small	58.5	78.7	65.67	82.8	68.7	75.09	16.3	9.82	23.3
YOLOv9	small	58.1	78.7	64.74	80.3	71.2	75.47	53.1	6.19	22.1
YOLOv10	small	56.1	75.6	62.81	77.6	67.7	72.31	36.8	8.04	24.4
YOLOv11	small	59.6	79.8	67.27	83.0	72.2	77.36	35.0	9.41	21.3
YOLOv12	small	58.9	79.8	66.18	81.6	72.4	76.72	16.5	9.23	21.2
RetailDet(Ours)	**small**	**65.9**	**82.5**	**72.67**	**85.4**	**75.1**	**79.92**	**18.7**	9.68	43
YOLOv5	medium	60.7	80.0	67.45	85.1	70.8	77.29	13.5	22.12	52.5
YOLOv6	medium	61.6	80.0	68.40	82.8	72.2	77.14	18.6	51.25	158.3
YOLOv8	medium	61.7	80.7	68.78	83.1	72.6	77.50	22.5	23.2	67.4
YOLOv9	medium	61.4	81.0	68.36	81.9	73.8	77.64	54.8	16.57	76.5
YOLOv10	medium	60.1	78.2	67.47	80.3	70.8	75.25	44.5	16.45	63.4
YOLOv11	medium	62.6	81.0	69.16	83.0	73.1	77.83	46.4	20.03	67.7
YOLOv12	medium	61.4	80.7	68.77	80.8	74.6	76.62	21.2	20.11	67.1
RetailDet(Ours)	**medium**	**66.3**	**82.1**	**73.29**	**85.2**	**74.9**	**79.72**	**20.8**	**14.86**	81.6
YOLOv5	large	63.3	80.8	69.88	83.2	74.2	78.44	23.9	47.91	114.2
YOLOv6	large	63.1	81.0	69.68	81.5	74.8	78.06	23.6	109.56	386.1
YOLOv8	large	64.0	80.9	70.92	84.6	72.5	78.08	36.3	39.43	145.2
YOLOv9	large	63.1	81.1	70.27	83.4	74.0	78.42	59.3	21.15	**82.7**
YOLOv10	large	64.6	78.8	70.98	82.6	72.2	77.0	55.8	25.72	126.3
YOLOv11	large	66.4	82.1	73.35	84.8	74.4	79.26	58.2	25.28	86.6
YOLOv12	large	63.1	81.9	70.47	83.7	75.0	79.05	37.3	26.34	88.6
RT-DETR [17]	large	64.7	81.5	72.07	81.6	76.2	78.8	29.4	31.99	103.4
DenseDet [34]	large	52.8	76.4	62.30	52.8	64.5	58.06	73.5	74.14	166
RTMDet	large	39.9	65.6	47.70	39.9	58.7	47.5	31.1	52.26	103.4
MAE-YOLO [36]	large	60.8	79.6	67.41	81.2	73.4	77.0	26.0	30.56	127.6
SSD-512	large	30.6	63.7	29.0	30.6	43.7	35.99	29.1	24.53	87.7
Faster RCNN Resnet101 [15]	large	40.4	63.1	44.5	40.4	47.7	43.75	45.5	60.35	121
Faster RCNN SwinB [16]	base	49.0	67.3	56.2	49.0	56.0	52.3	66.2	104	186
RetailDet(Ours)	**large**	**67.7**	**83.3**	**74.60**	**85.2**	76.2	**80.4**	25.1	20.79	140.8
RetailDet - N_{24}	**large**	**67.1**	**82.4**	**74.38**	**83.8**	**76.6**	**80.03**	**25.6**	**20.24**	**152.3**
RetailDet + N_{24}	**large**	**67.0**	**82.4**	**74.74**	**83.7**	**75.4**	**79.33**	**25.9**	**25.53**	**139.9**

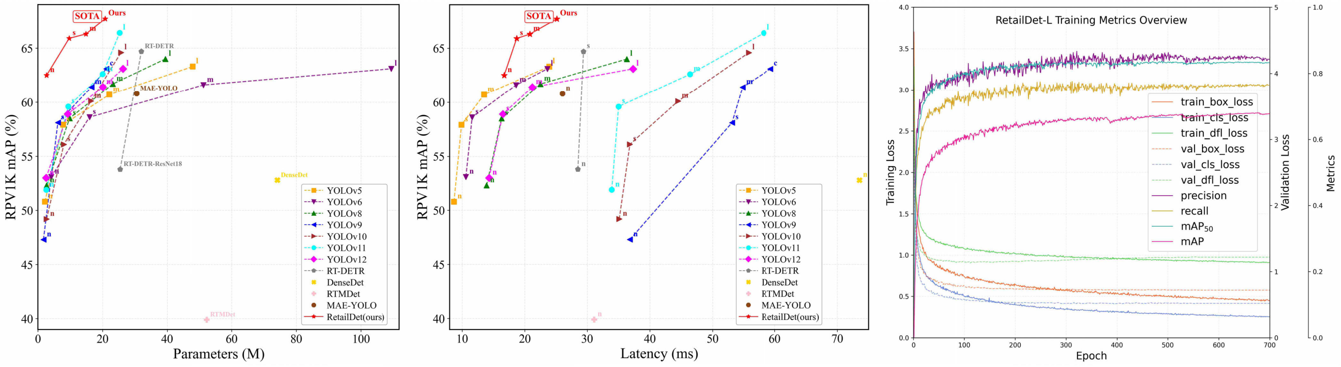


FIGURE 8. Comparison of detection models on the RPV1K dataset: Parameter-mAP (left), latency-mAP (center), and RetailDet-L (large) model training metric overview (right).

bimodal feature integration (Case2), MUD-FPN (Case3), and PKIB (Case4). The baseline model achieves a strong mAP of 66.4% with mAP_{50} and mAP_{75} reaching 82.1% and 73.35%. Adding RGB-D bimodal feature integration

TABLE 2. Performance comparison of different model configurations (✓indicates that the corresponding module is utilized).

modal	baseline(YOLOv11)	RGB-D bimodal	MUD-FPN	PKIB	mAP	mAP$_{50}$	mAP$_{75}$	Precision	Recall	F1
Case1	✓				66.4	82.1	73.35	84.8	74.4	79.26
Case2	✓	✓			65.7	81.5	72.09	84.8	75.6	79.3
Case3	✓	✓	✓		65.9	82.6	73.03	84.8	75.8	80.0
Case4	✓	✓	✓	✓	**67.7**	**83.3**	**74.6**	**85.2**	**76.2**	**80.4**

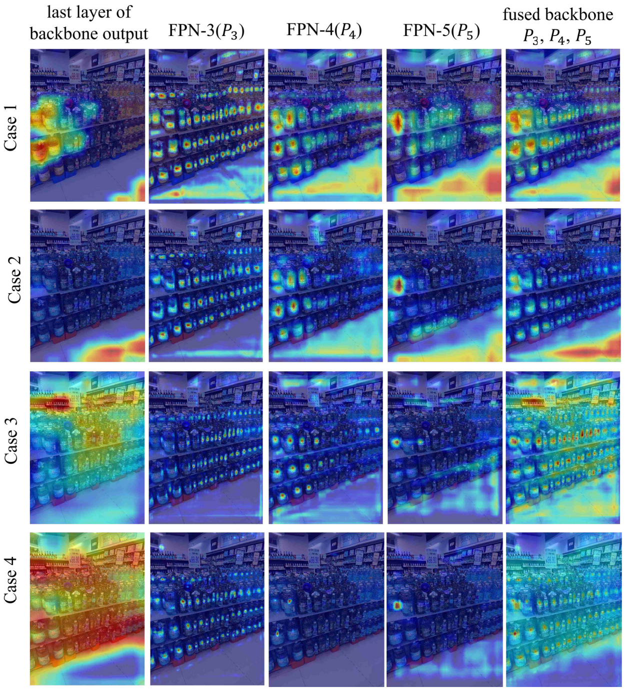

last layer of backbone output
FPN-3(P_3)
FPN-4(P_4)
FPN-5(P_5)
fused backbone P_3, P_4, P_5
Case 1
Case 2
Case 3
Case 4

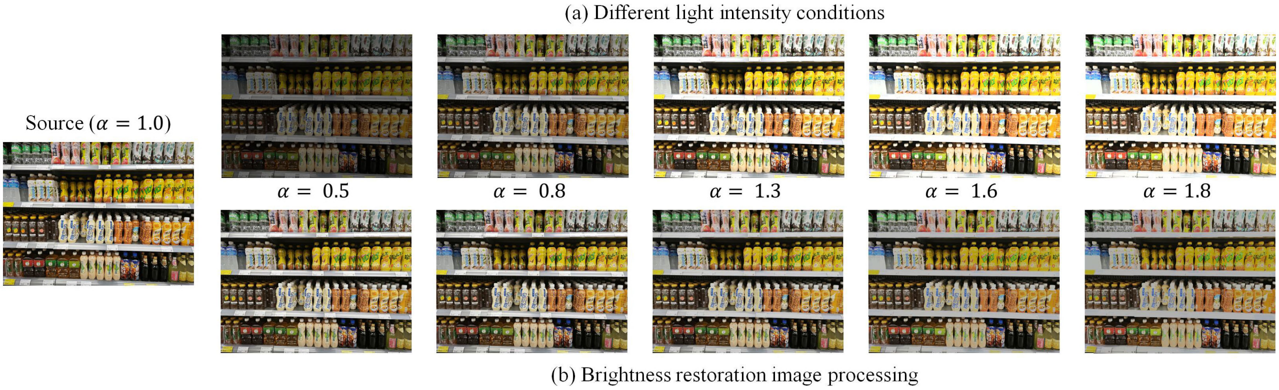
FIGURE 9. The feature heatmaps of the ablation experiments (Case 1 to Case 4) show a gradual change: background distractions (e.g., ceilings, floors) are increasingly suppressed, and the model focuses more on the primary targets (shelves). The heat responses in non-shelf areas weaken, while target-related features become more prominent and precise. These heatmaps visually supplement the ablation experiment results in Table 2, intuitively reflecting the numerical trends.

(Case2) maintains stable performance with an mAP of 65.7%, while improving the recall rate from 74.4% to 75.6%, demonstrating its effectiveness in balancing detection metrics. Integration of MUD-FPN (Case3) further enhances overall performance, with mAP slightly improving to 65.9% and precision maintaining at 84.8%, while the recall rate increases to 75.8%, resulting in an improved F1 of 80.0%. The complete model (Case4) with PKIB achieves optimal performance across all metrics: mAP increases to 67.7%, mAP$_{50}$ improves to 83.3%, and mAP$_{75}$ reaches 74.60%, with the highest precision of 85.1% and best recall rate of 76.2%, yielding an optimal F1 of 80.4%. The progressive improvements from Case1 to Case4 demonstrate our components' complementary nature. While the baseline model already shows strong performance, each additional component contributes to refined detection capability, with PKIB providing the final boost in overall performance. The consistent enhancement in both precision and recall metrics validates the effectiveness of our architectural design choices for retail scenarios.

Algorithm 2 Brightness Restoration Image Processing

```
Require: Input image $I$ with dimensions $W \times H$; RPV1K dataset $D$ containing $N$ images
Ensure: Enhanced image $I_{adj}$ with normalized brightness
Stage 1: Calculate target brightness
Initialize brightness Values list
for each image $I_{d_i} \in D$ do
    $I_{d_{HSV}} \leftarrow \text{RGB2HSV}(I_d)$    ▷ convert RGB to HSV
    $B_{d_i} \leftarrow \frac{1}{W \times H} \sum_{x,y} I^V_{d_{HSV}}(x, y)$    ▷ average brightness
    Add $B_{d_i}$ to brightness Values list
end for
$B_t \leftarrow \frac{1}{N} \sum_{i=1}^{N} B_{d_i}$    ▷ dataset average brightness
Stage 2: Brightness detection and adjustment
$I_{HSV} \leftarrow \text{RGB2HSV}(I)$    ▷ convert input image to HSV
$B_c \leftarrow \frac{1}{W \times H} \sum_{x,y} I^V_{HSV}(x, y)$    ▷ current brightness
$\alpha \leftarrow \frac{B_t}{B_c}$    ▷ Calculate adjustment factor
$I^V_{HSV} \leftarrow \text{Clip}(I^V_{HSV} \times \alpha, 0, 255)$    ▷ adjust brightness
$I_{adj} \leftarrow \text{HSV2RGB}(I_{HSV})$    ▷ convert back to RGB
return $I_{adj}$    ▷ return brightness-adjusted image
```

From the feature heatmaps (in Fig. 9), we observe that the model's focus on key areas becomes more precise and comprehensive as components are progressively integrated (Case 1 to Case 4). In Case 1, the backbone network's last layer heatmap captures the main targets, but with incomplete coverage. Case 2, incorporating RGB-D bimodal feature integration, shows heatmaps across FPN layers (FPN-3, FPN-4, FPN-5) with more consistent coverage of shelf products, demonstrating RGB-D bimodality's role in enhancing feature perception. With MUD-FPN addition in Case 3, the heatmap exhibits more uniform response in target areas and clearer feature capture at edges and details, indicating improved multi-scale feature fusion and transfer. Case 4's PKIB integration results in the fused heatmap (backbone and FPN combined) showing more complete highlighted coverage of shelf products, with enhanced feature complementarity among FPN layers, validating PKIB's effectiveness in strengthening feature associations and utilization efficiency. The heatmap evolution from Case 1 to Case 4 shows progression from local, single feature responses to global, multi-dimensional, and precise feature coverage, verifying each component's contribution to feature capture and fusion optimization. This qualitative progression aligns with the quantitative improvements shown in Table 2, confirming the



(a) Different light intensity conditions

(b) Brightness restoration image processing

FIGURE 10. Image processing visualization of different light intensity coefficients.

TABLE 3. Performance comparison of various models under different light intensity conditions without brightness restoration image processing. (The overall average brightness of the original dataset is set to 1.0. A smaller α value indicates darker lighting, while a larger value indicates higher brightness).

Model	Scale	$\alpha = 0.5$			$\alpha = 0.8$			$\alpha = 1.3$			$\alpha = 1.6$			$\alpha = 1.8$		
		mAP	mAP$_{50}$	F1	mAP	mAP$_{50}$	F1	mAP	mAP$_{50}$	F1	mAP	mAP$_{50}$	F1	mAP	mAP$_{50}$	F1
YOLOv5	nano	48.8	70.5	68	50.6	72.6	69.7	49.8	71.8	69.3	48.3	70.6	68.0	47.1	69.2	67.1
YOLOv6	nano	52	73.5	71.3	53	74.5	72.3	49.8	71.8	69.3	48.3	70.6	68.0	47.1	69.2	67.1
YOLOv8	nano	51.3	73.2	70.4	52.4	74.2	71.3	51.6	73.4	70.8	50	72.3	69.7	48.8	72	69.3
YOLOv9	nano	46.6	68	66.8	47.5	69	67.6	46.4	67.2	66.6	45.1	66.3	65.9	44	64.8	64.6
YOLOv10	nano	48	67.4	64.9	49.2	68.6	66.1	48.6	68.1	65.9	47.1	66.8	65.0	45.9	65.9	63.9
YOLOv11	nano	50.1	71.8	69.5	51.9	73.5	70.9	51.3	72.9	70.5	49.7	71.7	69.9	48.6	70.5	68.5
YOLOv12	nano	50.9	73.2	70.9	52.9	75	72.4	52.3	74.7	72.8	51	73.5	72.2	49.9	72.5	71.2
Faster RCNN SwinT	tiny	46.8	66.3	49.8	48.4	67.9	51.4	47.1	66.6	50.2	44.6	64.5	47.5	43.6	63	46.5
Cascade RCNN SwinT	tiny	49.5	66.9	52.1	50.7	67.5	53.2	49.1	66	51.6	46.4	64	49.0	44.9	63.4	47.6
RetailDet(Ours)	nano	**59**	**77.7**	**75.2**	**62.3**	**80.9**	**77.7**	**61.6**	**80.5**	**77.5**	**60.3**	**79.6**	**76.8**	**59.3**	**79**	**76.3**
YOLOv5	small	56.4	77.2	73.7	57.8	78.2	74.7	57.3	77.8	74.7	55.5	76.5	73.4	53.9	75.3	72.5
YOLOv6	small	57.8	77.7	64.9	58.5	78.4	76.6	58.2	78.5	76.5	56.9	77.5	75.7	55.6	76.4	74.8
YOLOv8	small	57.5	77.7	74.6	58.5	78.7	74.7	58	78.4	75.1	56.6	77.6	74.8	55.5	76.6	73.8
YOLOv9	small	56.9	77.3	74.1	57.9	78.5	75.1	57.7	78.5	75.3	56.1	77.5	74.3	54.7	76.3	73.45
YOLOv10	small	55.3	74.7	71.2	56.2	75.7	71.9	55.5	75.1	72.1	54.2	74.3	71.3	52.9	73.1	70.4
YOLOv11	small	58.2	**78.7**	**75.9**	59.4	79.7	76.9	59	79.2	76.1	57.6	78.3	75.9	56.3	77.3	74.8
YOLOv12	small	56.4	77.8	75.0	58.6	79.5	76.6	58.2	79.1	76.3	56.7	78.2	75.6	55.5	77.1	74.7
RetailDet(Ours)	small	**60.2**	76.7	75.2	**65.1**	**81.9**	**79.1**	**64.6**	**81.3**	**78.8**	**62.1**	**79.2**	**77.1**	**60.7**	**78.2**	**76.4**
YOLOv5	medium	59.9	79.6	76.5	60.7	80.1	77.4	60.3	79.9	77.4	59.2	79.3	76.7	58.1	78.7	76.6
YOLOv6	medium	60.9	79.2	76.9	61.7	80.1	77.5	61	79.6	77.3	59.8	78.7	76.6	58.6	77.9	75.8
YOLOv8	medium	60.5	80.1	76.7	61.6	80.7	77.5	61.2	80.7	77.2	59.7	79.7	76.4	58.6	78.9	75.8
YOLOv9	medium	60.6	79.9	76.9	61.5	80.9	77.5	60.9	80.4	77.1	59.4	79.4	76.1	58.2	78.5	75.7
YOLOv10	medium	58.9	77.1	74.5	60.1	78.2	75.3	59.5	77.7	74.6	58.0	77.0	74.5	56.7	75.8	73.5
YOLOv11	medium	**62.0**	**80.8**	**77.3**	62.8	**81.4**	78.4	61.9	80.3	77.6	60.5	79.5	76.5	59.6	79.0	75.9
YOLOv12	medium	60.2	80.1	77.1	61.4	80.6	77.7	60.3	80.1	77.2	59.0	79.3	76.7	57.9	78.3	75.9
RetailDet(Ours)	medium	61.2	77.5	75.8	**65.5**	81.2	**79.3**	**65.6**	**81.4**	**79.0**	**63.6**	**80.1**	**78.0**	**62.1**	**79.0**	**77.0**
YOLOv5	large	62.3	79.8	77.7	63.1	80.7	78.2	62.8	80.5	78.1	61.7	79.8	77.5	60.7	79.1	77.1
YOLOv6	large	62.1	80.3	78.1	62.9	80.8	78.1	62.9	81.1	78.2	61.5	79.9	77.3	60.4	79.1	76.5
YOLOv8	large	63.5	80.6	78.1	63.9	80.8	78.2	63.4	80.6	78.1	62.2	79.9	77.9	61.2	79.3	77.1
YOLOv9	large	62.9	81.0	78.3	63.3	81.3	78.4	62.4	80.6	78.5	60.9	79.6	77.5	59.8	78.8	77.1
YOLOv10	large	65.3	80.1	77.8	65.3	79.9	77.8	63.4	77.5	76.0	61.7	75.8	74.7	60.8	75.3	74.1
YOLOv11	large	**65.9**	**82.5**	**78.9**	66.5	**82.9**	79.0	65.8	82.4	78.9	**64.5**	**81.7**	78.5	**63.6**	**81.3**	**78.0**
YOLOv12	large	61.1	80.2	77.7	62.9	81.7	78.7	61.9	80.9	78.9	60.4	79.8	77.5	59.4	79.4	77.2
RT-DETR	large	64.1	81.0	78.4	64.7	81.6	78.8	64.4	81.6	78.8	63.2	81.1	**78.5**	62.2	80.3	77.6
DenseDet	large	46.0	69.5	52.1	48.2	71.6	54.3	48.3	71.6	54.0	46.0	69.2	51.6	44.7	68.1	50.2
RTMDet	large	39.2	64.4	48.2	40.3	65.7	49.8	38.9	64.6	48.5	37.4	62.3	46.7	36.4	61.0	45.5
MAE-YOLO	large	60.8	79.9	77.4	61.1	79.9	77.3	60.3	79.3	76.6	58.6	77.9	76.0	57.2	76.8	75.2
SSD-512	large	29.6	61.7	35.9	31.1	63.8	37.6	28.9	62.2	35.2	27.3	59.7	33.4	26.3	58.1	32.1
Faster RCNN Resnet101	large	37.7	60.5	41.5	40.0	62.4	43.5	39.5	61.8	43.2	37.7	59.4	40.9	36.6	58.5	40.0
Faster RCNN SwinB	large	48.2	67.0	51.7	49.2	68.2	52.6	47.6	65.9	51.0	45.5	63.8	48.9	44.0	62.4	47.4
RetailDet(Ours)	**large**	62.4	78.4	76.3	**66.9**	82.4	**80.1**	**66.5**	**82.6**	**80.2**	64.2	80.6	78.2	62.5	79.3	77.0
RetailDet - $\mathbf{N}_{24}$	large	60.9	76.9	75.4	66.1	81.8	79.7	66.2	81.8	79.7	63.8	80.3	78.3	62.3	79.2	77.3
RetailDet + $\mathbf{N}_{24}$	large	61.6	76.9	75.4	66.1	81.5	78.8	65.5	81.3	79.1	63.2	79.6	77.5	61.7	78.3	76.3

effectiveness of the model architecture design for feature perception in retail scenarios.

4) ROBUSTNESS TO LIGHT INTENSITY

To replicate diverse lighting scenarios, we implement an illumination variation architecture operating within the HSV color space [55], as illustrated in Fig. 10 (a) with different light intensity conditions. This methodology manipulates brightness while maintaining color integrity by exclusively modifying the Value (V) channel as follows:

$$V'(x, y) = \mathrm{Clip}(V(x, y) \cdot \alpha, 0, 255) \tag{29}$$

TABLE 4. Performance comparison of various models under different lighting intensity conditions when using brightness restoration image processing.

Model	Scale	$\alpha = 0.5$			$\alpha = 0.8$			$\alpha = 1.3$			$\alpha = 1.6$			$\alpha = 1.8$		
		mAP	mAP$_{50}$	F1	mAP	mAP$_{50}$	F1	mAP	mAP$_{50}$	F1	mAP	mAP$_{50}$	F1	mAP	mAP$_{50}$	F1
YOLOv5	nano	50.5	72.6	69.6	50.6	72.6	69.4	50.1	72.2	69.1	48.6	70.8	68.5	47.5	69.3	67.1
YOLOv6	nano	53.1	74.5	71.7	53.1	74.6	72.1	52.6	74.3	72.4	50.9	72.8	71.1	49.4	71.3	69.9
YOLOv8	nano	52.3	74.1	71.5	52.2	74.1	71.5	51.9	73.6	71.2	50.6	72.9	70.6	49.3	71.8	69.5
YOLOv9	nano	47.2	68.5	67.3	47.3	68.6	67.1	46.9	68.1	67.0	45.6	66.8	65.9	44.4	65.4	64.3
YOLOv10	nano	49.3	68.6	65.9	49.3	68.6	66.0	48.9	68.3	66.0	47.4	67.1	64.7	46.1	65.8	64.0
YOLOv11	nano	51.9	73.4	71.0	52.0	73.5	71.1	51.3	73.0	71.0	49.9	71.8	69.9	48.7	70.6	68.8
YOLOv12	nano	52.9	75.0	73.1	52.9	75.0	73.2	52.5	74.8	72.5	51.0	73.3	71.1	49.6	71.6	69.5
Faster RCNN SwinT	tiny	48.4	67.9	51.4	48.6	68.0	51.6	47.2	66.7	50.2	45.3	64.8	48.3	43.8	63.3	46.7
Cascade RCNN SwinT	tiny	50.4	66.9	52.9	50.4	67.4	53.0	49.8	66.9	52.4	47.3	65.0	49.9	45.8	63.0	48.3
RetailDet(Ours)	**nano**	**62.4**	**81.0**	**78.0**	**62.5**	**81.1**	**77.9**	**62.2**	**80.9**	**78.1**	**61.1**	**80.4**	**77.4**	**60.4**	**79.6**	**77.1**
YOLOv5	small	57.8	78.2	75.1	57.8	78.1	75.0	57.4	77.8	74.7	55.5	76.4	73.3	54.1	75.3	72.6
YOLOv6	small	58.7	78.6	76.7	58.6	78.5	76.5	58.3	78.4	76.3	56.4	76.8	75.2	55.0	75.3	74.3
YOLOv8	small	58.4	78.6	74.8	58.3	78.5	74.8	58.0	78.2	74.6	56.6	77.2	73.6	55.3	76.0	72.5
YOLOv9	small	58.0	78.7	75.6	58.0	78.8	75.5	57.9	78.1	75.6	56.4	77.5	74.5	55.0	76.3	73.5
YOLOv10	small	56.2	75.7	72.1	56.2	75.7	72.1	55.8	75.3	72.1	54.4	74.1	70.7	53.1	72.9	69.9
YOLOv11	small	59.6	79.8	77.1	59.5	79.7	77.0	59.0	79.5	76.9	57.7	78.4	75.7	56.4	77.3	74.9
YOLOv12	small	58.8	79.7	76.6	58.8	79.6	76.5	58.3	79.3	76.3	56.7	77.8	75.1	55.4	76.6	74.7
RetailDet(Ours)	**small**	**65.8**	**82.3**	**79.8**	**65.7**	**82.3**	**79.6**	**65.3**	**82.1**	**79.8**	**64.0**	**81.0**	**78.3**	**63.1**	**80.2**	**77.9**
YOLOv5	medium	60.6	79.9	77.5	60.6	80.0	77.3	60.4	80.1	77.5	59.2	79.2	76.5	58.0	78.3	76.1
YOLOv6	medium	61.5	79.8	77.1	61.5	79.8	77.0	61.1	79.6	77.2	59.5	78.4	76.7	58.1	77.6	75.7
YOLOv8	medium	61.7	80.7	77.4	61.6	80.7	77.3	61.6	80.7	77.3	61.4	80.8	77.6	59.8	79.8	76.5
YOLOv9	medium	61.4	81.1	77.5	61.4	81.0	77.4	61.0	80.7	77.6	59.3	79.4	76.6	58.1	78.5	75.8
YOLOv10	medium	60.0	78.1	75.3	60.1	78.1	75.2	59.7	77.9	74.8	58.2	77.0	74.5	56.9	75.9	74.0
YOLOv11	medium	62.6	81.2	78.3	62.7	81.1	78.2	62.1	80.6	78.0	60.5	79.2	76.5	59.4	78.4	75.7
YOLOv12	medium	61.3	80.4	77.4	61.3	80.4	77.3	61.0	80.4	77.3	59.3	79.5	76.6	58.1	78.2	75.6
RetailDet(Ours)	**medium**	**66.0**	**81.6**	**79.2**	**66.2**	**81.9**	**79.1**	**66.0**	**81.8**	**79.5**	**64.9**	**80.9**	**78.6**	**63.9**	**80.2**	**77.8**
YOLOv5	large	63.2	80.8	78.6	63.2	80.7	78.6	62.9	80.5	78.1	61.8	79.9	77.5	60.6	79.1	76.6
YOLOv6	large	63.0	80.9	77.9	63.0	80.9	77.8	62.7	80.7	78.3	61.1	79.7	76.8	59.5	78.3	75.9
YOLOv8	large	63.8	80.7	78.5	63.9	80.7	78.3	63.5	80.8	78.0	62.3	80.1	77.6	61.1	79.0	76.9
YOLOv9	large	63.1	81.0	78.3	63.0	81.0	78.4	62.7	80.8	78.4	61.1	79.5	77.4	60.0	78.8	76.9
YOLOv10	large	64.7	79.0	77.1	64.7	79.1	77.1	62.8	76.6	75.1	61.4	75.4	73.9	60.2	74.6	72.9
YOLOv11	large	66.1	82.5	79.2	66.2	82.7	79.2	65.9	82.6	79.0	64.5	81.7	78.3	63.5	81.2	77.7
YOLOv12	large	62.8	81.6	78.9	62.8	81.6	78.9	62.3	81.2	78.5	60.7	80.2	77.8	59.5	79.5	76.9
RT-DETR	large	64.6	81.6	79.1	64.7	81.5	78.8	64.4	81.6	79.0	63.0	80.9	78.4	61.8	80.1	77.6
DenseDet	large	48.7	71.8	54.7	48.8	71.9	54.8	47.9	71.0	53.9	48.1	69.4	51.8	44.7	67.9	50.5
RTMDet	large	39.8	65.4	49.6	39.7	65.4	49.5	39.6	65.1	49.1	38.1	62.9	47.4	37.3	61.4	46.2
MAE-YOLO	large	60.8	79.7	77.0	60.8	79.7	76.9	60.5	79.7	76.8	58.8	78.0	75.9	57.3	77.0	75.4
SSD-512	large	30.3	63.5	36.8	30.3	63.7	36.8	30.0	62.8	36.3	29.1	60.7	35.2	28.3	59.5	34.3
Faster RCNN Resnet101	large	40.3	62.8	43.6	40.3	62.3	43.6	39.6	62.1	42.9	37.9	60.0	41.4	36.7	58.8	40.1
Faster RCNN SwinB	large	48.9	67.5	52.2	48.9	67.6	52.3	48.2	67.0	51.5	46.2	65.1	49.6	44.8	63.6	48.1
RetailDet(Ours)	**large**	**67.5**	**83.1**	**80.5**	**67.5**	**83.0**	**80.4**	**67.2**	**83.0**	**80.6**	**66.0**	**82.3**	**80.1**	**65.1**	**81.6**	**79.1**
RetailDet - N$_{24}$	**large**	67.0	82.4	80.2	67.0	82.3	80.2	66.7	82.2	79.9	65.2	81.2	79.2	64.2	80.5	78.3
RetailDet + N$_{24}$	**large**	66.7	82.1	79.4	66.8	82.2	79.4	66.4	82.2	79.1	65.3	81.5	78.6	64.4	80.7	77.9

In this formulation, the parameter α represents the illumination coefficient ($\alpha \in [0.0, 2.0]$). We establish a comprehensive illumination spectrum comprising seven distinct levels: severely underexposed ($\alpha \leq 0.5$), underexposed ($\alpha \in (0.5, 0.8]$), slightly dim ($\alpha \in (0.8, 1.0]$), standard illumination ($\alpha = 1.0$), slightly bright ($\alpha \in (1.0, 1.3]$), overexposed ($\alpha \in (1.3, 1.6]$) and severely overexposed ($\alpha \in (1.6, 2.0]$), with these lighting variations demonstrated in Fig. 10 (a). This stratified approach facilitates systematic evaluation under controlled yet realistic lighting conditions. The Clip operation ensures that the transformed values $V'(x, y)$ remain within the valid intensity range of [0, 255]. Using brightness adjustment techniques in HSV color space, we evaluate RetailDet's robustness under varying illumination conditions across seven datasets with different illumination intensities. The comparative performance results across different light intensities are detailed in Table 3.

To further validate our approach, we use the standard illumination of the original dataset as a benchmark and apply the brightness recovery algorithm described in Algorithm 2 to datasets under various lighting conditions. The performance evaluation results after recovery processing are presented in Table 4, demonstrating the effectiveness of our proposed brightness recovery method.

Based on the experimental data in Table 3, our RetailDet model demonstrates good performance across different scales and lighting conditions. At the nano scale, when $\alpha = 0.8$, RetailDet achieves mAP = 62.3% and F1 = 77.7%, outperforming YOLOv11 at the same scale (mAP = 51.9%, F1 = 70.9%). At small and medium scales, RetailDet maintains its performance advantage, reaching mAP = 65.1% and 65.5% respectively at $\alpha = 0.8$. Particularly in the large-scale comparison, although YOLOv11 achieves a better F1 (F1 = 82.9%) than RetailDet (F1 = 80.1%) at $\alpha = 0.8$, in terms of lighting adaptability, when illumination changes from $\alpha = 0.8$ to $\alpha = 0.5$, YOLOv11's mAP drops from 66.5% to 65.9%, while RetailDet's drops from 66.9% to 62.4%; when illumination increases to $\alpha = 1.8$, YOLOv11 decreases to

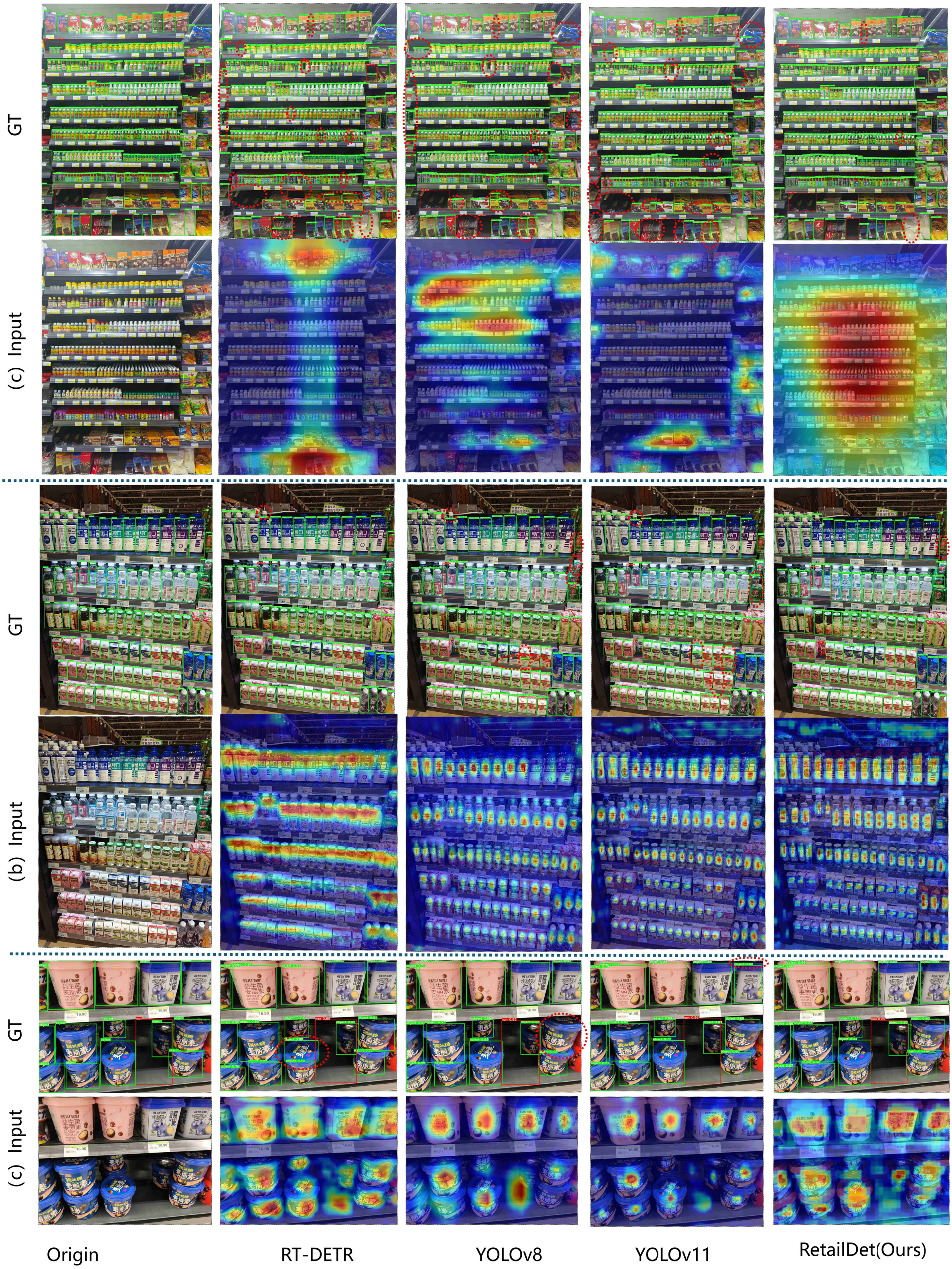


FIGURE 11. The detection results and heatmap visualizations of different models on the RPV1K dataset are as follows. Here, GT (ground truth) represents the real annotation information of the targets. From the feature heatmaps of densely arranged targets and long-tailed targets, RetailDet can better reflect the characteristics of long-tailed outer packaging, with fewer missed detections and accurate positioning. The dashed circles mark the areas with problems such as false positives, missed detections, and inaccurate bounding boxes.

63.6%, and RetailDet to 62.5%. These data indicate that while YOLOv11 shows slightly better performance under certain specific conditions, RetailDet remains competitive in terms of overall detection performance and lighting adaptability, maintaining relatively stable detection effectiveness across different lighting environments.

According to the experimental data in Table 4, after applying brightness restoration processing, the RetailDet model demonstrates significant performance advantages across different scale configurations. Specifically, at the nano scale, when $\alpha = 0.5$, RetailDet achieves performance levels of mAP = 62.4% and F1 = 81.0%, significantly outperforming YOLOv11 (mAP = 51.9%, F1 = 73.4%) and other comparative models under the same conditions. At the small scale, when $\alpha = 0.8$, RetailDet achieves detection performance of mAP = 65.7% and F1 = 82.3%, similarly surpassing other models' performance. At the medium scale, with $\alpha = 0.8$, RetailDet reaches performance metrics of mAP = 66.2% and F1 = 81.9%, maintaining stable performance advantages. Notably, at the large scale configuration, RetailDet achieves its best performance under $\alpha = 0.5$ conditions, reaching mAP = 67.5% and F1 = 83.1%, demonstrating stronger detection capabilities compared to other advanced models such as YOLOv11 (mAP = 66.1%, F1 = 82.5%). It is worth noting that under different lighting conditions (α ranging from 0.5 to 1.8), RetailDet maintains relatively stable detection accuracy, which fully demonstrates that brightness restoration processing effectively enhances the model's adaptability under different lighting environments, further validating the overall performance advantages of the RetailDet model.

In the retail scene detection of Fig. 11, there are obvious differences in the overall detection integrity and positioning accuracy among various models. RT-DETR has obvious missed detections in scenarios such as (a) the top of the shelf, (b) the dairy area, and (c) bottled products, with detection frames having large deviations from the real targets and heat maps being scattered and blurred. YOLOv8 has fewer missed detections than RT-DETR but still has omissions in areas such as the right side of (b) the dairy area and the gaps between (c) bottled products; moreover, its heat map has the problem of "over-focusing", and for example, the heat in (b) the dairy area is concentrated but does not accurately match the product boundaries, so its positioning accuracy needs to be improved. YOLOv11 performs relatively better: the detection of (a) the middle layer of the shelf is relatively complete, and there are no obvious false detections in (b) and (c) scenarios, though there are edge omissions in small target recognition (such as the trivial elements next to (c) bottled products) and the fitting degree of its heat map with the target distribution is also weaker than that of RetailDet (Ours). RetailDet (Ours) shows significant advantages, with detection frames covering comprehensively, very few missed or false detections, and in particular, the boundaries of (c) bottled products being accurately aligned with the real targets, the heat map clearly focusing on the target area, the positioning error being the smallest, and stronger adaptability in dense and diverse product scenarios.

From the perspective of the rationality of feature focusing and scene adaptability, the differences among various models are further highlighted. The heat map of RT-DETR is scattered without a clear focusing law, and its feature extraction generalization is poor, so it cannot accurately capture the key features of commodities in retail scenes, which directly affects the integrity of detection. Although YOLOv8 has a focusing trend, the heat in Fig. 11 (b) the dairy area covers non-target areas such as the blank edges of the shelf, resulting in redundant focusing. In Fig. 11 (c) the bottled commodity scenario, the filtering of background interference is also insufficient, and the purity and accuracy of feature extraction are low. The heat map of YOLOv11 is relatively concentrated, but in Fig. 11 (c) the bottled commodity scenario, the feature discrimination of the "commodity-background" boundary is weak, and the fitting degree between the focusing range and the actual distribution of targets is slightly inferior, leading to easy omissions in small target recognition. The heat map of RetailDet (Ours) is clearly focused and has a precise range, which can perfectly match the outline of the target commodities. In different retail scenarios such as Fig. 11 (a) the top/middle layer of the shelf, (b) multi-layer dairy products, and (c) dense bottled commodities, it can stably focus on the core area of commodities and filter out background interference. This advantage in feature extraction makes it more adaptable in complex retail environments with dense commodities, diverse displays, and variable light. Through better detection coverage, precise positioning, and reasonable feature focusing, it is significantly superior to other comparative models, providing more reliable technical support for retail business scenarios such as commodity display analysis and shelf inspection.

5) DISCUSSION

a: LIMITATIONS

Although RetailDet demonstrates certain detection capabilities in complex retail environments, it still exhibits several limitations. Analysis of the detection results indicates that the model is prone to missed detections and false positives when faced with extremely dense product arrangements (e.g., stacked large-packaged cooking oil), high visual similarity across different categories (e.g., mixed displays of beverages and personal care products), or significant occlusion (e.g., multi-layer snack shelves with less than half of the product visible). These challenges primarily arise from the difficulty in distinguishing features among dense small objects and interference caused by cross-category similarities. Future research may address these issues from three perspectives: (1) introducing multi-modal attention mechanisms to enhance fine-grained feature extraction and enable precise differentiation of similar products; (2) developing data augmentation strategies tailored for retail scenarios to simulate extreme density and occlusion, thus improving the model's generalization capability; and (3) exploring the integration of semi-supervised and self-supervised learning to leverage unlabeled retail data for optimized feature learning, thereby reducing missed detections and false positives and enhancing adaptability to more complex real-world retail environments.

b: ETHICS

RetailDet aims to enhance rather than replace retail workers, automating repetitive tasks like inventory counting so employees can focus on high-value customer service requiring emotional intelligence. This human-machine collaboration improves service quality while reducing physical labor burden. For technical implementation, RetailDet uses an edge-cloud collaborative architecture where terminal devices capture RGB images, edge nodes optimize transmission, and cloud servers handle depth generation and inference. This design reduces hardware costs and network load by transmitting only RGB data. While inference speed meets real-time requirements, concurrent multi-device operation may cause network congestion, which we address through intelligent scheduling and load balancing at edge nodes. Regarding data processing, we apply strict anonymization to retail images containing customer and employee information. The RPV1K dataset, as part of a corporate project, is currently only available through formal partnership agreements due to commercial licensing restrictions. To ensure research reproducibility, we plan to open-source our model architecture, pre-trained weights, and related code at https://github.com/bilychen88/RetailDet upon paper acceptance, providing researchers with a foundation to reproduce and extend our work.

V. CONCLUSION

In this paper, we propose an innovative solution for commodity detection and vacancy identification in smart retail scenarios. We construct the RPV1K retail scenario dataset and design an efficient two-stream network architecture RetailDet. The effectiveness of the proposed components is verified through comparative experiments and ablation studies. The results show that our RetailDet-large model achieves SOTA performance with a mAP of 67.7% and an F1 of 80.4%, while maintaining an inference latency of 25.1 ms with only 20.79 M parameters. Our method demonstrates substantial improvements over existing SOTA approaches, establishing a robust foundation for automated management in smart retail scenarios. Future work will explore advanced feature fusion strategies and lightweight model architectures, while extending the application scope to diverse retail environments.

IEEE PUBLISHING POLICY

The general IEEE policy requires that authors should only submit original work that has neither appeared elsewhere for publication, nor is under review for another refereed publication. The submitting author must disclose all prior publication(s) and current submissions when submitting a manuscript. Do not publish "preliminary" data or results. To avoid any delays in publication, please be sure to follow these instructions. Final submissions should include source files of your accepted manuscript, high quality graphic files, and a formatted pdf file. If you have any questions regarding the final submission process, please contact the administrative contact for the journal. author is responsible for obtaining agreement of all coauthors and any consent required from employers or sponsors before submitting an article.

The IEEE Access Editorial Office does not publish conference records or proceedings, but can publish articles related to conferences that have undergone rigorous peer review. Minimally, two reviews are required for every article submitted for peer review.

PUBLICATION PRINCIPLES

All authors are aware of and agree to publish.

DISCLOSURE OF INTERESTS

The authors declare no conflict of interest.

AUTHOR'S CONTRIBUTION

All authors read, were informed of, and approved the final manuscript. Contributions are as follows: Bidong Chen: Conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualization, writing original draft, writing review, and editing; Lingui Li: Data curation investigation, data curation, software, methodology, and validation; Yapeng Wang: Investigation, writing review and editing, supervision, funding acquisition, visualization, writing review, and editing; Rui Pedro Paiva: Investigation and supervision; Yuanda Lin: Visualization, review, and editing; Xu Yang: Investigation and formal analysis; and Han Zhu: review and editing.

REFERENCES

[1] L. Chen, S. Lin, X. Lu, D. Cao, H. Wu, C. Guo, C. Liu, and F.-Y. Wang, "Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3234–3246, Jun. 2021, doi: 10.1109/TITS.2020.2993926.

[2] Y. Han, B. Wang, T. Guan, D. Tian, G. Yang, W. Wei, H. Tang, and J. H. Chuah, "Research on road environmental sense method of intelligent vehicle based on tracking check," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 1, pp. 1261–1275, Jan. 2023, doi: 10.1109/TITS.2022.3183893.

[3] R. Tse, L. Monti, M. Im, S. Mirri, G. Pau, and P. Salomoni, "DeepClass: Edge based class occupancy detection aided by deep learning and image cropping," in *Proc. 12th Int. Conf. Digit. Image Process. (ICDIP)*, Jun. 2020, p. 13.

[4] Y. Zhang, W. Zhang, J. Yu, L. He, J. Chen, and Y. He, "Complete and accurate holly fruits counting using YOLOX object detection," *Comput. Electron. Agricult.*, vol. 198, Jul. 2022, Art. no. 107062, doi: 10.1016/j.compag.2022.107062.

[5] F. Bouhlel, H. Mliki, and M. Hammami, "Abnormal crowd density estimation in aerial images based on the deep and handcrafted features fusion," *Expert Syst. Appl.*, vol. 173, Jul. 2021, Art. no. 114656, doi: 10.1016/j.eswa.2021.114656.

[6] F. Achakir, N. Mohtaram, and A. Escartin, "An automated AI-based solution for out-of-stock detection in retail environments," in *Proc. 3rd Int. Conf. Electr., Comput., Commun. Mechatronics Eng. (ICECCME)*, Jul. 2023, pp. 1–6, doi: 10.1109/iceccme57830.2023.10253237.

[7] E. Goldman, R. Herzig, A. Eisenschtat, J. Goldberger, and T. Hassner, ''Precise detection in densely packed scenes,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5222–5231, doi: 10.1109/CVPR.2019.00537.

[8] B. Chen et al., ''RPV11K: A benchmark for joint product-vacancy detection in retail scenarios,'' 2025, doi: 10.21203/rs.3.rs-6428418/v1.

[9] J. Peng, C. Xiao, W. Xun, and Y. Li, ''RP2K: A large-scale retail product dataset for fine-grained image classification,'' *Data Brief*, Apr. 2020, Art. no. 109300, doi: 10.48550/arxiv.2006.12634.

[10] C. K. Reddy, L. Márquez, F. Valero, N. Rao, H. Zaragoza, S. Bandyopadhyay, A. Biswas, A. Xing, and K. Subbian, ''Shopping queries dataset: A large-scale ESCI benchmark for improving product search,'' in *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2022, pp. 3211–3221, doi: 10.48550/arxiv.2206.06588.

[11] M. Klasson, C. Zhang, and H. Kjellström, ''A hierarchical grocery store image dataset with visual and semantic labels,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 491–500, doi: 10.1109/WACV.2019.00058.

[12] X.-S. Wei, Q. Cui, L. Yang, P. Wang, L. Liu, and J. Yang, ''RPC: A large-scale and fine-grained retail product checkout dataset,'' *Sci. China Inf. Sci.*, vol. 65, no. 9, Sep. 2022, Art. no. 197101, doi: 10.1007/s11432-022-3513-y.

[13] F. Chen, H. Zhang, Z. Li, J. Dou, S. Mo, H. Chen, Y. Zhang, U. Ahmed, C. Zhu, and M. Savvides, ''Unitail: Detecting, reading, and matching in retail scene,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 705–722, doi: 10.1007/978-3-031-20071-7_41.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, ''You only look once: Unified, real-time object detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[15] R. Girshick, ''Fast R-CNN,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, ''Swin transformer: Hierarchical vision transformer using shifted windows,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[17] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, ''DETRs beat YOLOs on real-time object detection,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16965–16974, doi: 10.1109/CVPR52733.2024.k01605.

[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, ''SSD: Single shot MultiBox detector,'' in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[19] X. Wang, H. Jiang, M. Mu, and Y. Dong, ''A dynamic collaborative adversarial domain adaptation network for unsupervised rotating machinery fault diagnosis,'' *Rel. Eng. Syst. Saf.*, vol. 255, Mar. 2025, Art. no. 110662.

[20] Y. Cheng, J. Yan, F. Zhang, M. Li, N. Zhou, C. Shi, B. Jin, and W. Zhang, ''Surrogate modeling of pantograph-catenary system interactions,'' *Mech. Syst. Signal Process.*, vol. 224, Feb. 2025, Art. no. 112134.

[21] J. Yan, Y. Cheng, F. Zhang, N. Zhou, H. Wang, B. Jin, M. Wang, and W. Zhang, ''Multimodal imitation learning for arc detection in complex railway environments,'' *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–13, 2025.

[22] J. Tong, ''AOD-Net: A lightweight real-time fruit detection algorithm for agricultural automation,'' *J. Food Meas. Characterization*, vol. 19, no. 4, pp. 1–13, Apr. 2025.

[23] X. Huang, X. Wang, W. Lv, X. Bai, X. Long, K. Deng, Q. Dang, S. Han, Q. Liu, X. Hu, D. Yu, Y. Ma, and O. Yoshie, ''PP-YOLOv2: A practical object detector,'' *Pattern Recognit.*, Feb. 2021, Art. no. 108967, doi: 10.48550/arxiv.2104.10419.

[24] J. Redmon and A. Farhadi, ''YOLOv3: An incremental improvement,'' 2018, *arXiv:1804.02767*.

[25] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, ''Scaled-YOLOv4: Scaling cross stage partial network,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13029–13038, doi: 10.1109/CVPR46437.2021.01283.

[26] G. Jocher, ''Ultralytics YOLOv5 (Version 7.0) [Computer Software],'' 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[27] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, ''YOLOv6: A single-stage object detection framework for industrial applications,'' 2022, *arXiv:2209.02976*.

[28] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, ''YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

[29] G. Jocher, A. Chaurasia, and J. Qiu. (Jan. 2023). *Ultralytics YOLO (Version 8.0.0) [Computer Software]*. [Online]. Available: https://github.com/ultralytics/ultralytics

[30] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, ''YOLOv9: Learning what you want to learn using programmable gradient information,'' in *Proc. ECCV*, 2024, pp. 1–21, doi: 10.1007/978-3-031-72751-1_1.

[31] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, ''YOLOv10: Real-time end-to-end object detection,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 107984–108011.

[32] G. Jocher and J. Qiu. (Apr. 2024). *Ultralytics YOLO11 (Version 11.0.0) [Computer Software]*. [Online]. Available: https://github.com/ultralytics/ultralytics

[33] Y. Tian, Q. Ye, and D. Doermann, ''YOLOv12: Attention-centric real-time object detectors,'' 2025, *arXiv:2502.12524*.

[34] T. Rong, Y. Zhu, H. Cai, and Y. Xiong, ''A solution to product detection in densely packed scenes,'' 2020, *arXiv:2007.11946*.

[35] C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, and K. Chen, ''RTMDet: An empirical study of designing real-time object detectors,'' 2022, *arXiv:2212.07784*.

[36] Q. Liu, J. Lv, and C. Zhang, ''MAE-YOLOv8-based small object detection of green crisp plum in real complex orchard environments,'' *Comput. Electron. Agricult.*, vol. 226, Nov. 2024, Art. no. 109458, doi: 10.1016/j.compag.2024.109458.

[37] J. He, H. Chen, B. Liu, S. Luo, and J. Liu, ''Enhancing YOLO for occluded vehicle detection with grouped orthogonal attention and dense object repulsion,'' *Sci. Rep.*, vol. 14, no. 1, p. 19650, Aug. 2024, doi: 10.1038/s41598-024-70695-x.

[38] D. Papakiriakopoulos, K. Pramatari, and G. Doukidis, ''A decision support system for detecting products missing from the shelf based on heuristic rules,'' *Decis. Support Syst.*, vol. 46, no. 3, pp. 685–694, Feb. 2009.

[39] R. Pietrini, M. Paolanti, A. Mancini, E. Frontoni, and P. Zingaretti, ''Shelf management: A deep learning-based system for shelf visual monitoring,'' *Expert Syst. Appl.*, vol. 255, Dec. 2024, Art. no. 124635, doi: 10.1016/j.eswa.2024.124635.

[40] A. Milella, A. Petitti, R. Marani, G. Cicirelli, and T. D'orazio, ''Towards intelligent retail: Automated on-shelf availability estimation using a depth camera,'' *IEEE Access*, vol. 8, pp. 19353–19363, 2020, doi: 10.1109/ACCESS.2020.2968175.

[41] S. Ma, X. Chang, Y. Zhang, and L. Zheng, ''LoCount: Long-distance crowd counting based on LoRA signal,'' in *Proc. 19th Int. Conf. Mobility, Sens. Netw. (MSN)*, Dec. 2023, pp. 32–39, doi: 10.1109/MSN60784.2023.00019.

[42] X. Cai, Q. Lai, Y. Wang, W. Wang, Z. Sun, and Y. Yao, ''Poly kernel inception network for remote sensing detection,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 27706–27716, doi: 10.1109/CVPR52733.2024.02617.

[43] Q. Chen, Z. Zhang, Y. Lu, K. Fu, and Q. Zhao, ''3-D convolutional neural networks for RGB-D salient object detection and beyond,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 3, pp. 4309–4323, Mar. 2024, doi: 10.1109/TNNLS.2022.3202241.

[44] A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun, ''Depth pro: Sharp monocular metric depth in less than a second,'' 2024, *arXiv:2410.02073*.

[45] F. Chollet, ''Xception: Deep learning with depthwise separable convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.

[46] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, ''Boundary IoU: Improving object-centric image segmentation evaluation,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15329–15337, doi: 10.1109/CVPR46437.2021.01508.

[47] P. Kohli, L. Ladicky, and P. H. S. Torr, ''Robust higher order potentials for enforcing label consistency,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8, doi: 10.1109/CVPR.2008.4587417.

[48] Z. Zhang and M. R. Sabuncu, ''Generalized cross entropy loss for training deep neural networks with noisy labels,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 8792–8802.

[49] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12993–13000.

[50] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, and W. Zuo, "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 8574–8586, Aug. 2022, doi: 10.1109/TCYB.2021.3095305.

[51] Y. Li, S. Li, H. Du, L. Chen, D. Zhang, and Y. Li, "YOLO-ACN: Focusing on small target and occluded object detection," *IEEE Access*, vol. 8, pp. 227288–227303, 2020, doi: 10.1109/ACCESS.2020.3046515.

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[53] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.

[54] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[55] D. J. Bora, A. K. Gupta, and F. A. Khan, "Comparing the performance of LAB and HSV color spaces with respect to color image segmentation," 2015, *arXiv:1506.01472*.

BIDONG CHEN (Member, IEEE) is currently pursuing the Ph.D. degree in applied computer technology with Macao Polytechnic University, Macao, China, and the Ph.D. degree in informatics engineering with the University of Coimbra, Coimbra, Portugal. From 2018 to 2021, he worked on research related to general computer vision platforms and algorithms with the College of Information Science and Electronic Engineering, Zhejiang University, and the Intelligent Research Institute, Midea Group Company. His research interests include deep learning, embedded software and hardware, and natural language processing.

LINGUI LI is currently pursuing the B.E. degree in intelligent science and technology with the School of Modern Information Industry, Guangzhou College of Commerce, Guangzhou, Guangdong, China. His research interests include computer vision, deep learning, and natural language processing, with a focus on object detection and multimodal learning. He has participated in several research projects on intelligent systems and has experience in developing deep learning models for real-world applications.

YAPENG WANG (Member, IEEE) received the B.Eng. degree in telecommunication engineering and the B.Sc. degree in computer and its applications from North China Electric Power University, China, in 1998, and the M.Sc. degree in internet computing and the Ph.D. degree in electronic engineering from Queen Mary University of London, U.K., in 2002 and 2007, respectively. He joined the Faculty of Applied Sciences, Macao Polytechnic University, in 2021, as an Associate Professor. His current research interests include applied artificial intelligence, wireless communications, automatic speech recognition, nature language processing, medical image analysis, and machine learning.

RUI PEDRO PAIVA (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in informatics engineering from the University of Coimbra, Portugal, in 1996, 1999, and 2007, respectively. He is currently a Professor with the Department of Informatics Engineering, University of Coimbra, and a member of the Cognitive and Media Systems and Adaptive Computation Research Groups, CISUC. He has authored 2 books and over 140 papers in peer-reviewed journals and conferences. His research interests include music information retrieval, feature engineering, and machine learning applications in bio and musical signal processing. His algorithms won first place in ISMIR 2004 Audio Description Contest (melody extraction track) and MIREX 2012 Audio Mood Classification task.

YUANDA LIN is currently a Senior Engineer with Whale TV (Singapore), Fujian, focusing on providing intelligent video processing and streaming technologies for global OTT platforms. He leads the research and development of low-latency transmission and machine learning driven content systems. He focuses on industry innovation in the fields of video analytics and edge computing.

XU YANG (Member, IEEE) received the B.Eng. degree in telecommunication engineering from the University of Electronic Science and Technology of China, in 1997, and the M.Sc. degree in e-commerce engineering and the Ph.D. degree in electronic engineering from Queen Mary University of London, U.K., in 2003 and 2009, respectively. She joined the Faculty of Applied Sciences, Macao Polytechnic University, in 2013, as a Lecturer. Her current research interests include wireless communications, medical image analysis, machine learning, and AI applications.

HAN ZHU (Member, IEEE) received the B.S. degree in electronic information engineering from Hangzhou City University, Hangzhou, China, in 2019, and the M.S. degree in communication engineering from Macau University of Science and Technology, Macau, China, in 2021. He is currently pursuing the Ph.D. degree in computer applied technology with Macao Polytechnic University, Macau, and the Ph.D. degree in informatics engineering with the University of Coimbra, Coimbra, Portugal. His research interests include deep learning, security of communications, WSN, index modulation, and ambient backscatter communication.